ISBN: 972-8924-19-4 © 2006 IADIS

MINING AN ANOMALY: ON THE SMALL TIME SCALE BEHAVIOR OF THE TRAFFIC ANOMALY

JungHyun Kim
Dept. of Electronics and Computer Engineering, Hanyang University, Korea
17 Haeng-Dang Dong, Sung-Dong Gu, Seoul, Korea
junghyun@ece.hanyang.ac.kr

Soohan Ahn
Dept. of Statistics, City University of Seoul, Korea
Jeonnong-dong(Siripdae-1 gil 12-1), Dongdaemun-gu, Seoul 130-743, Korea
sahn@uos.ac.kr

Youjip Won
Dept. of Electronics and Computer Engineering, Hanyang University, Korea
17 Haeng-Dang Dong, Sung-Dong Gu, Seoul, Korea
yjwon@ece.hanyang.ac.kr

ABSTRACT

To detect the emergence of traffic anomaly at an early stage, it is mandatory to have in depth understanding of the small time scale behavior of the anomalous traffic. In this work, we perform comprehensive study of fine time scale behavior of the anomalous network traffic. We first examine the time domain and frequency domain behavior of aggregate traffic. Second, we perform flow level analysis for traffic anomaly. Third, we develop new anomaly detection technique called CDF(cumulative Distribution Function) based clustering. We propose to use cumulative distribution function of the packet interval. We compare the accuracy of CDF based clustering against existing techniques: MVA(Multi variate analysis) and PCA(Principle Components Analysis). Our small time scale study suggests that it is not necessary to have multi-variate analysis to detect traffic anomaly and in fact packet interval distribution alone can yield sufficient accuracy in detecting anomalous traffic and flow.

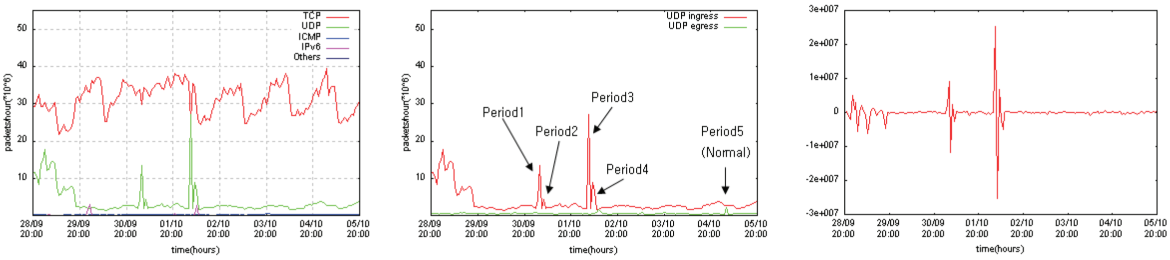
KEYWORDS

Anomaly Detection, Internet Measurement, Clustering, Principle Component Analysis

1. INTRODUCTION

The goal of this paper is to analyze characteristics of anomalous traffic which means a traffic associated with malicious behavior and to propose an efficient way of detecting its existence. A number of different tools have been applied to detect the existence of anomaly and to separate the anomalous behavior from the normal one: time series analysis (Brutlag 2000), frequency domain analysis (Barford et al 2002), principle component analysis(PCA) (Lakhina et al. 2004), wavelet theory (Crovella et al. 2003), and etc. A few works applied the information theoretic measure(*entropy*) to capture the traffic characteristics (Lee et al. 2001) (Xu et al. 2005). Recent works extend the idea of anomaly detection further to traffic classification (Karagiannis et al. 2005) (Lakhina et al. 2005) (Xu et al. 2005).

The objective of our work shares much of its motivation with the existing works in a sense that we intend to examine the anomalous behavior of the Internet backbone traffic and to characterize its behavior. The key contribution of these sort of works should be How to detect and prevent such behavior at an early stage. Given the proliferation of malicious traffic, the importance of the early detection cannot be emphasized further. For quick reaction to the emergence of the traffic anomaly, it is very important to understand the small time scale behavior of the aggregate traffic as well as individual flows. Many of the abovementioned traffic studies use the aggregated feature statistics at a certain time interval, e.g. at 5 min. While this level of aggregation is sufficient for observing and analyzing many of the traffic anomalies, we carefully believe that the characterization and analysis based upon this coarse-grained data may miss key information for real-time detection and prevention of anomalous traffic. As a complementary effort to the existing studies on traffic anomalies, we studied the behavior of the traffic at the small time scale.



(a) Entire Ingress Traffics

(b) UDP traffic only

(c) UDP Traffic after high-pass filter

Figure 1. Aggregate Traffic Behavior

2. DATA

We collect the packet trace from OC3 link of Internet backbone on one of the Exchange Point(IX) of Korea for one week(October 28 2004 to November 4 2004). Our packet trace data has high precision time stamp(10 µsec) which enables us to zoom into the traffic behavior in very fine time scale. Through this effort, we were able to carry out reliable and comprehensive traffic analysis for anomaly detection. We develop our own traffic analysis tool, *DMC Traff Mon*, for extracting and analyzing data in *libpcap* format data.

Fig. 1(a) illustrates the traffic volume in 1 hour time scale for a week. Our study is particularly focused on analyzing the UDP traffic anomaly observed in our trace. We can observe that in a certain period, UDP traffic radically increases and lasts for tens of minutes. Traffic anomaly is quite vague term and it is not trivial to present its clear definition. However, in our trace data(Fig. 1(b)), we can identify anomalous traffic region without much difficulty. We use simple high pass filter to extract the high frequency component of the traffic volume time series. It magnifies anomalous behavior more clearly(Fig. 1(c)).

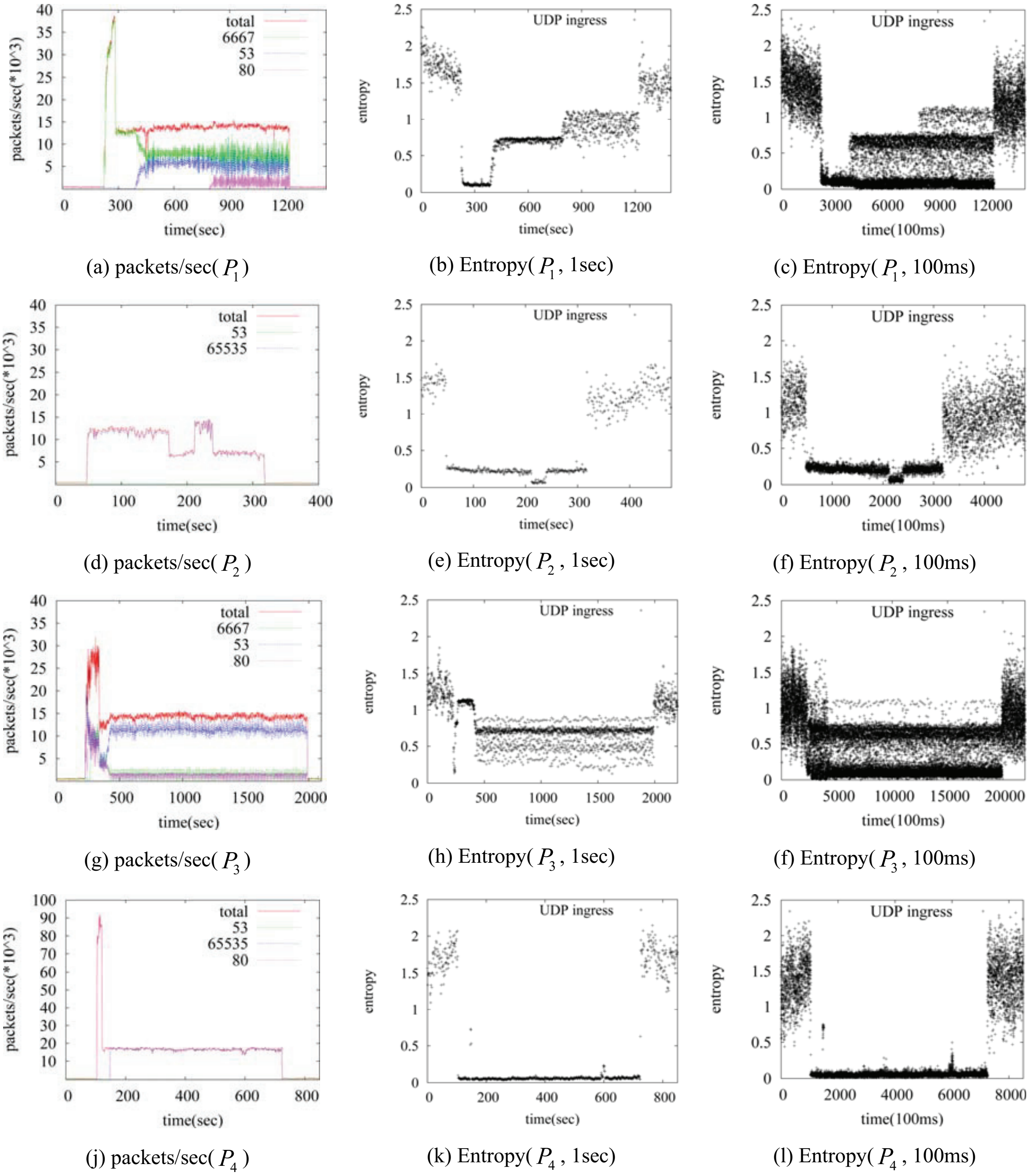
In this work, we analyze the small time scale behavior of anomalous UDP traffic and propose a novel scheme for detecting anomalous flow. The anomalous UDP traffic in our data is a sort of UDP flooding attack. we observed that UDP traffic which is intolerable volume went toward single victim.

3. VOLUME LEVEL ANALYSIS

3.1 Time Domain Behavior

Figures in Fig. 2 illustrate temporal behavior of the traffic in P_1, P_2, P_3 and P_4 in various aspects: packets/sec and entropy. It is found that the size of each peak in Fig. 1(b) is linearly proportional to the duration of the anomaly. More interestingly, the traffic anomalies bear very similar sustained data and packet rate across the periods(Fig. 2(a) - Fig. 2(j)).

Distribution of packet features, i.e. port numbers and IP addresses, is also important aspect of traffic. A number of works suggested to use information theoretic measure, *entropy*, to detect a certain change in underlying network traffic (Gu et al. 2005) (Lakhina et al. 2005). Entropy is a mathematical metric of the uniformity of a distribution and is defined by $H(P) = -\sum_i p_i \log p_i$ where $p_i = \frac{n_i}{N}.n_i$ and N denote the number of entities for class i and the total number of entities, respectively. The packets in anomalous traffic intervals are concentrated a small number of ports. As relatively larger fraction of packets converge to a few number of ports, the entropy of the packet distribution is expected to decrease.



(a) packets/sec(P_1)

(b) Entropy(P_1, 1sec)

(c) Entropy(P_1, 100ms)

(d) packets/sec(P_2)

(e) Entropy(P_2 , 1sec)

(f) Entropy(P_2 , 100ms)

(g) packets/sec(P_3)

(h) Entropy(P_3 , 1sec)

(f) Entropy(P_3 , 100ms)

(j) packets/sec(P_4)

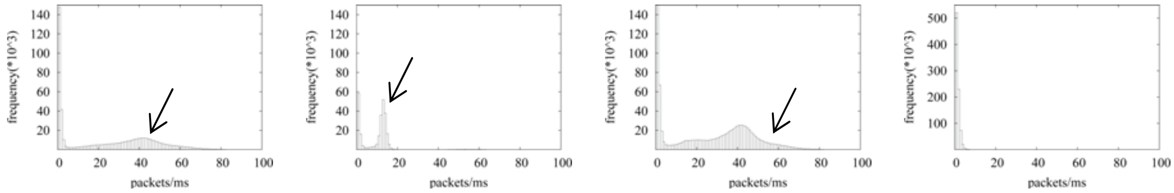
(k) Entropy(P_4 , 1sec)

(l) Entropy(P_4 , 100ms)

Figure 2. Time Domain Aspect of Traffic

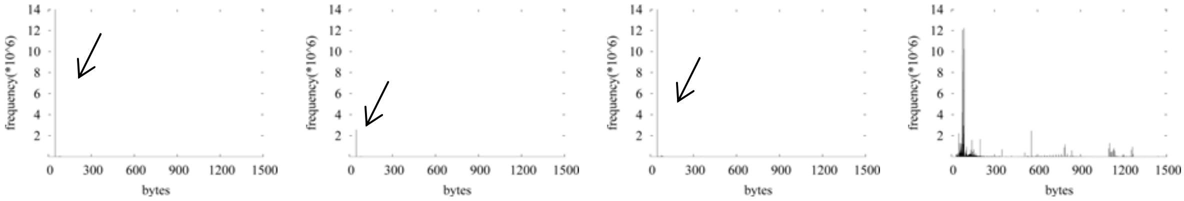
Each figure in Fig. 2(a) - Fig. 2(j) illustrates the aggregate traffic volume as well as the traffic volume for major destination ports. In P_1, P_2, P_3 and P_4, we can find that packets towards port 53(DNS), port 6667(IRC) and port 80(HTTP) constitute dominant fraction of the entire traffic. We examine the entropy aspect of the traffic for each period(Fig. 2(b) - Fig. 2(l)). We can see that entropy drops significantly with the start of anomaly. In fact, these packets converge to the same IP address.

Fig. 2(b) - Fig. 2(k) illustrate the entropy of the port distribution in 1 sec time scale while Fig. 2(c) - Fig. 2(l) illustrate the entropy in 100 msec time scale. In 1 sec time scale entropy data, the entropy increases in stepwise fashion as the number of destination ports increases. In 100msec time scale, we can find interesting phenomenon which has not been seen in 1 sec time scale entropy data. In Fig. 2(c), we cannot find the stepwise increase in entropy which we have observed in Fig. 2(b). Instead, new entropy band begins with the introduction of new destination port. This suggests that anomalous traffic has strong time dependent behavior in sub second time scale.



(a) Anomalous Period(P_1) (b) Anomalous Period (P_2) (c) Anomalous Period(P_3) (d) Normal Period(P_5)

Figure 3. Frequency Domain Aspect of Traffic (Packet Count)



(a) Anomalous Period(P_1) (b) Anomalous Period (P_2) (c) Anomalous Period(P_3) (d) Normal Period(P_5)

Figure 4. Packet Size Distribution of UDP Traffic

3.2 Frequency Domain Behavior

Figures in Fig. 3 illustrate the frequency representation of packet count process. Byte count process exhibit similar spectral behavior and we skip the respective graphs. These are for only UDP traffics. We partition the time axis into 1 msec time interval and measure the packets/msec and bytes/sec for each interval. Y-axis in these graphs is the number of the time bins for the respective bandwidth. In anomalous traffic region(figures from Fig. 3(a) to Fig. 3(c)), frequency domain representation of the bandwidth process has high frequency component. On the other hand, bandwidth spectrum in the normal traffic region(P_5, Fig. 3(d)) does not have high frequency components. This phenomenon is quite intuitive considering the fact that aggregate UDP traffic significantly increases during the traffic anomaly region. We found that these anomalies are triggered by few numbers of flows which have burstier packet interval.

The packet size is another important measure to represent the traffic characteristics. Figures in Fig. 4 illustrate packet size distributions in P_1, P_2, P_3 and P_5. It is found that packet size distribution in anomalous traffic region exhibits rather different behavior than the normal traffic region. In anomaly traffic regions(P_1, P_2 , P_3 and P_4), most UDP packets(more than 97%) are 48 Byte, which means that packet is actually empty.

4. FLOW LEVEL ANALYSIS

We define a flow as a series of packets which have same *source IP*, *destination IP*, *source port*, *destination port* and *protocol*. If packet interval is larger than 60 sec, we regard it as a start of new flow. We examine the details of a flow in each period. Fig. 5 illustrates the packets/flow distribution.

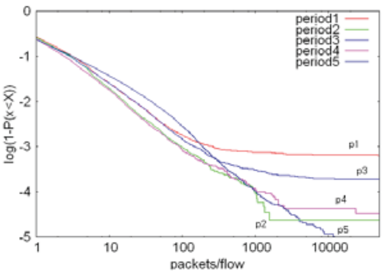


Figure 5. Packets/flow Distribution

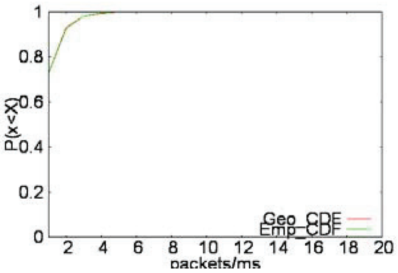
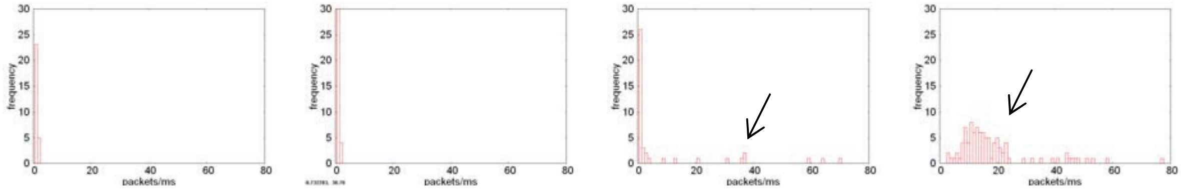


Figure 7. Empirical and Geometric Distribution of Packet Interval

It exhibits the stark contrast between the flows in anomalous traffic region(P_1, P_2, P_3 and P_4) and the flows in normal traffic region(P_5). The packets/flow distribution in anomalous traffic regions(P_1, P_2, P_3 and P_4) has much heavier tail than the distribution in P_5 and the $\log P(X > x)$ decays very slowly. It is found that there is not much difference between quantile statistics(byte count, packet count and flow duration) of normal region(P_5) and quantile statistics of anomalous traffic regions(P_1, P_2, P_3 and P_4) and existence of anomalous flows does not affect the quantile statistics. It implies that few number of flows trigger traffic anomaly and have significant impact on the aggregate behavior of the UDP traffic. Therefore, it is critical to have flow level understanding of traffic anomaly for quick and efficient diagnosis.

5. MINING ANOMALOUS TRAFFIC

5.1 Detecting Anomaly



(a) 2725.2 sec

(b) 2725.3 sec

(c) 2725.4 sec

(d) 2725.5 sec

Figure 6. Temporal Changes in Frequency Behavior (100 msec scale)

We propose an efficient statistical method to detect a time epoch where anomalous traffic exist. At first we denote the aggregated number of packets during a time period from $[t + 0.001(i-1)]$ sec to $[t + 0.001i]$ sec by $X_{t,i}$ where $t = 0,1,\cdots,$ and $i = 1,2,\cdots,1000$. We assume that $X_{t,i}$'s are mutually independent with distribution function F and probability mass function f to make a problem simple, which can be guaranteed when a packet arrival process is a Levy process. We also denote by $\hat{F}_t(x)$ the empirical distribution function in the t -th 1 sec-long time period, that is, $\hat{F}_t(x) = 0.001\sum_{i=1}^{1000} I(X_{t,i} \le x),\ x = 0,1,\ldots.$

The ideas is to exploit the changes in frequency domain behavior of the underlying traffic. Figures in Fig. 6 illustrate the shape-changes of the empirical probability mass function $\hat{f}_t$ corresponding to $\hat{F}_t(x)$. There are important observations in these figures. Fig. 6 plots the change in frequency domain behavior in 100 msec time scale. It takes two time slots, i.e. 200 msec, until the frequency domain representation exhibits anomalous characteristics. These graphs suggest an important information about detection latency. Given the speed of proliferation of malicious Internet traffic, it is very important to detect such an anomaly at its early stage. Fig. 6 suggests that the anomaly such as DDOS attack can be detected in subsecond time latency.

As shown in Fig. 3, the frequency domain behavior of the packet count process exhibits unimodal shape and has short tail in the period of normal traffic. In the period of anomalous traffic, the functions change and have bimodal shape and long tails. We investigate a method to detect the change point of the distribution functions. The technique proposed in this section is to use goodness-of-fit test which tests statistically if a given distribution function fits well the empirical data. For this, we assume through empirical study that given $X_{t,i} > 0$, $X_{t,i}$ follows Geometric distribution, that is,

$$\begin{aligned} f(x) &= P[X_{t,i} = x \mid X_{t,i} > 0] \\ &= p(1-P)^{X-1}, x = 1,2,\cdots \end{aligned} \tag{1}$$

where $p = 1/E[X_{t,i} \mid X_{t,i} > 0]$ and we estimate it using data in P_5, which is considered as there is no anomalous traffic. The estimated value is $p = (1.377)^{-1}$. It is shown in Fig. 7 that empirical distribution can be fitted with geometric distribution with sufficient accuracy. To detect an time epoch where anomalous traffic exist, we use the Smirnov-Von Mises test statistic defined as

$$W_t^2 = \int_{-\infty}^{\infty} (F(x) - \hat{F}_t(x))^2 dF(x) \tag{2}$$

$$\approx \frac{1}{n_t} \sum_{i=1}^{n_t} (F(x_t, i) - \hat{F}_t(x_t, i))^2 \tag{3}$$

where $x_{t,i}$ is a observed value of $X_{t,i}$ and n_t is the number of i's such that $x_{t,i} > 0$ for each t. For more details, refer to (Anderson et al. 1952). Our method is simple and determine a time epoch of anomalous traffic if the probability of that the test statistic is greater than a observed value is less than a given critical value α (for example 0.01 or 0.05). The distribution of W_t^2 can be computed using approximation or using Monte Carlo simulation. In this work, we use a probability table provided in (Anderson et al. 1952). They provide a table of probability $P(n_t W_t^2 \le z)$ for several values of z. Thus, we determine t as a time epoch of anomalous traffic if the p-value $P(n_t W_t^2 > n_t w_t^2) < \alpha$ t is a observed value of W_t^2.

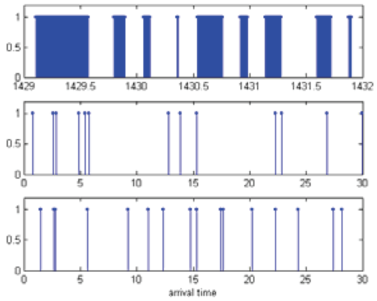


Figure 8. Packet Arrival Pattern in different flows

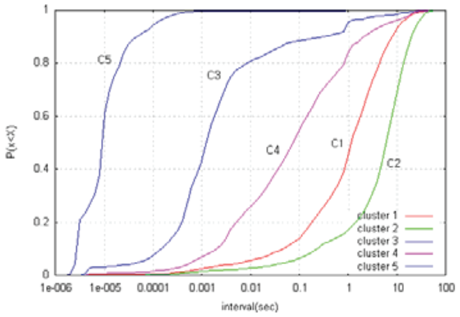


Figure 9. Empirical Distribution of Packet Inter-Arrival Time of each group

5.2 CDF Based Classification

For effective diagnosis at early stage, it is very important to quickly identify these flows and to take preventive action. We begin the study by investigating a way to characterize anomalous flows. Through physical examination, we identify 24,466,335 flows in P_1, P_2 and P_3. Among them, we extract flows which have more than 200 packets to characterize anomalous flows. We find 53 anomalous flows which generate single port or multi-port UDP flooding. Fig. 8 illustrates the packet arrival pattern in different types of flows. Packet interval in the top is generated by one of the anomalous flow. It has very bursty nature and we conjecture that the anomalous flow has widely different stochastic characteristics in its packet interval distribution. Exploiting this, we propose to use cumulative distribution function(CDF) of inter-arrival times of packets of a flow in classifying the flows. That is, we define n_i and $T_{i,k}, K = 1,\cdots,n_i - 1$, as the number of packets and the time elapsed between successive packets in i -th flow and

$$Y_{ij} = \sum_{k=1}^{n_i} I(T_k \le y_j), i = 1,\cdots,n \qquad (4)$$

$$y_j = 10^{-6+0.1(j-1)}, j = 1,\cdots,81$$

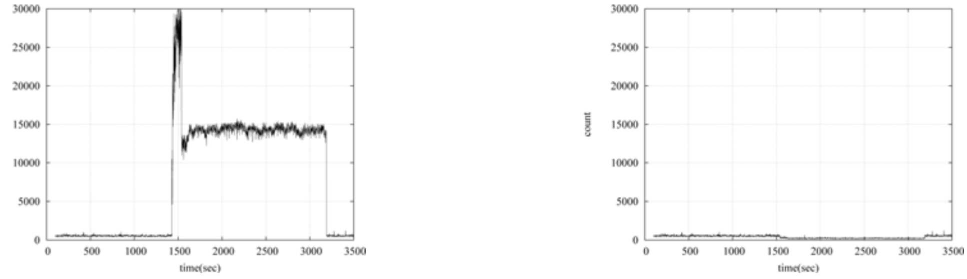
where n is the number of flows in P_1, P_2 and P_3, respectively. We perform k-means clustering on variables $(Y_{\cdot 1},\cdots,Y_{\cdot 81})$. Maximum number of clusters is set to 5. Fig. 9 illustrates the empirical distribution of inter-arrival times of packets of flows in each group. Cluster 5 consists of only anomalous flows.

Table 1. Comparison of CDF clustering and PCA clustering

	PC_1	PC_2	PC_3	PC_4	PC_5	*Total*
CDF_1	0	34	4	8	293	338
CDF_2	0	4	0	9	173	186
CDF_3	2	0	8	0	0	49
CDF_4	1	5	38	2	200	246
CDF_5	0	0	53	0	0	53
Total	3	43	103	19	714	882

6. CLASSIFICATION RESULTS

There are a number of different flow features for flow classification and we used CDF of packet intervals. We compare the accuracy of our CDF based clustering method against the other clustering metric used in the preceding works: average packet sizes (X_1), (standard deviation of packet sizes)/(mean of packet sizes)(X_2), mean of packet inter-arrival times(X_3) and (standard deviation of packet inter-arrival times)/(mean of packet inter-arrival times)(X_4). Although one of widely used variables used in flow analysis is the number of packets per flow, we do not use this because we have to wait till the end of flow to obtain this information and it does not satisfy the requirement of early detection. In CDF based clustering, we observe that 53 anomalous flows are all grouped into the same cluster. To remove the effect of correlations, though the variables are not highly correlated, we apply the principle component analysis(PCA) to select minimal set of orthogonal variables for classification. We apply k-means clustering analysis using variables obtained through the principle component analysis. We call this approach PCA clustering. We present clustering results in Table 1.



(a) Before removal(P_3)

(b) After removal(P_3)

Figure 10. Traffic before and after removing anomalous flows

Examining Table 1, we can see that PCA based clustering also performs reasonably well and it is not trivial to judge which of the two, CDF based clustering or PCA based clustering, is better than the others. However, data says it all. We closely examine the trace data and there exist total 53 anomalous flows. Based upon this empirical validation, we carefully conclude that distribution function of packet inter-arrival time of each flow is an efficient measure for clustering and classifying flows. We identified the anomalous flows with CDF based classification and removed them from the traffic. Fig. 10 illustrates aggregate traffic before and after removal of anomalous traffic. It can be seen that removal of flows makes the aggregate traffic behavior to normal.

7. CONCLUSION

Detecting and diagnosing anomalous traffic is ever challenging subject. While there have been numerous efforts in characterizing the anomalous traffic, there remain a number of issues which require further elaborate treatment. Significant fraction of preceding works deal with the traffic data aggregated at the fixed time interval. Aggregation and sampling based traffic study makes the study more efficient and faster. Further, via reducing the number of samples, we potentially can increase the dimension of the sample and can extract more inter-dimension correlation information. However, the study on coarse grain data may not be able to capture key information which is indispensable in quick detection of the emergence of the traffic anomaly. In this work, We find that the malicious flow susceptible DOS attach exhibits significantly different small time scale behavior from the normal flows. The difference is verified via statistical hypothesis test. We propose simple yet efficient clustering method which allows for early detection of malicious flows. This method is especially adequate to detect UDP flooding attack.

ACKNOWLEDGEMENT

This research was supported by the SRC/ERC program of MOST/KOSEF (Grant No. R11-2000-073-00000) and Second stage of Brain Korea 21 program..

REFERENCES

Anderson et al., 1952. *Asymptoic theory of certain goodness of fit criteria based on stochastic process. Annals of Mathematical Statistics.*

P. Barford et al., November 2002. A signal analysis of network traffic anomalies. *Proceedings of Internet Measurement Workshop*, Marseille, France.

J. Brutlag, 2000. Aberrant behavior detection in time series for network monitoring. *Proceedings of 14*th *USENIX System Administration Conference*.

M. Crovella et al., 2003. Graph wavelets for spatial traffic analysis. *Proceedings of INFOCOM '03*.

Y. Gu et al., 2005. Detecting anomalies in network traffic using maximum entropy estimation. *Proceedings of Internet Measurement Conference*, pages 345–350.

T. Karagiannis et al., August 2005. Blinc: Multilevel traffic classification in the dark. *Proceedings of SIGCOMM '05*, Philadelphia, USA.

A. Lakhina et al., August 2004. Diagnosing network wide traffic anomalies. *Proceedings of SIGCOMM'04*, OR, USA.

A. Lakhina et al., August 2005. Mining anomalies using traffic feature distributions. *Proceedings of SIGCOMM '05*, Pennsylvania, USA.

W. Lee et al., 2001. Information-theoreticmeasure for anomaly detection. *Proceedings of IEEE Symposium on Security and Privacy*.

K. Xu et al., August 2005. Profiling internet backbone traffic: Behavior models and applications. *Proceedings of SIGCOMM '05*, Philadelphia, USA.