Secure Computer Network: Strategies and Challengers in Big Data Era

Mercedes Barrionuevo[1], Mariela Lopresti[1], Natalia Miranda[1], and Fabiana Piccoli[1]

[1]*LIDIC, Universidad Nacional de San Luis, San Luis, Argentina*
{mdbarrio, omlopres, ncmiran, mpiccoli}@unsl.edu.ar

Abstract

As computer networks have transformed in essential tools, their security has become a crucial problem for computer systems. Detecting unusual values from large volumes of information produced by network traffic has acquired huge interest in the network security area. Anomaly detection is a starting point to prevent attacks, therefore it is important for all computer systems in a network have a system of detecting anomalous events in a time near their occurrence. Detecting these events can lead network administrators to identify system failures, take preventive actions and avoid a massive damage.

This work presents, first, how identify network traffic anomalies through applying parallel computing techniques and Graphical Processing Units in two algorithms, one of them a supervised classification algorithm and the other based in traffic image processing. Finally, it is proposed as a challenge to resolve the anomalies detection using an unsupervised algorithm as Deep Learning.

Keywords: Computer Network, Network Security, Anomalies and Attacks, Big Data, High Performance Computing, Machine Learning.

1 Introduction

In the last decade, World Wide Web (WWW) together its fast and growing development, has produced changes in information technologies. Although, it has brought great benefits in many areas, it also has some drawbacks.

Even though computer networks provide global connection and access for all information type, also provides malicious users with new tools for their destructive purposes. The costs of temporary or permanent damage caused by unauthorized access to computer systems have prompted organizations to implement complex systems to control the flow of data on their networks.

Threats to a data network are made by a packages set with specific characteristics to detect system vulnerabilities. These hazards represent risks for any organization and can be used to carry out attacks. Detecting possible attacks requires to count with several methods and strategies to classify traffic. This area is a wide interest problem, especially in emerging areas such as big or massive data.

Big Data is not a technology in itself, it is a work approach to obtain information (value) and benefits from large volumes of data. For this, it necessary take into account:

- How to capture, manage and take advantage of data.
- How to secure data and its derivatives, as well as its validity and reliability.
- How to share data to obtain improvements and benefits in the organization.
- How to communicate data (visualization techniques, formats and tools) to facilitate decision making and subsequent analysis.

A Big Data problem is defined through the 7 Vs: Volume, Velocity, Variety, Veracity, Validity, Visualization and Value. Each of these characteristics is briefly detailed in section 2.2.

For each of above aspects, an interdisciplinary work is necessary that includes areas of High Performance Computing (HPC), Image Processing and Machine Learning. The problem of traffic analysis and detection of possible network attacks fit to this kind of work and, in consequence, they can be considered a Big Data problem.

The use of data-intensive applications carry on users to consider new HPC configurations to work with huge amounts of data. In effect, the International Data Corporation (IDC) has incorporated the term High Performance Data Analytics (HPDA) [1] to represent Big Data analysis over HPC configurations. HPDA provides the ability to solve a new class of computational problems of data-intensive.

The objective of this work is to develop models for search of anomalies in the network traffic and detect attacks to data networks. These models search to identify patterns that deviate from normal behavior. For this, it is proposed to analyze the network traffic by

means of different techniques that allow to obtain concrete results and a near time to detect an attack and to make quick decisions.

This paper is organized as follows: the next section describes theoretical concepts need to this work. Section 3 explains the relation between analysis of network traffic and big data problems. In section 4, the main characteristics of proposed model are detailed, and in Section 5, some experimental results of performance are sketched. Finally, the conclusion and future works are drawn.

2 Background

This work involves many concepts, among them we emphasize computers networks traffic and its anomalies, extraction methods of packages characteristics, Machine Learning algorithms and Big Data. In this section, we describe each one of them.

2.1 Computer Networks Data Traffic

Network traffic provides information about what travels by network. The most common data types are log data, such as Internet Protocol (TCP/IP) records, event logs, internet access data, Network Management Protocol (SNMP) data reporting, among others [2]. This information is necessary for network security, specifically for anomalous events detection. Fig. 1 illustrates an example of TCP/IP traffic, the rows detail individual network traffic and the columns are specific characteristics of each traffic. In the example, the first column is a session index for each connection, and the second says when the connection has occurred [2].

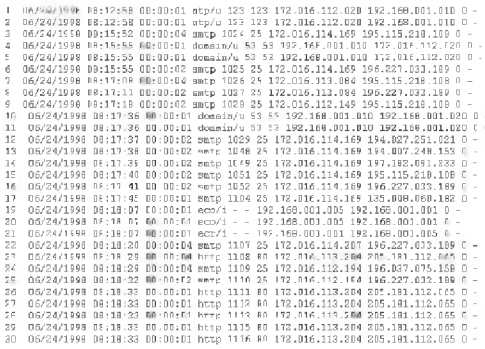

Figure 1: Example of TCP/IP Traffic.

Data traveling on network can provide important information about user and system behaviors. These data can be collected with some commercial products or specific software, for example TCP/IP data can be captured using different tools, called sniffers. Network traffic is composed of packets, flows and sessions. A packet is a data unit exchanged between a source and a destination on the Internet or another TCP/IP-based network; a network flow is a one-way packets sequence between two endpoints; and the session data represents communication between computers. A communication involves the interchange of multiple flows. Traditionally, an IP flow or tuple contains a set of attributes, for this work, the more important are: source and destination IP address, source and destination port, and protocol type. The protocol type, if you consider the 4-layer TCP/IP model, can be TCP or UDP (layer 3) or ICMP (layer 2). This information allows to establish a behavior baseline or normal pattern of network traffic, and in consequence to identify unexpected or unwanted conduct, called anomalous traffic. Therefore, an analysis strategy by anomalies bases on the traffic description in normal conditions to classifies as anomaly all patterns that move away from it. In order to obtain this dataset, there are several techniques, some of which are mentioned in the following subsection.

2.1.1 Analysis of a Network Packet

Studying the particular aspects of network traffic, it is necessary to extract only information from data packets and then process them. There are different techniques of extraction and processing, some of them are:

- Graphical representation of raw data: Generally, the representations are 2D and 3D scatter graphics, time-based graphics, histograms, pie charts, or diagrams.
- Statistical information and pattern extraction: They are based on average calculations, time distributions and probability distribution functions.
- Analysis based in rule (signatures), anomaly detection and policy: All traffic inspection analyzer that look for coincidences with a particular rule or signature belong this category. Rules are defined as values for certain fields in the header or a combination of several of them. These techniques are used in intrusion detection systems (IDS), such as Snort [1].
- Flow-based analysis: It focus on network traffic management as flow. The most of network information exchanged is oriented to connection (and non-oriented to packet), the analysis can take advantage of this. A clear example of typical network flow is a TCP connection, where the data exchanged are governed by the TCP state machine [3].

Each of these techniques is suitable for specific situations, also it is possible combine them. This work is based on flow analysis and rule-based analysis.

[1] www.snort.org

2.1.2 Usual Attacks

One of the biggest challenges for network administrators is to detect attacks on computer networks. An attack implies to take advantage of a computer system vulnerability (operating system, application software, or user's system) for unknown purposes but, usually, causing damage. Therefore, it is impossible to make a complete classification of all the actual attacks and possible weaknesses of the networks, even more when networks are connected to Internet. The denial of service (DoS) and distributed denial of service (DDos) attacks are of great interest today. A DoS attack comes from a single entity and its goal is to turn out unavailable the resources or services of a computer. There are different types, in particular this work has focused on the DoS attacks: Smurf, Fraggle and Land [4] [5], each one of them has the following characteristics:

- Smurf: This attack uses ICMP protocol to send a broadcast ping with a false source address. There are different ways to do a ping, they are:
  - Normal Ping: One or more ICMP echo requests are sent to a system, which responds with one or more ICMP echo replies. Thus, this operation verifies remote system.
  - Broadcast Ping: This ping sends an ICMP echo request to a broadcast address. Each system responds to sender, flooding it with ICMP echo replies.
  - Broadcast ping with false source: A broadcast ping is sent with victim's source address. Each system in network replies and floods the victim with answers. This operation is a combination of the two previous Ping.

  The pattern to recognize this attack type is to analyze to ICMP protocol, if source and destination IP addresses belong to the same network, and the destination address is a broadcast message.

- Fraggle: When you want to check if a system is working, you can use UDP-based tools instead of ICMP to inspect whether the system is listening by a specific port or not. This is commonly done with different types of vulnerability scans, that are used by attackers or security administrators. For example, if a system listens over Port 19 (TCP or UDP), when a connection is established, the system would respond with a constant character flow. Typically, the source system uses the TCP or UDP Port 7. When the source system begins to receive characters, it knows that the target system is operational and closes the connection. In a Fraggle attack, a broadcast packet is sent with a false address to the victim's port 19, if it has its port 19 open, it answers a constant flow of characters to victim. The pattern is similar to that of Smurf but in this case the protocol is UDP.

- Land: It's an attack using the TCP protocol. It creates an "infinite loop" which is caused by sending a SYN request with the same source and destination IP address. The victim computer responds to itself until a blocking state appears and it does not accept any new requests. Besides, as all processor resources are exhausted, the denial of service happens. To recognize this type of attack, it is necessary to analyze coincidence between the source and destination IP addresses, as well as the same ports.

As mentioned above, there are many other denial of service attacks, but the detection of patterns in them requires a deeper and detailed analysis that escapes this work.

2.2 Big Data

Big Data is defined as a large data set without an implicit organization and a particular structure. The large volume, velocity and variety of data make traditional databases be inadequate to store and quickly retrieve them, their processing capabilities are exceeded. At beginning of the big data era, the main generators of big data were scientific and physical applications, military experiments, simulations, NASA and super computers. Nowadays, there are many sources that contribute to a constant generation of big data, for examples airlines traffic, medical systems of assistance and insurance, stock exchanges, mechanisms and systems of electronic money, mobile applications. In 2012, approximately 500PB of medical care data were generated, 25.000PB will be expected for 2020. According to Stephen Gold, vice-president of Watson (IBM), 90% of all data was created in 2 years ago, 2.5PB are generated per day [6]. He compared Big Data with oil: if it is in the ground, it does not have much value, but it is different when you use it. In consequence Big Data becomes very interesting if you find ways to process and analyze data. There are studies that show how Big Data has contributes to several scopes, such as industries, markets, health, labor market, stock market, retail, real estate, education, finance, environmental research, genomics, sustainability, politics and biological research [7, 8].

Google has become the leader company by its search engine. BigTable, Hadoop and MapReduce are applications of real-world Big Data problems that have revolutionized industry and companies by providing great solutions [7]. Their parallel and distribution computing model provides them the ability to perform complex operations on very large data sets.

By all above said, large data volumes are produced constantly by a lot of applications. Therefore, it is important to understand and discuss all Big Data's V

in order to get information from these large data sets. At present, 7 Vs are considered to Big Data, they are:

1. ***Volume:***: The first V refers to size of data set, it can be measured in TB, PB or more. These data are generated from different sources as social networks, research studies, medical data, spatial images, crime reports, forecasts meteorological, natural disasters, applications, among other.

2. ***Velocity:*** Data are generated at high speed. Smart-phones or the World Wide Web (WWW) contribute with this.

3. ***Variety:*** A data set can contain structured, unstructured data or both. Besides , these data can be different types, for example: records, audios, videos, texts, images, etc. This feature implies a true complexity to work with them.

4. ***Veracity:*** The credibility is given by the accuracy and quality of data set. How sure are you of these data? Or How many data are really of a given type? Taking account that many unstructured data can come from Facebook posts, tweets, LinkedIn posts, etc. Do you trust what you see? This V concentrates the greatest preoccupation, because the objective is to process and analyze data to get good results.

5. ***Validity:*** In this case, the data accuracy is considered respect to intended use. In consequence, the same data set may be valid to one problem and not valid for another.

6. ***Visualization:*** It refers to way that data are presented or showed. Once data are processed, you need to represent them visually to be legible and accessible, beside to find patterns and hidden keys in relation of considered issue.

7. ***Value:*** The value is obtained when data are transformed into information, the information becomes knowledge and this last concludes in a decision.

In all cases, it is necessary considered that results from big data should not exceed the cost of its processing, administration and/or storage.

2.3 Machine Learning

Machine learning is the study of algorithms which can learn complex relationships or patterns from empirical data, and make accurate decisions. Machine learning includes broad range of techniques such as data pre-processing, feature selection, classification, regression, association rules, and visualization [9]. In big data analysis, machine learning techniques can help to extract information from large data sets, identifying hidden relationships.

Machine learning algorithms are classified into two groups: supervised and unsupervised learning. In the first kind, all data is labeled and the algorithms learn to predict output from input data. In second class, data are unlabeled and the algorithms learn to inherent structure from the input data.

The next sub-sections present two algorithms of machine learning, one of each class. Both can be applied in anomalies detection.

2.3.1 Supervised Classification: k-NN Algorithm

The classification process builds models capable of determining if an object, from its characteristics, is member or not of a category. A classification is supervised when, in advance, there is an already classified observations set, and it is known to which set belongs each observation. The algorithms dedicated to solve supervised classification problems usually operate with information provided by a set of samples, patterns, examples or training prototypes, all of them are representatives of each classes [10].

In particular, this paper uses a supervised classification algorithm based on neighborhood criteria. It is known as K nearest neighbor (k-NN). The k-NN method and its variants are based this intuitive idea: "Similar objects belong to the same class". The class is determined by the most similar objects. The similarity idea is formally reflected in distance concept, usually the Euclidean distance is used [11]. The calculation of the k-NN can be solved using parallel techniques, in [11] has been shown that the parallel implementation using GPU [12] obtains very good response times.

2.3.2 Neural Networks

A neural network (NN) consists of simple processing units: neurons, and their connections: directed and weighted links between two neurons i and j. Formally, a neural network is defined as follows: A NN is sorted triple (N,V,w) with two sets N, V and a function w, where N is neurons set and $V = (i,j)/i,j \in N$, these elements are connections between neuron i and j. The function $w : V \rightarrow \mathbb{R}$ defines the weight, $w((i,j))$, between i and j [13].

2.4 Processing Images

The great growth of data amount makes practically impossible for a person works all raw data and gets conclusions, trends and patterns. Data visualization techniques can help significantly to this process.

The main objective of visualization is an adequate perceptual representation of data, trends and underlying relationships between them. The visualization purpose is not only images creation but also the insight of large data amounts.

Informally, visualization is data or information transformation into images. A successful visualization can

reduce considerably the time consumed to understand data, to find relationships and to obtain information. To generate a visualization, it is necessary to map data in the Cartesian space (two or three dimensions). This space has to represent the relationships as intuitively as possible. Finding a good spatial representation of data not is a simple task, it is one of the most difficult tasks in visualization of abstract information [14].

2.5 High Performance Computing (HPC)

A Big Data Systems needs to process large data volumes and, consequently, demands greater computational capacity. For this reason, conventional computer systems are not suitable for proper processing. HPC techniques allow intensive computational operations and improve processing speed; involving different technologies such as distributed systems and parallel systems (computer clusters, cloud computing, graphic processing unit and massively parallel computers) [15, 16].

Graphical processing units (GPU) can be used to solve general purpose problems (General Purpose Graphical Processing Units, GPGPUs) [17]. When the GPU is used as a parallel computer, it necessary taken into account the processing units number, its memory structure and own programming model. The CUDA programming model, developed by NVIDIA, allows to use a GPU as a highly parallel computer, providing a high-level programming environment. It defines to GPU as a programmable co-processor of CPU with an architecture is formed by multiple streaming multiprocessors, each with several scalar processors.

In [18, 19, 20, 21], they use GPU to solve problems of massive data. They propose different techniques of data analysis in GPU. These techniques are included in areas such as machine learning, search and sorting, data mining, database, among others. This paper presents a big data problem, the anomalies detection in network traffic as soon as they happen. This work include several tasks, most of them can solve using GPU, and in consequence accelerate the solution.

3 A Big Data Problem: Analysis Network Traffic

The network traffic analysis in terms of information security is a Big Data problem. As it was said in previous sections, a problem of this type is defined by 7 Vs [22]. Anomalies or attacks detection in a Network involves to work with all circulating data in a network (volume), generated them to high speed (velocity), with heterogeneous nature (variety), belonging to normal or anomalous profiles (veracity), processed and filtered in base to significant characteristics, depending of particular system (validity), where their visual representation is possible in order to read them easily (visualization), and from them it is necessary to detect effectively and precisely attacks in a reasonable time (value).

Anomalies detection in data networks identifies unusual patterns, i.e. patterns not belong to the typical traffic in the network [23]. Detecting a possible attack requires technologies to classify traffic and associate data flows with those applications that generate them. Work data set grows at a high speed, it is much greater than processing capacity. Having the ability to handle large data volumes not only allows to know what is happening at this moment, but also to trace patterns over time.

When information is analyzed in real time, it is often easy to overlook some indicators. If information analysis is made in other context and by long time, it is possible to find other meanings. In this case, the main is to process network traffic, apply any techniques (with intelligent or not) and to obtain concrete results. All these have to be made in a short time to arrive relevant conclusions to detect real attacks and take quick decisions.

The next sections describe the first solutions developed, their advantages and drawbacks, and finally some experimental results that show their behavior.

4 First Solutions

The detecting task of possible packets with anomalous data in a computer network is very expensive. A good alternative can be combine images processing, machine learning techniques and HPC in a only one solution.

In [23, 24, 25] a detection of anomalies model to a local network called *P-SNADS* (Parallel-Supervised Network Anomalies Detection System) was presented. In figure 1, the *P-SNADS'* architecture is detailed. It has 2 stages, each of them well differentiated. These are:

- *Data Recollection:* This stage captures network traffic and organizes it in data flows. Its objective is to obtain data to work and implies the following sub-tasks:
  - Data Capture: This process catches network traffic. It uses a sniffer to do it, for example T-Shark, and it is activated at specific moments, for examples when there is more network traffic.
  - Data Selection: This step selects frames to be analyzed, according to the attacks considered, these are TCP, UDP and ICMP frames.
  - Feature Extraction: In this step, all interest fields of each frame are extracted to be analyzed. They are: source and destination IP address, source and destination port and protocol type (TCP, UDP or ICMP).

- Normalization: This step is fundamental, the tuples become an integer values vector. All address IP (IPv4) a.b.c.d are transformed according to equation (1)

$$(ax256)^3+(bx256)^2+(cx256)^1+(dx256)^0 \quad (1)$$

- *Anomalies Detection*: In this stage, the task is to find possible attacks in a network. It can be solved applying different techniques. Which one? It depends on tuples number to analyze:
  - If tuple is examined at a time, the k-NN algorithm is applied. It compares the collected tuples with the attacks signatures.
  - If a tuples set is analyzed, different images are generated. Afterwards, similarities are searched between each created images with those representing anomalies, previously defined and stored. To compare two images, the algorithm SIFT (Scale Invariant Feature Transform) is used (It detects and describes distinctive invariant features in images. For any object in an image, interesting points (key points) on object can be extracted to provide a "feature description" of object) [26]. If these key points match, a possible attack would be detected, otherwise the current traffic is not a recognized intrusion.

The attacks signature and anomalies image repository are in constant update, they have to include new attack or new features of already known.

The above stages demand a lot of computational resources. Therefore, HPC is considered to improve both solution, particularly GPGPU (General Purpose Graphics Programming Unit). The results obtained are satisfactory, the times of the parallel solutions to each stage are significantly lower than corresponding to sequential solution, a times reduction greater than 500x is achieved.

In [24, 25], known attacks by signatures are detected by the supervised learning technique k-NN. Each tuple of current traffic in network, is analyzed if matches with any of known attack signatures. As result, the K closest to each attack are obtained, determining the occurrence or not of them.

In both cases, there is a need to have signatures or patterns constantly updated, making it impossible to detect new attacks early. This paper intends to incorporate new techniques related to machine learning without supervision, which do not require intervention of network administrator, permanently he/she has to update attacks databases. For these learning techniques, by its nature, it is also possible to apply HPC in your computational solution.

5 Experimental Results Analysis

This section presents a experimental results analysis of fist solutions of P-SNADS. To catch network data traffic, T-Shark tool is used [2]. P-SNADS works with several samples, each of them has approximately 2000 frames. These frames belong to a local area network (LAN) of Networks Laboratory of Universidad Nacional de San Luis. Three attacks are simulated in a server and they pretend to deny the HTTP service. Once obtained the normalized frames, different databases are built with normal and anomalous traffic.

The results are contrasted with sequential solution. Its times are obtained of a PC with an AMD FX (tm) -6300 Six-Core Processor x6 processor with 7.8 GB memory and an 80.5 GB disk. For HPC solution, a GPU is used. It has the following features: Tesla K20c (2496 processors), 4.6 GB of Memory and 706 MHz of clock frequency and 2600 MHz of processor memory.

The k-NN module receives the databases as input data and performs an evaluation for different values of k. Four values of k are considered, they are 5, 7, 10 and 12.

The Table 1 shows the necessary milliseconds to obtain k-NN in its two versions: sequential and parallel. In the parallel version, the time includes the transfer times of the database to the GPU, *K-NN* calculation and results transfers (the k-NNs) to the CPU.

Table 1: Average times (in milliseconds) of *K-NN*.

K	Time (sequential)	Time (parallel)
5	1089,90	0,0904
7	1086,60	0,0907
10	1081,20	0,0935
12	1091,60	0,0978

As can be seen, the parallel times are significantly lower to the sequential time. In addition, the parallel version shows quasi-constant behavior, independently of k value.

When k-NN are computed, the next metric are considered: Positive Predictive Value (Precision - PPV),True Positive Rate (Recall - TPR), and F-measure (F) [24]

Figure 3 shows each metric to each attack considered: Smurf (a), Land (b) and Fraggle (c).

From observation of the above graphs, Smurf attack has an approximate accuracy of 72%, while for the others two are lower: Land= 41% and Fraggle = 52%. Regarding Recall, more significant results are obtained, to Smurf is 80%, Land is 72% and Fraggle is 75%. In base of these values, it possible to infer that

[2] https://www.wireshark.org/

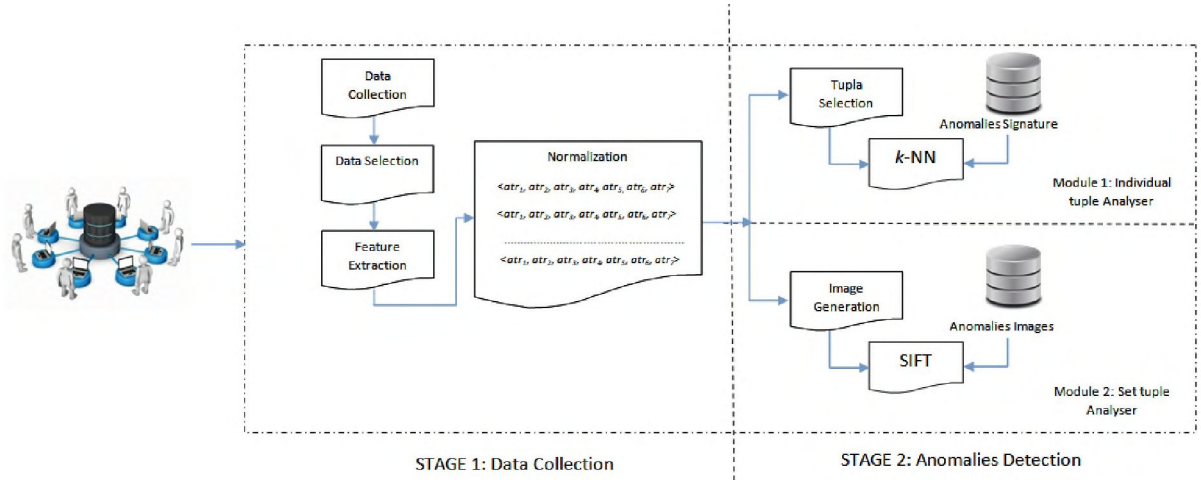

Figure 2: P-SNADS' Arquitechture

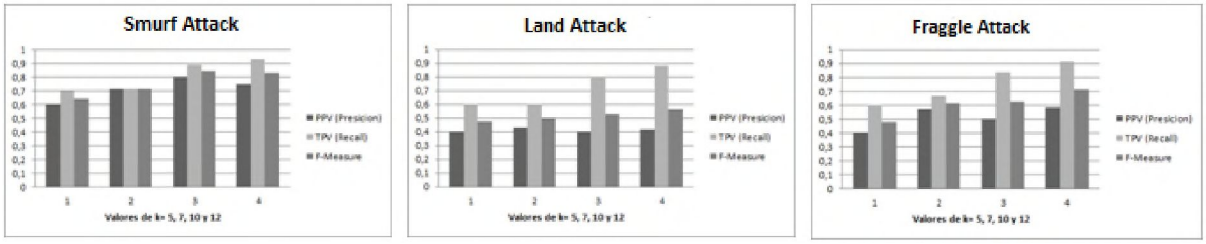

Figure 3: Results obtained in attacks: Smurf, Land and Fraggle.

the detection rate of anomalous traffic is high, particularly when k = 12. Finally, the F-measure also returns good results, particularly for k=12 no matter which attack is. The values obtained are 0.83 for Smurf; 0.56 for Land and 0.71 for Fraggle. It means that our model performs well in detecting attacks. Therefore, for metrics analyzed, P-SNADS achieves satisfactorily the objectives of this work.

For SIFT algorithm, different sizes images are considered, which sizes are between 50KB and 20MB.

The Table 2 shows how many milliseconds are needed to obtain all key points of an image using SIFT, and the Table 3 details the milliseconds quantity requires to determine the common key points between the images of traffic and the attack (this last process is known as MATCH process). In both tables, the times correspond to the sequential and parallel versions, and for all results, the time averages of several executions are displayed.

In both tables, 2 and 3, the times of parallel solutions are significantly lower than those of sequential versions, independently of image size. The times reduction is greater than 500x.

Table 2: SIFT average times in milliseconds.

Range of image size	SIFT (sequential)	SIFT (parallel)
10KB y 300KB	1780.92	14.045
300KB y 1MB	4650.075	20.615
1MB y 3MB	22084.63	64.6675
3MB y 5MB	40114.55	96.01

6 Conclusions and Future Works

In this paper, the P-SNADS model was presented, it is divided into two sequential stages with different objectives, one is the Data Collection and the other is Detection of Anomalies. This objective concentrates the greatest interest because it seeks to collaborate in the process of computer security detecting anomalous traffic in a computer network. To carry out this task, two different techniques are applied according to the nature of the attack.

When the attack has a signature, the supervised learning technique k-NN can be used, each captured traffic tupla is compared against known attacks firms. In base to results, the occurrence or not of an attack can be determined.

For attacks based in flow, tuple analysis is not a good solution. Techniques of image representation and processing are suitable. From captured traffic, different images are built in order to be able to compare with images that represent attack patterns. To establish an attack, the images comparison is made applying

Table 3: MATCH average times in milliseconds.

Range of image size	MATCH (sequential)	MATCH (parallel)
10KB y 300KB	637.09	4.635
300KB y 1MB	3325.09	8.05
1MB y 3MB	30421.66	26.885
3MB y 5MB	85513.74	71.12

SIFT algorithm.

The application of HPC techniques in several P-SADS' stages allows to improve execution times and, in consequence, early detect possible attacks to a network.

While both techniques, from the point of view of the 7 Vs, are applicable to Big Data problems, both are supervised methods, they need a constant updating of signatures or images that represent attacks.

The next step is to incorporate others Machine Learning techniques, such as Neural Networks and Deep Learning. These approaches have shown good performance in developing models and architectures for discovering patterns of malicious activities.

References

[1] Tulasi.B, R. S. Wagh, and B. S., "High performance computing and big data analytics - paradigms and challenges," *International Journal of Computer Applications*, vol. 116, Abril 2015.

[2] Y. Wang, *Statistical Techniques for Network Security: Modern Statistically-Based Intrusion Detection and Protection*. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing, 2008.

[3] "S. institute, transmission control protocol: Darpa internet program protocol specification. defense advanced research projects agency, information processing techniques office," Sept. 1981.

[4] D. Gibson, "Comptia security+: Get certified get ahead: Sy0-201 study guide createspace independent pub.," 2009.

[5] H. R. J. L., "Definición de un modelo de seguridad en redes de cómputo, mediante el uso de técnicas de inteligencia artificial. tesis presentada como requisito parcial para optar al título de magíster en ingeniería – automatización industrial. universidad nacional de colombia.," 2012.

[6] I. G. B. S. B. Analytics and Optimisation, "Analytics: el uso de big data en el mundo real.," in *Escuela de Negocios Saïd en la Universidad de Oxford.*

[7] C. L. P. Chen and C. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, 2014.

[8] D. S. Terzi, R. Terzi, and S. Sagiroglu, "Big data Analytics for Network Anomaly Detection from Netflow Data," *IEEE*, 2017.

[9] A. Y. Nikravesh, S. A. Ajila, C. H. Lung, and W. Ding, "Mobile network traffic prediction using mlp, mlpwd, and svm," pp. 402–409, June 2016.

[10] T. Hind, *Análisis Estadístico de Distintas Técnicas de Inteligencia Artificial en Detección de Intrusos*. PhD thesis, Universidad de Granada, 2012.

[11] N. Miranda, *Cálculo en Tiempo Real de Identificadores Robustos para Objetos Multimedia Mediante una Arquitectura Paralela GPU-CPU*. PhD thesis, Universidad Nacional de San Luis, 2014.

[12] P. M. F, *Computación de alto desempeño de GPU*. Editorial de la Universidad Nacional de La Plata (EDULP), 2011.

[13] A. M. Ghimeş and V. V. Patriciu, "Neural network models in big data analytics and cyber security," in *2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp. 1–6, June 2017.

[14] S. Martig, S. Castro, M. Larrea, S. E. D. Urribarri, M. Escudero, and L. Ganuza, "Herramientas de visualización para la exploración de datos," *IX Workshop de Investigadores en Ciencias de la Computación*, 2007.

[15] G. Hager and G. Wellein, *Introduction to High Performance Computing for Scientists and Engineers*. CRC Press, Inc., 1st ed., 2010.

[16] Y. You, S. L. Song, H. Fu, A. Marquez, M. M. Dehnavi, K. J. Barker, K. W. Cameron, A. P. Randles, and G. Yang, "MIC-SVM: designing a highly efficient support vector machine for advanced modern multi-core and many-core architectures," in *2014 IEEE 28th International Parallel and Distributed Processing Symposium, Phoenix, AZ, USA, May 19-23, 2014*, pp. 809–818, 2014.

[17] NVIDIA, "Nvidia cuda compute unified device architecture, c programming guide. version 7.5," 2015.

[18] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013* (K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, eds.), (Berlin, Heidelberg), pp. 411–418, Springer Berlin Heidelberg, 2013.

[19] S. Chen, J. Qin, Y. Xie, J. Zhao, and P.-A. Heng, "A fast and flexible sorting algorithm with cuda," in *Algorithms and Architectures for Parallel Processing* (A. Hua and S.-L. Chang, eds.), (Berlin, Heidelberg), pp. 281–290, Springer Berlin Heidelberg, 2009.

[20] Y. Chen, Z. Qiao, S. Davis, H. Jiang, and K.-C. Li, "Pipelined multi-gpu mapreduce for big-data processing," in *Computer and Information Science* (R. Lee, ed.), (Heidelberg), pp. 231–246, Springer International Publishing, 2013.

[21] S. Herrero-Lopez, "Accelerating svms by integrating gpus into mapreduce clusters," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1298–1305, Oct 2011.

[22] M. A. ud-din Khan, M. F. Uddin, and N. Gupta, "Seven v's of big data understanding big data to extract value," in *Conference of the American Society for Engineering Education.*

[23] M. Barrionuevo, M. Lopresti, N. Miranda, and F. Piccoli, "Un enfoque para la detección de anomalías en el tráfico de red usando imágenes y técnicas de computación de alto desempeño," *XXII Congreso Argentino De Ciencias de la Computación*, pp. 1166–1175, 2016.

[24] M. Barrionuevo, M. Lopresti, N. Miranda, and F. Piccoli, "An anomaly detection model in a lan using k-nn and high performance computing techniques," *Communications in Computer and Information Science*, pp. 219–230, January 2018.

[25] M. Barrionuevo, M. Lopresti, N. Miranda, and F. Piccoli, "P-sads: Un modelo de detección de anomalías en una red lan," *5to Congreso Nacional de Ingeniería Informática / Sistemas de Información Aspectos Legales y Profesionales y Seguridad Informática*, 2017.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, Nov 2004.