The Importance of Time in the Identification of Anomalous Situations by Means of MOVICAB-IDS

Álvaro Herrero, Emilio Corchado, and Bruno Baruque
Department of Civil Engineering, University of Burgos
C/Francisco de Vitoria s/n. C.P.: 09006. Burgos, Spain
Phone: +34 947259395, email: {ahcosio, escorchado, bbaruque}@ubu.es

ABSTRACT: Intrusion Detection Systems (IDSs) are a part of the computer security infrastructure of most organizations. They are designed to detect suspect patterns by monitoring and analysing computer network events. Different areas of artificial intelligence, statistical and signature verification techniques have been applied in the field of IDSs. Additionally, visualization tools have been applied for intrusion detection, some of them providing visual measurements of network traffic. As described in previous works, MOVICAB-IDS (MObile VIsualization Cooperative Agent-Based IDS) is a bio-inspired tool based on the use of unsupervised Neural Networks (NN), and provides the network administrator with a snapshot of network traffic, protocol interactions and traffic volume. It offers a complete and more intuitive visualization of the network traffic by depicting each simple packet. To improve the accessibility of the system, the administrator may visualize the results on a mobile device (such as PDA's, mobile phones or embedded devices), enabling informed decisions to be taken anywhere and at any time. It is a combination of a connectionist model and a multiagent system enriched by a functional and mobile visualization. The viability and effectiveness of MOVICAB-IDS has been shown in previous works. This paper focuses on the importance of the time-information dependence in the identification of anomalous situations in the case of the proposed model. Several experiments show that the connectionist method on which MOVICAB-IDS is based (that has never been applied to the IDS and network security field before the beginning of this research) can highlight the evolution of packets along time. That is, MOVICAB-IDS identifies anomalous situations by taking into account the time-related dimension among others and by using unsupervised bio-inspired models.

KEYWORDS: Unsupervised Learning, Connectionist Models, Exploratory Projection Pursuit, Multiagent Systems, Computer Network Security, Intrusion Detection.

INTRODUCTION AND PREVIOUS WORK

Intrusion Detection Systems (IDSs) are tools designed to monitor and analyse computer or network events in order to detect suspect patterns that may relate to a network or system attack. They have become a common complementary tool in the computer security infrastructure of most organizations.

Many different areas of Artificial Intelligence (such as Genetic Programming [1], Data Mining [2], [3] or Neural Networks (NN) [4], [5], [6] among others), statistical [7] and signature verification [8] techniques have been applied in the field of IDSs. There are several IDSs that can generate different alarms when an anomalous situation occurs, but they can not provide a general overview of what is happening inside a network. Various visualization techniques have been applied in the field of IDSs [4], [5], [9], [10], [11], [12] to tackle this issue. This kind of techniques has become an effective tool to address the complex problem of identifying anomalous situations in network traffic data. There is a great variety of these techniques. Apart from using a 2D or 3D visualization, some of them just summarize the data by port number [10], [11], while others show different measurements of network traffic [12], [13] or visualize alarm data [14], [15], [16] generated by different IDSs. Most of them use a glyph metaphor [14], [17] to encode information by changing different features (color, size, opacity, etc. in addition to the spatial coordinates). Some others use traditional graphs such as histograms [11] or histographs [13] among others.

A different approach is followed by MOVICAB-IDS, where each simple packet is visualized on its own. The model proposed in this paper offers a complete and more intuitive visualization of network traffic by depicting each packet and providing the network administrator with a snapshot of network traffic, protocol interactions, and traffic volume in order to identify anomalous situations.

Another important and quite novel feature in this work is the use of the time information variable. Normally, it is not used in the NN-based IDSs. In this model, due to the fact that we are applying neural projectionist techniques in the data analysis step, the time information variable provides an idea of how the traffic data flows. It helps to identify anomalous situations by taking into account such aspects as high packet density among others.

MODEL OVERVIEW

Our model is designed to split massive traffic datasets into segments and analyse them, thereby providing administrators with a visual tool to analyse the kinds of events taking place on the computer network. This tool provides an analysis of several subsequent segments as unique ones (simple segments) and also as an accumulated dataset.
MOVICAB-IDS (MObile VIsualization Connectionist Agent-Based Intrusion Detection System) may be defined as an IDS formed of different software agents [18] that work in unison in order to detect anomalous situations by taking full advantage of an unsupervised connectionist model [4], [5], [19], [20]. Five different steps are performed by MOVICAB-IDS in order to detect anomalous situations:

- 1st step.- Network Traffic Capture: captures packets travelling over the different network segments.
- 2nd step.- Data Pre-processing: the captured data is selected and pre-processed. A set of packets and features contained in the headers of the captured data is selected from the raw network traffic.
- 3rd step.- Segmentation: the data stream is divided into simple segments and accumulated ones (consisting of the addition of several consecutive simple segments).
- 4th step.- Data Analysis: a connectionist model is applied to analyse the data.
- 5th step.- Visualization: the projections are presented to the network administrator. The projections may be displayed on a different device than the one used for the first four steps. To improve the accessibility of the system, the administrator may visualize the results on a mobile device (as can be seen in Figure 3), enabling informed decisions to be taken anywhere and at any time.

THE UNSUPERVISED CONNECTIONIST MODEL

The data analysis task is based on the use of an unsupervised neural model called Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [19], [20]. CMLHL is based on Maximum Likelihood Hebbian Learning (MLHL) [21], [22] adding lateral connections [19], [20] which have been derived from the Rectified Gaussian Distribution [23]. The resultant NN can find the independent factors of a dataset but does so in a way that captures some type of global ordering in the dataset.
Considering an N-dimensional input vector (x), an M-dimensional output vector (y) and with W_{ij} being the weight (linking input j to output i), CMLHL can be expressed [14], [15], [16] as:

1. Feed-forward step:

$$y_i = \sum_{j=1}^{\mathbf{N}} W_{ij} x_j, \forall i \ . \tag{1}$$

2. Lateral activation passing:

$$y_i(t+1) = \left[y_i(t) + \tau(b - Ay) \right]^+ \ . \tag{2}$$

3. Feedback step:

$$e_j = x_j - \sum_{i=1}^{M} W_{ij} y_i, \forall j \ . \tag{3}$$

4. Weight change:

$$\Delta W_{ij} = \eta . y_i . sign(e_j) | e_j |^{p-1} \ . \tag{4}$$

Where: η is the learning rate, τ is the "strength" of the lateral connections, b the bias parameter, p a parameter related to the energy function [20], [21], [22] and A a symmetric matrix used to modify the response to the data. The effect of this matrix is based on the relation between the distances among the output neurons.

DATASETS

In the data pre-processing step, the system performs a data selection from all of the captured information. As a result, all of the above-mentioned datasets contain the following 5 variables extracted from the packet headers: protocol (all the protocols contained in the dataset have been codified), source port (the port number from where the source host sent the packet), destination port (the destination host port number to which the packet is sent), size (total packet size in Bytes) and timestamp (the time when the packet was sent). Showing the importance of the use of time information is the aim of this work. Following this idea, 2 different variations of each one of the datasets have been used. The difference between the variations is the inclusion or exclusion of time information.

The datasets have been analysed using unsupervised learning because in a real-life situation, there is no target reference with which to compare the response of the network. The use of this kind of learning is very appropriate for identifying unknown (0-day) attacks.

Our efforts have focussed on the study of two dangerous anomalous situations: MIB information transfers and port sweeps or scans. The MIB (Management Information Base) can be defined in broad terms as the database used by SNMP (Simple Network Management Protocol) to store information about the elements that it controls. This situation is a transfer of some or all the information contained in the SNMP MIB. This kind of transfer is potentially quite a dangerous situation because anybody who possesses some free tools, some basic SNMP knowledge and the community password (in SNMP v. 1 and SNMP v. 2) will be able to access all sorts of interesting and sometimes useful information. A port scan may be defined as series of messages sent to different port numbers to gain information on its activity status. These messages can be sent by an external agent attempting to access a host to find out more about the network services that this host is providing. A port scan provides information on where to probe for weaknesses, for which reason scanning generally precedes any further intrusive activity. On the other hand, there are network scans, in which the same port is the target for a number of computers. A network scan is one of the most common techniques used to identify services that might then be accessed without permission [11].

Finally, in order to test our model, 2 mutated datasets have been used. A mutation [24] testing model for numerical datasets has been applied to obtain these datasets. This model modifies different features of the information extracted from the packet headers. The modifications created by this model include changes in aspects such as: attack length (amount of time that each attack lasts), packet density (number of packets per attack), attack density (number of attacks per time unit) and time intervals between attacks. The mutations can also concern both source and destination ports. Mutations related with time dimension have been discarded.

In this work, 5 different datasets have been analysed by MOVICAB-IDS:

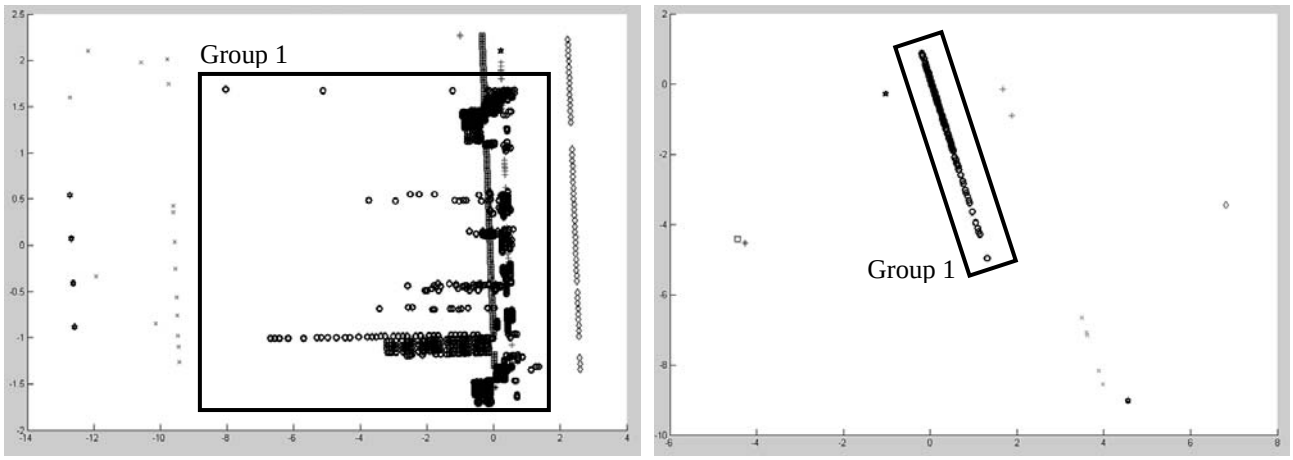
- Dataset 1: includes 3 sweeps, each one aimed at port numbers 161, 162 and 3750 respectively.
- Dataset 2: includes an MIB information transfer.
- Dataset 3: blends the examples contained in the datasets 1 and 2.
- Dataset 4: includes 2 sweeps (each one containing 5 packets) aimed at ports 4427 and 4439.
- Dataset 5: includes 2 sweeps (each one containing 30 packets) aimed at ports 1434 and 65788.

These datasets were generated ‘made-to-measure’ in a middle-size network where the “normal” traffic was known in advance. In addition to the SNMP packets, these datasets contain traffic related to other protocols installed in this network (“normal” traffic).

EXPERIMENTS AND RESULTS

As it is mentioned above, several experiments on each different dataset have been carried out. This section shows a comparison of the best results (for datasets 2, 3 and 5) obtained through the experiments. 2 different illustrations are shown for each dataset: the one on the left side (a) including time information and the one on the right side (b) excluding time information as a variable.

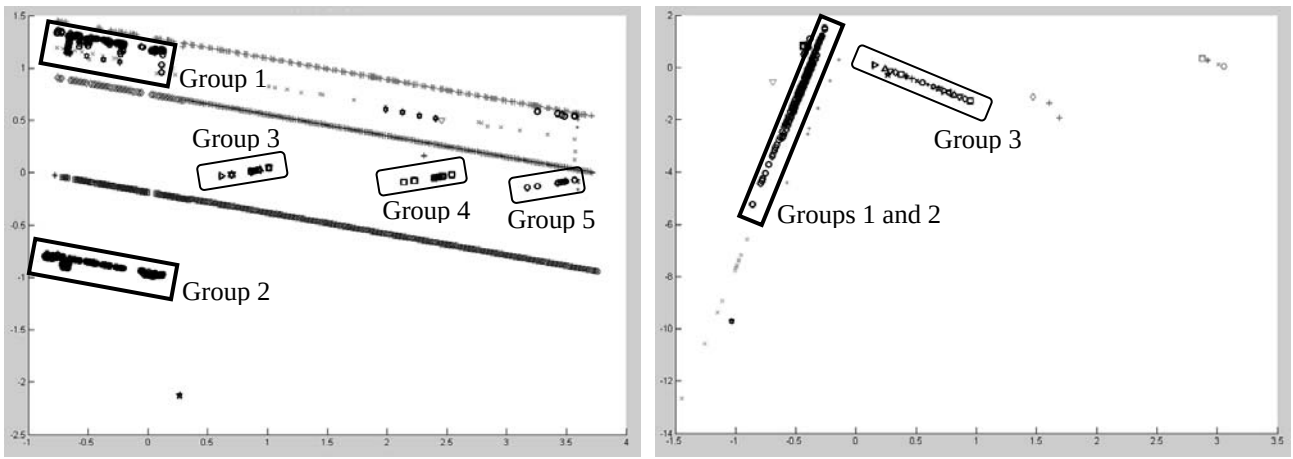
Different colours and shapes are used to depict each packet depending on the protocol it belongs to.



a) including time information b) excluding time information

Figure 1: MOVICAB-IDS projection for dataset 2.

Figure 1 shows the best projections obtained by MOVICAB-IDS for dataset 2. After visually analysing these results, we can clearly say that the inclusion of time information (Figure 1.a) allows the identification of the MIB transfer (Group 1) contained in dataset 2 in a clear way. As can be seen in the right illustration (Figure 1.b), the exclusion of time information concentrate the MIB transfer in a line (Group 1), what hinder the network administrator in the identification of this anomalous situation.



a) including time information b) excluding time information

Figure 2: MOVICAB-IDS projection for dataset 3.

By taking into account the variables previously described, a sweep can behave in different ways. The source port can be changed into a range in order to speed up the packet sending. In the case of a port scan, both the destination port and the protocol change for each packet. Dataset 3 includes 3 sweeps and each one of the packets contained in the sweep aimed at port 3750 belongs to a different protocol. Figure 2 shows the best projections obtained by MOVICAB-IDS for this dataset. It is easy to identify the 3 sweeps (Groups 3, 4 and 5) in Figure 2.a but only one (Group 3) of them (that whose packets belongs to different protocols) can be identified in Figure 2.b. As in the previous figure, time information allows the clear identification of the MIB transfer (Groups 1 and 2).

Finally, the projections for dataset 5 are shown in Figure 3. They confirm that excluding the time information impedes the identification of the MIB transfer and certain type of sweep (that whose packets belongs to the same protocol).

a) including time information b) excluding time information

Figure 3: Dataset 5 snapshot in a mobile platform emulator.

CONCLUSIONS

After analysing the results shown in the previous section, we can conclude that MOVICAB-IDS can deal with time information. Using this information allows us to identify some anomalous situations that would be unidentifiable otherwise. This research constitutes one of the first attempts to identify anomalous situations working at a single packet-level. That is, we do not use summarizing information (as TCP connections for example). On the contraty, we analyse each single packet.

Some IDSs based on supervised models need a "clean" (free of attacks) training dataset. This is not the case of MOVICAB-IDS, that can be trained with dataset containing anomalous situations.

Acknowledgments. This research has been supported by the MCyT project TIN2004-07033 and the project BU008B05 of the JCyL.

REFERENCES

[1] Abraham, A., Grosan, C., Martin-Vide, C.: Evolutionary Design of Intrusion Detection Programs. International Journal of Network Security (2006)

[2] Julisch, K.: Data Mining for Intrusion Detection: A Critical Review. Research Report RZ 3398, IBM Zurich Research Laboratory. Switzerland (2002)

[3] Lee, W., Stolfo, S.J.: A Framework for Constructing Features and Models for Intrusion Detection Systems. ACM Trans. on Inf. and System Sec. (TISSEC), Vol. 3(4). ACM Press, New York (2000)

[4] Corchado, E., Herrero, A., Sáiz, J.M.: Testing CAB-IDS through Mutations: on the Identification of Network Scans. Proc. of the Int. Conf. on Knowledge-Based & Intelligent Information & Engineering Systems (KES2006). LNCS, Vol. 4252. Springer-Verlag, Berlin Heidelberg New York (2006) 433-441

[5] Corchado, E., Herrero, A., Sáiz J.M.: Detecting Compounded Anomalous SNMP Situations Using Unsupervised Pattern Recognition. Proc. of the Int. Conf. on Artificial Neural Networks (ICANN 2005). LNCS, Vol. 3697. Springer-Verlag, Berlin Heidelberg New York (2005) 905-910

[6] Zanero, S., Savaresi, S.M.: Unsupervised Learning Techniques for an Intrusion Detection System. Proceedings of the ACM Symposium on Applied Computing (2004) 412-419

[7] Marchette, D.J.: Computer Intrusion Detection and Network Monitoring: A Statistical Viewpoint. Information Science and Statistics. Springer-Verlag, Berlin Heidelberg New York (2001)

[8] Roesch, M.: Snort - Lightweight Intrusion Detection for Networks. Proceedings of the 13th Systems Administration Conference (LISA '99) (1999)

[9] Goldring, T.: Scatter (and Other) Plots for Visualizing User Profiling Data and Network Traffic. Proceedings of the ACM Workshop on Visualization and Data Mining for Computer Security (2004)

[10] Muelder, Ch., Ma, K-L., Bartoletti: Interactive Visualization for Network and Port Scan Detection. Proceedings of the 8th International Symposium on Recent Advances in Intrusion Detection (RAID). Lecture Notes in Computer Science, Vol. 3858. Springer-Verlag, Berlin Heidelberg New York (2005)

[11] Abdullah, K., Lee, Ch., Conti, G., Copeland, J.A.: Visualizing Network Data for Intrusion Detection. Proceedings of the IEEE Workshop on Information Assurance and Security (2002) 100-108

[12] MRTG: The Multi Router Traffic Grapher, http://people.ee.ethz.ch/~oetiker/webtools/mrtg/

[13] Pin, R., Yan, G., Zhichun, L., Yan Ch, Watson, B.: IDGraphs: Intrusion Detection and Analysis Using Histographs. IEEE Workshop on Visualization for Computer Security 2005 (VizSEC 05) (2005) 39-46

[14] Komlodi, A., Rheingans, P., Utkarsha A., Goodall, J.R., Amit J.: A user-centered look at glyph-based security visualization. IEEE Workshop on Visualization for Computer Security 2005 (VizSEC 05) (2005) 21-28

[15] Abdullah, K., Lee, C.P.: Conti, G.; Copeland, J.A.; Stasko, J.: IDS RainStorm: Visualizing IDS Alarms. IEEE Workshop on Visualization for Computer Security 2005 (VizSEC 05) (2005) 01-10

[16] Koike, H., Ohno, K.: SnortView: visualization system of snort logs. Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security. Conference on Computer and Communications Security archive (2004) 143-147

[17] Erbacher, R.F.: Visual Traffic Monitoring and Evaluation. Proceedings of the Conference on Internet Performance and Control of Network Systems II (2001) 153-160

[18] Wooldridge, M.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. Gerhard Weiss (1999)

[19] Corchado, E., Han, Y., Fyfe, C.: Structuring Global Responses of Local Filters Using Lateral Connections. Journal of Experimental and Theoretical Artificial Intelligence, Vol. 15(4) (2003) 473-487

[20] Corchado, E., Fyfe, C.: Connectionist Techniques for the Identification and Suppression of Interfering Underlying Factors. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 17(8) (2003) 1447-1466

[21] Corchado, E., MacDonald, D., Fyfe, C.: Maximum and Minimum Likelihood Hebbian Learning for Exploratory Projection Pursuit. Data Mining and Knowledge Discovery, Vol. 8(3), Kluwer Academic Publishing (2004) 203-225

[22] Fyfe, C., Corchado, E.: Maximum Likelihood Hebbian Rules. Proceedings of the European Symposium on Artificial Neural Networks (2002) 143-148

[23] Seung, H.S., Socci, N.D., Lee, D.: The Rectified Gaussian Distribution. Advances in Neural Informa¬tion Processing Systems, Vol. 10 (1998) 350-356

[24] Vigna, G., Robertson, W., Balzarotti, D.: Testing Network-Based Intrusion Detection Signatures Using Mutant Exploits. ACM Conference on Computer and Communication Security (ACM CCS) (2004) 21–30