A Framework for Analysis of Anomalies in the Network Traffic

L.S. Silva[1], T.D. Mancilha[2], J.D.S. Silva[3], A.C.F. Santos[4], e A. Montes[5]

[1,2]Ground Segment Development Division - DSS
[3,4]Laboratory for Computing and Applied Mathematics – LAC
Brazilian National Institute for Space Research - INPE
C. Postal 515 – 12245-970 – São José dos Campos – SP - BRASIL
[5]Research Center Renato Archer – CenPRA
Rodovia Dom Pedro I, km 143,6 - 13069-901- Campinas – SP – BRASIL

E-mail: lilia@dss.inpe.br, thiago@dss.inpe.br, demisio@lac.inpe.br, adriana.ferrari@lac.inpe.br, antonio.montes@cenpra.gov.br

Keywords: anomaly detection, anomaly analysis, network traffic analysis, network traffic visualization, traffic anomalies.

Abstract

In this paper, a framework developed at INPE for analysis of anomalies in the network traffic is presented. The stages of this work, including the environment preparation for tests and development, attribute selection, data reduction and data visualization are described. Also, techniques and tools used in each stage will be approached as well as some results obtained from the applied techniques will be discussed.

1. Introduction

Tools for analyzing network traffic allow the detection of anomalies in the environment, including attacks and unusual events in the network, and enable fast execution of actions to avoid that the detected threats can propagate through the network.

The network traffic should be monitored in regular intervals to obtain data that will be analyzed by statistical or intelligent techniques in search of anomalies. The idea is storing data from normal traffic (historical data) for future comparison with real traffic (current data) to detect eventual anomalies. By observing the traffic and correlating it to its previous states, it may be possible to see whether the current traffic is behaving in a similar/correlated manner [12]. This approach is named anomaly detection.

In this paper, information about computer network data, including network session and session attributes will be described in the second section. A classification for network anomalies will be explained in the section 3. Network traffic data from session TCP/IP packets, more specifically, deriving from HTTP communication between client and server machines are explored in this work. These data belong to large datasets, they are found in several types, and stored in different scale and measure units. Considering this context, the use of techniques to reduce the very large dataset and tools to present data on the computer screen are necessary. The framework designed to detect network traffic anomalies and techniques used at INPE to reduce the traffic data volume for analysis as well as the tools applied to visualize the traffic will be presented in the section 4. Finally, conclusions of this work and next challenges will be approached in the section 5.

2. TCP/IP Network Traffic Data

Computer network traffic consists of packet arrival processes. However, given the huge number of packets involved in any computer network traffic, this would result in huge data sets.

Data packets travelling in the network carry useful information among the interconnected computers. A TCP/IP network packet has three parts [17] as shown in the Figure 1:

- IP (Internet Protocol) header;
- TCP (Transfer Control Protocol), UDP (User Datagram Protocol) or ICMP (Information Control Message Protocol) header, depending on the encapsulated transport protocol, and
- TCP (Transfer Control Protocol), UDP (User Datagram Protocol) or ICMP (Information Control Message Protocol) payload, depending on the network service in usage. For example, HTTP application uses TCP header and payload, while packets generated by Telnet network service contain UDP header and payload.

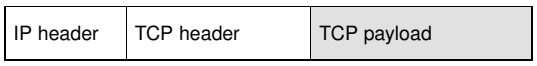


Figure 1. TCP/IP packet format

Payload data that are sent from source to destination computer are located in the packet payload field. Several strings in ASCII or hexadecimal or in both formats are found in the payload data. Among these strings, normal or malicious strings may exist. These are known attack patterns used by attackers to launch effective or attempted attacks and to discover machine vulnerabilities.

These malicious strings are known as attack signatures [14][16] and are found stored in signature-based intrusion detection systems, such as the popular Snort [14][15], which maintains its own signature base (Snort signatures) in permanent and reliable updating.

Intrusion detection methods are well-suited to collect and analyse detailed packet payload data. As such, many data mining methods proposed to detect intrusions rely on detailed data to mine for anomalies [18]. In contrast to work in anomaly detection with packet payload data, the objective of this work is to diagnose network anomalies using sampled packet header data. Data extracted from the network packet header are used for mapping the network traffic in the work in development at INPE.

A traffic attribute or feature is a field in the header of a packet (a primitive attribute) or can be constructed with basis on the primitive attributes (derived attributes) [5] [10]. Primitive attributes are directly obtained from packet header, such as: source and destination IP addresses, source and destination ports, and service type in use. Derived attributes carry stronger semantically information for the traffic mapping. In the work accomplished by the Minnesota University [10] three groups of features were used for their experiments: “content-based features”, extracted from raw *tcpdump* data using *tcptrace* software and “connection-based features” and “time-based features”, constructed by the researchers.

In this paper, nine attributes presented in the section 4.3 were used. Each set of nine attributes is stored as a register in a database and represents a unique network session or a connection record. A network session (or connection) can be defined as any sequence of packets characterizing an information exchange between two IP addresses, that contains information of beginning, middle and end, even so all communication be resident in a unique packet [2]. In the literature also is found the following definition: a connection is a sequence of TCP packets starting and ending at some well defined times, between which data flows to and from a source IP address to a target IP address under some well defined protocol [13].

Summarily, ‘network session’ can be uniquely identified by the combination of network data attributes. And a set of thousands network session represents ‘network traffic’ observed in different periods of time.

3. Network Traffic Anomalies

Anomalies on the network traffic can be defined as previously unseen (yet legitimate) traffic behaviors. A wide range of unusual events – some of which, but not all, may be malicious – known as traffic anomalies are commonplace in today’s computer networks [9]. Identifying, diagnosing and treating anomalies such as failures and attacks in a timely fashion are a fundamental part of day to day network operations.

Operators need to detect these anomalies as they occur and then classify them in order to choose the appropriate response. The principal challenge in automatically detecting and classifying anomalies is that anomalies can span a vast range of events: from network abuse (e.g., DoS attacks, scans, worms) to equipment failures (e.g., outages) to unusual customer behavior (e.g., sudden changes in demand, flash crowds, high volume flows), and even to new, previously unknown events. A general anomaly diagnosis system should therefore be able to detect a range of anomalies with diverse structure, distinguish between different types of anomalies and group similar anomalies. This is obviously a very ambitious goal [18].

Regardless of whether the anomalies in question are malicious or unintentional, it is important to analyze them for two reasons [11]:

- anomalies can create congestion in the network and stress resource utilization in a router, which makes them crucial to detect from an operational standpoint;
- some anomalies may not necessarily impact the network, but they can have a dramatic impact on a customer or the end user.

A significant problem when diagnosing anomalies is that their forms and causes can vary considerably: from

Denial of Service (DoS) attacks, to router misconfigurations, to the results of BGP policy modifications.

Despite a large literature on traffic characterization, traffic anomalies remain poorly understood. One of the reasons for this is that to identify anomalies requires a sophisticated monitoring infrastructure [11]. But the most ISPs (Internet Service Providers) only collect simple traffic measures, e.g., average traffic volumes, using SNMP. More adventurous ISPs do collect flow counts on edge links, but the processing the collected data is a demanding task. A second reason for the lack of understanding of traffic anomalies is that ISPs do not have tools for processing measurements that are fast enough to detect anomalies in real time. Thus, ISPs are typically aware of major events (worms or DoS attacks) after the fact, but are generally not able to detect them while they are in progress. A final reason is that the nature of network-wide traffic is high-dimensional and noisy, which makes it difficult to extract meaningful information about anomalies from any kind of traffic statistics.

An important challenge therefore is to determine how best to extract understanding about the presence and nature of traffic anomalies from the potentially overwhelming mass of network-wide traffic data [18]. A considerable complication is that network anomalies are a moving target. It is difficult to precisely and permanently define the set of network anomalies, especially in the case of malicious anomalies. New network anomalies will continue to arise over time; so an anomaly detection system should avoid being restricted to any predefined set of anomalies.

The current best practices for identifying and diagnosing traffic anomalies consist of visualizing traffic from different perspectives and identifying anomalies from prior experience. Different tools have been developed to automatically generate alerts to failures, but to automate the anomaly identification process remains a challenge [9].

Anomaly detection can be based on machine learning. The normal traffic behavior is modeled by a systematic method. The system traces significant deviation between monitored network traffic activities and the built model.

3.1 Known Traffic Anomalies

Traffic anomalies can be categorized [9] in four classes as follows:

- **Network**: network failure event or temporary misconfigurations resulting in a problem or outage. For instance: router software spontaneously stopped advertising one of the campus class B networks to campus BGP peers.
- **Attack**: Typically a Denial-of-Service event, usually flood based. For instance: an outbound flood of 40-byte TCP packets from a campus host that has had its security compromised and is being remotely controlled by a malicious party.
- **Flash**: A flash crowd [18] event. For instance: the increase in outbound traffic from a campus `ftp` mirror server following a release of RedHat Linux.
- **Measurement**: An anomaly that we determined not to be due to network infrastructure problems nor abusive network usage. For example: a campus host participating in TCP bulk data transfer with a host at another campus as part of a research project. Problems with the data collection infrastructure itself were also categorized as Measurement" anomalies. These include loss of flow data due to router overload or unreliable UDP NetFlow transport to the collector.

3.2 Attack Classes

In several papers [10][22], the anomalies classified as attacks fall into four main categories:

- **DoS (Denial of Service):** a class of attacks in which an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine. Examples are Apache2, Back, Teardrop and Smurf. Overloading servers, creation of mal-formed packets and exploration of service bugs to interrupt services are caused by this kind of attack.
- **Probing**: surveillance and other probing - a class of attacks in which an attacker scans a computer network to gather information or find known vulnerabilities. An attacker with a map of machines and services that are available on a network can use this information to look for exploits. Examples area Ipsweep, Nmap, Satan.
- **R2L (Remote to Local):** unauthorized access from a remote machine - a class of attacks in which an attacker sends packets to a machine over a network but who does not have an account on that machine; exploits some vulnerability to gain local access as a user of that machine. Examples are Dictionary, Ftp_write, Guest, Named.

- **U2R (User to Root):** unauthorized access to local super user (root) privileges - a class of attacks in which an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system. Examples are buffer overflow, Eject, Fdformat, Xterm.

4. Framework for Anomaly Analysis

The framework developed at INPE for analysis of anomalies in the network traffic comprises the following stages: tests and development environment preparation, data collection, attribute selection, data reduction and data visualization. Techniques and tools used in each development stage of this work will be approached in the next topics as well as some results obtained from the applied techniques will be discussed.

The goal is to use the implemented framework to analyze anomalies in the traffic behavior of two networks, production network and test network environment.

4.1 Development and Test Environment

The environment used to develop and test programming codes for network traffic analysis tasks contains hardware and software resources. Software resources includes: Java programming environment, databases, software for capture of packets, for reconstruction of network sessions, for data filtering and reduction and for graphical visualization of data. A computer for monitoring and capturing data operation is used and another machine for development and tests of applications is available. Additionally, server machines are appropriately installed for test purposes.

4.2 Data Collection

Data collected from two networks, production and controlled network, are used in the classification tests. The controlled network is a network used for launching of simulated attacks, traffic monitoring and tests. The production network involved in this work is an internal network at INPE whose data are captured by means of a network sensor installed outside its boundary and observed [4], as shown on Figure 2.

Traffic of both networks are monitored in ten minutes time intervals. In each time-window, data from packets passing through the networks are automatically recorded by sniffer tools like *tcpdump*, *ethereal* and scripts for data writing. In following, network packets are remounted in network session data and stored in database using the Recon system [2].

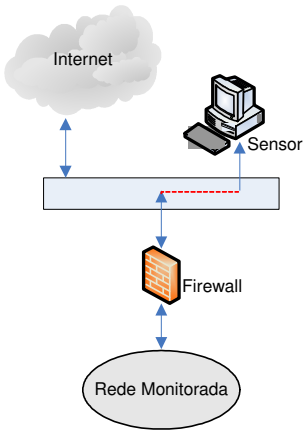


Figure 2: Network sensor placing in the monitored network

4.3 Attribute Selection

The analysis of traffic feature distributions is a powerful tool for the detection and classification of network anomalies. Many important kinds of traffic anomalies cause changes in the distribution of IP addresses or ports observed in the traffic [18]. For example, Table 1 lists a set of anomalies commonly encountered in backbone network traffic [18] by using four features: source and destination addresses and source and destination ports. Each of these anomalies affects the distribution of certain traffic features. In some cases, feature distributions become more dispersed, as when source addresses are spoofed in DoS attacks, or when ports are scanned for vulnerabilities. In other cases, feature distributions become concentrated on a small set of values, as when a single source sends a large number of packets to a single destination in an unusually high volume flow.

Table 1 - Qualitative effects on feature distributions by various anomalies.

Anomaly Label	Definition	Traffic Feature Distributions Affected
Alpha Flows	Unusually large volume point to point flow	Source address, destination address (possibly ports)
DOS	Denial of Service Attack (distributed or single-source)	Destination address, source address
Flash Crowd	Unusual burst of traffic to single destination, from a "typical" distribution of sources	Destination address, destination port
Port Scan	Probes to many destination ports on a small set of destination addresses	Destination address, destination port
Network Scan	Probes to many destination addresses on a small set of destination ports	Destination address, destination port
Outage Events	Traffic shifts due to equipment failures or maintenance	Mainly source and destination address
Point to Multipoint	Traffic from single source to many destinations, *e.g.,* content distribution	Source address, destination address
Worms	Scanning by worms for vulnerable hosts (special case of Network Scan)	Destination address and port

Thus, it can be noticed that the selection of relevant network attributes or features is an important role in the anomaly detection processes. Nine traffic attributes are selected and handled in this work, considering the kind of attacks to be launched on and detected in the network environment: medium size of network packets received by the client (in bytes); medium size of network packets received by the server (in bytes); number of packets received by the client; number of packets received by the server; small packet rate or packet with size less than 130 bytes (%); traffic direction; data bytes received by the client; data bytes received by the server; and session duration. All results in this work are based on analysis of these nine attributes from each session of the network traffic.

Recon system [2] is being used for reconstruction of network sessions, selection of attributes from the network traffic data and storage of session attributes in database as illustrated in the Table 2.

Table 2 – Data sample of attributes stored in a database table

codigo	hora	psize_cl	psize_sv	pnum_cl	pnum_sv	smallpkt	data_dir	brecv_cl	brecv_sv	duration
7000	08_20	129.6	174.4	5	5	0.8	0	648	872	0.084135
7001	08_20	443	131.67	6	6	0.75	0	2658	790	0.028824
7002	08_20	132.6	188.4	5	5	0.8	0	663	942	0.030568

4.

In order to reduce the traffic data, clustering techniques based on neural networks are being applied. In the first a SOM network was used. Traffic sessions with similar behaviours participate of the same cluster. After this, data clusters are exhibited in a graphical interface using the Matlab tool, as the example of the Figure 3.

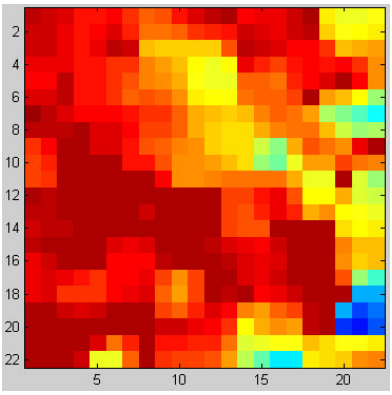


Figure 3: Clustering of network sessions for data reduction

In this test, data from the production network referent to the 0:00-23:59h time period was used. The characteristics of the SOM network used in the tests are: neurons matrix dimension= 22x22, neighborhood initial dimension=7, cluster similarity level=98%, gauss function and learning rate=0,5.

In this figure, different colors represent clusters of network sessions. From a total number of 800.000 sessions introduced for clustering, 655 session clusters were generated, that represents a significant data reduction.

4.4 Data Traffic Visualization

Visualization techniques are ways of creating and handling graphical representations of data. These representations are used in order to obtain better insight and understanding of the problem in study, because pictures can convey an overall message much better than a list of numbers [23]. There are several visualization techniques, such as: line graphs, histogram, bar chart, surface view, image display, scatter plot, isosurfaces, volume rendering and multiple line graphs with parallel coordinates [24], among others. The selection of the better visualization technique will depend on the type of data to be analyzed.

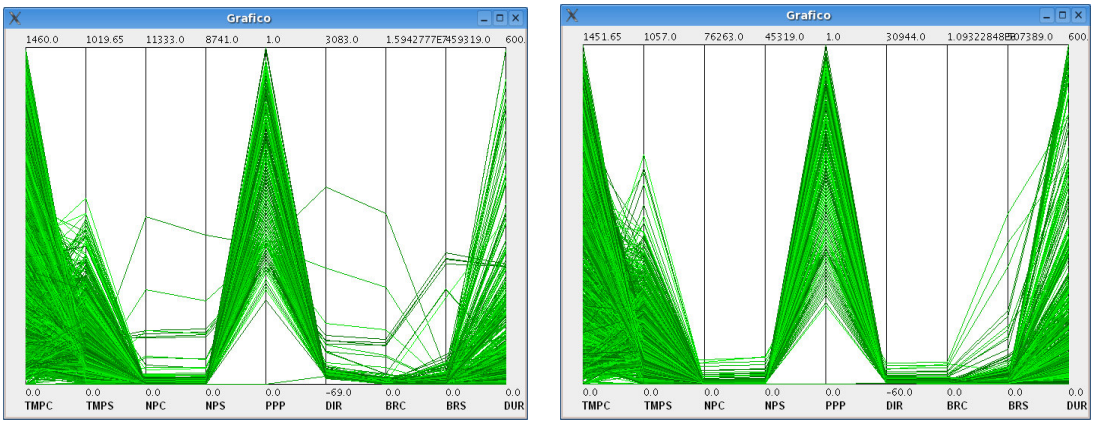
Some work has been conducted in order to graphically represent the network traffic [8][9][12][25]. A network traffic visualization tool named RGCom [1] is being developed at INPE. This application performs data reading from a database, data normalization, and data plotting in parallel coordinates on the computer screen. Graphical and database communication resources from Java programming environment are used in its implementation.

The graph produced by RGCom contains nine parallel coordinates with values of a determined attribute each one. Nine attribute points of a same session are plotted in that axis and they are interconnected, shaping a session line. All lines of one happened session in a selected by the user date and time interval are drawn and the resultant graph represents the network traffic behavior in that time.

The graphical interface of RGCom with options for users to select the session data table, time interval and attributes, is presented in the Figure 4.

Figure 4: RGCom Graphical Interface

Some graphical results from RGCom application on production network data, at two thirty-minutes interval of captured traffic data, are shown in the Figure 5.



(a) (b)

Figure 5: Graphical representation of the network traffic in two intervals of thirty-minutes

The figure 5(a) presents some points in eminence, such as a high rate of NPC (Number of Packets received by the Client) attribute values, indicating a possible anomaly in the network to be investigated, for example, data download or radio station listening. The figure 5(b) shows a high BRS (Bytes Received by the Server) rate, indicating a supposed data uploading to a server machine.

5. Conclusion

The mentioned framework for analysis of anomalies in the network traffic is implemented and operations of development and tests are being conducted in the proposed environment. Among these operations, traffic data are captured, filtered and stored in database. Clustering techniques based on neural network SOM are used for data reduction. The great advantage of the SOM clustering is to allow analysis of the medium behavior of the monitored traffic. RGCom tool is also being used. This application does not perform data reduction, but represents graphically the network traffic in given period of time. At moment, the most important objective of this tool is to provide the selection of attribute combinations that better describe a network traffic behavior, making the anomaly identification a task easier.

Next challenges involves to improve the RGCom tool, adding options such as selection of attributes and analysis of traffic in time-windows. For the SOM application, analysis of the session clusters generated have to be accomplished, the data clustering process will be changed in order to get best results and analysis of the medium behaviour of the network traffic need be performed. A significant data reduction was found with this technique, but the formed clusters have to be verified if they satisfactorily represent all traffic sessions with similar characteristics. Also, an efficient technique to minimize the network attribute set is a goal to be achieved. Besides, it is intended to model normal patterns and create routines to classify new traffic data based on this models.

References

[1] Mancilha, T.D, Silva, L.S, Salgado, A.E.M, Montes, A. and Paula, A. R. (2006), *Desenvolvimento em Java de uma Ferramenta de Visualização Gráfica do Tráfego de Rede*, X Encontro Latino Americano de Iniciação Científica – Universidade do Vale do Paraíba, São Jose dos Campos, SP.

[2] Chaves, M.H.P (2002), *Analise de Estado do Trafego de Redes TCP/IP para Aplicação em Detecção de Intrusão*, Dissertation for Master Degree in Applied Computing, INPE, São Jose dos Campos, SP.

[3] Chaves, C.H.P.C and Montes, A. (2005), *Detecção de Backdoors e Canais Dissimulados*, V Workshop dos cursos de Computação Aplicada (Worcap'2005), INPE, São José dos Campos, SP.

[4] Silva, L.S., Montes, A. and Silva, J.D.S (2005), *Evolução dos Trabalhos em Detecção de Anomalias na Rede*, V Workshop dos cursos de Computação Aplicada (Worcap'2005), INPE, São José dos Campos, SP.

[5] Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Dokas, P., Srivastava, J. and Kumar, V. (2004), *Detection and Summarization of Novel Network Attacks Using Data Mining*, available in: http://www.cs.umn.edu/research/minds/papers/raid03.pdf, accessed on july 2004.

[6] Kim, S.S. and Reddy, A. L. N. (2005), *A Study of Analyzing Network traffic as Images in Real-Time*, Department of Electrical Engineering ,Texas A&M University.

[7] Kim, S.S.; Reddy A. L. N. and Vannucci M. (2004), *Detecting Traffic Anomalies through Aggregate Analysis of Packet Header Data*, Proceedings of Networking 2004, LNCS 3042, pp 1047-1059, Athens, Greece.

[8] Kim, S.S. and Reddy, and A. L. N. (2005), *Modeling Network traffic as Images*, Proceedings of IEEE International Conference on Communications, Seoul Korea.

[9] Barford P., Kline J.; Plonka D. and Ron A. (2002), *A Signal Analysis of Network Traffic Anomalies*, Proceedings of ACM SIGCOMM Internet Measurement Workshop, Marseille, France.

[10] Lazarevic, A., Ertoz, L., Ozgur, A, Srivastava, J., and Kumar, V. (2003), *A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection*, Proceedings of Third SIAM Conference on Data Mining, San Francisco.

[11] Lakhina, A., Crovella M., and Christophe, D. (2004), *Diagnosing Network-Wide Traffic Anomalies*, Proceedings of the ACM SIGCOMM 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, Portland, Oregon, USA.

[12] Kim, S.S.; Reddy A. L. N. and Vannucci M. (2004), *Detecting Traffic Anomalies through Aggregate Analysis of Packet Header Data*, Proceedings of Networking 2004, LNCS 3042, pp 1047-1059, Athens, Greece.

[13] Rao, X., Dong, C., and Yang, S. (2003), *Statistic Learning and Intrusion Detection*, RSFDGrC 2003, LNAI 2639, pp. 652659.

[14] Caswell, B., J. Beale, J. C. Foster, and J. Posluns (2003), Snort 2 - *Sistema de Detecção de Intruso Open So*urce, Editora Alta Books, Rio de Janeiro.

[15] SNORT - http://www.snort.org/, page accessed on jan 2006.

[16] Northcutt, S., and J. Novak (2002), *Network Intrusion Detection*, Third Ed., New Riders.

[17] Stevens, W.R (2001), *TCP/IP Illustrated Vol. 1 The Protocols*, Addison Wesley, Indianapolis, USA.

[18] Lakhina, A., Crovella M., Christophe, D. (2005), *Mining Anomalies Using Traffic Feature Distributions*, Proceedings of the ACM SIGCOMM 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, Philadelphia, Pennsylvania, USA.

[19] Barford P. and Plonka D. (2001), *Characteristics of Network Traffic Flow Anomalies*, In Proceedings of ACM SIGCOMM Internet Measurement Workshop, San Francisco, CA.

[20] Ari I., Hong B., Miller E. L., Brandt S. A., Long D. E. (2004), *Modeling, Analysis and Simulation of Flash Crowds on the Internet*, Technical Report UCSC-CRL-03-15, University of California, Santa Cruz, CA.

[21] Bearavolu R., Lakkaraju K., Yurcik W., Raje H. (2003) *A visualization tool for situational awareness of tactical and strategic security events on large and complex computer networks*. In IEEE Military Communications Conference (Milcom), Urbana, IL, USA.

[22] Mukkamala, S. Janoski, G. and Sung, A. (2002), *Intrusion Detection Using Neural Networks and Support Vector Machines*, Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02), Honolulu, HI, USA.

[23] Henderson, S.J and Brodlie, K. (1996), *VisEd Visualization Education Tool*, http://www.siggraph.org/education/materials/HyperVis/vised/VisTech/vtmain.html accessed in 12/10/2006.

[24] Nascimento, H. A. D. and Ferreira, C. B. R. (2006), *Visualização de Informações – Uma Abordagem Prática. Coordenadas Paralelas*, pagina http://www.inf.ufg.br/funcomp/infovis/topico7.html acessada em 16/10/2006.

[25] Plonka D. (2000), *FlowScan: A Network Traffic Flow Reporting and Visualization Tool*, Proceedings of the USENIX 14th System Administration Conference, New Orleans, LA.