International Journal of Security and Its Applications
Vol.8, No.2 (2014), pp.87-94
http://dx.doi.org/10.14257/ijsia.2014.8.2.09

Detecting an Anomalous Traffic Attack Area based on Entropy Distribution and Mahalanobis Distance

Dolgormaa Bayarjargal and Gihwan Cho[1]

Div. of Computer Science and Engineering, Chonbuk Univ., Jeonju, S. Korea
{maatopaz, ghcho}@jbnu.ac.kr

Abstract

This paper deals with detecting an anomalous traffic area based on its distribution and distance measurements. On detecting anomalous traffic such as flooding attack traffic we should consider the packet attribute of the traffic. In order to identify that traffic, we compute entropy of selected packet attribute and Mahalanobis distance between normal and abnormal traffics. Chi-square test is used to evaluate the proposed method. Detection accuracy and performance are analyzed using real network traffic trace which consists the mostly backscatter of SYN flooding attack from LANDER project. The result of our proposed method indicates that it can show and offer significantly an accurate result.

Keywords: *Anomalous traffic, Flooding attack, Entropy, Mahalanobis distance*

1. Introduction

Anomalous traffic is defined as a pattern in the data that does not conform to the expected behavior [1]. Many anomalous traffic detection information entropy [6-8, 10] and distance measurement based [3, 5, 9] techniques have been proposed. When anomalies are result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear normal [1]. Hence detecting and defining a malicious region is still challenging. Due to the need to differentiate between normal and abnormal behavior, we designed an exact distribution and distance measurement method.

Our approach has three key features. First, our method for anomaly detection compute the entropy of selected packet attributes to define the attack area. Entropy is a measure describing the randomness or diversity of a selected set of features of traffic. Second, after defining suspicious traffic area, we calculate the Mahalanobis distance between defined attack area and normal area data based on entropy value, Mahalanobis distance considers correlation between the variables. If the estimated distance value is much bigger than defined threshold value, then it detected as an outlier. The threshold value is set to the most distant points from the mean of normal data. Third we employ chi-square test to ensure the robustness of proposed method.

On detecting anomalous traffic such as flooding attack traffic, we should consider the packet attribute of the traffic. To provide insight in to the detection accuracy and performance, we used 3 hour-long traffic trace dataset "DoS_80-20110715", which is provided by the LANDER project [11]. It was collected within a USC campus on the situation of attack or most likely backscatter because there is a web server running at the source address. With considering the traffic data features, we have focused the entropy of TCP flag and backscatter packet rate. Then we calculate the Mahalanobis distance between normal and abnormal traffics, and finally Chi-square test is used to evaluate the proposed method.

[1] Corresponding author

ISSN: 1738-9976 IJSIA

2. Related Work

Traditional anomaly detection methods based on pattern matching techniques cannot cope with the need for faster speed to manually update those patterns [2]. Most of previous entropy works considered IP size entropy [7, 8, 10]. Du and Abe [8] propose the IP packet size entropy which can detect DoS attacks beyond the volume based schemes' ability to detect. Nychis *et. al.*, [7] gives an understanding of flow-header features and behavioral distribution entropy. According to their study, entropies of source address and port distributions are strongly correlated with each other and provide very similar detection capabilities.

Thatte *et. al.*, [10] assumes attack traffic has some special IP packet size distribution. Because flooding attack has identical packet size. Such as, SYN flooding attack traffic consists of SYN packets with 40 bytes and an ICMP flooding attack consists with ICMP packets with 1500 bytes. Their proposed technique uses both packet rate and packet size statistics. Bellaiche and Gregoire [6] proposed the method which measures the balance of TCP handshake for detecting SYN flood attack. In other words, they focus on unusual TCP handshakes.

Mahalanobis distance based methods considered [3, 5, 9]. Wang *et. al.*, [3] proposed a payload based anomaly detector and they used Mahalanobis distance during the detection phase to calculate the new data against the pre-computed profile. Santiago-Paz *et. al.*, [5] proposed Entropy-Mahalanobis-based methodology to detect certain anomalies in IP traffic. Different from our proposed method, they focused source and destination IP address of entropy. And they used Mahalanobis distance to determine whether a given actual traffic slot is normal or abnormal.

Lazarevic *et. al.*, [9] studied the several anomaly detection schemes for identifying different network intrusions. According to their research their result Mahalanobis distance distribution method is effective in case of several normal behaviors are separately trained. In this paper our focus is on distinguishing normal and suspicious areas for anomaly detection and how it is distributed far from each other.

3. Traffic Anomaly Detection

In this section, we intend to propose an accurate detection of flooding attack area based on entropy distribution and Mahalanobis distance measurements. Finally Chi-square goodness fit test provides to ensure the robustness of proposed method.

3.1. Entropy

Shannon entropy or information entropy is a measure of the uncertainty associated with a random variable. So, in network traffic it used to calculate the distribution randomness of some attributes in the network packet.

$$H(t) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (1)$$

In our case p_i is the probability of each TCP flags to the total number of TCP packets in during defined interval time. We set the time interval to 1 minute. Here in equation (1), log_2 is the expected information value of TCP packets.

3.2. Mahalanobis Distance

Mahalanobis distance can be thought of as a metric for estimating how far each case is from the center of all the variables' distributions. This method used to identify and estimate the similarity of "unknown traffic" corresponds to "normal behavior". It takes into account not only the average value but also its variance and covariance of the variables measured. Instead of simply computing the distance from the mean values, it weights each variable by its standard deviation and covariance, so the computed value gives a statistical measure of how well the new example matches the training sample [3]. The Mahalanobis distance between the particular point p and μ is computed as [6]:

$$d_M = \sqrt{(p - \mu)^T \cdot \sum^{-1} \cdot (p - \mu)} \tag{2}$$

Where the p is the sample vector, μ denote the theoretical mean vector and Σ is the covariance matrix of the normal data. The mean and the common covariance matrix are generally unknown, must be estimated from normal data. The mean can be estimated as [5]:

$$\bar{\mu} = \frac{1}{N} \sum_{n=1}^{N} p_n \tag{3}$$

Then, the common covariance matrix Σ is estimated as [5]:

$$\sum = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T \tag{4}$$

In our case we chose the background traffic as "normal" data or "behavior". Mahalanobis distance not only expresses the distance between two points but also shows the features of a particular point. A point has got to at least more than one feature. We chose entropy of TCP flags and SYN/ACK packet rate per second (pkts/sec) as a point features. These features indicate that normal TCP packet distribution when SYN/ACK packet has equally low rate. Using Mahalanobis distance, "suspicious" traffic clearly classified from "normal behavior" when TCP packet distribution and SYN/ACK packet rate have sudden changes than expected.

Thus, it provides a relative measure of "normal" and abnormal traffic behaviors. As mentioned above, attack packet is considered as SYN/ACK packet. We set the threshold of data and all measured data points whose distances are greater than the threshold are detected as outliers.

3.3. Chi-square Test

Chi-square test measures how well a set of observations agree with that predicted by some hypothesized distribution.

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{5}$$

Here, O_i is observed frequency of specified packet attribute
E_i is expected frequency of specified packet attribute

In our case hypotheses can be:

H_0: No anomaly,

H_1: Presence of an anomaly in traffic

The chi- square test is used to analyze a contingency table consisting of rows and columns to determine if the observed cell frequencies differ significantly from the expected frequencies.

3.3.1. Calculating Expected Frequency

To find the expected frequencies, we assume independence of the row and columns. To get the expected frequency corresponding to each cell, we multiply row total and column total and divide by the overall total.

Table 1. Contingency Table of TCP Packets

TCP Packets	SYN/ACK	Other packet	Total
#of packets on t_0	a	b	a+b
#of packets before t_0	c	d	c+d
Total	a+c	b+c	a+b+c+d

Based on Table 1 values, expected frequency can be calculated as follows:

$$E_a = \frac{(a + c)(a + b)}{(a + b + c + d)} \tag{6}$$

3.3.2. Goodness of Fit

Definition: A goodness-of-fit test is inferential procedure used to determine whether a frequency distribution follows a claimed distribution. It is a test of the agreement or conformity the observed frequencies (O_i) and the expected frequencies (E_i) for several classes or categories [4].

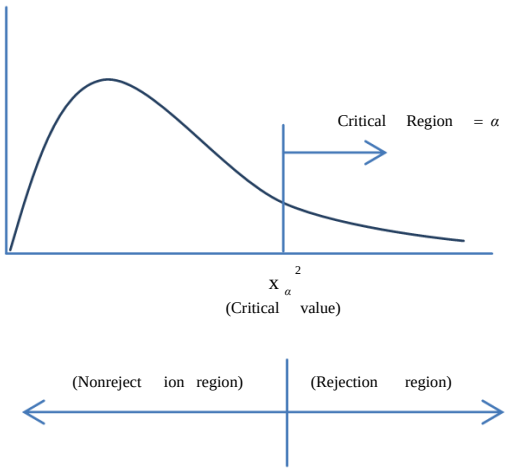

Figure 1. Critical Area of Chi-square

It also expresses fitness between the observed values and the expected values. The value of chi-square calculated by using equation (5) is compared with the critical value in order to test defined hypothesis is rejected or not. The critical value is given in chi-square distribution table. Figure 1 shows the critical area of Chi-square distribution.

4. Experiment and Evaluation

4.1. Entropy Analysis of TCP Packets based on Flags

The experimental dataset [11] contains one attack or possibly a backscatter from a DoS attack. Attack packets consist of many identical SYN/ACK, so we chose TCP flags as an entropy random variable. As we can see the Figure 2, it accurately detects the attack area same as mentioned in [11] the suspicious area is located between 20:44 to 21:02.

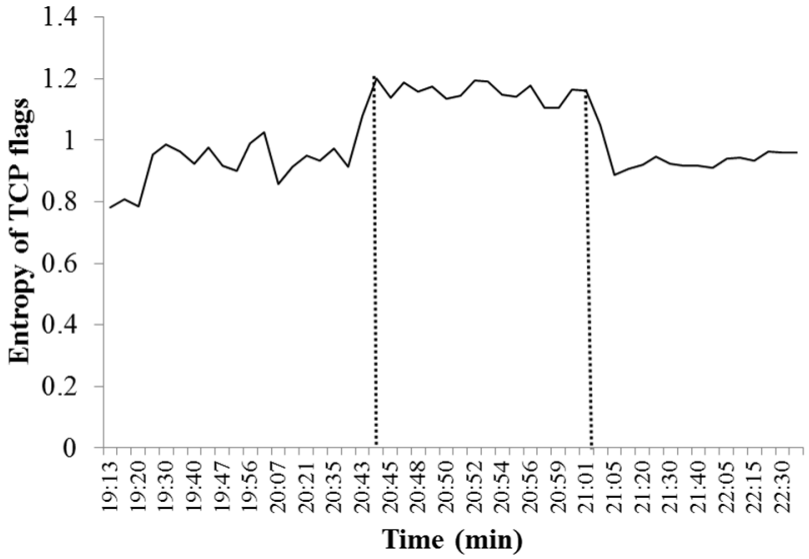

Figure 2. Entropy of TCP Flags

According to the dataset, most of TCP flags were ACK packets. There were around 4 million ACK packets in each interval and it did not change abruptly during the collecting dataset. Compared to ACK packets, SYN/ACK packets were firstly around 30 thousand in normal behavior. When attack happens SYN/ACK packets increased until 400 thousand packets. Thus, this sudden change of entropy is indicated as suitable property to judge the suspicious area. In other words, attack traffic does not manifest as large deviations in traffic volume entropy can accurately detect. When traffic is in normal behavior its entropy values are around 0.8 to 0.9. Then attack happens, entropy value increased until 1.2.

4.2. Entropy based Mahalanobis Distance Anomaly Detection

Mahalanobis distance measurement was used to identify the similarity between the normal area and the suspicious area which shown in Figure 2. Measurement results of Mahalanobis distance are shown in Table 2.

Table 2. Mahalanobis Distance between "Normal Behaviour" and Suspicious Traffic

Time	20:44	20:45	20:46	20:47	20:48	20:49	20:50	20:51	20:52
Distance	60.35	110.1	123.5	116.3	95.96	103.9	116.4	134.4	129.7

Time: (20:44~20:52)

Time	20:53	20:54	20:55	20:56	20:57	20:58	20:59	21:00	21:01
Distance	118.5	118.7	126.2	115.3	96.81	103.9	116.2	111.3	70.03

Time: (20:53~21:01)

According to the Table 2, suspicious points are much far from normal behavior. When traffic is in normal behavior SYN/ACK pkts/sec is between 80,13pkts/sec to 98.8pkts/sec. However, when a attack happens, it suddenly increased to over 1000 pkts/sec. Table 2 is illustrated as in Figure 3.

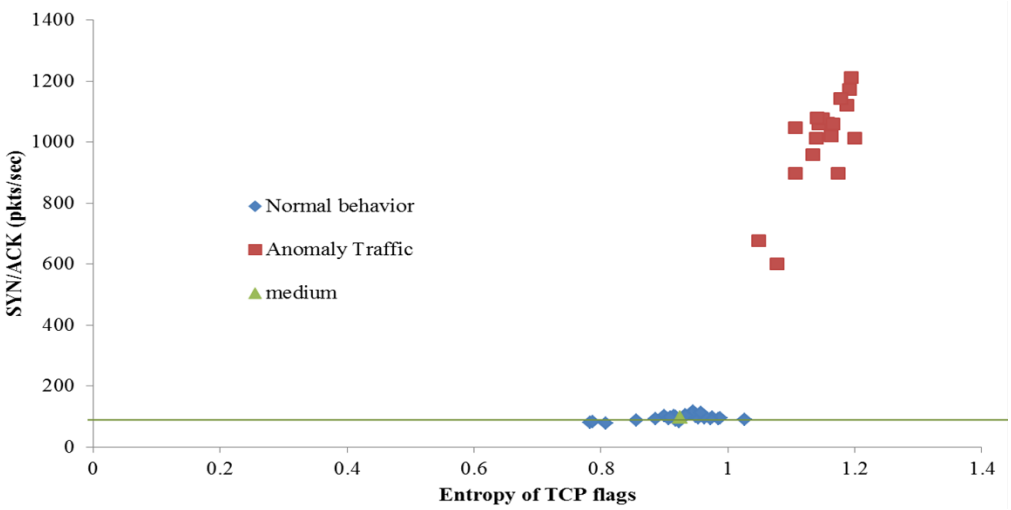

Figure 3. Distribution of Suspicious Traffic (Red Points) and "Normal" Traffic (Blue Points)

As shown in Figure 3 normal traffic data mean μ is (0.924, 96.08). It illustrated as green line. It means normal behavior of SYN/ACK packet rate per second should be an average of 96.08 and entropy of TCP flag should be 0.924. In other words, it expresses normal distribution ratio of TCP flags and incoming rate of SYN/ACK packet per second. We also set Mahalanobis distance threshold value as (0.14, 20) based on normal behavior.

4.3. Chi-square Goodness of Fit Test

Squared Mahalanobis distance of samples follows a Chi-square distribution with d degrees of freedom. Expected value is d. In our case, degree of freedom is 17 since there are 18 samples estimated. Using Chi-square distribution table, we can get the critical value. This critical value is compared with estimated Mahalanobis distance attack point values shown in Table 3. According to the Chi-square distribution table we chose critical value as $x._{900}^{2}$=10.085, because the attack area is comparatively smaller than whole traffic, only 17 minutes.

Table 3. Chi-square Distribution Table (Degree of Freedom=17)

df	$x^2_{.995}$	$x^2_{.990}$	$x^2_{.975}$	$x^2_{.950}$	$x^2_{.900}$	$x^2_{.100}$	$x^2_{.050}$	$x^2_{.025}$	$x^2_{.010}$
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409

Table 4. Comparison of Mahalanobis Distance Values and with Critical Value (Time 20:44~20:52)

Time	20:44	20:45	20:46	20:47	20:48	20:49	20:50	20:51	20:52
d_M>df	60.35 >10.085	110.1 >10.085	123.5 >10.085	116.3 >10.085	95.96 >10.085	103.9 >10.085	116.4 >10.085	134.4 >10.085	129.7 >10.085

As shown in Table 4, all estimated Mahalanobis distance values of attack points are bigger than critical value which shows that these points are an exactly outliers.

4. Implication and Conclusion

A previous work [11] made use of an adaptive threshold method to detect the attack traffic. It verifies packet/bit rate directed to a particular IP if it has more than 3 standard deviations higher than the mean packet/bit rate for this IP. However, there is considerable interest in traffic metrics for anomaly traffic detection. Hence we chose TCP flags as an entropy random variable. We intend to show the attack area and how far it is distributed from normal behavior.

In this paper, we have introduced distribution and distance measurement based anomaly traffic detection method which detects attack area accurately. First, dataset was analyzed and estimated the entropy of TCP packets based on its flags. Since previous entropy based SYN flooding detection method [9] concentrated on unusual TCP handshake which is only appropriate for DDoS flooding attack. Our measurement result shows that making use of TCP flag as an entropy random variable can effectively detect anomaly traffic. Second, Mahalanobis distance noteworthy contributes to show the distribution of attack traffic and normal traffic.

Mahalanobis distance shows not only distance between two points but also expresses features of a point. In our case, dataset contains many identical SYN/ACK packets which are identified as attack packets. Thus we chose it as a feature of Mahalanobis distance. Otherwise, SYN packet or any other packets can be one of the feature candidates to detect flooding attack. Using chi-square test, we evaluated that the proposed method exactly differentiates normal and abnormal traffics.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(KRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2042035).

References

[1] M. Chandola, B. Arinda and V. Kumar, Anomaly Detection: A Survey, ACM Computing Surveys, 41, 3, (2009) July, pp. 1501-1558
[2] F. Y. Leu and I. L. Lin, A DoS/DDoS Attack Detection System Using Chi-Square Statistic Approach, Systems, Cybernetics and Informatics, 8, 2, (2010), pp. 41-51
[3] K. Wang and S. J. Stolfo, Anomalous Payload-based Network Intrusion Detection, Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection, LNCS 3224, Springer, (2004) pp. 203-222
[4] Chi-Square, http://www.aaec.ttu.edu/faculty/eelam/3401
[5] J. Santiago-Paz, D. Torres-Roman and P. Velarde-Alvarado, Detecting Anomalies in Network Traffic Using Entropy and Mahalanobis, Proceedings of the 22nd CONIELECOMP, (2012) Feb 27-29, Cholula, Puebla, pp. 86-91
[6] M. Bellaiche and J. C. Gregoire, SYN Flooding Attack Detection Based on Entropy Computing, Proceedings of the GLOBECOM, (2009) November 30 – December 4; Honolulu, USA, pp. 1-6.
[7] G. Nychis, V. Sekar, D. G. Andersen, H. Kim and H. Zhang, An Empirical Evaluation of Entropy-based Traffic Anomaly Detection, Proceedings of the 8th ACM SIGCOMM, (2008); New York, USA, pp. 151-156
[8] P. Du and S. Abe, Detecting DoS Attacks Using Packet Size Distribution, Proceedings of the 2nd Biometics on Bio-Inspired Models of Network, Information and Computing Systems, (2007) December 10-12; Budapest, Hungary, pp. 93-96
[9] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur and J. Srivastava, Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection, Proceedings of the 3rd SIAM International Conference on Data Mining, (2003) May 1-3; San Francisco, USA, pp. 25-36
[10] G. Thatte, U. Mitra and J. Heidemann, Parametric Methods for Anomaly Detection in Aggregate Traffic, IEEE/ACM TRANSACTIONS on Networking (TON), 19, (2011), pp. 1148-1158
[11] DoS_80-20110715, available through PREDICT

Authors

Dolgormaa Bayarjargal, received the B.S. degree from Chonbuk National University, Jeonju, Korea, in 2012, in computer science and engineering. She is currently a M.S. degree candidate in Division of Electronics and Information engineering (computer engineering) at the Chonbuk National University. Her research interests include Internet network security and traffic anomaly detection.

Gihwan Cho, received the B.S. degree from Chonnam University, Gwangju, Korea, in 1985, and the M.S. degree from Seoul National University, Seoul, Korea, in 1987, both in computer science and statistics. He received Ph. D degree in computer science from University of Newcastle, Newcastle Upon Tyne, England, in 1996.

He worked for ETRI(Electronics and Telecommunications Research Institute), Daejeon, Korea, as a Senior Member of Technical Staff from Sep. 1987 to Aug. 1997, for the Dept. of Computer Science at Mokpo National University, Mokpo, Korea, as a full time lecture from Sep. 1997 to Feb. 1999. From Mar. 1999, he joined to the Division of Computer Science and Engineering at Chonbuk National University, Chonju, Korea, and he is currently serving as a professor. His current research interests include mobile computing, computer communication, security on wireless networks, wireless sensor networks, and distributed computing system. Prof. Cho is a member of IEEE, KIISE, KIPS, KMMS, KSII.

