ISSN:0975-9646

Assessing Deviations of Empirical Measures for Temporal Network Anomaly Detection: An Exercise

1. Amjan.Shaik, *CSE ,Ellenki College Of Engineering and Technology(ECET), Hyderabad.*
2.S.V.Achuta Rao , *CSE and IT , DJR Institute of Engineering and Technology (DJRIET), Vijayawada, India.*
3. Hymavathi. Bhadriraju , -*CSE, Bharath University (BU), Selaiyur, Chennai, India.*
4. Md.Reyaz,–*IT, Muffakham Jah College of Engineering and Technology (MJCET), Hyderabad, India.*
5.Nazeer. Shaik, *CSE, Moghal College Of Engineering and Technology(MCET), Bandlaguda, Hyderabad, India.*

Abstract
The Internet and computer networks are exposed to an increasing number of security threats. With new types of attacks appearing continually, developing flexible and adaptive security oriented approaches is a severe challenge. In this context, anomaly-based network detection techniques are a valuable technology to protect target systems and networks against malicious activities. However, despite the variety of such methods described in the literature in recent years, security tools incorporating anomaly detection functionalities are just starting to appear, and several important problems remain to be solved. This paper begins with an exercise of the most well-known anomaly-based detection techniques. Then, available platforms, systems under development and research projects in the area are presented. Finally, we outline the main challenges to be dealt with for the wide scale deployment of anomaly-based detectors, with special emphasis on assessment issues. Network anomaly detection is a vibrant research area. Researchers have approached this problem using various techniques such as artificial intelligence, machine learning, and state machine modeling. In this paper we introduce an internet traffic anomaly detection mechanism based on large deviations results for empirical measures. Using past traffic traces we characterize network traffic during various time-of-day intervals, assuming that it is anomaly-free. We present two different approaches to characterize traffic: (i)A model- free approach based on the method of types and Sanov's theorem, and (ii) A model-based approach modeling traffic using a Markov modulated process. Using these characterizations as a reference we continuously monitor traffic and employ large deviations and decision theory results to compare the empirical measure of the monitored traffic with the corresponding reference characterization, thus , identifying traffic anomalies in real-time. Our experimental results shows that applying our methodology (even short- lived) anomalies are identified with in a small number of observations. Through out, we compare the two approaches presenting their advantages and disadvantages to identify and classify temporal network anomalies. We also demonstrate how our framework can be used to monitor traffic from multiple network elements in order to identify both spatial and temporal anomalies. We validate our techniques by analyzing real traffic traces with time-stamped anomalies.

Keywords: Network traffic, Anomalies, Empirical measures, Network security, Anomaly detection.

I.INTRODUCTION

A network anomaly is a sudden and short-lived deviation from the normal operation of the network. Some anomalies are deliberately caused by intruders with malicious intent such as a denial-of-service attack in an IP network, while others may be purely an accident such as an overpass falling in a busy road network. Quick detection is needed to initiate a timely response, such as deploying an ambulance after a road accident, or raising an alarm if a surveillance network detects an intruder. Network monitoring devices collect data at high rates. Designing an effective anomaly detection system consequently involves extracting relevant information from a voluminous amount of noisy, high-dimensional data. Network intrusion detection systems are automated systems that detect intrusions in computer network systems. An anomaly behavior detecting-based intrusion detection system builds normal traffic model and uses this model to detect abnormal traffic patterns and intrusion attempts. The goal of this anomaly detection system is to determine whether an unknown network data item belongs to normal or to an intrusive pattern [1][7]. Current network intrusion detection methods provide low detection rates because of the multi-dimensional data problem. For example, a simple variant of single linkage clustering was applied to learn network traffic patterns on unlabelled noisy data [8].The KDD CUP 1999 dataset [9] was used and this approach achieved from 40% to 55% detection rate and from 1.3% to 2.3% false positive rate. Networks are complex interacting systems and are comprised of several individual entities such as routers and switches. The behavior of the individual entities contribute to the ensemble behavior of the network. The evolving nature of internet protocol (IP) networks makes it difficult to fully understand the dynamics of the system. Internet traffic was first shown to be composed of complex self-similar patterns by Leland *et al.* [1]. Multifractal scaling was discovered and reported by Levy-Vehel *et al.* [2]. To obtain a basic understanding of the performance and behavior of these complex networks, vast amounts of information need to be collected and processed. Often, network performance information is not directly available, and the information obtained must be synthesized to obtain an understanding of the ensemble behavior. There are two main approaches to studying or characterizing the ensemble behavior of the network: The first is the inference of the overall network behavior through the use of network probes [3] and the second by understanding the behavior of the individual entities or nodes. In the first approach, which is often referred to as network tomography [4], there is no assumption made about the network, and through the use of probe measurements, one can infer the characteristics of the network. This is a useful approach when characterizing non

cooperative networks or networks that are not under direct administrative control. In the case of a single administrative domain where knowledge of the basic network characteristics such as topology are available, an entity-based study would provide more useful information to the network administrator. Using some basic knowledge of the network layout as well as the traffic characteristics at the individual nodes, it is possible to detect network anomalies and performance bottlenecks. The detection of these events can then be used to trigger alarms to the network management system, which, in turn, trigger recovery mechanisms. The methods presented in this paper deal with entity-based measurements. The approaches used to address the anomaly detection problem are dependent on the nature of the data that is available for analysis. Network data can be obtained at multiple levels of granularity such as end-user-based or network-based. End-user-based information refers to the transmission control protocol (TCP) and user datagram protocol (UDP) related data that contains information that is specific to the end application. Network-based data pertains to the functioning of the network devices themselves and includes information gathered from the router's physical interfaces as well as from the router's forwarding engine. Traffic counts obtained from both types of data can be used to generate a time series to which statistical signal processing techniques can be applied [5], [6]. However, in some cases, only descriptive information such as the number of open TCP connections, source-destination address pairs, and port numbers are available. In such situations, conventional approaches of rule-based methods would be more useful [7]. Significant progress has been made in network monitoring instrumentation, automated on-line traffic anomaly detection is still a missing component of modern network security and traffic engineering mechanisms. Network anomaly detection approaches can be broadly grouped into two classes: signature-based anomaly detection where known patterns of past anomalies are used to identify ongoing anomalies for intrusion detection, and anomaly detection which identifies patterns that substantially deviate from normal patterns of operation. Earlier work has showed that systems based on pattern matching had detection rates below 80%. Furthermore, such systems need constant (and expensive) updating to keep up with new attack signatures. As a result, more attention has to be drawn to methods for traffic anomaly detection since they can identify even novel ,unseen types of anomalies. In contrast with other approaches , we are not trying to characterize the abnormal operation, mainly because it is too complex to identify all the possible anomalous instances, especially those that have never been observed. Instead we observe past system behavior and, assuming that it is anomaly-free, we obtain a statistical characterization of normal behavior. Then, using this knowledge we continuously monitor the system to identify time instances where system behavior does not appear to be normal. The novelty of our approach is, we characterize normal behavior and in how we assess deviations from it.

1.1 Scope

Automated on-line traffic anomaly detection is still a missing component of modern network security and traffic engineering mechanisms. We introduce an internet traffic anomaly detection mechanism based on large deviations results for empirical measures. Using past traffic traces we characterize network traffic during various time-of-day intervals, assuming that it is anomaly-free.

1.2 Purpose

We focus on anomaly detection and in particular on statistical anomaly detection, where statistical methods are used to assess deviations from normal operation. Our main contribution is the introduction of a new statistical traffic anomaly detection framework that relies on identifying deviations of the empirical measure of some underlying stochastic process characterizing system behavior. We note that the words "traffic" and "router" are purposefully absent from the previous paragraph. Rather, we use the generic term "system". This is to indicate that our approach can be easily adapted to identify anomalies in any trace of system activity we would like to monitor ,e.g., access to various application ports, IP source-destination addresses, system calls, etc.

1.3 Features

We demonstrate two methods to characterize normal behavior:

(i) A model-free approach employing the method of types to characterize the type (i.e., empirical measure) of an Independent and Identically-Distributed (IID) sequence of appropriately averaged system activity,

(ii) A model-based approach where system activity is modeled using a Markov Modulated Process (MMP).

Given these characterizations, we employ the theory of Large Deviations (LD) and decision theory results to assess whether current system behavior deviates from normal. LD theory provides a powerful way of handling rare events and their associated probabilities with an asymptotically exact exponential approximation. The key technical results we rely upon are Sanov's theorem in the model-free approach, a related result for the empirical measure of a Markov process for the model-based case, and Hoeffding's [8] composite hypothesis testing rule for assessing deviations from normal activity.

The model-free approach aggregates traffic over short time intervals to which we will refer to as time buckets. Although the correlation between samples in short time scales is significant, it reduces rapidly between aggregates over a time bucket. Hence, we consider the sequence of traffic aggregates over a time bucket as an IID sequence and employ the method of types to characterize its distribution. Our model-based approach uses an MMP process to model legitimate traffic during some time-of-day interval. Earlier work has shown that MMP models can accurately characterize network traffic, at least for the purposes of estimating important quality-of-service metrics.

II. RESEARCH BACKGROUND

Related Work and Contribution

Most methods of network anomaly detection are based on network traffic models. Brutlag uses as an extension of the Holt-Winters forecasting algorithm, which supports incremental model updating via exponential smoothing [1]. Hajji uses a Gaussian mixture model, and develops an algorithm based on a stochastic approximation of the Expectation-Maximization (EM) algorithm to obtain estimates of the model parameters [2]. Yamanishi et al. also assume a hierarchical structure of Gaussian mixtures in developing the "SmartSifter" tool, but uses different algorithms for updating the model parameters [3]. They use a variant of the Laplace law in the discrete domain, and a modified version of the incremental EM algorithm in the continuous domain. They test their algorithm to detect network intrusion on the standard ACM KDD Cup 1999 dataset. Lakhina et al. apply Principal Component Analysis (PCA) to separate IP network data into disjoint "normal" and "anomalous" subspaces, and signal an anomaly when the magnitude of the projection onto the anomalous subspace exceeds a threshold [4]–[6]. Huang et al. build on Lakhina's centralised PCA method of anomaly detection from [6], and develop a framework where local PCA analysis and stochastic matrix perturbation theory is used to develop an adaptive, distributed protocol [7]. Researchers have recently begun to use machine learning techniques to detect outliers in datasets from a variety of fields. Gardener et al. use a One-Class Support VectorMachine (OCSVM) to detect anomalies in EEG data from epilepsy patients [8]. Barbar´a et al. have proposed an algorithm to detect outliers in noisy datasets where no information is available regarding ground truth, based on a Transductive Confidence Machine (TCM) [9]. Transduction is an alternative to induction, in that instead of using all the data points to induce a model, one is able to use a small subset of them to estimate unknown properties of test points. Ma and Perkins present an algorithm using support vector regression to perform online anomaly detection on time series data in [10]. Ihler et al. present an adaptive anomaly detection algorithm that is based on a Markov-modulated Poisson process model, and use Markov Chain Monte Carlo methods in a Bayesian approach to learn the model parameters [11]. An example of a machine learning approach to network 1 anomaly detection is the time-based inductive learning achine (TIM) of Teng et al. [12]. Their algorithm constructs a set of rules based upon usage patterns. An anomaly is signaled when the premise of a rule occurs but the conclusion does not follow. Singliar and Hauskrecht use a support vector machine to detect anomalies in road traffic [13]. They use statistics collected by a sophisticated network of sensors including microwave loops and lasers, and design a detector for road traffic incidents.

III METHODOLOGY

In this section we discuss our model-free approach and provide the structure of an algorithm to detect temporal network anomalies. As noted in the introduction we focus on traffic at points of interest in the network, even though our approach is general enough to be applied to any trace of system activity. We assume that the *traffic trace* we monitor (in bits/bytes/packets/flows per time unit), corresponding to a specific time-of-day interval, can be characterized by a stationary model over a certain period ie, a month if no technological changes ,e.g., link bandwidth upgrades have taken place.

Consider a time series X1……Xn, of traffic activity . Let Yb the *partial sum o*r aggregate traffic over the time bucket starting at (t-1)b and containing samples. The crucial assumption we make is that is an IID. sequence for some appropriate bucket size . This is a reasonable assumption in many settings as temporal correlations tend to become weaker over longer time intervals. We quantize the values of the partial sums mapping them to the finite set of cardinality . For the rest of the paper, we will be referring to as the *underlying alphabet*. The quantization is done as follows: we let be the range of values takes, divide it into subintervals of equal length, and map to for .To select the appropriate size of the alphabet we follow the approach of [10] and use the so called Akaike's Information Criterion (AIC).

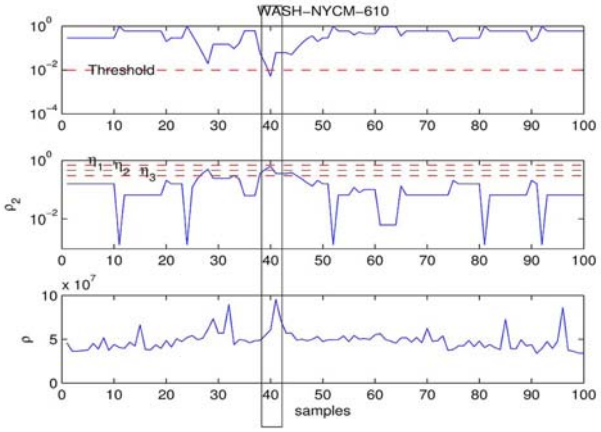

Figure 1: Graphical Representation For Model-free method

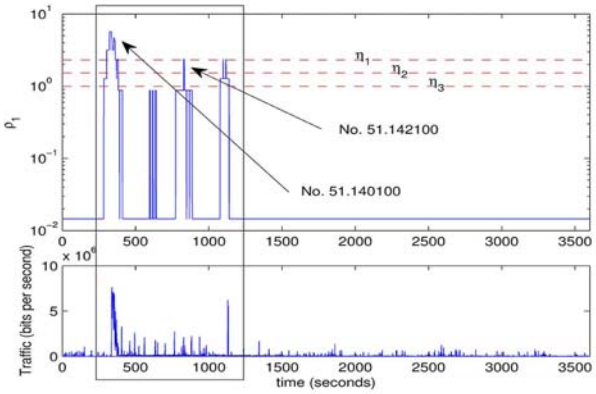

Figure 2: Graphical Representation For Model-based method.

Classifications

Client Model

A client is an application or system that accesses a remote service on another computer system, known as a server, by way of a network. The term was first applied to devices that were not capable of running their own stand-alone programs, but could interact with remote computers via a network. These dumb terminals were clients of the time-sharing mainframe computer.

Server Model

In computing, a server is any combination of hardware or software designed to provide services to clients. When used

alone, the term typically refers to a computer which may be running a server operating system, but is commonly used to refer to any software or dedicated hardware capable of providing services.

Network Model

Generally, the channel quality is time-varying. For the ser-AP association decision, a user performs multiple samplings of the channel quality, and only the signal attenuation that results from long-term channel condition changes are utilized our load model can accommodate various additive load definitions such as the number of users associated with an AP. It can also deal with the multiplicative user load contributions.

Empirical Measures for Anomaly Detection

As we mentioned before, the size of the alphabet and the number of states of the MMP for the Abilene data set is small when only temporal information is considered. Thus, it is easy to monitor subnets of PoPs (of low dimensionality) by specifying the group of PoPs of interest and the role of each PoP ,origin or destination. We present results for two case studies with different spatial characteristics. We apply our framework to: (a) flows that originate (end) from (at) PoPs that are 1-hop neighbors and (b) flows that originate (end) from (at) PoPs that are many hops away from each other. In the first case study, the flows originate (end) at the Sunny Valley (SNVA) PoP with destination (originating from) the PoPs in its vicinity. We demonstrate instances of the identification of anomalies applying the model-free and the model based methods, respectively. The values of the parameters for the two methods are obtained from the temporal anomaly detection examples. It is worth noticing that the detection rate reached 100% and the false alarms rate was very low ,lower than the values when only temporal anomalies were studied. This is due to two main reasons: (a) instantaneous high values in the time-series of observations that do not necessarily indicate attacks are smoothed due to time averaging, and (b) attacks may have temporal and spatial correlation.

Congestion Traffic Minimization

We demonstrate two different approaches, a model-free and a model-based one. The model-free method works on a longer time-scale processing traces of traffic aggregates over a small time interval. Using an anomaly-free trace it derives an associated probability law. Then it processes current traffic and quantifies whether it conforms to this probability law. The model-based method constructs a Markov modulated model of anomaly-free traffic measurements and relies on large deviations asymptotic and decision theory results to compare this model to ongoing traffic activity. We demonstrated a rigorous framework to identify traffic anomalies providing asymptotic thresholds for anomaly detection. In our experimental results the model-free approach showed a somewhat better performance than the model-based one. This may be due to the fact that the former gains from the aggregation over a time-bucket in addition to the fact that the latter one requires the estimation of more parameters, hence, it may introduce a larger modeling error. For future work, it would be interesting to analyze the robustness of the anomaly detection mechanism to various model parameters. Since we monitor the detailed distributional characteristics of traffic and do not rely on the mean or the first few moments we are confident that our approach can be successful against new types of temporal and spatial anomalies. Our technique is of low implementation complexity and is based on first principles, so it would be interesting to investigate how it can be embedded on routers or other networking devices.

Analysis of Network Client Server

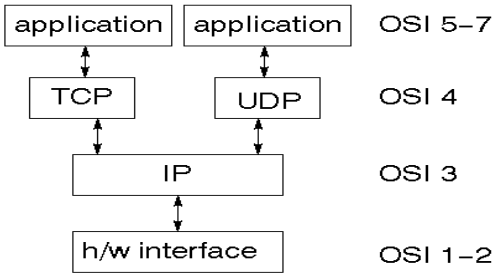

Figure 3: Diagram for TCP/IP layers

IV. EXPERIMENTAL RESULTS

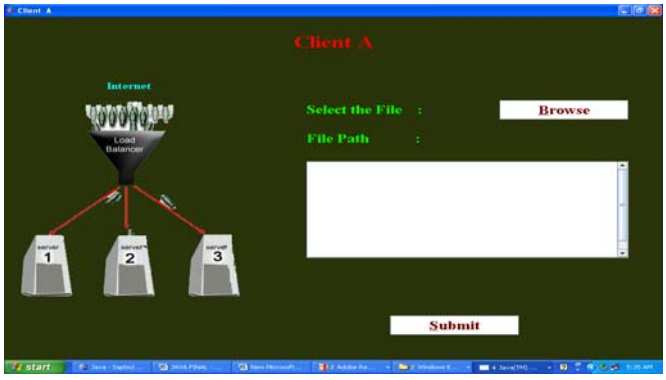

Figure 4: First Client

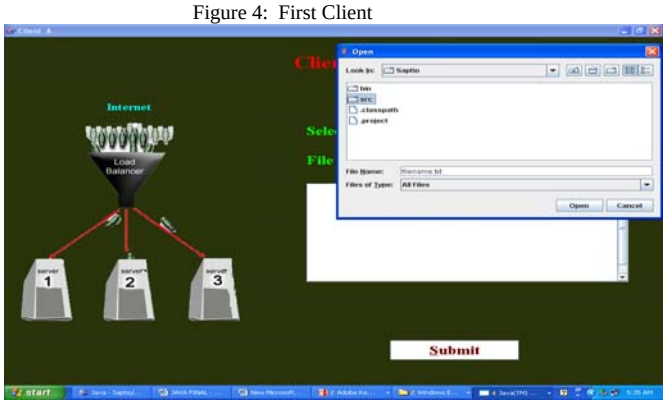

Figure 5: Selecting the required file from the network

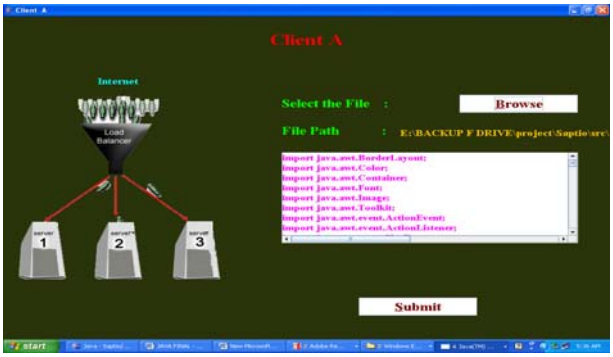

Figure 6: Sending the selected file

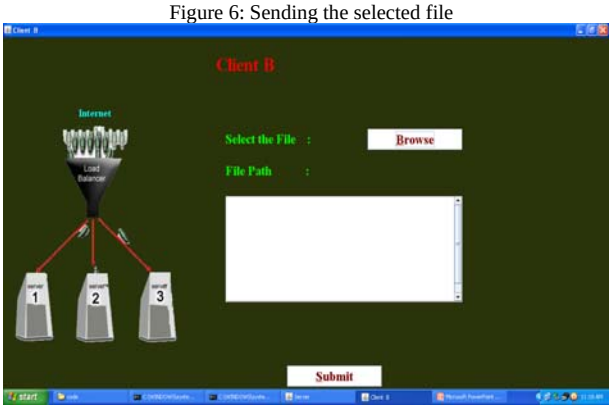

Figure 7: Second client

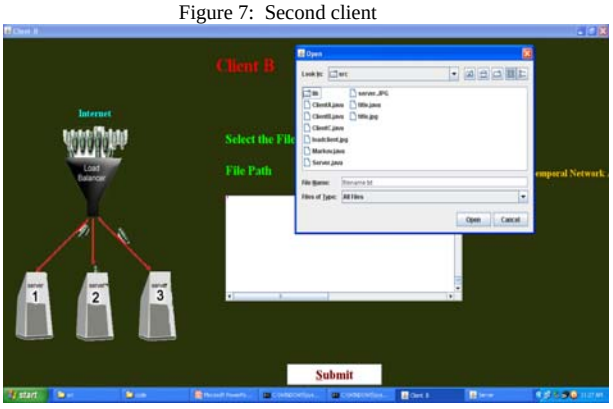

Figure 8: Selecting the required file from the network

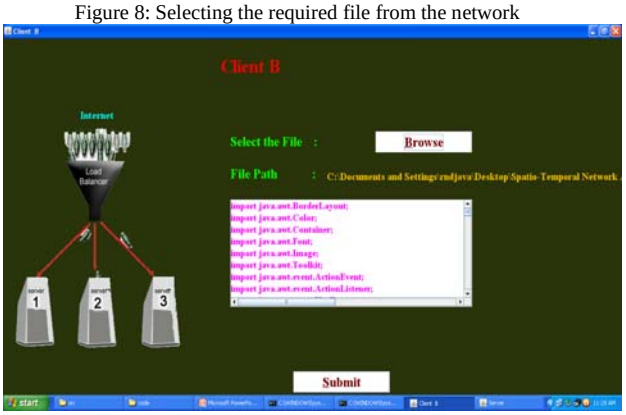

Figure 9: Sending the selected file

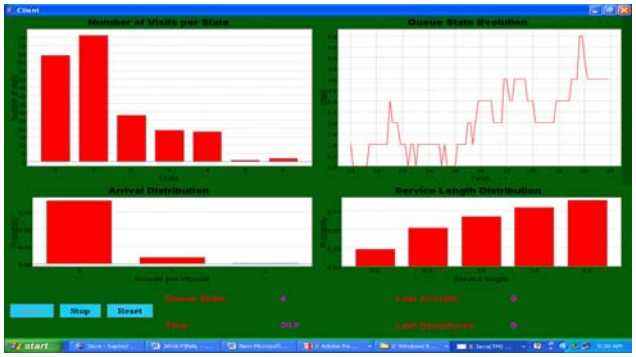

Figure 10: Graphical Representation

ACKNOWLEDGEMENTS

The authors thankful to the Ellenki College of Engineering and Technology - Research and Development Cell for collecting the data and preparation of this paper.

CONCLUSION

We introduced a general distributional fault detection scheme able to identify a large spectrum of temporal anomalies from attacks and intrusions to various volume anomalies and problems in network resource availability. We then showed how this framework can be extended to incorporate spatial information, resulting in robust temporal anomaly detection in large scale operational networks. Although most of the proposed anomaly detection frameworks are able to identify temporal or spatial anomalies, we are able to identify both as we preserve both the temporal and spatial correlation of network feature samples. In our approach presently when we send a data from client to server, it takes more time to retrieve large data. In future by using some more algorithms there is a scope that the data in huge amounts can be retrieved faster from any client. And the hacking of data can be prevented by following some new algorithms like encryption, digital signature.

REFERENCES

[1]Axelsson S. Research in intrusion detection systems: a survey.Technical report. Chalmers University of Technology. Goteborg
1998.

[2]Axelsson S. The Base-rate fallacy and its implications for the difficulty of intrusion detection. ACM Transactions on Information
and System Security 2000.

[3]Bermu´ dez-Edo M., Salazar-Herna´ndez R., Dı´az-Verdejo J.E.,Garcı´a-Teodoro P. Proposals on assessment environments for anomaly-based network intrusion detection systems. LNCS 4347; 2006. p. 210–21.

[4]Bridges S.M., Vaughn R.B. Fuzzy data mining and genetic algorithms applied to intrusion detection. In: Proceedings of the National Information Systems Security Conference; 2000.

[5]Cansian A.M., Moreira E., Carvalho A., Bonifacio J.M.Network intrusion detection using neural networks. In: International Conference on Computational Intelligence and Multimedia Applications (ICCMA'97); 1997.

[6]Cohen W.W. Fast effective rule induction. In: Proceedings 12th International Conference on Machine Learning; 1995. p. 115–23.

[7] M. Roesch, "Snort—Lightweight intrusion detection for networks," in *LISA '99: Proc. 13th USENIX Conf. System Administration*, Seattle, WA, Nov. 1999, pp. 229–238.

[8] V. Paxson, "Bro: A system for detecting network intruders in realtime," *Computer Networks*, vol. 31, no. 23–24, pp. 2435–2463, 1999.

[9] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in *Proc. ACM SIGCOMM Workshop on Internet Measurement*, Marseille, France, Nov. 2002, pp. 71–82.

[10] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 DARPA off-line intrusion detection evaluation," *Computer Networks*, vol. 34, no. 4, pp. 579–595, 2000.
[11] I. Paschalidis and S. Vassilaras, "Model-based estimation of buffer overflow probabilities from measurements," in *Proc. ACM SIGMETRICS 2001/Performance 2001 Conf.*, Cambridge, MA, Jun. 16–20, 2001, pp. 154–163.
[12] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proc. ACM SIGCOMM*, Portland, OR, Aug. 2004, pp. 219–230.
[13] A. Lakhina, M. Crovella, and C. Diot, "Characterization of networkwide anomalies in traffic flows," in *Proc. ACM SIGCOMM Internet Measurement Conf.*, Taormina, Italy, Oct. 2004, pp. 201–206.
[14] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Information Theory*, Budapest, Hungary, 1973, pp. 267–281.
[15] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1989.
[16] I. C. Paschalidis and G. Smaragdakis, "A large deviations approach to statistical traffic anomaly detection," in *Proc. 45th IEEE Conf. Decision and Control*, San Diego, CA, 2006, pp. 1900–1905.

ABOUT THE AUTHORS

Amjan Shaik is working as a Professor and Head, Department of Computer Science and Engineering at Ellenki College of Engineering and Technology (ECET), Hyderabad, India. He has received M.Tech. (Computer Science and Technology) from Andhra University. Presently, he is a Research Scholar of JNTUH Hyderabad. He has been published and presented 34 Research and Technical papers in International Journals , International Conferences and National Conferences. His main research interests are Software Engineering, Software Metrics, Computer Networks and Software Quality.

S.V. Achuta Rao is working as a Professor and Head, Department of CSE and IT at DJR Institute of Engineering and Technology (DJRIET), Vijayawada, India. He has received M.Tech. (Computer Science and Engineering) from JNTU, Kakinada, India. Presently, he is a Research Scholar of Rayalaseema University (RU), Kurnool, India. He has been published and presented good number of Research and technical papers in International and National Conferences. His main research interests are Data Mining, Networking, Image Processing, Software Engineering and Software Metrics.

Hymavathi. Bhadriraju is working as a Lecturer , Department of Computer Science and Engineering at Bharth University, Chennai, India. She has received M.Tech (CSE) from Bharth University, Chennai.,India. She has presented number of Technical papers in National Conferences. Her research interests are Software Engineering, Software Testing, Computer Networks, Network Security and Programming Languages.

Md. Riyazuddin is working as an Assistant Professor , Department of IT at Muffakham Jah College of Engineering and Technology (MJCET), Hyderabad, India. He has received M.Tech.(Information Technology) from JNTUH Hyderabad. He has resented number of technical papers in International and National Conferences. His research interests are Computer Networks, Data Mining, Information Security and Software Engineering.

Nazeer.Shaik is working as an Assistant Professor , Department of Computer Science and Engineering at Moghal College of Engineering and Technology (MCET), Hyderabad, India. He has received M.Tech (CSE) from Bharth University, Chennai. He has presented number of Technical papers in National Conference. His research interests are Software Engineering, Software Project Management, Computer Networks and Mobile Computing.