ITL Bulletin

ADVISING USERS ON INFORMATION TECHNOLOGY

TESTING INTRUSION DETECTION SYSTEMS

Elizabeth B. Lennon, Editor
Information Technology Laboratory
National Institute of Standards and Technology

Introduction

In government and industry, intrusion detection systems (IDSs) are now standard equipment for large networks. IDSs are software or hardware systems that automate the process of monitoring the events occurring in a computer system or network, analyzing them for signs of security problems. Despite the expansion of IDS technology in recent years, the accuracy, performance, and effectiveness of these systems is largely untested, due to the lack of a comprehensive and scientifically rigorous testing methodology. This ITL Bulletin summarizes NISTIR 7007, *An Overview of Issues in Testing Intrusion Detection Systems*, by Peter Mell and Vincent Hu of NIST's Information Technology Laboratory, and Richard Lippmann, Josh Haines, and Marc Zissman of the Massachusetts Institute of Technology Lincoln Laboratory. The Defense Advanced Research Projects Agency (DARPA) sponsored the work.

The lack of quantitative IDS performance measurements can be attributed to some challenging research barriers that must be overcome before the necessary tests can be created. NISTIR 7007 outlines the quantitative measurements that are needed, discusses the obstacles to the development of these measurements, and presents ideas for research in IDS performance measurement methodology to overcome the obstacles. NISTIR 7007 is available online at http://csrc.nist.gov/publications/nistir/index.html.

Who Needs Quantitative Evaluations?

The results of quantitative evaluations of IDS performance and effectiveness would benefit many potential customers. Acquisition managers need this information to improve the process of system selection, which is often based only on the claims of the vendors and limited-scope reviews in trade magazines. Security analysts who review the output of IDSs would like to know the likelihood that alerts will result when particular kinds of attacks are initiated. Finally, R&D program managers need to understand the strengths and weaknesses of currently available systems so that they can effectively focus research efforts on improving systems and measure their progress.

Measurable IDS Characteristics

Listed below is a partial set of measurements that can be made on IDSs. These measurements are quantitative and relate to performance accuracy.

- ***Coverage.*** This measurement determines which attacks an IDS can detect under ideal conditions. For signature-based systems, this would simply consist of counting the number of signatures and mapping them to a standard naming scheme. For non-signature-based systems, one would need to determine which attacks out of the set of all known attacks could be detected by a particular methodology. The number of dimensions that make up each attack makes this measurement difficult. Another problem with assessing the coverage of attacks is determining the importance of different attack types. In addition, most sites are unable to detect failed attacks seeking vulnerabilities that no longer exist on a site.
- ***Probability of False Alarms.*** This measurement determines the rate of false positives produced by an IDS in a given environment during a particular time frame. A false positive or false alarm is an alert caused by normal non-malicious background traffic. Some causes for Network IDS (NIDS) include weak signatures that

Continued on page 2

ITL Bulletins are published by the Information Technology Laboratory (ITL) of the National Institute of Standards and Technology (NIST). Each bulletin presents an in-depth discussion of a single topic of significant interest to the information systems community. **Bulletins are issued on an as-needed basis** and are available from ITL Publications, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8901, Gaithersburg, MD 20899-8901, telephone (301) 975-2832. To be placed on a mailing list to receive future bulletins, send your name, organization, and business address to this office. You will be placed on this mailing list only.

Bulletins issued since February 2002

alert on all traffic to a high-numbered port used by a backdoor; search for the occurrence of a common word such as *help* in the first 100 bytes of SNMP or other TCP connections; or detection of common violations of the TCP protocol. They can also be caused by normal network monitoring and maintenance traffic generated by network management tools. It is difficult to measure false alarms because an IDS may have a different false positive rate in each network environment, and there is no such thing as a *standard* network. Also important to IDS testing is the receiver operating characteristic (ROC) curve, which is an aggregate of the probability of false alarms and the probability of detection measurements. This curve summarizes the relationship between two of the most important IDS characteristics: false positive and detection probability.

- ***Probability of Detection.*** This measurement determines the rate of attacks detected correctly by an IDS in a given environment during a particular time frame. The difficulty in measuring the detection rate is that the success of an IDS is largely dependent upon the set of attacks used during the test. Also, the probability of detection varies with the false positive rate, and an IDS can be configured or tuned to favor either the ability to detect attacks or to minimize false positives. One must be careful to use the same configuration during testing for false positives and hit rates.
- ***Resistance to Attacks Directed at the IDS.*** This measurement demonstrates how resistant an IDS is to an attacker's attempt to disrupt the correct operation of the IDS. One example is sending a large amount of non-attack traffic with volume exceeding the processing capability of the IDS. With too much traffic to process, an IDS may drop packets and be unable to detect attacks. Another example is sending to the IDS non-attack packets that are specially crafted to trigger many signatures within the IDS, thereby overwhelming the human operator of the IDS with false positives or crashing alert processing or display tools.
- ***Ability to Handle High Bandwidth Traffic.*** This measurement demonstrates how well an IDS will function when presented with a large volume of traffic. Most network-based IDSs will begin to drop packets as the traffic volume increases, thereby causing the IDS to miss a percentage of the attacks. At a certain threshold, most IDSs will stop detecting any attacks.
- ***Ability to Correlate Events.*** This measurement demonstrates how well an IDS correlates attack events. These events may be gathered from IDSs, routers, firewalls, application logs, or a wide variety of other devices. One of the primary goals of this correlation is to identify staged penetration attacks. Currently, IDSs have only limited capabilities in this area.
- ***Ability to Detect Never-Before-Seen Attacks.*** This measurement demonstrates how well an IDS can detect attacks that have not occurred before. For commercial systems, it is generally not useful to take this measurement since their signature-based technology can only detect attacks that had occurred previously (with a few exceptions). However, research systems based on anomaly detection or specification-based approaches may be suitable for this type of measurement.
- ***Ability to Identify an Attack.*** This measurement demonstrates how well an IDS can identify the attack that it has detected by labeling each attack with a common name or vulnerability name or by assigning the attack to a category.
- ***Ability to Determine Attack Success.*** This measurement demonstrates if the IDS can determine the success of attacks from remote sites that give the attacker higher-level privileges on the attacked system. In current network environments, many remote privilege-gaining attacks (or probes) fail and do not damage the system attacked. Many IDSs, however, do not distinguish the failed from the successful attacks.
- ***Capacity Verification for NIDS.*** The NIDS demands higher-level protocol awareness than other network devices such as switches and routers; it has the ability of inspection into the deeper level of network packets. Therefore, it is important to measure the ability of a NIDS to capture, process, and perform at the same level of accuracy under a given network load as it does on a quiescent network.
- ***Other Measurements.*** There are other measurements, such as ease of use, ease of maintenance, deployments issues, resource requirements, availability and quality of support, etc. These measurements are not directly related to the IDS performance but may be more significant in many commercial situations.

IDS Testing Efforts to Date

IDS testing efforts vary significantly in their depth, scope, methodology, and focus. Evaluations have increased in complexity over time to include more IDSs and more attack types, such as stealthy and denial of service (DoS) attacks. Only research evaluations have included novel attacks designed specifically for the evaluation and evaluated the performance of anomaly detection systems. Evaluations of commercial systems have included measurements of performance under high-traffic loads. Traffic loads were generated using real high-volume background traffic mirrored from a live network and also with commercial load-testing tools.

Academic, research laboratories, and commercial organizations have all been active in IDS testing efforts. The University of California at Davis and

ITL Bulletins Via E-Mail

We now offer the option of delivering your ITL Bulletins in ASCII format directly to your e-mail address. To subscribe to this service, send an e-mail message from your business e-mail account to listproc@nist.gov with the message **subscribe itl-bulletin**, and your name, e.g., John Doe. For instructions on using listproc, send a message to listproc@nist.gov with the message **HELP**. To have the bulletin sent to an e-mail address other than the From address, contact the ITL editor at 301-975-2832 or elizabeth.lennon@nist.gov.

IBM Zurich developed prototype IDS testing platforms. MIT Lincoln Laboratory performed the most extensive quantitative IDS testing to date, developing an intrusion detection corpus that is used extensively by researchers. The Air Force Research Laboratory focused on testing IDSs in real-time in a more complex hierarchical network environment. The MITRE Corporation investigated the characteristics and capabilities of network-based IDSs. The Neohapsis Laboratories/Network Computing magazine collaboration involved the evaluation of commercial systems. The NSS Group evaluated 15 commercial IDSs and one open-source IDS in 2000 and 2001, and issued a detailed report and analysis. Lastly, Network World Fusion magazine reported a more limited review of five commercial IDSs. See NISTIR 7007 for a complete description of these testing efforts.

IDS Testing Issues

- ***Difficulties in Collecting Attack Scripts and Victim Software.*** The difficulty of collecting attack scripts and victim software hinders progress in developing tests. It is difficult and expensive to collect a large number of attack scripts. While such scripts are widely available on the Internet, it takes time to find relevant scripts to a particular testing environment. Once a script is identified, our experience is that it takes roughly one person-week to review the code, test the exploit, determine where the attack leaves evidence, automate the attack, and integrate it into a testing environment.
- ***Differing Requirements for Testing Signature-Based vs. Anomaly-Based IDSs.*** Although most commercial IDSs are signature-based, many research systems are anomaly-based, and it would be ideal if an IDS testing methodology would work for both of them. This is especially important for comparison of the performance of upcoming research systems to existing commercial ones. However, creating a single test to cover both types of systems presents some problems.
- ***Differing Requirements for Testing Network-Based vs. Host-Based IDSs.*** Testing host-based IDSs presents some difficulties not present when testing network-based IDSs. In particular, network-based IDSs can be tested in an off-line manner by creating a log file containing TCP traffic and then replaying that traffic to IDSs. Since it is difficult to test a host-based IDS in an off-line manner, researchers must explore more difficult real-time testing. Real-time testing presents problems of repeatability and consistency between runs.
- ***Four Approaches to Using Background Traffic in IDS Tests.*** Most IDS testing approaches can be classified in one of four categories with regard to their use of background traffic: testing using no background traffic/logs, testing using real traffic/logs, testing using sanitized traffic/logs, and testing using simulated traffic/logs. While there may be other valid approaches, most researchers find it necessary to choose among these categories when designing their experiments. Furthermore, it is unclear which approach is the most effective for testing IDSs since each has unique advantages and disadvantages.

See NISTIR 7007 for a complete discussion of these issues.

Recommendations for IDS Testing Research

Research recommendations for IDS testing focus on two areas: improving datasets and enhancing metrics.

- ***Shared Datasets.*** There is a great need for IDS testing datasets that can be shared openly between multiple organizations. Few datasets exist that have even semi-realistic data or have the attacks within the background traffic labeled. Without shareable datasets, IDS researchers must either expend enormous resources creating proprietary datasets or use fairly simplistic data for their testing.
- ***Attack Traces.*** Since it is difficult and expensive to collect a large set of attacks scripts for the purposes of IDS testing, a possible alternative is to use attack *traces* instead of real attacks. Attack traces are the log files that are produced when an attack is launched and that specify exactly what happened during the attack. Such traces usually consist of files containing network packets or systems logs that correspond to an instance of an attack. Researchers need a better understanding of the advantages and disadvantages of replaying such traces as a part of an IDS test. In addition, there is a great need to provide the security community with a large set of attack traces. Such information could be easily added to and would greatly augment existing vulnerability databases. The resulting vulnerability/attack trace databases would aid IDS testing researchers and would provide valuable data for IDS developers.
- ***Cleansing Real Data.*** Real data generally cannot be distributed due to privacy and sensitivity issues. Research into methods to remove the confidential data within background traffic while preserving the essential features of the traffic could enable the use of such data within IDS tests. Such an advance would alleviate the need for researchers to expend additional effort creating expensive simulated environments. Another problem with real background data is that it may contain attacks about which nothing is known. It is possible, however, that such attacks could be automatically removed. One idea is to collect a trace of events in the real world and use a simulation system to produce data similar to those in the collected trace.

Who we are

The Information Technology Laboratory (ITL) is a major research component of the National Institute of Standards and Technology (NIST) of the Technology Administration, U.S. Department of Commerce. We develop tests and measurement methods, reference data, proof-of-concept implementations, and technical analyses that help to advance the development and use of new information technology. We seek to overcome barriers to the efficient use of information technology, and to make systems more interoperable, easily usable, scalable, and secure than they are today. Our website is http://www.itl.nist.gov/.

- ***Sensor and Detector Alert Datasets.*** Some intrusion correlation systems do not use a raw data stream (like network or audit data) as input, but instead rely upon alerts and aggregated information reports from IDSs and other sensors. Researchers need to develop systems that can generate realistic alert log files for testing correlation systems. A solution is to deploy real sensors and to *sanitize* the resulting alert stream by replacing IP addresses. Sanitization in general is difficult for network activity traces, but it is relatively easy in this special case since alert streams use well-defined formats and generally contain little sensitive data (the exception being IP addresses and possibly passwords).
- ***Real-Life Performance Metrics.*** Receiver operating characteristic (ROC) curves are created by stepping through alerts emitted by the detector in order of confidence or severity. The goal is to show how many alerts must be analyzed to achieve a certain level of performance and, by applying costs, to determine an optimal point of operation. The confidence or severity-based ROC curve, however, is not a good indicator of how the IDS will perform with an intelligent human administrator sitting at the console. The human administrator does not consider the IDS alerts alone, but makes use of additional information such as network maps, user trouble reports, and learned knowledge of common false alarms when considering which alerts to analyze first. Thus the *alert ordering* used as a basis of the ROC is often not realistic. A further problem is that few current detection systems output a continuous range of scores but instead output only a few priorities (low/medium/high). Thus the ROC consists of only a few very coarse points. It might be useful to use alert type, source, and/or destination IP address along with severity or confidence to order a set of IDS alerts for the purpose of estimating cost and performance of a detector. This new technique could produce a curve that could provide a much more realistic basis for comparing attack detection and false alarm performance, and for estimating the cost of using the intrusion detection product at various levels of performance.
- ***New Technologies.*** Newly evolving IDS technologies include *meta-IDS* technologies that attempt to ease the burden of cross-vendor data management; *IDS appliances* that promise increased processing power and more robust remote management capabilities; and *Application-layer* technologies that filter potential attack traffic to downstream scanner on dedicated network segments. These new directions focus on new technologies for enterprises or service providers and represent examples of research efforts to solve the difficulties of false positives, traffic bottlenecks, and distinguishing serious attacks from nuisance alarms.

Conclusion

While IDS testing efforts to date vary significantly and have become increasingly complex, the lack of a comprehensive and scientifically rigorous testing methodology to quantify IDS performance has hindered the development of needed tests. NIST believes that a periodic, comprehensive evaluation of IDSs could be valuable for acquisition managers, security analysts, and R&D program managers. However, because both normal and attack traffic vary widely from site to site, and because normal and attack traffic evolve over time, these evaluations will likely be complex and expensive. To enable evaluations to be conducted more efficiently, NIST recommends that the community find ways to create, label, share, and update relevant datasets containing normal and attack activity.

Disclaimer:
Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards and Technology
100 Bureau Drive, Stop 8900
Gaithersburg, MD 20899-8900

Official Business
Penalty for Private Use $300

Address Service Requested

PRSRT STD
POSTAGE & FEES PAID
NIST
PERMIT NUMBER G195