MOVICAB-IDS: Visual Analysis of Network Traffic Data Streams for Intrusion Detection

Álvaro Herrero[1], Emilio Corchado[1], and José Manuel Sáiz[1]

[1] Department of Civil Engineering, University of Burgos, Spain
{ahcosio, escorchado, jmsaiz}@ubu.es

Abstract. MOVICAB-IDS enables the more interesting projections of a massive traffic data set to be analysed, thereby providing an overview of any possible anomalous situations taking place on a computer network. This IDS responds to the challenges presented by traffic volume and diversity. It is a connectionist agent-based model extended by means of a functional and mobile visualization interface. The IDS is designed to be more flexible, accessible and portable by running on a great variety of applications, including small mobile ones such as PDA's, mobile phones or embedded devices. Furthermore, its effectiveness has been demonstrated in different tests.

Keywords: Unsupervised Learning, Neural Networks, Exploratory Projection Pursuit, Multiagent Systems, Computer Network Security, Intrusion Detection.

1 Introduction and Previous Work

In the context of a computer network, an IDS (Intrusion Detection System) can roughly be defined as a tool that is designed to detect suspicious patterns that may be related to a network or system attack. To do so, a Network IDS (NIDS) analyses the events occurring along the computer network. Such tools have now become very necessary additions to reinforce security infrastructure.

Many different forms of Artificial Intelligence (such as Genetic Programming [1], Data Mining [2], [3] or Neural Networks [4], [5], [6] among others), and 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. Some of them (The Multi Router Traffic Grapher [12] for example) offer visual measurements of network traffic. MOVICAB-IDS goes further and offers a complete and more intuitive visualization of network traffic by depicting each simple packet and providing the network administrator with a snapshot of network traffic, protocol interactions, and traffic volume, generally in order to identify anomalous situations.

Knowledge discovery, pattern recognition, data mining and other such techniques, deal with the problem of extracting interesting classifications, clusters, associations and other patterns from data. Furthermore, the existence of laptops, palmtops,

handhelds, embedded systems, and wearable computers is making ubiquitous access to a large quantity of distributed data a reality. Advanced analysis of distributed data for extracting useful knowledge is the next natural step in the increasingly interconnected world of ubiquitous and distributed computing.

We have therefore extended our agent-based IDS model [4], [5] to make it accessible from any wireless device, such as a palmtop, laptop or mobile phone to give more accessibility to network administrators, enabling permanent mobile visualization, monitoring and supervision of their networks.

The remaining five sections of this paper are structured as follows: section 2 contains an overview of MOVICAB-IDS; data management is then explained in depth in section 3; some experimental results are described in section 4; the model is evaluated in section 5; finally, section 6 puts forward a number of conclusions and pointers for future work.

2 MOVICAB-IDS

Our model is designed to split massive traffic data sets 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 also provides an analysis of several subsequent segments as unique ones (simple segments) and also as an accumulated data set.

Thus, MOVICAB-IDS (MObile VIsualization Connectionist Agent-Based Intrusion Detection System) may be defined as an IDS formed of different software agents [13] that work in unison in order to detect anomalous situations by taking full advantage of an unsupervised connectionist model [4], [5], [14], [15], [16].

To detect anomalous situations, MOVICAB-IDS performs the following functions:

- 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. (See Sect. 3)
- 3rd step.- Segmentation: the data stream is divided into simple segments and accumulated ones (consisting of the addition of several consecutive simple segments). (See Sect. 3.1)
- 4th step.- Data Analysis: a connectionist model is applied to analyse the data. (See Sect. 2.1)
- 5th step.- Visualization: the projections of both, simple and accumulated segments, are presented to the network administrator for the analysis and monitoring.

The visualization step 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, enabling informed decisions to be taken anywhere and at any time.

2.1 The Unsupervised Connectionist Model

The data analysis task is based on the use of a neural Exploratory Projection Pursuit (EPP) [17], [18] model called Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [14], [15], [16]. It was initially applied in the field of Artificial Vision [14], [15] to identify local filters in space and time. In MOVICAB-IDS it is applied in the field of Computer Network Security. CMLHL is based on Maximum Likelihood Hebbian Learning (MLHL) [19], [20] adding lateral connections [14], [15] which have been derived from the Rectified Gaussian Distribution [21]. The resultant net can find the independent factors of a data set but does so in a way that captures some type of global ordering in the data set.

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 [15], [19], [20] 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.

3 Data Stream and Data Sets

As previously mentioned, NIDSs have to deal with the practical problem of high volumes of quite diverse data [22]. To deal with the problem of high diversity, MOVICAB-IDS splits the traffic into different groups, taking into account the protocol (either UDP, TCP, ICMP…) over IP. For the sake of simplicity, only UDP traffic is considered in this work due to its potential dangers.

Once the data set is classified by the protocol over IP, our model is based on the analysis of five main numerical variables (timestamp, source and destination port,

packet size and protocol) existing on the packets headers. The capability of these variables to identify different anomalous situations has already been demonstrated [4], [5].

Then, MOVICAB-IDS divides the pre-processed data sets as follows:

- Equal simple segments. Each simple segment contains all the packets whose timestamps are between its initial and final limits. As can be seen in Fig. 1, there is a time overlap between each consecutive simple segments. This is done because anomalous situations could conceivably take place between simple segment S_x and S_{x+1} (the next segment following S_x). In this case, it would be necessary to consider some packets twice in order to visualize the end of the anomalous situation and the evolution between simple segments. Both the length (time duration) of the simple segments and the overlap time can be set up by the administrator.
- Accumulated segments. Each one of these segments contains several consecutive simple ones (Fig. 1). The main considerations are, firstly, to present a long-term picture of the evolution of network traffic to the network administrator and, secondly, to allow the visualization of attacks lasting longer than the length of a simple segment. The number of simple segments making up the accumulated segments is configurable.

3.1 Fragmentation

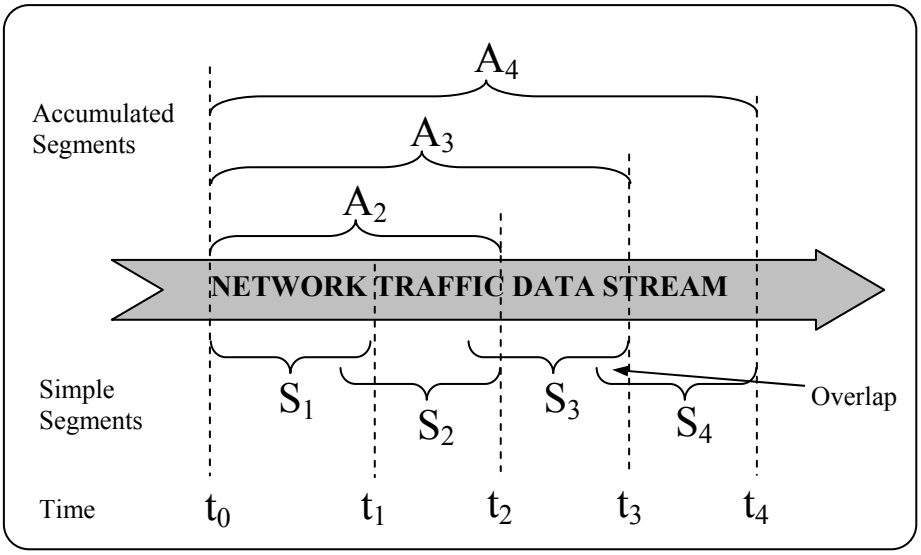

Fig. 1. Data stream fragmentation. Each data set is divided into several simple segments (e.g. S_1, S_2 and so on) and accumulated ones (e.g. A_2, A_3 …).

Fig. 1 shows the fragmentation system used by MOVICAB-IDS. In this study we have fixed a length of 10 minutes for each simple segment, and 2 minutes of overlap between consecutive segments. All these values can be fixed to make the system more suitable for the administrator by taking into account issues such as the traffic volume (packets), the available calculus power, and so on.

Table 1 describes the data sets used in this work.

Table 1. Data sets description

Data set	# packets	Initial timestamp (ms)	Final timestamp (ms)
S_1	3122	1	600000
S_2	3026	480000	1080000
S_3	3052	960000	1560000
S_4	9673	1440000	2040000
S_5	10249	1920000	2520000
S_6	3584	2400000	3000000
S_7	3051	2880000	3480000
S_8	2818	3360000	3960000
…			
A_2	5553	1	1080000
A_3	8036	1	1560000
A_4	17079	1	2040000
A_5	20227	1	2520000
A_6	23169	1	3000000
A_7	25450	1	3480000
A_8	27787	1	3960000
…			
A_{13}	49464	1	6360000

Datasets from A_2 to A_{13} have been developed to show the evolution of accumulated segments from the starting point of the capture.

Two main anomalous situations are distributed throughout different segments in these data sets. These situations can be very dangerous and are related to Simple Network Management Protocol (SNMP) [4], [5]: a network scan (a sweep to two different destination port to check whether SNMP service is active) and MIB (Management Information Base) information transfers. The MIB stores potentially sensitive information on elements controlled by the SNMP.

4 Experiments and Results

To perform the following experiments, we have used a very powerful server. It is equipped with 64 Hyper-Threading Xeon processors and 12 GB of memory.

Depending on the protocol, MOVICAB-IDS uses different colours and shapes to depict the packets, leading to a more intuitive visualization for the administrator.

The data sets were generated ‘made-to-measure’ and are known. They 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.

In the following figures (Fig. 2 and Fig. 3), we show some examples of how our system performs when applied to simple segments of 10 minutes. Fig. 2 (for data set S_1) is an example of normal traffic with no anomalous situations as all the packets

evolve in "normal" parallel directions over time [4], [5]. On the other hand, Fig. 3 (for data set S_4) shows how the system identifies an anomalous situation related to a MIB information transfer [4], [5]. This situation (Groups 1 and 2 in Fig. 2) is identified as anomalous due to its high temporal concentration of packets in comparison to a "normal" one [4], [5], that is visualized as smooth, straight lines running in parallel to each other as can be seen in Fig. 2.

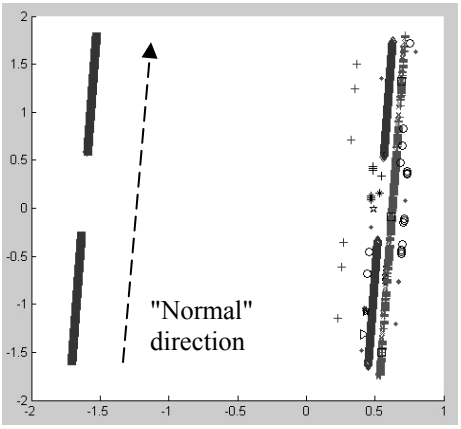

Fig. 2. Visualization of S_1.

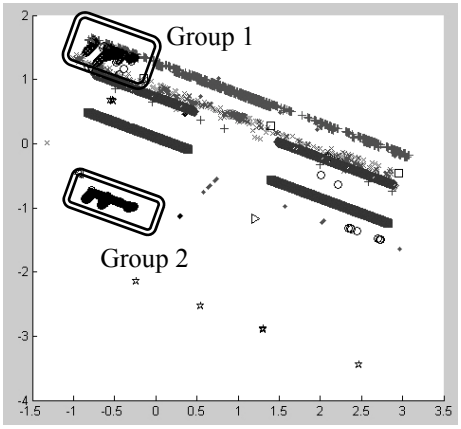

Fig. 3. Visualization of S_4.

The following two figures show the evolution of a 20 minute-long accumulated segment (A_2 - Fig. 4) and then an 80 minute-long one (A_8 - Fig. 5). As can be seen, the same network scan can be identified in both data sets in which it is contained (Groups 1 and 2 in Fig. 4, and Group 1 in Fig. 5). Additionally, A_8 includes a MIB information transfer (Groups 2 and 3 in Fig. 5).

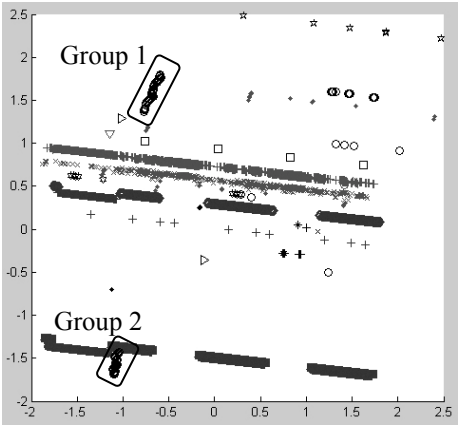

Fig. 4. Visualization of A_2.

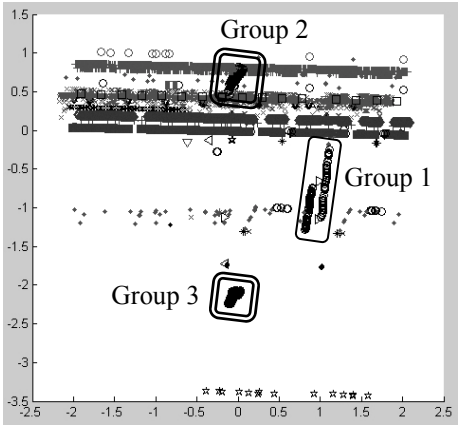

Fig. 5. Visualization of A_8.

Finally, an example of the visualization step for A_{13} is shown in Fig. 6. This data set includes both a network scan (Group 1) and two MIB information transfers (Groups 2-3 and 4-5). An emulator was used to test the visualization on a mobile platform.

Fig. 6. MOVICAB-IDS visualization of data set A_{13}.

In these examples, the administrator can easily identify a network scan represented by its evolution along a non-parallel direction to the normal one while an MIB transfer is characterized by its high packet density.

5 Evaluation

Up until the present, there has been no specific evaluation technique for numerical IDSs. We have therefore used a novel mutation-based method to evaluate the performance of MOVICAB-IDS. In general, a mutation can be defined as a random change. In keeping with this idea, this evaluation modifies different features of the numerical information. Thus, both the destination ports and the number of packets (included in the scan) in data set A_2 have been mutated. As can be seen in Fig. 7,

MOVICAB-IDS detects the mutated scans (Groups 1 and 2). Once again, these anomalous situations are identified by their evolution along a non-parallel direction that intersects with the normal one that represents the rest of the traffic. The goal is to test the system in real-life situations that differ from those used to train the model and which might be generated by a hacker.

Moreover, the statistical technique known as Principal Component Analysis (PCA) [23] was applied to data set S_4 (as can be seen in Fig. 8) for comparison purposes. This technique, already used in the field of IDSs [9], failed to detect the MIB information transfer contained in the data set.

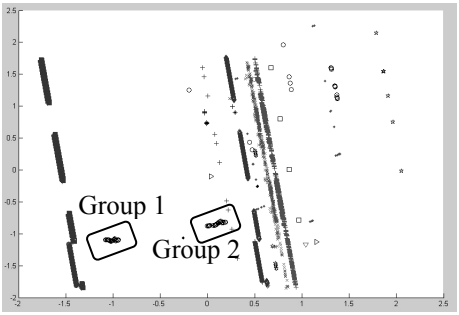

Fig. 7. Visualization of mutated A_2.

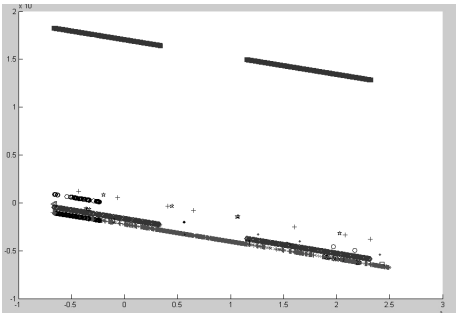

Fig. 8. PCA visualization of S_4.

6 Conclusions and Future Work

We have presented an IDS which offers network administrators greater accessibility using any mobile devices to its visualization features. The system can deal with a high-volume network traffic data stream by pre-processing and splitting it into simple and accumulated segments. Simple segments are characterized by a time overlap with the preceding and the following simple segment in order to prevent any short anomalous situation passing by unnoticed at the very end or at the beginning of a simple segment. To provide a continual analysis of the network traffic, this IDS also studies accumulated segments for more general purposes to monitor and analyse traffic. To achieve this, we have performed the experiments using a supercomputer, which allow us the possibility of increasing the segment length.

This system can be used in combination with other IDS to overcome their limitations (e.g: identification of 0-day attacks).

Further work will focus on the study of different anomalous situations to extend the model to cover several protocols, and the application of different learning rules in the Analysis Step.

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 Transactions on Information and System Security (TISSEC), Vol. 3(4). ACM Press, New York (2000) 227 – 261
4. Herrero, A., Corchado, E., Sáiz, J.M.: A Cooperative Unsupervised Connectionist Model Applied to Identify Anomalous Massive SNMP Data Sending. Proceedings of the International Conference on Natural Computation (ICNC). Lecture Notes in Computer Science, Vol. 3610. Springer-Verlag, Berlin Heidelberg New York (2005) 778-782
5. Corchado, E., Herrero, A., Sáiz J.M.: Detecting Compounded Anomalous SNMP Situations Using Unsupervised Pattern Recognition. Proceedings of the International Conference on Artificial Neural Networks (ICANN 2005). Lecture Notes in Computer Science, 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, T.: 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. Wooldridge, M.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. Gerhard Weiss (1999)
14. 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
15. 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
16. Corchado, E., Corchado, J.M., Sáiz, L., Lara, A.: Constructing a Global and Integral Model of Business Management Using a CBR System. Proceedings of the 1st International Conference on Cooperative Design, Visualization and Engineering (CDVE). Lecture Notes in Computer Science, Vol. 3190. Springer-Verlag, Berlin Heidelberg New York (2004) 141-147
17. Friedman J., Tukey. J.: A Projection Pursuit Algorithm for Exploratory Data Analysis. IEEE Transaction on Computers, Vol. 23 (1974) 881-890
18. Hyvärinen A.: Complexity Pursuit: Separating Interesting Components from Time Series. Neural Computation, Vol. 13(4) (2001) 883-898
19. 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
20. Fyfe, C., Corchado, E.: Maximum Likelihood Hebbian Rules. Proceedings of the European Symposium on Artificial Neural Networks (2002) 143-148
21. Seung, H.S., Socci, N.D., Lee, D.: The Rectified Gaussian Distribution. Advances in Neural Information Processing Systems, Vol. 10 (1998) 350-356
22. Dreger, H., Feldmann, A., Paxson, V., Sommer, R.: Operational Experiences with High-Volume Network Intrusion Detection. Proceedings of the ACM Conference on Computer and Communications Security. ACM Press, New York, USA. (2004) 2-11
23. Oja, E.: Neural Networks, Principal Components and Subspaces. International Journal of Neural Systems, Vol. 1 (1989) 61-68