ANOMALY DETECTION BASED ON DIVERSE APPROACHES IN NETWORK TRAFFIC

Anup Bhange*

Amber Syad**

Satyendra Singh Thakur***

ABSTRACT

Network anomaly detection is a vibrant research area. Researchers have to come nearer this problem using various methods. Earliest work in traffic analysis has revealed that modem network produces traffic stream that are self similar over several time scales from microseconds to minutes. Anomaly detection focuses at getting the existence of anomalous patterns in network traffic. Automatic detection of such patterns can give network administrators with an additional source of information to analyze network behavior or finding the root cause of network faults. Network always undergoes from the traffic anomaly such as router rate change, device restart or the worm attack. The early detection of unusual anomaly in the network is a main factor to fast recover and avoidance of future serious problem to offer a stable network transmission. This paper discusses 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 discusses a method to recognize anomalies in network traffic, based on a nonrestricted α-stable first-order model and statistical hypothesis testing.

Keywords*: Statistical approach, α-Stable Model, anomaly detection, EM algorithm.*

*Department of CSE, Bhopal, Patel Institute of Technology.

**Assistant Professor, Department of CSE, Bhopal, Patel Institute of Technology.

***Assistant Professor, Department of CSE, Bhopal, Patel college of Science Technology.

I. INTRODUCTION

Early Network anomaly detection has become serious to service provider in their pledge to sustain **a** given Service level agreement. Hence, there has been much interest in studying adaptive event detection Schemes with application to communication networks.

The network has been extensively used in every area. Different applications, new protocols and new type of networks have made the network changes greatly every day. Also there are more and more kinds of viruses and network attacks in the nowadays network. The statistical approach would be a better solution than the signature based approach to satisfy the needs. However, traffic of unusual kinds of protocol has different kind of statistical distribution. Paper discusses the Gaussian Mixture Model to estimate the combined statistical model. The Gaussian Mixture Model can be a model with less residual in the network distribution of combined traffic of unusual type. The paper uses the EM algorism to approximate the value of different Gaussian distribution.

Anomaly identification focuses at finding the presence of anomalous patterns in network traffic. Automatic identification of such patterns can give network administrators with an additional source of information to detect network behavior or getting the basic cause of network faults. Research appeals in anomaly recognition typically pursue a four-stage method, in which the first three stages define the recognition approach, while the last stage is dedicated to authenticate the method. So, in the first stage, traffic data are collected from the network (data acquisition). Second, data are analyzed to take out its most relevant features (data analysis). Third, traffic is classified as normal1 or abnormal (inference); and fourth, the whole method is validated with various types of traffic anomalies (validation). Following the aforesaid four-stage method, it can be mention that data acquisition is typically carried out by electing one or more routers from time to time, so that traffic data are collected and stored for subsequent analysis in the second stage. Some researchers sample data at the packet level, gathering information from headers, latencies, etc., while others favor to use aggregated traffic as the source of information. Sampling data at the packet level supplies more information, but at the cost of a higher computational load and dedicated hardware must be utilized. Aggregated traffic, at the other end, provides less information from which to choose for the presence or absence of anomalies, but is a simpler method and does not require any special hardware.

In the data analysis stage, various techniques can be applied to take out interesting features from current traffic. Some of them include wavelets, statistics-based measurements, and

statistical models. Out of these techniques, the use of statistical models as a means to take out important features for data analysis has been found to be very interesting, since they permit for a robust analysis even with small sample sizes. Moreover, with a traffic model, its set of parameters can be used as extracted traffic features, since any traffic sample is determined by the model parameters.

Nevertheless, anomaly detection is still often based (at least partially) on classical models, such as Gamma distributions. The fact that these models do not account for high variability may have a negative impact on capturing traffic features and, as a consequence, on recognizing anomalies. Several methods have been used in the inference stage as well. Classification methods based on neural networks, statistical tests, information theory, to quote a few, can be found in anomaly detection literature. There appears to be a common point in all of them, though. The inference stage takes its decisions on the existence of a reference traffic window, which permits the classification way to assess whether the current traffic window is normal or abnormal. How the reference window is selected not only has an impact on the final normal versus abnormal classification rate, but it also decides the exact definition of a traffic anomaly. Some approaches assume that an anomaly is an abrupt change in some of the features extracted from traffic, so the reference window is simply the previous-to-current traffic window. In the validation stage, researchers give quality measures about the recognition capability of their technique according to a decided criterion, which is typically the detection rate in terms of false positives and false negatives (i.e., the fraction of normal traffic patterns incorrectly classified as anomalous and the fraction of anomalous traffic patterns incorrectly classified as normal, respectively), although some others favor other quality measures. This paper discusses an anomaly recognition approach based on α-stable distributions which does not require network administrators decide reference traffic windows and it is capable to detect flood and flash-crowd anomalies regardless of the presence or absence of unexpected changes in network traffic.

II. RELATED WORKS

In recent years, a huge amount of work has been done in the network anomaly detection. Machine learning approaches have been widely used on detecting network anomalies recently such as the n nearest-neighbor methods [1], also the Neural Network [2], support vector machines [3]. Some other techniques like the genetic computation [4], Bayesian networks [5], outlier detection [6], Y-means clustering algorithm [7], Probability Statistics [8-10] have been adapted in the anomaly detection and analysis work. However the machining learning

method cannot be proven secure [11]. Also most of these approaches should analysis huge amount of source data. These techniques also have difficulty in deciding the scope of parameter standard, the lack of flexibility and high rate of false alarm, etc. Other methods are bases on the network anomaly signature [12]. 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 network, the signature based approach cannot be a good solution for the network anomaly [13]. 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.

III. STATISTICAL APPROACH

As the network traffic of different protocol has different feature and distinctive statistical distribution. The distribution of the source data is discussed and observed that the residual of the statistic result is not normal. The statistical data with Gaussian distribution should be with the normal residual.

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 taken 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 ($\alpha 1, \alpha 2, \alpha 3$) 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. 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 overlap with each other to form the Gaussian mixture distribution.

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. In input data 1440 is the circular length. The calculation time delay can be adjusted. The time cost changes greatly with the time delay. Here consider the value is 100.

The calculation window consequence should be:

$window_n : tn \rightarrow tn + 1440$

$window_{n+1} : tn + 100 \rightarrow tn + 100 + 1440$

4. Iterative Algorism

Through the application of EM, the mean value of these Gaussian distribution.can be obtained.

Step 1:

Calculate the $E[z_{ij}]$ for each hidden variablez_{ij}. Assume the current $h =< \mu 1, \mu 2 \cdots \mu j >$.

$E[z_{ij}] = p(x = xi \mid \mu = \mu j)/(\Sigma^{2}_{n=1}\ p(x = xi \mid \mu = \mu n))$

Step 2:

The maximum similarity approach is considered to calculate the h` =< μ`, μ`, · · · μ` >, the $E[z_ij]$ is taken as an estimation of z_ij. then replace the h =< μ1, μ2 · · · μj > with h` =< μ`, μ`, · · · μ` >.

$\mu j \leftarrow (\Sigma^{m}_{i=1}\ E[z_{ij}\ ]\ xi) / (\Sigma^{m}_{i=1}\ E[z_{ij}\ ])$

To repeat these two steps, estimated mean value μj of distribution j can be easily achieved

IV. DISCUSSION ON TIME SERIES ANALYSIS

1. 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.

2. 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.

V. 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 [35], 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 in [36].

At this point, it should be clear that any model for instantaneous traffic marginals 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 [24].

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 to be

larger than physically possible, again, with a non-negligible probability. This also influences the Gamma distribution.

1. α-Stable Model

α-stable distributions can be considered as a superset of Gaussian functions and originate as the solution to the Central Limit Theorem when second order moments do not present [17], that is, when data can suddenly vary by large amounts as time passes by. This fits nicely to the high variability property seen in network traffic. Moreover, α-stable distributions have an asymmetry parameter which allows their PDF to change from totally left-asymmetric to totally right-asymmetric, while genuine Gaussian distributions are always symmetric. This factor makes α-stable distributions naturally adaptable to the first traffic property (compact support) even when average traffic is virtually 0 or very near the maximum theoretical network throughput. In addition, α-stable distributions give an explanation to the restriction imposed in [36] about the requirement to aggregate many traffic traces for them to converge to a Gaussian distribution. According to the Generalized Central Limit Theorem [27], which includes the infinite variance case, the sum of n α-stable distributions is another α-stable distribution, although not necessarily Gaussian.

VI. 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, based on a nonrestricted α-stable model and statistical hypothesis testing.

REFERENCES

1. E. Eskin,A. Arnold,M. Prerau,L. Portnoy,and S. Stolfo, A geometric framework for unsupervised anomaly detection: Detectingintrusions in unlabeled data. Kluwer, 2002.
2. Manikopoulos C, Papavassiliou S, ANetwork intrusion and fault detection: A statistical anomaly approach. IEEE Communications Magazine, 2002, 40 (10):7682.
3. 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.

4. W. Lu and I. Traore, Detecting new forms of network intrusions using genetic programming. Computational Intelligence,20(3):475-494, Aug.2004.
5. 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.
6. 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.
7. Y. Guan, A. A. Ghorbani, and N. Belacel, An unsuper vised clustering algorithm for intrusion detection. In Proc. of the Sixteenth Canadian Conference on Artificial Intel ligence (AI 2003), pages 616-617, Halifax, Canada, May 2003. Springer.
8. Staniford S, Hoagland JA, McAlerney JM, Practical automated detection of stealthy portscans. Journal of Computer Security, 2002, 10 (1/2):105136.
9. Mahoney VM, A machine learning approach to detecting attacks by identifying anomalies in network traffic. Melbourne: Florida Institute of Technology, 2003.
10. Wang K, Stolfo SJ, Anomalous payload-based network intrusion detection. In: Jonsson E, Valdes A, Almgren M, eds. Proc. of the 7th Int'l Symp. On Recent Advances in Intrusion Detection (RAID 2004). LNCS 3224, Heidelberg: Springer- Verlag, 2004. 203222.
11. M. Barreno, B. Nelson, R. Sears, A. D.Joseph, and J. D.Tygar, Can machine learning be secure? In ASIACCS '06: Proceedings of the 2006 ACM Symposium on Information, computer and communications security, pages 16-25, New York,NY, USA, 2006. ACMPress.
12. F. Feather, D. Siewiorek, R. Maxion, Fault detection in an Ethernet network using anomaly signature matching. ACM SIGCOMM Computer Communication Review, 1993.
13. Ilker Onat, Ali Miri, AReal-TimeNode-Based Traffic Anomaly Detection Algorithm for Wireless Sensor Networks. Proceedings of the 2005 Systems Communications (ICW'05) 2005.
14. A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, and P. Abry, "Non-Gaussian and Long Memory Statistical Characterizations for Internet Traffic with Anomalies," IEEE Trans. Dependable and Secure Computing, vol. 4, no. 1, pp. 56-70, Jan. 2007.
15. M. Thottan and C. Ji, "Anomaly Detection in IP Networks," IEEE Trans. Signal Processing, vol. 51, no. 8, pp. 2191-2204, Aug. 2003.

16. Y. Gu, A. McCallum, and D. Towsley, "Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation," Proc. Internet Measurement Conf., Oct. 2005.
17. A. Lakhina, M. Crovella, and C. Diot, "Diagnosing Network-Wide Traffic Anomalies," Proc. ACM SIGCOMM '04, pp. 219-230, Aug. 2005.
18. A. Ray, "Symbolic Dynamic Analysis of Complex Systems for Anomaly Detection," Signal Processing, vol. 84, no. 7, pp. 1115-1130, 2004.
19. S.C. Chin, A. Ray, and V. Rajagopalan, "Symbolic Time Series Analysis for Anomaly Detection: A Comparative Evaluation," Signal Processing, vol. 85, no. 9, pp. 1859-1868, 2005.
20. A. Wagner and B. Plattner, "Entropy Based Worm and Anomaly Detection in Fast IP Networks," Proc. 14th IEEE Int'l Workshops Enabling Technologies: Infrastructures for Collaborative Enterprises, pp. 172-177, June 2005.
21. M. Ramadas, S. Ostermann, and B. Tjaden, "Detecting Anomalous Network Traffic with Self-Organizing Maps," Proc. Sixth Int'l Symp. Recent Advances in Intrusion Detection, pp. 36-54, 2003.
22. S.T. Sarasamma, Q.A. Zhu, and J. Huff, "Hierarchical Kohonen Net for Anomaly Detection in Network Security," IEEE Trans. Systems, Man and Cybernetics, Part B: Cybernetics, vol. 35, no. 2, pp. 302-312, Apr. 2005.
23. L. Kleinrock, Queueing Systems, Volume 2: Computer Applications. John Wiley and Sons, 1976.
24. W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson, "Self-Similarity through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," IEEE/ACM Trans. Networking, vol. 5, no. 1, pp. 71-86, Feb. 1997.
25. G. Samorodnitsky and M.S. Taqqu, Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance. Chapman & Hall, 1994.
26. F. Simmross-Wattenberg, A. Trista´n-Vega, P. Casaseca-de-la Higuera, J.I. Asensio-Pe´rez, M. Martı´n-Ferna´ndez, Y.A. Dimitriadis, and C. Alberola-Lo´pez, "Modelling Network Traffic as α -Stable Stochastic Processes: An Approach Towards Anomaly Detection," Proc. VII Jornadas de Ingenierı´a Telema´tica (JITEL), pp. 25-32, Sept. 2008.
27. G.R. Arce, Nonlinear Signal Processing: A Statistical Approach. John Wiley and Sons, 2005.

28. J. Jiang and S. Papavassiliou, "Detecting Network Attacks in the Internet via Statistical Network Traffic Normality Prediction," J. Network and Systems Management, vol. 12, no. 1, pp. 51-72, Mar. 2004.
29. W. Yan, E. Hou, and N. Ansari, "Anomaly Detection and Traffic Shaping under Self-Similar Aggregated Traffic in Optical Switched Networks," Proc. Int'l Conf. Comm. Technology (ICCT '03), vol. 1, pp. 378-381, Apr. 2003.
30. Internet Traffic Archive, http://ita.ee.lbl.gov/, 2011.
31. Waikato Internet Traffic Storage, http://wand.cs.waikato.ac.nz/wits/, 2011.
32. Cooperative Assoc. for Internet Data Analysis, http://www.caida.org/, 2011.
33. B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, "Sketch-Based Change Detection: Methods, Evaluation, and Applications," Proc. Internet Measurement Conf. (IMC), pp. 234-247, Oct. 2003.
34. G. Cormode and S. Muthukrishnan, "What's New: Finding Significant Differences in Network Data Streams," IEEE/ACM Trans. Networking, vol. 13, no. 6, pp. 1219-1232, Dec. 2005.
35. A. Papoulis, Probability, Random Variables, and Stochastic Processes, third ed., McGraw-Hill, 1991.
36. W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," IEEE/ACM Trans. Networking, vol. 2, no. 1, pp. 1-15, Feb. 1994.
37. J.R. Gallardo, D. Makrakis, and L. Orozco-Barbosa, "Use of α-Stable Self-Similar Stochastic Processes for Modelling Traffic in Broadband Networks," Performance Evaluation, vol. 40, pp. 71-98, 2000.
38. "Apache JMeter," The Apache Jakarta Project, Apache Software Foundation, http://jakarta.apache.org/jmeter/, 2011.
39. A. Stavrou, G.F. Cretu-Ciocarlie, M.E. Locasto, and S.J. Stolfo, "Keep Your Friends Close: The Necessity for Updating an Anomaly Sensor with Legitimate Environment Changes," Proc. ACM/CSS Workshop Security and Artificial Intelligence (AISec), 2009.
40. G.F. Cretu-Ciocarlie, A. Stavrou, M.E. Locasto, and S.J. Stolfo, "Adaptive Anomaly Detection via Self-Calibration and Dynamic Updating," Proc. 12th Int'l Symp. Recent Advances in Intrusion Detection (RAID), Sept. 2009.

41. G. Macia´-Ferna´ndez, J. Dı´az-Verdejo, and P. Garcı´a-Teodoro, "Evaluation of a Low-Rate DoS Attack against Application Servers," Computers and Security, vol. 27, pp. 335-354, 2008.
42. Huang Kai, Qi Zhengwei, Liu Bo Network Anomaly Detection Based on Statistical Approach and Time Series Analysis "2009 International Conference on Advanced Information Networking and Applications Workshops"
43. Federico Simmross, Juan Ignacio, Pablo Casaseca-de-la-Higuera, Ioannis A. Dimitriadis" Anomaly Detection in Network Traffic Based on Statistical Inference and α-Stable Modeling" IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 8, NO. 4, JULY/AUGUST 2011
44. Ciro D' Apice, Olga GalaktionovaModeling of Network Traffic"2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)"