An Approach on Empirical Measures for Temporal Network Anomaly Detection

[1] **A.Pavana Deepthi** [2] **K.Radhika** [3] **D.Jamuna**

Abstract:

In this paper, we introduce an Internet traffic anomaly detection mechanism.. 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. Throughout, 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 security, statistical anomaly detection, method of types,

1. INTRODUCTION:

Although 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 70%. 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 this work 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. In contrast with other approaches this project is 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 in the way we characterize normal behavior and in how we assess deviations from it. More specifically, we propose 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 sequence of appropriately averaged system activity, and (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 composite hypothesis testing rule for assessing deviations from normal activity. 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.).In this paper, however, we focus on two case studies: (a) three different representations (bytes, packets and flows) of sampled origin–destination flow data from a backbone network, and (b) the aggregate traffic that arrives to or originates from the border router of some local area network (LAN) we wish to monitor. Traffic has diurnal variations which are primarily due to human activity. However, for relatively short time-scales (e.g., of about an hour), and especially during busy hours, stationary models can be appropriate. 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 i.i.d. 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.

This proposed paper focuses 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. 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 in the way we characterize normal behavior and in how we assess deviations from it.

2. 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.

2.1. 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 standalone programs, but could interact with remote computers via a network. These dumb terminals were clients of the time-sharing mainframe computer.

2.2. 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.

2.3. 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.

2.4. EMPIRICAL MEASURES FOR ANOMALY DETECTION:

As was 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). It presents results for two case studies with different spatial characteristics. It applies 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. It illustrates 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. 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/or spatial correlation.

2.5. CONGESTION TRAFFIC MINIMIZATION:

Two different approaches, a model-free and a model-based one are provided here. 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. A rigorous framework to identify traffic anomalies providing asymptotic thresholds for anomaly detection is provided here. 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 (emerging) temporal and spatial anomalies. Our method is of low implementation complexity (only an additional counter is required), and is based on first principles, so it would be interesting to investigate how it can be embedded on routers or other network devices.

3. FRAME WORK:

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. 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.

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 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.

4. Approaches to network anomaly detection:

4.1. Statistical Approaches for Network Anomaly Detection:

General steps involved in statistical anomaly detection. The first step is to preprocess or filter the given data inputs. This is an important step as the types of data available and the time scales in which these data are measured can significantly affect the detection performance. In the second step, statistical analysis and/or data transforms are performed to separate normal network behaviors from anomalous behaviors and noise. A variety of techniques can be applied here, e.g., Wavelet Analysis, Covariance Matrix analysis, and Principal Component Analysis. The main challenge here is to find computationally efficient techniques for anomaly detection with low false alarm rate. In the final step, decision theories such as Generalized Likelihood Ratio (GLR) test can be used to determine whether there is a network anomaly based on the deviations observed. In a broader context, statistical anomaly detection can also be viewed from the machine learning perspective, where the goal is to find appropriate discriminate functions that can be used to classify any new input data vector into the normal or anomalous region with good accuracy for anomaly detection. One subtle difference between statistical anomaly detection and machine learning based methods is that statistical approaches generally focus on statistical analysis of the collected data, whereas machine learning methods focuses on the “learning” part.

4.2. Discrete Algorithms for Network Anomaly Detection:

In many cases, network anomaly detection involves tracking significant changes in traffic patterns such as traffic volume or the number of traffic connections. Due to the high link speed and the large size of the Internet, it is usually not scalable to track the per-flow status of traffic. In Duffield et al. proposed packet sampling to monitor flow-level traffic. By limiting the number of flows that need to be monitored, sampling can partially solve the scalability problem at the cost of anomaly detection performance. In this area, one important issue is to investigate the tradeoff between the amount of sampled information and the corresponding performance.

Simple sampling cannot fully solve the scalability problem as any packets or flows that are not sampled may contain important information about anomalies. Furthermore, it is likely that this information can only be recovered if these packets or flows are sampled and stored. Thus, sometimes a large number of flows (based on different combinations of source and destination IP addresses) may need to be sampled to achieve an anomaly detector with good performance. To address the disadvantages of sampling approaches, there has been extensive research in data streaming algorithms for anomaly detection in high-speed networks. Specifically, using streaming techniques, the anomaly detection problem can be formulated as a heavy-hitter detection problem or a heavy-change detection problem. In the heavy-hitter detection problem, the goal is to identify the set of flows that represent a significantly large proportion of the ongoing traffic or the capacity of the link . In the heavy-change detection problem, the goal is to detect the set of flows that have drastic change in traffic volume from one time period to another.

5. Conclusion:

In this paper 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. More specifically, we propose two methods to characterize normal behavior: a model-free approach employing the method of types to characterize the type of an independent and identically-distributed sequence of appropriately averaged system activity, and a model-based approach where system activity is modeled using a Markov Modulated Process (MMP).

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. Finally we proposed approaches to network anomaly detections.

References

[1]E. S. Al-Shaer and H. H. Hamed. “Discovery of policy anomalies in distributed firewalls”. In IEEE Infocom, 2004.

[2].Al-Shaer and H. Hamed, “Conflict classification and Analysis of Distributed Firewall policies”, IEEE J SEL AREA COMM, 2005

[3].Chotipat Pornavalai and Thawatchai Chomsiri."Firewall Rules Analysis", International Technical Conference on Circuits/Systems, Computers & Comm. (ITC-CSCC 2004), JULY 2004.

[4].Thawatchai Chomsiri, Chotipat Pornavalai: Firewall Rules Analysis, International Conference on Security &

Management, SAM 2006, Las Vegas, Nevada, USA, June 26-29, 2006.

[5].Deri Luca and Suin Stefano and Maselli Gaia (2003) Design and implementation of an anomaly detection system: An empirical approach. In Proceedings of Terena TNC .

[6].Y. Bartal, A.J. Mayer, K. Nissim, A. Wool, Firmato: A novel firewall management toolkit, in: Proceedings of the IEEE Symposium on Security and Privacy, 1999

[7].Errin W. Fulp. "Optimization of network firewall policies using ordered sets and directed acyclical graphs". Technical report, Computer Science Department, Wake Forest University, 2004

[8].Yu Gu, Andrew McCallum and Don Towsley. "Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation", Tech. rep., Department of Computer Science, UMASS, Amherst, 2005.

[9].F. Cuppens, N. Cuppens, and J. Garc´ıa. "Detection and removal of firewall misconfiguration". In International conference on Communication, Network and Information Security (CNIS2005), Phoenix, AZ, USA, November 2005. IASTED.

[10].E. Al-Shaer and H. Hamed. "Firewall Policy Advisor for Anomaly Detection and Rule Editing." IEEE/IFIP Integrated Management Conference (IM'2003), March 2003.

[11].Salem, O., Vaton, S. and Gravey, A. (2007). A novel approach for anomaly detection over high-speed networks. In, Proceedings of EC2ND.

[12]. Shin Ando, Clustering Needles in a Haystack: An Information Theoretic Analysis of Minority and Outlier Detection, Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, p.13-22,October28-31, 2007.

Authors Information:

[1] **A.Pavana Deepthi** pursuing M.Tech CSE from jaya prakash Narayana College of Engineering & Technology. MCA from Villa Mary College for women, Hyderabad. Her areas of interest include Network Security, Software Engineering and Sensor Network.

[2]**K.RadhikaD.**, Working as Associate Professor CSE Dept. Jayaprakash Narayan College of Engineering, Mahabubnagar, M.Tech(SE) from B.Tech from Sri Sai jyothi college, JNTUH, Hyderabad. B.Tech (CSE) from Jayaprakash Narayan College of Engineering, Mahabubnagar Experience 8 Years in Teaching Profession. Her areas of Interest are in Wireless Sensor Networks

[3] **Prof.D.Jamunna**, Working as Professor & Head of CSE Dept. Jayaprakash Narayan College of Engineering, Mahabubnagar, M.Tech(SE) from School of Information Technology, JNTUH, Hyderabad. BE (CSE) from Vijayanagara Engineering College, Bellary. Experience 17 Years in Teaching Profession. Her areas of Interest are in Wireless Sensor Networks, Data Mining, and Networking and guided M. Tech and B. Tech Students IEEE Projects.