

Information Theory and Data-Mining Techniques for Network Traffic Profiling for Intrusion Detection

Pablo Velarde-Alvarado¹, Rafael Martinez-Pelaez¹, Joel Ruiz-Ibarra¹,
Victor Morales-Rocha²

¹Autonomous University of Nayarit, Area of Basic Sciences and Engineering, Ciudad de la Cultura "Amado Nervo", Tepic, Nayarit, Mexico

²Department of Electrical and Computer Engineering, Autonomous University of Ciudad Juarez, Ciudad Juarez, Mexico

Email: pvelarde@uan.edu.mx, rafael_uvm@hotmail.com, joelruizibarra@gmail.com, victor.morales@uacj.mx

Received June 2014

Abstract

In this paper, information theory and data mining techniques to extract knowledge of network traffic behavior for packet-level and flow-level are proposed, which can be applied for traffic profiling in intrusion detection systems. The empirical analysis of our profiles through the rate of remaining features at the packet-level, as well as the three-dimensional spaces of entropy at the flow-level, provide a fast detection of intrusions caused by port scanning and worm attacks.

Keywords

Intrusion Detection, Traffic Profiling, Entropy, and Network Worms

1. Introduction

Network Intrusion Detection Systems [1], or NIDSs, have become an important component to detecting attacks against information systems. However, they offer only a limited defense. For instance, a signature-based NIDS (S-NIDS) monitors packets on the network and compares them against a database of signatures or attributes from known malicious threats. A weakness of this type of NIDS is that there will always be a lag between a new threat being discovered and the signature for detecting that threat being applied to the NIDS. During that lag time the NIDS would be unable to detect the new threat.

A second type of NIDS is the anomaly-based NIDS (A-NIDS), which monitors network traffic and compares it against an established baseline, [2]. The baseline helps to identify what is normal behavior for that network. If a deviation from the established baseline reaches a specified threshold, an alarm is generated. Therefore, anomaly detection techniques have the potential to detect new and unforeseen types of attacks. Traditional anomaly based IDSs, employ algorithms that focus primarily on changes in the traffic volume at specific points on the network, and promptly alert the operator of a sudden increase. However, such systems can be evaded through sophisticated attacks that focus on compromising significant hosts, causing them a collapse of memory or CPU

and maintaining a level of traffic within the normal threshold.

Recently, a new generation of anomaly based IDSs have emerged, which focus on gaining knowledge in the structure and composition of the traffic and not just its volume. Such systems are based on the fact that the malicious activities affect the natural randomness of the network, e.g., they change significantly the entropy of the network [4] [5]. The composition of traffic is related to its probability distribution, and can be characterized by its entropy; a malicious activity changes that composition and the shape of the distribution and therefore its entropy. By means of entropy measures to a set of traffic features, we can establish the profiles of normal activity of the network and determine intrusions to the system.

This paper presents an analysis at the packet and the flow level on traces obtained through measurements conducted in a campus network under real attacks of the Blaster [6] and Sasser [7] worms, as well as a port scan attack to the proxy server of that network. The captured traces during a week of normal operation, helped to develop a profile of normal behavior that is useful to be compared to attack conditions.

The paper is organized as follows. In Section 2, we present our profiling approach and the context of this paper. Section 3 describes the test environment; Sections 4 and 5 explain the methodology: the rate of remnant items and spaces of entropy and results. Section 6 gives concluding remarks.

2. Profiling Approach

We propose two methods for the creation of profiles based on entropy. The analysis applies primarily to the packet-level for the method of the rate of remnant elements and to the flow-level for the spaces of entropy. Initially, there is a set of captured traffic traces corresponding to five days in typical working hours in an academic LAN. The traces have been inspected to be considered free of anomalies, so they may serve as a baseline.

We use traffic features to build the profiles. A traffic feature is a field in a header of a packet (at the packet level) or a field in a five-tuple (at the flow level), respectively. Four fields will be used: source address (*srcIP*), destination address (*dstIP*), source port (*srcPrt*), and destination port (*dstPrt*).

After the feature extraction, an essential part in the builder profile block is the measurement of entropy. For a discrete set of symbols $\{a_1, a_2, a_3, \dots, a_n\}$ with probabilities p_i , $i = 1, 2, \dots, n$, the entropy of the discrete distribution of a random variable X associated, is a measure of randomness in the set of symbols and represented as

$$H(X) = \sum_{i=1}^n p_i \log_2 p_i, \quad 0 \leq H(X) \leq H_{MAX} = \log_2 n \quad (1)$$

The relative uncertainty (RU) provides a measure of variety or uniformity that is independent of the sample size. For a random variable X RU is defined as, [3],

$$RU = \frac{H(X)}{H_{MAX}}, \quad 0 \leq RU(X) \leq 1 \quad (2)$$

$RU(X) \approx 1$ means that observed values of X are closer to being uniformly distributed, thus less distinguishable from each other, whereas $RU(X) \approx 0$ indicates that the distribution is highly concentrated.

3. Experimental Platform

The worm propagation and port scanning were carried out on academic LAN which is subdivided into four subnets (192.168.1.0, 192.168.2.0, 192.168.4.0, and 10.253.253.0). There are 100 hosts running Windows XP SP2 mainly. One router (192.168.1.1) connects the subnets with 10 Ethernet switches and 18 IEEE 802.11b/g wireless access points. The data rate of the core network is 100Mbps. A sector of the network is left vulnerable for worm propagation, with ten no Windows XP service pack installed at all (192.168.1.104 - 113). In the experiments Blaster and Sasser worms where released in the vulnerable sector. The scanning port attack was observed on the proxy server (192.168.4.253).

Data-Set and Tools

The benign traffic traces in typical work hours for a period of five days were labeled with a number from one to five. The anomalous traffic for port scanning attack was labeled as 6-P1. Blaster and Sasser worm attacks were labeled as 6-P2 and 6-P4, respectively. The data-set was collected by a network sniffer tool based on libpcap

library used by tcpdump, [8]. All traces were cleaned to remove spurious data using plab, a platform for packet capture and analysis, [9]. Traces were split into segments using tracesplit which is a tool that belongs to Libtrace, [10]. The traffic-files in ASCII format suitable for MATLAB processing were created with ipsumdump, [11]. The flow generation was done with flow analyzer, a tool based on perl and developed by us.

4. Rate of Remnant Elements

We define a traffic trace χ of a duration t_D seconds with a total of N packets, χ is divided into M non-overlapping slots of $t_d = t_D/M$ seconds each one. The i -th slot has W_i packets for $i = 1, 2, \dots, M$. In each i -slot, four features are extracted that we associate with a value of r , namely $r = 1$ for source IP address, $r = 2$ for destination IP address, $r = 3$ for source TCP port, and $r = 4$ for destination TCP port. Let S be a finite sequence of $r = 1$ values or IP source addresses in a slot- i . This sequence with elements in an alphabet set A , is a function from $\{1, 2, 3, \dots, |A|\}$ to A for some $|A| \geq 0$. The generated sequence S is denoted by $(a_1, a_2, a_3, \dots, a_{W_i})$, and the length of S is W_i . The elements of S belong to an alphabet A with cardinality $n = |A|$. From A an ordered set $A^{(o)} = \{a_1^{(o)}, a_2^{(o)}, \dots, a_n^{(o)}\}$ is created, $A^{(o)}$ contains the n -source IP addresses in decreasing order sorted by frequency. With the associated frequencies of A , we define a probability mass function (pmf)

$$\Pr(X_i^{r=1}, j) = p_j(a_j^{(o)}) = \begin{cases} f_j & 1 \leq j \leq n \\ 0 & \text{rest} \end{cases} \quad (3)$$

where $f_1 \geq f_2 \geq f_3 \geq \dots \geq f_n$. Ordered set $A^{(o)}$ is transferred to an iterative process P to create l subsets of $A^{(o)}$ denoted as $A^{(o,k)}$, $1 \leq k \leq l$. This family of l subsets is shown in Equations (4)-(6) and holds $A^{(o,k)}/A^{(o,k+1)} = \{a_k^{(o)}\}$

$$A^{(o)} = A^{(o,1)} = \{a_1^{(o)}, a_2^{(o)}, a_3^{(o)}, \dots, a_n^{(o)}\} \quad (4)$$

$$A^{(o,1)} = \{a_2^{(o)}, a_3^{(o)}, a_4^{(o)}, \dots, a_n^{(o)}\} \quad (5)$$

$$A^{(o,l)} = \{a_l^{(o)}, a_{l+1}^{(o)}, a_{l+2}^{(o)}, \dots, a_n^{(o)}\} \quad (6)$$

When in a k -iteration, the relative uncertainty of a partial pmf reaches a threshold β , i.e., $RU(X_i^r, k) > \beta$, we say that the iterative process P reached its latest iteration, and hence, $k = l$. An estimator of relative uncertainty for a discrete random variable X_i^r in the k -iteration is defined in terms of its partial pmf as:

$$RU(X_i^r, k) = \frac{\hat{H}(P(X_i^r))}{\hat{H}_{MAX}} = \frac{\sum_{j=k}^n p_j(a_j^{(o)}) \log_2 p_j(a_j^{(o)})}{\log_2(n-k)} = \frac{\sum_{j=k}^n f_j \log_2 f_j}{\log_2(n-k)} \quad (7)$$

Selecting a $\beta \approx 1$, the resultant subset $A^{(o,l)}$ is closer to being uniformly distributed. Then, for a given β , and a number l of iterations carried out, it is possible to calculate the remnant R_i^r for a subset $A^{(o,l)}$. Generalizing this for an i -slot and a r -traffic feature we have the rate of remnant elements:

$$R_i^r = \begin{cases} n & \text{when all } p_j(a_j^{(o)}) = \frac{1}{n}, n \geq 1 \\ n-l & \text{for } l \geq 1 \end{cases} \quad (8)$$

In other words, R_i^r is the cardinality of the subset $A^{(o,l)}$. We found that this feature under normal conditions presents regularities that allow creating behavioral traffic profiles. **Table 1** summarizes the R_i^r behavior with $\beta = 0.95$ for our data-set.

Through of mean, variance, the intensity factor (σ^2/μ), and maximum value we can define a threshold for normal behavior of R_i^r . For instance, by averaging the means of R_i^r and its maximum values during benign traffic, we can define an average threshold of 28.5 with a maximum of 114.8 units. We denoted these thresholds for each r by $T(R_i^1) = (28.5; 114.8)$, $T(R_i^2) = (31.7; 115.6)$, $T(R_i^3) = (92.0; 342.6)$, and $T(R_i^4) = (132.3; 542.2)$.

Figure 1 shows the patterns R_i^1 and R_i^2 for benign traffic in Trace 5 and their variation are inside of standard behavior mentioned above.

Table 1. Values of mean, variance, and intensity factor for the rate of remnants.

Trace	Mean				Variance				Intensity Factor			
	<i>srcIP</i>	<i>dstIP</i>	<i>srcPrt</i>	<i>dstPrt</i>	<i>srcIP</i>	<i>dstIP</i>	<i>srcPrt</i>	<i>dstPrt</i>	<i>srcIP</i>	<i>dstIP</i>	<i>srcPrt</i>	<i>dstPrt</i>
1	29.1	33.1	103.5	136.5	328.2	377.5	4502	8818	11.29	11.42	43.5	64.61
2	29.5	35.1	88.8	138.6	566.5	688.1	3830	12,918	19.23	19.62	43.16	93.24
3	31.1	33.0	96.2	141.7	497.9	595.2	4598	11,972	15.99	18.06	47.78	84.51
4	31.1	35.9	103.2	141.9	732.8	727.9	6209	18,868	23.59	20.26	60.16	133
5	21.5	21.5	69.0	102.88	277.8	351.1	3208	9191	12.94	16.33	46.5	89.3
6-P1	14.5	16.6	76.1	74.5	145.1	166.8	4916	4958	10.0	10.05	64.6	66.53
6-P2	27.9	19,618	3107	1005	3464	6.2e07	1.7e06	4.3e04	12.42	3209	549.7	43.24
6-P4	3149	5214	3045	2243	97,058	1.2e06	2.4e05	2.4e05	310.9	237.5	90.14	111.24

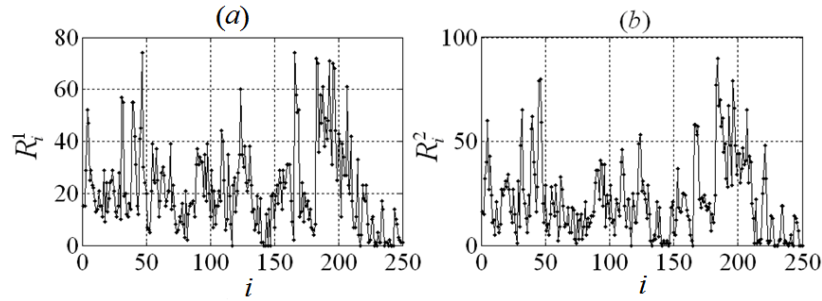


Figure 1. Rate of remnants for (a) *srcIP* and (b) *dstIP* for standard traffic in Trace 5 in typical work hours. ($t_d = 60$ s, and $\beta = 0.95$).

An anomaly related with a port scan attack directed to the proxy server was possible to detect it since the first slot that appeared (*i.e.* $i = 2, 3$) in trace 6-P1. The attack was carried out across a large number of TCP packets with source addresses supplanted. The growth of R_i^1 is possible to observe in **Figure 2(a)** and is far away from $T(R_i^1) = (28.5; 114.8)$.

R_i^r patterns during worms attacks are presented in **Figure 2(b)** and **Figure 2(c)**. There is a significant increase in the amount of remnants for $r = 3$ during the Blaster and Sasser Worm propagation. It is important to note that the anomaly detection is done from the earliest slots that the intrusion appears.

5. Three-Dimensional Spaces of Entropy

The construction of a space of entropy is carried out at flow level, and through these spaces is possible to create profiles of behavior for the traffic of a network. Four three-dimensional spaces are generated for each one of the features extracted from the flows. We define a traffic trace χ of a duration t_D seconds that is divided into M non-overlapping slots of $t_d = t_D/M$ seconds each one. In an i -slot K_i flows are generated with a given inter-flow gap (IFG). All the flows for each slot are stored on indexed text files. The traffic features used in this technique are the flow's fields and are identified as $r = 1$ for source IP address, $r = 2$ for destination IP address, $r = 3$ for source TCP port, and $r = 4$ for destination TCP port.

Once that flows in an i -slot are generated, they should be clustering according to a r -flow feature. For instance, with a cluster key or pivot $r = 1$ the flows are aggregated into those flows that share the same source IP address. The number of clusters depends on $|A_i^{r=1}|$, where $A_i^{r=1}$ is the alphabet set of all source IP addresses seen in the slot i . Thus, each cluster has flows with the same source IP address, but the rest of fields or features ($r = 2, 3, 4$) have freedom of variation. In this context, we can estimate the entropy for each $r = 2, 3, 4$ of each cluster. If we join these three values and associate them with a coordinate, we have a cloud of data points in a 3-D Euclidean space, where the axis are $(\hat{H}_{srcPrt}, \hat{H}_{dstPrt}, \hat{H}_{dstIP})$ for $r = 1$. Finally, the $|A_i^{r=1}|$ points in the slot i , that is, $(\hat{H}_{srcPrt}, \hat{H}_{dstPrt}, \hat{H}_{dstIP})_1, (\hat{H}_{srcPrt}, \hat{H}_{dstPrt}, \hat{H}_{dstIP})_2, \dots, (\hat{H}_{srcPrt}, \hat{H}_{dstPrt}, \hat{H}_{dstIP})_{|A_i^{r=1}|}$ are plotted in the 3D-space.

When we apply this procedure to the rest of cluster keys and all slots, we get four spaces of entropy.

Figure 3 shows the spaces of entropy for three. First, in **Figure 3(a)**, we see the point pattern for Trace-1, which corresponds to normal traffic conditions being typical for Traces 1 - 5. **Figure 3(b)** and **Figure 3(c)** show

a marked difference with regard to benign traffic, since the data points move away from positions typically observed.

The characterization of the spaces of entropy represented by the vector $\mathbf{X}^r \in \mathbb{R}^3$ for a cluster key r was realized applying initially a technique of multivariable analysis, the Principal Component Analysis. PCA provides a roadmap for how reduce a complex data-set to a lower dimension $\mathbf{Z}^r \in \mathbb{R}^d, d \leq 3$ to reveal the sometimes hidden, simplified structure that often underlie it. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component, PCA 1), the second greatest variance on the second coordinate, and so on, [12].

To obtain accurate analysis estimation, tools like Kernel Density Estimation (KDE) may be useful. This analysis was applied on data-points at the slot level on the PCA 1 by using a Gaussian kernel of 200 points, with a bandwidth of $h = 1.06\sigma J^{1/5}$, given by the Silverman's criteria, where J is the number of observations and σ is the standard deviation of the set of observations. KDE shows that the traffic slots have Gaussian bimodality behavior in its pdf. Each mode was labeled as principal mode and far mode respectively, as shown in **Figure 4**.

Figure 4(a) shows that the densities of PCA 1 in benign traffic traces present a clear regularity in their shapes. This pattern of behavior changes drastically for anomalous traffic traces, see **Figure 4(b)** and **Figure 4(c)**.

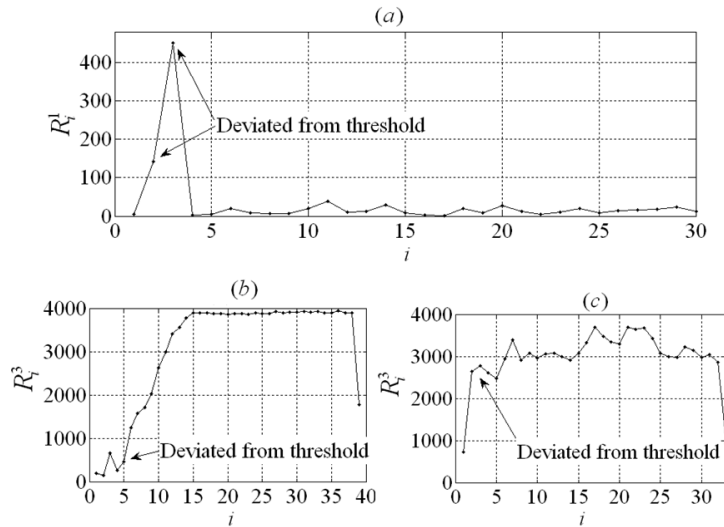


Figure 2. Rate of remnant for: (a) *srcIP*, under port scan attack using spoofed IP addresses which is observable in slots $i = 2$ and $i = 3$ on Trace 6-P1 (b) *srcPrt* during Blaster Worm on Trace 6-P2. (c) *srcPrt* during Sasser Worm on Trace 6-P4. All cases $t_d = 60$ s and $\beta = 0.95$.

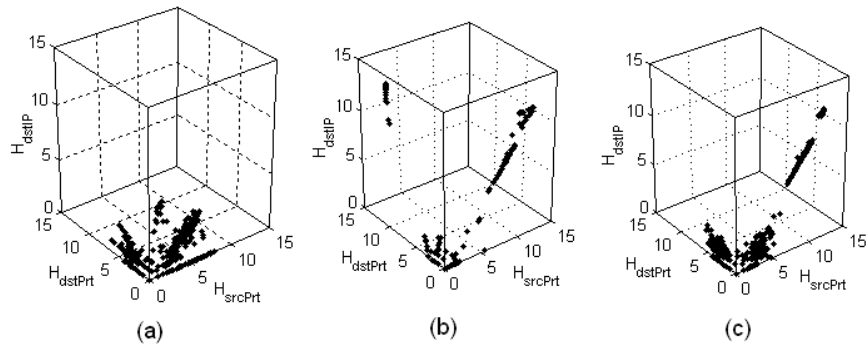


Figure 3. Spaces of entropy for three *srcIP* cluster keys: (a) Traffic Trace-1 (benign traffic) in typical working hours. (b) Anomalous traffic, Trace 6-P2 (Blaster Worm) during 38 min period. (c) Anomalous traffic, Trace 6-P4 (Sasser Worm) during 38 min period.

This procedure was applied to i -slot level for a given r -, feature. The transformed data points are denoted by Z_i^r . Densities of Z_i^r under normal conditions presents bimodality, the second mode (far mode) is situated on the left side of the main mode, at an average of minus five units. On the other hand, the empirically calculated average variance of Z_i^r for benign traffic is three units.

During analysis of the trace P1 was discovered that slots two and three had a variance of 0.32 and 0.74 units, respectively. These values are anomalous with regard to the threshold of three units. Thus, an anomaly (**Figure 5(a)**) was early detected in Trace 6-P1. The analysis of these traffic slots with Wireshark, was able to identify a port scan attack.

In another case, **Figure 5(b)** shows the first three anomalous slots, where far mode moves away from the typical value. The far mode displace to -9 , -11 and -13 units, representing an anomaly with regard to the threshold of -5 . This anomalous behavior in the far mode was caused by the spread of the Blaster worm.

Similarly, the Sasser worm attack caused anomalous values in the variance of the transformed data-points, these anomalous values were close to 0.5 units.

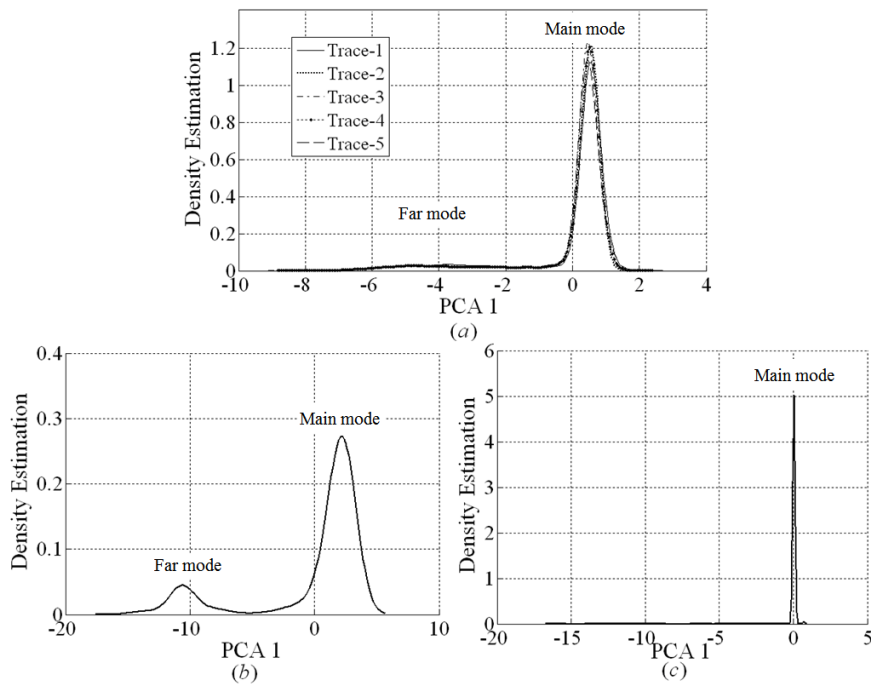


Figure 4. Kernel density estimation for PCA1 for: (a) Traces 1-5 (Benign Traffic), (b) Trace 6-P2 (Blaster attack), and (c) Trace 6-P4 (Sasser attack).

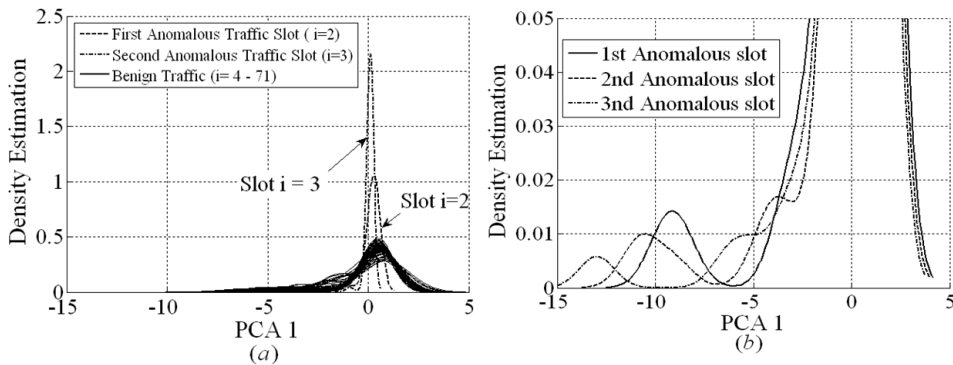


Figure 5. Anomaly Detection on PCA1 (a) Anomaly caused by port scan detected in first slot in Trace 6-P1, (b) Anomaly caused by the Blaster worm that deviate the far mode out of the typical position.

In all three cases studied, the attack leads to changes in traffic patterns that are detected from the first slots containing the malicious traffic.

6. Conclusions and Future Work

The generation of behavioral profiles based on entropy offers an effective support for the Intrusion Detection Systems. The results of this study in a campus network show that under the Blaster and Sasser worm attacks as well as the port scanning, an A-NIDS employing profiles generated by the Rate of remnant elements or Three-Dimensional Spaces of Entropy methodologies can provide a rapid response detecting deviations from an established baseline in the early slots that the attack appears.

As a future work, we will investigate the effect on variation of the slot duration t_d , smaller values of slot duration represent faster response times, but also represent a smaller data set where to obtain representative traffic features, finding the optimum value is an important objective design.

Acknowledgements

This research was supported by the National Counsel of Science and Technology (CONACyT) of Mexico as part of the project SEP-2011 Ciencia Basica No. 167859.

References

- [1] Fung, C. and Boutaba, R. (2013) *Intrusion Detection Networks: A Key to Collaborative Security*. Auerbach Publications.
- [2] Kruegel, C., Valeur, F. and Vigna, G. (2005) *Intrusion Detection and Correlation*. Advances in Information Security. Springer.
- [3] Xu, K., Zhang, Z. and Bhattacharyya, S. (2005) Profiling Internet Backbone Traffic: Behavior Models and Applications. *SIGCOMM*, **2005**, 22-26.
- [4] Nucci, A. and Bannerman, S. (2007) Controlled Chaos. *IEEE Spectrum*, **44**, 42-48.
<http://dx.doi.org/10.1109/MSPEC.2007.4390022>
- [5] Velarde-Alvarado, P., Vargas-Rosales, C., Torres-Roman, D. and Munoz-Rodriguez, D. (2008) Entropy Based Analysis of Worm Attacks in a Local Network. *Research in Computing Science*, **34**, 225-235.
- [6] Copley, D., Hassell, R., Jack, B., Lynn, K., Perme, R. and Soeder, D. (2003) ANALYSIS: Blaster Worm. eEye Digital Security Research. <http://research.eeye.com/html/advisories/published/AL20030811.html>
- [7] Ukai, Y. and Soeder, D. (2004) ANALYSIS: Sasser. eEye Digital Security Research. <http://research.eeye.com/html/advisories/published/AD20040501.html>
- [8] Jacobson, V., Leres, C. and McCanne, S. Tcpcap/libpcap. <http://www.tcpdump.org/>
- [9] A. Peppo, plab. Tool for Traffic Traces. <http://www.grid.unina.it/software/Plab/>
- [10] Trac Project. Libtrace. <http://www.wand.net.nz/trac/libtrace>
- [11] E. Kohler, ipsumdump. Traffic tool. <http://www.cs.ucla.edu/~kohler/ipsumdump>
- [12] Jolliffe, I.T. (2002) *Principal Component Analysis*, Series: Springer Series in Statistics. 2nd Edition, Springer, XXIX, 487 pp. 28.