Dependable Security: Testing Network Intrusion Detection Systems

Carrie Gates
CA Labs, CA
carrie.gates@ca.com

Carol Taylor
University of Idaho
ctaylor@cs.uidaho.edu

Matt Bishop
University of California Davis
bishop@cs.ucdavis.edu

Abstract

The Network security systems have unique testing requirements. Like other systems, they need to be tested to ensure that they perform as expected, and to specify the conditions under which they might fail. However, un-like other systems, the data required to perform such testing is not easily or publicly available. In this paper we present the requirements for appropriate network traces for testing such systems, along with the challenges of creating public network traces. We make recommendations for tackling these challenges and suggest approaches to developing a public suite of network traces for use by the security community.

1. Introduction

Security is an important component of a dependable system and, like any component, requires appropriate and comprehensive testing. In the case of systems that are intended to provide security, such testing extends not only to the security of the system itself but also to the reliability of its results. That is, how well does the system perform and what are the conditions under which it fails to detect security events?

In this paper we focus on network-based security detectors, such as worm detectors, scan detectors, intrusion detection systems and behavioral analysis systems. Each system proposes to detect particular types of activity given a set of network traffic. The success of these systems in distinguishing security events from benign network traffic is dependent on having training data available that is representative of the network where the sensor is deployed.

Yet, testing the effectiveness of these types of systems with respect to a given network environment is early impossible given the absence of benchmark data sets or testing standards. So, it is not possible to compare the performance, accuracy or efficiency of two systems within a particular type of environment. Gates and Taylor [3] identified the need for a standard set of network traces for testing intrusion detection systems.

Improving the accuracy of these security sensors is critical for increasing the overall dependability of a network since failure to detect security incidents could negatively impact the entire system. Thus, it is our view that the lack of a standardized testing methodology with publicly available data is an emerging problem which should be addressed by the dependable system research community. As a first step to developing standardized testing strategies for network security systems, we argue for the development of a suite of public network traces that can be used for testing security systems.

This section introduced the problem and provided the justification for the worthiness of the problem. In Section 2 we present some background information on existing public data sets and their limitations. Section 3 describes the characteristics required of such data sets in order to be beneficial to the network security community. In Section 4 we present the challenges inherent in developing test sets that meet these requirements. Suggestions on an initial approach to the development of appropriate data sets are provided in Section 5. We conclude in Section 6 with a discussion of the benefits of having such data sets available.

2. Background

In 1998, MIT's Lincoln Labs developed a set of network traces with the goal of comparing different based intrusion detection systems [5]. The approach used was to model traffic from an air force base, generating fake, attack-free, test sets that matched the statistical characteristics of the base. Attacks were performed on an isolated network with the traffic captured and injected into the test sets. Thus sets that contained known attacks were created, and payload was available since actual network traffic was not used (and so there were no privacy concerns). Training sets were made available to the security community, and trained detection systems were submitted for evaluation and comparison. The trained systems were tested using similarly developed sets, but with attacks

that were not necessarily provided in the training set. The data sets from this and subsequent evaluations were made publicly available.

The Lincoln Labs data was the first public data set created, for the purpose of testing Intrusion Detection Systems and thus served a valuable purpose. However, there were serious flaws with the data set, as identified by McHugh [9] and Mahoney and Chan [6]. For example, the proportion of attack traffic to legitimate traffic was not representative of actual network traffic, nor were the synthesized traffic levels representative of actual traffic [9]. Given the synthesized nature of the non-attack traffic, it did not contain the miscellaneous misconfigurations and spurious traffic typically found in internet traffic (see Bellovin's analysis of network traffic [1], Pang et al.'s analysis of "background radiation" [12], and Mahoney and Chan's description of testing their intrusion detection system on network traces [7]). Additionally, the synthesized traffic did not capture all of the characteristics of normal traffic [6]. Despite these limitations, and despite the age of the data (these data sets were created before peer-to-peer traffic, Slammer, and many other worms and viruses), the Lincoln Labs data sets are still commonly used for testing network security systems.

A more recent publicly-available data set consisting solely of attacks was recently developed by Massicotte et al. [8]. This data set was used to test the capabilities of Bro [13] and Snort [15] to detect each attack in the absence of any other confounding traffic. While this is a useful data set, network security detectors, particularly those based on anomaly detection or behavioral analysis, require testing against normal network conditions, including both legitimate traffic and the usual background radiation.

Another data set that is publicly available was produced by LBNL/ISCI and contains anonymized traffic captured from inside an enterprise [10], with the anonymization approach described by Pang et al. [11]. This traffic was captured for two internal subnets and attack traffic filtered at the border is not available. Additionally, only packet headers have been made available, without any corresponding payload.

The CRAWDAD (Community Resource for Archiving Wireless Data At Dartmouth) project at Dartmouth University [16] has also provided traffic traces captured from their wireless network since 2001. This data was collected to specifically analyze wireless data, and therefore does not contain any wired data or border network traces. Additionally, it contains only packet headers and not the full payload.

The PREDICT (Protected Repository for the Defense of Infrastructure Against Cyber Threats) project [4] will provide data sets to security researchers. However, researchers must apply for data access and abide by any restrictions on the data sets to which they are granted access. Access is limited to researchers who are physically located within the United States. PREDICT is currently not available as of 2007.

3. Required Characteristics

Considering the drawbacks of the Lincoln Labs data and other data sets described in the previous section, we derived attributes that should be supported by any publicly available data whose purpose is testing security systems. We identified five characteristics that any suite of network traces requires if it is to be beneficial for the development, testing and comparison of network security systems:

1. Current: Threats are constantly evolving. Fore example, denial-of-service attacks became common in February 2000, Code Red and Nimbda were released in mid-2001 making worms a household word, and phishing started getting noticed in late 2003. Additionally, legitimate network traffic is also constantly evolving. For example, peer-to-peer traffic started becoming common in 2000, and RSS 2.0 feeds in 2003. Thus, any network data used in a testing suite will need to be current in order to reflect traffic and attack trends.
2. Labeled: When a network trace is used by a security system for testing purposes, it is important that the events of interest to the tester are labeled in the trace so that the true and false positive and negative rates can be determined. Having these rates will also allow for the comparison of algorithms that purport to detect the same types of events.
3. Comprehensive: Any suite of network traces will need to represent a variety of traffic and attack types in order to demonstrate applicability across different network topologies and traffic volumes, For example, different sizes of networks (/24, /16, /8) will need to be represented, in addition to different types of networks (e.g., university, corporate, government).
4. Real: It is important that any network traces be gathered from in-use networks, rather than simulated. This is because simulating network data is difficult [14] and prone to generating incorrect artifacts [6]. Additionally, the

correct balance between traffic types (e.g., the volume of legitimate traffic to scans, attacks and background radiation) needs to be maintained as this can potentially affect the performance of the detector [9].

5. Payload: While some systems require only packet headers for analysis, it is not possible to confirm the analysis without using the payload. Additionally, many systems, particularly signature-based systems, require payload access in order to determine if an attack is present.

4. Challenges

Identifying the required characteristics of data sets for testing security systems was based on limitations of existing data sets. However, creating data sets that exhibit these desirable attributes presents certain difficulties which we describe in this section.

1. Current: Any suite of test sets must continually evolve in order to remain current. However, changing the test set often results in problems such as keeping the data labeled and the ability to do historical comparisons between systems in the published literature. We recommend that both historical traces be available along with newer data to ensure that current traffic trends are represented.
2. Labeled: One of the advantages of simulating background traffic and injecting attacks is that the data is then labeled so that both attack and background traffic are well defined. However, when using real network traffic, some other approach is required to distinguish attack from normal traffic.
3. Comprehensive: Given that legitimate traffic and potentially attack traffic vary by network size and organization (e.g., university, government, small business, Google, etc.), network traces will need to be collected from multiple sites in order to meet the other data requirements outlined above which will require their co-operation. Privacy issues and appropriate anonymization to prevent the leaking of any information is a huge challenge. Publishing real network data will require lawyers to ensure that the appropriate safeguards and agreements are in place.
4. Real: While it is possible to acquire test sets from a variety of networks and time periods, how does one confirm that the sets are representative for a given site? That is, how does one determine that all of the attacks of interest can be found in the test set, and that all of the forms of legitimate traffic that might impact a security system are represented?
5. Payload: The use and release of payload information is accompanied by a myriad of privacy concerns. Given that real network traffic is required, it is vitally important that the payload information be anonymized in order to protect the organization, but that the anonymization does not result in changing the overall traffic characteristics. To date, no such anonymization approach has been developed.

5. Approach

We propose the development of a set of testing traces that initially addresses four of the five requirements through a community-based approach. Given that the traffic needs to be real, we would need the co-operation of multiple sites in providing traffic traces. This could start on a smaller scale by collaborating with sites that have already released traffic traces, such as LBNL/ICSI [10] and the CRAWDAD project at Dartmouth College [16]. By demonstrating the value of such a data repository, we anticipate that other organizations will be willing to submit traffic traces. However, inclusion of a complete set of traces will likely require active solicitation of organizations.

The ability to anonymize the data is central to this repository. Contributors may require different levels of anonymization. Research is on-going into anonimization of publicly available data sets [2]. In short, the contributors and the repository must develop compatible threat models to the degree of anonymization desired.

We aim to have a balanced approach to keeping data current. Our initial goal is to generate new data sets at least every two years, but not more often than every year. Old data sets will still be available, as they will have the most information available (to be discussed in more detail in the next paragraph) and so, despite their age, might still provide the most value for testing.

In terms of labeling data, we rely on the data set users. In return for access to the data sets, we will ask that researchers help in label the data. As they use their detectors on the data, their results will show where they thought different types of attacks or anomalies occurred in the data. While any one system will not necessarily find all events of interest, nor have no false

positives, it is hoped that over time the agreement between systems (voting) can be used to determine labels for the data. The older data sets will therefore have the best labels, while newer data sets will still need to develop more in-depth labels and analysis of the data.

We propose to initially ignore the requirement for payload. However, given the privacy concerns, much research still needs to be performed on anonymizing payload before this requirement can be addressed. Even with anonymizing approaches being available, the legal issues involved in providing payload information might be prohibitive.

6. Concluding Comments

The benefits of having a set of standardized network data sets for the testing of intrusion detection and other security systems are many. Primarily, however, a set of traces:

- provides for reproducibility of results, and allows
- independent verification of sensor effectiveness
- allows for proper comparisons between different algorithms so that researchers can determine how their algorithms perform with respect to others
- improves the quality of security components by providing a variety of traces for both testing and development

7. References

[1] S.M. Bellovin. Packets found on an internet. Technical report, AT&T Bell Laboratories, May 1992.

[2] Rick Crawford, Matt Bishop, Bhume Bhumiratana, Lisa Clark, and Karl Levitt. Sanitization models and their limitations. In Proceedings of the New Security Paradigms Workshop, 2006.

[3] Carrie Gates and Carol Taylor. Challenging the anomaly detection paradigm: A provocative discussion. In Proceedings of the 2006 New Security Paradigms Workshop, Schloss-Dagstuhl, Germany, September 2006.

[4] RTI International. Protected repository for the defense of infrastructure against cyber threats (PREDICT). http://www.predict.org/, 2006.

[5] Richard P. Lippmann, Fried D, Graf I, Haines J, Kendall K, McClung D, Weber D, Webster S, Wyschogrod D, Cunningham R, Zissman M. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. Proceedings of the DARPA Information Survivability Conference and Exposition, January 2000.

[6] Matthew V. Mahoney and Philip K. Chan. An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. In Proceedings of the Sixth International Symposium on Recent Advances in Intrusion Detection, Pittsburgh, PA, USA, September 2003.

[7] Matthew V. Mahoney and Philip K. Chan. Learning rules for anomaly detection of hostile network traffic. In Proceed. Third IEEE International Conference on Data Mining, 2003.

[8] Fr´ed´eric Massicotte, Fran¸cois Gagnon, Yvan Labiche, Lionel Briance, and Mathieu Couture. Automatic evaluation of intrusion detection systems. In Proc. of 22nd Annual Computer Security Applications Conf., Miami, FL, 2006.

[9] John McHugh, Alan Christie, and Julia Allen. Defending yourself: the role of intrusion detection systems. IEEE Software, pages 42 – 51, 2000. September/October.

[10] Ruoming Pang, Mark Allman, Mike Bennett, Jason Lee, Vern Paxson, and Brian Tierney. A first look at modern enterprise traffic. In Proceedings of the ACM SIGCOMM/USENIX Internet Measurement Conference, pages 15 – 28, Berkeley, CA, October 2006.

[11] Ruoming Pang, Mark Allman, Vern Paxson, and Jason Lee. The devil and packet trace anonymization. ACM Computer Communications Review, 36(1):29 – 38, January 2006.

[12] Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry Peterson. Characteristics of internet background radiation. In Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement, pages 27 – 40, Taormina, Sicily, Italy, October 2004.

[13] Vern Paxson. Bro: A system for detecting network intruders in real-time. In Proceedings of the 7th USENIX Security Symposium, San Antonio, Texas, January 1998. Usenix Association. [14] Vern Paxson and Sally Floyd. Why we don't know how to simulate the internet. In Proceedings of the 29th conference on Winter simulation, pages 1037–1044, Wisconsin, 1997. ACM Press.

[15] Martin Roesch. Snort — lightweight intrusion detection for networks. In Proceedings of the 13th Systems Administration Conference, pages 229 –238, Seattle, WA, USA, November 1999. Usenix Association.

[16] CRAW-DAD. http://crawdad.cs.dartmouth.edu,2006. Last visited: February 2007.