Estimating Original Flow Length from Sampled Flow Statistics

Weijiang Liu, Jian Gong, Wei Ding, and Yanbing Peng

Department of Computer Science and Engineering,
Southeast University, 210096 Nanjing, Jiangsu, China
{wjliu, jgong, wding, ybpeng}@njnet.edu.cn

Abstract. Packet sampling has become an attractive and scalable means to measure flow data on high-speed links. Passive traffic measurement increasingly employs sampling at the packet level and makes inferences from sampled network traffic. This paper proposes a maximum probability method that estimates the length of the corresponding original flow from the length of a sampled flow. We construct the probability models of the original flow length distributions of a sampled flow under the assumptions of various flow length distributions, respectively. Through recovery analyzing with different parameters, we obtain a consistent linear expression that reflects the relationship between the length of sampled flow and that of the corresponding original flow. Furthermore, after using publicly available traces and traces collected from CERNET to do recovery experiments and comparing the experiment outcomes and theoretic values calculated with Pareto distributions, we may conclude that the maximum probability method calculated by using the Pareto distribution with 1.0 can be used to estimate original flow length in the concerned network.

1 Introduction

With the rapid increase of network service types and user number, measuring the volume of network traffic and network performance is becoming more difficult. Packet sampling has become an attractive and scalable means to measure flow data. Sampling entails an inherent loss of information. There are some studies in [1-4] that use statistic inference to recover information as much as possible. However, to our knowledge the above studies are not available to estimate length of the original flow by the given sampled flow length. The original flow length is very important for many applications, e.g., Resources Required for Collecting Flow Statistics, Characterizing Source Traffic and Determining thresholds for setting up connections in flow-switched networks. This paper proposes a Maximum Probability (MP) method that estimates the length of the corresponding original flow from the length of a sampled flow. This method is simple to calculate and easy to implement. The main contributions of this paper include:

1) Maximum Probability (MP) method that estimates the length of the corresponding original flow from the length of a sampled flow is proposed.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part IV, LNCS 3994, pp. 120–127, 2006.

2) A linear expression that reflects the relationship between the length of sampled flow and that of the corresponding original flow is obtained. Although they are different for different Pareto parameters, the differences are very small.

3) We conclude that the value calculated by using the Pareto distribution with 1.0 can be used to estimate original flow length in the concerned network.

The rest of this paper is organized as follows. In the next section, we review some elementary concepts on flow and sampling. In Section 3 we construct the probability models of the original flow length distributions of a sampled flow under the assumptions of various flow length distributions, respectively. In Section 4, we propose the Maximum Probability (MP) method and compare the estimating results of uniform distribution, Pareto distributions and empirical distributions. We conclude with some a summary our contributions in Section 5.

2 Some Elementary Concepts

There are a number of different ways to implement this, e.g., independent sampling of packets with probability $p = 1/N$, and periodic selection of every Nth packet from the full packet stream. In both cases we will call N the sampling period, i.e., the reciprocal of the average sampling rate. Although the length distributions by random and periodic sampling can be distinguished, the differences are, in fact, sufficiently small [2].

Definition 1. A flow is defined as a stream of packets subject to flow specification and timeout.

When a packet arrives, the specific rules of flow specification determine which active flow this packet belongs to, or if no active flow is found that matches the description of this packet, a new flow is created. In this paper, the flow interpacket timeout is 64 seconds. A general flow is a stream of packets subject to timeout and having the same source and destination IP addresses, same source and destination port numbers (not considering protocol). In this paper, we will use the term original flow to describe the above flow. A flow length is the number of packets in the flow.

Definition 2. A sampled flow is defined as a stream of packets that are sampled at probability $p = 1/N$ from an original flow.

3 Probability Distribution of Original Flow Length

In this paper, sampling probability is $p = 1/N$. For a specific original flow F, let X_F denote the number of packets in F, Y_F denote the number of packets in the sampled flow from F. The conditional distribution of Y_F, given that $X_F = l$, follows a binomial distribution

$$Pr[Y_F = k | X_F = l] = B_p(l, k) = \binom{l}{k} p^k (1-p)^{l-k}.$$

For an original flow F, let $Pr[Y_F = y, X_F = x]$ denote the probability of $X_F = x$ and $Y_F = y$, by the conditional probability formula,

$$Pr[X_F = x|Y_F = y] = \frac{Pr[Y_F = y, X_F = x]}{Pr[Y_F = y]} = \frac{Pr[Y_F = y|X_F = x]Pr[X_F = x]}{Pr[Y_F = y]} \tag{1}$$

and by the complete probability formula, we obtain:

$$Pr[Y_F = y] = \sum_{i=y}^{\infty} Pr[Y_F = y|X_F = i]Pr[X_F = i] = \sum_{i=y}^{\infty} B_p(i, y)Pr[X_F = i] \tag{2}$$

3.1 Uniform Distribution

Suppose original flows lengths satisfy uniform distribution, that is, $Pr[X_F = k] = Pr[X_F = k + 1]$, for all $k = 1, 2, \cdots$. Moreover,

$$\sum_{i=k}^{\infty} B_p(l, k) = \sum_{i=k}^{\infty} \binom{l}{k} p^k (1-p)^{l-k} = 1/p = N$$

hence $Pr[Y_F = y] = \sum_{i=y}^{\infty} B_p(i, y)Pr[X_F = i] = Pr[X_F = y] \sum_{i=y}^{\infty} B_p(i, y) = Pr[X_F = y]/p$, so $Pr[Y_F = y|X_F = x] = pB_p(x, y)$ and we obtain the following results:

Lemma 1. *The probability that a sampled flow of length k is sampled from an original flow of length l is $Pr[X_F = l|Y_F = k] = \binom{l}{k} p^{k+1}(1-p)^{l-k}, l = k, k + 1, \cdots$.*

Let $a_1 = \frac{B_p(l,k)}{B_p(l-1,k)} = 1 + \frac{kN+1-l}{(l-k-1)N}$. For $l \leq kN$, since $a_1 > 1$, hence $B_p(l, k)$ is increasing as l increases. For $l > kN+1$, since $a_1 < 1$, hence $B_p(l, k)$ is decreasing as l increases. At $l = kN+1$, $a_1 = 1$ means that $B_p(l, k)$ is maximized at $l = kN$ and $l = kN + 1$. We have

Lemma 2. *The probability $Pr[X_F = l|Y_F = k]$ is maximized at $l = kN, kN+1$. It is increasing as l increases for $l < kN + 1$ and decreasing as l increases for $l > kN + 1$.*

3.2 Pareto Distribution

Assume original flow lengths satisfy Pareto distribution. Its probability mass function is

$$Pr[X_F = x] = \beta\alpha^{\beta}/x^{\beta+1}, \ \alpha, \beta > 0, \ x \geq \alpha \tag{3}$$

where β is called Pareto parameter. Hence Equation (2) can be written as:

$$Pr[Y_F = y] = \sum_{i=y}^{\infty} B_p(i, y)\beta\alpha^{\beta}/i^{\beta+1}, \ y \geq \alpha$$

Lemma 3. *Under the assumption that original flow lengths satisfy Pareto distribution, the probability that a sampled flow of length $y(\geq \alpha)$ is sampled from an original flow of length x is*

$$Pr[X_F = x|Y_F = y] = \frac{B_p(x,y)/x^{\beta+1}}{\sum_{i=y}^{\infty} B_p(i,y)\beta\alpha^{\beta}/i^{\beta+1}}.$$

4 Maximum Probability Method (MP)

The purpose of the Maximum Probability method (MP) is to estimate the length of the corresponding original flow from the length of a sampled flow. The point we are trying to make is that MP estimates the length of the original flow according to maximum probability. MP contains three steps:

i) Computing probability. Given a sampled flow with fixed length k, compute $Pr[X_F = l|Y_F = k]$ for $l = k, k+1, \cdots$.

ii) Finding maximum probability. Define $mp = \max_{l \geq k}\{Pr[X_F = l|Y_F = k]\}$.

iii) Estimating length. Let $\overline{l} = \min_{l}\{mp = Pr[X_F = l|Y_F = k]\}$, we write our estimate of the length of the original flow as $\overline{l}$.

Below we apply the MP method to different flow length distributions.

4.1 Uniform Distribution

Let the length of original flows be uniform distribution. By Lemma 2, we can use MP to obtain a linear expression as

$$\overline{X}_F = \frac{Y_F}{p} \tag{4}$$

where $\overline{X}_F$ is the estimate of the length of the original flow, Y_F is the length of the sampled flow, p is sampling probability. The estimates of the original flow lengths for $p = 0.1, 0.05, 0.01$ are shown in Figure 1(a). From this figure we can observe the linear relationship between the length of original flow and that of sampled flow under uniform distribution.

4.2 Pareto Distribution

Let the length of original flows be Pareto distribution. By Lemma 3, we can compute the results with $\beta = 0.5, 0.75, 1.0, 1.5$, respectively, and obtain an expression that reflect the relationship between the estimated length of the original flow and the sampled flow length as

$$\overline{X}_F = \frac{Y_F}{p} - n(p,\beta)\ for\ Y_F \geq 1/p \tag{5}$$

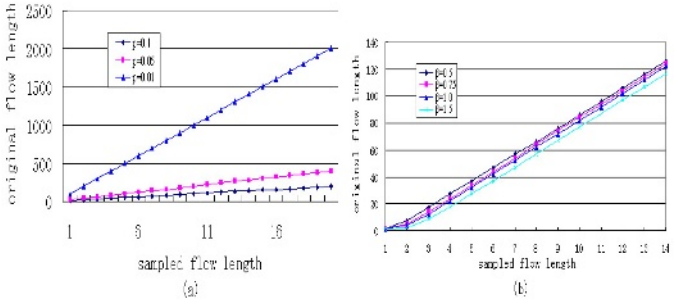

Fig. 1. (a) MP estimate of original flow length under uniform distribution. (b) MP estimate of original flow length under Pareto distribution.

where $\overline{X}_F$ is the estimate of the length of the original flow, Y_F is the length of the sampled flow, p is sampling probability, $n(p, \beta)$ is a positive integer involving p and β as defined in following subsection. In Figure 1(b), with sampling probability $p = 0.1$, and Pareto parameter $\beta = 0.5, 0.75, 1.0, 1.5$, respectively, we can observe that, for a sampled flow with fixed length Y_F, $\overline{X}_F$ is increasing as β decreases. It is minimized at $\beta = 1.5$, maximized at $\beta = 0.5$, medial at $\beta = 1.0$. Though there are differences for different parameter, the estimates show very similar tendencies for all parameters. Therefore, we may conclude that $\overline{X}_F$ calculated by using the parameter $\beta = 1.0$ can be used as approximations under unknown parameter values in the concerned network.

4.3 Finite Pareto Distribution

In the concerned network, due to constraint of measurement time, the lengths and numbers of flows all are finite. Let M denote the maximum original flow length in an original flow distribution, the probability density function in Equation (3) is

$$Pr[X_F = x] = \gamma / x^{\beta+1}, \;\; x = \alpha, \cdots, M. \tag{6}$$

where $\gamma = \frac{\beta\alpha^\beta}{\sum_{x=\alpha}^{M} \beta\alpha^\beta / x^{\beta+1}}$, $0 < \gamma < 1$. Equation (6) can be extended so that β can be extended to real field, that is, β can be zero and negative. We call Equation (6) extended Pareto distribution. For $\beta = 0$, Equation (6) is

$$Pr[X_F = x] = \gamma / x, \;\; x = \alpha, \cdots, M. \tag{7}$$

For $\beta = -1$, Equation (6) is

$$Pr[X_F = x] = \gamma, \;\; x = \alpha, \cdots, M. \tag{8}$$

Equation (8) is uniform distribution, therefore we call uniform distribution as degenerate Pareto distribution.

We now consider how the finite constraint impacts the conditional probability. Due to the constraint of finite number of flows, the conditional probability in

Lemma 3 is written by

$$Pr_M[X_F = x|Y_F = y] = \frac{B_p(x,y)/x^\beta}{\sum_{i=y}^{M} B_p(i,y)/i^{\beta+1}} \tag{9}$$

Obviously $Pr_M[X_F = x|Y_F = y] = \rho(y)Pr[X_F = x|Y_F = y]$, where $\rho(y) = \frac{\sum_{i=y}^{M} B_p(i,y)/i^{\beta+1}}{\sum_{i=y}^{\infty} B_p(i,y)/i^{\beta+1}} < 1$ is a function with variable y. Hence, for a fixed y, the above two probabilities are maximized at the same x. Therefore Equations (4) and (5) are still valid, we rewrite them as consistent form

$$\overline{X}_F = \frac{Y_F}{p} - n(p,\beta) \ \ for \ Y_F \geq 1/p \tag{10}$$

where $\overline{X}_F$ is the estimate of the length of the original flow, Y_F is the length of the sampled flow, p is sampling probability, $n(p,\beta)$ is a binary function with variables p and β whose value domain is integer set. Here $n(p,\beta)$ is a binary function with variables p and β whose value domain is integer set. Function $n(p,\beta)$ has the following properties:

1) It is a monotone decreasing function on variable p, that is , for fixed β, is decreasing as p increases.

2) It is a monotone increasing function on variable β, that is , for fixed p, is increasing as β increases.

3)$n(p,-1) = 0$, for any $p(0 < p < 1)$

For example, $n(0.1,-1) = 5$, $n(0.1,0) = 10$, $n(0.1,0.5) = 14$, $n(0.1,1.0) = 18$, $n(0.1,1.25) = 21, n(0.1,1.5) = 23, n(0.1,2.0) = 27, n(0.1,3.0) = 36$.

4.4 Comparing for Different Distributions

We use six traces to verify the MP method. The first three traces [5], all containing packets during a 10 minute period, were collected with a Dag3.2E 10/100 MBit/sec Ethernet card at the outside of the firewall servicing researchers at Bell Labs via a 9 MBits/sec link to the Internet in May 2002. The other three Traces, either of which contains packets during a 10-minute period too, were collected at Jiangsu provincial network border of China Education and Research Network (CERNET) in disjoint time interval on April 17, 2004. The backbone capacity is 1000Mbps; mean traffic per day is 587 Mbps. For each trace, we sample at $p = 0.1, 0.05, 0.01$ respectively. Then we use MP to estimate the length of original flow. We find the estimated lengths are very close at same sampling probability for the six traces. For clear display, we only show the estimates in three experiments at sampling probability $p = 0.1$ in Figure 2(a). As shown in Figure 2(a), the estimates are very close.

Figure 2(b) illustrates the estimates for uniform distribution, Pareto distribution ($\beta = 1.0$) and Experiment 1(For first trace, sampling at $p = 0.1$, then we use

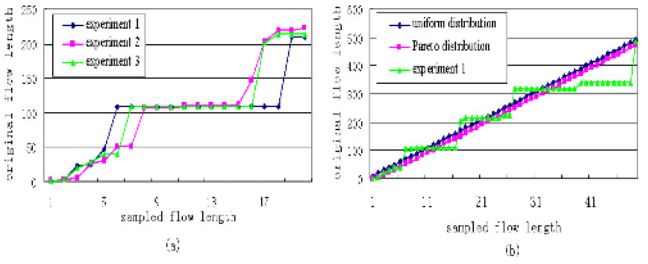

Fig. 2. (a) MP estimate of original flow length under empirical distribution. (b) Comparing MP estimates of original flow length under different distributions.

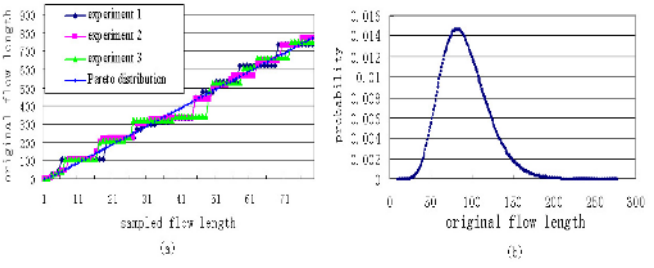

Fig. 3. (a): Comparing MP estimates of original flow length under empirical distributions and Pareto distribution. (b) Probability distribution of original flow length, given length 10 of sampled flow at sampling probability p=0.1.

MP to estimate the length of original flow. Similarly Experiment 2 for the third trace, Experiment 3 for the fifth trace). We can observe that the estimates for uniform are slightly big but the estimates for Pareto distribution with $\beta = 1.0$ fit those for Experiment 1 very well. From Figure 3(a), we can see that the estimate results of several experiments all move up and down at the point of estimate for Pareto distribution $\beta = 1.0$. Therefore we can use the Pareto distribution with $\beta = 1.0$ as theoretical distribution of the three empirical distributions.

4.5 Difficulties and Applications

When we use MP to estimate the original flow length of a sampled flow, we must find the one that makes the probability maximized. However, the maximized probability sometimes is still very small. For example, consider a sampled flow with length 10 at sampling probability 0.1. We calculate the probability by using Pareto distribution with 1.0 to estimate its original flow length. Figure 3(b) displays the probability distribution of the original flow length of the sampled flow. Within it we can observe that the probability at length 82 is maximized with value 0.0147. Although 0.0147 is maximum value, it is too small. It reflects the difficulty and uncertainty associated with an estimate. To improve certainty, we can estimate the confidence interval for original flow length subject to interval width (as small as possible). Suppose that the estimated length may fall into an

interval with width 70, we can sum for some values and obtain a maximum probability $Pr[51 \leq x \leq 120|y = 10] = 0.800017$. Therefore, we can use MP to estimate the boundary of length for a specific flow.

5 Conclusions

This paper proposes a naive method (MP) to estimates original flow length from the sampled flow. For different Pareto parameters, we obtain a consistent linear expression that reflects the relationship between the length of sampled flow and that of the corresponding original flow. In the concerned network, the length distributions of flows collected in any time interval do not satisfy Pareto distributions with fixed parameter strictly, but they can follow a Pareto distribution with parameter in interval [0.5, 1.5] approximately. The value 1.0 is the middle value of interval [0.5,1.5] exactly. Theory analysis and experiment results show that it is a reasonable choice using parameter 1.0 to calculate at the condition of unknown parameter value.

Acknowledgement

This work is supported in part by the National Grand Fundamental Research 973 Program of China under Grant No.2003CB314804; the National High Technology Research and Development Program of China (2005AA103001); Jiangsu Planned Projects for Postdoctoral Research Funds.

References

1. Duffield, N.G., Lund, C. , Thorup, M.: Properties and Prediction of Flow Statistics from Sampled Packet Streams. ACM SIGCOMM Internet Measurement Workshop 2002,159-171, November 2002.
2. Duffield, N.G., Lund, C. , Thorup, M.: Estimating Flow Distributions from Sampled Flow Statistics. IEEE/ACM Transation on Networking, **13**(2005) 325-336.
3. Tatsuya Mori, Masato Uchida, Ryoichi Kawahara: Identifying Elephant Flows Through Periodically Sampled Packets. ACM SIGCOMM Internet Measurement Conference 2004,115-120.
4. Noriaki Kamiyama: Identifying High-Rate Flows With Less Memory. IEEE Infocom 2005,2781-2785 ,March 2005.
5. NLANR: Abilene-I data set,http://pma.nlanr.net/Traces/long /bell1.html.