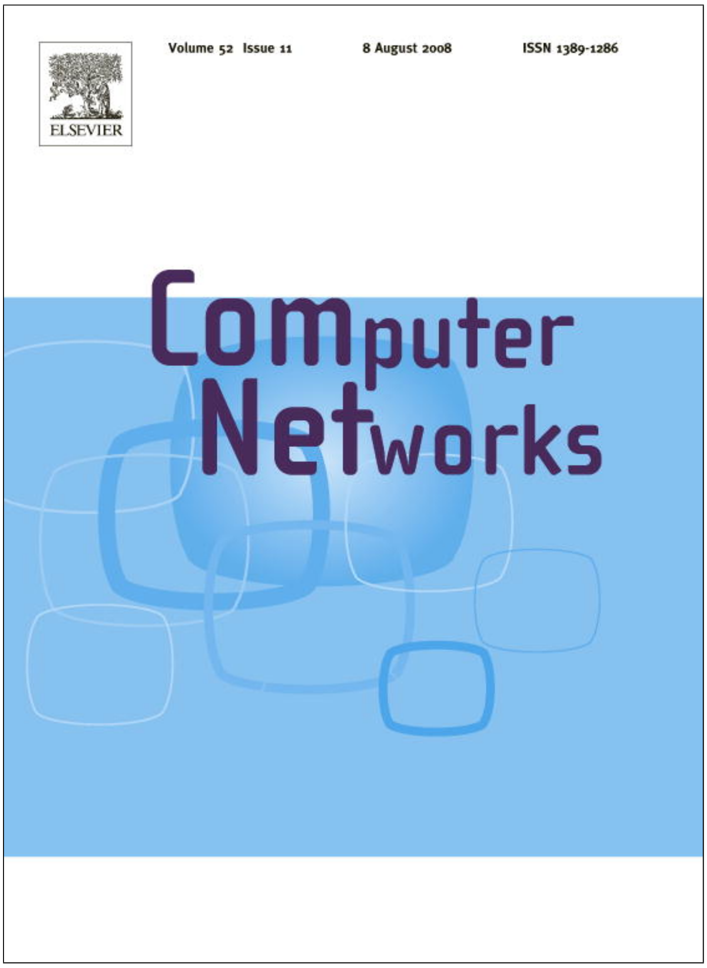
Volume 52 Issue 11
8 August 2008
ISSN 1389-1286
ELSEVIER
Computer Networks

Computer Networks 52 (2008) 2221–2226

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Double sampling for flow measurement on high speed links ☆

W.J. Liu [a,*], J. Gong [b]

[a] School of Computer Science and Technology, Dalian Maritime University, 1 Linghai Road, Dalian, Liaoning 116026, China
[b] School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu 210096, China

ARTICLE INFO

Article history:
Received 18 May 2007
Received in revised form 22 March 2008
Accepted 9 April 2008
Available online 15 April 2008

Responsible Editor: A. Marshall

Keywords:
Packet sampling
IP flows
Network measurement
Statistic inference

ABSTRACT

Traffic measurement and monitoring are an important component of network QoS management and traffic engineering. With high speed Internet links, efficient and effective packet sampling techniques for traffic measurement are not only desirable, but increasingly becoming a necessity. 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. However, it meets difficulty in estimating the original flow distribution. To circumvent the problem, we propose and analyze a double sampling technique for flow measurement. In particular, we rewrite the expectation maximization (EM) algorithm that estimates flow distribution for double sampling. Using real network traffic traces, we show that the proposed double sampling technique indeed produces the desired accuracy in estimating the flow distribution.

1. Introduction

Packet collection tools have been developed since the inception of packet switched networks as a means to debug protocol stacks or network interfaces. There are in fact many ways of collecting packets on a link, based either on a software or hardware solution, for both online and offline analysis. For instance one can use hardware equipment such as a line tester or a protocol analyser to generate real time counts of link layer faults or packet arrivals. One can also use software tools such as tcpdump to investigate IP packets on a LAN. These measurement techniques are non-intrusive, in the sense that they do not modify the traffic, and are often referred to as passive measurements. They differ from active measurement techniques where artificial traffic is injected in the network, for instance to estimate link bandwidth.

1.1. Related work

Passive traffic measurement increasingly employs sampling at the packet level to control the consumption of resources in measurement. Many high end routers form flow statistics from only a sampled substream of packets in order to limit the consumption of memory and processing cycles involved in flow cache lookups.

In 1993, Claffy et al. [1] studied systematic, stratified random and simple random sampling method by packet or time. Systematic sampling involves deterministically selecting one in every k packets of the data set. Stratified random sampling is similar to systematic sampling, except that rather than selecting the first packet from each bucket, a packet is selected randomly from each bucket. Simple random sampling uniformly selects n packets from the total population at random. They compared above three methods on $1/T$ time-driven and $1/N$ packet-driven mechanisms and showed that time-driven technique performs worse than packet-driven one. The method of sampling by packet's content was discussed in [2]. Cozzani and Giordano [2] showed how to manage the trade-off between the sampling rate and attained estimation accuracy, representing a theoretical/quantitative method

☆ This work is supported in part by 973 Program of China under Grant No. 2003CB314804, the National High Technology Research and Development Program of China (2005AA103001).
* Corresponding author. Tel.: +86 41184723716.
E-mail addresses: wjliu@newmail.dlmu.edu.cn (W.J. Liu), jgong@njnet.edu.cn (J. Gong).

1389-1286/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2008.04.003

usable by network operators to design control instrumentation.

The packet sampling working group (PSAMP), which was founded by IETF in 2003, is chartered to define a standard set of capabilities for network elements to sample subsets of packets by statistical and other methods [3]. The method and application of Trajectory Sampling was considered in [4,5]. The key idea of Trajectory Sampling is to sample packets based on a hash function computed over the packet content. Using the same hash function will yield the same sample set of packets in the entire domain, and enables us to reconstruct packet trajectories. Furthermore, sampling techniques have been employed in network products such as Cisco's Netflow [6] and NetranMet [7].

In fact, two different sampling rules are included in the above work: *packet sampling*, which acts directly on individual packets and is ignorant of flows, and *flow sampling*, where entire flows of packets are retained or discarded at once. *Packet sampling* without per-packet processing can be easily implemented. Hohn and Veitch in [8] discussed the inaccuracy of estimating flow distribution from sampled traffic, when the sampling is performed at the packet level. For flow sampling, its disadvantage is that all packets must be processed before deciding to be retained or discarded. This is a challenge on high speed links.

In order to overcome aforementioned shortcomings of packet sampling and flow sampling, we propose a novel sampling method, *double sampling*, that combines packet sampling and flow sampling. Double sampling consists of packet sampling and flow sampling:

1. First, employ packet sampling to obtain a substream.
2. Then, deploy flow sampling to select a small number of packets from the substream.

1.2. Contribution and outline

This paper presents a novel sampling method, double sampling, that overcomes the shortcomings of packet sampling alone and flow sampling alone. This method is not only simple to implement but also scalable for high speed links. This method displays its advantage in estimating the distribution of flow length, i.e., the accuracy of flow distribution measurement is dramatically improved.

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 describe double sampling in detail. In Section 4, we give methods for estimating the length and total bytes of original flows. Then we analyze the estimation methods. Finally, we give the EM algorithm that estimates flow distributions from sampled flow statistics by double sampling. In Section 5 we make some experiments to observe estimation accuracy. We conclude in Section 6.

2. Some elementary concepts

This section considers packet sampling alone. Within the functional requirement of sampling packets at a given rate, a number of different implementations are possible. Implementations include independent sampling of packets with probability $1/N$, 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. An IP flow is a set of packets, that are observed in the network within some time period, and that share some common property known as its key. The fundamental example is that of so-called raw flows: a set of packets observed at a given network element, whose key is the set of values of those IP header fields that are invariant along a packet's path. Examples are the raw flows observed at a router, where the flow key distinguishes individual source and destination IP address, and TCP/UDP port numbers. There are at least a few definitions for the term flow depending on the context of research. In this study, we employ the one adopted in [9] which stems from the packet train model by Jain and Routhier [10].

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

In most cases, we call flow specification as flow identifier. 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 s. 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. The frequency of flows with k packets is the number of flows that contain k packets.

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. Double sampling

In this section, we will discuss the methodology of double sampling. Double sampling consists of packet sampling and flow sampling:

Step 1 Independent and identically distributed (i.i.d.) packet sampling consists of, for each packet in an independent manner, retaining the packet with probability $1/n$ or discarding it with probability $(n-1)/n$.

Step 2 Independent and identically distributed (i.i.d.) flow sampling consists of, for each flow whose packet is selected in Step 1 in an independent manner, retaining the flow with probability $1/m$ or discarding it with probability $(m-1)/m$.

There are several ways to implement Step 1. In probabilistic sampling, the router makes a pseudorandom decision whether to sample each packet. In implementations, the decision could, for example, be governed by a pseudorandom number generator with well-known properties (see e.g. [11]). Periodic (or deterministic) sampling can be used too, e.g., every nth packet is selected.

We use a hash function over flow identifier (the part of packet's header) to implement Step 2. The same hash function is used throughout measurement interval, so that we are ensured that all packets of a flow are either sampled, or discarded. The choice of an appropriate hash function will obviously be crucial to ensure that this subset is not statistically biased in any way. For this, the sampling process, although it is a deterministic function of the packet's header, has to resemble a random sampling process. We can choose a hash function that maps each item in the universe of flows to a random number uniform over the range $\{1, 2, \ldots, m\}$. In practice, reasonable hash functions appear to behave adequately, e.g. [12]. A packet is sampled if its hash value is equal to the specific integer, e.g., 1. Because the hash function is perfectly random, flow sampling rate is $1/m$, i.e., a flow in every m flows is sampled.

For double sampling, we denote sampling period N by $n * m$, written as $N = n * m$. For example, if $n = 5, m = 10$, then N is written as $5 * 10$.

4. Estimation methodology

4.1. Estimation of the length and total bytes of flow

Denote observed flows by $f_k (k = 1, 2, \ldots)$ and the total number of packets arrived of all flows by N_{total} within the measurement interval. For a flow f_k, let N_k be the number of packets belonging to it; and denote the jth packet of the flow by $p_{kj} (j = 1, 2, \ldots, N_k)$ and its packet size by x_{kj}. Similarly, denote the number of sampled packets belonging to flow f_k by n_k, and the lth packet in the sampled packets by $s_{kl} (l = 1, 2, \ldots, n_k)$, its packet size by X_{kl}. In the sequence of arrival, we arrange all packets as $p_i (i = 1, 2, \ldots, N_{\text{total}})$, which corresponds to some p_{kj} uniquely, i.e., exists a map: $i(k,j) \mapsto i$.

If $n_k > 0$, the total bytes of flow f_k is estimated as

$$\hat{x}_k = n \sum_{l=1}^{n_k} X_{kl}. \tag{1}$$

Eq. (1) shows how to estimate the total bytes of a sampled flow, i.e., at least a packet is sampled.

If $n_k > 0$, the total number of packets in flow f_k is estimated as

$$\widehat{N}_k = n n_k. \tag{2}$$

Eq. (2) shows how to estimate the total number of packets in a sampled flow, i.e., at least a packet is sampled.

The total bytes of all flows in the whole measurement interval x is estimated as

$$\hat{x} = m \sum_{k \in \{k | n_k > 0\}} \hat{x}_k. \tag{3}$$

Eq. (3) shows how to estimate the total bytes of all flows.

The total packet number of all flows in the whole measurement interval N_{total} is estimated as

$$\widehat{N}_{\text{total}} = m \sum_{k \in \{k | n_k > 0\}} \widehat{N}_k. \tag{4}$$

Eq. (4) shows how to estimate the total number of packets in all flows.

4.2. Estimation analysis

Let $w_i (i = 1, \ldots, N)$ be random variables taking the value 1 (indicating that the packet was sampled) with probability $1/n$ and 0 (indicating that the packet was not sampled) with probability $1 - 1/n$. For independent packet sampling, the w_i are independent.

Similarly, let $u_k (k = 1, 2, \ldots)$ be random variables taking the value 1 (indicating that the flow was sampled) with probability $1/m$ and 0 (indicating that the flow was not sampled) with probability $1 - 1/m$. For independent flow sampling, the u_i are independent.

Lemma 1. *The mean and variance of random variable w_i are $E(w_i) = 1/n$ and $\text{Var}(w_i) = \frac{n-1}{n}$, respectively.*

Lemma 2. *The mean and variance of random u_k are $E(u_k) = 1/m$ and $\text{Var}(u_k) = \frac{m-1}{m}$, respectively.*

Lemma 3. *The random variables w_i, u_k are mutually independent.*

Lemma 4. $E(w_i u_k) = \frac{1}{mn}, \text{Var}(w_i u_k) = \frac{mn-1}{(mn)^2}$.

Theorem 1. $E(\widehat{N}_k | n_k > 0) = N_k, \text{Var}(\widehat{N}_k | n_k > 0) = (n-1) N_k$.

Proof. We can write the random variable $\widehat{N}_k = n n_k = n \sum_{j=1}^{N_k} w_{i(k,j)} u_k$. By $n_k > 0$, we have $u_k = 1$. Hence $\widehat{N}_k = n \sum_{j=1}^{N_k} w_{i(k,j)}$, conditional on $n_k > 0$. Therefore, $E(\widehat{N}_k | n_k > 0) = n \sum_{i=1}^{N_k} E(w_{i(k,j)}) = N_k, \text{Var}(\widehat{N}_k | n_k > 0) = n^2 \sum_{i=1}^{N_k} \text{Var}(w_{i(k,j)}) = (n-1) N_k$. □

Theorem 2. $E(\hat{x}_k | n_k > 0) = \sum_{j=1}^{N_k} x_{kj}, \text{Var}(\hat{x}_k | n_k > 0) = (n-1) \sum_{j=1}^{N_k} x_{kj}^2$.

Proof. We write the random variable $\hat{x}_k = n \sum_{l}^{n_k} X_{kl} = n \sum_{j=1}^{N_k} w_{i(k,j)} u_k x_{kj}$. By $n_k > 0$, we have $u_k = 1$. Hence $\hat{x}_k = n \sum_{j=1}^{N_k} w_{i(k,j)} x_{kj}$, conditional on $n_k > 0$. Therefore, $E(\hat{x}_k | n_k > 0) = n \sum_{i=1}^{N_k} E(w_{i(k,j)}) x_{kj} = \sum_{j=1}^{N_k} x_{kj}$, $\text{Var}(\hat{x}_k | n_k > 0) = n^2 \sum_{i=1}^{N_k} \text{Var}(w_{i(k,j)}) x_{kj}^2 = (n-1) \sum_{j=1}^{N_k} x_{kj}^2$. □

Theorem 3. $E(\widehat{N}_{\text{total}}) = N_{\text{total}}, \text{Var}(\widehat{N}_{\text{total}}) = (mn-1) N_{\text{total}}$.

Proof. We write the random variable

$$\widehat{N}_{\text{total}} = m \sum_{k \in \{k | n_k > 0\}} \widehat{N}_k = mn \sum_{k=1} n_k = mn \sum_{k=1} \sum_{j=1}^{N_k} w_{i(k,j)} u_k.$$

By Lemma 4, we have

$$E(\widehat{N}_{\text{total}}) = \sum_{k=1} N_k = N_{\text{total}},$$

$$\text{Var}(\widehat{N}_{\text{total}}) = (mn)^2 \sum_{k=1} \sum_{j=1}^{N_k} \text{Var}(w_{i(k,j)} u_k)$$
$$= (mn-1) N_{\text{total}}. \quad \square$$

Theorem 4. $E(\hat{x}) = x, \mathrm{Var}(\hat{x}) = (mn-1)\sum_k\sum_j x_{kj}^2$.

Proof. We write the random variable

$$\hat{x} = m\sum_{k\in\{k|n_k>0\}}\hat{x}_k = mn\sum_k\sum_{l}^{n_k}X_{kl} = mn\sum_k\sum_{j=1}^{N_k}w_{i(k,j)}u_kx_{kj}.$$

By Lemma 4, we have

$$E(\hat{x}) = \sum_k\sum_{j=1}^{N_k}x_{kj} = x,$$

moreover,

$$\mathrm{Var}(\hat{x}) = (mn-1)\sum_k\sum_j x_{kj}^2.$$

These theorems show the estimators are unbiased. □

4.3. Estimation of flow length distributions

Let $g = \{g_j : j = 1,2,\ldots,n\}$, where g_j is the frequency of sampled flows with j packets, denote a set of sampled flow length frequencies after double sampling, and let $f = \{f_i : i = 1,2,\ldots,n,\ldots\}$, where f_i is the frequency of original flows with i packets, denote a set of original flow length frequencies. Our objective is inferring $\{f_i\}$ from $\{g_j\}$.

Firstly, we make recovery of flow sampling as follows:

$$\acute{g}_j = m * g_j, \quad j = 1,2,\ldots,n. \tag{5}$$

We regard $\acute{g} = \{\acute{g}_j : j = 1,2,\ldots,n\}$ as sampled flow length frequencies with packet sampling rate $1/n$ from original flows. In practice, measured sampled flow length distributions are smoother, so some effective manner of smoothing would be required for long flows. According to Eq. (5), the inferred distribution of flow lengths would be concentrated on length j. However, we know that flow length distributions have the property of being heavy-tailed, i.e., the number of long flows is very few. For example, for flow sampling rate $1/m = 1/100$, if we only sample a flow of length $j = 10{,}000$ in length interval [9900, 11,000], i.e., $g_{10{,}000} = 1$, and $g_j = 0$, for $j = 9900,\ldots,11{,}000$, we should not think that there are 100 flows of length 10,000, rather than think that there are 100 flows in length interval [9900, 11,000]. So we estimate as 100 flows of different lengths, not 100 flows of the same length. After they are smoothed, we obtain

$$\bar{g} = \{\bar{g}_j, j = 1,2,\ldots\}. \tag{6}$$

Under independent sampling of packets with probability $p = 1/n$, the number of packets j sampled from an original flow of i packets follows the binomial distribution $B_p(i,j) = \binom{i}{j}p^j(1-p)^{i-j}$. Let $\gamma = \sum_j\bar{g}_j$, and ϕ_i denote the frequencies of original flows of length i conditional on at least one of its packets being selected, and $\sum_i\phi_i = 1$. Our aim is to estimate $\phi = \{\phi_i\}$ from the frequencies $\{\bar{g}_j\}$. We now derive an expression for log-likelihood $L(\phi)$ to obtain $\bar{g}_i$ given ϕ. Here, $c_{ij} = B_p(i,j)/(1-B_p(i,0))$ is the probability that packets are sampled from a flow of length i, conditional on $j \geqslant 1$, i.e., that the flow is sampled. For any j, its probability function is $(\sum_{i\geqslant j}\phi_ic_{ij})^{\bar{g}_j}$. Hence we obtain the likelihood function $\prod_{j\geqslant 1}(\sum_{i\geqslant j}\phi_ic_{ij})^{\bar{g}_j}$. Therefore the logarithm of likelihood function is

$$L(\phi) = \sum_{j\geqslant 1}\bar{g}_j\log\sum_{i\geqslant j}\phi_ic_{ij}. \tag{7}$$

Now we adopt the EM algorithm in [13]; the standard form is as follows.

Starting with an initial value $\phi^{(0)}$, for example, $\phi^{(0)} = \{\frac{\bar{g}_i}{\gamma}\}$, the algorithm finds $\sup\{L(\phi) : \phi \in \Delta\}$, by iterating between the following two steps ($k = 0,1,\ldots$):

E step Let f_{ij} denote the frequencies of original flows of length i from which j packets are sampled. Thus $\bar{g}_j = \sum_i f_{ij}$, while $\bar{f}_i = \sum_j f_{ij}$ is the frequency of original flows of length i at least one of whose packets is sampled. Form the complete data likelihood function assuming known f_{ij}

$$L_c(\phi) = \sum_{i\geqslant j\geqslant 1}f_{ij}\log\phi_ic_{ij}. \tag{8}$$

Form the expectation $Q(\phi,\phi^{(k)})$ of $L_c(\phi)$ conditional on the known frequencies $\bar{g}_j$, according to a distribution $\phi^{(k)}$:

$$Q(\phi,\phi^{(k)}) = \sum_{i\geqslant j\geqslant 1}E_{\phi^{(k)}}[f_{ij}|\bar{g}]\log\phi_ic_{ij}. \tag{9}$$

M step Define $\phi^{(k+1)} = \mathrm{argmax}_{\phi\in\Delta}Q(\phi,\phi^{(k)})$. From the Legendre equations in the maximization of $Q(\phi,\phi^{(k)})$ we have $\phi_i^{(k+1)} = \frac{E_{\phi^{(k)}}[f_{ij}|\bar{g}]}{\gamma}$. Through direct computation of the above conditional expectation we obtain

$$\phi_i^{(k+1)} = \frac{1}{\gamma}\sum_{i\geqslant j\geqslant 1}\frac{\phi_i^{(k)}c_{ij}\bar{g}_j}{\sum_{l\geqslant j}\phi_l^{(k)}c_{lj}}. \tag{10}$$

Iterate steps E and M until some termination criterion is satisfied. Let $\bar{\phi}$ denote the termination point. We write our estimation of original flows as $f_i = \bar{\phi}_i\gamma/(1-B_p(i,0))$.

5. Evaluation

In this section we apply the estimators derived in the previous section to experimental traffic traces. We infer the flow statistics from the sampled versions of the traces, and we compare them with the unsampled flow statistics of the original traces.

For flow estimating, we adopt estimated relative difference (ERD) as our evaluation metric. Suppose that A is the actual number and E is the number estimated. Then ERD is defined as follows: $\mathrm{ERD} = \frac{A-E}{A}$. For flow length distribution estimating, we adopt the weighted mean relative difference (WMRD) from [13] as our evaluation metric. 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 $\mathrm{WMRD} = \frac{\sum_i|n_i-\hat{n}_i|}{\sum_i\left(\frac{n_i+\hat{n}_i}{2}\right)}$.

5.1. Data considerations

We use 20 traces in our experiments. The first 10 traces, each of which contains packets during 1 min period, is from the first publicly available 10 Gigabit Internet backbone packet header trace from NLANR: Abilence III data set [14]. It was collected on June 1st, 2004 at the OC192c

Packet-over-SONET link from Internet2's Indianapolis (IPLS) Abilene router node towards Kansas City (KSCY). The other 10 Traces, either of which contains packets during 1 min 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 1000 Mbps; mean traffic per day is 587 Mbps. For each trace, we make double sampling at double sampling period $n * m = 2 * 2, 2 * 5, 2 * 10, 5 * 2, 5 * 5, 5 * 10, 10 * 2, 10 * 5, 10 * 10$, respectively.

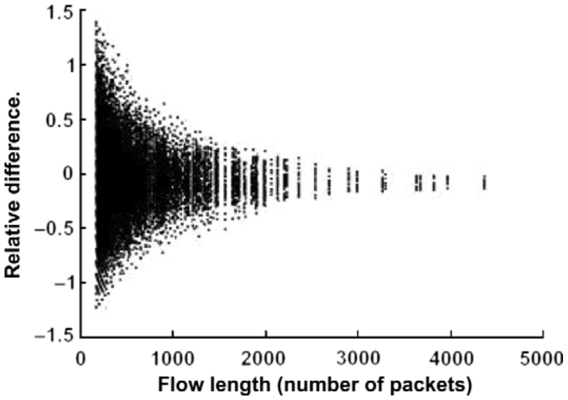

Fig. 1. Estimated relative differences of sampled flows in double sampling period $10 * 10$ for a trace.

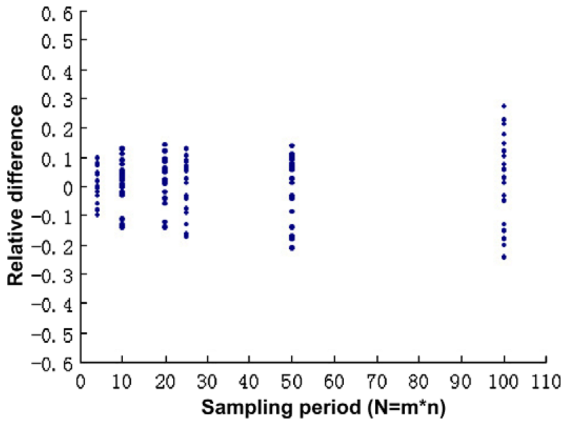

Fig. 2. Estimated relative differences of total packet numbers in different sampling period for 20 traces.

5.2. Estimation comparison

Firstly, the estimated relative differences of packet numbers of sampled flows are computed and shown in Fig. 1. Fig. 1 reflects the estimated relative differences of packer numbers of sampled flows for some trace. We can see that the points are symmetrically distributed around coordinate y. Thus, the estimation is unbiased in accordance with Theorem 1. As shown in Fig. 2, the estimation of total packet number conforms to the property of being unbiased showed by Theorem 3. For space limit, we omitted the corresponding traffic bytes results which are similar to Figs. 1 or 2.

We then compare double sampling with packet sampling alone in estimating flow distributions. For same trace in same sampling period ($N = n * m, nm$), we run double sampling and packet sampling respectively. Then we use EM algorithm to estimate flow distributions from sampled statistics, respectively. Comparing with the actual flow distributions, we find that the estimated results by double sampling are always more accurate than those by packet sampling. This conforms to the conclusion that inversion based on flow sampling performs well [8]. Fig. 3 compares the two estimators of Jiangsu trace derived by double sampling and packet sampling at sampling period $N = 10 * 10, 100$. Observe that estimated result by packet sampling is much worse. Table 1 shows the flow length distribution estimation of double sampling is much more accurate that of packet sampling alone.

5.3. Estimation accuracy and scalability

In this subsection, we consider how to choose n and m for fixed $N = n * m$. In fact, this is the choice on the estimation accuracy and on scalability with link speed. For fixed N, n is decreasing as m increases, and vice versa. When

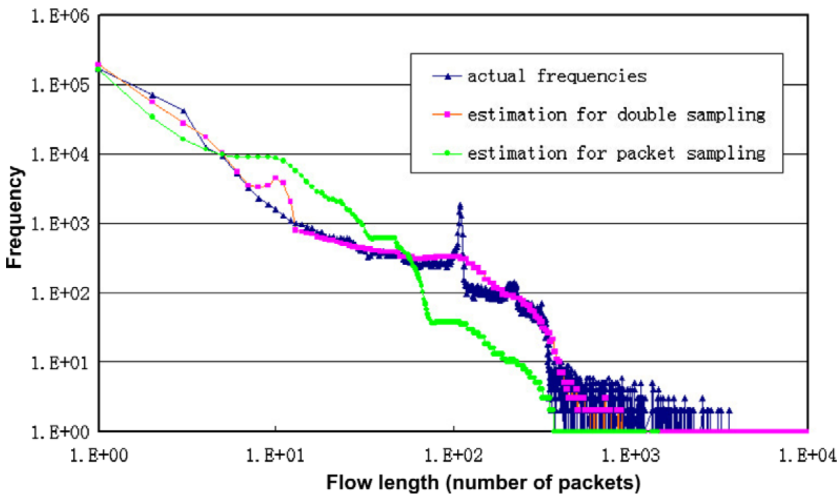

Fig. 3. Comparison of double sampling and packet sampling alone at sampling period $N = 10 * 10, 100$ for Jiangsu trace.

Table 1
WMRD of flow length distribution estimation for double sampling and packet sampling alone

Trace	Sampling period $n * m, N$	WMRD of double sampling	WMRD of packet sampling alone
Abilence III	2 * 2,4	4%	7%
	2 * 5, 10	5%	18%
	5 * 2, 10	9%	18%
	5 * 5, 25	11%	22%
	5 * 10, 50	13%	26%
	10 * 10, 100	20%	37%
Jiangsu	2 * 2, 4	5%	12%
	2 * 5, 10	8%	23%
	5 * 2, 10	10%	23%
	5 * 5, 25	11%	28%
	5 * 10, 50	13%	34%
	10 * 10, 100	19%	39%

$n = 1$, double sampling becomes flow sampling alone. In this case, scalability with link speed is worst because each packet is processed, but the estimation of flow length distributions is best. On the contrary, double sampling becomes packet sampling alone when $m = 1$. The scalability is best, but the estimation accuracy is worst. Therefore, choosing proper m and n is the trade-off between the accuracy and the scalability in practical application. Double sampling is very flexible in making appropriate trade-offs between the accuracy and the scalability.

6. Conclusions

Double sampling has been proposed to overcome the shortcomings of packet sampling alone and flow sampling alone. This method is not only simple to implement but also scalable for high speed links. Theoretical analysis demonstrates that the estimation of sampled flow is unbiased. In experiments, the advantage of double sampling is shown in estimating the distribution of flow length; i.e., the accuracy of flow distribution measurement is dramatically improved.

References

[1] K.C. Claffy, G.C. Polyzos, H.-W. Braun, Application of sampling methodologies to network traffic characterization, ACM SIGCOMM (September) (1993) 194–203.
[2] I. Cozzani, S. Giordano, A passive test and measurement system: traffic sampling for QoS evaluation, in: Global Telecommunications Conference, 1998. GLOBECOM 1998. The Bridge to Global Integration, vol. 2, IEEE, pp. 1236–1241.
[3] Packet Sampling (psamp). <http://www.ietf.org/html.charters/psamp-charter.html>, 2005.02.02.
[4] N.G. Duffield, M. Grossglauser, Trajectory sampling for direct traffic observation, IEEE/ACM Trans. Network. 9 (3) (2001) 280–292.
[5] N.G. Duffield, M. Grossglauser, Trajectory sampling with unreliable reporting, IEEE Infocom 2004, March 2004.
[6] Sampled Cisco. <http://www.cisco.com/en/US/products/sw/iosswrel/ps1829/products_feature_guide09186a0080081201.html>, 2002.12.8.
[7] NeTraMet Version 4.4. <http://www2.auckland.ac.nz/net/Accounting/ntm.Release.note.html>, 2002.12.
[8] Nicolas Hohn, Darryl Veitch, Inverting sampled traffic, in: Internet Measurement Conference 2003. October 27–29, Miami Beach, Florida, USA. pp. 222–233.
[9] K.C. Claffy, H.W. Braun, G.C. Polyzos, A parameterizable methodology for Internet traffic flow profiling, IEEE JSAC 13 (1995) 1481–1494.
[10] R. Jain, S.A. Routhier, Packet trains-measurements and a new model for computer network traffic, IEEE JSAC 4 (1986) 986–995.
[11] P. L'Ecuyer, Efficient and portable combined random number generators, Communications of the ACM, 31 (1988) 742–749 and 774.
[12] M.V. Ramakrishna, Practical performance of Bloom Filters and parallel free-text searching, Communications of the ACM 32 (10) (1989) 1237–1239.
[13] N.G. Duffield, C. Lund, M. Thorup, Estimating flow distributions from sampled flow statistics, ACM SIGCOMM (2003) 325–336.
[14] NLANR: Abilene-III data set. <hppt://pma.nlanr.net/Special/ipls3.html>.

Weijiang Liu, male, born in 1969, received the Ph.D., degree in computation mathematics from Jilin University, Changchun, China, in 1998. From June, 2004 to June, 2006, he was a postdoctoral fellow of Post Doctoral Station for Computer Science and Technology, Southeast University, China. He is currently a professor in school of Computer Science and Technology, Dalian Maritime University, China. He has published more than 40 papers and his current research interests include network measurement and network performance.

Jian Gong, male, born in 1957, received the B.S. degree in computer science from Nanjing University, and the M.S. and the Ph.D., degrees in computer science from Southeast University, China. He is a professor in School of Computer Science and Engineering, Southeast University. He has published more than 100 papers and his current research interests include network measurement, network performance and network security.