Quantifying Host-based Application Traffic with Multi-Scale Gamma Model

Yosuke Himura
The Univ. of Tokyo
him@hongo.wide.ad.jp

Kensuke Fukuda
NII / PRESTO, JST
kensuke@nii.ac.jp

Kenjiro Cho
IIJ
kjc@iijlab.net

Hiroshi Esaki
The Univ. of Tokyo
hiroshi@wide.ad.jp

1. INTRODUCTION

The modeling and characterization of Internet traffic is essential for simulations, traffic classification, anomaly detection and QoS. However, characterization of application traffic should be done from more various perspectives to realize practical utilization of the characteristics. For example, the following question is still open: "what kind of application should be characterized by which model?".

In this paper, to quantify behaviors of application traffic, we apply the multi-scale gamma statistical model to a massive amount of dataset. This model was originally proposed in [2] for characterizing aggregated traffic. It approximates a histogram of the number of packets during a unit time Δ as a gamma distribution, which is uniquely determined by two parameters: α and β. We aim to show the applicability of these parameters to application-based methods such as traffic classification and anomaly detection. Our preliminary findings are the following. (1) Application-specific traffic has specific ranges of α and β. For example, (a) DNS traffic shows $\alpha \in [0.1, 10]$ and $\beta \in [1, 10]$, and (b) scan traffic represents $\alpha \in [1, 50]$ and $\beta \in [0.1, 10]$. (2) α and β of legitimate events are generally stable on a link, but are highly variable with link upgrades. In contrast, anomaly events are characterized by higher α and β.

2. CHARACTERIZATION OF APPLICATION TRAFFIC WITH MULTI-SCALE GAMMA MODEL

Let Δ be the time bin to aggregate packets, and let $X_\Delta(t)$ be the number of packets which arrive during t-th Δ. The histogram of $X_\Delta(t)$ is approximated by a gamma distribution $f_{\alpha,\beta}(X) = \frac{1}{\beta\Gamma(\alpha)}\big(\frac{X}{\beta}\big)^{\alpha-1}\exp\big(-\frac{X}{\beta}\big)$, where $\Gamma(x)$ is the gamma function. α decides the shape of the histogram (e.g., the shape is exponential for $\alpha \leq 1$), whereas β determines the scale of the histogram. In this paper, we empirically set $\Delta = 1s$. Fig.1 shows an example of fitting by this model. The left figure represents time-series data; x-axis is time t, and y-axis is $X_\Delta(t)$. The right figure illustrates the histogram of $X_\Delta(t)$ (boxes) and the gamma distribution estimated by the moment method (line).

In order to characterize application traffic, we need to identify application traffic from aggregated traffic. For practical analyses, we do our analysis at the host-level by classifying specific hosts of representative applications into nine categories. We use classification heuristics based on protocol types, TCP flags, port numbers, communication network structures. The nine categories are DNS, MAIL, P2P, WWWS (www server), WWWC (www client), FLOOD (flooding attack), SCAN (scanning activity), OTHER (other identified application such as FTP and SSH), UNKNOWN (unidentified host). For the characterization of application traffic, we use first seven categories and concentrate on the hosts which have more than 1000 packets in a trace for obtaining reliable statistics.

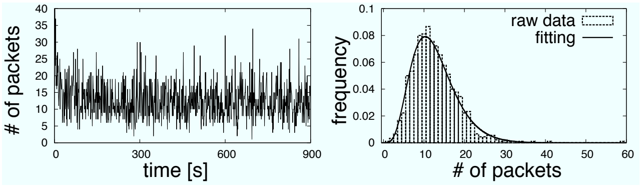

Figure 1: An example of gamma model fitting. Left: a traffic from a DNS server. Right: the histogram of $X_\Delta(t)$ and an estimated gamma distribution.

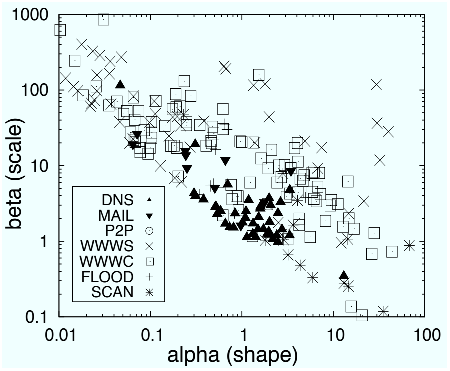

Figure 2: Correlation among application type, α and β (on 2007-09-16). Application-specific traffic has specific ranges of the parameters.

For the evaluation dataset, we use the MAWI traffic traces[1]. These real data are measured on a trans-pacific link between Japan and U.S. everyday from 2001. These data are composed of 15 minutes pcap traces (from 14:00-15:00, JST). The link was upgraded two times: from 18 Mbps to 100 Mbps (in June and July 2006), and from 100 Mbps to 150 MBps (in June 2007).

PAM2009, April 1-3, 2009, Seoul, Korea.
.

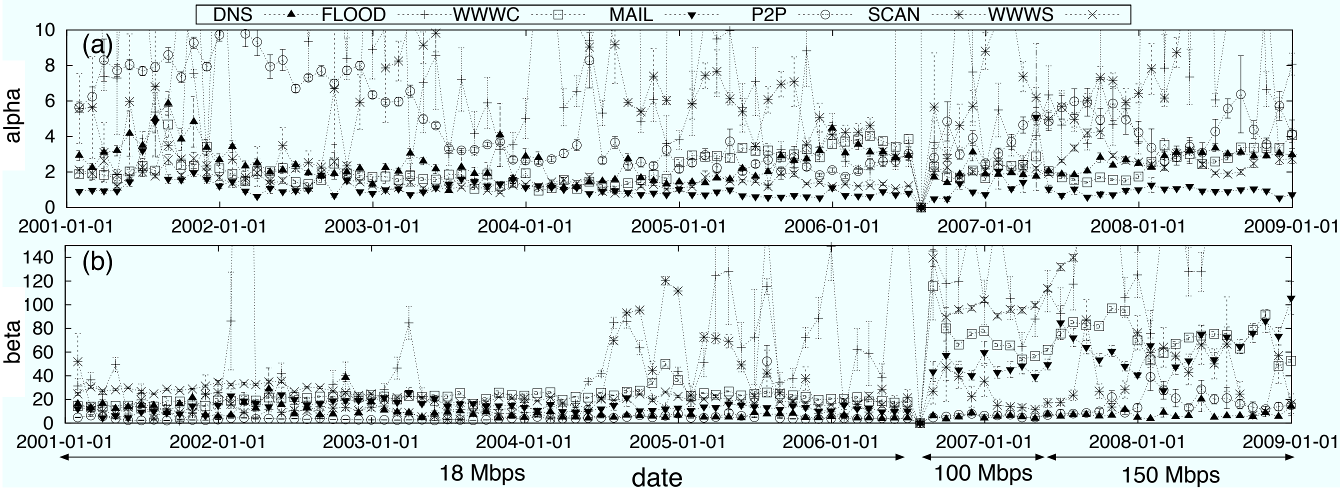

Figure 3: Time evolution of α and β (from 2001-01-01 to 2008-12-31). Each points has aggregated average value for 1 month and includes its standard error.

3. RESULTS

Correlation between application type and the parameters: Fig.2 illustrates the correlation among application type, α and β with a trace from 2007-09-16. The x-axis is α, the y-axis is β, and each symbol stands for corresponding type of application traffic. This figure indicates that type of application traffic depends on α and β, so the difference in application traffic is characterized by the multi-scale gamma model. For example, SCAN symbols are located on $\alpha \in [1, 50]$ and $\beta \in [0.1, 10]$. Thus, the scanning activity produces Gaussian and low-scale histogram of $X_{\Delta}(t)$. From time-series perspective, the scanner hosts continuously and constantly send packets. For another example, DNS symbols are concentrated on $\alpha \in [0.1, 10]$ and $\beta \in [1, 10]$. So, DNS represents various shape and scale of the histogram in this time-scale. On the other hand, the WWW hosts (WWWC and WWWS) produce the wide range of α and β. A plausible explanation of this effect is a mixture of interactive file-transfer and video streaming in the same protocol (HTTP). To summarize, α and β can be used as features for identification of malicious outlier traffic and classification of application traffic.

Time evolution of the parameters: Fig.3 represents the time evolution of α and β from 2001-01-01 to 2008-12-31. For each symbol (application traffic), the x-axis is date and the y-axis is the average value of the parameter which includes its standard error (s.e.) at the aggregation level = 1 month. When the link bandwidth was 18 Mbps, α and β are generally stable except for P2P, SCAN and FLOOD. P2P shows decreasing α, which indicates that the shape of the histogram changes from Gaussian to exponential; This implies changes in the dominant form of P2P software (e.g., from hybrid P2P to pure P2P, or from file transfer to video streaming), and does not imply upgrades of other links on the path because traditional applications (e.g., DNS) show stable α. Conversely, SCAN and FLOOD produce higher α and β. One reason of these higher values is massive outbreaks of worms; For example, the massive outbreak of Sasser worm (from June 2004 to June 2005) obviously affected β. Thus, we can detect anomaly traffic by observing α and β. On the other hand, when the 18 Mbps link was replaced with 100 Mbps one, β highly increased. Since the higher packet processing speed makes $X_{\Delta}(t)$ highly variable, the scale of the histogram should be enlarged. In contrast, β of DNS was not affected by the bandwidth. Since DNS hosts frequently send answer packets and do not transfer large-size files, the hosts constantly and continuously sends a certain amount of packets, so the histogram of $X_{\Delta}(t)$ is independent of bandwidth. In summary, α and β of legitimate events are mainly stable on a link, and become higher according to the bandwidth upgrade. Anomaly events represent high and variable α and β, and we can detect anomaly traffic by observing them.

4. SUMMARY

In this paper, we characterized application traffic by the multi-scale gamma model, and we obtained the following findings: (1) Application-specific traffic has specific ranges of α and β. (2) α and β of legitimate events are generally stable on a link, but are highly variable with link bandwidth upgrades. In contrast, anomaly events are characterized by higher α and β.

The goal of our study is to discover the characteristics which quantify differences in application traffic, and to show the applicability of the characteristics to application-based methods such as traffic classification and anomaly detection. In the future, we will improve our classification method for more detail and appropriate characterization and use traffic traces from other links, investigating the effects of Δ.

5. REFERENCES

[1] K. Cho, K. Mitsuya, and A. Kato. Traffic Data Reposity at the WIDE Project. *USENIX 2000 FREENIX Track*, June 2000.

[2] A. Scherre, N. Larrieu, P. Owezarski, P. Borgnat, and P. Abry. Non-Gaussian and Long Memory Statistical Characterisations for Internet Traffic with Anomalies. *IEEE Transactions on Dependable and Secure Computing, vol. 4, no. 1*, pages 56–70, October 2006.