Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 7 (2013) 360 – 369

Procedia Technology

The 2013 Iberoamerican Conference on Electronics Engineering and Computer Science

Entropy and flow-based approach for anomalous traffic filtering

Rafael Zempoaltecatl-Piedras[a], Pablo Velarde-Alvarado[b], Deni Torres-Román[a]

[a]*Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV), Unidad Guadalajara, Av. Del bosque 1145, Col. El Bajío, Zapopan, C.P. 45019, Jalisco, México.*
[b]*Universidad Autónoma de Nayarit, Ciudad de la Cultura "Amado Nervo" s/n, C.P. 63155, Tepic, Nayarit, México.*

Abstract

Research Tools in Anomaly-based Intrusion Detection are highly dependent on appropriate traffic trace data. Traditional datasets present several issues such as: removal of sensitive information (anonymization) and insufficient number or volume of attack instances, which limit their quality for the design and evaluation of A-NIDSs. In this paper, we present a method for anomalous traffic filtering which can be used for generating anomaly-free traffic traces. The sanitized dataset can be used to improve the computation of the behaviour profiles during the training stage. The proposal is based on the construction and statistical analysis of the flow-level entropy space for the identification of outliers using three entropy estimators. Empirical results showed that the new traffic traces of the sanitized dataset have a distributional similarity among them greater than that presented among the original datasets.

Keywords: flow-level traffic filtering; entropy; network security; A-NIDS.

1. Introduction

The current trend for connectivity to converged IP networks implies a sustained increase in the amount of information running through this medium. However, it also provides a means for initiating malicious activities as may be attacks on networks. The steady increase in incidents that threaten the security requires innovative methods for detecting anomalous activities that the current Network Intrusion Detection Systems (NIDSs) must be able to identify, [1]. The current Internet connections to high-speed networks involve traffic of gigabits per second which engages intense analysis trying to examine and identify the behavior at packet level. Currently, there is an approach for aggregated information of related packets of network traffic in the form of flows in order to reduce the analysis on the amount of packets, [2]. Hence, the flows supply information and patterns about the traffic instead of packet analysis, [3]. The presented proposal is based on the flow-level analysis. A flow is defined as a stream of packets that are observed at a given interface that share the following five characteristics: source and destination IP address, source and destination port number, and the same protocol.

* Corresponding author. Tel: +52 (33) 3777-3600
E-mail addresses: rafa.piedras@yahoo.com (Rafael Zempoaltecatl-Piedras), pvelarde@uan.edu.mx (Pablo Velarde-Alvarado), dtorres@iteso.mx (Deni Torres-Román).

Selection and peer-review under responsibility of CIIECC 2013
doi:10.1016/j.protcy.2013.04.045

The proposed method, see section 5 and 6, is intended to improve the performance of the A-NIDSs as the training phase relies on better set of datatraces. Then, the requirement to have a reliable training phase is better fulfill when the datatraces are anomaly free. Furthermore, from the point of view of performance evaluation of new methods for A-NIDS purpose is important to have benchmark of traffic traces. However, there is a lack of suitable traffic traces for this purpose. Thus, if the proposed methodology is good enough, it can be extended to generate this type of datasets, [4].

2. Related Work

As the performance of the A-NIDS depends on the training process significantly, the aim of an anomaly-free dataset is to achieve the expected result for training and evaluating. The scarcity of training repositories and research on the field of sanitizing has led to develop innovative techniques regarding the duty of traffic filtering.

The method of *micro-models* in [5] involves small data slices to detect and remove any anomaly not considered within the normal model of the training dataset. The micro-model is an approach considering small time intervals testing the dataset at packet level, however does not consider anomalies spread on several data slices.

The Static sanitization and Dynamic sanitization handle anonymized datasets with an approach of the content and the structured data, [6]. The raw data is transformed into an appropriate structured format. The sanitization takes the structured data, converts it into a representation and applies the sanitization rules similar to signature-based detection.

The Database Partitioning considers a labeling process to obtain a dataset attack-free according to the rulesets getting clean, anomalous and attacks subsets, [7]. The traffic is filtered using an old set of rules and then an up-to-date set of rules for the same tool depending on the database of anomalies but recent and known ones could not be detected.

3. Network Intrusion Detection Systems

NIDSs monitor the behavior of the network traffic in order to raise alerts when possible intrusions or suspicious patterns are observed. NIDSs must be highly reliable and have low false alarm rate, [8]. For A-NIDSs, the training represents the period of greatest need to have a reliable system to detect anomalies. The efficiency of the training and evaluation involves having a repository of highly unfailing captured traffic datasets which must be anomaly-free, reflecting the quality of this step and directly impacts the performance of the A-NIDS.

Currently, we have identified three kinds of sources of datasets that can be used for training and evaluation of NIDSs, they are:

1. **Synthetic dataset**. This type of datasets are generated artificially, the best known are Knowledge Discovery and Data Mining (KDD) Cup 1999 and MIT-DARPA datasets. However, since they were created many years ago, they do not reflect current security threats such as botnets, SQL injection, and worm attacks.
2. **Anonymized dataset**. These kinds of datasets are captured in real networks and are constantly updated. However, for reasons of privacy, certain details of the traces are modified by a process called *anonymization*. This affects the utility of the trace within methods based on *traffic feature distributions*. Examples of these are the Measurement and Analysis on the WIDE Internet (MAWI) repository, [9], the

National Laboratory for Advanced Network Research (NLANR), [10], and The Cooperative Association for Internet Data Analysis (CAIDA), [11].

3. **Pseudo synthetic dataset**. The traces are generated in a controlled environment, where models are defined by the behavior of a real network. However, the issue with this alternative is that it generates specific traces according to the modeled network, [12].

Our proposal involves that a network generates its own datasets. The method is based on the traffic representation by the Method of Entropy Spaces (MES), on which Principal Component Analysis (PCA) is applied and unsupervised techniques for identifying anomalous traffic as well, [13]. Thus, generated anomaly-free dataset provide better results in the training phase of the A-NIDS, which will impact positively on the performance in terms of accuracy and false alarms rate, [14]. Hence, the dataset will be unique and real, keeping reliability of the network.

4. Entropy and Estimators

Randomness, diversity, and uncertainty are present in the behavior of traffic features of all data networks for which the entropy is a useful tool to characterize the dynamics of certain traffic variables in terms of information theory. The Shannon entropy allows a measure of information according to the content of the dataset, [15]. Mathematically, for a discrete random variable (r.v.) X, the Shannon entropy is given by (1) where $p(x_k)$ represents the probability distribution of the r.v. and M is the cardinality of the alphabet of X.

$$H(X) = -\sum_{k=1}^{M} p(x_k) \ln p(x_k) \tag{1}$$

The Naïve approximation of the probability p takes to get an estimate of the entropy in (1) known as Naïve estimator. However this estimator has a high bias presented in (2) where N indicates the number of elements.

$$\hat{H}_{Naïve}(X) = -\sum_{k=1}^{M} \frac{n_k}{N} \ln\left(\frac{n_k}{N}\right) \tag{2}$$

The Balanced estimator subsequently introduced in [16] constitutes a low-bias method of entropy estimating which is defined by (3) below.

$$\hat{H}_{Bal}(X) = \frac{1}{N+M} \sum_{k=1}^{M} \left[(n_k + 1) \sum_{j=n_{k+2}}^{N+2} \frac{1}{j} \right] \tag{3}$$

Last, the Balanced-II estimator presented in [17] improves the computing cost and processing resources of (3) reducing the second summation as part of partial harmonic series as follows in (4).

$$\hat{H}_{Bal-II}(X) = \frac{1}{N+M} \sum_{k=1}^{M} \left[(n_k + 1) \ln\left(\frac{N+2}{n_k+1}\right) \right] \tag{4}$$

The variations in the probability distribution of k elements of a dataset will likewise change the characteristics of the traffic behavior for the same dataset. Hence, entropy estimation plays a very important role in the identification of those characteristics.

An entropy estimation comparison between (2), (3) and (4) can be given in terms of Mean Square Error (MSE) for one particular kind of known discrete probability distribution such as Poisson since it is a well known and common in telecommunications and its entropy function is simple, which is given by (5), where k is the number of occurrences and λ is the Poisson parameter.

$$H_{Poisson}(X) = \lambda[1-\log(\lambda)] + e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k \log(k!)}{k!} \quad (5)$$

Thus, we can compare the performance of the estimators in terms of MSE. Figure 1 shows the comparison for the case of a Poisson distribution with $\lambda = 1$, $k = 5$, $M = 5$ and sample size N, from 10 to 100. Under these conditions, the estimator Balanced-II results on lower MSE when the length $N < 20$. For larger *N* the MSE result converges for the three estimators.

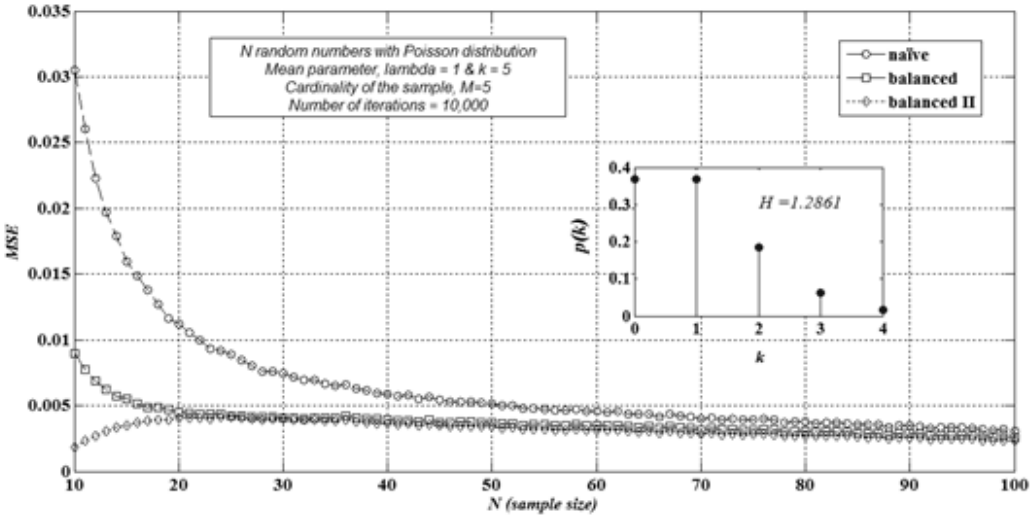

Figure 1. Poisson entropy estimation comparison

5. Entropy Spaces

MES is a proposed method to perform a graphical abstraction of a network traffic trace, [18]. This abstraction is intended to generate a three dimensional representation comprised of data-points. Data-points patterns reflect the behavior of the trace.

The traffic representation by the entropy spaces considers that a captured traffic χ could be split onto not overlapped m slots of maximum duration t_d seconds. On a particular *i*-slot, K_i flows are generated. Four traffic features r are taken from the flow fields: r_1 source IP address (srcIP), r_2 destination IP address (dstIP), r_3 source port number (srcPrt) and finally r_4 destination port number (dstPrt). For each i-slot subsequently the flows are clustered according the defined r feature. In this way, four cluster keys can be defined. Particularly, for a cluster key r_1, the clusters are formed of flows that containing the same source IP address and variability on the remaining flows features, i.e. r_2, r_3 and r_4 which are denoted as free dimensions. The number of forming clusters depends on the cardinality or alphabet size $|\mathbb{A}_i^r|$ of registered source IP addresses into the i-slot.

The entropy space, used in this work, is a three-dimensional representation of the captured traffic χ flows. This space involves data-points representing the entropy of traffic flows clusters. The entropy estimation of the three free dimensions on a cluster k for r_1 is represented by the vector as follows in (6):

$$\left(\hat{H}_{r_3}, \hat{H}_{r_4}, \hat{H}_{r_2}\right)_k \quad (6)$$

For the entropy space, (6) is considered as a data-point. The data-points of the *i*-slot yield to $k = 1, \ldots, |\mathbb{A}_i^r|$ generating points cloud data. In the same manner for the m slots of the trace χ will complete the space.

Our technique involves analyzing the behavior of the data-points in order to identify those furthest from the group that begin to form outliers. Later they will be evaluated in forensic analysis at the packet level to observe the behavior and corroborate the anomalous activity that generated the displacement of the data-points in the entropy space. After the identification of the data traffic related to anomalous data points, proceed to the traffic filtering.

The method of PCA allows transforming a set of correlated variables with each other into a set of uncorrelated variables called principal components. The principal components result in a system of which the maximum variance of the data occurs on the first axis (PCA 1), the second largest variance on the second axis (PCA 2) and so on, [19].

By applying PCA allows us to obtain scores that get an amplitude value and drive to dictate an outlier based on the highlighted value and therefore allow a dataset to identify outliers and likewise will maintain a threshold based on the registered values obtained from the repository.

6. Experimental Results

6.1 Scenario

The scenario of analysis is the campus network of the Universidad Autónoma de Nayarit (UAN), which consists of approximately 7,000 hosts, including wired and wireless local networks. The access to Internet through high-speed E3 connection as shown in Figure 2. The inside network is divided by VLANs according to campus distribution, facilities and offices.

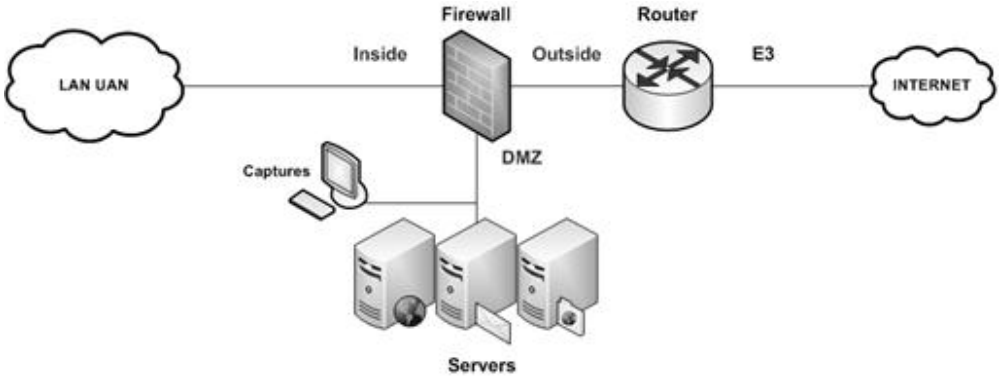

Figure 2. UAN Network general diagram

6.2 Repository

The trace repository for analysis consists of datasets obtained in the scenario above described regarding to November 07, 08, 09, 10, 11 and 18 of 2011 shown in table 1.

Table 1. Traffic traces corresponding to dataset repository UAN

Trace (YY/MM/DD)	Time period	Size (GB)	Packets
UAN-111107	07:00:00-15:00:00	2.7	5,110,923
UAN-111108	07:00:00-15:00:00	2.5	4,047,738
UAN-111109	07:00:00-15:00:00	2.6	4,216,903

UAN-111110	07:00:00-15:00:00	2.7	5,152,525
UAN-111111	07:00:00-15:00:00	2.4	4,718,816
UAN-111118	07:00:00-15:00:00	2.0	3,493,639

6.3 Graphic Results

The results were obtained through the implementation of scripts running Matlab and Perl as well as free linux tools such as Wireshark and tcpdump. Thus, on the entropy spaces focusing on cluster key of source IP address the outliers are showed as the most distant points of the cluster, identified via Wireshark. Figure 3 shows the entropy space analysis for the raw trace UAN-111108 using the Naïve, Balanced and Balanced-II estimators respectively.

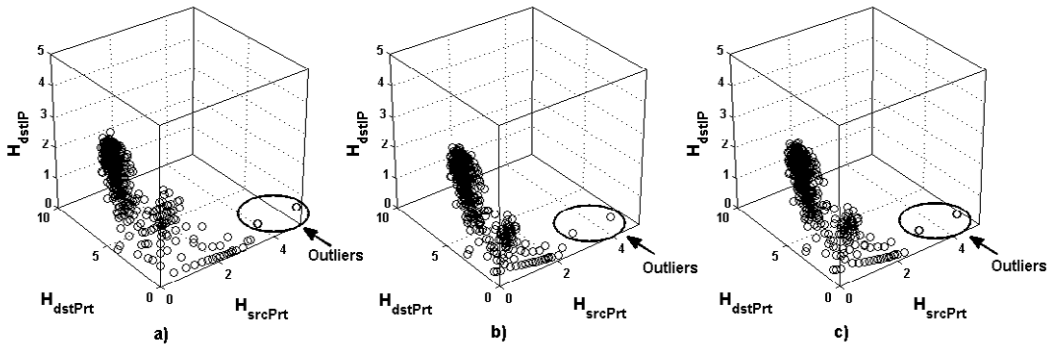

Figure 3. Entropy space for raw trace UAN-111108 Cluster Key srcIP.
Entropy Estimator a)Naïve, b)Balanced and c)Balanced-II

The comparison of the entropy spaces for the Naïve estimator regarding the Balanced and Balanced-II estimators, shows that the second one presents a greater concentration of data points than the other ones. This dispersion of Naïve estimator does not allow further highlight outliers in comparison with Balanced and Balanced-II. Through filtering and sanitation methods the outliers, which correspond to anomaly behavior as distant parts containing SQL injection attacks, present in the trace UAN-111107 are discarded for the entropy space shown in Figure 4. This comparison shows again that Balanced and Balanced-II estimators have the highest concentration of data points.

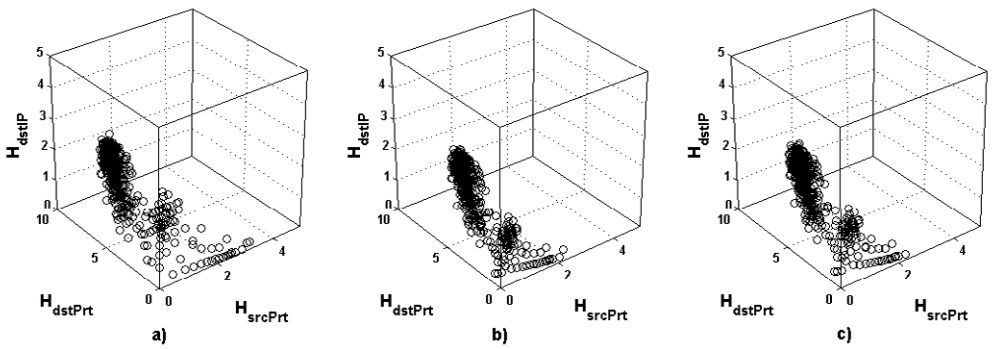

Figure 4. Entropy space sanitized trace UAN-111108 Cluster Key srcIP.
Entropy Estimator a)Naïve, b)Balanced and c)Balanced-II

The PCA analysis for the entropy space of the trace UAN-111108 is shown in figure 5. As the graphs in the 3D space, there are plots for Naïve, Balanced and Balanced-II estimators left respectively.

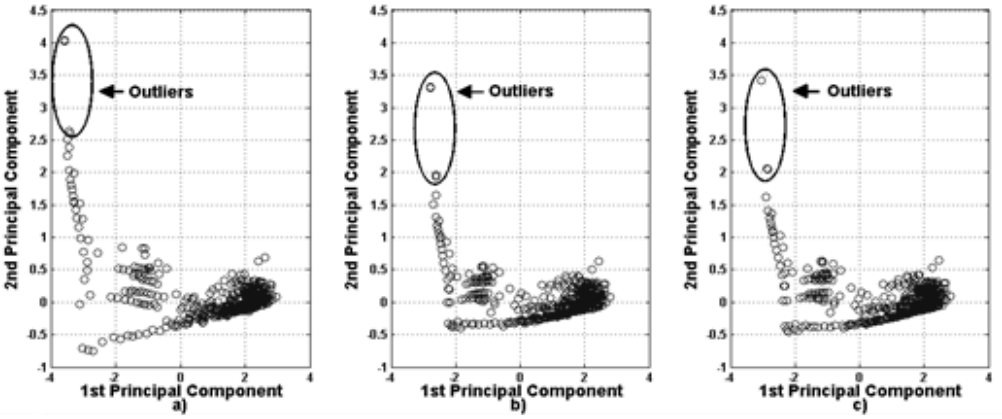

Figure 5. PCA for raw trace UAN-111108. Entropy Estimator a)Naïve, b)Balanced and c)Balanced-II

The figure 6 shows the PCA analysis for sanitized trace UAN-111108. By filtering the outliers is very important to highlight the spaces amplitudes derived from the estimators Balanced and Balanced-II. Unlike what obtained by Naïve the width reductions are mostly notable.

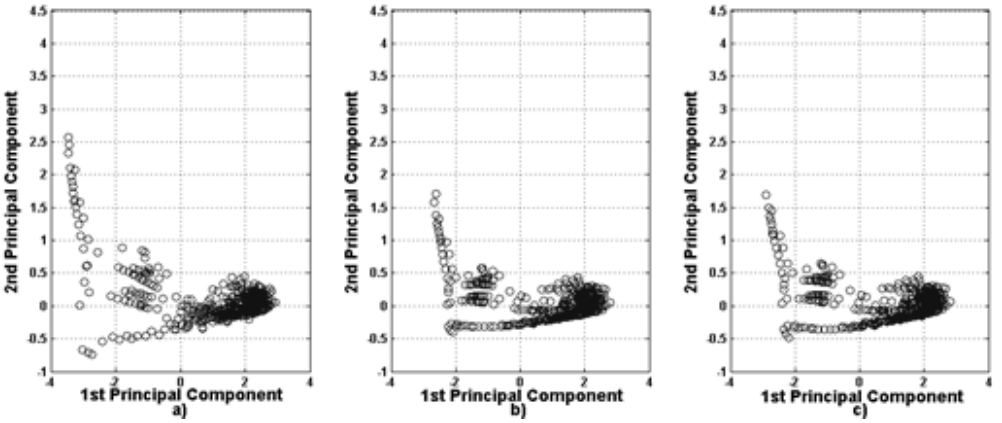

Figure 6. PCA for sanitized trace UAN-111108. Entropy Estimator a)Naïve, b)Balanced and c)Balanced-II

The table 2 contains a summary of the threshold values for the amplitudes for the traces repository (raw traffic traces) used for this analysis. An additional graphical diagnostic may be conducted using the Q-Q plots. It allows confirming the sanitization procedure by filtering out the traffic anomalies of the traces generating a new set of data-points regarding to similar statistical properties.

Table 2. Traffic traces PCA amplitude thresholds.

Trace (YY/MM/DD)	Naïve	Balanced	Balanced-II
UAN-111107	2.9	1.1	1.1
UAN-111108	2.6	1.7	1.7
UAN-111109	2.4	1.6	1.6
UAN-111110	1.8	1.2	1.3
UAN-111111	2.1	1.2	1.2
UAN-111118	2.3	1.6	1.7

The figure 7 shows a Q-Q plot regarding the raw trace UAN-111107 with the raw one UAN-111110 containing both outliers. Similarly, the plots contain the results of the three estimators in the same order.

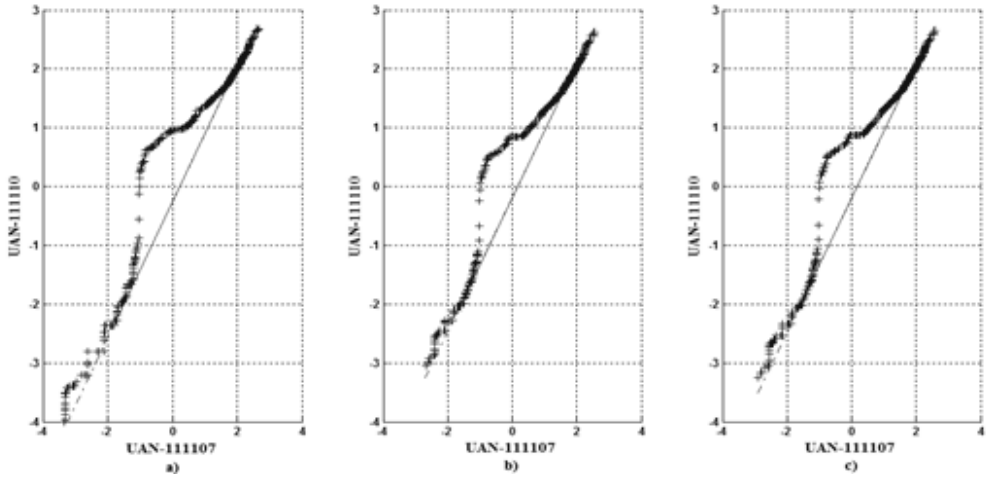

Figure 7. Q-Q Plot for raw trace UAN-111107 1st Principal Component with raw trace UAN-111110 1st Principal Component. Entropy Estimator a)Naïve, b)Balanced and c)Balanced-II

Finally the figure 8 displays the Q-Q plot for the sanitized trace UAN-111107 and the sanitized one UAN-111110. The straight line represents that the two distributions of the two traces, which are free of anomalies, are equal.

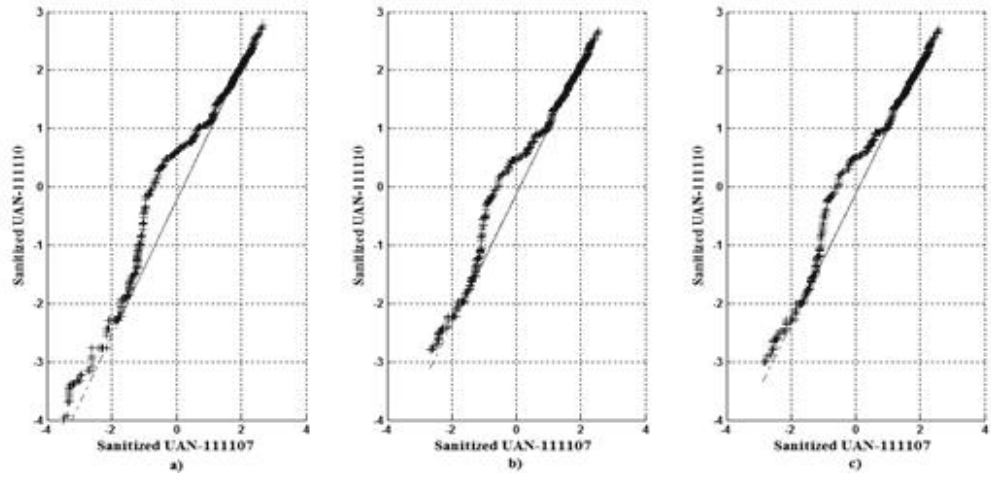

Figure 8. Q-Q Plot for sanitized trace UAN-111107 1st Principal Component with sanitized trace UAN-111110 1st Principal Component. Entropy Estimator a)Naïve, b)Balanced and c)Balanced-II

7. Conclusions

Through the related outliers, the Method of Entropy Spaces (MES) has been useful for graphically detecting anomalous traffic, mainly for the entropy estimators Balanced and Balanced-II. The MES was evaluated in a real scenario regarding to a campus network with real traffic. The method for anomalous traffic filtering to create training datasets for A-NIDSs is of great importance in order to achieve good performance in detecting anomalies and avoiding false events. The obtained results were tested with the Q-Q plot method in order to compare the traces

with anomalies versus the sanitized ones.

Moreover, a more simple 2D analysis is provided by the PCA analysis of the entropy spaces. Although, a criterion for the threshold needs further studies, an average amplitude value could be set as a limit for anomalies regarding to the used entropy estimator as well as the analyzed traces of the particular network. Given the scarcity of suitable and publicly available datasets for training/evaluation, an alternative for the generation of experimental datasets from a production network can significantly contribute to the intrusion detection research field

8. Acknowledgements

Pablo Velarde-Alvarado and Deni Torres-Roman are grateful with CONACyT for its support through its “Ciencia Básica CB-2011” Research funding for Project number 167859. Rafael Zempoaltecatl-Piedras is very grateful with CONACyT for supporting him to pursuit his M.S. degree at CINVESTAV, Guadalajara Campus.

9. References

[1] RSA 2012 Cybercrime Trends Report. http://www.rsa.com
[2] Stephen N., and Judy N., Network Intrusion Detection, New Riders, 2003.
[3] Alaidaros H, Mahmuddin M, Mazari A. An Overview of Flow-Based and Packet-Based Intrusion Detection Performance in High Speed Networks.
[4] Velarde-Alvarado, P., Vargas-Rosales, C., Toral-Cruz, H. and, Ramirez-Pacheco, J. An Information-theoretic Approach to Traffic Traces Sanitization for Intrusion Detection Purposes.
[5] G. F. Cretu, A. Stavrou, M. E. Locasto, S. J. Stolfo, and A. D. Keromytis. Casting out Demons: Sanitizing Training Data for Anomaly Sensors. 2008. IEEE Computer Society.
[6] M. Bishop, B. Bhumirbh, R. Crawford, K. Levitt. How to Sanitize Data. 2004. IEEE Computer Society.
[7] M. Bermúdez, R. Salazar, J. Díaz, and P. García. Proposals on Assessment Environments for Anomaly-Based Network Intrusion Detection Systems. CRITIS'06 Proceedings of the First international conference on Critical Information Infrastructures Security.
[8] Denning, D. An Intrusion Detection Model. Proceedings of the seventh IEEE Symposium on Security and Privacy, May 1986, pp. 119-131.
[9] MAWI Working Group Traffic Archive. http://mawi.wide.ad.jp/mawi/
[10] The National Laboratory for Applied Network Research (NLANR) Project. http://www.nlanr.net/
[11] CAIDA Data - Overview of Datasets, Monitors, and Reports. http://www.caida.org/data/overview/
[12] Shiravi, A., Shiravi, H., Tavallaee, M. and, Ghorbani, A. A. “Toward developing a systematic approach to generate benchmark datasets for intrusion detection” Elsevier Computers & Security, vol. 31, 357-374 (2012).
[13] Velarde-Alvarado, P., Vargas-Rosales, C., Toral-Cruz, H. and, Ramirez-Pacheco, J. “Characterizing Flow Level Traffic Behavior with Entropy Spaces for Anomaly Detection” submitted to "Building Next-Generation Converged Networks: Theory and Practice" to be published by CRC Press, USA 2012.
[14] Bermudez-Edo, M., Salazar-Hernandez, R., Diaz-Verdejo, J. and, Garcia-Teodoro, P. “Proposals on Assessment Environments for Anomaly-based Network Intrusion Detection”, Lect. Notes in Computer Science Springer-Verlag, vol. 4347, 210 – 221 (2006).
[15] C. Shannon. A mathematical theory of communication. Bell System Technical Journal, 27. pp. 379-423 and 623-656. (1948).
[16] J. Bonachela, H. Hinrichsen, M. Muñoz. Entropy Estimates of Small Data sets. Journal of Physics A. Mathematical and Theorical No.41. April 2008.
[17] P. Velarde-Alvarado, C. Vargas-Rosales, D. Torres-Román, A. Martinez-Herrera. Detecting Anomalies in Network Traffic Using the Method of Remaining Elements. IEEE Communicatiosn Letters, vol. 13 no. 6 2009.

[18] Velarde-Alvarado, P., Vargas-Rosales, C., Toral-Cruz, H. and, Ramirez-Pacheco, J. (in press) “Characterizing Flow Level Traffic Behavior with Entropy Spaces for Anomaly Detection” submitted to "Building Next-Generation Converged Networks: Theory and Practice" to be published by CRC Press, USA (2012).
[19] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Jiang Zhang. Face Recognition Using Laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 3. pp. 328-340. 2005.