A Hybrid Sampling Approach for Network Flow Monitoring

Guang Cheng[1], Jian Gong[1], Yongning Tang[2]

[1] College of Computer Science & Engineering, Southeast University, Nanjing, P.R.China, 210096
[2] DePaul University's School of Computer Science, Telecommunications and Information Systems, Chicago IL USA 60604
gcheng@njnet.edu.cn

Abstract

Online flow distribution monitoring is critical in intrusion detection. However, high-speed traffic monitoring is significantly challenging for a monitoring system with limited resources (e.g., memory and processing cycles). *Flow and packet sampling techniques are commonly adopted to tackle this problem. Flow sampling can reduce the variance of the estimators in short flows; However, it increases the estimated error for the heavy-tailed flow. On the other hand, passive sampling presents an opposite results. In this paper, we propose a novel flow sampling approach by taking advantage of both packet and flow sampling techniques. An effective flow estimator is also introduced to estimate flow distributions. Extensive simulations are conducted with real traffic data from CERNET backbone network traffic traces to evaluate the system performance and compare it with other traffic sampling approaches.*

Keywords: *Packet Sampling, Flow Sampling, Hybrid Sampling, Flow Distributions.*

1 Introduction

It is important to measure flow distribution online in intrusion detection. Flow information (e.g., the number of flows and the length of each flow) presents the critical information in detecting potential security threats such as traffic worms and DDoS attacks. However, high-speed traffic monitoring is significantly challenging for a monitoring system with limited resources (e.g., memory and processing cycles) several sampling methods have been proposed to control the resource consumption of the monitoring system. Packet sampling and flow sampling are the two most commonly adopted approaches.

Sampling entails an inherent loss of information, and statistical inference is commonly adopted to recover the lost information. In a packet sampling approach, if the packet sampling ratio is p, then the probability that the flow with the length of k is sampled is $1-(1-p)^k$, which shows that the probability of the sampled heavy-tailed flows is larger than that of the short flows, and the probability of a sampled packet also influences the probability of the sampled flows. The packet sampling method will lose a significant amount of short flow data, but retains the heavy-tailed flow data. Thus, the estimated error of short flows is larger than that of the heavy-tailed flows. Further, it is difficult to estimate the number of original flows. On the other hand, in a flow sampling approach, since each flow has an equal probability of occurrence q, the long flows have the same probability as short flows. The number of the heavy-tailed flows is very small, so the probability of the heavy-tailed flows being sampled is also very small, and their estimated error can easily exceed the predefined threshold. Additionally, the flows that are kept by the sampling process are the same as the original flows, all the marginal flow properties, and in particular the flow distribution can be readily estimated from the observed sampling traffic.

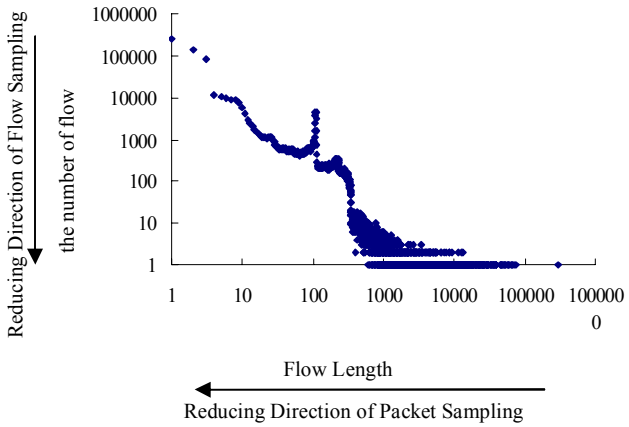

Fig. 1. Traffic Sampling Method

Figure 1 shows the reducing direction for different traffic sampling methods, where the X axis shows flow length, and Y axis is the number of active flows. A packet sampling method reduces the length of the flow in X axis direction. If the length of an original flow reduces to 0, then the flow will not sampled, and the record of the flow will be disappeared. Flow sampling method reduces the number of flows. If the number of flows of a particular length reduces to 0, then the flow records of the length will disappear. Thus, if a point approaches to the Y axis, then it may disappear from the figure when using packet sampling. Similarly, if a point approaches to X axis, then it will disappear when using flow sampling. Figure 2 (a) shows that the length distribution of short flows in the flow sampling set is very similar to that of the original traffic, but the distribution of the heavy-tailed flows has lost its original character. Conversely, the

heavy-tailed distribution of the packet sampling set in the figure 2(b) is very similar to that of the original traffic in the figure 1, but the short flows in the packet sampling set is not an accurate estimate of the original traffic.

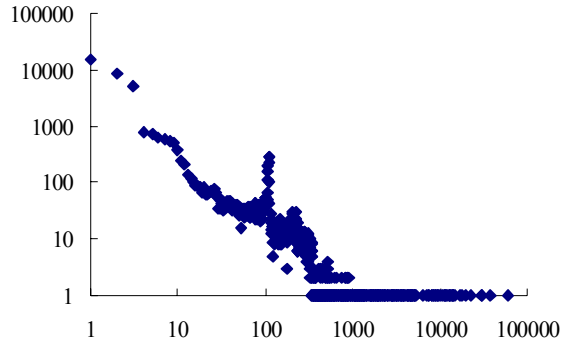

Fig. 2. (a) Distribution of Flow Sampling set

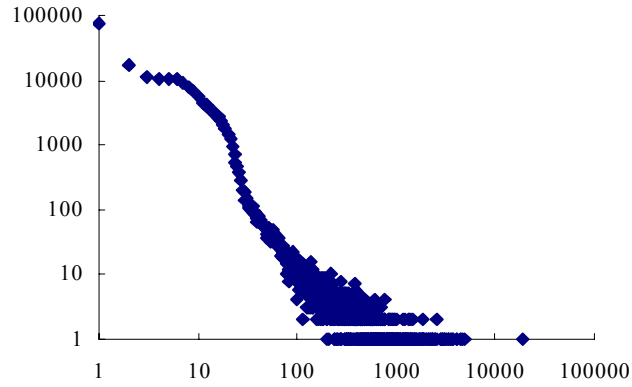

Figure 2 (b) Distribution of Packet Sampling set

In order to improve the estimate, accuracy for both the short flows and heavy-tailed flows, we propose a hybrid sampling approach to optimize the trade-off between the overall estimate accuracy of flow distribution and the available system resources. Our approach can intelligently estimate the short flows by adopting flow sampling, and recover the heavy-tailed flow distribution by using packet sampling. The rest of this paper is organized as follows. In section 2, we discussed the related work. In section 3, we classify three types of flows: namely short flows, heavy-tailed flows, and middle flows, and present different estimation methods accordingly. We analyze the computational complexity of our algorithm. Section 4 shows the evaluation results on estimation accuracy and computational complexity of our system by using real traffic from the CERNET backbone. We conclude our work in Section 5.

2. Related Work

The IETF Packet Sampling working group (PSAMP) [1] is chartered to define a standard set of capabilities for network elements to sample subsets of packets by statistical and other methods. Furthermore, sampling techniques have been employed in network products such as Cisco's Netflow [2].

The problem of detecting the flow distribution using sampling technique includes the detecting heavy-tailed flows, the number of flows, and the frequent items. Hohn [3] has proved that it is very hard to estimate the flow size distribution accurately from sampled traffic. Duffield [4] develop a model to estimate the distribution of flow, but he didn't give an accuracy proof, he also studied the statistical properties of packet-level sampling using real-world Internet traffic traces. This is followed by [5] in which the flow distribution is inferred from the sampled statistics. After showing that the simple scaling of the flow distribution estimated from the sampled traffic is in general not accurate, the authors propose an EM algorithm to iteratively compute a more accurate estimation. This scaling method is simple, but it uses the sampling properties of SYN flows to estimate TCP flow frequencies; The EM algorithm does not rely on the properties of SYN flows and hence is not restricted to TCP traffic, but its versatility comes at the cost of computational complexity. Ribeiro [6] proposes a systematic approach, using a fisher information metric and a Cramer-Rao bound, to understand the contributions that different types information within sampled packets have on the quality of flow-level estimates.

Estan [7] presented a family of bitmap algorithms for counting active flows. The measured flow numbers and the distribution of their lengths can be used to evaluate gains in deployment of web proxies [8], and determine thresholds for setting up connections in flow-switched networks [9].

Estan [10] has proposed a different packet sampling scheme in order to better capture the statistics of longer flows. Estan gives two algorithms to detect heavy-tailed traffic: sampled & hold and multistage filter. Kumar [11] proposed a novel SCBF that performs per-flow counting without maintaining per-flow state in and an algorithm for estimation of flow size distribution [12]. Raspall [13] present a shared-state sampling algorithm to detect large flows in high-speed networks, this algorithm can achieve a decrease in the detection probability for small flows, without affecting the detecting probability for large flows.

This paper presents a novel method for estimation of flow size distributions from sampled flow and sampled packet Statistics. Our method can be used not only to estimate TCP flows but also can be extended to general flows. Our primary contribution is to demonstrate that we can accurately estimate flow distribution is in practice using a combination of packet sampling and flow sampling. We present this work using five steps: The first step establishes the statistics of the original flow. The second step classifies short flows and heavy-tailed flows

based on the sampling theory. (We also define middle flows between short flows and heavy-tailed flows.) The third step estimates the full distribution of short flows using flow sampling, and the full distribution of heavy-tailed flows using packet sampling. The fourth step recovers the distribution of middle flows by the method of least squares, based the flow sampling set. Finally, we estimate the distribution of flow size.

3. Hybrid Estimated Flow Distribution Algorithm

In this section, we propose a flow classification model for estimating the total number of active flows and accordingly categorizing the monitored flows into the short, middle and heavy-tailed classes. Then we design a hybrid system to estimate the short and middle flow distribution by using the flow sampling set, and the heavy-tailed flow distribution by using the packet sampling set. Finally, we discuss algorithm complexity.

3.1 The Design of Flow Number Estimator

Assuming the number of the active flows are n, m and x in the flow sampling set , the packet sampling set , and the intersection set X ($X = \Phi \cap \Psi$) respectively. The number of flows in the original traffic set is N. So we have: $\Omega \supseteq \Psi$, $\Omega \supseteq \Phi$, and accordingly there are N-n flows in the set , which don't belong to set (because the element in set is sampled from the set randomly). The length distribution g(i) in the set is the approximately same as the length distribution f(i) in the set , thus we have equation (1).

$$g(i) \approx f(i) \tag{1}$$

We can consider that the set X is sampled from the set with the probability p, which is equal to the sampling probability which the set is sampled from the set .

The number n-x of flows in the set are not in the set X, so n-x flows are not sampled from the set X, and N-m flows in the set are not sampled from the set . n-x, and N-m can be computed by the equation below:

$$n - x = \sum_{i=1}^{\infty} g(i) \cdot n \cdot (1-p)^i = n \cdot \sum_{i=1}^{\infty} g(i) \cdot (1-p)^i$$

$$\mathrm{N - m} = \sum_{i=1}^{\infty} \mathrm{f(i)} \cdot \mathrm{N} \cdot (1 - \mathrm{p})^{\mathrm{i}} = N \cdot \sum_{i=1}^{\infty} f(i) \cdot (1-p)^i$$

According to the above two equations and equation (1), we can obtain a flow number estimator as shown in the equation (1), where M is the size of the longest flow, $\hat{N}$ is the estimated value of the number of the active flows, and m is the number of the active flows in the packet sampling set, x expresses the number of the active flows in the intersection set between the packet sampling set and flow sampling set, and f means the ratio of the flow sampling to the original traffic.

$$(n-x)/n = (\hat{N} - m)/\hat{N}, \ f = m/x,$$
$$\hat{N} = n \cdot m / x = n \cdot f \tag{2}$$

Theorem 1: From the hybrid sampling set produced with independent random packet sampling and flow sampling at a flow sampling rate which is 1/f, so the estimator of the number of flows with a relative standard deviation is $\sqrt{1/n + 1/x + 1/m}$.

Proof:

Let N be the total number of flows sent during the measurement interval. With a flow sampling rate of p, the expected number of flows n. The number of those sampled has a binomial distribution with mean n=N*p, and variance p(1-p)N. Since we get the estimate for the number of flows in the original traffic by multiplying the number of sampled flows by f, f=m/x. Let the packet sampling ratio be q, and the flow sampling ratio of x from m 1/f. So the variance of m is q(1-q)N, and the variance of x is 1/f(1-1/f)qN. $\hat{N} = p \cdot N \cdot m / x$, so the standard variance of the estimate N will be

$$V(\hat{N}) = p \cdot (1-p) \cdot N \cdot (m/x)^2 +$$
$$(pN \cdot m / x^2)^2 \cdot 1/f(1-1/f)qN + (pN/x)^2 \cdot q(1-q)N$$
$$= N/p + N^2/x + N^2/m$$

, so its relative stand deviation is $\sqrt{1/n + 1/x + 1/m}$.

3.2 Estimated Short Flows Distribution Based on Flow Sampling

Assuming the ratio of the flows with length i in the original traffic and the flow sampling set are f(i) and g(i) respectively, and the flow sampling probability is p, so all flows with length i are considered a subset of the original traffic set. Equation (3) defines a new variable Y_{ij} to record the i^{th} flow in the traffic set whose length is equal to j or isn't j.

$$Y_{ij} = \begin{cases} 1, & if\ the\ length\ of\ the\ i_{th}\ flow\ is\ j. \\ 0, & if\ the\ length\ of\ the\ i_{th}\ flow\ isn't\ j. \end{cases} \tag{3}$$

Let $Y_i = \sum_{j=1}^{N} Y_{ij}$, $f(i) = \overline{Y}_i = Y_i / N$, $g(i) = \overline{y}_i = \sum_{j=1}^{n} y_{ij} / n$.

In fact, we can be easy to know than Y_i is the number of flows with i packets, f(i) is the frequency of flows with i packets in the original traffic set, and g(i) is the frequency of flows with i packets in the flow sampling traffic set. We have the following theorems.

Theorem 2: g(i) is a estimator of f(i) in the original traffic, and its estimated variance is $V(g(i)) = E(g(i) - f(i))^2 = \frac{f(i) * (1 - f(i))}{n} \left(\frac{N-n}{N-1} \right)$.

Proof:

$$E(g(i)) = E(\bar{y}_i) = \bar{Y}_i = f(i)$$

Since $\sum_{j=1}^{N} Y_{ij}^2 = N \cdot f(i)$, $\sum_{j=1}^{n} y_{ij}^2 = n \cdot g(i)$, we have

$$S^2 = \frac{1}{N-1} \cdot \sum_{j=1}^{N} (Y_{ij} - \bar{Y}_i)^2 = \frac{1}{N-1} \cdot \left(\sum_{j=1}^{N} Y_{ij}^2 - N\bar{Y}_i^2 \right)$$

$$= \frac{1}{N-1} (N \cdot f(i) - N \cdot f(i)^2)$$

$$= \frac{N}{N-1} f(i)(1 - f(i))$$

Then we can infer the following.

$$V(g(i)) = E(g(i) - f(i))^2 = \frac{S^2}{n} \left(\frac{N-n}{N} \right)$$

$$= \frac{f(i) * (1 - f(i))}{n} \frac{N-n}{N-1}$$

Theorem 3: $v(g(i)) = s_{g(i)}^2 = \frac{N-n}{(N-1)n} g(i) \cdot (1 - g(i))$ is the estimation of V(g(i)).

Proof:

$v(g(i)) = \frac{s^2}{n} \left(\frac{N-n}{N} \right)$ is the estimation of the variance of the mean $\bar{y}$, thus we can derive the equation below from Theorem 1.

$$s^2 = \frac{n}{n-1} g(i) \cdot (1 - g(i)),$$

so $v(g(i)) = \left(\frac{N-n}{(n-1)N} \right) \cdot g(i) \cdot (1 - g(i))$.

Theorem 4: The estimation of the number of the active flows is:

$$flow_{\min} = \frac{\mu_\alpha^2}{r^2} / (1 + \frac{\mu_\alpha^2}{r^2} \cdot \frac{1}{n}) \approx \frac{\mu_\alpha^2}{r^2} \quad (4)$$

r is the relative error of the estimated ratio p, that is $P_r \left(\left| \frac{Np - NP}{NP} \right| \le r \right) = P_r(|p - P| \le rP$

So r is also the relative error of Y=PN, then we can obtain n,

$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$, $n_0 = \frac{\mu_\alpha^2}{r^2} \cdot \frac{1-p}{p}$, if n_0/N is very small, n_0 can be considered to be the estimate of n. So, $n = \frac{\mu_\alpha^2}{r^2} \cdot \frac{1-p}{p}$, $flow_{\min} = p \cdot n$. $flow_{\min}$ stands for the minimum number of flow sampling to estimate the flow distribution of the extent of the defined precision. $flow_{\min} = \frac{\mu_\alpha^2}{r^2} / (1 + \frac{\mu_\alpha^2}{r^2} \cdot \frac{1}{n}) \approx \frac{\mu_\alpha^2}{r^2}$.

Thus, the number of the flows with length i can be estimated from the flow sampling immediately. Accordingly, we have the following equation (5).

$$\hat{Y}(i) = y(i) \cdot f, \quad y(i) \ge flow_{\min} \quad (5)$$

If y(i) >= $flow_{min}$, y(n_{min}+1) < $flow_{min}$ (i=1,…,n_{min}), then the flow length is equal to $i \in [1, n_{\min}]$. We can use equation (5) to estimate the distribution of short flows. For example, if the relative error of the estimated flow length is 25%, and the given confidence level α= 90%, then flowmin=44. $flow_{\min} = \frac{\mu_\alpha^2}{r^2} = \frac{1.645^2}{0.25^2} = 44$.

3.3 Estimated Distribution of Heavy-tailed Flow on Packet Sampling

Assuming the packet ratio of the i[th] flow in the original traffic and packet sampling set are f(i) and g(i) respectively, and the packet sampling probability is p=1/n, all packets sampled one of ith flow are considered a subset of the original traffic set. Thus, we can define each packet in equation (6). Equation (6) defines a new variable Y_j to record the j[th] packet in the traffic set which is belong to a flow with i packets or not.

$$Y_{ji} = \begin{cases} 1, & \textit{if the } j^{th} \textit{ packet belongs to the } i^{th} \textit{ flow.} \\ 0, & \textit{if the packet doesn't belong to flow i.} \end{cases} \quad (6)$$

$$Y_i = \sum_{j=1}^{N} Y_{ij} \quad , \quad f(i) = Y_i / N = \bar{Y}_i \quad , \quad g(i) = \bar{y} = \sum_{i=1}^{n} y_i / n \quad .$$

Using equation (4), we can compute the minimal length of the packet sampling flow $packet_{\min} = \frac{\mu_\alpha^2}{r^2}$ to estimate heavy-tailed flows by packet sampling, where $packet_{min}$ represents the minimal length of the packet sampling flow. Thus, we can use packet sampling to estimate the heavy-tailed length of the original traffic immediately.

Let the packet sampling ratio be 1/n, then the estimation of flow length in the original traffic set is $packet_{min}$=*n. Let F_k be an original flow with k packets, S_l is a sampled flow with l packets from F_k, then S_l follows a binomial distribution $B_p(k,i) = \binom{k}{i} p^i (1-p)^{k-i}$. If the probability p is less than 0.1, then the binomial distribution is similar to a Poisson distribution. $p(x=i) = \frac{e^{-\lambda} \cdot \lambda^i}{i!}$, where $\lambda = k \cdot p = k / n$. If the length of sampling flow is i, then the probability p(x=i) reaches its maximum.

$$p(x = i-1) < p(x = i) > p(x = i+1)$$

$$\frac{e^{-\lambda} \cdot \lambda^{i-1}}{(i-1)!} < \frac{e^{-\lambda} \cdot \lambda^i}{i!} > \frac{e^{-\lambda} \cdot \lambda^{i+1}}{(i+1)!} \quad , \quad \lambda > i > \lambda - 1 \quad ,$$

$$i = \lfloor \lambda \rfloor = \lfloor k/n \rfloor$$

Let the length of a flow be k, then the length of the maximal probability in sampling flows is $\lfloor k / n \rfloor$.

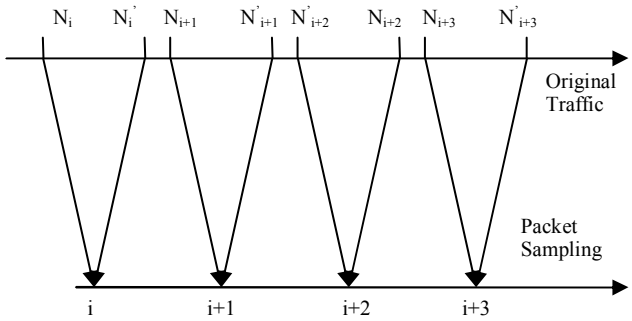

Fig. 3. the relationship between packet sampling and original sampling in the heavy-tailed flows.

Where N_i = i*n, N_i' = i*n+(n-1), The original flow of length in the area [N_i, N_i'] can be sampled to length i, which is the maximal probability. Let the length of an original flow be k, k=m*n+l, where m is a positive integral number, $n > l \geq 0$. The packet sampling flow length of the maximal probability is $\lfloor k/n \rfloor = \lfloor (m*n+l)/n \rfloor = m$, so the length of its original flow is $k = m \cdot n + l, \quad l = [0, n-1]$. Based on the maximal probability theory, the length m of sampling flow is sampled from the original flow set $K_m = [m \cdot n, m \cdot n + (n-1)]$. In the set K_m, the probability that an original flow of length m*n+i can be sampled into a flow of length k is

$$p_{m,i} = \frac{e^{-k/n} \cdot (k/n)^m}{m!} = \frac{e^{-(m \cdot n+i)} \cdot (m \cdot n+i)}{m!}, \quad i \in [0, n-1]$$

$w_{m,i}$ is the weight that a sampled flow of length i is sampled from an original flow of length m*n+i in the set K_m.

$$w_{m,i} = \frac{p_{m,i}}{\sum_{i=0}^{n-1} p_{m,i}} = \frac{e^{-(m \cdot n+i)} \cdot (m \cdot n+i)/m!}{\sum_{i=0}^{n-1} e^{-(m \cdot n+i)} \cdot (m \cdot n+i)/m!} = \frac{e^{-i} \cdot (m \cdot n+i)}{\sum_{i=0}^{n-1} e^{-i} (m \cdot n+i)} \tag{7}$$

If we let the number of the length m flows in the packet sampling set be g(m), then the estimated number of the length m*n+i flow in the original traffic set is

$$\hat{n}_i = \lfloor g(m) \cdot w_{m,i} \rfloor \tag{8}$$

If $s = \sum_{i=0}^{n-1} \hat{n}_i$, then g(m)-s flow aren't estimated in the equation (8). In this paper, we distribute the g(m)-s flow into the area K_m randomly, based on the weight equation (7). The cumulative function of the weight is $W_{m,k} = \sum_{i=0}^{k} w_{m,i}$, so $W_{m,n-1} = 1$. Let variable be a random function that produces random data in the area [0, 1]. If $W_{m,i-1} < random() \leq W_{m,i}$, then 1 is added to x(i). The above procedure is computed g(m)-s times, so we can estimate the number of the original flow in the area of K_m in the equation (9).

$$\hat{N}(m \cdot n+i) = \hat{n}_i + y(i) = \lfloor g(m) \cdot w_{m,i} \rfloor + y(i), \quad i \in [0, n-1] \tag{9}$$

3.4 Estimated Medial Length Flow

If the length of short flow is less than $flow_{min}$, then we can use equation (5) to estimate the length distribution, and if the length of heavy-tailed flows is larger than $packet_{min}$*n, then we can use the equation (9). In the section, we will analyze the estimation method that the length of flow is in the range of A=($flow_{min}$, $packet_{min}$ * n). If

$$\sum_{i=k}^{l-1} n_i < flow_{\min}, \quad s = \sum_{i=k}^{l} n_i \geq flow_m, \quad i \in [m+1, packet_{\min} * (n-1) + (n-1)],$$

or, $$s = \sum_{i=k}^{l} n_i \leq flow_{\min}, \quad l = packet_{\min} * (n-1) + (n-1).$$

Where n_i is the number of length i flows in the range of A. The estimated number of the active flows in the range of [k,l] is $\hat{n}_i = s \cdot f$, The distribution of flow length is heavy-tailed, so $P[X > x] \sim x^{-\alpha}$, $x \to \infty, 0 < \alpha < 2$. The simplest heavy-tailed distribution is the Pareto distribution. and its probability function is $p(x) = \alpha k^{\alpha} x^{-\alpha-1}$, α, k>0, x≥k, and its cumulative function is $F(x) = P[X \leq x] = 1 - (\frac{k}{x})^{\alpha}$. Using the probability p(x), we can obtain the differential coefficient equation $p'(x) = -\alpha k^{\alpha} (\alpha+2) x^{-\alpha-2}$.

$$\Delta = \frac{p'(x+k)}{p'(x)} = \frac{-\alpha k^{\alpha}(\alpha+2)(x+k)^{-\alpha-2}}{-\alpha k^{\alpha}(\alpha+2)x^{-\alpha-2}} = \frac{(x+k)^{-\alpha-2}}{x^{-\alpha-2}}$$

$$= \left(1+\frac{k}{x}\right)^{-\alpha-2} = 1-(\alpha+2)\frac{k}{x} + 0(\frac{k}{x})$$

If x is larger than k, then $\Delta \approx 1$. So we can use a linear that approaches the length distribution in a small range, that is, we can use the method of least squares to approach to point (i, $\hat{n}_i$), $i \in [k,l]$.

$$\hat{N}(i) = a + b \cdot i \tag{10}$$

These points (i, $\hat{n}_i$), $i \in [k,l]$ are close to the equation (10), so we can compute the parameter a, and b in the equation (10) by the below equation.

$$\sum_{i=l}^{k} \Delta_i^2 = \sum (a + b \cdot i - \hat{n}_i)^2 = \min \quad , \quad a = \hat{n}_i - b \cdot i \quad ,$$

$$b = \frac{\sum (i - \bar{i})(\hat{N}(i) - \hat{n}_i)}{\sum (i - \bar{i})^2}$$

3.5 Algorithm Complexity

In the active flow number estimation algorithm, if the number of the active flow in the flow sampling set is n, then the size of the packet sampling set is m. Accordingly, the time complexity of the algorithm is O(n+m). For the estimation algorithm for the short flows as shown in Equation (5), we only recover the length distribution of

flows whose length is less than $flow_{min}$, thus the time complexity of the algorithm is $O(flow_{min})$. For the estimation algorithm for the heavy-tailed flows as shown in Equation (9), it computes the packet sampling flow from $packet_{min}$ to m, if every k packet samples one packet, thus, the time complexity of the algorithm is $O(k*(m-packet_{min}))$.

The estimation algorithm for the middle flows as shown in Equation (10) estimates the flow with the length between $k*packet_{min}$ and $flow_{min}$. The distributed length space is divided into several subsections such that each subsection is estimated by the method of least squares, Thus, the time complexity of the algorithm is $O((packet_{min}*k-n_{min})^2)$.

Using the methods above, we can obtain the total time complexity is $O(n+m)+ O(flow_{min})+ O(k(m-packet_{min}))+ O((packet_{min}*k-flow_{min})^2)$, Conclusively, the total time complexity of approach is $O(n^2+m)$.

4 Performance Evaluation

We conducted extensive experiments to evaluate our system performance. We collected packets from the CERNET backbone network with 1Gbps bandwidth. The 48 bytes packet header of every packet was intercepted. We measured two groups of original traffic from the CERNET backbone network, deterministically took one in every 16 or 32 packets (systematic sampling) as packet sampling sets from the original traffic, and got the flow sampling set by mask length 4 bits and 5 bits to obtain the flow sampling rate $1/2^4=1/16$, and $1/2^5=1/32$. Table 1 is the detailed information of the measured and sampled data.

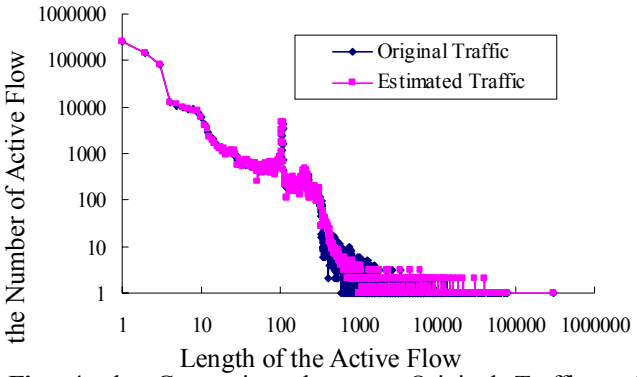

Fig. 4. the Comparison between Original Traffic and Estimated Traffic Using the First Measured Traffic Data. The estimated distribution using our method is compared with the length distribution of original traffic.

In this paper, we define IP flow as a set of packets with the same 5-tuple {IP protocol; source and destination address; source and destination port}, and with a measured duration. The packet sampling method used is to take one in every n packets from the original traffic. We define a hash function, whose input is the 5-tuple fields {IP protocol; source address; destination address; source port; destination port}, and its output is an 8 bit flow. We use a defined m bits mask to match the hash value to get a sampling rate $1/2^m$.

Table 1. this table is the detailed information of the measured and sampled data.

Sequence	1	2	3	4
Duration	00:00:18-00:03:03	10:00:06-10:05:26	00:00:18-00:03:03	10:00:06-10:05:26
Sampling Length	16	16	32	32
Mask Length	4	4	5	5
Total Packet #	32097160	64239632	32097160	64239632
Packet # of flow sampling	1978792	3836233	1142185	2041066
Packet # of packet sampling	2006150	4015022	1003062	2007483
Packet # in the interaction	123599	240465	35626	64403

Table 2. the Estimation Result of Flow Distribution

Sequence	1	2	3	4
Total Flow #	678028	1926296	678028	1926296
Flow # in Packet Sampling	203624	462271	158691	315998
Flow # in Flow Sampling	42640	119894	22231	60527
Flow # in Interaction Set	12781	28728	5599	9747
Estimation Of Flow #	679330	1929250	630087	1962287

Table 2 is the estimated result of our method, and compares it with the EM algorithm [5]. Figure 4 and Table 3 show that the estimation accuracy of our algorithm is close enough to that of the original traffic and is much more accurate than that of EM algorithm. In the Table 3, two performance metrics is introduced. FlowError metrics is to estimate accuracy of the number of flows, and WMRD metrics is to estimate the distribution accuracy. Equation (11) and Equation (12) give the two metrics.

FlowError metrics: We use a FlowError metrics as the evaluation metric to estimate accuracy of the number of flows. Suppose the number of original flows is N, and our estimation of this number is $\hat{N}$. The value of FlowError is given by:

$$FlowError = \left| N - \hat{N} \right| / N \times 100\% \qquad (11)$$

WMRD metrics: We use WMRD as a evaluation metric of distribution estimation accuracy. Suppose the number of

original flows of length i is n_i and our estimation of this number is $\hat{n}_i$. The value of WMRD is given by:

$$WMRD = \frac{\sum_i |n_i - \hat{n}_i|}{\sum_i (\frac{n_i - \hat{n}_i}{2})} \times 100\% \quad (12)$$

Table 3. Estimated Accuracy between EM and Hybrid Algorithm

Sequence	1	2	3	4
Ratio of packet Sampling	3861343	7610790	2109621	3984146
Sampling ratio in EM	4012160	8029954	2006081	4014978
Theory Error by Theo. 1	1.01%	0.67%	1.52%	1.11%
FlowError	0.19%	0.15%	7.07%	1.87%
WMRD of Hybrid Sampling	3.8%	1.6%	11.4%	3.1%
WMRD of EM Algorithm	24%	18%	25%	19%

5 Conclusion

We present a flow estimation algorithm using hybrid sampling technique, which combining both flow and packet sampling, to estimate the flow distributions. Two kinds of sampling were used, I.I.D. packet sampling, and I.I.D. flow sampling, with a given sampling probability. Exact theoretical estimated techniques were derived. As we proved, the flows that are kept by the flow sampling procedure are identical to the original ones, and the intersection flow set between flow and packet sampling set has the same distribution as in packet sampling set, we can count the number of active flows in the flow sampling set, the packet sampling set, and the intersection set between packet sampling and flow sampling respectively. According to the active flow numbers in the three sets, we can infer the number of active flows in the original traffic set, and the sampling ratio between flow sampling set and original sampling set, using sampling theory. To achieve the required accuracy, we take full advantage of flow sampling and packet sampling, and use the flow sampling to estimate the distribution of short and heavy-tailed flows. We recover the distribution of flows between short flows and heavy-tailed flows using the method of least squares based on the flow sampling set. Finally, we used CERNET backbone traffic to analyze the performance of the algorithms and compare them with others algorithms. The experimental results show that our algorithms outperformed both packet and flow sampling approaches on the estimation accuracy of flow distributions.

Acknowledgments. This work has been supported by the Key Project of Chinese Ministry of Education under Grant No. 105084, the National Grand Fundamental Research 973 program of China under Grant No. 2003cb314804, the Natural Science Fundamental of Jiangsu Province under Grant No. BK2006092, and the Excellent Youth Teacher of Southeast University Program under Grant No. 4009001018, and the Natural Science Fundamental of China under Grant No. 50609006.

References

1.Packet Sampling (psamp), http://www.ietf.org/html.charters/psamp-charter.html, 2002.12.
2. Cisco IOS NetFlow, http://www.cisco.com/warp/public/732/Tech/netflow/index.shtml
3. Nicolas Hohn, Darryl Veitch. Inverting sampled traffic. In IMC' 03: Proceedings of the 3rd ACM SIGCOMM conference on Internet Measurement, NY, USA, 2003.
4. Duffield, N.G., Lund, C., Thorup, M.: Properties and Prediction of Flow Statistics from Sampled Packet Streams, ACM SIGCOMM Internet Measurement Workshop 2002, Marseille, France, November 6-8, 2002.
5. Duffield, N.G., Lund, C. , Thorup, M.: Estimating Flow Distributions from Sampled Flow Statistics. ACM SIGCOMM . 2003, Karlsruhe,Germany. August 25-29. 325-336.
6. Bruno Ribeiro, Don Towsley, Tao Ye, Jean Bolot, Fisher Information of Sampled Packets: an Application to Flow Size Estimation, Rio de Janeiro, Brazil, IMC 06, Oct.2006.
7. Estan, C. andVarghese, G., Bitmap algorithms for counting active flows on high speed links, in Proc. ACM SIGCOMM Internet Measurement Conference, Oct. 2003.
8. Feldmann, A. , Caceres, R. , Douglis, F. , Glass, G., Rabinovich, M.: Performance of Web Proxy Caching in Heterogeneous Bandwidth Environments, in Proc. IEEE INFOCOM 99, New York, NY, March 23-25, 1999.
9. Feldmann, A., Rexford, J., and C¡äaceres, R.: Efficient Policies for Carrying Web Traffic over Flow-Switched Networks, IEEE/ACM Transactions on Networking, vol. 6, no.6, pp.673-685, December 1998.
10. Estan, C. andVarghese, G., New Directions in Traffic Measurement and Accounting, SIGCOMM2002
11. Abhishek Kumar, Jun Xu, Li Li, and Jia Wang, Space Code Bloom Filter for Efficient Traffic Flow Measurement, Proceedings of ACM/USENIX Internet Measurement Conference, Miami, FL, October 2003.
12. Abhishek Kumar, Minho Sung, Jun (Jim) Xu and Jia Wang, Data streaming algorithms for efficient and accurate estimation of flow size distribution, ACM SIGMETRICS 2004.
13. Frederic Raspall, Sebastia Sallent, Josep Yufera, Shared-State Sampling, Rio de Janeiro, Brazil, IMC 06, Oct.2006.