Holt-Winters Statistical Forecasting and ACO Metaheuristic for Traffic Characterization

Marcos V. O. de Assis*, Luiz F. Carvalho*, Joel J. P. C. Rodrigues†, Mario Lemes Proença Jr*.
*Computer Science Department, State University of Londrina (UEL), Londrina, Brazil
†Instituto de Telecomunicações, University of Beira Interior, Portugal
{mvoassis, luizfcarvalhoo}@gmail.com, joeljr@ieee.org, proenca@uel.br

Abstract—Due to modernization, expansion of computer networks has become an inevitable process. However, this growth is also accompanied by increased complexity, which makes it necessary to use resources that assist the management of these networks. In this paper, we propose a traffic characterization using two-dimensional flow analysis for modeling the behavior traffic pattern, here called Digital Signature of Network Segment Using Flow Analysis (DSNSF). To accomplish this task we have used the improved Holt-Winters forecasting and Ant Colony Optimization metaheuristic methods. The DSNSF obtained by each model are compared to a real traffic of packets and bits and then subjected to specific evaluations in order to measure its accuracy.

Keywords—DSNSF; Holt-Winters; Ant Colony Optimization; NetFlow; Traffic Characterization; Network Management.

I. INTRODUCTION

Nowadays computer networks perform an essential role due to the agility and efficiency they provide to message and information exchange, improving people's interaction. Enterprise environment computer networks are composed by a large number of devices and, hence, depends on its fully operation. Due to the enormous amount of information, the manual traffic analysis of huge networks can become complex, error-prone and, consequently, inefficient for network administrators.

Thus, efficient management methods and tools capable of detecting changes on the normal behavior of computer networks in an agile and independent of human supervision way are needed [1]. These aspects characterize the so-called Autonomic Management [2], which can be defined as the automation of the anomaly detection processes in computer networks. Therefore, network administrators can quickly correct failures or interrupt attacks pointed out by alarms, enabling the network operation to be minimally damaged.

In recent years, several studies have been developed in this research field, where different methods and techniques are applied aiming to provide efficient means of anomalies detection in large scale networks. A strong approach of these studies addresses the networks traffic characterization [3], [4], [5], [6]. Through different types of methods (statistical, metaheuristic, among others), there are developed models that characterize the operation of these networks, process here referred as Digital Signatures of Network Segments based on Flow analysis (DSNSF). The DSNSF can be defined as followed pattern or a set of basic information that describe the traffic profile of network segment.

This paper aims to analyze the effectiveness of using two different methods in generating the DSNSF. Methods used are Holt-Winters for Digital Signature (HWDS), a modification of the classic statistical method of forecasting Holt-Winters, and Ant Colony Optimization for Digital Signature (ACODS), meta-heuristic which aims to optimize the clustering processes used in pattern extraction.

The DSNSFs use flow analysis for characterizing the described network. An IP flow can be defined as a unidirectional sequence of N IP packets with similar attributes. These attributes (origin and destination IP addresses and ports, number of packets and bits, protocols, and others) record all kinds of IP communication performed through the network, enabling characterization processes loyal to the real behavior.

Previous studies, although plentiful, are mostly focused on the analysis of a single flow characteristic of the networks. However, according to [7], the use of multiple characteristics (or dimensions) simultaneously helps the characterization process of networks behavior, especially when verified the existence of a correlation between them.

The DSNSFs generated by the mentioned methods characterize the network through two dimensions of the analyzed flows: bits and packets per second. For testing the efficiency of these methods, we collected real flows from the Federal University of Technology Paraná - Campus Toledo. We used data concerning two months of flows collection for the initial training of the methods, followed by two weeks used in the developed efficiency tests.

The remainder of this paper is composed as follows: Section II explains the two characterization methods; Section III exposes the obtained results by comparing the two methods on the DSNSF generation; finally, Section IV concludes the paper.

II. GENERATION OF DSNSF

In this section, we present two methods to create the DSNSF using bits and packets per second, collected from flows generated by network’s assets. This paper presents a study
where the flow attributes bits and packets are used for composition of a normal traffic pattern. The methods belong to different groups of algorithms. The first is a derivation of the prediction model Holt-Winters and the second is based on the metaheuristic Ant Colony Optimization. The description of both methods will present the aspects that make them suitable for the DSNSF construction.

A. Holt-Winters for Digital Signature

The Holt-Winters is a statistical method of forecasting applied to time series characterized by the presence of linear trend and seasonality (periodicity), which is based on the Exponential Weight Moving Average method (EWMA). This method works dividing the analyzed data in three parts, each being represented by an equation of the EWMA type. They are the baseline ("a"), the linear trend ("b") and the seasonality trend ("c") [8]. Thus, the prediction $y'_{t+1}$, obtained through the analysis of $y_n$, is given by:

$$y'_{t+1} = a_t + b_t + c_{t+1-s}$$  (1)

In equation (1), the variable "s" indicates the analyzed time interval, "n" is the number of time windows which the analyzed time series have (seasonal cycle), "$y_n$" is the value of the time series at the instant $t$ and "$y'_{t+1}$" is the calculated forecasting for the instant $t+1$. The three parts that compose the forecasting are obtained through EWMA equations:

$$a_t = \alpha (y_t - c_{t-s}) + (1-\alpha)(a_{t-1} + b_{t-1})$$  (2)

$$b_t = \beta (a_t - a_{t-1}) + (1-\beta) b_{t-1}$$  (3)

$$c_t = \gamma (y_t - a_t) + (1-\gamma) c_{t+1-s}$$  (4)

The EWMA equations are characterized by their division into two parts, each receiving a weight at the final result. This weight is assigned through the use of the variables $\alpha$, $\beta$ and $\gamma$, enabling the creation of smoother forecasts. The values assigned to the variables $\alpha$, $\beta$ and $\gamma$ directly influence the behavior of the prediction. They are named "smoothing coefficients", and their values must necessarily belong to the interval $0 < \alpha, \beta, \gamma < 1$. This adjustability allows a faster or slower conversion of the DSNSF, according to the specific requirements of the analyzed network segment.

In this work, we used a modification of the Holt-Winters traditional method, named Holt-Winters for Digital Signature (HWDS), aiming to achieve better results related to the creation of DSNSF. Thus, modifications were made in the equations that describe the components baseline (2) and linear trend (3). To achieve a better adaptation of the normal networks' behavior, a similar approach to the one used for seasonal trend (4) is implemented. Therefore, instead of using the time intervals immediately before the analyzed ones for the forecasting calculation, the time intervals that are equal to the current and relating to the prior seasonal cycle are used:

$$a_t = \alpha (y_t - c_{t-s}) + (1-\alpha)(a_{t-s} + b_{t-s})$$  (5)

$$b_t = \beta (a_t - a_{t-s}) + (1-\beta) b_{t-s}$$  (6)

The proposed change is justified because, in similar days, equal time series windows tend to be similar, which does not occur with consecutive windows of a same time series, improving the DSNSFs creation results. The basic operation of the HWDS method can be observed at the Algorithm 1.

### Algorithm 1 – HWDS used to DSNSF creating

**Require:** Set of bits and packets collected from historic database, preset value for the smoothing coefficients.

**Ensure:** $\mu$: Vector representing the bits and packet sets of a day, which is divided in 288 intervals of 5 minutes, in other words, it is the DSNSF values for the next day.

1: for $i = 1 : 288$ do
2: Calculate the baseline for $i$ (5)
3: Calculate the linear trend for $i$ (6)
4: Calculate the seasonal trend for $i$ (4)
5: Calculate the forecasting for $i+1$ (1)
6: $\mu_i \leftarrow$ calculated forecast
7: end for
8: return $\mu$

The creation of DSNSF from the HWDS prediction method occurs dynamically, i.e., the generated DSNSF undergoes changes every new day processed. Thus, data are used relating two different days for the prediction calculus of how the third one should behave: the current day data and the DSNSF generated for the prior day. This characteristic allows for the method to adapt to gradual changes occurred in a simplified way, without the need to perform any adjustments in the method.

B. Ant Colony Optimization for Digital Signature

The Ant Colony Optimization (ACO) is based on the principles of swarm intelligence, which are defined population of agents competing and globally asynchronous, cooperating to find an optimal solution [9][10].

The ACO for Digital Signature (ACODS) described in this paper, is presented in Algorithm 2 and aims to optimize the efficiency of clustering minimizing the objective function value $J$ (7), in other words, it seeks solutions to the grouping data in a way that allows the extraction of patterns, behaviors and characteristics [11]. Thus, this ensures that each element i will be grouped to the best cluster $j$ in which $j = 1, ..., K$. In addition, it enables the construction of solutions that are not given by local optimal, which is the existing problem in some clustering algorithms.

$$J = \sum_{i=1}^{E} \sum_{j=1}^{K} \sum_{a=1}^{A} (x_{ia} - c_{ja})^2$$  (7)

in which $E$ represents the amount flows to be clustered and $A$ indicates data dimensionality, i.e., number of features to be processed. The collected flows are divided in 5 minute intervals, resulting in 288 data sets throughout the day. The variable $x_{ia}$ denotes value of the feature $a$ of flow $i$ and $c_{ja}$ stores value of cluster's center $j$ at a dimension.
Algorithm 2 – ACODS used to DSNSF creating

Require: Set of bits and packets collected from historic database, number of clusters
Ensure: $\mu$: Vector representing the bits and packet sets of a day arranged in 288 intervals of 5 minutes, i.e. the DSNSF

1: for $i = 1 : 288$ do
2: while stopping condition is not reached do
3: Create solution
4: Update pheromone trail
5: Evaluate solutions through the objective function (7)
6: end while
7: Calculate the center of each cluster of the best solution found
8: for $j = 1 : number$ of clusters do
9: if number of elements in the cluster $C_j < \lambda$ then
10: Discard the cluster $C_j$
11: end if
12: end for
13: $\mu_i \leftarrow weight$ed average between the clusters
14: end for
15: return $\mu$

The result of Algorithm 2 is the value that describes the combination of the most representative clusters. To obtain this value, the weighted average is calculated between the clusters. Thus, the result will be closer to the cluster center with the highest number of elements, i.e., the cluster that best represents the data behavior collected at intervals of five minutes.

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

III. RESULTS OBTAINED

The analyzed flows were collected from Federal University of Technology Paraná – Toledo Campus. The flows are exported from the main gateway of a Softflowd application, which is a network analyzer capable of exporting flows in NetFlow v9 pattern. After collection, these data are saved in LOG files every five minutes and used for DSNSF creation.

In this work, we use a data set consisting of ten weeks of 2012 for the creation and evaluation of DSNSFs. Flows belonging to March and April months are used for HWDS and ACODS training, whereas data from 1st to 18th of May are used to measure the effectiveness of the DSNSFs traffic characterization. It is important to highlight that May 1st, first day used for the mentioned measurements, is a national holiday. Thus, even with the behavior of the analyzed network segment being anomalous in this day, we decided to keep it in the analysis to emphasize the adaptability of the methods to similar situations.

The methods discussed and used in DSNSFs generation are adjusted through specific parameters, which objectives to define their behavior in order to optimize the obtained results. HWDS, as discussed previously, has its operation based on the smoothing coefficients, which must be small to ensure that historical information has a greater influence than new entries in the generated DSNSF. These coefficients were defined with values $\alpha = 0.28$, $\beta = 0.0035$ and $\gamma = 0.1$, based on exhaustive efficiency tests and prior works [8][12]. Furthermore, the variable "s" representing the seasonal cycle of the time series has the value 288, because the collected data were analyzed as five minute intervals, generating 288 different intervals each day. In the ACODS, the variable $\lambda$ and the number of agents are configured by taking into account the number of elements to be clustered. The $\lambda$ variable represents 5% of the total whereas the number of agents is 15%. The number of clusters, $K$, initially used, was chosen by the Silhouette method of interpretation and validation [13], achieving better results when this variable value is 3.

Once trained, the HWDS and ACODS methods were used to represent 18 days' behavior of May. Fig. 1 and 2 presents the real traffic collected on 16th May along with the DSNSFs generated by the discussed methods for the attributes bits/s and packets/s, respectively. As can be seen, the traffic characterization generated by both methods can, in general, efficiently describe the behavior observed at the analyzed day, where there is a greater use of the network resources during the periods from 7:30 to 12:00 hours and from 13:00 to 18:00 hours. The results obtained through the use of the HWDS and ACODS methods are visually very similar. Furthermore, it can be observed a high correlation between the analyzed attributes, since the trends presented by bits/s DSNSFs follow the packets/s DSNSFs ones of the same day.

As a means of measuring in parallel the efficiency of the discussed methods in generating DSNSFs we adopted three different metrics, which were fully applied for 3 weeks of data intended for testing: calculus of the Correlation Coefficient (CC), Normalized Mean Squared Error (NMSE) and Symmetric Mean Absolute Percentage Error (sMAPE) [14].

The Correlation Coefficient calculus is a metric that provides the similarity degree of two random variables, in this case, the actual observed traffic and the generated DSNSF. Its results may vary between -1 to 1, where 1 indicates a total correlation, -1 an inverse total correlation and 0 a null correlation. Fig. 3 and 4 presents the obtained results through this technique, applied over the test set. As can be observed, the correlation coefficients of both methods are very similar, achieving values higher than 0.75. These values indicate a high correlation between the DSNSFs generated and the observed movements of the actual traffic.

However, correlation tests only indicate if the trends followed by the movement of the observed traffic and the generated DSNSF are related, measuring this relation degree. To measure the distance between them it is used the Normalized Mean Square Error (NMSE), metric that indicates how close the generated DSNSF is from the actual traffic.
The 0 value means a null error between the inputs and, therefore, values close to 0 indicate good traffic characterizations.

Fig. 5 and 6 present the obtained results through the NMSE usage over days of the test sets for bits/s and packets/s, respectively. As shown, both methods presents good error rates, achieving averages NMSE values of 0.5, and the HWDS method fared slightly better.

For accuracy analysis of time series predictions, the most commonly used metrics are relating to errors. In addition to NMSE, it is used the Symmetric Mean Absolute Percentage Error (sMAPE) [14], a modification of the traditional techniqe MAPE, which considers the errors in a symmetrical way, allowing a better analysis of the obtained outcomes.

The result of this measure is a percentage error value, which does not have preset values to describe good or bad predictions. For this paper’s environment of development, it was observed that sMAPE values between 10% and 20% represent very good forecasts, while values between 20% and 30% describe relatively good predictions on the average. Fig. 7 and 8 present the obtained percentage errors at the days of the test sets for bits/s and packets/s through the sMAPE technique. As observed, both methods achieved, on the average, percentage errors between 20% and 30%, indicating that the generated DSNSFs are capable of describing the observed traffic relatively well. Regarding the HWDS and ACODS methods, the differences between their results are small, where HWDS fared 0.7% better on Packets/s DSNSFs, and ACODS fared 2.2% better on bits/s DSNSFs.
In terms of computational complexity between the two methods, we used asymptotic notation to indicate the number of executed instructions for DSNSF creation. The ACODS complexity is given first by partitioning a set $E$ of initial data by $K$ centers of $A$ dimensions. Using the population of ants to assist the search of the best centers for the collation of data and, like all the ants are compared with each other in pursuit of the final solution, a quadratic complexity is added. Taking number of iterations $I$ into account as stopping criterion of the algorithm, we have a final complexity of $O(\text{EKAM} I^2)$. For generating a DSNSF of one day, this process is repeated $N$ times, where $N$ is the number of analyzed windows at the time series which describes the behavior of the current day (algorithms input). Although a maximum of interactions $I$ is defined, ACODS quickly converges to solution.

The HWDS presents an initial $\theta(N)$ complexity, which is the result of the calculation variables: baseline, linear trend, seasonal trend and the forecasting. For traffic characterization this process is repeated $N$ times, resulting in final complexity of $\theta(N)$. Thus, scalability, regarding the increased amount of flows attributes processed and network segments analyzed, does not affect significantly HWDS performance. However, for the ACODS method, the network administrator should estimate to what extent it is convenient to increase the amount of data computed for the DSNSF generation.

IV. CONCLUSIONS

In this paper we presented two methods for traffic characterization (DSNSF) in an autonomic way using two-dimensional flow analysis, aiming to help network management. The first method is the improved Holt-Winters forecasting, named HWDS, and the second one is derived from the Ant Colony Optimization metaheuristic, called ACODS. We explored three evaluation techniques to compare the accuracy of the discussed methods. The obtained results of the correlation and NMSE tests between the DSNSFs and the real traffic were superior to 0.75 and not higher then 0.5, respectively, for both methods. Furthermore, the results of the time series prediction test, sMAPE, present errors between 20% and 30%, which denotes a good adaptability. Thus, the generated DSNSFs show that the characterization can follow the trends and make predictions of network traffic effectively, demonstrating that the discussed methods are promising cores for autonomic tools of network management and anomaly detection. Although the results of these tests are similar for both methods, the HWDS presents a lower computational complexity than the ACODS.

In future works, we intend to develop a tool for anomaly detection using the generated DSNSFs. We also plan to increase the amount of analyzed flow attributes and explore the correlation between them for more effective detection.

ACKNOWLEDGMENT

This work was supported by CAPES, due to the concession of scholarships; SETI/Fundação Araucária and MCT/CNPq for Rigel and Bellatrix Projects' financial support; 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 PEst-OE/EEI/LA0008/2011 Project.

REFERENCES