Usefulness of DARPA Dataset for Intrusion Detection System Evaluation

Ciza Thomas Vishwas Sharma N. Balakrishnan
Indian Institute of Science, Bangalore, India

ABSTRACT

The MIT Lincoln Laboratory IDS evaluation methodology is a practical solution in terms of evaluating the performance of Intrusion Detection Systems, which has contributed tremendously to the research progress in that field. The DARPA IDS evaluation dataset has been criticized and considered by many as a very outdated dataset, unable to accommodate the latest trend in attacks. Then naturally the question arises as to whether the detection systems have improved beyond detecting these old level of attacks. If not, is it worth thinking of this dataset as obsolete? The paper presented here tries to provide supporting facts for the use of the DARPA IDS evaluation dataset. The two commonly used signature-based IDSs, Snort and Cisco IDS, and two anomaly detectors, the PHAD and the ALAD, are made use of for this evaluation purpose and the results support the usefulness of DARPA dataset for IDS evaluation.

Keywords: Intrusion Detection Systems (IDS), DARPA dataset

1. INTRODUCTION

The MIT Lincoln Laboratory under DARPA and AFRL sponsorship, has collected and distributed the first standard corpora for evaluation of computer network Intrusion Detection Systems (IDS). This DARPA evaluation dataset is used for the purpose of training as well as testing the intrusion detectors. These evaluations contributed significantly to the intrusion detection research by providing direction for research efforts and an objective calibration of the technical state-of-the-art. They are of interest to all researchers working on the general problem of workstation and network intrusion detection.[1]

In the DARPA IDS evaluation dataset, all the network traffic including the entire payload of each packet was recorded in tcpdump format and provided for evaluation. In these evaluations, the data was in the form of sniffed network traffic, Solaris BSM audit data, Windows NT audit data (in the case of DARPA 1999) and file system snapshots and tried to identify the intrusions that had been carried out against a test network during the data-collection period. The test network consisted of a mix of real and simulated machines; background traffic was artificially generated by the real and simulated machines while the attacks were carried out against the real machines. Taking the DARPA 1999 dataset for further discussion, the dataset consists of weeks one, two and three of training data and weeks four and five of test data. In training data, the weeks one and three consist of normal traffic and week two consists of labeled attacks.

The classification of the various attacks found in the network traffic is explained in detail in the thesis work of Kendall[2] with respect to DARPA intrusion detection evaluation dataset[1] and is explained here in brief. The attacks fall into five main classes namely, Probe, Denial of Service(DoS), Remote to Local(R2L), User to Remote(U2R) and the Data attacks. The Probe or Scan attacks automatically scan a network of computers or a DNS server to find valid IP addresses (ipsweep, lsdomain, mscan), active ports (portsweep, mscan), host operating system types (queso, mscan) and known vulnerabilities (satan). The DoS attacks are designed to disrupt a host or network service. These include the Solaris operating system crash (selfping), active termination of all TCP connections to a specific host (tcpreset), corruption of ARP cache entries for a victim not in others' caches (arppoison), crash the Microsoft Windows NT web server (crashiis) and crash Windows NT (dosnuke).

In R2L attacks, an attacker who does not have an account on a victim machine gains local access to the machine (guest, dict), exfiltrates files from the machine (ppmacro) or modifies data in transit to the machine (framespoof). New R2L attacks include an NT power point macro attack (ppmacro), a man-in-middle web browser attack (framespoof), an NT trojan-installed remote administration tool (netbus), a Linux trojan SSH server (sshtrojan) and a version of a Linux FTP file access-utility with a bug that allows remote commands to run on a local machine (ncftp). In U2R attacks, a local user on a machine is able to obtain privileges normally reserved for the unix super user or the windows NT administrator. The Data attack is to exfiltrate special files which the security policy specifies should remain on the victim hosts. These include secret attacks, where a user who is allowed to access the special files exfiltrates them (ntfsdos, sqlattack).

With the increase in the network traffic and the introduction of new applications and attacks, continuous improvement is required to make the IDS evaluation dataset a key element to keep it valuable for researchers. The user behavior also shows great unpredictability and changes over time. Modeling the network traffic is an immensely challenging undertaking because of the complexity and intricacy of human behaviors. The DARPA dataset models the synthetic traffic from a session level. Evaluating the proposed IDS with DARPA 1999 dataset may not be representative of the performance with more recent attacks or with other attacks against different types of machines, routers, firewalls or other network infrastructure. All these reasons have caused a lot of criticisms against this IDS evaluation dataset.

In the absence of any other dataset for IDS evaluation it becomes reasonable to analyze the shortcomings and also its importance and strengths for such a critical evaluation. Since this dataset was made publicly available eight years back, the IDSs that were developed after this time were taken for analyzing whether the dataset has become obsolete. The analysis shows that the inability of the IDSs far outweigh the limitations of the dataset. This paper is supposed to give enough support to IDS researchers using the DARPA dataset in their evaluations.

A paper talking in similar lines as the paper presented here is by Brugger and Chow[3] . They have analyzed the DARPA 1998 dataset using Snort and have concluded that any sufficiently advanced IDS should be able to achieve good false positive detection performance on the DARPA IDS evaluation dataset.

This paper is organized as follows. Section 2 discusses the criticisms against the DARPA dataset for IDS evaluation. There are a lot of factors that support the use of DARPA dataset for IDS evaluation. Section 3 discusses those factors and section 4 covers the experimental evaluation results and their discussion. Section 5 concludes the paper.

2. CRITICISMS AGAINST THE DARPA IDS EVALUATION DATASET

The main criticism against the DARPA IDS evaluation data set is that the test bed traffic generation software is not publicly available, and hence it is not possible to determine how accurate the background traffic inserted into the evaluation is. Also the evaluation criteria does not account for system resources used, ease of use, or what type of system it is.[4]

The other popular critiques to the DARPA IDS evaluation data set are by McHugh[5] and by Mahoney and Chan.[6] McHugh[5] presents his criticism on the procedures used in building the dataset and in performing the evaluation. In his critique of DARPA evaluation, McHugh questioned a number of their results, starting from usage of synthetic simulated data for the background and using attacks implemented via scripts and programs collected from a variety of sources. In addition, the background data does not contain the background noise like the packet storms, strange packets, etc. Hence, the models used to generate background traffic were too simple with the DARPA dataset, and if real background traffic was used, the false positive rate would be much higher. Mahoney and Chan[6] comments on the irregularities in the data, like the obvious difference in the TTL value for the attacks as well as the normal packets, which makes even a trivial detector showing appreciable detection rate. They have conducted an evaluation of anomaly-based network IDS with an enhanced version of the DARPA dataset created by injecting benign traffic from a single host.

All the above criticisms have been well researched comments and these works have made it clear that there remain several issues unsolved in design and modeling of the resultant dataset. However the comment made in the thesis work of Pickering,[4] that benchmarking, testing and evaluating with the DARPA dataset is useless unless serious breakthroughs are made in machine learning is not agreed upon by the authors of this paper. The DARPA dataset has the drawback that it was not recorded on a network connected to the Internet. Internet traffic usually contains a fairly large amount of anomalous traffic that is not caused by any malicious behavior.[7] Hence the DARPA dataset being recorded in a network isolated from the Internet might not include these types of anomalies. The unsolved problems clearly remain. However, in the lack of better benchmarks, vast amount of the research is based on the experiments performed on the DARPA dataset. The general thought that even with all the criticisms, the DARPA dataset is still rigorously used by the research community for evaluation of IDSs bring to the fore the motivation for this paper.

3. FACTS IN SUPPORT OF THE DARPA IDS EVALUATION DATASET

A dataset that is seen to be used for IDS evaluation other than the DARPA dataset is the Defcon Capture The Flag (CTF) dataset. Defcon is an yearly hacker competition and convention. However, this dataset has several properties that makes it very different from the real world network traffic. The differences include an extremely high volume of attack traffic, the absence of background traffic, and the availability of a very small number of IP addresses.

The non-availability of any other dataset that includes the complete network traffic was probably the initial reason to make use of the DARPA dataset for IDS evaluation by researchers. Also the experience while trying to work with the real data traffic was not good; the main reason being the lack of the information regarding the status of the traffic. Even with intense analysis the prediction can never be 100 percent accurate because of the stealthiness and sophistication of the attacks and the unpredictability of the non-malicious user. It involves high cost if an attempt is made to properly label the network connections with raw data. Hence most of the research work that used the real network data were not able to report the detection rate or other evaluation metrics for comparison purpose.

Mahoney and Chan[6] comment that if an advanced IDS could not perform well on the DARPA dataset, it could also not perform acceptably on realistic data. Hence before thinking of discarding the DARPA dataset, it is wise to see whether the state-of-the-art IDSs perform well, in the sense that it detects all the attacks of the DARPA dataset. The two commonly used signature-based IDSs were made use of for this evaluation purpose, the Snort version 2.3.4 and the CISCO IDS 4215 and also two anomaly detectors, the PHAD and the ALAD.

McHugh et al.[8] comments that despite its shortcomings, the Lincoln evaluation indicates that even the best of the research IDS systems falls far short of the DARPA goals for detection and false-alarm performance. The Air Force Rome Lab has built a real-time testbed based on the system used by Lincoln to generate its offline evaluation data. They have used this system to evaluate a few of the research systems, with results similar to those obtained in the offline evaluation. All of the evaluations performed to date indicate that IDSs are only moderately successful at identifying known intrusions and quite a bit worse at identifying those that have not been seen before. This renders automatic response to intrusions, a goal of both the research and commercial communities, a dubious prospect.

The poor performance of some of the IDSs evaluated in this work for illustrative purposes were expected. This was mainly due to the general impression that the dataset used was old and hence not appropriate for the current IDS evaluation. Assuming that the dataset is not generalized and hence counting it as a drawback of the dataset, fine tuning of the IDSs to the dataset was considered. Snort has a main configuration file that allows one to add and remove preprocessor requirements as well as the rules files included. This is where the limit of fragmentation to be taken into notice and also whether packet reconstruction is required or not is typically specified. By improving the Snort rule-set, Snort can be customized to perform better in certain situations using the DARPA dataset. Thus we tried manipulating the benchmarking system.

4. RESULTS AND DISCUSSION

4.1 Test Setup

The test setup for the experimental evaluation consisted of three Pentium machines with Linux Operating System. The experiments were conducted with the simulated IDSs Snort, PHAD, and ALAD and also the Cisco IDS4215, distributed across a single subnet observing the same domain. This collection of heterogeneous IDSs was to examine how the different IDSs perform in the presence of DARPA attacks.

4.2 Data Set

MIT-DARPA dataset (IDEVAL 1999)[1] was used to train and test the performance of Intrusion Detection Systems. The network traffic including the entire payload of each packet was recorded in tcpdump format and provided for evaluation. The data for the weeks one and three were used for the training of the anomaly detectors PHAD and ALAD and the weeks four and five were used as the test data. The DARPA 1999 test data consisted of 190 instances of 57 attacks which included 37 Probes, 63 DoS attacks, 53 R2L attacks, 37 U2R/Data attacks with details on attack types given in Table 1.

Table 1. Attacks present in DARPA 1999 dataset

Attack Class	Attack Type
Probe	portsweep, ipsweep, queso, satan, msscan, ntinfoscan, lsdomain, illegal-sniffer
DoS	apache2, smurf, neptune, dosnuke, land, pod, back, teardrop, tcpreset, syslogd, crashiis, arppoison, mailbomb, selfping, processtable, udpstorm, warezclient
R2L	dict, netcat, sendmail, imap, ncftp, xlock, xsnoop, sshtrojan, framespoof, ppmacro, guest, netbus, snmpget, ftpwrite, httptunnel, phf, named
U2R	sechole, xterm, eject, ps, nukepw, secret, perl, yaga, fdformat, ffbconfig, casesen, ntfsdos, ppmacro, loadmodule, sqlattack

4.3 Experimental Evaluation

The IDS Snort was evaluated with the DARPA 1999 dataset and the results are shown in Table 2. It can be noted in Table 2 that some of the attacks for a certain attack type may get detected whereas some from the same attack type may not get detected. This is the reason for the same attack type to appear in both the rows of the table.

The performance of PHAD and ALAD on the same dataset are given in Tables 3 and 4 respectively.

Table 2. Attacks detected by Snort from the DARPA'99 dataset

Attacks detected by snort	teardrop, dosnuke, portsweep, ftpwrite, sechole, yaga, phf, netcat, land, satan, ppmacro, nc-setup, imap, nc-breakin, ncftp, sshtrojan guessftp, tcpreset, secret, selfping, ls, named, dosnuke, crashiis, sqlattack, neptune, xlock, xsnoop, ntinfoscan, httptunnel, udpstorm, loadmodule
Attacks not detected by snort	ps, portsweep, crashiis, sendmail, ftpwrite netcat, nfsdos, sshtrojan, casesen, xterm1 back, guesspop, xsnoop, pod, snmpget, land eject, guesstelnet, syslogd, dict, guessftp, secret, netbus, crashiis, processtable, sqlattack, smurf, httptunnel, loadmod, mailbomb, fdformat, ntfsdos, arppoison, sechole, queso apache2, warez, arppoison ffbconfig, named

Table 3. Attacks detected by PHAD from the DARPA'99 dataset

Attacks detected by PHAD	fdformat, teardrop, dosnuke, portsweep, phf, land, satan, neptune
Attacks not detected by PHAD	loadmodule, anypw, casesen, ffbconfig, ppmacro, eject, ntfsdos, perl, ps, fdformat, sechole, sqlattack, sendmail, nfsdos, sshtrojan, arppoison, xlock, guesspop, xsnoop, snmpget, processtable, guesstelnet, guestftp, netbus, mailbomb, crashiis, secret, smurf, httptunnel, loadmod, land named, warez

The duplication in both the rows as appeared in Table 2 is avoided in the rest of the tables to the maximum extent possible by taking the majority of detection or missing from a certain attack type.

The attacks detected by the Cisco 4215 IDS is given in the Table 5.

4.4 Discussion

The experimental evaluation gave rise to certain questions:

- Since the DARPA attack signatures were known as the most popular dataset for the IDS evaluations at the time of their design and development, why is it that a 100% detection becomes impossible?
- Why is it not possible to have zero false alarms with a signature-based IDS like Snort?
- The anomaly detectors are also inferior in attack detection and high with false alarms even when they were thoroughly trained on the normal dataset. Why is it that none of the learning algorithms that learn from

Table 4. Attacks detected by ALAD from the DARPA'99 dataset

Attacks detected by ALAD	casesen, eject, fdformat, ffbconfig, yaga, phf, ncftp, guessftp, crashiis, ps
Attacks not detected by ALAD	loadmodule, anypw, perl, sqlattack, sendmail, sshtrojan, guesspop, snmpget, xlock, netbus, secret, smurf, httptunnel, sqlattack, sechole, xsnoop, land, mailbomb, sqlattack, ppmacro, warez, named, nfsdos, sechole, xterm, loadmodule, arppoison, processtable, arppoison

Table 5. Attacks detected by Cisco IDS from the DARPA'99 dataset

Attacks detected by Cisco IDS	portsweep, land, crashiis, ppmacro, mailbomb, netbus, sechole, sshtrojan, imap, phf
Attacks not detected by Cisco IDS	dosnuke, ps, ftpwrite, yaga, sendmail, guesspop, xsnoop, snmpget, guesstelnet, secret, smurf, httptunnel, loadmod, ps, arppoison, sqlattack, warez, sqlattack, fdformat, processtable, arppoison, named, satan, nc-setup, nc-breakin, ncftp teardrop, nfsdos, xlock, guessftp

the normal traffic behavior learn successfully when there is no shortage for the normal traffic data from the dataset or even otherwise?

The Snort is designed as a network IDS; extremely good at detecting distributed port scans and also fragmented attacks which hide malicious packets by fragmentation. The pre-processor of snort is highly capable of defragmenting the packets. Matching the alert produced by Snort with the packets in the dataset by means of timestamp might sometimes cause a miss. This is mainly because of the time gap within ±10seconds between the two. However, the Table 2 shows that the DARPA 1999 dataset does in fact model attacks that Snort has trouble detecting or the Snort's signature database is still not updated with those signatures. Isn't it reasonable to think that the attacks for which the signatures are not available with an IDS like Snort, which has a regularly updated rule set, are the ones that still exist undetected? The attackers also are vigilant of the detection trend and hence can't we think that some of the latest attacks are variants of those undetected attacks since those attacks were successful in terms of detection avoidance. Or can't we say that if an IDS is capable of detecting those attacks in addition to the ones detected by the Snort, it is a better performing IDS than Snort? Or is it reasonable to think of changing the test bed when the IDS is suboptimum in performance on that test bed?

In a study made by Sommers et al.,[9] after comparing the two IDSs Snort and Bro, they comment that Snort's drop rates seem to degrade less intensely with volume for the DARPA dataset. They have also concluded in the paper that Snort's signature set has been tuned to detect DARPA attacks. Even then, if we cannot detect all the attacks of this eight year old dataset, it clearly shows the inability of reproducing the signatures of all the available attacks in the dataset of a signature-based IDS. This shows the inability of the IDSs rather than the deficiency of the dataset.

Preprocessing of the DARPA dataset is required before applying to any machine learning algorithm. With the anomaly based IDSs, PHAD and ALAD, we tried to train them by mixing the normal data from an isolated network along with the week 1 and week 3 training dataset. Even then, the algorithms produced less than 50% detection and around 100 false alarms for the entire DARPA test dataset. Again, there are enough reasons to think of the failure on the part of the learning algorithms.

The usual reasoning for the poor performance of the anomaly detectors is that the training and the test data are not correlated; but that happens in real networks as well. The normal user behavior changes so drastically from what the algorithm has been trained with, and hence we expect the machine learning algorithms to be extremely sophisticated and learn the changing behavior. Hence the uncorrelated test bed is good for evaluating the performance of the learning algorithms. Then again, it is the failure on the part of the learning algorithm rather than the dataset if the anomaly detectors are performing poorly. This concludes that the DARPA dataset, even though old, still carries a lot of novelty and sophistication in attacks.

The Cisco IDS is a network-based Intrusion Detection System that uses a signature database to trigger intrusion alarms. As any other network IDS, the Cisco IDS also has only a local view. This feature-gap is pointed out indirectly in a white paper from Cisco systems;[10] "...does not operate properly in an asymmetrically routed environment." This means that Cisco IDS cannot detect attacks over multiple routes.

Thus it can be generalized that the main reasons for the poor performance of the IDSs with the DARPA 1999 IDS evaluation dataset are:

- The training and testing datasets are not correlated for R2L and U2R attacks and hence most of the pattern recognition and machine learning algorithms except for the anomaly detectors that learn only from the normal data, will perform very bad in detecting of the R2L and the U2R attacks.
- The normal traffic in real networks and also in the dataset are not correlated and hence the trainable algorithms are expected to generate a lot of false alarms.
- None of the network based systems did very well against host based U2R attacks.[11]
- The DoS and the R2L attacks have a very low variance and hence difficult to detect with a unique signature for a signature-based IDS or to observe as an anomaly by an anomaly detector.[12]
- Several of the surveillance attacks probe the network and retrieve significant information, and they go undetected, by limiting the speed and scope of probes.[11]
- The dataset provides a large sample of computer attacks embedded in normal background traffic; several realistic intrusion scenarios conducted in the midst of normal background data.
- Many threats and thereby the exploits that are available on the computer systems and networks are undefined and open-ended.

These are some of the reasons for the poor performance of the IDSs and these limitations have to be overcome by sophisticated detection techniques for an improved and acceptable IDS performance.

Also, we have seen that Snort performs exceptionally well in detecting the R2L attacks and the probes, PHAD performs well in detecting the probes and ALAD performs well in detecting the U2R attacks. This shows clearly that each IDS is designed to focus on a limited region of the attack domain rather than the entire attack domain. Hence IDSs, as we know are limited in their performance at the design stage itself.

When we analyze certain IDS alerts, the doubt arises as to whether it is justifiable to say that the IDS detects the particular attack. Take for instance the attacker executes the command : $./exploit. In the real dataset especially for per packet model, it will get translated to many packets, with the first packet containing "$", second packet containing ".", third packet containing "/", fourth packet containing "e" . Is it justifiable to say that the IDS detects the particular attack when the IDS detects the 4th packet as anomalous? It depends on the implementation of the IDS. Some IDS buffer the data, before matching it against the stored patterns. In that case, it is able to see the whole string "$./exploit" and hence detects the anomaly. For an IDS that analyzes on a per packet basis, it is able to find some anomalous pattern in one packet before the connection is terminated and then flags is as an anomalous connection. If the aim is to find intrusive "connections", then any packet corresponding to intrusive connection, detected as malicious, should be good.

5. CONCLUSION

The whole world has a growing interest in network security. The DARPA's sponsorship, the AFRL's evaluation and the MIT Lincoln Laboratory's support in security tools have resulted in a world class IDS evaluation setup that can be considered as a groundbreaking intrusion detection research. In this paper, two of the commonly used signature-based IDSs Snort and Cisco IDS and two anomaly detectors, the PHAD and the ALAD are evaluated using the DARPA 1999 data set. The results were found to support the usefulness of DARPA dataset for IDS evaluation. The DARPA evaluation dataset has been found to have the required potential in modeling the attacks that appear commonly on the network traffic. If a system is evaluated on the DARPA data set, then it cannot claim anything more in terms of its performance on the real network traffic. Hence this dataset can be considered as the base line of any research. The paper is concluded by commenting that it can be used to evaluate the IDSs in the present scenario, even though any effort to make the dataset more "real" and therefore fairer for IDS evaluation is to be welcomed.

REFERENCES

1. DARPA intrusion detection evaluation, http://www.ll.mit.edu/IST/ideval/data/data_index.html
2. K. Kendall, A database of computer attacks for the evaluation of intrusion detection sytsems, Thesis, MIT, 1999
3. S. T. Brugger, J. Chow, An assessment of the DARPA IDS evaluation dataset using Snort, Tech. Report, CSE-2007-1, Nov. 2005
4. K. J. Pickering, Evaluating the viability of intrusion detection system benchmarking, Bachelor Thesis, University of Virginia, US, 2002
5. J. McHugh, Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA IDS evaluations as performed by Lincoln Laboratory, ACM Transactions on Information and System Security, vol.3, No.4, Nov. 2000
6. M. V. Mahoney, P. K. Chan, An analysis of the 1999 DARPA /Lincoln Laboratory evaluation data for network anomaly detection, Technical Report CS-2003-02
7. S. M. Bellovin, Packets found on an Internet, Technical report, AT&T Bell Laboratories, May, 1992
8. J. McHugh, A. Christie, J.Allen, Defending Yourself: The Role of Intrusion Detection Systems, IEEE software, Sep/Oct. 2000
9. J. Sommers, V. Yegneswaran, P. Barford, Toward comprehensive traffic generation for online IDS evaluation, Technical Report, University of Wisconsin
10. SAFE: A security blueprint for enterprise networks, White paper, Cisco Systems, 2000
11. R. Durst, T. Champion, B. Witten, E. Miller, L. Spagnuolo, Testing and evaluating computer intrusion detection systems, Communications of the ACM, vol.42, No.7, Jul. 1999
12. W. Lee, S.J.Stolfo, A Data Mining framework for building intrusion detection models, IEEE Symposium on Security and Privacy, 1999