The Stability of the Internet Traffic Features

Bin Zhang[1, a *], Yuandong Mao[2, b], Mei Zhang[2, b], Yun Yu[1, c], Guoquan Jiang[1, d] and Bo Deng[1, e]

[1]Nanjing Telecommunication Technology Research Institute, Nanjing, China

[2]The First Middle School of Xinxiang Country, Henan, China

[a]zhang_bin163@163.com, [b]657697622@qq.com [c]yuyun56@163.com, [d]jianggq2001@163.com, [e]myhspace@163.com,

*The corresponding author

Keywords: Traffic feature; Entropy; Stability

Abstract. In this paper, we present a statistical analysis of traffic features at the packet level. We show that all traffic features demonstrate similar approximately power-law distribution for different time and interval at minute time scale except for the packet size. We observe that feature entropy and independent feature symbols number for fixed packets number are relatively stable in a short time interval, which is very useful for traffic anomaly detection.

Introduction

Network traffic analysis has become increasingly important for various network management functions such as traffic modeling, traffic engineering and anomaly detection and response. In 1999, Vern Paxson published a presentation titled 'Why Understanding Anything About The Internet Is Painfully Hard' [1]. In his presentation, Vern Paxson describes the Internet as: "ubiquitous diversity and change: over time, across sites, how the network is used, and by whom".

Prior work has focused primarily on distributions concerning the flow size distribution (given a flow size s, find the number of flows with size s) is studied in [2-4], where a flow is a sequence of packets that share the same five tuples of (source IP, port, destination IP, port and protocol), Distributional aspects such as entropy (e.g. entropy of the packet distribution over various IP addresses and ports) has also been a subject of current interest [5-8].

Despite many work on feature distributions concerning the flow size, little attention has been paid on traffic feature distributions involving source and destination addresses and ports, etc; Although entropy of traffic features has been studied in anomaly detection area, little work correlates traffic feature distributions to their entropy values. To our knowledge, putting traffic feature entropy value distributions on a very long time zone and analyzing their stability is also an untouched area.

In this paper, we focus on six traffic features: source IP address (SIP), destination IP address (DIP), source port (SP), destination port (DP), packet size (PS) and flow symbol (FS). FS is defined as a symbol to indicate different flows, i.e., if the packets have the same five tuple they have the same FS. For studying the traffic features we define a stochastic process $X=(Xs , : s = 1 , 2 , . . .)$ to stand for a specific traffic feature symbol series of consecutive packets of a traffic trace.

We examine each traffic feature at the packet level in this work and try to answer the following basic questions: What does the distribution of each traffic feature looks like at different time and interval? Does it exist some characters that do not change with time and observation point?

To answer these questions, we measure, using statistical parameters, the features in four traffic datasets, two collected in 2007 and 2010 at one of the link between TUNET(the Tsinghua University campus Network) and CERNET(the China Educational & Research Network), the other two in 2007 and 2010 at CERNET international link from China to USA.

Related Work

Some papers [2-4] focus on tracking flow size distribution. Some works [10, 11] disclose other characters of traffic feature. Eddie et al. [10] find that traffic IP Addresses can be well modeled by a multi-fractal Cantor by observing their structure. Nahur et al. [11] find the long range mutual information phenomenon of FS.

There are also a lot of papers [5-8] that use traffic feature entropy to analyze traffic, but none of them to put the feature entropy in a long time zone to observe the changing pattern. Finally, Alice et al.[12] find the contribution of some traffic feature in discriminating the protocol classes is the same in different network locations and if it does not change in time. Different from [12], our work tries to find the more general rule for the traffic features.

Datasets

Our traces are collected from CERNET and TUNET. CERNET is the first and the largest nationwide education and research computer network in China, and is one of China's seven major backbone networks. TUNET is the biggest campus network in China and also one of the biggest campus networks in the world. Two datasets are collected in 2007 and 2010 at one of the OC-48 link between TUNET and CERNET, the other two in 2007 and 2010 at CERNET international OC-48 link from China to USA. For accurate analysis, instead of using netflow sampled flow-level data all traffic traces we used are collected by our monitoring system like IPMON [9]. All traces are fully captured packet-level data without sampling. Only packet headers are captured for saving spaces.

Feature Distribution

We define symbol length for a specific feature similar as the flow length, i.e., the same symbol's number for an interval or consecutive packets number. From statistical analysis of different traffic traces we find that the symbol length distributions of traffic features (SIP, DIP, SP, DP, FS) all approximately follow a power-law distribution for different time and interval at minute time scale, but the PS is an exception. We plot the frequency versus the DIP symbol length in log-log scale for different consecutive packets number in Fig. 1 a)-e). We plot the same figure for PS in Fig. 1 f)-g).

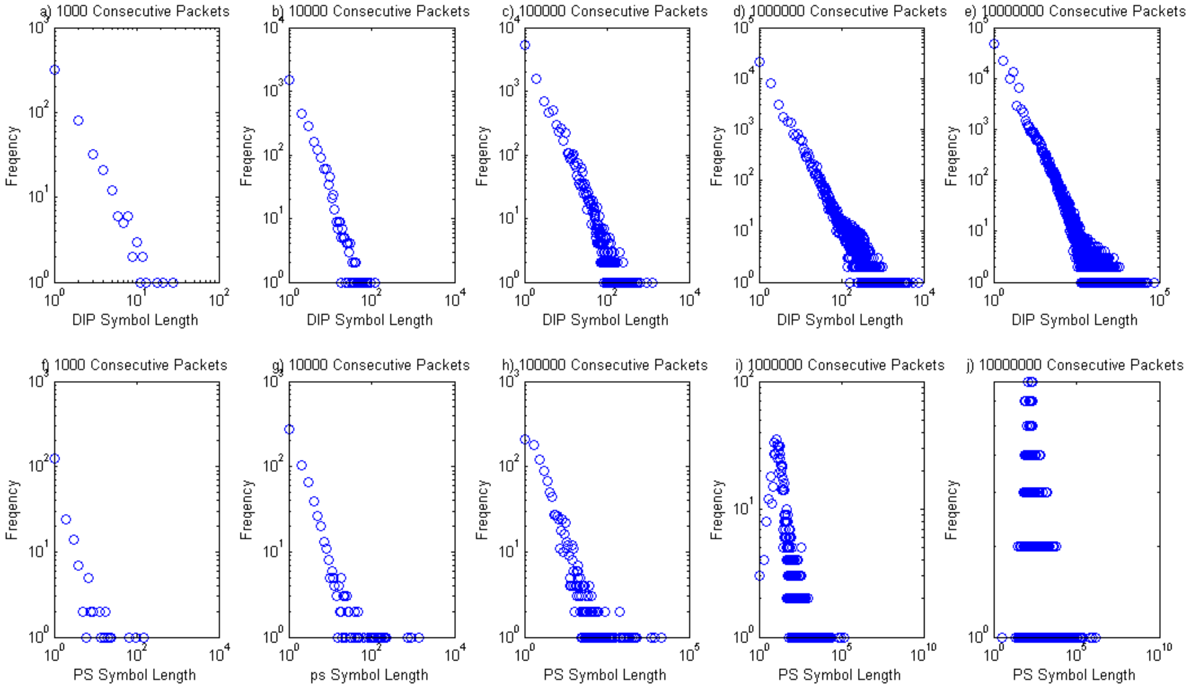

Figure 1. The symbol length distribution

Comparing the figures we can see the difference. The power law does not change for X (DIP) distribution with different consecutive packets number, while the distribution law begins to change for X (PS) with gradually adding the statistical packets number (see Fig. 1 i),j)). Others (SIP, SP, DP, FS) are similar to DIP, i.e., they have similar power-law distributions as DIP---the main difference among them is the slope value. We here neglect their figures due to limited space. The reason for the changing distributions maybe X (PS) value is restricted between 40 (no payload) and 1500(MTU) in ethernet for ipv4, while other features value has a wider range (232 for IP addresses, 28 for ports, 283 for flow symbols). We will leave further analysis on feature distribution for our future works.

Feature Analysis

Methodology of Computing Entropy. Shannon introduced information entropy to capture the degree of dispersal or concentration of a distribution of a sample. We start with an empirical process $X = \{n_i;\ i = 1, \quad N\}$, meaning that feature symbol i occurs n_i times in the sample. Then the sample entropy is defined as:

$$H(X) = -\sum_{i=1}^{N} (\frac{n_i}{S}) \log(\frac{n_i}{S}) \tag{1}$$

Where S is the total number of observations of X. N is the total independent feature symbols number in all observations. The value of sample entropy lies in the range *(0; logN)*. The metric takes on the value *0* when the distribution is maximally concentrated, i.e., all observations are the same. It takes on the value *logN* when the distribution is maximally dispersed, i.e., $n_1 = n_2 = \ldots = n_N$.

Hence, entropy can be computed on a sample of consecutive packets. We compute each entropy of traffic features in Table 1 using the same methodology as [5] which sets a sliding window of a fixed size, *W=10000*. That is, *S=10000*. We compute the first *S* packets feature entropy for a traffic trace and then move to the next *S* packets. We define the consecutive *S* packets as a packet unit. *N* is the total independent feature symbols number for all observations in a packet unit. Let us take the feature---SIP as an example to illustrate, if there are *K* distinct SIP number in a packet unit, then the independent SIP symbols number---*N* is *K*.

Entropy Stability. From measuring and analyzing traffic feature entropy values of traffic traces from different time and locations, we find all features entropy values keep relatively stable for a very short time interval(at second time scale for our traces). Fig. 3 presents the consecutive 200 entropy values distribution for all traffic features. From Fig. 2 we can see all feature entropy values keep relatively stable, hence the current entropy value range can be predicted by a few previous entropy values. Sharp entropy value changes (out of the range) may indicate an anomaly. Many anomaly detectors [5-7] are based on this similar observations.

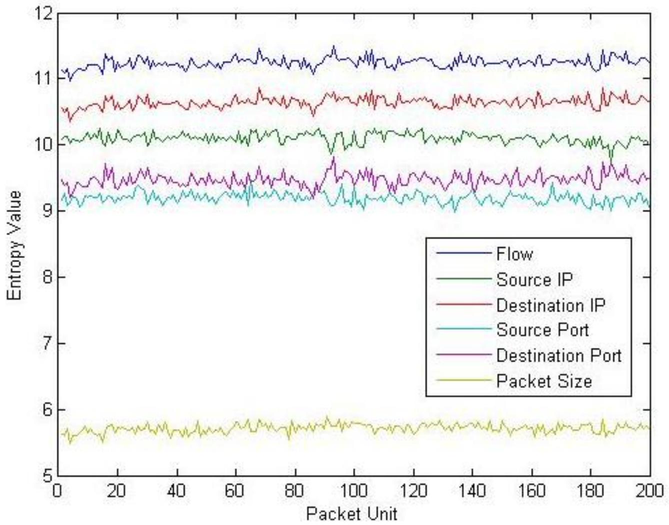

Figure 2. Entropy values of 200 consecutive packet unit

Conclusions

In this paper we presented a statistical analysis for measuring the packet-level features of internet traffic. The analysis to different traces has shown that all traffic features have similar approximately power-law distributions for different time and interval at minute time scale except for the packet size. The feature entropy values keep relatively stable for a short time interval because adjacent packet unit traffic shows more similar distributions than the farther one.

Acknowledgements

This work was supported by the China Postdoctoral Science Foundation under No. 2015M582832, the Jiangsu Post-doctor Research Fund of China under No. 1402138C

References

[1] V. Paxson. “Why Understanding Anything About The Internet Is Painfully Hard,” UCB Berkeley MIG Seminar, April 1999.

[2] N. Duffield, C. Lund, and M. Thorup. Estimating flow distributions from sampled flow statistics. In SIGCOMM, 325–336, 2003.

[3] A. Kumar, M. Sung, J. J. Xu, and J. Wang. Data streaming algorithms for efficient and accurate estimation of flow size distribution. In SIGMETRICS/ Performance, 177–188, 2004.

[4] L. Yang and G. Michailidis. Sampled based estimation of network traffic flow characteristics. In INFOCOM, 1775–1783, 2007.

[5] L. Feinstein, D. Schnackenberg, R. Balupari, D. Kindred. Statistical approaches to DDoS attack detection and response. In DARPA Information Survivability Conference and Exposition, Vol.1, 303 - 314, 2003

[6] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. SIGCOMM CCR, 35(4):217–228, 2005.

[7] G. Nychis, V. Sekar, DG. Andersen, H. Kim, H. Zhang. An Empirical Evaluation of Entropy-based Traffic Anomaly Detection. In IMC, 151 - 156, 2008.

[8] A. Lall, V. Sekar, M. Ogihara, J. Xu, and H. Zhang. Data streaming algorithms for estimating entropy of network traffic. SIGMETRICS Perform. Eval. Rev., 34(1):145–156, 2006.

[9] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, S.C. Diot. Packet-level traffic measurements from the Sprint IP backbone. IEEE Network, Nov.-Dec. 2003, 17(6):6-16.

[10] E. Kohler, J. Liy, V. Paxson, S. Shenker. Observed Structure of Addresses in IP Traffic. In Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment, 253 - 266, 2002.

[11] N. Fonseca, M. Crovella, K. Salamatian. Long range mutual information. ACM SIGMETRICS Performance Evaluation Review, Aug. 2008, 36(2): 32-37.

[12] A. Este, F. Gringoli, L. Salgarelli. On the Stability of the Information Carried by Traffic Flow Features at the Packet Level. SIGCOMM CCR, 39(3):13-18, Jul. 2009.