Research Journal of Applied Sciences 11 (10): 1076-1082, 2016
ISSN: 1815-932X

Anomaly Detection in Network Traffic Using Stream Data Mining: Review

Wesam S. Bhaya and Suad A. Alasadi
College of Information Technology, University of Babylon, Babylon, Iraq

Abstract: Malicious activities on the Internet are commonly shown in Internet traffics. Anomalies like DDos, Worm, flooding attack, etc are defined as any deviation from the normal and something which are outside the usual range of variations. These anomalies consume network resources and lead to security issues such as confidentiality, integrity and availability. Anomaly detection is one of the data mining tasks which are the analysis of large volumes of data to determine items, events or observations which do not belong to unexpected patterns. Data mining techniques can be used in anomaly detection such as k-means clustering, artificial neural networks. Network traffic is an example of data stream which characterize as continues, massive and rapid sequence of data. Thus Mining such application need techniques which are different from traditional data mining techniques. These techniques must be able to process data which is continues, high speed and you can look at only once. This paper shows overview of anomaly detection framework, the growing field of data stream and presents techniques of stream data mining which are used for anomaly detection in network traffic.

Key words: Network security, anomaly detection, stream data mining

INTRODUCTION

Recently computer networks are becoming articular points in our daily life, the internet also increase importance in our society. These improvement have led to increase the network consuming where traffic are changed and anomalous activity traffic are generally showed but difficult to be diagnosed. In order to protect these networks from attacks, it is important to monitor and analyze network flows. Anomaly detection aims to discover anomalies which effect the infrastructure of the network. Anomalies may be intentional such as attacks or not malicious such as failures, misconfigurations. Thus, detecting such anomalies accurately has become an important problem for the network community to solve (Gu *et al*., 2005; Marnerides *et al*., 2014; Morkhade and Bartere, 2013).

Anomaly can be defined as any deviation from the normal and something which is outside the usual range of variations (Bereziñski *et al*., 2014). Anomaly detection methods are broadly classified into two categories: signature-based, anomaly based detections (Sarinnapakorn *et al*., 2014).

Signature-based anomaly detection works by monitoring network packets and comparing it with a database of signatures or patterns from known malicious attacks. Usually signatures is a combination of packet header and packet content to determine the anomalous traffic flows. Anomaly based detection in the other hand, works by monitoring packets on the network and comparing it with a baseline profile of normal packets. This profile recognize what is normal for the network such as: the normal bandwidth usage, the common protocols used and detect the anomaly as any packet which is different from the normal profile (Kumar, 2007).

Data mining can be used for anomaly detection. Since, it works to extract features from network traffic; it can be used to distinguish between common legitimate and attack traffics.

Computer network, sensor network are examples of applications that generate data stream which is massive (terabytes in size), continues and rapid sequence of data elements. Traditional data mining need multi scans of data to be processed, this is not realistic for data stream. Thus traditional mining techniques cannot be applied for data stream (Gupta and Rajput, 2013).

Stream data mining differ from traditional mining techniques as it can process data which is large, continues, infinite and only one scan of data. This study previews stream data mining techniques which are used for anomaly detection (Gaber *et al*., 2005).

MATERIALS AND METHODS

General frameworks for anomaly detection: One of the important data analysis task is anomaly detection. It is used to detect abnormal data from a given dataset. There are different approaches of anomaly detection; Fig. 1 shows the general framework for anomaly detection (Agrawal and Agrawal, 2015).

Corresponding Author: Wesam S. Bhaya, College of Information Technology, University of Babylon, Babylon, Iraq

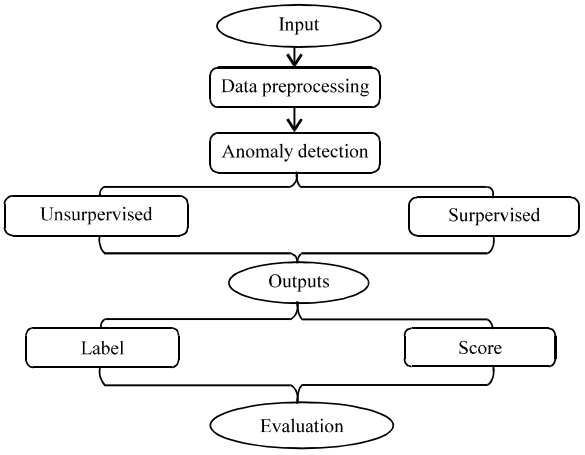

Fig. 1: Basic network anomaly detection model

The first step for anomaly detection is preprocessing which is essential to remove irrelevant and noise data, fill the missing values and integrate input data types. For example IP address is hierarchical while the protocol is categorical. The preprocessing stage depend on the applied anomaly detection model. Then, the anomaly detection stage are performed on the data. The model for detection may be (supervised or unsupervised). Final the evaluation stage are used to evaluate the output of model which can be scores or labels using benchmark dataset.

Review of stream mining: The recent advances and growing the number of applications that enable the capture of data in different measurements with different fields need intelligent data processing and online analysis (Gaber *et al*., 2005). Real time surveillance systems, computer network traffics, sensor networks are an example of the software that generate massive stream of data. These systems generate massive, continues and rapid sequence of data stream (Kholghi and Keyvanpour, 2011).

Data mining is the process of discovering models for data (Rajaraman and Ullman, 2011). Its techniques are appropriate for simple and structure data such as relational database. The advance developments in database and data collection make data with complex forms like semi-structured and non-structured, hypertext and multimedia. This led to the appearance of stream data mining (Gupta and Rajput, 2013).

Data stream characterized as any data with continues change sequence that need to store or process. Sensor network, computer network are examples of systems that generate seam data which are massive, temporal order, fast change and infinite.

These features had been led to appear challenges in data streams such as storage, querying and mining. Traditional mining techniques require multiple scans of data. So they are not suitable for stream data applications. These techniques can be modified in order to fit data streams (Hao *et al*., 2009).

Data stream mining is the process of extract information in form of models and patterns from continues stream data. It is not technique but it is a function that is combined with other data mining techniques to produce better and quicker results. Data stream mining has specific uses. The first use is the handling of data stream which is change continuously like credit fraud detection, web mining, intrusion detection. The second use is the processing of large data that cannot be stored in memory all at once (Esmaeili and Almadan, 2011). The basic framework of stream mining is explained in Fig. 2.

Preprocessing techniques for data stream mining: For extract knowledge in form of patterns and models from data stream, it is important to modify methods that can able to analyze and process stream in multiple dimension, multiple level, one pass and in online form. It is also necessary for methods to process data with efficient space and time complexity (Kholghi and Keyvanpour, 2011).

The solution of data stream mining are categorize into data based and task based. These types are shown in Fig. 3. Data based techniques work by reducing the data volume by summarizing the total dataset or selecting subset of dataset to be processed. While task based techniques work by modifying the existing methods or adapting new one to reduce the complexity of stream processing.

Data based techniques: This type works by summarizing the entire data or selecting some parts of dataset to be analyzed. The following subsections are examples of this technique.

Sampling: It is the process of determining subset of input stream to be analyzed using the probability for data item. It is one of the fundamental statistical techniques that has been used for long time. The main problems of sampling in data stream analysis is the unknown size of dataset. Another problem is that it may be important in some applications to check for anomalies such as in traffic analysis. Thus this technique may not be the right choice for anomaly detection (Gupta and Rajput, 2013; Brauckhoff *et al*., 2006).

Load shedding: It defined as the process of dropping a sequence of items. The advantage of using it is that it is used in querying stream successfully. But in the other hand, it has the same problems of sampling. It cannot be

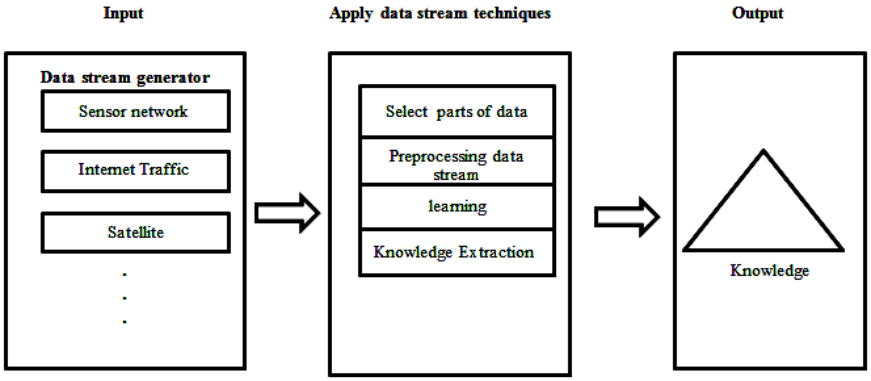

Fig. 2: General framework of stream data mining

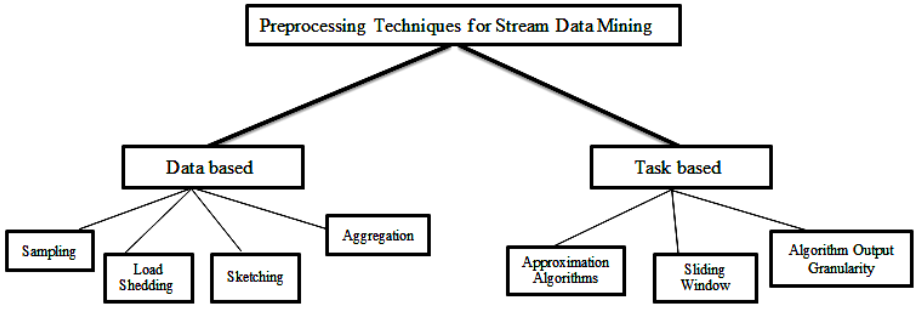

Fig. 3: Stream data preprocessing methods

used with algorithms of data mining since it drops chunks of data which might be important patterns in time series analysis (Gaber *et al*., 2005).

Sketching: It is the process of projecting subset of features randomly. It works by vertically sample the observed stream. It has been applied in aggregate queries to compare variable data streams. The main drawback of this technique is the accuracy. Thus it is difficult to be used in data stream analysis. PCA is a good solution in streaming applications (Kholghi and Keyvanpour, 2011).

Synopsis Data Structures (SDA): SDA summarizes the input stream using the available summarization methods for analysis purposes. Wavelet, histograms and moments are such examples. The problem of synopsis data is that it is not represent all the features of the dataset. Thus when we use such data structures, the produced answers will be approximate (Aggarwal, 2007).

Aggregation: It is the process of applying statistical metrics like means, median and variance on data which are necessary to summarize the input streams. The resulted aggregated data are used in data mining algorithms. The main drawback of this technique is that its performance is attenuated with highly fluctuating data distributions (Aggarwal, 2007).

Task based techniques: This type includes methods that work to reduce the computational complexity of stream data processing. It includes techniques that work by modifying existing techniques or invent new ones. The following subsections are examples of this technique.

Approximation algorithms: The main using of this technique in the designing of algorithms. The solutions that can obtain by these algorithms are approximate and have error rates. Approximation algorithms considered the direct solution for data streams.

Sliding window: Sliding window is an advance technique which is recently used to give detailed analysis for the most recent data items. It is used to produce approximate answer for querying data stream. Each new incoming data is analyzed using the summarization of the old one, then the old items are removed and replaced with new items. There are two kinds of sliding window: (count based, time based) window. In first type, the latest N items are stored.

While in second type, the items can stored if it generated or received in the last time units (Gupta and Rajput, 2013; Aggarwal, 2007).

Algorithm output granularity: This technique is considered as the first recommended approach for analyzing data which is used with high frequency data rates. It needs time and space requirements to do well. AOG characterized by the ability of mining data stream, adaptation of resources and streams and merging the generated structures which are running out of memory. AOG can used in mining techniques such as association, clustering and classification (Kholghi and Keyvanpour, 2011).

RESULTS AND DISCUSSION

Stream mining based anomaly detection techniques: Anomalies are patterns in a dataset whose behaviors are different from normal behavior or whose belong to unexpected behaviors. It is termed as outliers. The naming of anomaly may be belonging to malicious activity or intrusions. The abnormal behaviors which are found in dataset form an articular point for detecting anomalies and important topic for analysis.

There are different methods for anomaly detection based on stream mining. This paper reviews different types of techniques for anomaly detection based on stream data mining. These techniques will be explained in the subsection below.

Clustering based anomaly detection: In data mining, clustering is one of the fundamentals tools for data analysis. It is used to detect distributions and patterns from the dataset. There is need for using clustering in data stream, because it is efficient in obtaining summaries from data streams, it also considered a power tool for outlier analysis.

Clustering is defined as the process of grouping a set of points into clusters according to a distance measure. The result of clustering is set of clusters,each cluster will have set of points with small distance from one another and with large distance from other clusters. In the other hand, Outlier detection is defined as the process of finding points which are dissimilar to the general structure of data. Both these problems have been studied in the data mining, database and statistics community over the years (Rajaraman and Ullman, 2011; Ahmed *et al*., 2016).

Munz, introduced method for anomaly detection using K mean clustering algorithm. The training data which was unlabeled are split into normal and abnormal clusters. The resulting centroids are used as patterns for anomaly detection in new observed traffic. The complexity of the this algorithm is O(K n t) where K is the number of clusters, n is the number of objects to be classified and t is the number of iterations. Finally, this paper is evaluated using DARPA'98 dataset and real traffic.

Syarif used Expectation Maximization (EM) clustering algorithm for anomaly detection. EM clustering is an extension to k mean clustering. It depends on the mean of cluster to assign a point to a cluster which is similar to. The researcher compared the performance of detection using EM algorithms with Kmean, improved k mean and K median. The experimental results showed that EM clustering performs high accuracy compared to them; it was 78.06% for k median it was 76.71%, for improved k Mean it was 65.40% and for k mean it was 57.81%.

Chen and Li (2011) proposed an clustering algorithm for anomaly detection that employed enhanced DBSCAN. it is density based clustering algorithm which is used for processing massive data. It can discover clusters with arbitrary shapes. Clusters forms as dense regions from set of points and separated by regions of low density. Enhanced DB scan algorithm is used to enhance the performance of DBSCAN such as using R* tree to determine the neighbors of an object. The results indicated that the proposed algorithm for anomaly detection has high detection rates with low false positive rates with DARPA dataset.

Papalexakis *et al*. (2012), used co-clustering for anomaly detection. It is the process of getting set of rows and columns which are similar from dataset matrix. The researcher used two algorithms of co-clustering algorithm (soft and hard co-clustering). The first algorithm also called Sparse Matrix Regression, this algorithm does not find clusters for whole dataset but it is an efficient in finding parameters which are good to indicate for ambiguity in dataset and producing pure clusters. The second type also called Information Theoretic, this algorithm works on the entire dataset and creates clusters for normal and abnormal connections for full dataset. The results with the KDD 1999 Cup, showed that co-clustering is a powerful, unsupervised method for separating normal from abnormal connections.

Miller *et al*. (2011), presented clustering algorithm based stream data mining for anomaly detection which is a modified version of the DenStream algorithm. This algorithm considered each new packet as a new point to be clustered. If a packet is assigned to cluster, it is classified as normal. Else, it is classified as anomaly.

Tan proposed an anomaly detection algorithm which is Streaming HSTrees that satisfied the key requirements for mining data streams. The results showed that HS-Trees with the regular model update is robust in evolving data stream. The detection rate of HS-Trees is identical to the oracle-informed Hoeffding Tree (an

optimistic baseline). The runtime of HS-Trees outperforms Hoeffding Tree. The performance of HS-Trees is robust against different parameter settings.

Classification based anomaly detection: George (2012), suggested an approach for anomaly detection using machine learning algorithms. The researcher used PCA as dimensionality reduction and SVM as classification algorithms. The proposed approach was evaluated using KDD dataset. The results showed the decrease in execution time when using PCA with SVM. Precision and recall values also showed that SVM with PCA is more accurate in detecting network anomalies.

Nagar *et al.* (2014), implemented new technique for anomaly detection using SVM and Decision tree. The researcher first clustered data traffic using SVM and then classified the traffic using decision tree. The proposed approach provided high accuracy, less time complexity and error rate.

Pradhan *et al.* (2012), introduced an approach for anomaly detection using neural network. The researcher aimed to enter user behavior as a parameter in ID using backpropagation neural network. neural network is a successful method for training and learning ID System. Backpropagation neural network is able to classify normal traffic and detect anomalies without need huge number of training data. The proposed method was evaluated using DARPA dataset and showed that neural network can classify traffic correctly. The researcher got 88% classification rate for known and unknown attacks.

Aneetha and Bose (2012), suggested new approach for anomaly detection using neural network and clustering algorithms which are modified SOM and k mean clustering. The modified SOM is used to create network and K mean clustering is used to group nodes in the network using similarity measures. This approach is compared with other neural network methods and the result showed high detection rate and low false alarm rate.

Murugan and Rajan (2014), used Bayesian network for IDS. Bayesian network enables from finding the probabilistic between variables. Thus the researcher used it to proposed an anomaly based ID which capable of knowing the probabilistic relationships between attacks and categorizing attack types.

Shamshirband *et al.* (2014), proposed an IDS using Fuzzy Qlearning (FQL) to protect the network from DDOS attack. The researcher used CAIDA and KDD to evaluate the algorithm. The results showed that FQL has more accurate in detection in compare with using Fuzzy Logic Controller and Q-learning algorithm individually.

Hussein (2014), used one of the efficient data mining algorithms called Very Fast Decision Tree (VFDT) for anomaly detection. The researcher used VFDT for anomaly detection in order to classify the connections as normal or abnormal. VFDT is a high-performance data mining system which is based on Hoeffding trees and it deals with data streams. The proposed system worked on two modes, on-line and offline modes. The researcher used KDDCUP99 dataset. The experimental results indicated that the proposed system which used VFDT achieved a high classification accuracy rate of 93%.

Time series based anomaly detection: Wu and Shao (2005) presented a method for detecting the anomalies using time series analysis. This method analyzed the abrupt change of time series data using Auto Regression (AR) model. AR modeled the abrupt change of time series data and performed sequential hypothesis test to detect anomalies. Using time correlation and location correlation, the method determined anomalous activity and its time and location corectly.

Yasami *et al.* (2008), proposed a novel approach for online network based anomaly detection using Stochastic Learning Automata (SLA). The researcher studied time based analysis for packet anomaly detection and used a learning algorithm for modeling a normal time series data of broadcast traffic behavior. The researcher used an indicator of abnormality as any deviation of observed time series data of broadcast traffic from the learned automaton. Experimental results showed that it has high performance and low percentage of (FP).

Wuzuo and Weidong (2010) proposed an online detection of network traffic using degree distributions. The approach used degree distributions to profile the features of traffic and then computed the entropy to alarm the changes of degree distribution which is useful for knowing the normal and abnormal traffic by using adaptive threshold. The result showed that this scheme is feasible and efficient for online anomaly detection.

Nayyar proposed approach for detecting anomalies in real-time. The main idea of this approach is that it classifies patterns in sliding window and identifies the class of anomaly to a pattern belong to. When it classify a data in the sliding window as anomaly, It will dispatched to algorithm to predict which data points in sliding window are anomalies. Experimental results showed good accuracy rate for detecting the anomalies.

CONCLUSION

In this study, different types of anomaly detection techniques using stream data mining are explained. These techniques are categorized into three major categories (clustering , classification, time based anomaly detection). Each category can differentiate between normal and abnormal behavior of packets on the basis of comparison

between them. When performance of our data or service is deviate from the normal distribution, it is considered as anomaly. Once anomaly is detected in the data it can be removed using any suitable detection technique. Anomaly detection is an interesting area of computer and network security. It is considered as one of the fundamental problems of data mining. In this paper, we have summarized anomaly detection techniques along with various research direction and application domains.

ACKNOWLEGMENTS

I want to acknowledge IT college to its support to accomplish this study.

REFERENCES

Aggarwal, C.C., 2007. Data Streams: Models and Algorithms. Vol. 31, Springer, Berlin, Germany, ISBN- 13:978-0-387-47534-9, Pages: 354.

Agrawal, S. and J. Agrawal, 2015. Survey on anomaly detection using data mining techniques. Procedia Comput. Sci., 60: 708-713.

Ahmed, M., A.N. Mahmood and M.R. Islam, 2016. A survey of anomaly detection techniques in financial domain. Future Gener. Comput. Syst., 55: 278-288.

Aneetha, A. and S. Bose, 2012. The combined anomaly detection using neural networks and clustering techniques. Comput. Sci. Eng. Int. J. (CSEIJ.), 2: 37-46.

Berezinski, P., J. Pawelec, M. Ma³owidzki and R. Piotrowski, 2014. Entropy-based internet traffic anomaly detection: A case study. Proceedings of the Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, June 30-July 4, 2014, Springer, Brunow, Poland, ISBN: 978-3-319-07012-4, pp: 47-58.

Brauckhoff, D., B. Tellenbach, A. Wagner, M. May and A. Lakhina, 2006. Impact of packet sampling on anomaly detection metrics. Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, October 25-27, 2006, ACM, Rio de Janeiro, Brazil, ISBN:1-59593-561-4, pp: 159-164.

Chen, Z. and Y.F. Li, 2011. Anomaly detection based on enhanced DBS can algorithm. Procedia Eng., 15: 178-182.

Esmaeili, M. and A. Almadan, 2011. Stream data mining and anomaly detection. Int. J. Comput. Appl., 34: 38-41.

Gaber, M.M., A. Zaslavsky and S. Krishnaswamy, 2005. Mining data streams: A review. ACM. Sigmod Rec., 34: 18-26.

George, A., 2012. Anomaly detection based on machine learning: Dimensionality reduction using PCA and classification using SVM. Int. J. Approach Comput. Appl., 47: 5-8.

Gu, Y., A. McCallum and D. Towsley, 2005. Detecting anomalies in network traffic using maximum entropy estimation. Proceedings of the 5th Conference on Internet Measurement, October 19-21, 2005, Berkeley, CA., USA., pp: 32-32.

Gupta, N. and I. Rajput, 2013. Stream data mining: A survey. Intl. J. Eng. Res. Appl., 3: 1113-1118.

Hao, M.C., U. Dayal, D.A. Keim, R.K. Sharma and A. Mehta, 2009. Visual analytics of anomaly detection in large data streams. Proceedings of the Conference on Visualization and Data Analysis 2009, January 18, 2009, Hewlett-Packard Labs, Palo Alto California, pp: 72430B-72430B.

Hussein, N.A., 2014. Design of a network-based anomaly detection system using VFDT algorithm. Master Thesis, Eastern Mediterranean University, Famagusta, Northern Cyprus. http://i-rep.emu.edu.tr:8080/xmlui/handle/11129/1671

Kholghi, M. and M. Keyvanpour, 2011. An analytical framework for data stream mining techniques based on challenges and requirements. Int. J. Eng. Sci. Technol. (IJEST.), 3: 2507-2513.

Kumar, S., 2007. Survey of current network intrusion detection techniques. http://www.cs.wustl.edu/~jain/cse571-07/ftp/ids.pdf.

Marnerides, A.K., F.A. Schaeffer and A. Mauthe, 2014. Traffic anomaly diagnosis in internet backbone networks: A survey. Comput. Networks, 73: 224-243.

Miller, Z., W. Deitrick and W. Hu, 2011. Anomalous network packet detection using data stream mining. J. Inf. Secur., 2: 158-168.

Morkhade, M.S.S. and M. Bartere, 2013. Survey on data mining based intrusion detection systems. Int. J. Appl. Innovation Eng. Manage. (IJAIEM.), 2: 338-343.

Murugan, S. and S. Rajan, 2014. Detecting anomaly IDS in network using bayesian network. IOSR. J. Comput. Eng., 16: 1-7.

Nagar, M., P. Shraddha and J. Maurya, 2014. Detection and classification of network anomalies using SVM and decision tree. Int. J. Comput. Sci. Inf. Technol., 5: 2338-2341.

Papalexakis, E., A. Beutel and P. Steenkiste, 2012. Network anomaly detection using Co-clustering. Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, August 26, 2012, ACM, Washington, USA., ISBN: 978-0-7695-4799-2, pp: 403-410.

Pradhan, M., S.K. Pradhan and S.K. Sahu, 2012. Anomaly detection using artificial neural network. Int. J. Eng. Sci. Emerging Technol., 2: 29-36.

Rajaraman, A. and J.D. Ullman, 2011. Mining of Massive Datasets. Cambridge University Press, UK., ISBN-13: 978-1107015357, Pages: 326.

Shamshirband, S., N.B. Anuar, M. Laiha, M. Kiah and S. Misra, 2014. Anomaly detection using fuzzy q-learning algorithm. Acta Polytech. Hungarica, 11: 5-28.

Wu, Q. and Z. Shao, 2005. Network anomaly detection using time series analysis. Proceedings of the Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services-(ICAS-ISNS'05), October 23-28, 2005, IEEE, Shanghai, China, ISBN:0-7695-2450-8, pp: 42-42.

Wuzuo, W.A.N.G. and W.U. Weidong, 2010. Online detection of network traffic anomalies using degree distributions. Intl. J. Commun. Network Syst. Sci., 3: 1-6.

Yasami, Y., S.P. Mozaffari and S. Khorsandi, 2008. Stochastic learning automata-based time series analysis for network anomaly detection. Proceedings of the International Conference on Telecommunications (ICT08), June 16-19, 2008, IEEE, Tehran, Iran, ISBN:978-1-4244-2035-3, pp: 1-6.