DOI: 10.5769/C2013007 or http://dx.doi.org/10.5769/C2013007

Greatest Eigenvalue Time Vector Approach for Blind Detection of Malicious Traffic

Danilo Fernandes Tenório[2], João Paulo C. L. da Costa[1,2], and Rafael Timóteo de Souza Júnior[1,2]

(1) Laboratory of Array Signal Processing
(2) Laboratory for Decision-Making Technologies (LATITUDE)
Department of Electrical Engineering
University of Brasilia (UnB)
URL: www.pgea.unb.br/~lasp

***Abstract*—Recently, blind techniques have been applied to detect malicious traffic and attacks in honeypots. The honeypot traffic can be divided into legitimate and malicious traffic, where the legitimate traffic corresponds to DHCP, broadcasting, and synchronization. In practice, other servers connected to the network may be also targets for attacks and malicious traffic. Therefore, it is crucial to develop detection techniques for malicious traffic for such computers. In this paper, we propose a solution that blindly detects malicious traffic for any computer connected to the network. We validate our proposed solution considering two types of malicious traffic: synflood and portscan.**

***Keywords*—Eigenvalue Decomposition; Model Order Selection; Detection.**

I. Introduction

Nowadays one of the greatest challenges in Internet is security assurance, obtained by integrity, availability and confidentiality of data. There are several ways to provide security, taking into account both technical aspects, through the use of safety equipments or systems, as administrative and personal, related to establishment of a security policy and awareness campaigns. Regarding safety equipments or systems, we can use for instance firewall, Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS).

Several methods have been proposed for identifying and characterizing malicious activities. Classical methods typically employ data mining [1] [2] and regular file parsing [3] for detecting patterns which indicate the presence of specific attacks in the analyzed traffic. Recently, automatic blind malicious traffic detection techniques have been developed for honeypots [4] [5]. However, the honeypot traffic is simpler since there are no legitimate applications running.

The use of Model Order Selection (MOS) Schemes to detect highly correlated components as significant network activities; and identifying malicious activities in honeypot network flow datasets without any previous information or attack signatures by applying model order selection schemes has been proposed in [4].

In this work, we propose an automatic blind malicious traffic scheme to be used in any server of a network. Inspired by [4] [5], we model a real network traffic data into three components: the legitimate signals, the malicious signals and the noise.

Our proposed scheme is based on the eigenvalue decomposition, however, in contrast to [4] [5], we consider the time variation of the greatest eigenvalue. We show that based on this variation, attacks such as portscan and synflood can be detected.

This paper is organized as follows. In Section II, we define the notation used in this paper. In Section III, we discuss about data log, and how we model it as signals and noise. In Section IV, we characterize the portscan and synflood attacks. In Section V, we propose our scheme for the normalized and nonnormalized case. In Section VI, we explain in details the experiments with real data, and evaluate several MOS scheme presenting experimental results which attest the validity of our approach. In Section VII, we make our concluding remarks.

II. notation

In this paper the scalars are denoted by italic letters (*a, b, A, B, α, β*), vectors by lowercase bold letters (**a**, **b**), matrices by uppercase bold letters (**A**, **B**), and $a_{i,j}$ denotes the (*i, j*) elements of the matrix **A**. The superscripts ^{T} and $^{-1}$ are used for matrix transposition and matrix inversion, respectively.

III. Data Collection

The log information of a computer connected to the network is formed by timestamp, protocol, source IP address, source port, destination IP address, destination port and additional information, depending on the type of transport protocol used.

In order to exemplify these collected data, we consider the following TCP traffic log:

21:00:34.099289 IP 192.168.1.102.34712 > 200.221.2.45.80: Flags [S], seq 2424058224, win 14600, options [mss 1460, sackOK,TS val 244136 ecr 0,nop,wscale 7], length 0

and the UDP traffic log:

21:24:42.484858 IP 192.168.1.102.68 > 192.168.1.1.67: BOOTP/DHCP, Request from 00:26:9e:b7:82:be, length 300

In this paper, we consider only the following information from the log data timestamp, port type and port number.

A. Data Model

The reduced log data is divided into q time slots of N samples, where each sample is collected in a certain time period. Each element $x_{m,n}^{(q)}$ represents the number of times that the port m appears at the n-th time period, at the q-th time slot.

The collected data at the q-th time slot is represented by $\mathbf{X}^{(q)} \in \mathbb{R}^{M \times N}$, where M represents the amount of ports, and N represents the amount of time samples. The matrix $\mathbf{X}^{(q)}$contains all traffic (signal, noise and attack), and we can model it as:

$$\mathbf{X}^{(q)} = \mathbf{S}^{(q)} + \mathbf{N}^{(q)} + \mathbf{A}^{(q)} \quad (1)$$

where $\mathbf{S}^{(q)}$ is the matrix that represents the legitimate traffic, $\mathbf{N}^{(q)}$ represents the noise, and $\mathbf{A}^{(q)}$ the malicious traffic.

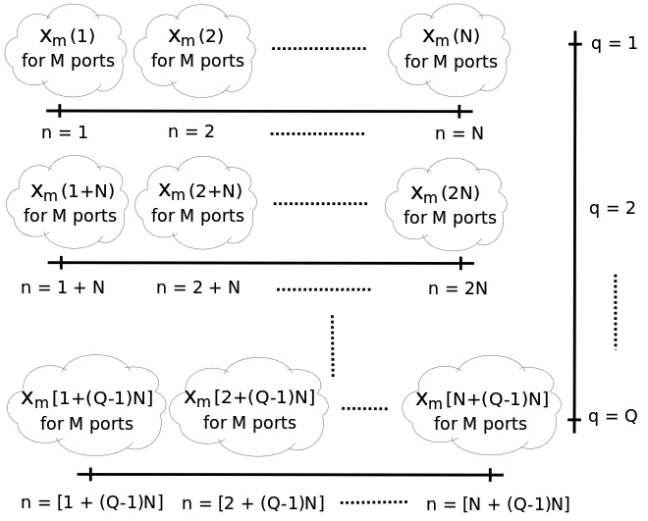

Figure 1. Traffic matrix $\mathbf{X}^{(q)}$, for q = 1, 2, …, Q.

Our goal in this paper is to detect the rank of the matrix $\mathbf{A}^{(q)}$, given only the matrix $\mathbf{X}^{(q)}$. Thereby, if the rank $\{\mathbf{A}^{(q)}\} \neq 0$, we have a malicious traffic; otherwise, if rank$\{\mathbf{A}^{(q)}\} = 0$, there is no malicious traffic.

IV. Characterization of Portscan and Synflood Attacks

In this section, we show important properties of the portscan and synflood. These properties are important to explain the validity of the proposed solution.

In the Fig. 2, the portscan transmits only two packets for each TCP port and one packet for each UDP port. Note that there is a high correlation since the traffic is equal.

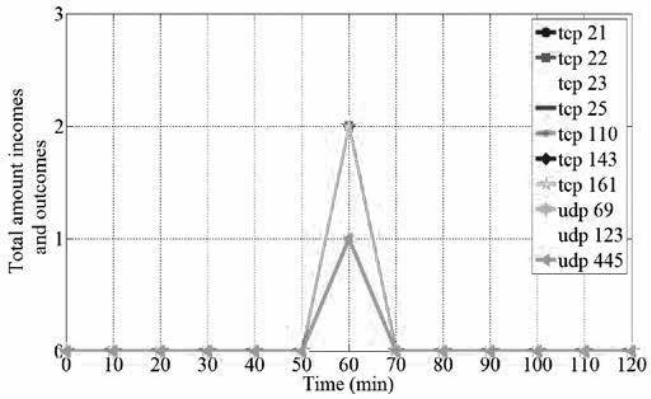

Figure 2. Malicious traffic over M ports vs n time slots (M = 10 and n = 120). This traffic profile represents the traffic characterized by portscan, consisting of TCP and UDP portscanning.

In the Fig. 3, the synflood attack consists of sending hundreds of packets with the SYN flag active in a short period of time. In our case, considering this attack, if a server with port 80 open, the server is overloaded and may cause the unavailability of the service rendered by it. In a time interval of ten minutes, there were more than two hundred ten thousand packages related to the attack, an unusual traffic in a data network, especially by the fact of being concentrated in a short period of time.

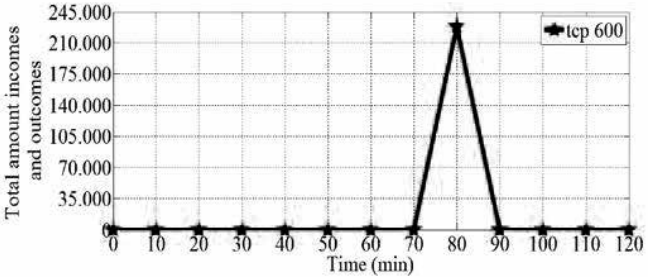

Figure 3. Malicious traffic over M ports vs n time slots (M = 1 and n = 120). This traffic profile represents the traffic characterized by synflood.

II. Proposed Solution

Basically, the model order of a dataset is estimated as the number of main correlated components with energy significantly higher than the rest of uncorrelated components. In other words, the model order can be characterized by a power gap between main components and the noise components. In the context of network traffic, the principal components are represented by outstanding network activities, such as highly correlated network connections which have, for example, the same destination port [4]. The efficacy and efficiency of methods based on Principal Component Analysis (PCA) depend on the MOS scheme adopted, since each scheme has different probabilities of detection for different kinds of data [6].

We consider two cases, one case normalizing $\mathbf{X}^{(q)}$, and the other one nonnormalizing it. The purpose of this was to adapt the solutions to the characteristics of portscan and synflood

attacks. Thus, we built two correlation matrices. One from the normalized case $\mathbf{R}_{xx}^{'(q)} \in \mathbb{R}^{M\times M}$, and other from the nonnormalized case $\mathbf{R}_{xx}^{(q)} \in \mathbb{R}^{M\times M}$.

A. Normalized case

In detecting portscan we have low values associated with this attack (Fig. 2), but repeatedly, since we are scanning in multiple ports (correlated traffic). Then, when we normalize the $\mathbf{X}^{(q)}$matrices, the portscan can be collected.

The normalization of the vectors can be obtained by the following equation:

$$\mathbf{x}_m^{'(q)} = \frac{\mathbf{x}_m^{(q)} - \bar{\mathbf{x}}_m^{(q)}}{\sigma_m^{(q)}} \tag{2}$$

for $q = 1, ..., Q$ where $\bar{\mathbf{x}}_m^{(q)}$ is the mean of $\mathbf{x}_m^{(q)}$ and $\sigma_m^{(q)}$ is the deviation standard of $\mathbf{x}_m^{(q)}$.

Once obtained the vectors $\mathbf{x}_m^{'(q)}$, we can construct the matrix $\mathbf{X}^{'(q)}$, and then determine the correlation matrix in order to find the eigenvalues.

$$\mathbf{R}_{xx}^{'(q)} = \frac{1}{N}\ \mathbf{X}'^{(q)}\mathbf{X}'^{(q)^{\mathrm{T}}} \tag{3}$$

where N is the number of the sample time.

The eigenvalue decomposition of $\mathbf{R}_{xx}^{'(q)}$ is given by:

$$\mathbf{R}_{xx}^{'(q)} = \mathbf{E}'^{(q)}\mathbf{\Lambda}'^{(q)}\mathbf{E}'^{(q)\mathrm{T}} \tag{4}$$

where $\mathbf{\Lambda}^{'(q)}$ is a diagonal matrix with the eigenvalues, and the matrix $\mathbf{E}^{'(q)}$ has the eigenvectors corresponding to each eigenvalue. However, for our model order selection schemes, only the eigenvalues are necessary.

By selecting only the main diagonal of the matrix $\mathbf{\Lambda}^{'(q)}$, via diag{$\mathbf{\Lambda}^{'(q)}$}, and by ranging $q = 1, ..., Q$, we can build a matrix $\mathbf{C}' \in \mathbb{R}^{M\times Q}$.

Assuming that the eigenvalues $\lambda_m^{'(q)}$ are in the descending order, i. e., $\lambda_1^{'(q)} > \lambda_2^{'(q)} > \cdots > \lambda_{m-1}^{'(q)} > \lambda_m^{'(q)}$, the first column of the matrix $\mathbf{C}'$ has the Greatest Eigenvalue Time Vector (GETV).

As shown in the Section VI, by using the GETV in the model order selection schemes, it's possible to detect the presence of malicious traffic even if applications are running.

B. Nonnormalized case

In detecting synflood we have a huge uncorrelation traffic. Thus, differently of portscan we have in this case only one traffic, but with a high value. So, we cannot normalizing this traffic since the normalization would cause an abrupt reduction of the high value associated with this attack, causing it to disappear.

Thus, in order to detect the synflood, we cannot normalize the matrix $\mathbf{X}^{(q)}$. However, except by normalization, the whole procedure to find the GETV is equal to the one shown is Subsection V.A.

C. GETV combined with Model Order Selection Schemes

Each model order selection scheme has different characteristics. We used the following method in our simulations: AIC [7] [8], MDL [7] [8], EDC [8] [9], RADOI [10], EFT [11] [13] and SURE [12].

The EFT and EDC models showed the satisfactory results for our scheme. In case of EDC, the information criterion is a function of the geometric mean, *g(i)*, and arithmetic mean, *a(i)*, of the *i* smallest eigenvalues. Note that $c_{1,q}$ and $c_{1,q}^{'}$, $q = 1, ..., Q$, should be in descending order.

The estimate of the model order d can be represented by $\hat{d}$, through the following expressions:

$$\hat{d} = \operatorname{argmin}(J(i)) \tag{5}$$

$$J(i) = -2N(Q - i\ +1\ )\log\left(\frac{g(i)}{a(i)}\right) + (i-1)p(Q,i,N) \tag{6}$$

where $p(Q,i,N) = [2Q\ -(i-1)]\sqrt{N\log(\log(N))}$.

To use the (6) we have firstly to put the vector of eigenvalues in ascend order.

For the EFT based schemes, i.e. R-D EFT II, R-D EFT, M-EFT, and EFT, we have has to compute the threshold coefficients, as shown in [13]. Without the threshold coefficients, the EFT based schemes cannot be applied. By computing these coefficients and applying the EFT it's possible to find the model order of our scheme.

VI. Simulations

In this section, we describe the performed experiments in order to validate our proposed scheme for detecting portscan and synflood attack in a computer.

A. Data Analysis

For this simulation we used a computer (based on Linux operational system) performing common tasks (web access mainly) during an interval of three hours. The application tcpdump was used to capture the network traffic, as shown in Fig 4.

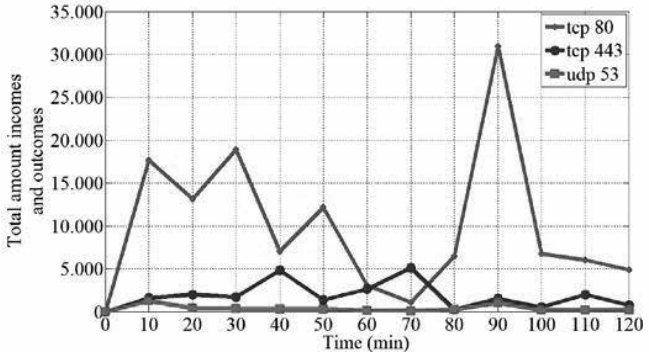

Figure 4. Signal traffic over M ports vs n time slots (M = 3 and n = 120). This traffic profile represents only the signal, without any kind of attack or noise.

At 21:54h we conducted with the portscan, in order to simulate an attacker who wants to know the status of the following ports: TCP 21, 22, 23, 25, 110, 143 and 161; and UDP 69, 123, and 445. At the range time from 22:10h to 22:19h we conducted with the synflood attack (log below), in order to simulate an attacker who wants to cause a Denial of Service (DoS), causing unavailability of services.

22:10:04.986927 IP 192.168.1.104.64263 > 192.168.1.102.600: Flags

[S], seq 3652238756, win 1365, length 0

22:10:04.986961 IP 192.168.1.102.600 > 192.168.1.104.64263: Flags [R.], seq 0, ack 3652238757, win 0, length 0

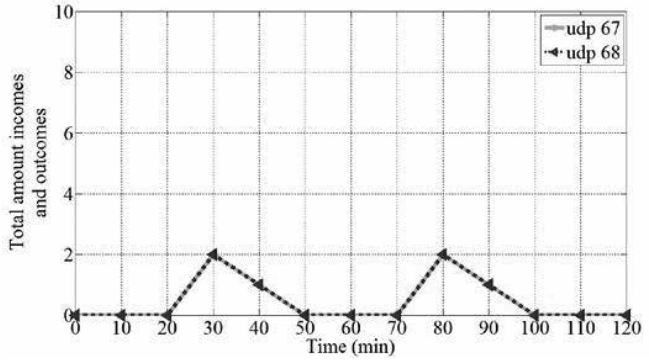

Figure 5. Noise traffic over M ports vs n time slots (M = 2 and n = 120). This traffic profile represents only the noise, consisting of udp 67 and udp 68 traffic.

The Fig. 4 shows the signal traffic consisting of requests and responses on TCP port 80, TCP port 443 and UDP port 53. The TCP port 80 is associated with unencrypted web access, the TCP port 443 to encrypted web access, and UDP port 53 is associated with name resolution, DNS.

The noise traffic (log below) is formed by UDP port 67 and UDP port 68, associated with the Dynamic Host Configuration Protocol (DHCP). It can be seen in Fig. 5.

21:24:42.484858 IP 192.168.1.102.68 > 192.168.1.1.67: BOOTP/DHCP, Request from 00:26:9e:b7:82:be, length 300

21:24:42.487652 IP 192.168.1.1.67 > 192.168.1.102.68: BOOTP/DHCP, Reply, length 548

B. Eigenvalue Decomposition (EVD)

As described in Section III, the total simulation time of 120 minutes was fragmented into Q = 6 periods of N = 20 minutes each, where each period we use the sampling time of 1 minute. As the simulation began at 21:00h, the first period goes from 21:00h until 21:20h (T_1), the second from 21:20h until 21:40h (T_2), the third from 21:40h to 22:00h (T_3), the fourth from 22:00h until 22:20h (T_4), the fifth from 22:20h until 22:40h (T_5), and finally the sixth from 22:40h to 23:00h (T_6). Thus, it was possible to build Q = 6 matrices $\mathbf{X}^{(q)}$ of the total traffic (signal + noise + attack). Obviously not every period there is attack, only at T_4 occurred the synflood attack (Fig. 3), and at T_3 the portscan (Fig. 2).

Once we have the $\mathbf{X}^{(q)}$ matrices for each period, it is now possible to obtain the correlation $\mathbf{R}_{xx}^{'(q)}$ and $\mathbf{R}_{xx}^{(q)}$ matrices, related to each matrix $\mathbf{X}^{(q)}$. With that it was possible to obtain the set of eigenvalues for that correlation matrices, generating a total of 2Q vectors of eigenvalues: 6 vectors related to $\mathbf{R}_{xx}^{'(q)}$, built from the normalization of $\mathbf{X}^{(q)}$ (Fig. 6), and 6 vectors related to $\mathbf{R}_{xx}^{(q)}$, built from the nonnormalizing of $\mathbf{X}^{(q)}$ (Fig. 7).

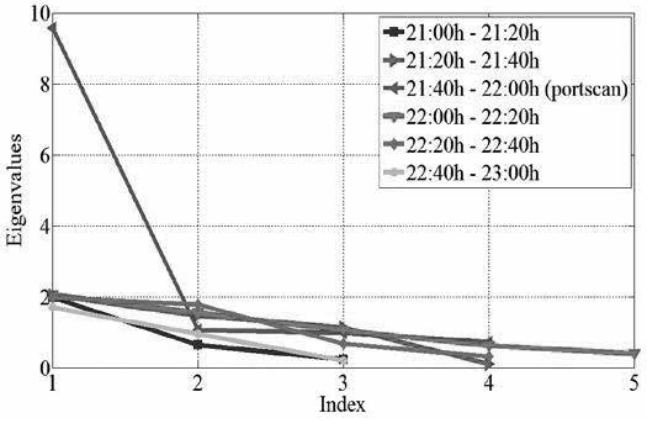

Figure 6. Eigenvalues of the normalized case over each time slot. In this figure is the greatest eigenvalue related to the portscan is much greater than the others.

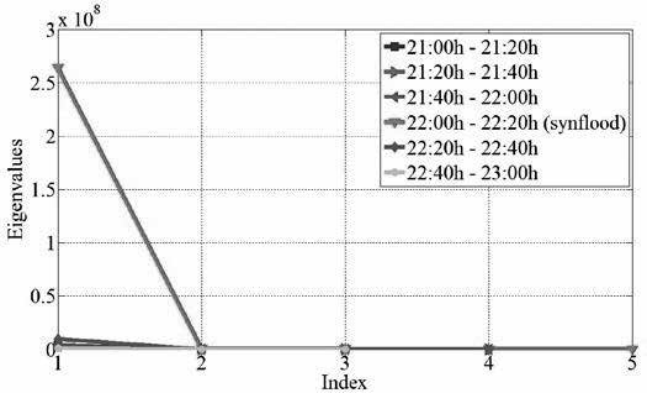

Figure 7. Eigenvalues of the nonnormalized case over each time slot. In this figure is possible to see the eigenvalue related to the synflood (purple line index 1).

Calculating the eigenvalues of each $\mathbf{R}_{xx}^{'(q)}$ and $\mathbf{R}_{xx}^{(q)}$ matrices, we can reduce the size of our matrix $\mathbf{X}^{(q)}$, and get some interesting conclusions, derived from the eigenvalue decomposition properties of the correlation matrices, such as: the eigenvectors associated with each eigenvalue are orthogonal to each other, and also linearly independent; and the eigenvalues are real and nonnegative.

C. Applying Model Order Selection to Data Analysis

Although the variation of the eigenvalues related to the attacks, the job is not complete until we find a model that applies to this scheme. The estimation of the model order by visual inspection is performed by following subjective criteria such as considering only the eigenvalues greater than one and visually identifying a large gap between two consecutive eigenvalues. Then, to let this work the most complete and objective as possible we tested several MOS approach, like AIC, MDL, EDC, RADOI, EFT and SURE.

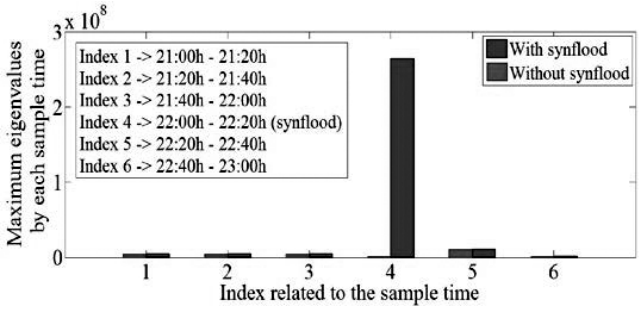

Figure 8. Greatest eigenvalue time vector approach related to the nonnormalized case. It is possible to see the eigenvalue related to the synflood, much greater than the other ones (brown bar index 4).

Before we show the results obtained with the application of the models, we will discuss the input values for each MOS approach. The Fig. 8 and Fig. 9 show the greatest eigenvalues obtained in each period. Thus, we applied the Greatest Eigenvalue Time Vector (GETV) approach.

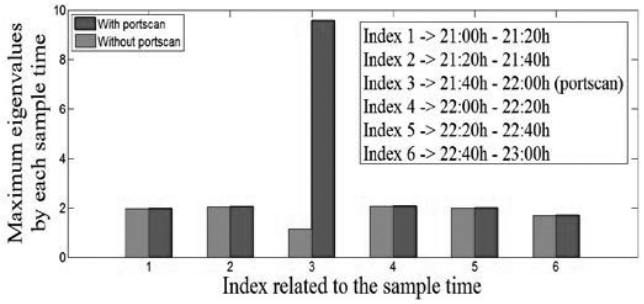

Figure 9. Greatest eigenvalue time vector approach related to the normalized case. It is possible to see the eigenvalue related to the portscan, greater than the other ones (brown bar index 3).

The method consists of selecting the largest eigenvalues of the q = 6 time slots, and apply them to the existing model order selection schemes in order to detect malicious traffic.

In Figs. 8 and 9, we show the greatest eigenvalues of the q = 6 time slots. In Fig. 8 we have the nonnormalized case, used to detect synflood attack. In this figure it is possible to compare the values of the eigenvalues with and without the attack. We can see clearly how the component related to the attack stands out from the rest. In Fig. 9 we have the normalized case, used to detect the portscan. In this figure it is possible to compare the values of the eigenvalues with and without the attack.

TABLE I. Nonnormalized Case

Case	Method					
	AIC	*MDL*	*EDC*	*RADOI*	*EFT*	*SURE*
With synflood	2	1	1	4	1	14
Without synflood	4	1	0	1	0	13

The tables I and II were obtained after we apply the methods cited in our scheme. The table I give us the results in the nonnormalized case, and the table II in the normalized case.

TABLE II. Normalized Case

Case	Method					
	AIC	*MDL*	*EDC*	*RADOI*	*EFT*	*SURE*
With portscan	1	1	1	1	1	6
Without portscan	0	0	0	1	0	1

According to the tables, we see the various model orders and found two that stood out. The Efficient Determination Criterion (EDC) and Exponential Fitting Test (EFT) showed us the correct order models, which is equal to one, indicating that there was an attack. The methods are efficient in detecting both attacks. The efficiency of the model as it behaves when there is no attack, in this case showing that the model order is zero, indicating that there is no attack, neither portscan nor synflood.

VII. Conclusion

In this paper we propose The Greatest Eigenvalue Time Vector (GETV) approach for detecting portscan and synflood in a network traffic flow data collected at a computer. First we showed the data log used, and the propose of the model for network flow data, in order to verify the validity of our approach through simulation results with real log files collected at a computer. Several model order selection methods were experimented with the simulation data, showing that EDC and EFT yields the best results for this type of data.

Since our proposed scheme is blind, it does not require previous collection of data and learning periods.

References

[1] Y. H. Hu "Parallel eigenvalue decomposition for toeplitz and related matrices," International Conference Acoustics, Speech, and Signal Processing (ICASSP'89), 1989, pp. 1107 -1110, vol. 2, 1989.

[2] H. T. Wu, J. F. Yang and F. K. Chen. "Source number estimators using transformed gerschgorin radii," IEEE Transactions on Signal Processing, 43(6):1325–1333, 1995.

[3] R. R. Nadakuditi and A. Edelman. "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," IEEE Transactions of Signal Processing, 56:2625–2638, 2008.

[4] B. M. David, J. P. C. L. da Costa, A. C. A. Nascimento, M. D. Holtz, D. Amaral, and R. T. Sousa Júnior. "Blind automatic malicious activity detection in honeypot data," pp. 02-04, ICoFCS 2011.

[5] J. P. C. L. da Costa, E. P. de Freitas, B. M. David, A. M. R. Serrano, D. Amaral, and R. T. Sousa Júnior. "Improved blind automatic malicious activity detection in honeypot data," ICoFCS 2012.

[6] J. P. C. L. da Costa, "Parameter estimation techniques for multidimensional array signal processing," Shaker, First edition, 2010.

[7] J. P. C. L. da Costa, A. Thakre, F. Roemer, and M. Haardt, "Comparison of model order selection techniques for high-resolution parameter estimation algorithms," in Proc. 54th International Scientific Colloquium (IWK'09), Ilmenau, Germany, Oct. 2009.

[8] M. Wax and T. Kailath, "Detection of signals by information theoreticcriteria," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-33, pp. 387–392, 1985.

[9] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai. "On detection of the number of signals in presence of white noise," J. Multivar. Anal., 20:1- 25, 1986.

[10] E. Radoi and A. Quinquis, "A new method for estimating the number of harmonic components in noise with application in high resolution radar," EURASIP Journal on Applied Signal Processing, pp. 1177–1188, 2004.

[11] A. Quinlan, J.-P. Barbot, P. Larzabal, and M. Haardt, "Model order selection for short data: An exponential fitting test (EFT)," EURASIP Journal on Applied Signal Processing, 2007, Special Issue on Advances in Subspace-based Techniques for Signal Processing and Communications.

[12] M. O. Ulfarsson and V. Solo, "Rank selection in noisy PCA with SURE and random matrix theory," in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Las Vegas, USA, Apr. 2008.

[13] J. P. C. L. da Costa, F. Roemer, F. A. de Castro Junior, R. F. Ramos, S. Schwarz, and L. Sabirova, ``Ilmenau Package for Model Order Selection and Evaluation of Model Order Estimation Scheme of Users of MIMO Channel Sounders,'' XXIX Simpósio Brasileiro de Telecomunicações (SBrT'11), Curitiba, Brazil.