行政院國家科學委員會專題研究計畫精簡報告

網路鑑識系統研究與實作

Design and Implementation of Network Forensic System

計畫編號：NSC 98-2221-E-009-063-MY2

執行期間：98/08/01 ~99/07/31

主持人：李程輝教授 國立交通大學電信工程研究所

參與人員：謝景融、黃郁文、黃迺倫、李韋儒、陳曉薇、劉永祥、蔡承潔

國立交通大學電信工程研究所

中文摘要

由於網際網路的快速發展，近年來網路安全已成為大家所關注的主要領域。為了提升網路攻擊的偵測效率，在此研究中我們提出基於熵（Entropy）的網路行為模式建立演算法。此演算法包含兩個階段：第一階段目的是，以系統化的方式先將正常網路行為的封包，轉換成一個「相應不確定性」(Relative Uncertainty)的時間序列，再記錄此序列的機率分佈（Probability Distribution)；在第二階段，使用卡方適合度檢驗法（Chi-Square Goodness-of-Fit Test）偵測異常網路行為，本階段會觀測短期網路行為所建立的機率分佈，並與第一階段所建構出的長期網路行為比較，由於卡方適合度檢驗法是量測兩個機率分佈差異程 度的一種方法，故應用此法在這個階段。最後使用KDD CUP 1999的數據來驗證本研究所提出之演算法，實驗結果顯示此演算法，在選擇適當特徵集合的前提下，可達到高準確率及低計算複雜度的偵測結果。

關鍵字：行為特徵；熵；卡方；異常偵測

Entropy-Based Profiling of Network Traffic for Detection of Security Attacks

Tsern-Huei Lee
Department of Communication Engineering
National Chiao Tung University
Taiwan
Email: tlee@banyan.cm.nctu.edu.tw

Jyun-De He
Department of Communication Engineering
National Chiao Tung University
Taiwan
Email: bryan7404@gmail.com

Abstract—**Network security has become a major concern in recent years. In this research, we present an entropy-based network traffic profiling scheme for detecting security attacks. The proposed scheme consists of two stages. The purpose of the first stage is to systematically construct the probability distribution of Relative Uncertainty for normal network traffic behavior. In the second stage, we use the Chi-Square Goodness-of-Fit Test, a calculation that measures the level of difference of two probability distributions, to detect abnormal network activities. The probability distribution of the Relative Uncertainty for short-term network behavior is compared with that of the long-term profile constructed in the first stage. We demonstrate the performance of our proposed scheme for DoS attacks with the dataset derived from KDD CUP 1999. Experimental results show that our proposed scheme achieves high accuracy if the features are selected appropriately.**

Keywords-profiling; entropy; chi-square; anomaly detection

I. Introduction

With the rapid growth of Internet, there is increasing size and complexity of Internet traffic data. In the meanwhile, the damage of cyber attacks on the Internet is getting more and more severe. Therefore, network security is becoming an important issue for network users. Traditional network protection mechanism such as firewall is not enough to detect fast-changing attacks at the present time. Intrusion detection system is one of the major devices that has recently developed to detect and prevent different types of attacks.

The techniques adopted in intrusion detection are generally classified into two types: misuse detection and anomaly detection. Misuse detection is a technique which detects attacks with signatures. For accurate detection, the signature database of misuse detection systems must be updated frequently. Misuse detection systems are in general unable to detect new security attacks. Anomaly detection is a technique which profiles normal behaviors at the beginning, and compares network activities with normal behavior profiles to detect possible security attacks. Anomaly detection is based on the observation that the network activities during attacks are often quite different from the activities under normal usage. Statistics such as mean, variance, or even probability distribution were adopted as metrics for detecting attacks [1]. Compared with misuse detection, the major advantage of anomaly detection is that it does not require a database of signatures and can detect and prevent the outbreak of new attacks.

Kim et al. [2] proposed an optimized intrusion detection system using Principle Component Analysis (PCA) and Back-propagation Neural Network (BNN) based on Genetic Algorithm (GA). The research seeks to not only decrease dimension of features but also figure out intrinsic feature set. They used the KDD CUP 1999 data to validate the proposed approach for detecting DoS attack. The results show that the feature dimension decreases to 10 dimensions and the highest detection rate is about 91.00%. In this paper, we present a new scheme which achieves higher accuracy with lower complexity.

One important step of anomaly detection is data-processing (or data-profiling), a process which transforms original Internet packet information (e.g. protocol type, service type, port number, IP address) into "traffic behavior patterns" [3]. There are many possible methods for data-profiling. In our research, we use "entropy-based scheme" to create traffic behavior patterns for our data-profiling system and analyze them with "Chi-Square Goodness-of-Fit Test" [4]. The "particular distribution" of a specific number of packets is described using the entropy-based scheme. And the detection module detects attacks with the famous Chi-Square Goodness-of-Fit Test after the data-profiling process.

In Section II we introduce background including entropy, Relative Uncertainty, and Chi-Square Goodness-of-Fit Test. In Section III we present our proposed scheme and explain the details of the procedure. The dataset used in experiments is described in Section IV. Simulation results are contained in Section V. Finally, we draw conclusion in Section VI.

II. Background

A. Entropy and Relative Uncertainty

Entropy is an indication which measures the "observational variety" contained in the data [5]. Consider a random variable X that may have N_X discrete values. If we randomly observe X for m times, there would generate a probability distribution on X,

$$p(x_i) = m_i / m \ , \ x_i \in X \tag{1}$$

where m_i represent the number of times we observe X taking the value x_i.

The *entropy* of X is defined as

$$H(X) = -\sum_{x_i \in X} p(x_i)\log_2 p(x_i) \qquad (2)$$

$$0 \le H(X) \le H_{\max}(X) = \log_2 \min\{N_X, m\} \qquad (3)$$

where $H_{\max}(X)$ is the *maximum entropy* and by convention $0\log 0 = 0$ (unobserved possibilities do not enter the measure).

In [6], the *Relative Uncertainty* (RU) is defined as the *standardized* entropy and is given by

$$RU(X) = \frac{H(X)}{H_{\max}(X)} = \frac{H(X)}{\log \min\{N_X,\ m\}}, \qquad (4)$$
$$0 \le RU(X) \le 1.$$

Obviously, if all the observed values are the same, i.e., $p(x) = 1$, for some $x \in X$, then we have $RU(X) = 0$. On the other hand, if all the observed values are different, meaning that there is the highest level of variety in the observed data, then it holds that $RU(X) = 1$. In general, $RU(X) \ll 1$ indicates that the data distribution is more skewed, and $RU(X) \cong 1$ means that the values of the observed data are close to being uniformly distributed. In Section III we use above definitions and properties to convert original packet information into *behavior profiles*.

B. *Chi-Square Goodness-of-Fit Test*

The Chi-Square Goodness-of-Fit Test is a hypothesis test which compares two probability distributions to decide the degree of difference [4]. Some definitions as needed for our study.
The null and alternative hypotheses for the test are:

H_0: The variable has the specified distribution, and

H_a: The variable does not have the specified distribution.

The Chi-Square Goodness-of-Fit Test is to compute the test statistics expressed as

$$\chi^2 = \sum_i (O_i - E_i)^2 / E_i \qquad (5)$$

where O_i is the observed frequency and E_i is the expected frequency from the regular distribution for event *i*. The significance level *α* is a threshold which is decided based on the extent of computer vulnerability. For highly secure computer networks, *α* is chosen to be small so that results are *statistically significant* at *α* level. The degree of freedom is given by $df = I - 1$, where I is the number of possible values for the variable. The degree of freedom determines the exact shape of a chi-square distribution. Given a significance level, there is a corresponding threshold which is the decisive value to determine if the null hypothesis is true. In other words, if the chi-square value of two distributions computed by equation (5) exceeds the threshold, then the distributions of two observed data are declared to be different at significance level *α*. For example, in Fig. 1 the degree of freedom is 7, the significance level is 0.05, and the corresponding threshold is 14.067. If we perform the Chi-Square Test and obtain a value 20.0, then, at the 5% significance level, the data provide sufficient evidence to conclude that the observed distribution differs from the expected distribution.

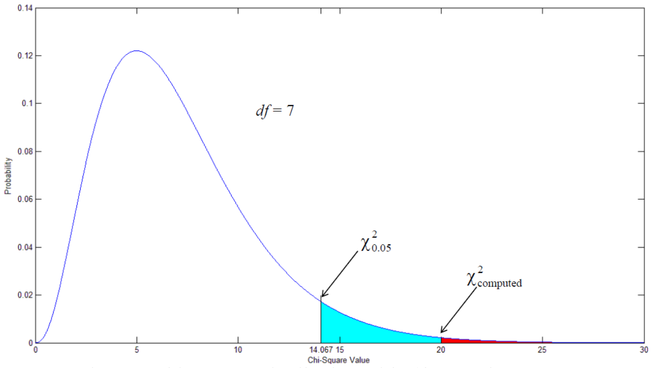

Fig. 1. Chi-Square Distribution with *df* = 7 and α = 0.05.

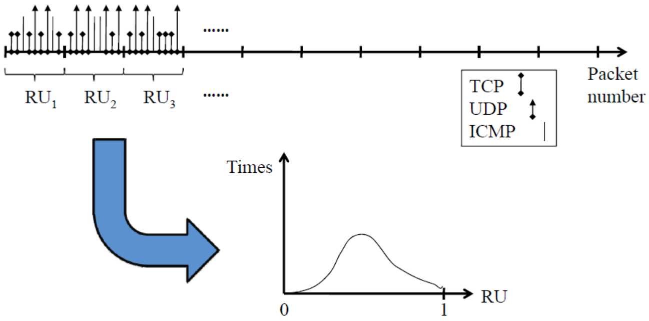

Fig. 2. The Relative Uncertainty Based Distribution of Protocol Types.

III. Our Proposed Scheme

This section contains the main ideas of data-profiling and anomaly detection with Chi-Square Goodness-of-Fit Test. A general Internet packet header includes information such as protocol type, port number, and IP address. Such information can be used to derive statistics that are related to the behaviors of Internet network, and such behaviors can be categorized as normal behaviors or abnormal behaviors. The abnormal behaviors may be regarded as attacks or intrusions. There are two issues. How do we scientifically build the profiles of the behaviors of a network system? And how do we identify attacks from the profiled data?

A. *Relative Uncertainty Based Distribution*

To address the first issue, we develop a methodology which uses the concept of *Relative Uncertainty*. As an example, assume that there are three types of protocols: TCP, UDP, and ICMP, they are recorded in packet headers. Hence, we observe sequential protocol types as a time series for a series of packets. This time series of protocol types can be transformed into a time series of Relative Uncertainty. As mentioned before, the Relative Uncertainty represents the observational variety in the network traffic. A major advantage of using Relative Uncertainty for data profiling is that it can find the same messages hiding in many features simultaneously without concerning the different units of the features.

In Fig. 2, there is a protocol type series extracted from the header of observed packets. The Relative Uncertainty is calculated every N packets. (In Fig. 2, we have $N = 9$.) A series of Relative Uncertainty is obtained after this process. The value of N is determined as follows.

We assume that the two adjacent values in the Relative Uncertainty series should not differ a lot for normal behavior. Therefore, the Mean Manhattan Distance, which can describe the absolute difference of the adjacent values in the Relative Uncertainty series, is adopted in determining the value of N. Define the Manhattan Distance of the Relative Uncertainty series as

$$MD_{(j,j+1)} = \sum_{k}^{K} \left| RU_j^k - RU_{j+1}^k \right| \qquad (6)$$

where K is the number of dimension (or features), and the Mean Manhattan Distance as

$$Mean\ MD = \frac{1}{J-1} \sum_{j}^{J-1} MD_{(j,j+1)} \qquad (7)$$

where J is the total number of index. The value of N is selected to minimize the Mean Manhattan Distance. After the RU series is generated, we construct the probability distribution of the series as the long-term profile of network behavior.

B. Chi-Square Test Based Anomaly Detection

To decide whether or not network behavior is normal during a specific time period, we collect network activities during that period, construct its profile (i.e., the distribution of the RU series), and compare the profile with that derived from normal behavior. We adopt the well-known Chi-Square Test to compare two distributions.

In our proposed approach, the normal behavior profile is constructed off-line with data collected for a long period of time without any attack. Since it is constructed by a long time observation, the profile is likely to be a stable distribution. The meaning of short-term profile is a model of dynamic behaviors which is generated by monitoring the short time behaviors of a network system during a specific period of time.

Fig. 3 shows the technique of Chi-Square Test based anomaly detection. The expected distribution is equivalent to the lone-term profile in this case and the observed distribution is the same as the short-term profile. Assume the expected distribution has been generated by observing long time behaviors of normal activities of a network system. We first apply a sliding window to compute the values of Relative Uncertainty which are transformed from online collection of network activities. The computed Relative Uncertainty values are then used to construct the observed distribution. Finally, we compare the expected distribution to observed distributions by using Chi-Square Goodness-of-Fit Test. Clearly, the process gives a sequence of chi-square values. If a chi-square value is greater than the pre-determined threshold, the activities during the period of time the chi-square value is computed are regarded as abnormal.

IV. Data Set

We use the data set of KDDCUP 1999 [7] built for the world-wide competition of designing intrusion detection systems. The data set has 41 features which can be grouped into 3 categories, namely, *Basic Feature*: those which can be extracted from packet header without inspecting the payload; *Content Feature*: those generated by accessing the payload of the original packet; and *Time based Traffic Feature*: those traffic features computed using a 2 second time window.

In our study, we focus on the denial-of-service attack (DoS attack) that is characterized by an obvious attempt by attackers to prevent legitimate users of a service from using that service. The basic and time based traffic features are suitable to detect the DoS attacks [8]. Therefore, we select 23 features that are chosen from the basic features and time based traffic features, as indicated in Table 1.

Table 1. 23 Features of the Dataset.

Label	Feature	Type of attribute
A	protocol_type	symbolic
B	Service	symbolic
C	src_bytes	numerical
D	dst_bytes	numerical
E	count	numerical
F	srv_count	numerical
G	serror_rate	numerical
H	srv_serror_rate	numerical
I	rerror_rate	numerical
J	srv_rerror_rate	numerical
K	same_srv_rate	numerical
L	diff_srv_rate	numerical
M	srv_diff_host_rate	numerical
N	dst_host_count	numerical
O	dst_host_srv_count	numerical
P	dst_host_same_srv_rate	numerical
Q	dst_host_diff_srv_rate	numerical
R	dst_host_same_src_port_rate	numerical
S	dst_host_srv_diff_host_rate	numerical
T	dst_host_serror_rate	numerical
U	dst_host_srv_serror_rate	numerical
V	dst_host_rerror_rate	numerical
W	dst_host_srv_rerror_rate	numerical

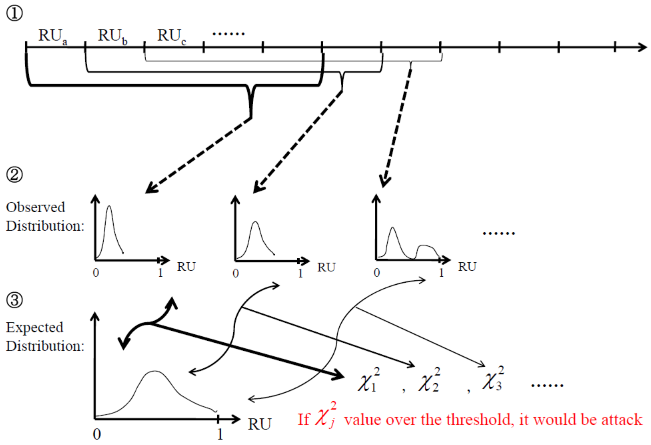
①
RU_a
RU_b
RU_c
……
②
Observed Distribution:
0
1
RU
0
1
RU
0
1
RU
……
③
Expected Distribution:
0
1
RU
χ_1^2 , χ_2^2 , χ_3^2 ……
If χ_j^2 value over the threshold, it would be attack

Fig. 3. The Process of Chi-Square Test Based Anomaly Detection.

V. Simulation Results

In this section, we evaluate the performance of our proposed behavior-based anomaly detection algorithm for KDD 1999 data set. First at all, we decide the size of N that minimizes the Mean Manhattan Distance. We request the number of elements in the Relative Uncertainty series of the long-term profile to be at least 100, because the Chi-Square Goodness-of-Fit Test is based on an assumption of large sample size. The result is $N = 24$.

Table 2. Confusion Matrix and Performance-Evaluation Method.

		Actual value		
		Bad	Good	total
Prediction outcome	Bad	(A) True Positive	(C) False Positive	(A) + (C)
	Good	(B) False Negative	(D) True Negative	(B) + (D)
	total	(A) + (B)	(C) + (D)	(A)+(B) +(C)+(D)

True Positive Rate (TPR) = A / (A+B)
False Positive Rate (FPR) = C / (C+D)
Accuracy (ACC) = (A+D) / (A+B+C+D)

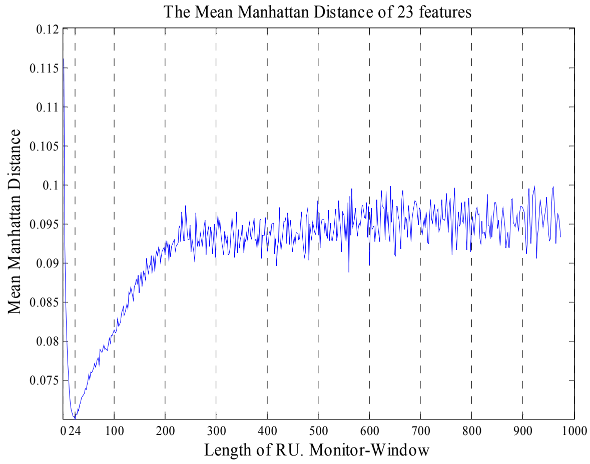

Fig. 4. Mean Manhattan Distance vs. the Length of Relative Uncertainty Monitor-Window.

Table 3. The Maximum Accuracy of Features that is Larger than 90%.

α	0.5%			0.1%			0.01%		
Feature	**ACC (%)**	**TPR (%)**	**FPR (%)**	**ACC (%)**	**TPR (%)**	**FPR (%)**	**ACC (%)**	**TPR (%)**	**FPR (%)**
C	94.28	98.45	22.60	94.89	98.21	18.56	95.43	98.24	15.94
D	95.17	98.94	20.02	95.91	98.71	15.35	96.55	98.64	11.86
M	94.03	97.68	20.74	94.59	97.69	18.01	95.18	97.55	14.44
N	94.80	97.00	14.33	94.95	96.57	11.76	95.00	95.65	7.71
R	95.97	98.11	12.71	96.01	97.93	11.81	96.19	98.51	13.26
S	94.20	97.30	18.63	94.40	96.92	16.03	94.44	96.60	14.51

In Table 2, there are the definitions of True Positive, False Positive, False Negative, True Negative, True Positive Rate (detection rate), False Positive Rate, and Accuracy. To evaluate our proposed scheme, we select one feature of the set at a time in this simulation. The top six features ranked by the accuracy are src_bytes (C), dst_bytes (D), srv_diff_host_rate (M), dst_host_count (N), dst_host_same_src_port_rate (R), and dst_host_srv_diff_host_rate (S). These features can be used to detect DoS attacks effectively.

Table 3 shows the accuracy, true positive rate, and false positive rate of the features at different significance levels. We study the accuracy for different significance levels. Results show that the accuracy increases while the significance level decreases. Note that a smaller significance level results in a larger threshold, which decreases false positive rate and increases false negative rate. In our experiment, the false negative rate increases by $\pm 1\%$ and the false positive rate reduces by $3 \sim 4\%$.

Table 4. Correlation Coefficient Matrix.

	C	D	M	N	R	S
C	1.0000	0.7448	0.6512	0.8037	0.7739	0.7082
D	0.7448	1.0000	0.8192	0.7259	0.6960	0.6242
M	0.6512	0.8192	1.0000	0.6717	0.6366	0.5863
N	0.8037	0.7259	0.6717	1.0000	0.9036	0.8684
R	0.7739	0.6960	0.6366	0.9036	1.0000	0.8483
S	0.7082	0.6242	0.5863	0.8684	0.8483	1.0000

Table 4 shows the correlation coefficient matrix evaluated from the Relative Uncertainty time series of the six features listed in Table 3. They are highly correlated with each other. In other words, using a single feature with the highest accuracy should suffice for detection of DoS attacks.

The true positive rate of our proposed scheme is higher than that (i.e., 91%) of the scheme presented in [2]. Besides, our scheme uses only one feature. Our study shows that transforming the original data sequence into a sequence of Relative Uncertainties could be an effective solution for detecting network attacks with low computation complexity.

VI. Conclusion

In this paper, we proposed a novel, two-stage approach for detecting network attacks. In the first stage, normal behavior profiles are constructed based on Relative Uncertainty. In the second stage, the Chi-Square Goodness-of-Fit Test is performed for the distributions obtained from behavior profiling and network activities collected online. We demonstrated the effectiveness of our proposed scheme with the KDD 1999 dataset for DoS attacks. Simulation results show that our proposed scheme achieves lower complexity and higher accuracy than previous schemes. Based on the experimental results, we believe that the proposed scheme could be a good choice for network behavior profiling and attack detection.

References

[1] T.-Q. Zhu and P. Xiong, “Optimization of membership functions in anomaly detection based on fuzzy data mining,” *in Proc. ICMLC International Conference Machine Learning and Cybernetics*, 2005.

[2] D. S. Kim, H.-N. Nguyen, T. Thein, and J. S. Park, “An Optimized Intrusion Detection System Using PCA and BNN,” *in Proc. Information and Telecommunication Technologies*, 6th Asia-Pacific Symposium, p.p. 356-359, 10-10 Nov. 2005

[3] K. Xu, F. Wangm S. Bhattacharyya, and Z.-L. Zhang, “A Real-time Network Traffic Profiling System,” *in Proc. DSN Dependable Systems and Networks*, 2007.

[4] R. Goonatilake, A. Herath, S. Herath, and J. Herath, “Intrusion Detection Using the Chi-square Goodness-of-fit Test for Information Assurance, Network, Forensics and Software Security,” *JCSC Journal of Computing Sciences in Colleges*, VOL. 23, p.p. 255-263, issue 1, October 2007.

[5] T. Cover and J. Thomas, “Elements of Information Theory,” ser. Wiley Series in Telecommunications, New York, Wiley, 1991.

[6] K. Xu and Z.-L. Zhang, “Internet Traffic Behavior Profiling for Network Security Monitoring,” *IEEE Transactions on Networking*, VOL. 16, NO. 6, December 2008.

[7] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[8] M. F. Abdollah, A. H. Yaacob, S. Sahib, I. Mohamad, and M. F. Iskandar, “Revealing the Influence of Feature Selection for Fast Attack Detection,” *IJCSNS International Journal of Computer Science and Network Security*, VOL.8, No.8, August 2008.