Journal of Network and Computer Applications 64 (2016) 1–11

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Network anomaly detection using IP flows with Principal Component Analysis and Ant Colony Optimization

Gilberto Fernandes Jr.[a,*], Luiz F. Carvalho [b], Joel J.P.C. Rodrigues [a,c], Mario Lemes Proença Jr.[b]

[a] Instituto de Telecomunicações, University of Beira Interior, Rua Marquês d'Ávila e Bolama, Covilhã 6201-001, Portugal
[b] Computer Science Department, State University of Londrina (UEL), Londrina, Brazil
[c] University of Fortaleza (UNIFOR), Fortaleza, Ceará, Brazil

ARTICLE INFO

Article history:
Received 21 August 2014
Received in revised form
15 June 2015
Accepted 18 November 2015
Available online 6 February 2016

Keywords:
Traffic characterization
Anomaly detection
Network management
Principal Component Analysis (PCA)
Ant Colony Optimization (ACO)
Dynamic Time Warping (DTW)

ABSTRACT

It is remarkable how proactive network management is in such demand nowadays, since networks are growing in size and complexity and Information Technology services cannot be stopped. In this manner, it is necessary to use an approach which proactively identifies traffic behavior patterns which may harm the network's normal operations. Aiming an automated management to detect and prevent potential problems, we present and compare two novel anomaly detection mechanisms based on statistical procedure Principal Component Analysis and the Ant Colony Optimization metaheuristic. These methods generate a traffic profile, called Digital Signature of Network Segment using Flow analysis (DSNSF), which is adopted as normal network behavior. Then, this signature is compared with the real network traffic by using a modification of the Dynamic Time Warping metric in order to recognize anomalous events. Thus, a seven-dimensional analysis of IP flows is performed, allowing the characterization of bits, packets and flows traffic transmitted per second, and the extraction of descriptive flow attributes, like source IP address, destination IP address, source TCP/UDP port and destination TCP/UDP port. The systems were evaluated using a real network environment and showed promising results. Moreover, the correspondence between true-positive and false-positive rates demonstrates that the systems are able to enhance the detection of anomalous behavior by maintaining a satisfactory false-alarm rate.

1. Introduction

Computer networks have become increasingly indispensable due to the services offered by the constant development of communication technologies. However, this progress is pegged to the growing management complexity of these networks. It implies the responsibility of the network administrator to detect anomalies like flash crowds, network elements failures, misconfigurations, malicious attacks, among others, in order to ensure they do not cause significant impact on the quality of connections or interrupt services provided to end users (Zhang et al., 2014; Zarpelao et al., 2009).

In literature, anomaly detection methods can be classified in two ways: Signature-based and profile-based. Signature-based systems use a prior knowledge about the characteristics of each kind of anomaly to identify potential incidents previously known. Moreover, profile-based approaches create a network profile representing the traffic normal behavior, and traffic anomalies are detected from deviations with respect to this profile. Although there may be higher false alarms rates, profile-based systems are more promising due to its flexibility, as they can detect unknown anomalies (Patcha and Park, 2007; Bhuyan et al., 2014).

This paper presents two profile-based systems for anomaly detection composed of two stages: (i) creating a model which characterize the normal traffic behavior through historical data, called Digital Signature of Network Segments using Flow Analysis (DSNSF) and (ii) detection of behavior deviations with activation of multilevel alarms. Thus, our approach is able to work proactively, detecting anomalous network's traffic in an automatic manner, without human supervision.

The first proposed traffic characterization method is Principal Component Analysis for Digital Signature (PCADS), which is based on Principal Component Analysis, a statistical method for dimensionality reduction. PCA will be used as a mechanism to identify the network traffic time intervals with a medium variance among the input historical data, allowing it to describe what could be the normal behavior of a network segment. We also present a modification of Ant Colony Optimization metaheuristic, named Ant Colony Optimization for

* Corresponding author. Tel.: +351 910438982.
E-mail addresses: gil.fernandes6@gmail.com (G. Fernandes Jr.), luizfcarvalhoo@gmail.com (L.F. Carvalho), joeljr@ieee.org (J.J.P.C. Rodrigues), proenca@uel.br (M.L. Proença Jr.).

http://dx.doi.org/10.1016/j.jnca.2015.11.024
1084-8045/

Digital Signature (ACODS), a clustering approach which seeks near-optimal solutions to grouping data through self-organized agents.

The detection mechanism is constructed over an adaptation of the pattern matching technique Dynamic Time Warping (DTW) (Sakoe and Chiba, 1978), called Adaptative DTW (ADTW). This technique is used to recognize shifted behavior between the DSNSF and real traffic series through time alignment, enabling improved analysis of sudden events and those that occur along the time. This aims to improve the accuracy and reduce false alarms rates in anomaly detection.

The entirety of this work was accomplished through the analysis and extraction of seven IP flow features present in a historical database of a real network: bits, packets and number of flows transmitted per second, source IP address, destination IP address, source Port and destination Port. This multidimensional monitoring enables more effective analysis than using only one attribute. According to Molnar and Moczar (2011), many previous works are focused on the analysis of a single flow attribute for anomaly detection, making it unfeasible for recognition of complex attacks. Furthermore, multidimensional analysis of correlated attributes is effective due to its ability to better describe the behavior of a network and, consequently, infer when this behavior is anomalous (Zhou et al., 2009).

The main contributions of this paper consist in presenting two different approaches for a profile-based anomaly detection system. They are compared regarding the traffic characterization, anomaly detection accuracy results and their computational complexity. To evaluate the proposed systems, we tested them using real traffic data extracted from a core switch at State University of Londrina (UEL).

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 discusses the proposed anomaly detection methodology. Section 4 introduces the proposed traffic characterization and anomaly detection methods. In Section 5, the methods are evaluated and compared. This work is concluded in Section 6.

2. Related work

Lakhina et al. (2004, 2005) were one of the pioneers in using the Principal Component Analysis (PCA) for anomaly detection. Their works present an anomaly detection method based on an efficient subspace separation of all network traffic through PCA. The anomalous subspace, which is noisier and contains the significant traffic spikes, is separated from the normal network-wide traffic, which is dominated by predictable traffic, containing the normal network behavior. After that, it is possible to accurately diagnose volume anomalies when the magnitude of the projection onto the anomalous subspace exceeds an associated Q-statistic threshold.

Callegari et al. (2011) used the main idea of the PCA method developed by Lakhina et al. (2004) to add some novelties to it in order to make great improvements. The authors use entropy along with the Kullback–Leibler (K–L) divergence to construct the time series from the aggregate flows and detect anomalous behavior. According to them, it can result in a better performance with more stability to the detection system. Also, Kanda et al. (2010) combine sketches (random traffic projections) and PCA to create a new method to detect and identify the source IP addresses associated with the traffic anomalies in the backbone traces measured at a single link.

Abadi and Jalali (2006) have proposed the AntNag algorithm, which is considered one of the first approaches that use Ant Colony Optimization (ACO) for intrusion detection. The authors assume that intruders usually trigger attacks taking advantage of multiple security vulnerabilities. The algorithm is able to discover the set of all possible attack scenarios as a directed graph, called Network Attack Graph (NAG). An exploit is represented by an edge and every complete path in the graph describes an attack scenario. The ACO aims to minimize the set of exploits that must be eliminated to ensure a safer network scenario.

Most researches of anomaly detection employ signature based strategy or supervised-learning. These approaches have several drawbacks, such as the need for labeled data; an external supervisor is essential and bad results from unknown anomaly detection. So, Mazel et al. (2011) introduce an unsupervised approach to detect and characterize network anomalies. This approach, initially, acts using a clustering technique, combining Sub-Space Clustering with Evidence Accumulation or Inter-Clustering Results Association, to blindly identify anomalies in traffic flows.

A Particle Swarm Optimization-based clustering approach to finding out anomalous behavior of the network based on traffic volume is used in Lima et al. (2010). For this purpose, the centers of clusters found during the clustering process of both the actual traffic and those of normal behavior are compared. Authors perform clustering in the traffic collected by SNMP agents and its respective DSNS by using the *K*-means algorithm combined it with the PSO algorithm in order to improve the calculation and solutions of clusters centroids. A Digital Signature of Network Segment (DSNS) created from SNMP objects is used to provide thresholds for normal behavior of network traffic. In this manner, a PSO alarm system verifies the existence of anomaly in discrete time intervals by checking the Euclidian distance between the sampled traffic movement and its respective clustering centroid, triggering an alarm wherever this distance exceeds a threshold value λ. Although this work achieved excellent results, it is limited by the use of SNMP objects. Using flow-based network management like in our approach, a more detailed traffic analysis can be performed, since it presents a greater variety of information about network behavior.

Qin et al. (2011) observed that the abnormal behaviors over the internet, such equipment malfunction or network resource abuse, can cause significant changes in the normal flow patterns. Therefore, Qin et al. proposed a network monitoring model based on traffic flow analysis that extracts only four features of the flow records to capture the traffic patterns. This approach extracts the abnormal behavior by measuring those features using a Blind Source Separation (BSS) method, and it is effective in identifying abnormal behaviors not related to significant changes in traffic volumes.

Tartakovsky et al. (2013) developed a novel multi-cyclic anomaly detector based on a changepoint detector procedure, called Shiryaev–Roberts (SR), which has exact multi-cyclic optimality. Changepoint detectors are statistical methods that aim fastest manners to identify a change in the state of a stochastic process or time series. Authors performed tests using a real SIN flood attack, outperforming CUSUM, another common changepoint detector.

Jiang et al. (2013) perform a multi-scale analysis in order to build a novel anomaly detector for high-speed backbone networks. Author divide this in three steps. Their first action is to use wavelet transform to decompose network traffic into multiple scales since they observed that burst variations in that decomposition are results of unusual changes of network traffic (i.e. anomalies). After that, their second step is to perform a Principal Component Analysis on the continuous wavelet transform previously done, constructing a mapping function. Then, they extract unusual features of the network traffic by using this mapping function and PCA. In another work from Jiang et al. (2014), they present a system to network-wide traffic detection based on the transform domain analysis. Authors take into account the network topology information and network-wide traffic in order to accurately detect the abnormal traffic. To do that, they

classify origin–destination flows in the network into different groups in accordance with common destination addresses since these flows share some relative characteristics. Then, the inherent properties of network-wide traffic are extracted by introducing the transform domain analysis. Through this process, it was verified that the anomalous network traffic exhibits the distinct high-frequency features in the transform domain.

Although Rawat et al. (2004) propose an intrusion detection system; its methodology is similar to what is proposed in this work. Authors based on the assumption that abnormal activities (intrusions) differ from normal activities substantially. Thus, any normal execution of a process follows a pattern that can be profiled as the normal behavior, and then, any deviations observed in it may be signed as intrusions. Therefore, to measure the similarity between processes, Rawat et al. proposed the use of Kendall Tau distance, while considering three key issues (occurrence of individual system call, frequency of this system call in the process, and the position of this call in the process), which led to promising results regarding false positive rates.

The idea of using data mining techniques combined with expert systems in designing effective anomaly-based IDS is presented by Sodiya et al. (2004). The objective is to extract patterns from audit data consistent and useful user behavior and then maintain these behaviors in normal profiles. The detection of anomalous events occurs from the comparison of the current record and the profile database stored by the system. This knowledge base was acquired from the previous designs and the problems of current intrusion detection systems in the literature. If the activities and parameters expressed by this current record are not equal to the normal profile, an alarm is generated, informing the occurrence of an anomaly and disrupting the event instantaneously. However, if an activity is not found to be intrusive, the record is kept in the normal profile and may be used for future detections.

3. Methodology

A profile-based detection system has its efficiency fully related on traffic behavior characterization. Proença et al. (2005) and Assis et al. (2014) observed that the traffic network is currently composed of cycles consisting of bursts which have particular

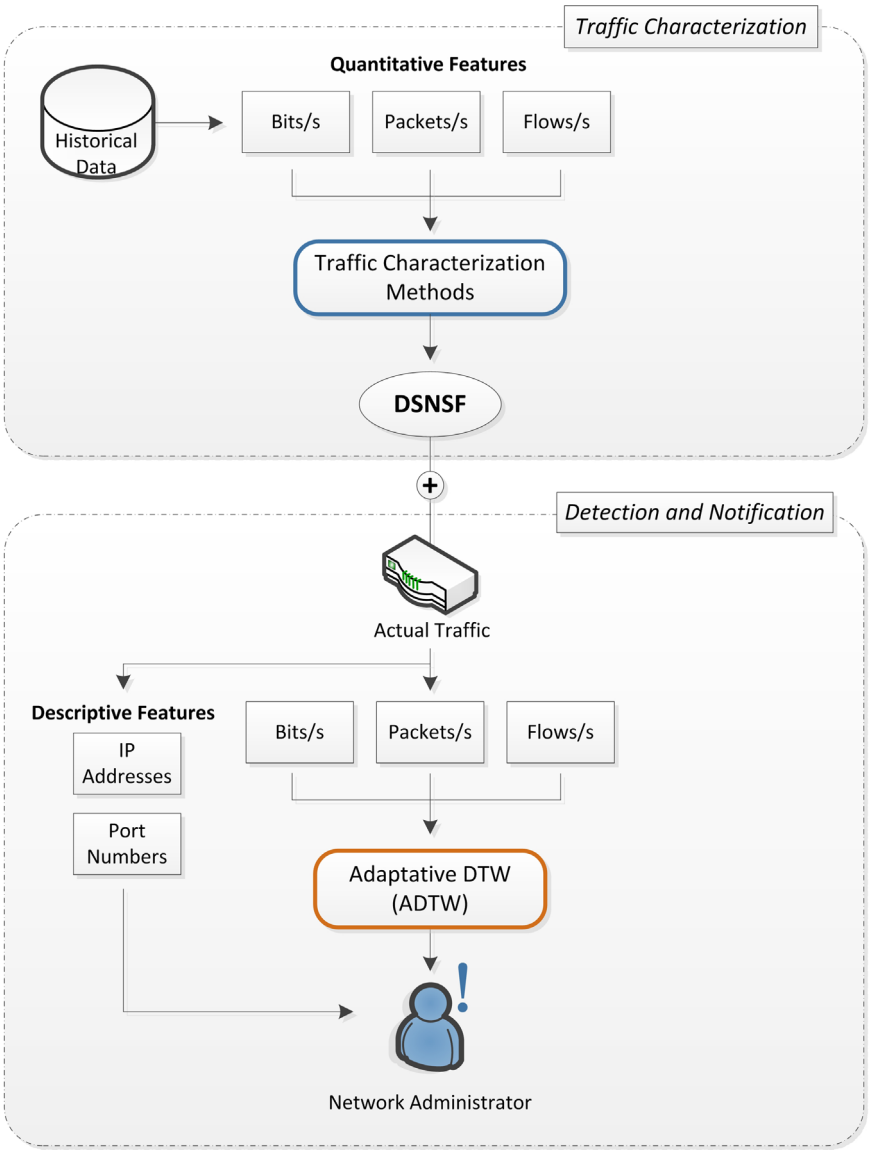

Fig. 1. Modules of anomaly detection methodology.

characteristics of its use. These behaviors are directly affected by working hours and the workdays period of people who use the network. Therefore, our paper contributes by recognizing these behaviors and their characteristics to be created a traffic digital signature called Digital Signature of Network Segment using Flow analysis (DSNSF). Such signature is responsible for harboring information about the normal traffic behavior, being adopted for anomaly detection by recognizing behavioral deviances which clashe from the usual.

The proposal presented in this work is detailed in Fig. 1 and objectively evaluates two systems able to characterize the network normal behavior analyzing the historical traffic, stored in an IP flow collector. In this manner, our methodology is divided in two modules. The first one, Traffic Characterization, performs feature extraction of analyzed flow attributes to generate the DSNSF. For this purpose, a multidimensional analysis of the attributes bits, packets, and flows transmitted per second is performed by the PCADS and ACODS methods. The discussed methods belong to different groups of algorithms, so the creation of digital signatures in both methods occurs differently due to their peculiarities.

In the second module, named Detection and Notification, the previously generated digital signatures are compared to actual network traffic passing by the network device (switch, router). Abrupt changes of great magnitude in the traffic attributes indicators, which greatly clash with the normal profile represented by DSNSF, can be considered unusual events. Thus, those shifted behaviors detection is accomplished by the Adaptative Dynamic Time Warping (ADTW) algorithm proposed in this paper. Information regarding the IP source and destination addresses, and origin and destination ports, are used in the analysis, which are necessary to provide detailed reports to the network manager about the anomalies found.

Each day has a DSNSF, which is generated using historical traffic activity from previous weeks. Thus, the signature is automatically adjusted to the traffic behavior by introducing new behaviors every week. The importance of the historical basis is directly related to Local Search approach. Assuming that a behavior is present in most of the analyzed days and maintains it regularly, it absolutely will be added to DSNSF. Otherwise, it could be characterized as a sporadic event and its inclusion in the normal pattern would entail numerous false alarms during the detection of anomalies.

A weakness of the profiles-based detection is that anomalies can be slowly introduced in the normal pattern until they become a legitimate event. For example, a SYN flood attack that keeps happening every day at the same time interval. As it becomes more frequent, the DSNSF will incorporate it as a normal behavior if nothing is done to interrupt the frequent attacks. However, we solved this limitation by offering the administrator relevant information, like IP addresses and TCP/UDP port numbers, which helps with the solution of such problems. In this manner, if these anomalous events are interrupted, they are not included in the historical basis and consequently will not compose DSNSF.

The use of DSNSF along with anomalies recognition provides significant information to the solution of attacks or failures. Furthermore, by using the DSNSF, it is possible to constantly monitor the traffic, due to its ability to provide information about the use of resources exploited by network services. This information assists the network administrator in quick decision-making, promoting the network reliability and availability of the services offered by it.

4. Traffic characterization and anomaly detection

In this section, we present two traffic characterization methods used for DSNSF creation: The PCADS statistical method and the ACODS metaheuristic. Also, we describe our anomaly detection approach based on thresholds deriving out of digital signature.

4.1. PCADS—Principal Component Analysis for Digital Signature

The Principal Component Analysis for Digital Signature (PCADS) (Fernandes et al., 2013) is based on Principal Component Analysis (PCA) to create a DSNSF for traffic characterization of a network segment. The PCA was first introduced in 1901 by Karl Pearson (Pearson, 1901). It is a statistical procedure used to reduce the dimensionality of a multivariate problem by analyzing the variance of each variable among all input dimensions. Then, the input data can be represented by a reduced set of dimensions without much loss of information (Jolliffe, 2002).

In classic PCA, the input is a $n \times p$ matrix, composed of p columns representing the dimensions (variables) and n lines, as the n samples of each variable.

The data collected from flow records are arranged in such a way that the traffic movement of each day is denoted by a three vectors containing a total of bits, packets and flows corresponding to the 24 h of each day. Thus, the input matrix of PCADS is constructed. The p dimensions will be the p traffic movements (days) chosen as a basis to generate the DSNSF, and the n lines will be the n samples of bits, packets or number of flows per second extracted from the flow records.

Algorithm 1. PCADS algorithm used for DSNSF creation.

Require: Set of bits/s, packets/s or flows/s collected from historic database arranged in three $n \times p$ matrices.

Ensure: μ: a vector representing the bits/s, packets/s or flows/s sets of a day, arranged in 1440 intervals of 1 min, i.e. the DSNSF.

1: **for** t = 1–1440 do
2: Normalize the input data (Zero Mean Matrix)
3: Calculate the covariance matrix
4: Calculate the eigenvectors and eigenvalues
5: Choose the eigenvector that has an associated intermediate eigenvalue
6: $\mu_t \leftarrow eigenvector_t \times zero\_mean\_matrix_t$
7: **end for**
8: **return** μ

The PCA method used in the PCADS model is presented in Algorithm 1. Since each traffic period has its own characteristics and seeking to prevent a period interferes in another one, PCADS is performed in a time window of one minute (t=24 h/1 min=1440 time intervals).

First of all, it is required to move the origin to the mean of the data set, by subtracting off the mean from each column in the input matrix. This is called mean deviation form (or zero mean matrix), and it is important because it eases the covariance matrix calculation and avoids distorted results due to differences in mean link utilization.

Then, with the zero mean matrix, the algorithm calculates the covariance matrix, which is used to compute two important structures, the eigenvectors, and eigenvalues. Each dimension has an associated eigenvector, which points toward the variance of data, and an eigenvalue, a numerical value which indicates the significance of its associated dimension among the others.

The eigenvectors with the highest eigenvalues are called principal components, and they are chosen by classic PCA to compose a new reduced dataset. PCADS creates a digital signature using only one principal component (eigenvector). However, instead of selecting the eigenvector with the highest eigenvalue, we chose an eigenvector with an intermediate value of eigenvalue. This is

because the largest eigenvalue represents the dimension with the largest variance among all components of the dataset, and by creating a digital signature based on a component of high variance, it will clash with the normal traffic pattern of the network segment. After some tests, it was observed that an intermediate variance component will produce a more uniform signature in order to prevent possible disparities (anomalies) in the training dataset to generate noise in the DSNSF.

In step six of Algorithm 1, for each time interval t, the selected eigenvector is multiplied with the input matrix in mean deviation form to produce a result of only one dimension. Finally, after computing the principal component analysis for each time interval t, the output is the DSNSF.

4.2. ACODS—Ant Colony Optimization for Digital Signature

The foraging behavior of ants combined with the ability to find the shortest path between their colony and food source has inspired the creation of a very promising metaheuristic, the Ant Colony Optimization (ACO). It is based on the principles of swarm intelligence, which are defined as a population of agents competing and globally asynchronous, cooperating to find an optimal solution (Dorigo et al., 2006).

According to Jiang and Papavassiliou (2006), the ant colonies' habits living in groups is essentially similar to the grouping of data. Algorithms based on the behavior of ants have natural advantages in the application of cluster analysis. It is because these methods are particularly suitable to perform exploratory data analysis using self-organization of individuals during the construction of a solution. Thus, we introduce the Ant Colony Optimization for Digital Signature (ACODS), a modification of the Ant Colony Optimization metaheuristic for DSNSF creation using the clustering approach.

Ants, using statistics and probabilities, travel through the search space represented by a graph $G(V, E)$, in which V is a finite set of all nodes and E is the edges set. These agents are attracted to more favorable locations to optimize an objective function, in other words, those in which the concentration of pheromone deposited by ants which previously went through the same path is higher. In this paper, we assume that the paths are formed between the center of a cluster (centroid) and each flow which will be clustered.

The activities performed by ACODS for the DSNSF creating are classified into three categories:

- *Build solutions*: This step consists of the movement of ants concurrently and asynchronously by the states of the problem. It is determined by moving agents from one node to another neighbor in the graph structure.
- *Local Search*: It aims to test and evaluate solutions created by ants through a local search. In our model, this activity is used to remove not unpromising portions of solutions.
- *Pheromone Update*: This is the process in which the pheromone trails are modified. The trails' values can be incremented (when ants deposit pheromones in the edge or connections between the used nodes) or can be decremented. The increased concentration of pheromone is an essential factor in the algorithm implementation since it directs ants to seek new locations more prone to acquire a near-optimal solution.

The ACODS presented in this paper is shown Algorithm 2 and aims to optimize the efficiency of clustering, minimizing the objective function value J (1), in other words, it seeks solutions to the grouping data in a way that allows the extraction of patterns, behaviors and characteristics (Colanzi et al., 2010). Thus, this ensures that each flow i will be grouped to the best cluster j in which $j=1,\ldots, K$. In addition, it enables the construction of solutions that are not given by local optimal, which is the existing problem in some clustering algorithms.

$$J = \sum_{i=1}^{E}\sum_{j=1}^{K}\sqrt{\sum_{a=1}^{A}\left(x_{ia}-c_{ja}\right)^2} \tag{1}$$

in which E represents the number of flows to be clustered and A indicates data dimensionality, i.e., the number of flow features to be processed. The collected flows are divided into 1 min intervals, totaling 1440 data sets throughout the day. It is necessary to generate the DSNSF with data granularity sufficient to be used for ADTW and to provide effective anomaly detection. The variable x_{ia} denotes the value of the feature a of flow i and c_{ja} stores value of the center of cluster j at a dimension.

Algorithm 2. ACODS algorithm used for DSNSF creation.

Require: Set of bits/s, packets/s or flows/s collected from historic database, number of clusters
Ensure: μ: Vector representing the normal behavior for bits/s, packets/s or flows/s sets of a day, arranged in 1440 intervals of 1 min, i.e. the DSNSF.

```
for t=1–1440 do
  while Stopping condition is not reached do
    Create Solution
    Evaluate solutions through the objective function
    Update the pheromone trail
end while
Calculate center of each cluster of the best solution found
for j=1 to number of cluster do
    if number of elements in the cluster C_j < γ then
        Discard the cluster C_j
    end if
  end for
  X_i → Weighted average among the clusters
end for
return X
```

The result of Algorithm 2 is the value which describes the combination of the most representative clusters. To obtain this value, the weighted average is calculated among them. Thus, the result will be closer to the cluster center with the highest number of elements, i.e., the cluster which best represents the data behavior collected at intervals of one minute. Additional information about the characterization process is presented in Assis et al. (2013).

After the clustering process, groups of similar data are formed. Due to the high similarity of network traffic, most data presents similar behavior. Thus, on *Local Search* step, the clusters formed by small amounts of data that greatly deviate from the pattern should be rejected from the signature construction. Therefore, a lower limit is set, γ, which determines the minimum allowable proportion of flows grouped into a cluster. If any cluster has fewer objects than stipulated by γ, it is dropped from the final solution, as well as flows belonging to it. This strategy ensures an uninvolved or minimized, in the worst case scenario, of the anomalous traffic in the DSNSF composition.

4.3. Anomaly detection—adaptive Dynamic Time Warping

In order to find traffic behaviors which are different from DSNSF, a similarity measure should be adopted. The Euclidean distance between each point of the same index has been widely used in time series for this purpose (Esling and Agon, 2012). However, this metric is not suitable for identifying shifts in data sequence. Thus, given two time series, one of them shifted on the

time axis, it is possible for the calculation of the Euclidean distance to consider totally different series. Believing that normal traffic behavior can suffer such displacements due to the changes in the schedule of users' activities, we developed an adaptive similarity measure to fit these situations.

Dynamic Time Warping (DTW) is a pattern matching technique widely used in speech recognition utilized to find an optimal alignment between two series, where one may present alterations being partially elongated or shortened relative to other, along the time axis (Sakoe and Chiba, 1978). Assuming the analysis of DSNSF series denoted by $X=\{x_1, x_2, \ldots, x_n\}$ and real traffic series $Y=\{y_1, y_2, \ldots, y_m\}$, the DTW result can be given by a correlation factor between the two series, calculated after alignment. For this measure, when result is closer to zero, the input sequences are more similar. Another way to obtain the DTW result is by a graphical representation, provided using a matrix of size $n \times m$ where the axes denote the analyzed series. Using this approach, the algorithm creates an optimal path alignment, ω, between the input sequences, minimizing the distortion D expressed by

$$D = \sum_{n=1}^{nm} d(X(n), Y(m)), \tag{2}$$

where $m = \omega(n)$ and each element (i, j) contains the distance d between the points (x_i, y_j), calculated as observed in Sakoe and Chiba (1978).

The DTW calculation of the optimal alignment to compare the time series using is given by four basic steps:

Step 1: Create the solution matrix S, which must consists of n rows and m columns, wherein each element in row i and column j represents the modulus of difference between each interval of comparative series, since n corresponds to the DSNSF length and m corresponds to the length of the time series that describes the real traffic.

Step 2: Establish the Accumulated Distance matrix (DA), composed of n rows and m columns. This matrix is given by the sum of its own values with the upper element of the solution matrix as follows:

$$AD_{i,j} = AD_{i-1,j} + S_{i,j} \text{for} i > 1, j > 1 \tag{3}$$

Step 3: Create of the dimensions movement matrix, composed of n rows and m columns. This matrix must be initiated by assigning the value zero to the last element of the first column. Therefore, an iteration towards bottom-up should be performed in AD matrix, to know which the lowest value is. If the lowest value is below the current element of the iteration, the movement matrix must be filled with value 1; if the lowest value in AD is on the left, the matrix motion should receive the value 3. Finally, if the smallest value is in the left inferior diagonal or the values are equal, the attributed value to the movement matrix current element is 2.

Step 4: Create the best path matrix w. For this purpose, the movement matrix is analyzed from the last element of the first row. Therefore, it is selected element with a smaller distance, d, between other elements, as suggested by

$$d = \min(|w_{i,j} - w_{i-1,j}|, |w_{i,j} - w_{i,j-1}|, |w_{i,j} - w_{i-1,j-1}|), \text{ for } i > 1, j > 1 \tag{4}$$

The Adaptive Dynamic Time Warping (ADTW) approach for anomaly detection is performed at preset time intervals of one minute and consists of two steps. The first one comprises the similarity calculation, S_t, between real traffic and DSNSF at time interval t, using the conventional DTW algorithm. Even small shifts in a series are verified, the results indicate a good match between them because of the time alignment. Until then, only the correspondence found between the shapes of analyzed time series were verified.

In the second step, the distance between the series is calculated, Δ_t, considering their amplitudes. Thus, a subtraction between the average values of both time series is made at the same interval t, as shown in Eq. (5). The result used in the detection of significant changes in network traffic with respect to normal model is calculated normalizing the multiplication between vectors S and Δ as follows:

$$\Delta_t = \overline{Y_t} - \overline{X_t} \tag{5}$$

$$R = \frac{S \times \Delta}{\max(S \times \Delta)'} \tag{6}$$

in which $\Delta=\{\Delta_1, \Delta_2, \ldots, \Delta_t, \ldots\}$ and $S=\{S_1, S_2, \ldots, S_t, \ldots\}$. The goal is to provide a measure based in both form and distance of the series in which they are complementary, e.g., it may be that the result of S_t is close to zero, but the distance between the series at interval t is accentuated. It could be a consequence of a failure or misconfiguration, affecting the normal use of the network, since traffic presents normal behavior but a different intensity. Figure 2(a) exemplifies analysis by ADTW for comparing the series, in contrast to the approach of the Euclidean distance shown in Fig. 2(b).

To improve the detection system efficiency, real traffic movement and DSNSF are evaluated in the same time window t, which comprises a one-minute interval. This approach allows recognition of both punctual anomalies as those which occur over time. Additionally, only an alarm is generated in a time window, ensuring that the administrator is alerted only in event of situations which actually deserve attention.

The flow attributes are analyzed separately, checking the correspondence with the DSNSF created for each of them. A significance coefficient $\Phi=20\%$ is used as a threshold for error between the real traffic and DSNSF at interval t, i.e., R_t. This value is set to compensate for possible inaccuracies occurred during the calculation of r, as well as the small variations of the legitimate use of the network. Moreover, the choice of this value occurred by checking several other thresholds, and this proved to be the most suitable for our application, as can be seen in the session results.

5. Results and discussion

In this section, we evaluate the traffic characterization method used to create the DSNSF, and the accuracy of the presented anomaly detection algorithm.

Aiming to validate whether the proposed methods can operate in a real network environment, we collected IP flows from a core switch at State University of Londrina (Brazil) network, which is composed of about 7000 interconnected devices. Due to the large traffic volume, it was used a sampling rate 1:256, implemented by the collection protocol, sFlow (Phaal et al., 2001).

The collection period comprises two months, starting on September 10 and ending on November 9, 2012. To ease the evaluation, the data set was separated into two groups: The traffic data of first weeks were used by ACODS and PCADS as historical

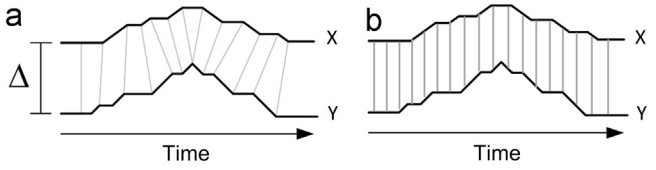

Fig. 2. Comparison schemes of two time series: (a) by using ADTW and (b) by comparing time series using Euclidian distance.

information for DSNSF creating and the workdays of following period—from October 15 to November 9—was used for traffic characterization and anomaly detection evaluation.

Figure 3 illustrates the DSNSFs of November 8 for the three traffic attributes studied in this paper compared with the real traffic observed, each of them describing the 24h of the day. As we can observe, the digital signature curves generated by both PCADS (Fig. 3(a)) and ACODS (Fig. 3(b)) could estimate efficiently the normal behavior of that network segment as there is a great adjustment between the DSNSF and the real traffic.

To measure the accuracy of each method on DSNSFs generation, we adopted two different evaluation metrics: Normalized Mean Square Error (NMSE) and normalized Correlation Coefficient (CC). The Normalized Mean Square Error (NMSE) (Poli and Cirillo, 1993), evaluates the difference between the expected and what was actually verified. This measures' limit is the value zero, which indicates the situation where the expected value is exactly equal to the verified. Thus, higher values of this metric indicate more distant results from the expected.

The normalized Correlation Coefficient (Benesty et al., 2009) indicates the degree of correlation between two variables, as well as the direction of this correlation (positive or negative). The values obtained are within the range of -1 to $+1$. Value 1 indicates total correlation, score 0 (zero) shows that the two variables are not correlated, and -1 specifies a full inverse correlation, that is, where a variable increases, the other decreases, and vice versa.

Figure 4(a) depicts results obtained using the NMSE metric for the traffic of bits/s. As can be seen, PCADS and ACODS achieved similar results, obtaining small errors. October 15 and November 2 presented accentuated errors since they are national holidays, where the network traffic behavior differs from its normal pattern. Another anomalous behavior can be noted in October 30, in which a large traffic volume is observed in all flow attributes analyzed, with peaks up to 56% in excess of the traffic forecasted by DSNSF. It is due to the result of a public tender of the university, which was released on this day, causing a large number of accesses to the university server. Likewise, Fig. 4(b) exhibit NMSE outcomes for the traffic of packets/s and flows/s, and we can observe that both methods reached low errors and resembling results again.

Now, Fig. 5 shows the results relating to bits/s pointed by normalized Correlation Coefficient. The results for bits/s (Fig. 5(a)) ranged between 0.8 and 1 for the two methods. Furthermore, the results relating to packets/s and flows/s (Fig. 5(b)) have lower correlation values, achieving a mean of 0.7. Both PCADS and ACODS produced similar results, which can be classified as strong correlation, as we can observe in Benesty et al. (2009), where authors points out that it occurs in cases where correlation coefficient values are above 0.7.

To properly evaluate our anomaly detection system, we used a tool to artificially inject anomalous events in the real traffic. We simulated anomalous situations in our data set by using a tool named *Scorpius*, which was developed by our network research group (Scorpius, 2013). *Scorpius* is a tool which simulates network anomalies in real traffic flow data, like DoS, DDoS, Port Scan and Flash Crowd, used by our group to help in testing anomaly detection systems. The anomalies are injected in the collected flow data without real and direct intervention in the network, preserving it from impacts caused by anomalies.

Thus, we simulated DoS, DDoS and Flash Crowds in our real data set in order to create a template containing all infected time intervals, aiming to compare it with the alarms generated by our proposed system.

In order to measure the overall efficiency of proposed detection system, we use the Receiver Operating Characteristics (ROC) graph and the Accuracy measure. A ROC graph is defined as a technique to measure the performance of classifiers, being widely used in signal detection theory to describe the trade-off between hit rates (true-positive rates) and false alarm rates (false-positive rates). true-positive rate (TPR) describes correctly detected signals while false-positive rate (FPR) describes how often a signal was detected

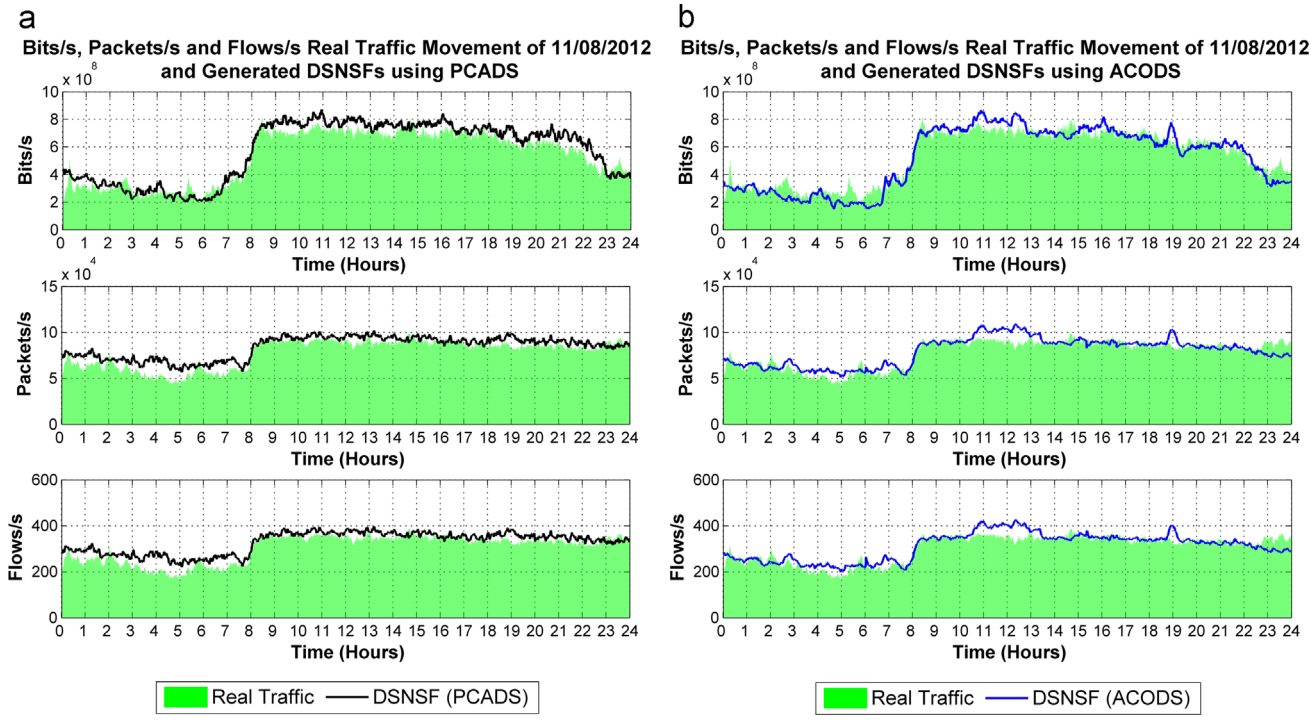

Fig. 3. Traffic characterization example comparing the DSNSFs of bits, packets and number of flows transmitted per second with the real traffic movement observed at November 8 for both PCADS (a) and ACODS (b) methods.

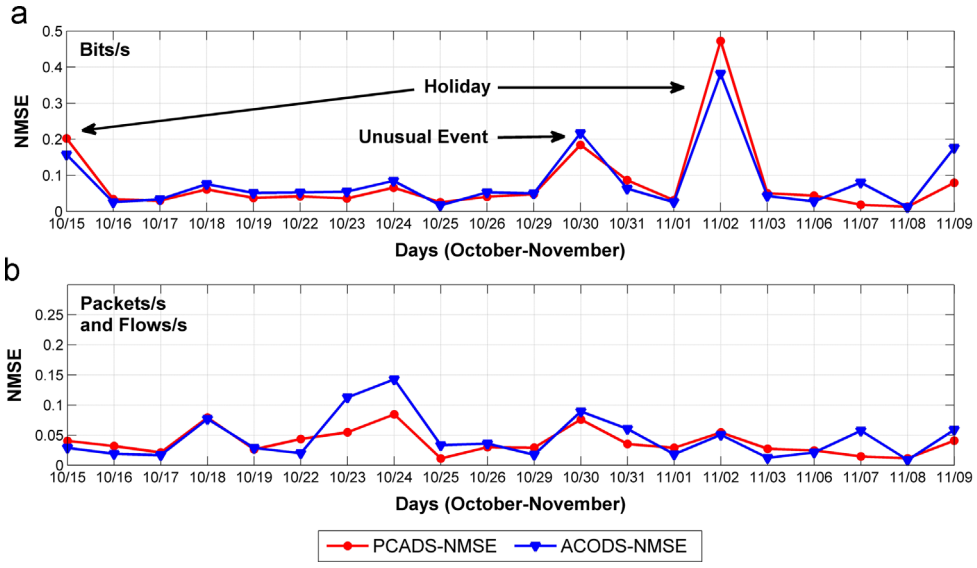

Fig. 4. NMSE indices between the generated DSNSFs and the real traffic movement of analyzed days.

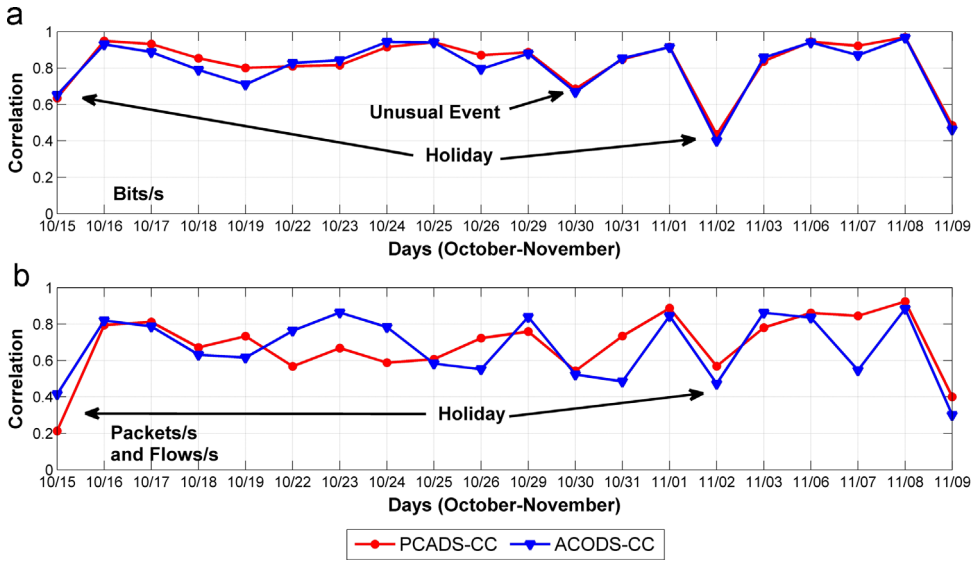

Fig. 5. Correlation Coefficients between the generated DSNSFs and the real traffic movement of analyzed days.

wrongly (Fawcett, 2005). Then, the Accuracy measure describes the overall hit rate, i.e., the hit rate of both classes (positive and negative).

Figure 6 shows the ROC graphs calculated for both PCADS (Fig. 6(a)) and ACODS (Fig. 6(b)) for the four weeks selected to study (from October 15 to November 9) with the artificial anomalous behaviors. Also, we tested the detection using different Φ values, aiming to verify whether it is the best threshold for detecting events that differentiate it from the DSNSF. As seen, our approach was able to recognize a higher percentage of intervals containing anomalous traffic behavior using smaller values of coefficient significance. It is important since Φ cannot be large enough so that anomalous behaviors are classified as normal, because only anomalies which cause great impact on the flow attributes behavior would be recognized.

Several other values were examined for Φ; however, it is clear that values lower than 15% have worse results for TPR and FPR. This occurs due to the reduction of threshold, allowing minimal traffic variation in relation to DSNSF, including legitimate behaviors, to be wrongly characterized as anomalies. On the other hand, when this value is higher than 20%, anomalies are detected only when their behavior deviates greatly from the traffic pattern established by DSNSF. Through the ROC curve analysis, it can be inferred that such a situation causes lower rates of true-positive, and this deficiency is enhanced while the value of Φ increases. In general, as can be seen in the zooms in Fig. 6, the one based on PCADS performed better than ACODS, with reduced false-positive rates, reaching a trade-off of 92% TPR with 21% FPR, as ACODS reaches 92% TPR with 24% FPR.

In Fig. 7, the Accuracy measure for the same study period of four weeks and the same comparison using different Φ values performed in Fig. 6 is shown. This measure is the proportion of true results (true-positive and true-negative). Both systems produced better accuracy rates when Φ=20%, perceiving that for Φ values above or below this threshold, the accuracy rate begins to decrease. So, ω=20%, both systems achieved worthy results, obtaining an average accuracy rate of 96%.

After an anomalous event is detected, our detection system can provide the network manager a detailed report about that time interval, containing information like IP source and destination addresses, and origin and destination ports. These descriptive

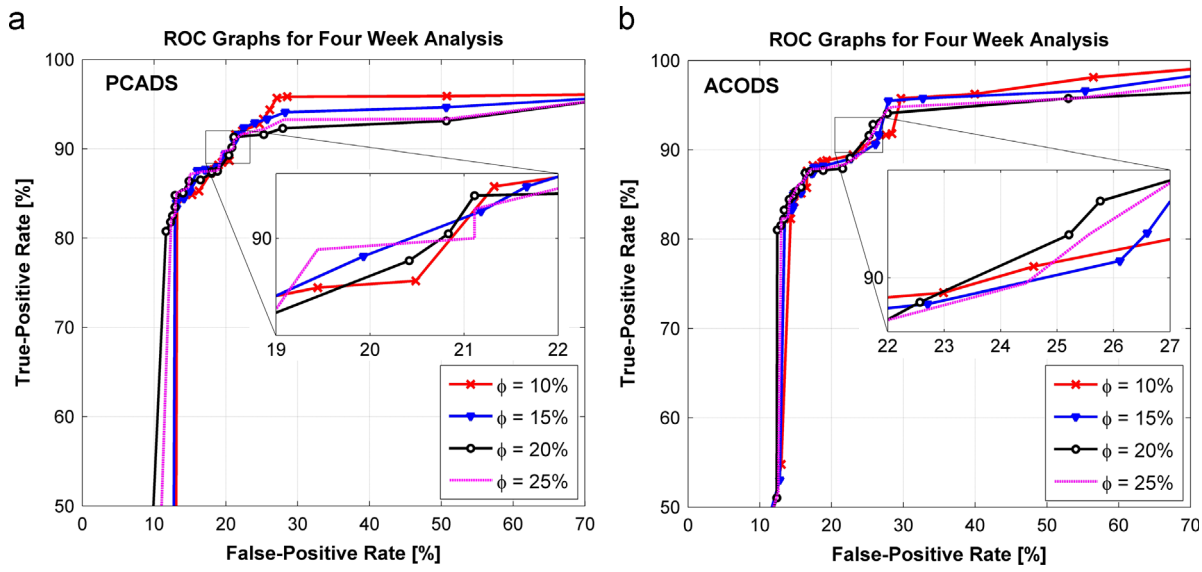

Fig. 6. ROC curve of workdays from October 15 to November 9 for both PCADS and ACODS using different Φ values.

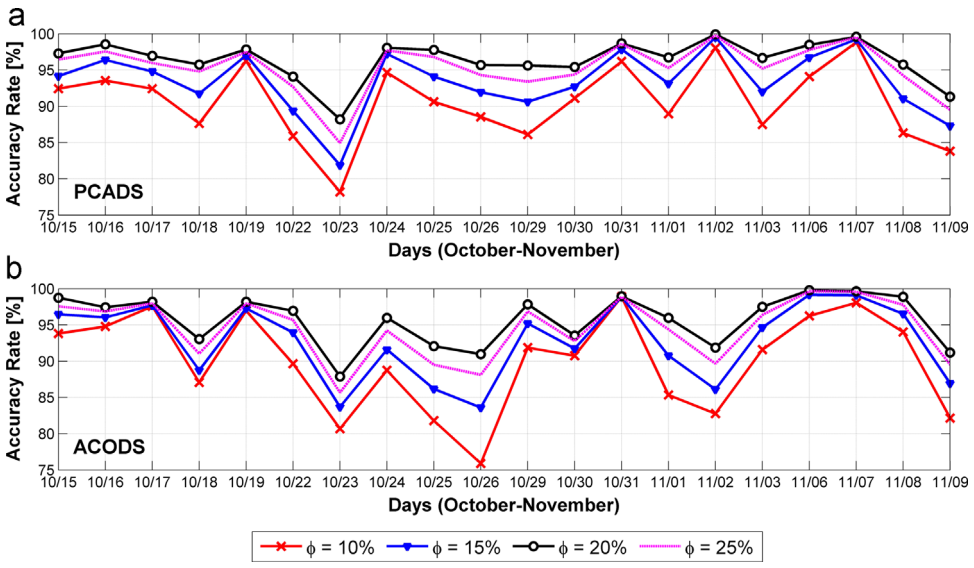

Fig. 7. Accuracy Rate of four weeks of tests for both PCADS and ACODS using different Φ values.

attributes may unveil where the problem ensued, or who caused it, as well as what kind of application was targeted.

To demonstrate this service, we selected two days with artificial anomalies from our data set, both using fictitious IP addresses and Port numbers. Figure 8 presents a top 3 list of traffic statistics of a Flash Crowd (a) and a DDoS attack (b). The top 1 destination IP addresses and Ports identified by our system are actually the fictitious attributes used in the anomaly simulation. We can observe that an anomalous situation affects a large traffic flows proportion when compared to normal activity not just related to volume, but likewise in descriptive attributes.

5.1. Complexity analysis

The computational complexity of PCADS algorithm is limited by the asymptotic notation $O(np^2)$. This is because the costly operation of PCADS is the calculus of all the principal components of a $n \times p$ matrix X, which is achieved by solving the symmetric eigenvalue problem for a covariance matrix X^T (Lakhina et al., 2004). Then, in order to solve this problem, it is necessary to compute the Singular Value Decomposition (SVD), a method used to obtain the eigenvectors and eigenvalues of a matrix X. Therefore, the computational complexity of a complete SVD of a $n \times p$ matrix is limited by $O(np^2)$ (Parlett, 1998; Biglieri and Yao, 1989).

The computational complexity of ACODS algorithm is presented as an asymptotic notation based on the number of instructions executed. Firstly, the split of an initial data set of E elements in K centers of size A, results in $O(EKA)$. By using the ant population to aid in searching for the best data grouping centers, and as they are all compared to each other to find the final solution, a quadratic complexity is added, resulting in $O(EKAM^2)$. At last, by taking into account the number of iterations I as stopping condition of the algorithm, we have a final complexity of $O(EKAM^2I)$. To generate a DSNSF of 24h, this procedure is repeated 1440 times in a loop. Although a maximum number of iterations I is set, ACODS converges rapidly to the solution.

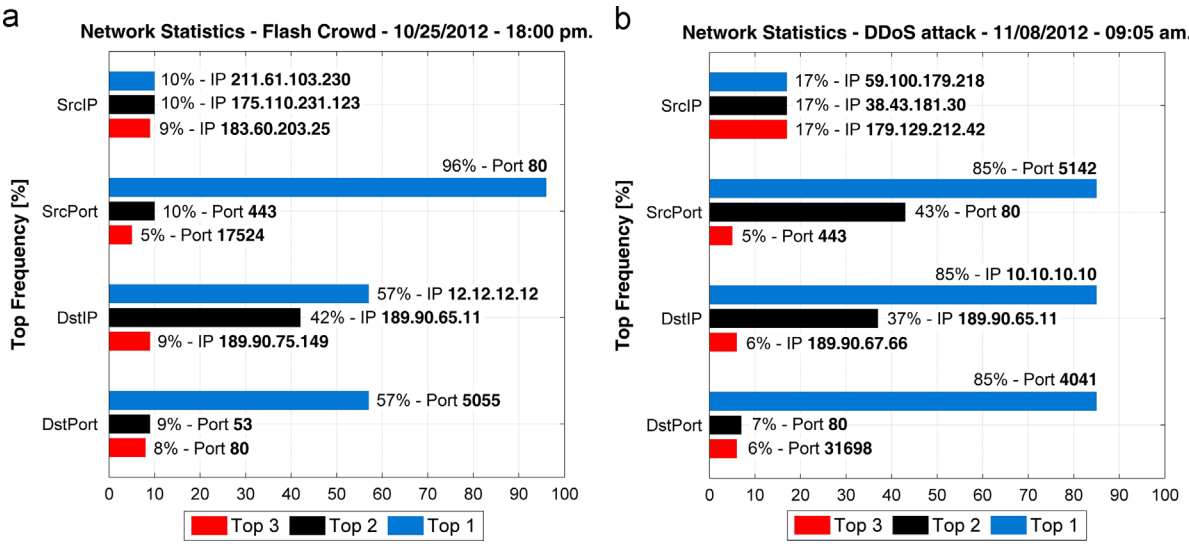

Fig. 8. Network traffic statistics from two kinds of anomalies.

Regarding the ADTW procedure, the computational complexity is the same as the traditional Dynamic Time Warping algorithm, which is $O(E^2)$ (Esling and Agon, 2012).

6. Conclusion and future work

In this paper, we present and evaluate two profile-based anomaly detection systems to help network management. The major contribution consists in the application and contextualization of Principal Component Analysis, Ant Colony Optimization and Dynamic Time Warping methods to an environment of pattern recognition and anomaly detection. It also stands out as a contribution the analysis of seven IP flow attributes, where: (i) three quantitative attributes—bits, number of packets and flows—are used in order to characterize the network traffic through DSNSF generation, a key to effectively identify anomalous behaviors and (ii) four descriptive attributes—source IP, destination IP, TCP/UDP source port and TCP/UDP destination port—which are used by the Information Module to provide the network manager information needed to identify the problem and take specific measures against it.

Regarding Traffic Characterization module, we compared two different methods, PCADS and ACODS. According to NMSE and Correlation Coefficient results, both accomplished similar results, leading to good traffic predictions, since we can verify small errors between the DSNSF and the real traffic in Figs. 4 and 5.

In the Detection and Identification module, the Adaptive DTW (ADTW) algorithm investigated in this work had satisfactory performance pertaining to false alarm rates. Concerning that subject, both systems produced better results when adjusting the ADTW Φ value to 20%. Moreover, by analyzing the ROC graphs and Accuracy rates, PCADS performed better than ACODS. Moreover, the correspondence between true-positive and false-positive rates demonstrates that the systems are able to enhance the detection of anomalous behavior by maintaining a satisfactory false-alarm rate. In addition, as presented in Fig. 8, our anomaly detection methodology can supply the network administrator with important traffic statistics in order to help in problems solution, aiming for accurate and fast anomaly detection.

Therefore, we conclude that the proposed methodology, by using PCADS, ACODS and ADTW, is suitable to help network management, detecting traffic anomalies and consequently, supplying availability and reliability to networks and their provided services.

In future work, we plan to increase the number of analyzed flow attributes, explore the correlation between them for a more effective detection and investigate different classes of anomalies.

Acknowledgments

This work was supported by CNPq due to the concession of scholarship by process 249794/2013-6; SETI/Fundação Araucária for Betelgeuse Project's financial support process 41939.410.32989.30092013; *Instituto de Telecomunicações*, Next Generation Networks and Applications Group (NetGNA), Portugal; and by National Funding from the FCT-Fundação para a Ciência e a Tecnologia through the UID/EEA/500008/2013 Project.

References

Abadi M, Jalali S. An ant colony optimization algorithm for network vulnerability analysis. Iran J Electr Electron Eng 2006;6(3):106–20.

Assis MVO, Rodrigues JJPC, Proença Jr. ML. A seven-dimensional flow analysis to help autonomous network management. Inf Sci 2014;278:900–13. http://dx.doi.org/10.1016/j.ins.2014.03.102.

Assis MVO, Carvalho LF, Rodrigues JJPC, Proença Jr ML. Holt-winters statistical forecasting and aco metaheuristic for traffic characterization. In: 2013 IEEE International Conference on Communications (ICC'13), 9 – 13 June 2013. Budapest, Hungary.

Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise reduction in speech processing. Berlin, Heidelberg: Springer; 1–4.

Biglieri E, Yao K. Some properties of singular value decomposition and their applications to digital signal processing. Signal Process 1989;18(3):277–89. http://dx.doi.org/10.1016/0165 – 1684(89)90039-X.

Bhuyan MH, Bhattacharyya DK, Kalita JK. Network anomaly detection: methods, systems and tools. *Commun Surv Tutor* IEEE 2014;16(1):303–36. http://dx.doi.org/10.1109/SURV.2013.052213.00046.

Callegari C, Gazzarrini L, Giordano S, Pagano M, Pepe T. A novel PCA-based network anomaly detection. In: 2011 IEEE International Conference on Communications (ICC), 5 – 9 June 2011. Kyoto, Japan. IEEE; p. 1–5.

Colanzi TE, Assuncao WKG, Pozo ATR, Vendramin ACBK, Pereira DAB. Empirical studies on application of genetic algorithms and ant colony optimization for data clustering. In: 2010 XXIX International Conference of the Chilean Computer Science Society (SCCC), November 2010; p. 1–10.

Dorigo M, Birattari M, Stutzle T. Ant colony optimization. Comput Intell Mag IEEE 2006;1(4):28–39.

Esling P, Agon C. Time-series data mining. ACM Comput Surv 2012;45(1):1–34. http://dx.doi.org/10.1145/2379776.2379788.

Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2005;27(8):861–74.

Fernandes Jr G, Zacaron AM, Rodrigues JJPC, Proença Jr ML. Digital signature to help network management using principal component analysis and *K*-means

clustering. In: 2013 IEEE International Conference on Communications (ICC'13), 9 – 13 June 2013. Budapest, Hungary.
Jiang D, Xu Z, Zhang P, Zhu T. A transform domain-based anomaly detection approach to network-wide traffic. J Netw Comput Appl 2014;40:292–306. http://dx.doi.org/10.1016/j.jnca.2013.09.014.
Jiang D, Yao C, Xu Z, Qin W. Multi-scale anomaly detection for high-speed network traffic. Trans Emerg Telecommun Technol 2013;26(3):308–17. http://dx.doi.org/10.1002/ett.2619.
Jiang J, Papavassiliou S. Enhancing network traffic prediction and anomaly detection via statistical network traffic separation and combination strategies. Comput Commun 2006;29(10):1627–38.
Jolliffe I. Principal component analysis. New York, NY: Springer Verlag; 2002.
Kanda Y, Fukuda K, Sugawara T. Evaluation of anomaly detection based on sketch and pca. In: Global Telecommunications Conference (GLOBECOM 2010), December 2010. IEEE; p. 1–5.
Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies. In: Proceedings of the 2004 conference on applications, technologies, architectures, and protocols for computer communications, SIGCOMM'04. New York, NY, USA: ACM; 2004. p. 219–30.
Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies. In: Proceedings of the 2005 conference on applications, technologies, architectures, and protocols for computer communications, SIGCOMM'05. New York, NY, USA: ACM; 2005. p. 217–28.
Lima MF, Sampaio LDH, Zarpelão BB, Rodrigues JJPC, Abrao T, Proença ML. Networking anomaly detection using dsns and particle swarm optimization with re-clustering. In: Global Telecommunications Conference (GLOBECOM 2010), December 2010. IEEE; p. 1–6.
Mazel J, Casas P, Labit Y, Owezarski P. Sub-space clustering, inter-clustering results association amp; anomaly correlation for unsupervised network anomaly detection. In: 2011 7th International Conference on Network and Service Management (CNSM), October 2011; IEEE. p. 1–8.
Molnar S, Moczar Z. Three-dimensional characterization of internet flows. In: 2011 IEEE International Conference on Communications (ICC), 5–9 June 2011. Kyoto, Japan. IEEE; p. 1–6.
Parlett BN. The symmetric eigenvalue problem. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.; http://dx.doi.org/10.1137/1.9781611971163.
Patcha A, Park J-M. An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput Netw 2007;51(12):3448–70.
Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag 1901;2(6):559–72.
Phaal P, Panchen S, McKee N. InMon Corporation's sFlow: a method for monitoring traffic in switched and routed networks. 2001. Available at: ⟨http://www.ietf.org/rfc/rfc3176.txt⟩ [accessed 8.10.13].
Poli AA, Cirillo MC. On the use of the normalized mean square error in evaluating dispersion model performance. Atmos Environ Part A. Gen Top 1993;27(15):2427–34.
Proença ML, Zarpelao BB, Mendes LS. Anomaly detection for network servers using digital signature of network segment. In: Advanced industrial conference on telecommunications/service assurance with partial and intermittent resources conference/e-learning on telecommunications workshop, Telecommunications, 2005, 17 – 20 July 2005; IEEE. p. 290–5.
Qin T, Guan X, Li W, Wang P, Huang Q. Monitoring abnormal network traffic based on blind source separation approach. J Netw Comput Appl 2011;34(5):1732–42.
Rawat S, Gulati VP, Pujari AK. Frequency- and ordering-based similarity measure for host-based intrusion detection. Inf Manag Comput Secur 2004;12(4):411–21.
Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process 1978;26(1):43–9.
Scorpius-sFlow Anomaly Simulator. UEL Network Research Group. 2013. Available on: ⟨http://redes.dc.uel.br/scorpius⟩. [last access 19.08.14].
Sodiya AS, Longe HOD, Akinwale AT. A new two-tiered strategy to intrusion detection. Inf Manag Comput Secur 2004;12(1):27–44.
Tartakovsky AG, Polunchenko AS, Sokolov G. Efficient computer network anomaly detection by changepoint detection methods. *IEEE J Sel Top Signal Process* 2013;7(1):4–11. http://dx.doi.org/10.1109/JSTSP.2012.2233713.
Zarpelao BB, Mendes LS, Proenca ML, Rodrigues JJPC. Parameterized anomaly detection system with automatic configuration. In: Global Telecommunications Conference (GLOBECOM 2009), November 30, 2009–December 4, 2009. IEEE; p. 1–6.
Zhang J, Li H, Gao Q, Wang H, Luo Y. Detecting anomalies from big network traffic data using an adaptive detection approach. *Inf Sci* 2014, . http://dx.doi.org/10.1016/j.ins.2014.07.044.
Zhou CV, Leckie C, Karunasekera S. Decentralized multi-dimensional alert correlation for collaborative intrusion detection. J Netw Comput Appl 2009;32(5):1106–23.