ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 10, December 2012

Anomaly Exposure in Network Traffic and its collision: a Survey

Saurabh Ratnaparkhi
M.Tech Scholar
Dept. of CSE
CMJ University
Shilong

Anup Bhange
Asst.Prof
Dept. of IT
KDKCollege of Engg.
Nagpur

ABSTRACT:

Antagonistic network traffic is frequently "diverse" from kind traffic in ways that can be classified without knowing the way of the attack.
Anomaly exposure is a significant problem that has been investigates within diverse research areas and application domains. Many anomaly exposure techniques have been specifically residential for certain application domains, while others are more generic. These survey papers provide a ordered and complete overview of the research on anomaly exposure. We have grouped existing techniques into different categories based on the underlying approach adopted by each technique. For each group we have recognized key assumptions, which are used by the method to distinguish between normal and anomalous behavior.
For each category, we discuss a basic anomaly exposure technique, and then show how the different accessible techniques in that group are variants of the basic technique. Here our paper focus on anomaly exposure in network and its impact over the network and classify the different technique for detecting anomaly in network.

Keywords: ***Gaussian Mixture Model, EM Algorithm, Time Slice Window***

1. INTRODUCTION:

Malicious mistreatment of the Internet is commonly seen in to-day's Internet traffic. Anomalies such as worms, port scans, denial of service attacks, etc. can be found at some time in the network traffic. These anomalies dissipate network re-sources, because recital deprivation of network devices and end hosts, and lead to security issues about all Internet users. Thus, exactly identify such anomalies has become a significant problem for the network society to solve. Anomaly exposure refers to the problem of counting patterns in data that do not be conventional to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains of these, anomalies and outliers are two terms used most commonly in the context of anomaly exposure; Sometimes interchangeably. Anomaly exposure finds general use in a wide diversity of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities.

The significance of anomaly exposure is due to the fact that anomalies in data interpret to significant (and often critical) actionable information in a wide assortment of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an Un-authorized destination [Kumar (1)].
An anomalous MRI image may signify presence of malignant tumors [2]. Anomalies in credit card trans-action data could point to credit card or identity theft [Aleskerov et (3)] or anomalous readings from a space craft sensor could indicate a fault in some part of the space craft [Fujimaki et al. (4)]. Detecting outliers or anomalies in data has been deliberate in the statistics community as early as the 19thcentury [Edgeworth (5)]. In excess of time, a variety of anomaly exposure technique has been developed in several research communities. Many of this technique have been purposely developed for certain request domains, while others are more generic. This survey tries to give an ordered and complete overview of the research on anomaly detection.
It assist a better sympathetic of the diverse directions in which research has been done on this topic, and how method developed in one area can be applied in domains for which they were not intended to begin with the cost of the information dispensation and Internet convenience falls, more and more organizations are becoming susceptible to a wide variety of cyber threats. According to a new survey by CERT/CC [6], the rate of cyber attacks has been extra than repetition every year in recent times. There-fore, it has become ever more significant to make our in-formation systems, particularly those used for critical purpose in the military and commercial sectors, resistant to and tolerant of such attacks. The most widely deployed methods for detecting cyber terrorist attacks and protecting against cyber terrorism employ signature-based detection techniques. Such technique can only detect formerly known attacks that have a matching signature, since the signature database has to be physically revised for each new type of attack that is discovered. These limitations have led to an increasing interest in intrusion detection techniques based on data mining [7].
Intrusion detection techniques generally fall into one of two group; misuse detection and anomaly detection. In misuse detection, each case in a data set is labeled as 'normal' or 'intrusive' and a learning algorithm are skilled over the labeled data. These methods are able to mechanically retrain

intrusion detection models on dissimilar input data that comprise new types of attacks, as long as they have been labeled appropriately. Research in misuse recognition has focused mainly on classification of network intrusions using various standard data mining algorithms [7,8], rare class predictive models, connection rules [2, 5] and cost sensitive modeling. Unlike signature-based intrusion detection systems, models of misuse are created mechanically, and can be more complicated and precise than physically created signatures. A key advantage of misuse recognition method is their high degree of correctness in detecting known attacks and their variations. Their noticeable drawback is the inability to detect attacks whose instances have not yet been observed.

Anomaly detection approaches, on the other hand, construct models of normal data and detect departure from the normal model in observed data. Anomaly exposure functional to intrusion detection and computer security has been an active area of research since it was originally proposed by Denning [9]. Anomaly detection algorithms have the benefit that they can detect new category of intrusions as deviations from normal usage [9, 10]. In this problem, given a set of normal data to train from, and given a new piece of test data, the objective of the intrusion detection algorithm is to decide whether the test data belong to "normal" or to an anomalous behavior. However, anomaly detection method suffers from a high rate of false alarms. This occurs primarily because before unseen (yet legitimate) system behaviors are also recognized as anomalies, and hence bunting as potential intrusions. This paper focuses on a full comparative study of several anomaly detection schemes for identifying different network intrusions.

2. RELATED WORK:

Network intrusion exposure systems such as SNORT[11] and Bro[12] use hand written rules to identify signatures of known attacks, such as a exact string in the application payload, or doubtful behavior, such as server requests to unused ports. Anomaly detection systems such as SPADE, [13] ADAM, and NIDES learn a statistical model of normal network traffic, and flag divergence from this model. Models are frequently based on the sharing of source and destination addresses and ports per transaction (TCP connections, and sometimes UDP and ICMP packets).

For example, SPADE offers four probability models (estimated by average frequency) of incoming TCP connections:

- P(destination-address, destination-port)
- P(source-address, destination-address, destination-port)
- P(source-address, source-port, destination-address, destination-port)
- Bayes network approximation of the above.

Lower probabilities consequence in higher anomaly scores, since these are most probably more likely to be hostile. ADAM is a classifier which can be trained on both known attacks and on (presumably) attack-free traffic. Pattern which does not equal any cultured sort are flagged as anomalous. ADAM also models address subnets (prefixes) in calculation to ports and personality address. NIDES, like SPADE and ADAM, models ports and addresses, flagging dissimilarity between short and long term behavior.

SPADE, ADAM, and NIDES use frequency-based models, in which the prospect of an event is predictable by its average frequency during training. PHAD [14], ALAD [15], and LERAD [16] use time-based models, in which the prospect of an event depends as a substitute on the time since it last occur. For each quality, they gather a set of allowed values (anything observed at least once in training), and flag novel values as anomalous.

purposely, they assign a score of a new valued attribute, where t is the time since the quality was last anomalous (during either training or testing), n is the numeral of teaching explanation, and r is the size of the set of allowable values. Note that r/n is the average rate of anomalies in training; thus characteristic with high n/r are not likely to produce anomalies in difficult and ought to score high. The factor t create the model time-dependent, acquiescent higher scores for aspect for which there have been no anomalies for a long time.

In recent years, a enormous amount of work has been done in the network anomaly exposure. Machine learning approaches have been extensively used on identify network anomalies freshly such as the n adjacent neighbor methods [17], also the Neural Network [18], support vector machines [19]. Some other techniques like the genetic computation [20], Bayesian networks [21], outlier detection [22], Y - means clustering algorithm [23], Probability Statistics have been adapted in the anomaly detection and analysis work. However the machining learning method cannot be established secure. Also most of these approaches should analysis huge amount of source data. These techniques also have difficulty in choose the extent of parameter standard, the lack of suppleness and high rate of false alarm, etc. Other methods are bases on the network anomaly signature. Different applications, new protocols and new type of networks have made the network changes greatly every day. Also with the expansion of wireless network and ADHOC net work, the signature based approach cannot be a good solution for the network anomaly. The statistical analysis assumes to be a best approach.

A lot of statistical technique has been adapted in the network traffic analysis and anomaly detection. Because the network traffic of different protocol has dissimilar characteristic and distinctive statistical distribution. So the approach to model the combined traffic with one distribution process will not work well except for special application. Some research has proved the network traffic has the pattern of self- similar. A lot of work has been carried out to analysis the network traffic with self similar process.

2.1 Gaussian Mixture Model

Analyzing the source data, the network traffic cannot be described as a Gaussian distribution. The distribution of a Gaussian should be like the shape

of ellipse and its residual should be normal. The Gaussian mixture model probability density function is a weighted average of several Gaussian distribution. Here it is take n the Gaussian mixture model with three single Gaussian distribution as an example.

$p(x) = \alpha 1\, g(x ; \mu 1, \theta 1) + \alpha 2\, g(x ; \mu 2, \theta 2) + \alpha 3\, g(x ; \mu 3, \theta 3)$

The parameter list (α1, α2, α3) must satisfy the following condition:

$\alpha 1 + \alpha 2 + \alpha 3 = 1$

The single Gaussian mixture distribution can be represented as:

$g(x ; \mu, \sigma 2) = (2\pi)-d/2\ \sigma-d$

$\exp[-((x - \mu)T(x - \mu))/2\ \sigma 2\ ]$

The more Gaussian models, the more precise the Gaussian mixture model will be. In the approach discussed here it finds the amount of Gaussian distribution will influence the time
cost and performance of approach.

2.2 EM Algorithm:

EM is an iterative method for estimating the value of some unidentified quantity, given the values of some correlated, identified quantity. The method is to first consider that the quantity is represented as a value in some parameterized probability distribution. The EM procedure is studied below: Initialize the distribution parameters Repeat until convergence:

E- Step: approximate the Expected value of the unknown variables, given the current parameter estimate

$Q(h\ `|h) = E[\ln p(Y|h`)|h, X]$

M- Step: re- estimate the distribution parameters to maximize the similarity of the data, given the expected estimates of the unknown variables

$h \leftarrow \arg\max Q(h\ `|h)$

At here, the EM algorism is used to estimate the mean value of different Gaussian distribution which overlaps with each other to form the Gaussian mixture distribution.

2.3 Time Slice Window:

The method of the mixture of Gaussian model is considered to match the network traffic distribution. Then the EM algorism is used to estimate the mean value of each Gaussian distribution. Considering the data of time series, the data should be partitioned with time slot. It is called the"window". The size of the windows should be decided. From the input data the network traffic illustrate the circular fluctuation of date and night. So the time slot window should be the integral times of the 24 hours. I n input data 1440 am the circular length. The calculation time delay can be adjusted. The time cost change s greatly with the time delay. Here consider the value is 100.

The calculation window consequence should be:

windown: $tn \rightarrow tn + 1440$

windown+1

: $tn + 100 \rightarrow tn + 100 + 1440$

2.4 The K and D Indicators Approach:

The index denotes the relationship between highest value, lowest value of recent days and the value of the last day. This index can reflect the sudden increase or decrease of the network traffic.

The calculation approach is listed below:

$k(n) = 100 * [(C(n) - L5)/(H5 - L5)]$

$D(n) = 100 * (H3/L3)$

In the formulation the C(n) is the value of time stamp n; L5 is the lowest value in the most recent 5 times. H5 is the highest value in the most recent 5 times. H3 is the sum of (C- L5) in three times. L3 is the sum of (H5 - L5) in three data points.

The K line is more susceptible to the change of the new coming data than the D line. So if the K line passes through the D line, a fluctuation of network traffic is specified. So an alarm will be triggered. At other end, the next cross would be the signal of normal which means the anomaly has passed away.

2.5 The Up and Low Bound Method:

After calculating the mean value μj , all of them are added into one and check whether the value varies greatly. If so it is considered that there may be some traffic anomaly in the network traffic.

$Z\,up(t) = x(t)` + k *$

$r(t)z\,down(t) = x(t)` - k *$

$r(t)\ x(t)`$ is the mean value of mean value μj in the latest m
samples;

$x(t)` = (x(t) + x(t-1) + x(t-2) + \cdots + x(t - m + 1)) / m$

r(t) is the standard deviation of mean value μj sum in the latest m samples;

$Ai = (x(t - i) - x(t)`)$
2

k is a weighting factor of fluctuation.

The zup(t) represents the upper limit of mean value μj sum according to its tendency. The zdown (t) represents the down limit of mean value μj sum according to its tendency. If the value crosses the line a alert will be submitted. The k would be a configurable parameter which associated with the fluctuation range of the normal network traffic behavior.

3. APPLICATIONS OF ANOMALY DETECTION

In this section we discuss several applications of anomaly detection. For each application domain we discuss the following four aspects:

-The notion of anomaly.
-Nature of the data.
-Challenges associated with detecting anomalies.
-Existing anomaly detection techniques.

4. NETWORK TRAFFIC MODELS

Conventionally, network traffic has been modeled as a Poisson process. Indeed, the Poisson model has been successfully used in telephone networks for many years, and so it was inherited when telecommunication networks became digital and started to send information as data packets. Also, this model has a simple mathematical expression [23], and has only one parameter, λ, which is in turn very intuitive (the mean traffic in packets per time unit).

Several authors have proposed network traffic behavior and presented other models that overcome the limitations which are inherent to Poisson processes, the most notable ones probably being that the Poisson model has a fixed relationship between mean and variance values (both are equal to λ), and that it does not account for high

variability or long- range dependence Some proposed models are usually based on the assumption that network traffic is self - similar in nature, as originally stated .

At this point, it should be clear that any model for instantaneous traffic marginal's must be flexible enough to adapt to some properties observed in traffic, namely:

1. Let C(t) be the amount of traffic accumulated at time t. Then, C(t) <=C(t+1) and C(t+1)-C(t)<=M, where M is the network maximum transmission rate.
2. The fact that at time t there is a certain amount of traffic C (t) does not imply in any way that at time t +1 the amount of traffic lies anywhere near C(t). This is equivalent to say

Network traffic exhibits the high variability property.

The latter property is also identified as the "Noah effect" or the infinite variance syndrome.

At the other side, the first aforementioned property states the clear fact that network traffic has compact support between 0 and the M. Compact support creates symmetric distributions (Gaussian distributions are symmetric) inappropriate. Accordingly, if traffic data concentrate near the maximum transmission rate, a symmetric model would allow traffic increments.

5. CHALLENGES:

-Defining a normal region which encompasses every possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous observation which lies close to the boundary can actually be normal, and vice-versa.

-|When anomalies are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear like normal, thereby making the task of defining normal behavior more difficult.

-In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future.

-The exact notion of an anomaly is different for different application domains. For example, in the medical domain a small deviation from normal (e.g., fluctuations in body temperature) might be an anomaly, while similar deviation in the stock market domain (e.g., fluctuations in the value of a stock) might be considered as normal. Thus applying a technique developed in one domain to another is not straightforward.

-Availability of labeled data for training/validation of models used by anomaly detection techniques are usually a major issue.

-Often the data contains noise which tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

6. CONCLUSIONS:

This paper has presented idea about the statistical anomaly identification of network traffic. Here paper studied a statistical approach to analysis the distribution of network traffic to recognize the normal network traffic behavior. The EM algorism is discussed to approximate the distribution parameter of Gaussian mixture distribution model. Another time series analysis method is studied. This paper also discussed a method to recognize anomalies in network traffic.

REFERENCES:

[1]Chandola, V., Boriah, S., and Kumar, V. 2008. Understanding categorical similarity measures for outlier detection. Tech. Rep. 08-008, University of Minnesota. Mar

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar Anomaly Detection: A Survey August 15, 2007

[3] Aleskerov, E., Freisleben, B., and Rao, B. 1997. Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of IEEE Computational Intelli-gence for Financial Engineering. 220{226}

[4] Fujimaki, R., Yairi, T., and Machida, K. 2005. An approach to spacecraft anomaly detection problem using kernel feature space. In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY, USA, 401{410.

[5] Edgeworth, F. Y. 1887. On discordant observations. Philosophical Magazine 23, 5, 364{375}

[6] 1. Successful Real-Time Security Monitoring, Riptech Inc. white paper, September 2001

[7] 2. W. Lee, S. J. Stolfo, Data Mining Approaches for Intrusion Detection, Proceedings of the 1998 USENIX Security Symposium, 1998

[8] 4. J. Luo, Integrating Fuzzy Logic With Data Mining Methods for Intrusion Detection, Master's thesis, De-partment of Computer Science, Mississippi State University, 1999.

[9] 7. D.E. Denning, An Intrusion Detection Model, IEEE Transactions on Software Engineering, SE-13:222-232, 1987

[10] 8. H.S. Javitz, and A. Valdes, The NIDES Statistical Component: Description and Justification, Technical Report, Computer Science Laboratory, SRI Interna-tional, 1993

[11] Roesch, Martin, "Snort - Lightweight Intrusion Detection for Networks", Proc. USENIX Lisa '99, Seattle: Nov. 7-12, 1999.

[12] Paxson, Vern, "Bro: A System for Detecting Network Intruders in Real-Time", Lawrence Berkeley National Laboratory Proceedings, 7'th USENIX Security Symposium,Jan. 26-29, 1998, San Antonio TX,

[13]SPADE, Silicon Defense, http://www.silicondefense.com/software/spice/

[14] Mahoney, M., P. K. Chan, "PHAD: Packet Header Anomaly Detection for Identifying Hostile Network Traffic", Florida Tech. technical report 2001-04, http://cs.fit.edu/~tr/

[15] Mahoney, M., P. K. Chan, "Learning Nonstationary Models of Normal Network Traffic for Detecting Novel Attacks ", Edmonton, Alberta: Proc. SIGKDD, 2002, 376-385

[16] Mahoney, M., P. K. Chan, "Learning Models of Network Traffic for Detecting Novel Attacks", Florida Tech. technical report 2002-08, http://cs.fit.edu/~tr/ Mahoney, M., P. K. Chan, "Learnin

[17] E. Eskin,A. Arnold,M. Prerau,L. Portnoy,and S. Stolfo, A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. K luwer, 2002.

[18] Manikopoulos C, Papavassiliou S, ANetwork intrusion and fault detection: A statistical anomaly approach . IEEE Communications Magazine, 2002, 40 (10):7682.
[19] S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, Modeling intrusion detection system using hybrid intelligent systems. Journal of Network and Computer Applications, 30(1):114-132, January 2007
[20] W. Lu and I. Traore, Detecting new forms of network intrusions using genetic programming. Computationa Intelligence,20(3):475- 494, Aug.2004.
[21] D. Barbara, N. Wu, and S. Jajodia, Detecting novel net work intrusions using bayes estimators. In Proceedings of the First SIAM International Conference on Data Mining (SDM 2001), Chicago, USA,April 2001.
[22] W. Lu and I. Traore, A novel unsupervised anomaly detection framework for detecting network attacks in real - time. In4th International Conference on Cryptology and Network Security (CANS), Xiamen, Fujian Province, China, December2005.
[23] A. Papoulis, Probability, Random Variables, and Stochastic Processes, third ed., McGraw- Hill, 1991.