LETTER Special Section on New Technologies and their Applications of the Internet

Impact of Censoring on Estimation of Flow Duration Distribution and Its Mitigation Using Kaplan-Meier-Based Method

Yuki SAKAI[†a)], *Nonmember*, Masato UCHIDA[††b)], Masato TSURU[††c)], *Members*, *and* Yuji OIE[††d)], *Fellow*

SUMMARY A basic and inevitable problem in estimating flow duration distribution arises from "censoring" (i.e., cutting off) the observed flow duration because of a finite measurement period. We extended the Kaplan-Meier method, which is used in the survival analysis field, and applied it to recover information on the flow duration distribution that was lost due to censoring. We show that the flow duration distribution from a short period of actual traffic data with censoring that was estimated using a Kaplan-Meier-based method can approximate well the flow duration distribution calculated from a sufficiently long period of actual traffic data.

key words: *flow duration distribution, censoring, Kaplan-Meier method*

1. Introduction

To achieve efficient network design, management, and control, evaluating various traffic characteristics by network measurement is necessary. Flow duration is one of the most important characteristics because it is often used for traffic classification to provide services that are secure and have a high quality in a network [1], [2]. Therefore, the characteristics of flow duration distribution have been investigated extensively. For example, the appropriateness of Pareto and log-normal distributions as a statistical model for characterizing flow duration distribution has been evaluated [3]–[6]. In addition, an active measurement method for estimating flow duration distribution has been proposed under an assumption that flow duration distribution follows a log-normal distribution [7].

A basic and inevitable problem in estimating flow duration distribution arises from "censoring" (i.e., cutting off) the observed flow duration because of a finite measurement period [6]. That is, an observed flow duration might be shorter than the true flow duration, which we do not really know, due to censoring. The impact on the flow duration distribution of information lost by censoring increases as the number of long-lived flows transmitted over the network increases. This problem has become critical because the number of elephant flows, which are defined as huge and long-lived flows, is increasing in current networks [6]. Although this problem might be solved if the measurement period was sufficiently long, the measurement cost would be staggering. Therefore, using the observed flow duration data even if it is censored is more efficient for drawing conclusions about the true flow duration.

We investigated the estimation of flow duration distribution from a short period of traffic data with censoring in the observed flow duration. We looked at the data analytic approach called the Kaplan-Meier method, which can take into account censored data. The Kaplan-Meier method is a non-parametric statistical procedure used in the survival analysis field for analyzing censored data for which the outcome variable of interest is the time of an event. By event, we mean any designated experience of interest that may happen to an individual, such as birth or death. In the context of flow duration, the birth and death time correspond to the start and end time of a flow, respectively. We show that the flow duration distribution estimated from a short period of traffic data with censoring can approximate well the flow duration distribution calculated from a sufficiently long period of traffic data.

This paper is organized as follows. In Sect. 2, we explain the classification of flows and our method for estimating flow duration distribution based on the Kaplan-Meier method. In Sect. 3, we first show the impact of censoring on estimating flow duration distribution. Then, we show the accuracy of estimating flow duration distribution using our Kaplan-Meier-based method. Section 4 is the conclusion.

2. Analysis Method

2.1 Classification of Flows

We define "flow" as a set of packets identified by a five tuple (flow key) consisting of source IP address, destination IP address, source port number, destination port number, and protocol field value. However, if two succeeding packets with the same flow key are separated by a time gap exceeding a pre-defined timeout period δ, they are considered to belong to separate flows. We use $\delta = 15$ [sec] considering into the default value of TCP connection timeout. In addition, we eliminate the flows that are composed of only one packet, such as ARP, ICMP, DNS, and NTP packets, from the objective of analysis.

For each flow f, let us denote *the start time of observation* by $t_s(f)$ and *the end time of observation* by $t_e(f)$, where $t_s(f)$ and $t_e(f)$ are determined as follows (see Fig. 1).

Manuscript received February 18, 2009.

Manuscript revised May 25, 2009.

[†]The author is with the Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, Iizuka-shi, 820–8502 Japan.

[††]The authors are with Network Design Research Center, Kyushu Institute of Technology, Iizuka-shi, 820–8502 Japan.

a) E-mail: yuki@infonet.cse.kyutech.ac.jp
b) E-mail: m.uchida@ndrc.kyutech.ac.jp
c) E-mail: tsuru@ndrc.kyutech.ac.jp
d) E-mail: oie@cse.kyutech.ac.jp

DOI: 10.1587/transinf.E92.D.1949

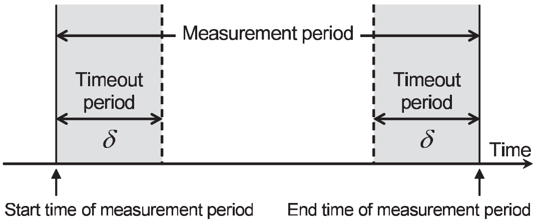


Fig. 1 Condition of observation.

- If a SYN (FIN) packet of flow f is observed during the measurement period, then the timestamp attached to this observed packet is set to $t_s(f)$ ($t_e(f)$). In this case, we consider that *the start (end) time of the flow* has been *detected*.
- Otherwise, the timestamp attached to the firstly (lastly) observed packet of flow f is set to $t_s(f)$ ($t_e(f)$). In this case, if the first (last) packet of flow f is observed during the headed (tailed) timeout period, we consider that *the start (end) time of the flow* has "not" been *detected*. Otherwise, we consider that *the start (end) time of the flow* has been *detected*.

Here, the observed flow duration of flow f is defined as $d(f) = t_e(f) - t_s(f)$.

In consideration of the above conditions for flow observation, flows are classified as follows.

- Flows of which both start and end time are detected. These are referred to as non-censored (*nc*) flows. The set of *nc* flows observed during the measurement period is denoted by $\mathcal{F}_{nc}$.
- Flows of which either start or end time is detected. These are referred to as one-side-censored (*oc*) flows. The set of *oc* flows observed during the measurement period is denoted by $\mathcal{F}_{oc}$. The *oc* flows are divided into right-censored (*rc*) and left-censored (*lc*) flows. The *rc* (*lc*) flows are those of which the start (end) time is detected but the end (start) time is not.
- Flows of which neither start nor end time is detected. These are referred to as both-side-censored (*bc*) flows. The set of *bc* flows observed during the measurement period is denoted by $\mathcal{F}_{bc}$.

The union set of $\mathcal{F}_{nc}$, $\mathcal{F}_{oc}$, and $\mathcal{F}_{bc}$ is denoted by $\mathcal{F}_{nc+oc+bc}$ (i.e., $\mathcal{F}_{nc+oc+bc} = \mathcal{F}_{nc} \cup \mathcal{F}_{oc} \cup \mathcal{F}_{bc}$), which is equivalent to the set of all flows observed during the measurement period. Note that the sets $\mathcal{F}_{nc}$, $\mathcal{F}_{oc}$, and $\mathcal{F}_{bc}$ are disjointed from each other.

2.2 Application of Kaplan-Meier Method

Consider a random variable D representing flow duration. The distribution function of D is defined as $F(t) = \Pr\{D \leq t\}$, and the complementary cumulative distribution (ccd) function (or survival function) of D is defined as $S(t) = 1 - F(t) = \Pr\{D > t\}$.

The Kaplan-Meier method is a non-parametric statistical procedure for estimating the ccd function $S(t)$ [8]. That is, the Kaplan-Meier method does not assume the model of underlying true distribution. If we follow the original Kaplan-Meier method faithfully, only the *nc* and *rc* flows should be used to estimate the ccd function because only the forward time direction is considered in the original method. However, both forward and backward time directions are considered in our analysis because the difference of time direction does not have any particular meaning in the context of flow duration. This is also because we expect that we can fully make use of the information about flow duration included in the measurement data by considering both time directions. Therefore, the *nc*, *rc*, and *lc* flows are used to estimate the ccd function of flow duration. Here, note that the *nc* flows have to be duplicated two-fold because they are analyzed in both forward and backward time directions. The set of duplicated *nc* flows is denoted by $\mathcal{F}_{dnc}$, where $|\mathcal{F}_{dnc}| = 2|\mathcal{F}_{nc}|$ holds. The union set of $\mathcal{F}_{dnc}$ and $\mathcal{F}_{oc}$ is denoted by $\mathcal{F}_{dnc+oc}$ (i.e., $\mathcal{F}_{dnc+oc} = \mathcal{F}_{dnc} \cup \mathcal{F}_{oc}$).

In our analysis, we use the estimated ccd function of flow duration, which is based on the Kaplan-Meier method, as follows:

$$S_{KM}(t) = \begin{cases} 1, & \text{if } t < t_{\min}, \\ \displaystyle\prod_{\{d(f) | d(f) \leq t, f \in \mathcal{F}_{dnc}\}} \frac{n(d(f)) - m(d(f))}{n(d(f))}, & \text{if } t \geq t_{\min}, \end{cases}$$

where

$$t_{\min} = \min\{d(f) | \ f \in \mathcal{F}_{dnc}\},$$
$$n(x) = |\{f | \ x \leq d(f), \ f \in \mathcal{F}_{dnc+oc}\}|,$$
$$m(x) = |\{f | \ x = d(f), \ f \in \mathcal{F}_{dnc}\}|,$$

and $d(f)$ denotes the duration of flow f. The original Kaplan-Meier method is given by replacing $\mathcal{F}_{dnc}$ and $\mathcal{F}_{dnc+oc}$ with $\mathcal{F}_{nc}$ and $\mathcal{F}_{nc+rc}$ (= $\mathcal{F}_{nc} \cup \mathcal{F}_{rc}$), respectively, in the above formulation. Although we omit the details, the stability of the estimation result obtained by our method is much better than that by the original method in the analysis of actual traffic data.

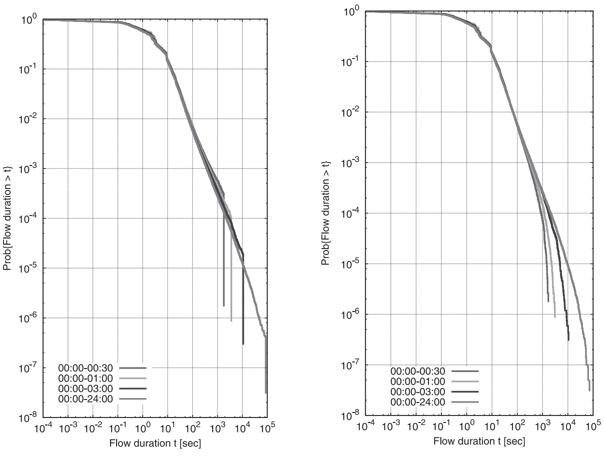
3. Analysis of Actual Traffic Data

3.1 Traffic Data

In our analysis, we used one-way traffic traces provided by the Widely Integrated Distributed Environment (WIDE) project [9]. The traces were measured on the trans-Pacific line and were available from the Measurement and Analysis on the WIDE Internet (MAWI) traffic archive [10]. The measurement line was 100 Mbps. More detailed information about the traces is given in [9], [10]. In the following section, we use the 24-hour trace measured on June 10, 2007.

3.2 Impact of Censoring

Evaluation results of the impact of censoring on estimating



(a) $S_{\mathcal{F}_{nc+oc+bc}}(t)$

(b) $S_{\mathcal{F}_{nc}}(t)$

Fig. 2 Comparison between $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$.

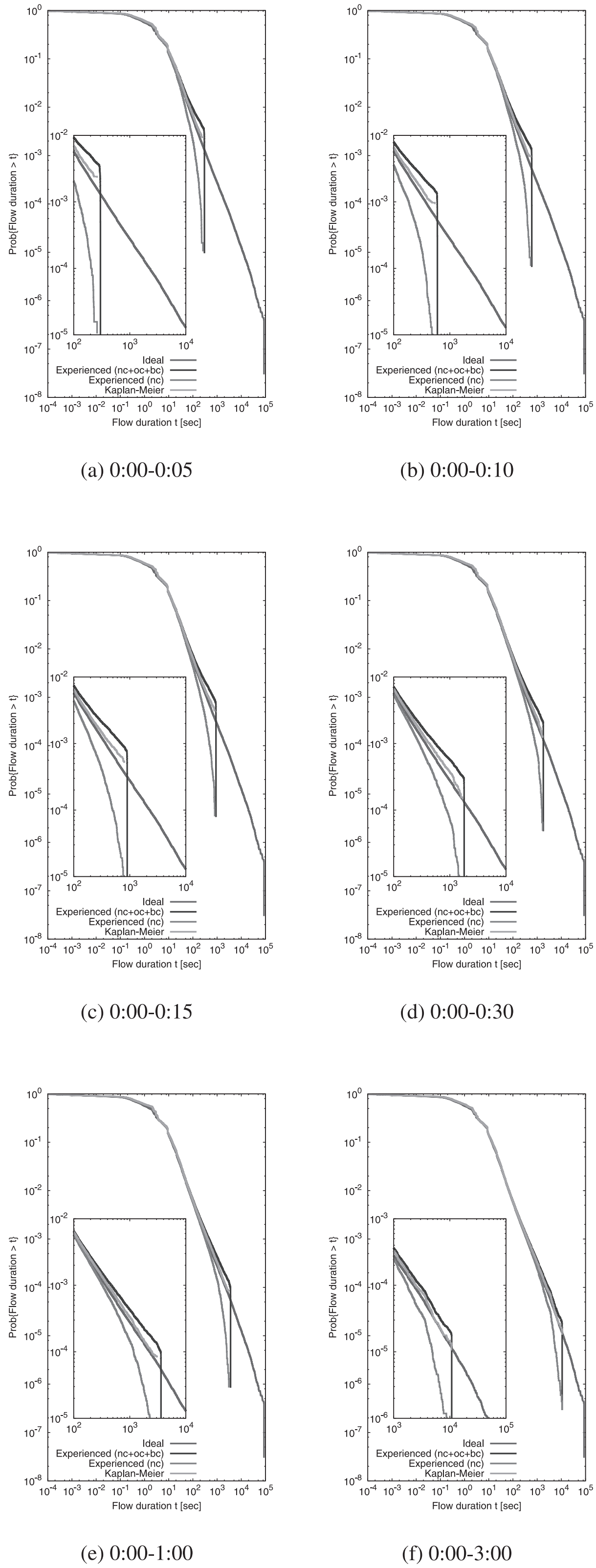
flow duration distribution are provided in this section. We consider two kinds of simple experienced ccd functions of flow duration that are defined as follows:

$$S_{\mathcal{F}_{nc+oc+bc}}(t) = \frac{1}{|\mathcal{F}_{nc+oc+bc}|} \sum_{f \in \mathcal{F}_{nc+oc+bc}} I(d(f) > t),$$

$$S_{\mathcal{F}_{nc}}(t) = \frac{1}{|\mathcal{F}_{nc}|} \sum_{f \in \mathcal{F}_{nc}} I(d(f) > t),$$

where $I(\cdot)$ is an indicator function taking a value of one when the expression in the parentheses is true and zero otherwise. It is important to note that both $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$ have drawbacks as well. That is, the experienced ccd function $S_{\mathcal{F}_{nc+oc+bc}}(t)$ would not be accurate because the set $\mathcal{F}_{nc+oc+bc}$ includes the censored flows that have durations shorter than the true value. The experienced ccd function $S_{\mathcal{F}_{nc}}(t)$ would also not be accurate because the set $\mathcal{F}_{nc}$ does not include the censored flows that might have some information about flow duration distribution. Note that $S_{KM}(t)$, $S_{\mathcal{F}_{nc+oc+bc}}(t)$, and $S_{\mathcal{F}_{nc}}(t)$ coincide with each other if both $\mathcal{F}_{oc}$ and $\mathcal{F}_{bc}$ are empty.

The experienced ccd functions $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$, where the measurement periods are 0:00–0:30, 0:00–1:00, 0:00–3:00, and 0:00–24:00 of June 10, 2007, respectively, are shown in Figs. 2 (a) and 2 (b). The upper end of $S_{\mathcal{F}_{nc+oc+bc}}(t)$ in Fig. 2 (a) is sharply skewed compared with that of $S_{\mathcal{F}_{nc}}(t)$ in Fig. 2 (b). This is because $S_{\mathcal{F}_{nc+oc+bc}}(t)$ includes the censored flows that cannot last longer than a measurement period. However, from these figures, we can see that both $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$ converge to the same ccd function, which is believed to be the true ccd function, as the length of the measurement period increases. In other words, the impact of censoring increases as the length of the measurement period decreases. These results show that a sufficiently long measurement period is needed to estimate the true ccd function based on $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$.



(a) 0:00-0:05

(b) 0:00-0:10

(c) 0:00-0:15

(d) 0:00-0:30

(e) 0:00-1:00

(f) 0:00-3:00

Fig. 3 Comparison of $S_{KM}(t)$ (Kaplan-Meier), $S_{\mathcal{F}_{nc+oc+bc}}(t)$ (Experienced (nc+oc+bc)) and $S_{\mathcal{F}_{nc}}(t)$ (Experienced (nc)) for various measurement periods. As a basis for comparison, $S_{\mathcal{F}_{nc+oc+bc}}(t)$ for measurement period of 0:00–24:00 is plotted (Ideal). Magnified plots are given inside each figure.

3.3 Estimation of Flow Duration Distribution

In the previous section, we found that the simple experienced flow duration distribution $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$ can not work well if the measurement period is short. Therefore, it is necessary to consider an estimation method that does not need a long measurement period, the cause of high measurement cost. In this section, we demonstrate that using our Kaplan-Meier-based method to estimate flow duration distribution is much more accurate than using the conventional method that uses $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$ for the estimation even if the measurement period is short.

Figures 3 (a), 3 (b), 3 (c), 3 (d), 3 (e), and 3 (f) show the estimated ccd functions based on the Kaplan-Meier method $S_{KM}(t)$ and the experienced ccd functions $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$, where the measurement periods are 0:00–0:05, 0:00–0:10, 0:00–0:15, 0:00–0:30, 0:00–1:00, 0:00–3:00 of June 10, 2007, respectively. In these figures, the experienced ccd functions $S_{\mathcal{F}_{nc+oc+bc}}(t)$ for the measurement period of 0:00–24:00 are plotted as a basis for comparison.

As shown in these figures, the accuracy of the flow duration distribution of $S_{KM}(t)$ is better than that of $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$. The above results mean that the Kaplan-Meier method can be applied for estimating flow duration distribution by using short periods of traffic data with censoring. Although the improvement to $S_{\mathcal{F}_{nc+oc+bc}}(t)$ and $S_{\mathcal{F}_{nc}}(t)$ in estimation accuracy given by the Kaplan-Meier method (i.e., $S_{KM}(t)$) seems to be small, we believe this improvement can be significant in precise resource management. For example, the accurate knowledge about the flow duration distribution is of importance to design an appropriate size of flow management table that would be used in various flow-based traffic engineering mechanisms especially on very high-speed networks such as Carrier Grade NAT (CGN) [11] and Deep Packet Inspection (DPI).

In this paper, the analysis results on one data set from MAWI [10] were presented as an example due to space limitation. However, the authors have confirmed the similar results on several data sets from MAWI.

4. Conclusion

We first showed that the impact of censoring on estimating flow duration distribution is not negligible if the measurement period is short. Then, we showed that estimating flow duration distribution using a Kaplan-Meier-based method is much more accurate than using the conventional method based on the experienced flow duration distribution even if a short period of traffic data with censoring is used for the analysis. The application of the Kaplan-Meier method to other flow statistics will be considered in our future work.

Acknowledgments

This work was partly supported by the National Institute of Information and Communications Technology, and the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (S) (No. 18100001).

References

[1] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, “Class of service mapping for QoS: A statistical signature-based approach to IP traffic classification,” Proc. Internet Measurement Conference (IMC’04), pp.135–148, Sicily, Italy, Oct. 2004.

[2] A. Lakhina, M. Crovella, and C. Diot, “Characterization of network-wide anomalies in traffic flows,” Proc. Internet Measurement Conference (IMC’04), pp.201–206, Sicily, Italy, Oct. 2004.

[3] S. Ata, M. Murata, and H. Miyahara, “Analysis of network traffic and its application to design of high-speed routers,” IEICE Trans. Inf. & Syst., vol.E83-D, no.5, pp.988–995, May 2000.

[4] J. Charzinski, “Internet client traffic measurement and characterisation results,” Proc. 13th International Symposium on Services and Local Access (ISSLS’00), Stockholm, Sweden, June 2000.

[5] A.B. Downey, “Evidence for long-tailed distributions in the Internet,” Proc. ACM SIGCOMM Internet Measurement Workshop (IMW’01), pp.229–241, San Francisco, CA, USA, Nov. 2001.

[6] T. Mori, M. Uchida, and S. Goto, “Flow analysis of Internet traffic: World wide Web versus peer-to-peer,” Systems and Computers in Japan, vol.36, no.11, pp.70–81, 2005.

[7] T. Asaka, K. Ori, and H. Yamamoto, “Method of estimating flow duration distribution using active measurement,” IEICE Trans. Commun., vol.E86-B, no.10, pp.3030–3038, Oct. 2003.

[8] E.L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” J. American Statistical Association, vol.53, no.282, pp.457–481, 1958.

[9] “Widely integrated distributed environment project.”
http://www.wide.ad.jp/

[10] “Measurement and analysis on the WIDE Internet.”
http://tracer.csl.sony.co.jp/

[11] T. Nishitani and S. Miyakawa, “Carrier grade network address translator (NAT) behavioral requirements for unicast UDP, TCP and ICMP.”
http://tools.ietf.org/id/draft-nishitani-cgn-00.txt