Anomaly Detection Using DSNS and Firefly Harmonic Clustering Algorithm

Mario H. A. C. Adaniya*, Moisés F. Lima*, Joel J. P. C. Rodrigues†, Taufik Abrão* and Mario Lemes Proença Jr.*
*Department of Computer Science, State University of Londrina (UEL), Londrina, Brazil
†Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal
E-mails: {mhadaniya, moisesflima}@gmail.com, joeljr@ieee.org, {taufik, proenca}@uel.br

***Abstract*—The networks are becoming an essential part of society life and anomalies may represent a loss in network performance. Modeling the traffic behavior pattern is possible to predict the behavior expected and characterize an anomaly. We proposed a hybrid clustering algorithm, Firefly Harmonic Clustering Algorithm (FHCA), for network volume anomaly detection by the combined forces of the algorithms K-Harmonic means (KHM) and Firefly Algorithm (FA). Processing the Digital Signature of Network Segment (DSNS) data and real traffic data, it is possible to detect and point intervals considered anomalous with a trade-off between the 80% true-positive rate and 20% false-positive rate.**

I. Introduction

The Internet has brought to our daily life easy and new ways to do tasks as searching and gathering information, to communicate and spread ideas and others small gestures that are changing our daily life. To prevent possible failures and loss of performance the infrastructure providing theses services must be monitored. An anomaly behavior can be caused from a simple programming error in some software to hardware failure, among many other causes that affect directly the network operation.

In [1], the authors grouped the existing techniques into different categories according to the key concept. The authors defined a straightforward anomaly detection approach as defining a region representing normal behavior and declaring any observation in the data which does not belong to this normal region as an anomaly. In our work, we consider an volume anomaly anything that is outside a threshold value of the Digital Signature of Network Segment (DSNS) generated through GBA (Automatic Backbone Management) tool presented in [2]–[4] and briefly described in section III. The anomaly context adopted is described in section IV.

In this work, our proposal to detect volume anomalies is the Firefly Harmonic Clustering Algorithm (FHCA), which is the joint of DSNS, K-Harmonic means (KHM) [6] and Firefly Algorithm (FA) [7]. The FHCA cluster the DSNS data and the network traffic sample generating K_D and K_T centers respectively. Assuming the DSNS as normal traffic predicted, the comparison between the K_D and K_T is used to labeled if there is an anomaly or not. Exploring the advantages of KHM and the FA, the proposed algorithm applied to detection of volume anomalies in real network traffic achieve a trade-off between the 80% true-positive rate and 20% false-positive rate.

In section II is discussed some related works founded in literature using heuristic, clustering and both techniques applied to anomaly detection. Section III describes the GBA tool and the DSNS. Section IV describes the anomaly context adopted. Section V is the proposed algorithm. Section VI present the results achieved by the proposed algorithm. Section VII present the conclusion and futures improvements.

II. Related Work

Techniques of anomaly detection can be classified according to: statistical, based on classifier, machine learning and using finite state machines [8]. Our model adopt the Digital Signature of Network Segment generated by GBA (Automatic Backbone Management) tool as normal traffic predicted and compare to real traffic sample fitting our model into the based on classifier category according to [8]. Considering another aspect in anomaly detections techniques in [9], the authors are concerned with categorizing the techniques of how to work with the data.

A hybrid solution combining clustering and heuristics is found in literature as in [10], where the authors make use of the Bee Algorithm to overtake the K-means (KM) local optima problem resulting in a better performance than the KM. In [11], the authors present a

new performance function to K-Harmonic Means (KHM) adopting a new distance measurement resulting in better clusters. In [12] the authors proposed the Simulated Annealing K-Harmonic Means Clustering (SAKHMC) and the results show that the CPU times for SAKHMC is greater than the KM, but the objetive function value for the SAKHMC is significantly greater then KM or KHM.

In [13] the author proposed the SFK-means, a joint of K-means, Fuzzy K-means and Swarm K-means to improve the local convergence and high false alarms. In [14], the authors proposed network intrusion detection with unsupervised outlier detection using random forest to classification and classified the data set in trees. In [15], the authors proposed an inverse manner to detect anomalies. From inside to outside, aiming to prevent unwanted activities from affecting other networks.

Our proposal is the Firefly Harmonic Clustering Algorithm (FHCA) an optimized clustering algorithm using the Digital Signature of Network Segment (DSNS) to detect volume anomalies and for a given threshold value the system triggers alarms to the network administrator. This threshold value is calculated with the ratio between the centers generated by the FHCA for the DSNS and real traffic samples.

III. Traffic Characterization: BLGBA and DSNS

The first step to detect anomalies is to adopt a model that characterizes the network traffic efficiently, which represents a significant challenge due to the non-stationary nature of network traffic. Large networks traffic behavior is composed by daily cycles, where traffic levels are usually higher in working hours and are also distinct for workdays and weekends. Thus, the GBA tool is used to generate different profiles of normal behavior for each day of the week, meeting this requirement. These behavior profiles are named Digital Signature of Network Segment (DSNS), proposed by Proença in [2] [3] [4].

The DSNS can be defined as a set of information that constitutes the traffic profile of a network segment or server. This information includes data such as traffic volume or number of errors, among others. The DSNS was generated by a model that performs a statistical analysis of the history of data collected in the SNMP objects, taking into account the exact moment of the collection. The period of this history can range from 4 to 12 weeks and the generation of DSNS is performed for each second of the day, every day of the week. As a result, we have a DSNS, which has an individual behavior profile for each day of the week. This approach is much closer from the ideal since it is possible to identify different behaviors on the network movement for each day of the week.

Figure 1 shows charts containing workdays of one week of monitoring of UEL network. Data were collected from SNMP object *udpInDatagrams*, at the University's Proxy server. The data collected are represented in green and the respective DSNS values by the blue line. Working hours (from 8 a.m. to 6 p.m.) present traffic levels higher and it is possible to observe a great adjustment between the DSNS and the real traffic.

IV. Anomaly Definition

In this section, we define anomaly in our context. To label if an interval found by the proposed algorithm can be classified as an anomaly or not, we adopt the following definition. Given $\mathbf{d}$, which represents the Digital Signature of Network Segment (DSNS) data, it can be described as a vector with N positions, the position index is related to the timestamp of collect value and $\mathbf{d}$(index) = the value collected.

$J = N/\Delta$, where Δ is the hysteresis interval adopted. For example, one day have 24 hours resulting in 86400 seconds. As explained in section III, we collect data from every second throughout the day, $N = 86400$, assuming intervals of 300 seconds each interval, J = 86400/300 = 288. The J value is be the total lenght of intervals, and $\mathbf{d_j}$ represent the values of interval J. We can define:

$$[\mathbf{d}_1, \ldots, \mathbf{d}_j, \ldots, \mathbf{d}_J] = \mathbf{d}, \tag{1}$$

The λ parameter is a representation of the variation occurred in DSNS. We can adopt a constant value based on prior knowledge of the network or using a statistical measure. In our work, λ is described by:

$$\lambda = \frac{\frac{1}{J}\sum_{j=1}^{J}\sigma(\mathbf{d}_j)}{\max[\sigma(\mathbf{d}_j)]}; \tag{2}$$

where σ is the standard deviation of $\mathbf{d}_j$ and max[] returns the highest σ value from all intervals.

The real traffic must go through the DSNS on a different scale and certain deviation is tolerable natural. In figure 2, the lines drawn represent the acceptable range created from DSNS. It is possible to observe that the traffic (red line) follows, in most of the time, the DSNS (blue line) and is inside the threshold range most of it. Depending on the network segment and the MIB objects collected data, the parameters may vary because of the

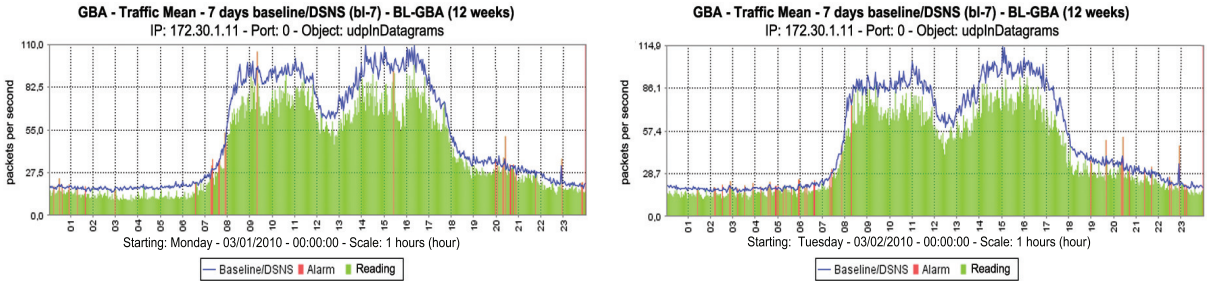

Figure 1. DSNS and real traffic collected from GBA from 03/01/2010 to 03/02/2010.

volume. For example, on a HTTP server the traffic is measure by the IP address of destination and origin, and it is different from a Firewall, where all traffic passes before entering and/or leaving a network segment. As the volume passing by through the Firewall is larger than a HTTP server, the λ is different.

To determine if a $\mathbf{d}_j$ is an anomaly or not, the equation 3 describes:

$$A(J) = \begin{cases} 0, \lambda_{DOWN} < \mathbf{d}_j < \lambda_{UP} \\ 1, c.c. \end{cases} \quad (3)$$

V. Firefly Harmonic Clustering Algorithm

One of the most classical clustering algorithm is the K-Means (KM), which partitioning n data in k centers, where each k center have high similarity with the n data grouped around. The problem in finding K center locations for N data can be defined as an optimization problem to minimize the equation 4 [12]:
$X = \{x_1, \ldots, x_n\}$: the data to be clustered;
$C = \{c_1, \ldots, c_k\}$: the set of cluster centers;

$$KM(X, C) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \|x_i^{(j)} - c_j\|^2 \quad (4)$$

Due to partitioning strategy in the initialization, the KM creates spots with local density data presenting a strong association between points and centers preventing the centers from moving out of a local density of data [16]. To address this problem, Zhang proposed the K-Harmonic means (KHM) in [6]. The main idea is calculate a weight function, $w(x_i)$, to recalculate the centers where each point present a certain weight to the final result and a membership function, $m(c_j|x_i)$, to determine the relationship strenght of x_i in relation to center c_j. Applying KHM, the optimization function becomes 5:

$$KHM(X, C) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}} \quad (5)$$

where p is an input parameter of KHM and assume $p \geq 2$.

We make use of a heuristic designed by Yang [7] to find the solution to our equation 5. The Firefly Algorithm (FA) is based on the behavior of the fireflies and the characteristics of light emitted. One important features is the emission intensity of light from a firefly is proportional to the objective function, i.e., $I(X) \propto KHM(X, C)$, but the intensity with which light is perceived by the firefly decreases with the distance between the fireflies.

The objective of the joint KHM and FA procedures is to minimize the problem of initialization observed in KM and to escape local optimal solutions resulting in a more efficient algorithm to cluster the network traffic samples, in order to efficiently detect volume anomalies from MIB objects. We named the proposed algorithm: Firefly Harmonic Clustering Algorithm (FHCA). The pseudo code is presented in Algorithm 1.

Algorithm 1 Firefly Harmonic Clustering Algorithm

Initialize the algorithm with randomly the initial centers;
while (i < IterationNum)||(error < ErrorAcepted(KHM(x,C)))
Calculate the objective function value according to equation (5);
For each data point x_i compute the membership value according to equation:

$$m(c_j|x_i) = \frac{\|x_i - c_j\|^{-p-2}}{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}$$

For each data point x_i, calculate the weight function according to equation:

$$w(x_i) = \frac{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}{(\sum_{j=1}^{k} \|x_i - c_j\|^{-p})^2}$$

For each center c_j, recompute its location based on the equations above:

$$c_j = \frac{\sum_{i=1}^{n} m(c_j|x_i) w(x_i) x_i}{\sum_{i=1}^{n} m(c_j|x_i) w(x_i)}$$

end while
Post process results and visualization

We run FHCA for all the intervals from the Digital Signature of Network Segment (DSNS) ($\mathbf{d_J}$) and the

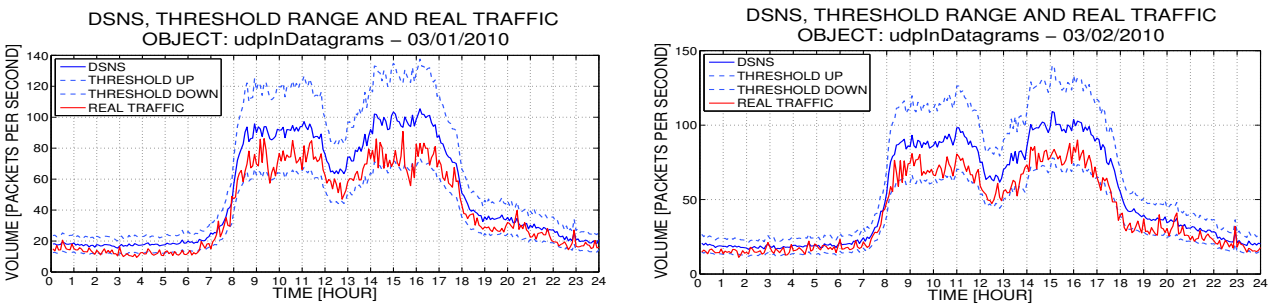

Figure 2. DSNS, the threshold range and real traffic of *udpInDatagrams* SNMP object, on Web-Server of State University of Londrina.

network traffic samples ($\mathbf{t_J}$). The algorithm results is the generation of K centers for each interval, $\mathbf{K_J^d}$ for centers from the DSNS and $\mathbf{K_J^t}$ for network traffic samples. Assuming the DSNS as the normal behavior for our network, we calculate $\mathbf{M_J} = \sigma(\mathbf{K_J^t})/\sigma(\mathbf{K_J^d})$, this ratio between the normal and the real traffic is the amount used for detecting volume anomalies.

VI. Results

To validate the proposed algorithm the results of the Firefly Harmonic Clustering Algorithm (FHCA) is presented using collect data from the Proxy server of the network environment from State University of Londrina (UEL). A week starting from 03/01/2009 (Monday) until 03/07/2010 (Sunday) and the MIB object *updIn-Datagrams* which represent the total number of input datagrams received from interfaces.

The first parameter tested was Δ interval length in a set (300, 600, 900 and 1200 seconds). The precision is the percentage of corrected data classified throughout all the data classified, and the value describing the best curve is $\Delta = 300$ as shown in figure 3.

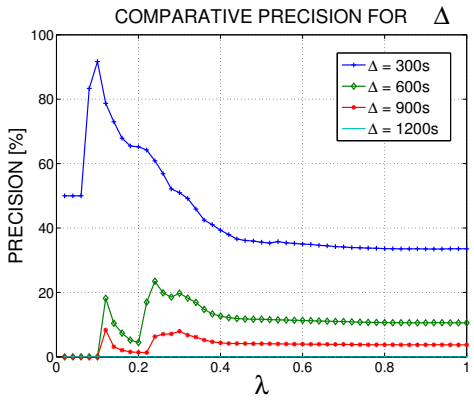

Figure 3. Comparison of precision rate for values of Δ.

Another important characteristic to define in a cluster algorithm is the number of centers. The values tested were $K = [2, 3, 4, 5]$. In figure 4 is presented the result. To quantify the results we adopted the True-positive rate (TPR) which describes the successes of FHCA algorithm classifying. As $K = 2$ assumes the best TPR, it is the value adopted in the next setup analysis and simulations results.

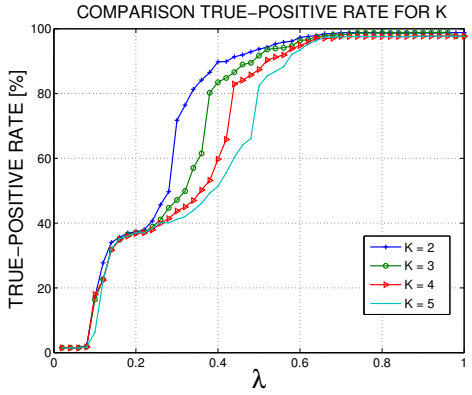

Figure 4. Comparison of true-positive rate for K.

In figure 5, it is demonstrated the Receiver Operating Characteristics (ROC) graph for the week under study considering both algorithms. The ROC graph is constructed with the TPR and False-positive rate (FPR) [18]. FPR describes how much the interval pointed by the FHCA was classified wrongly. The trade-off between TPR and FPR achieved by KM when FPR = 20% is less efficient than the rate achieved by the FHCA. This indicates a slight gain in calculating the centers for the FHCA.

The accuracy of the algorithm measure the degree of closeness of the algorithm measurements are from of its actual (true) value. In figure 6 it is shown a comparative of the KM and FHCA accuracy results. Increasing λ the FHCA shows a slight advantage over KM.

VII. Conclusions

We proposed a new algorithm named Firefly Harmonic Clustering Algorithm (FHCA) for volume anomaly de-

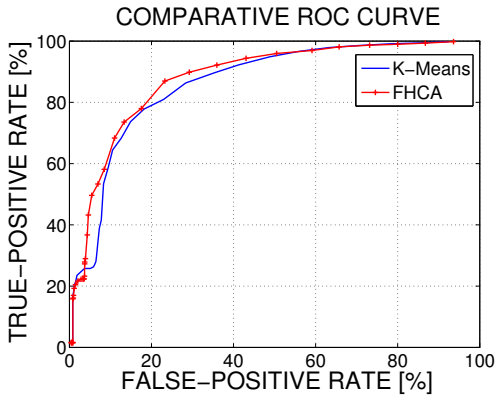

Figure 5. Comparative ROC Curve between K-Means and FHCA.

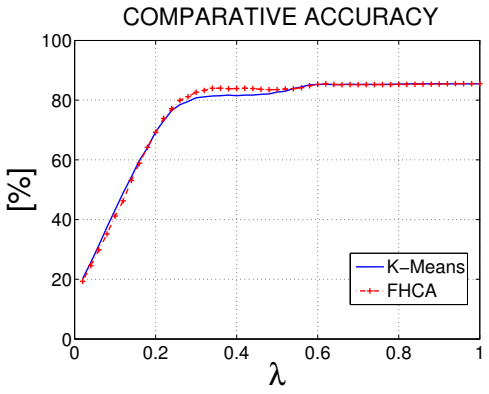

Figure 6. Accuracy for KM and FHCA.

tection using Digital Signature of Network Segment (DSNS) achiving satisfactory results in precision and accuracy with true-positive rates in 80% and false-positive rates in 20%. For future work, the goal is add more objects to analyze and combine strength with technique such as Principal Component Analysis (PCA) or Support Vector Machine (SVM).

ACKNOWLEDGEMENTS

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) through a post-graduate master's degree level and by SETI/Fundação Araucária and MCT/CNPq by the financial support for the Rigel Project and partially supported by the 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

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey." *ACM Computing Surveys.*, vol. 41, no. 3, 2009.

[2] M. L. Proença Jr., C. Coppelmans, M. Botolli, and L. de Souza Mendes, *Security and reliability in information systems and networks: Baseline to help with network management.* Springer, 2006, pp. 149–157.

[3] M. Lima, L. Sampaio, B. Zarpelã ando, J. Rodrigues, T. Abrã ando, and M. Proenç anda, "Networking anomaly detection using dsns and particle swarm optimization with reclustering," in *GLOBECOM 2010, 2010 IEEE Global Telecommunications Conference*, December 2010, pp. 1 –6.

[4] B. B. Zarpelão, L. de Souza Mendes, and M. L. P. Jr., "Anomaly detection aiming pro-active management of computer network based on digital signature of network segment." *Journal of Network and Systems Management (JNSM)*, vol. 15, no. 2, pp. 267–283, 2007.

[5] S. Shanbhag and T. Wolf, "Anombench: A benchmark for volume-based internet anomaly detection," in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, December 2009, pp. 1–6.

[6] B. Zhang, M. Hsu, and U. Dayal, "K-harmonic means - a data clustering algorithm," Hewlett-Packard Laboratories, Tech. Rep. HPL-1999-124, 1999.

[7] X.-S. Yang, *Nature-Inspired Metaheuristic Algorithms*. Luniver Press, 2008.

[8] W. Zhang, Q. Yang, and Y. Geng, "A survey of anomaly detection methods in networks," in *Computer Network and Multimedia Technology, 2009. CNMT 2009. International Symposium on*, January 2009, pp. 1–3.

[9] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 51, pp. 3448–3470, August 2007.

[10] D. T. Pham, S. Otri, A. A. Afify, M. Mahmuddin, and H. Al-Jabbouli, "Data clustering using the bees algorithm," in *Proc 40th CIRP Int. Manufacturing Systems Seminar, Liverpool*, 2007.

[11] R. Jingbiao and Y. Shaohong, "Research and improvement of clustering algorithm in data mining," in *Signal Processing Systems (ICSPS), 2010 2nd International Conference on*, vol. 1, July 2010, pp. V1–842 –V1–845.

[12] Z. Güngör and A. Ünler, "K-harmonic means data clustering with simulated annealing heuristic." *Applied Mathematics and Computation*, vol. 184, no. 2, pp. 199–209, 2007.

[13] R. Ensafi, S. Dehghanzadeh, R. Mohammad, and T. Akbarzadeh, "Optimizing fuzzy k-means for network anomaly detection using pso," in *AICCSA 2008. IEEE/ACS International Conference on Computer Systems and Applications*, Apr. 2008, pp. 686 – 693.

[14] J. Zhang and M. Zulkernine, "Anomaly based network intrusion detection with unsupervised outlier detection," in *Communications, 2006. ICC '06. IEEE International Conference on*, vol. 5, June 2006, pp. 2388 –2393.

[15] K. Limthong, P. Watanapongse, and F. Kensuke, "A wavelet-based anomaly detection for outbound network traffic," in *Information and Telecommunication Technologies (APSITT), 2010 8th Asia-Pacific Symposium on*, June 2010, pp. 1–6.

[16] F. Yang, T. Sun, and C. Zhang, "An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization." *Expert Syst. Appl.*, vol. 36, no. 6, pp. 9847–9852, 2009.

[17] S. Luke, *Essentials of Metaheuristics*. Lulu, 2009.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2005.