International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.2, Issue No.12, pp : 1173-1177 1 Dec. 2013

Review of Anomaly Detection based on Classifying Frequent Traffic Patterns

Mayank Nagar, Mr.Sandeep Kumar, Proff.Udaypal Reddy

#CSE, RGPV,IES College Of Technology Bhopal(M.P.),India
Author: nagar.mayank22@gmail.com.,

Co guide:*sandeep.kumar@outlook.in*

ABSTRACT: *As internet grow rapidly and numerous applications use it as a larger communication media. Many organizations are Internet dependent for their working methodology. This will lead a larger amount of network traffic. Network traffic anomaly is kind of status that the traffic activities deviated from its normal condition behaviors. The characteristic of network traffic anomaly is that it explodes unexpectedly without any warning. It will lead immense damage to networks and network apparatus in a minimum time. Detection of network traffic anomalies is vital for network operators as it facilitates to classify security incidents and to observe the availability of network related services. In this paper we are going to present various methods or schemes for anomaly detection. In this paper we also focus on various schemes that are applicable to enhance the network performance.*

1. INTRODUCTION

From last few years cyber attacks are increases speedily over the network. as a result, the anomalies detection on network traffic data also has been considered lengthily. As network traffic increases it may lead damage and increases the chances of attacks also. Besides to anomaly detection, anomaly classification i.e., automatically recognizing the kind of a detected anomaly has been slightly matter of worry [1]. Protecting networks from various attacks is a vital apprehension of computer security. Inclusive compilation and truthful explanation of traffic information are core problems in network traffic anomaly detection. As network traffic may lead to variety of information exchange and sensitive data transfer. Although it is also well known that the dependency of network are also emerging rapidly. Due to this the network condition are very crucial now a days and it will become more complicated in forthcoming time. This traffic may lead to massive damage of network system and its related resources. Network behaviour examine is comes under Anomaly detection.

Normal functionality of anomaly detection is to categorize the traffic either traffic is normal or anomalous. The abnormal behaviors that may occur in the network or system are identified through the description and analysis of the network traffic, and send alerts to the administrator. This procedure is distinct as network traffic anomaly detection [2]. But this will not so much efficient to detect behaviour of traffic easily and earlier. Comprehensive collection and truthful explanation about the traffic related information are main troubles in network traffic anomaly detection. The abnormal traffic has carried out huge destruction to the network, and there are many more network traffic anomalies along with the speedy recognition of network applications. Hence, to distinguish anomaly quickly and accurately and make sensible response has become the striking and important aim in the current academic and industrial circles [2]. Generally network anomaly detection methodology relies on the investigation of network traffic and the characterization of the vibrant statistical belongings of traffic normality, with the intention of accurately and timely detects network anomalies. Anomaly detection is based on the theory that perturbation of standard behavior suggests the existence of anomalies, attacks, faults, etc [3].

Based on the natural complexity in characterizing the normal network performance, the difficulty of anomaly detection may be categorized as model based and non-model based. According to model based anomaly detectors, it is assumed that a identified model is accessible for the normal behavior of definite specific aspects of the network and any divergence from the norm is supposed an anomaly. Network behaviors that cannot be characterized by any model for such condition non-model based approaches are used. Non-model based approaches can be auxiliary classified based on the unambiguous implementation and accuracy constraints that have been imposed on the detector.

Let's discuss some common approaches for network anomaly detection.

A. Statistical Approaches for Network Anomaly Detection

Statistical approach uses some steps for detecting network anomaly. The first step is to pre-process or filter the specified data inputs. This is a significant step as the types of data presented and the time scales in which these data are measured can significantly distress the detection performance [4]. In the second step, statistical examination and/or data transforms are performed to take apart normal network behaviors from anomalous behaviors and noise. A diversity of techniques can be applied here, like Covariance Matrix analysis, Wavelet Analysis, and Principal Component Analysis. The primary challenge here is to find computationally proficient techniques for anomaly detection with low false alarm rate. At last in final step, decision theories for instance Generalized Likelihood Ratio (GLR) test can be used to determine whether there is a network anomaly based on the deviations observed.

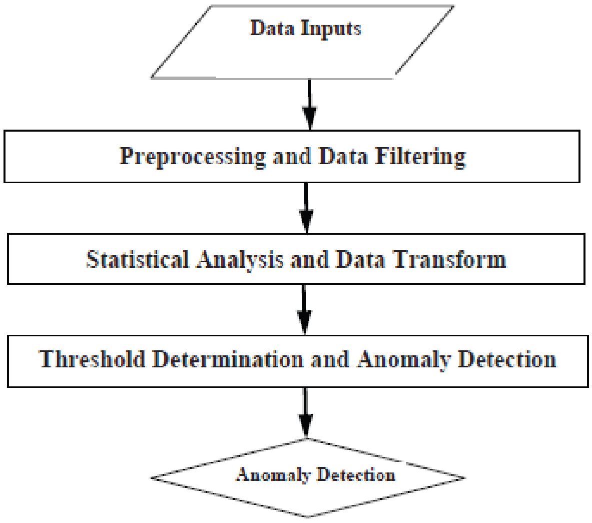
Data Inputs
Preprocessing and Data Filtering
Statistical Analysis and Data Transform
Threshold Determination and Anomaly Detection
Anomaly Detection

Figure 1: Statistical Approach for Network Anomaly Detection [4].

In a larger context, statistical anomaly detection can also be inspected from the machine learning point of view, where the objective is to find appropriate discriminant functions that can be accessed to classify any new input data vector into the normal or anomalous region with excellent accuracy for anomaly detection. One restrained difference among 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. Some of them are discussed below.

Change-Point Detection: Change-point detection is the difficulty of discovering time points at which properties of time-series data change. This includes a broad range of real-world problems and has been vigorously conversed in the community of statistics and data mining. A representative statistical formulation of change-point detection is to consider probability distributions from which data in the past and present intervals are generated, and observe the intention time point as a change point if two distributions are significantly different. In these methods, the logarithm of the probability ratio between two successive intervals in time-series data is observed for detecting change points.

Wavelet Analysis: Wavelet analysis has been applied to modeling of non-stationary data series because it can characterize the scaling properties in the temporal and frequency domains. The wavelet transform can get arbitrary signal characteristic of time-frequency domain, which can help to explore the transient abnormal phenomenon from normal signals and demonstrate its components. Researchers used wavelet analysis to detect anomaly just based on the differences between the normal and anomalous traffic signals in the frequency domain.

Covariance Matrix Analysis: By employing covariance matrix analysis has been show to be a powerful anomaly detection method. Each component in the covariance matrix corresponds to the correlation among two monitored features at different sample sequences. The norm profile of the normal traffic can then be described by the mathematical expectation of all covariance matrices constructed from samples of the normal class in the training dataset. The covariance matrix method is extended, where the sign of the covariance matrices is used directly for anomaly detection.

Principal Component Analysis: Principal Component Analysis (PCA) is a dimensionality-reduction method of mapping a set of data points onto new coordinates. The spirit of PCA based anomaly detection is to separate the normal behavior from anomalies through dimensionality-reduction. The basic idea of using PCA for traffic anomaly detection is that: the *k*-subspace obtained through PCA corresponds to the normal behavior of the traffic, whereas the remaining (*n*-*k*) subspace corresponds to either the anomalies or the anomalies and the noise. Each new traffic measurement vector is projected on to the normal subspace and the anomalous subspace. Afterwards, different thresholds can be set to classify the traffic measurement as normal or anomalous. Later on the source of the inconsistent traffic can then be pinpointed by determining the ingress and egress points of different traffic flows.

B. Discrete Algorithms for Network Anomaly Detection

In numerous cases, network anomaly detection involves tracking noteworthy changes in traffic patterns such as traffic amount or the number of traffic connections. Due to the high link speed and the large volume of the Internet, it is generally not scalable to trace the per-flow status of traffic. By limiting the number of flows that require to be monitored, sampling can incompletely solve the scalability problem at the cost of anomaly detection performance. However, simple sampling cannot completely solve the scalability problem as any packets or flows that are not sampled may contain important information about anomalies. Furthermore, it is expected that this information can only be recovered if these packets or flows are sampled and stored. Specifically, using streaming techniques, the anomaly detection trouble can be formulated as a heavy-hitter detection problem or a heavy-change detection problem. In the heavy-hitter detection problem, the main target is to recognize 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.

Heavy-Hitter Detection: In the perspective of network anomaly detection, the object of heavy-hitter detection is to efficiently recognize the set of flows that represent a significantly huge proportion of the link capacity or of the active traffic with small memory requirements and limited state information. The challenge with heavy-hitter detection is that data processing cannot be done on a per-flow basis due to the bulky bandwidth of the current network. Thus, stream algorithms based on summary structures are used to resolve this problem with assured error bounds. In data mining, there has been widespread research of algorithms for heavy-hitter detection.

Heavy-Change Detection: The purpose of heavy-change detection is to efficiently discover the set of flows that have drastic change in traffic volume from one time period to another with small memory requirements and limited information. It can be formulated just like to the heavy-hitter detection problem. Clearly heavy-change detection is a harder problem than heavy-hitter detection. In heavy-change detection, one data stream computation technique called sketch has shown immense potential. The fundamental idea is to précis the input streams so that per-flow investigation can be avoided. The sketch-based techniques uses a small amount of memory and has constant prerecord update and modernization costs, thus it can be used for change detection in high-speed networks with a large number of flows.

2. BACKGROUND

The numerous attacks on network infrastructure, using a variety of forms of denial of service (DoS) attacks and worms, have lead to an increased necessitate for developing techniques for analyzing and monitoring network traffic. If proficient analysis tools were available, it could become feasible to identify the attacks, anomalies and obtain action to restrain them before much time to propagate across the network. The motivation for this work came from a need to decrease the likelihood that an attacker may hijack the campus machines to stage an attack on a third party. A campus may desire to prevent or limit misuse of its machines in staging attacks, and probably bound the liability from such attacks. Traffic anomalies, such as flash crowds, denial-of-service attacks, port scans, and the spreading of worms, can have damaging effects on Internet services. Detecting and diagnosing these anomalies is decisive to network operators, who must take counteractive action to assuage congestion, warn affected users and block attacks.

3. RELATED WORK

Ignasi Paredes-Oliva et al [1] proposed Practical Anomaly Detection based on Classifying Frequent Traffic Patterns. They introduce a novel scheme and build a system to sense and classify anomalies that are based on an elegant blend of frequent item-set mining with decision tree learning. This scheme automatically identifies and classifies anomalies in high-speed networks using traffic flow data, like Net Flow. They combines normally two techniques on is from data mining and another is machine learning [1].

In this method frequent item-set mining (FIM) is used to find a set of *frequent item-sets* (FIs) firstly. A frequent item-set is a great set of flows that have one or more flow features in common. Secondly builds a decision tree to categorize frequent item-sets as anomalous or benevolent and to determine their specific type in anomalous case. Instinctively, they decompose observed traffic into distinct groups (FIs) of related traffic flows that allows us to categorize each FI with high accuracy [1].

Frequent Item-set Mining: Frequent item-set mining (FIM) is a well-known data mining method that focuses on finding items that arise frequently together in a certain dataset. A set of items will be measured frequent if they emerge together at least as many times as a specified threshold, which is described as *minimum support*. Concerning FIM to network traffic permits us to decide groups of numerous flows sharing a certain combination of features [1].

They had implemented an anomaly detection and classification system and organized it in a construction network, where it effectively monitors two 10 Gb/s links. Furthermore, a particularly promising characteristic of used classifier is that it has been trained using traffic traces from the European backbone network of G´EANT and has been used successfully to detect and classify anomalies in a substantially different regional network [1].

Yingjie Zhou Guangmin Hu Weisong He recommended Using Graph to Detect Network Traffic Anomaly [2]. In this a network traffic anomaly detection method based on time-series graph mining. It perfectly and completely describes the relationships among multi-time series which are used in traffic anomaly detection by time-series graph, and can efficiently detect the network traffic anomaly; especially DDos attacks [2].

In year 2011, J. A. Barria and S. Thajchayapong proposed Detection and Classification of Traffic Anomalies using Microscopic Traffic Variables [5]. They recommend a novel anomaly detection and classification algorithm that explicitly utilizes the chronological changes in discrepancy and the changes in spatial covariances of microscopic traffic variables, explicitly relative speed, inter-vehicle time gap and lane changing. This method concerns a novel method using the smallest eigenvalue of covariance matrix to imprison changes in microscopic characteristics as well as to assess their severity. The performance of the projected algorithm is also assessed under partial availability of individual vehicle information. There analysis framework is based on a distributed traffic monitoring system that could rely on locally shared information amongst neighbouring vehicles to compute microscopic traffic variables and assess road traffic situation on a freeway segment [5].

In year 2012, Adaniya et al [6] proposed Anomaly detection using DSNS and Firefly Harmonic Clustering Algorithm. They recommended a new algorithm named Firefly Harmonic Clustering Algorithm (FHCA) for quantity anomaly detection using Digital Signature of Network Segment (DSNS) achieving satisfactory results in precision and accuracy with true-positive rates in 80% and false-positive rates in 20%. The first step to identify anomalies is to accept a model that describes the network traffic efficiently, which represents a considerable challenge due to the non-stationary environment of network traffic. Large networks traffic performance is composed by daily cycles, where traffic levels are typically higher in working hours and are also dissimilar for workdays and weekends. Thus, the GBA tool is used to produce different profiles of standard behavior for each day of the week, convention this requirement. The DSNS can be describing as a set of information that comprises the traffic profile of a network segment or server. This information embraces data such as traffic volume or number of errors, among others. The DSNS was produced by a model that performs a statistical analysis of the record of data collected in the SNMP objects, taking into account the accurate moment of the collection [6].

Curtis Storlie et al [7] proposed Graph Based Statistical Analysis of Network Traffic. They suggest a graph-based method for analyzing traffic patterns in a huge computer network in permutation with novel statistical methods for determining time-related anomalies in data with diurnal trends. They model the traffic as a graph and extort the sub graphs corresponding to individual sessions and use them to develop a statistical model for the network traffic. The aim of this analysis is to find out patterns in the network traffic data that might indicate intrusion activity or other malicious behavior. They also described a statistical method for analyzing TSG decompositions that obtains into account the diurnal patterns of the communications and computes on that basis a predictive model for prospect traffic that can be used to perceive anomalies [7].

In year 2012, Iwan Syarif et al [8] presented unsupervised clustering approach for network anomaly detection. They describe the advantages of using the anomaly detection scheme over the misuse detection technique in detecting unidentified network intrusions or attacks. It also examines the performance of a variety of clustering algorithms when applied to anomaly detection. They implement and compare the performance of five different clustering algorithms in our anomaly detection module which are k-Mean, improved k-Mean, k-Medoids, EM clustering and distance-based outlier detection algorithms [8].

Their results shows that the misuse detection procedure achieves a very good performance result with more than 99% correctness when detecting identified intrusion but it fails to perfectly detect data set with a large number of unidentified intrusions where the highest accuracy result is only 63.97%. In contrast, the anomaly detection scheme shows promising outcomes where the distance-based outlier detection technique outperforms the other three clustering algorithms with the exactness of 80.15%, pursued by EM clustering (78.06%), k-Medoids (76.71%), enhanced k-Means (65.40%) and k-Means (57.81%). Supplementary experiment shows that the distance-based outlier detection achieved very well in detecting probing attacks (83.88%) and DoS attacks (82.21%) but it be unsuccessful to detect R2L attacks (42.44%) and U2R attacks (52.73%) [8].

In the same year 2012, Monowar Hussain Bhuyan et al [9] recommended Towards an Unsupervised Method for Network Anomaly Detection in Large Datasets. They present an efficient tree based subspace clustering procedure (TreeCLUS) for searching clusters in network intrusion data and for detecting recognized as well as unidentified attacks without using any labeled traffic or signatures or training. In this scheme they also introduce a multi-objective cluster labeling method to label each constant cluster as normal or anomalous. The major attractions of this anticipated scheme are: (i) TreeCLUS do not require the number of clusters apriori, (ii) it is liberated from the limitation of using any proximity measure, (iii) CLUSLab is a multi-objective cluster labeling procedure including effective unsupervised characteristic clustering technique for identifying overriding feature subset for each cluster, and (iv) TreeCLUS exhibits a high recognition rate and a low false positive rate, particularly in case of probe, U2R, and R2L attacks. The proposed method is established superior as compared with other competing network anomaly detection techniques [9].

Again in same year Anup Bhange and Sumit Utareja proposed Anomaly Detection and Prevention in Network Traffic based on Statistical approach and α-Stable Model [10]. As per their research proposals in anomaly detection characteristically follow a four-stage approach. In this scheme the first three stages define the detection mechanism, while the last stage is dedicated to authenticate the scheme. Consequently, in the first stage, traffic data are collected from the network (known as data collection). Second, data are analyzed to extort its most relevant features (i.e. data analysis). Third, traffic is classified as normal or abnormal (inference); and fourth, the entire approach is validated with different types of traffic anomalies [10].

Manmeet Kaur Marhas, Anup Bhange, Piyush Ajankar recommended Anomaly Detection in Network Traffic: A Statistical Approach [11]. They present a statistical scheme to investigate the allocation of network traffic to be acquainted with the normal network traffic performance. The potential to sense unknown attacks is the potency of statistical anomaly detection systems. Anomaly detection systems originate a model of the normal behavior of a network or system and detect divergence from this normal profile. This enables them to discover acknowledged and mysterious malicious activities likewise. The normal profile has been derived based on different Information such as system calls on a single host, payload byte patterns in received traffic, or volume and entropy Information over the traffic in a whole network [11].

Hao Zhang et al [12] proposed User Intention-Based Traffic Dependence Analysis for Anomaly Detection. They explain a novel approach that can be used for detecting anomalous traffic on a host. This scheme investigates direct and indirect dependencies in how a user interacts with applications and how applications respond to the user's requests following the specifications of the applications. By enforcing an application's correct responses to user actions, they are capable to identify vagabond events. Vagabond events are nothing but to outbound network events that are not generated by any user actions and may hence be due to anomalies [12].

This work aims to demonstrate the feasibility of user intention-based dependence analysis for detecting suspicious network connections of a host in a concrete web browser setting. They enforce correct system behaviors, as opposed to anomalous characteristics. Their user intention-based traffic dependence analysis produces structures in network events. These structures across outbound requests enable improved context-aware security analysis. Dependence analysis on network flows builds a traffic-dependency graph based on the observed network events and user actions [12].

Analyzing the dependencies between network traffic and user activities has not been systematically investigated as a general approach for anomaly detection. Traffic dependency graph captures the causal relations of user actions and network events for improving host integrity. Result indicated that the feasibility of enforcing HTTP traffic dependencies [12].

Venkatesh Saligrama and Manqi Zhao proposed Local Anomaly Detection [13]. They recommended a novel graph-based statistical conception that combines the idea of temporal and spatial locality. This notion provide itself to an elegant characterization of optimal decision rules and in turn suggests corresponding empirical rules based on local nearest neighbor distances. They also show complex scoring proposal overcomes the inherent resolution issues of alternative multi-comparison approaches that are based on fusing the outcomes of location-by-location comparisons [13].

4. CONCLUSION

Network traffic anomaly refers to the status that the traffic behaviors deviated from its normal behaviors. It can bring great damage to networks and network equipments in a short time. Existing traffic anomaly detection methods usually treat time-varying traffic information as a one-dimensional signal, and detect traffic anomaly through a variety of signal analysis methods. In this paper we give general review of such techniques that are used to detect and classify Anomalies. But it still required improvements.

REFERENCES

i. *Ignasi Paredes-Oliva, Ismael Castell-Uroz, Pere Barlet-Ros, Xenofontas Dimitropoulos and Josep Sol´e-Pareta "Practical Anomaly Detection based on Classifying Frequent Traffic Patterns", IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS-2012), pp. 49- 54, 2012.*

ii. *Yingjie Zhou Guangmin Hu Weisong He "Using Graph to Detect Network Traffic Anomaly", International Conference on Communications, Circuits and Systems (ICCCAS-2009), pp. 341 – 345, 2009.*

iii. *V. Chatzigiannakis, G. Androulidakis, K. Pelechrinis, S. Papavassiliou and V. Maglaris "Data fusion algorithms for network anomaly detection: classification and evaluation", Third International Conference on Networking and Services (ICNS-2007), pp. 50, 2007.*

iv. *Thottan M., Ji C. "Anomaly Detection in IP Networks", IEEE Trans. Signal Processing, Special Issue of Signal Processing in Networking, Vol. 51, No. 8, pp. 2191-2204, 2003.*

v. *J. A. Barria and S. Thajchayapong "Detection and Classification of Traffic Anomalies using Microscopic Traffic Variables", IEEE Transactions on Intelligent Transportation Systems, Volume: 12, Issue: 3, pp. 695 – 704, 2011.*

vi. *Mario H. A. C. Adaniya, Moises F. Lima, Joel J. P. C. Rodrigues, Taufik Abrao and Mario Lemes Proenc "Anomaly detection using DSNS and Firefly Harmonic Clustering Algorithm", IEEE International Conference on Communications (ICC-2012), pp. 1183 – 1187, 2012.*

vii. *Hristo Djidjev, Gary Sandine, and Curtis Storlie "Graph Based Statistical Analysis of Network Traffic", Los Alamos National Lab, MLG '11 San Diego, CA, USA-2011.Onlineavailableat: http://users.cis.fiu.edu/~lzhen001/activities/KDD2011Program/workshops/MLG/doc/paper_10.pdf*

viii. *Iwan Syarif, Adam Prugel-Bennett, Gary Wills "Unsupervised clustering approach for network anomaly detection", Networked Digital Technologies Communications in Computer and Information Science, Volume 293, pp 135-145, 2012.*

ix. *Monowar Hussain Bhuyan, Dhruba Kr. Bhattacharyya and Jugal K. Kalita "Towards an Unsupervised Method For Network Anomaly Detection in Large Datasets", Computing and Informatics, Vol. 1, issue 4, pp. 1–32, 2012.*

x. *Anup Bhange and Sumit Utareja "Anomaly Detection and Prevention in Network Traffic based on Statistical approach and α-Stable Model", International Journal of Advanced Research in Computer Engineering & Technology, ISSN: 2278 – 1323, Volume 1, Issue 4, pp. 690 – 698, June 2012.*

xi. *Manmeet Kaur Marhas, Anup Bhange, Piyush Ajankar " Anomaly Detection in Network Traffic: A Statistical Approach", International Journal of IT, Engineering and Applied Sciences Research (IJIEASR), ISSN: 2319-4413, Volume 1, No. 3, pp. 16 – 20, December 2012.*

xii. *Hao Zhang, William Banick, Danfeng Yao and Naren Ramakrishnan "User Intention-Based Traffic Dependence Analysis for Anomaly Detection", IEEE Symposium on Security and Privacy Workshops (SPW-2012), pp. 104 – 112, 2012.*

xiii. *Venkatesh Saligrama and Manqi Zhao "Local Anomaly Detection", in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS- 2012), pp. 969 – 983, 2012.*