Thirteenth International Conference
Financial and Actuarial Mathematics – FAM-2020

Databases of anomalous traffic for estimation of intrusion detection

Pavel Stoynov
Technicla University – Sofia,
Faculty of Computer Science and Technology
Sofia, Bulgaria

Abstract: Intrusion Detection Systems (IDS) are IT security systems for detecting hostile activity in IT networks with different use (financial, trade, administration, scientific etc.). When these systems also prevent the hostile activity, they are called Intrusion Detection and Prevention Systems (IDPS). The models for intrusion detection can be divided into two groups:

1. Signature verification. In this case, the system detects previously registered intrusion signature.

2. Anomaly detection. In this case, The system looks for abnormal behavior.

Contemporary IDS usually include both signature verification and anomaly detection realized by rule-based expert system and statistical module correspondingly.

Anomaly detection IDS are mainly in research stage today.

This paper aims at presenting some popular databases of anomalous traffic for estimation of Intrusion Detection Systems (IDS). KDD'99, MetroSec and ADFA-IDS are presented.

Key words: Intrusion Dtection Systems (IDS), statistical evaluation of IDS, KDD'99, MetroSec, and ADFA-IDS

Introduction

Intrusion Detection Systems (IDS) are IT security systems for detecting hostile activity in IT networks with different use (financial, trade, administration, scientific etc.). When these systems also prevent the hostile activity, they are called Intrusion Detection and Prevention Systems (IDPS). They can be divided into two groups:

1. Host-based – for detecting hostile activity on host systems. They analyze information from operation systems and application files captured by software agents.

2. Network based – for detecting hostile activity on networks. They evaluate information from network communication, analyzing the stream of packets captured by sensors.

The models for intrusion detection can be divided into two groups (Gotseva, Trifonov and Stoynov, 2019):

1. Signature verification. In this case, the system detects previously registered intrusion signature.

2. Anomaly detection. In this case, The system looks for abnormal behavior.

Most IDS commercial tools follow the signature verification detection system. However, they have some drawbacks (false alarms, signature of the attack can't be easily discovered, IDS should be periodically updated with new signatures).

Contemporary IDS usually include both signature verification and anomaly detection modules.

For statistically evaluating Intrusion Dtection Systems IDS in network information security, researchers need a documented attack database which database generally represents the ground truth (Rindberg et al., 2008). The evaluation itself counts the number of true positives, true negatives, false positives, and false negatives.

This paper aims at presenting some popular databases of anomalous traffic for estimation of Intrusion Dtection Systems (IDS). The databases KDD'99, MetroSec and ADFA-IDS are presented.

Database KDD'99

Up to now, the most used database is KDD'99 originating from DARPA project on the off-line analysis of intrusion detection systems, at the MIT' Lincoln laboratory MIT (2008). A small network was set-up for this purpose, the aim being to emulate a US Air Force base connected to the Internet. The background traffic was generated by injecting attacks in well defined points of the network, and collected with TCPDUMP. Traffic was grabbed and recorded for 7 weeks, and served for IDS calibration. Once calibration was performed, 2 weeks of traces containing attacks were used to evaluate the performance of IDS under study. This DARPA work in 1998 used a wide range of intrusion attempts, tried to simulate a realistic normal activity, and produced results that could be shared between all researchers interested in such topic. After some changes, mainly dealt with the use of more furtive attacks, the use of target machines running Windows NT, the definition of a security policy for the attacked network, and tests with more recent attacks, this database is used nowadays under the name KDD'99.

McHugh (2001) published a strong criticism on the procedures used when creating the KDD'99 database, especially on the lack of verification of the network realism compared to an actual one. Mahoney et Chan (2003) reviewed into detail the database and showed that the traces were far from simulating realistic conditions, and therefore, that even a very simple IDS can exhibit very high performance results, performances that it could never reach in a real environment.

Despite all the disclaimers about this KDD'99 database for IDS evaluation, it is still massively used. This is mainly due to the lack of other choices and efforts for providing new attack databases.

METROSEC project data

The limitations of the KDD'99 database motivated MetroSec project from the French ACI program on Security & Computer science, 2004-2007 (MetroSec, 2008) to produce controlled and documented traffic traces, with or without anomalies, for testing and validating intrusion detection methods Owezarski (2007).

The experimental platform for creating the MetroSec trace database uses the Renater network, the French network for education and research. Renater is an operational network which is used by a significantly large community in its professional activity. The laboratories involved in the generation of traces are ENS in Lyon, LIP6 in Paris, IUT of Mont-de-Marsan, ESSI in Nice, and LAAS in Toulouse. Traffic is captured at these different locations by workstations equipped with DAG cards Owezarski (2007).

Anomalies studied and generated in the framework of the MetroSec project consist of more or less significant increases of traffic in terms of volume. Two kinds of anomalies are considered: anomalies due to legitimate traffic (like flash crowds- FC) and anomalies due to illegitimate traffic, as flooding attacks, generated by several classical attacking tools.

The anomalies due to illegitimate traffic were DDoS attacks, launched by using flooding attacking tools (IPERF, HPING2, TRIN00 et TFN2K), with changes frequently the attack characteristics and parameters (duration, DoS flow intensity, size and rate of packets) in order to create different profiles for attacks to be detected afterwards Owezarski (2007).

Database ADFA-IDS

Another recent dataset is the Intrusion Dataset (IDS) of Australian Defence Force Academy (ADFA) - ADFA-IDS (Creech and Hu, 2017). It is provided by the University of Arizona Artificial Intelligence Lab and intended as representative of modern attack structure and methodology to replace the older datasets KDD and UNM (Creech and Hu, 2013; Tran et al., 2012).

The version of ADFA-IDS used is released on March 27th, 2017 (Haider et al., 2017) as an update to the original ADFA-IDS made publicly available in 2013 (Creech and Hu, 2013; Creech and Hu, 2014; Haider et al., 2015). ADFA IDS includes independent datasets for Linux and Windows environments.

ADFA-LD (Linux dataset) was generated on a Ubuntu Linux 11.04 host OS with Apache 2.2.17 running PHP 5.3.5. FTP, SSH, MySQL 14.14, and TikiWiki were started (Xie and Hu, 2013; Xie et al., 2014).

Table 1 below shows the payloads and vectors used to attack the Ubuntu OS and generate the dataset.

Table 1. The payloads and vectors used to attack
the Ubuntu OS and generate the dataset

PAYLOAD/EFFECT	VECTOR
password bruteforce	FTP by Hydra
password bruteforce	SSH by Hydra
add new superuser	Client side poisoned executable
Java based meterpreter	Tiki Wiki vulnerability exploit
Linux meterpreter payload	Client side poisoned executable
C100 Webshell	PHP remote file inclusion vulnerability

ADFA-WD (Windows dataset) was genearted on a Windows XP Service Pack 2 host OS with the XP default firewall enabled for all attacks, and file sharing enabled, a network printer configured, wireless and Ethernet networking (Haider et al., 2016a; Haider et al., 2016b). Norton AV 2013 was used to scan certain payloads. FTP server, web server and management tool, and streaming audio digital radio package were activated.

A target ratio of 1:10:1=normal:validation:attack data was used to guide collection and structuring activities.

The vectors used here are: TCP ports, web based vectors, browser attacks, and malware attachments

The effects are: Bind shell, reverse shell, exploitation payload, remote operation, staging, system manipulation, privilege escalation, data exfiltration, and back-door insertion.

References:

Creech, G. and J. Hu (2013) Generation of a new IDS test dataset: Time to retire the KDD collection, 2013 IEEE Wireless Communications and Networking Conference, WCNC 2013, pp.4487-4492.

Creech, G. and J. Hu (2014) A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns, IEEE Transactions on Computers, vol. 63, issue 4, April 2014, pp. 807-819.

Creech, G. and J. Hu (2017) ADFA IDS Dataset, University of Arizona Artificial Intelligence Lab, AZSecure-data, Director Hsinchun Chen.

Gotseva, D., R. Trifonov, P. Stoynov (2019) Neural Networks for Intrusion Detection – Proceedings of the International conference HiTtech-2019, 10-12 October, Sofia, Bulgaria.

Haider, W., J. Hu, and M. Xie (2015) Towards reliable data feature retrieval and decision engine in hostbased anomaly detection systems, Proc. Of the 2015 10th IEEE Conference on Industrial Electronics and Applications, ICIEA 2015, pp. 513-517.

Haider, W., J. Hu, X. Yu, and Y. Xie (2016a) Integer data zero-watermark assisted system calls abstraction and normalization for host based anomaly detection systems, Proc. Of the 2nd IEEE International Conference on Cyber Security and Cloud Computing, 2016, pp. 349-355.

Haider, W., G. Creech, Y. Xie, and J. Hu (2016b) Windows Based Data Sets for Evaluation of Robustness of Host Based Intrusion Detection Systems (IDS) to Zero-Day and Stealth Attacks. Future Internet 8.3 (2016): 29.

Haider, W., J. Hu, S. Slay, B. P. Turnbull and Y. Xie (2017) Generating Realistic Intrusion Detection System Dataset based on Fuzzy Qualitative Modeling," Journal of Network and Computer Applications (JNCA), 2017, DOI: 10.1016/j.jnca.2017.03.018.

Mahoney, M., P. Chan (2003) An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection. In Recent Advances in Intrusion Detection (RAID 20003), pages 220–237, September 2003.

McHugh, J. (2001) 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, 2001.

Metrosec (2008) http://www.laas.fr/METROSEC.

MIT (2008) Lincoln Laboratory.
http://www.ll.mit.edu/mission/communications /ist/corpora/ideval, 2008.

Rindberg, H., M. Roughan and J. Rexford (2008) The need for simulation in evaluating anomaly detectors. Computer Communication Review, 38(1):55–59, January 2008.

Owezarski, P. (2007) Contribution of the French METROSEC project to traffic anomalies detection, Colloque STIC, Paris, France, 5-7 November 2007.

Tran, Q., F. Jiang, and J. Hu (2012) A real-time NetFlow-based intrusion detection system with improved BBNN and high-frequency field programmable gate arrays, Proc. Of the 11th IEEE International Conference on trust, Security and Privacy in Computing and Communications, TurstCom 2012, 2012, pp. 201-208.

Trifonov, R., G. Tsochev, R. Yoshinov, S. Manolov, G. Pavlova (2017) Conceptual model for cyber intelligence network security system. International Journal of Computers. Vol. 11, 2017.

Xie, M. and J. Hu (2013) Evaluating host-based anomaly detection systems: A preliminary analysis of ADFA-LD, Proc. Of the 6th International Congress on Image and Signal Processing, CISP 2013, pp. 1711-1716.

Xie, M., J. Hu, X. Yu, and E. Chang (2014) Evaluating host-based anomaly detection systems: Applications of the frequency-based algorithms to ADFA-LD," LNCS, vol. 8792, 2014, pp.542-549.