This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2007 proceedings.

Sampled Based Estimation of Network Traffic Flow Characteristics

Lili Yang

Department of Statistics
University of Michigan
Ann Arbor, MI 48109

George Michailidis

Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Abstract— **In this paper, we consider the problem of non-parametric estimation of network flow characteristics, namely packet lengths and byte sizes, based on sampled flow data. We propose two different approaches to deal with the problem at hand. The first one is based on single stage Bernoulli sampling of packets and their corresponding byte sizes. Subsequently, the flow length distribution is estimated by an adaptive expectation-maximization (EM) algorithm that in addition provides an estimate for the number of active flows. The estimation of the flow sizes (in bytes) is accomplished through a random effects regression model that utilizes the flow length information previously obtained. A variation of this approach, particularly suited for mixture distributions that appear in real network traces, is also considered. The second approach relies on a two-stage sampling procedure, which in the first stage samples flows amongst the active ones, while in the second stage samples packets from the sampled flows. Subsequently, the flow length distribution is estimated using another EM algorithm and the flow byte sizes based on a regression model. The proposed approaches are illustrated and compared on a number of synthetic and real data sets.**

I. INTRODUCTION

Several network management and engineering tasks, such as quality of service provisioning, usage based accounting, traffic profiling and control, fault detection and service level agreement verification (SLA), require traffic measurements [5], [4]. However, the collection of the necessary information on every packet becomes prohibitive in terms of processing capacity, cache memory and required bandwidth in today's high speed links. Packet sampling techniques have emerged as a scalable alternative to address this problem, as witnessed by recent recommendations in the Internet Engineering Task Force working groups [7] and its implementation in high-speed routers [1]. An overview of applications where sampling proves useful for passive Internet measurements is given in [4].

Understanding the characteristics of traffic flows is crucial for allocating the necessary resources (bandwidth) to accommodate users demand. The problem of using sampled flow statistics in order to estimate the number of active flows in a link and their packet length and to a lesser extent bytes distribution has attracted some attention in the literature (see for example [3], [11], [8], [6] and references therein). In this paper, we extend our work [14] on estimating flow length (in terms of number of packets of flows) distributions to flow size (measured in bytes) distributions and also focus on mixture distributions that characterize many real network traces. Specifically, the maximum likelihood estimator for the number of active flows and the flow length (in packets) and size (in bytes) distribution is obtained based on sampled packet data. While several papers have looked at estimating the flow length distribution, the flow size one has hardly received any attention in the literature. Further, the length and size of an original flow is estimated from that of the sampled one, through their posterior means. The maximum likelihood estimator is calculated through an adaptive expectation-maximization (EM) algorithm.

Experience with real network traffic traces shows that the actual flow length and consequently flow size distribution are bimodal, with one mode corresponding to the short flows and the second one to long ones. An example of such a scenario with data obtained from the link that connects the campus of the University of North Carolina at Chapel Hill to the Internet is shown in Figure 1. Further, in many cases the number of short flows significantly outnumbers that of long ones. Hence, the corresponding sampling distribution would exhibit similar features. For low sampling rates, this implies that the sampling distribution could be modeled as a mixture distribution with two components. Estimation techniques that do not take into consideration the structure of the underlying distribution perform poorly. We obtain a maximum likelihood estimator for such distributions through a two-stage EM algorithm, by considering a profile likelihood approach. Further, we also propose a maximum likelihood estimator based on an alternative two stage sampling scheme, which in the first stage samples flows and in the second stage samples packets from the already selected flows. Such a sampling scheme avoids the problem of length bias sampling, where longer flows have a higher chance of being selected and hence represented in the sampled data.

The remainder of the paper is organized as follows: in section II, the problem at hand is formulated and the maximum likelihood estimator for the flow length and size distributions introduced, together with estimates of the original flow lengths and sizes. In Section III, the problem of estimating mixture distributions from sampled data is introduced. Two estimation schemes, one based on a two stage EM algorithm and the other on a two stage sampling scheme are discussed. In Section IV, the proposed algorithms are evaluated on synthetic and real data sets. Finally, some concluding remarks are drawn in Section V.

0743-166X/07/$25.00 ©2007 IEEE

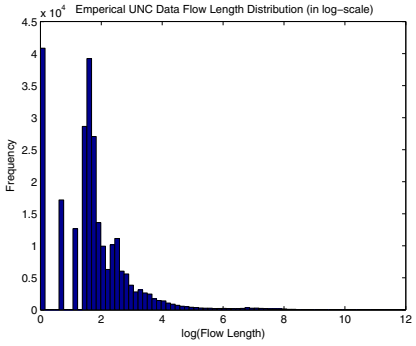

Fig. 1. Flow length distribution

II. PROBLEM FORMULATION

Suppose that on a network link there are M active flows, comprised of $N_m, m = 1, ..., M$ packets each. The number of packets in each flow is referred to as the *flow length*. Further, the payload of each packet consists of $Z_m^i,\ i = 1, ..., N_m$ bytes and hence the size of the m-th flow in bytes is given by $B_m = \sum_{i=1}^{N_m} Z_m^i$, which is referred to as the *flow size*. Packets and their byte sizes are sampled according to a Bernoulli sampling scheme [2]; i.e. each packet is selected with probability p, independent of any other characteristics and its size Z_m^i observed and recorded. Sampled packets can be assigned to a particular flow, by observing its flow key obtained from information available in the packet header [4]. Therefore, the observed data are sampled flow lengths $n_1, n_2, ..., n_r$, and their corresponding flow sizes $b_1, b_2, ..., b_r$, with $b_k = \sum_{i=1}^{n_k} Z_k^i$, where r is the number of sampled flows. Obviously, there is no information about the composition of active flows for which none of their packets are contained in our sample. It is worth noting that an online implementation of such a sampling scheme yields biased samples, since it is more likely to obtain packets from longer flows. We examine this issue in more detail in Section III and investigate an alternative sampling scheme. The objective of this study is threefold: (i) estimate non-parametrically the flow length distribution F of the link, and in addition estimate the original length of sampled flows $N_i, i = 1, 2, ..., r$, (ii) estimate the flow size (expressed in bytes) distribution G and similarly estimate the original flow sizes $B_i, i = 1, 2, ..., r$, and (iii) estimate the number of active flows M in the link. In addition, we would like to provide uncertainty assessments in the form of confidence intervals for these quantities of interest. The proposed model is described next; we start with a brief summary of the component for the packets that is similar to that used in [3], [14].

Let ϕ_i denote the probability that a flow contains i packets and let $\phi = \{\phi_i\}$. Further, let $g_j,\ j = 0, 1, ..., J$ be the frequency of sampled flows of size j, with J being the total number of different sampled flow sizes in all the observed flows. Note that g_0 is not observed, since it corresponds to the frequency of unsampled flows.

Let $M = \sum_{j=0}^{J} g_j$; it can be seen that it will be an estimate for the true number of flows in the link. Further, $r = \sum_{j=1}^{J} g_j$ gives the total number of sampled flows. Let c_{ij} denote the probability of having j packets sampled, given that the true flow length is i packets. We then have that $p_{ij} = \phi_i c_{ij}$ represents the probability that an original flow contains i packets and $j \leq i$ of them have been sampled. Finally, let f_{ij} be the frequency of flows of length i with j packets sampled; it can then be seen that $g_j = \sum_i f_{ij}$.

We can then postulate the following joint model for the number of original flows and the probability that they contain i packets:

$$L(\phi_i, M) = \binom{M}{g_0, g_1, ...g_J} \prod_{j \geq 0} (\sum_{i \geq j} \phi_i c_{ij})^{g_j} \qquad (1)$$

where $M = \sum_{j=0}^{J} g_j$. The objective becomes to maximize this likelihood function subject to the following constraints: $\sum_i \phi_i = 1$, and $\phi_i \geq 0$, where $i \in \{i(0), i(1), ..., i(J)\}$, with $i(j)$ denoting the length of a flow being i packets when j of them have been sampled. In the present setting, we choose $i(0) = \frac{1}{2p}$ instead of 0, since an original flow containing 0 packets is rather meaningless. Also, the possible flow lengths values, denoted by S_I, are restricted to integers closest to j/p.

The above formulation *jointly* estimates the number of active flows (by treating M as a nuisance parameter) and the packet length density. The problem of maximizing the likelihood function given in (1) is computationally hard, due to presence of the constraints. However, by using the Expectation-Maximization algorithm [9] the task becomes manageable. The main idea behind the EM algorithm is to impute the frequencies of all the active flows whose length is i packets (E-step) and subsequently maximize the likelihood over the parameters of interest (M-step). In [14], it is shown that the k-th iteration these two steps are as follows:

E-step:

$$E_{\phi^{(k)}}(f_{ij} | g_j, j = 1, 2, ..., J) = g_j p_{i|j}, \qquad (2)$$

where

$$p_{i|j} = \frac{\phi_i^{(k)} c_{ij}}{\sum_{l \in S_I, l \geq j} \phi_l^{(k)} c_{lj}} \qquad (3)$$

is the probability that the flow with sample size j contains actually i packets in total.

M-step:

$$\phi_i^{(k+1)} = \frac{\sum_{i \geq j \geq 1} g_j p_{i|j} + \hat{g_0}^{(k)} p_{i|0}}{\sum_{i \in S_I} (\sum_{i \geq j \geq 1} g_j p_{i|j} + \hat{g_0}^{(k)} p_{i|0})}, \qquad (4)$$

where $p_{i|j}$ is defined as the probability that for a flow of length i, j of its packets have been sampled.

Based on the estimates of the flow length distribution $\hat{\phi}$, an estimate of the the posterior probability distribution of a flow being of length i given that k of its packets have been sampled is obtained by

$$f(i|k) = \frac{c_{ik} \hat{\phi}_i}{\sum_{i \in S_I} c_{ik} \hat{\phi}_i}. \qquad (5)$$

The posterior mean, which is the weighted average of all possible flow lengths, is given by

$$\hat{N(k)} = \mathrm{E}(N(k)) = \sum_{i \in S_I} i f(i|k). \quad (6)$$

and is a good estimator for the original length of a sampled flow.

The next task is to estimate both the flow size distribution G (in bytes) and the size of the active flows $B_i, , i = 1, 2, ..., r$. The following observation motivates the proposed estimation procedure. One can view the sampling process of flow sizes as a one-stage cluster sampling scheme [2]; therefore, the classical estimator is given by $\hat{B}_k = (\sum_{i=1}^{n_k} Z_k^i)/p$. Subsequently, one can estimate non-parametrically the distribution G, using the estimates B_k as data. The above procedure works well for light tailed distributions, but proves problematic for heavy tailed ones, due to the inherent length biasedness of the sampling scheme.

In order to overcome this deficiency, we propose an estimation procedure based on the following regression model:

$$b_j = \gamma_0 + \gamma_1 j + \epsilon, \text{ for all } j;$$

i.e. the sampled size of the flow in bytes is a linear function of the number of observed packets from that flow. The parameter γ_0 corresponds to an estimate in bytes of non-sampled flows, while γ_1 captures the correlation between average flow length and size. By obtaining the least squares estimates $\hat{\gamma}_0$ and $\hat{\gamma}_1$, we can subsequently estimate (predict) the flow sizes from

$$\hat{B_j} = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{i}(j),$$

where $\hat{i}(j)$ is the estimated number of packets for a flow whose sampled length is j packets.

An extension of the regression model is given by

$$b_j = \gamma_0^j + \gamma_1 j + \epsilon,$$

where the baseline estimate γ_j^0 becomes flow specific. The corresponding estimate of the flow size becomes

$$\hat{B_j} = \hat{\gamma}_0^j + \hat{\gamma}_1 \hat{i}(j).$$

We proceed to estimate the distribution of the flow sizes G, using the estimates $\hat{B}_j$ that provide a collection of possible flow sizes with corresponding support S_B. A naive non-parametric estimate [15] of G would be based on the relative frequencies of the data, but in practice exhibits poor performance. A better estimator is based on an adaptive EM algorithmic scheme and provides superior estimates of the flow size distribution, as demonstrated in section IV. Since the sampling scheme for the flow sizes is Bernoulli, the previously described EM algorithm is also suitable for the flow size distribution G. However, when we restricted attention to flow lengths only, the collection of possible flow lengths S_I (i.e. the support of the distribution F) remained fixed throughout the iterations. In the original EM algorithm, ϕ was estimated based on S_I, which is not the same as $\hat{i}|j$; however, the support S_B of the flow size distribution depends only on $\hat{i}|j$. Hence, the difference in the supports of the two distributions make it difficult to utilize flow length information in the estimation of the flow size distribution. However, by updating each $i|j$ in S_I by $\hat{N}(j)$ as in equation (6), where $i^{(0)}|j$ is initialized by j/p as before, the two supports are made comparable. Consequently, by updating the support S_I within the iterations, the support S_B of the flow size distribution coincides with the updated one for the flow length S_I. It can be shown that $\hat{\phi}$ is the maximum likelihood estimate of the flow byte distribution as well.

In Figure 2, the quantile-quantile plot of the original flow size vs the estimated one (sampling rate $p = 0.01$) for 1,000 Poisson flows of mean length 5,000 and uniform size distribution with support $[1200, 1500]$ using the adaptive EM algorithm is shown. It can be seen that with the exception of the tails of the distribution, the estimate is highly accurate.

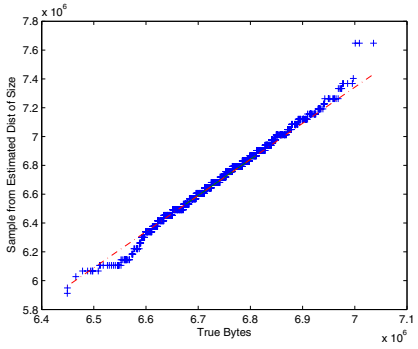

Fig. 2. Quantile-quantile plot of the true vs the estimated flow size for 1,000 Poisson flows with uniformly distributed byte sizes

A. Statistical Inference

We briefly discuss next some of the asymptotic properties of the derived maximum likelihood estimator. First, notice that the parameter space Ω of the posited multinomial model is closed and compact. In addition, the model satisfies the following: for all $\omega \neq \theta$ in the parameter space Ω, the distribution $P_\omega \neq P_\theta$. In order to establish consistency, the following condition needs to be satisfied [9]

$$\mathrm{E}_\theta \sup_\Omega W < \infty, \quad (7)$$

where $W(\omega) = \log[\frac{L(\omega)}{L(\theta)}]$. In this context, θ is the parameter vector of the true number of flows M and probabilities ϕ, while $\omega = [\tilde{M}, \tilde{\phi}] \in \Omega$ is another set of values of these parameters. The above condition holds for the model, since

$$\mathrm{E}_\theta \sup_\Omega W = \mathrm{E}_\theta \sup_\Omega [l(\omega) - l(\theta)] = \sup_\Omega l(\omega) - \mathrm{E}_\theta l(\theta)$$

$$l(\omega) = \log[\binom{\tilde{M}}{g_0, g_1, ...g_J}] + \sum_{j \geq 0} g_j \log(\sum_{i \geq j} \tilde{\phi}_i c_{ij}) \quad (8)$$

with both $\tilde{\phi}_i$ and c_{ij} being between 0 and 1. Further, $\sum_{i \geq j} \tilde{\phi}_i c_{ij}$ gives the estimated probability of having a sampled flow of length j under a model with parameter vector ω. Hence, $\log(\sum_{i \geq j} \tilde{\phi}_i c_{ij}) \leq 0$, which implies that the second component of (8) is bounded above by 0. On

the other hand, the first component of (8) is also finite assuming that $n \in \Omega$ is upper-bounded. Consequently, the log-likelihood for a finite number of original flows, $l(\omega)$, is always less than ∞. $\mathrm{E}_\theta l(\theta) = \sum_x log(h(x))h(x)$, where x is a multinomial random vector with parameter θ, and $h(x)$ is the corresponding density function. Therefore, $\mathrm{E}_\theta l(\theta) = \sum_x [\log(\begin{matrix} M \\ x \end{matrix}) + \sum_j x_j \log(\sum_i \phi_i c_{ij})]h(x)$. Since $\sum \phi_i = 1$, and $c_{ij} > 0$, $\sum_i \phi_i c_{ij}$ never reaches 0. Hence, $\mathrm{E}_\theta l(\theta) > -\infty$, followed by $\mathrm{E}_\theta \sup_\Omega W < \infty$. We can then conclude that the maximum likelihood estimator converges to the true parameter vector θ; i.e. $(\hat{\phi}, \hat{M}) \rightarrow^p \theta$

Given the consistency result, we then need to calculate the Fisher Information Matrix for the parameters of interest, in order to obtain the asymptotic distribution for the estimated flow length distribution, and subsequently calculate the variance of $\hat{N(k)}$. We implemented Louis' [10] procedure for obtaining the observed information matrix when using the EM Algorithm. Gradient vectors of the complete log-likelihood take the following form:

$$S = (\frac{f_{i(o)}}{\phi_{i(o)}}, ..., \frac{f_{i(J)}}{\phi_{i(J)}})'. \tag{9}$$

Then, the observed Information Matrix is given by

$$I_{obs} = \sum_{i=1}^{r} S(\hat{f}_i, \hat{\phi}, \hat{M}) S(\hat{f}_i, \hat{\phi}, \hat{M})', \tag{10}$$

where $\hat{f}_i$ is calculated from $E_{\hat{\phi}}[\sum_j f_{ij} | f_{ij} \in R]$ with f_{ij} being the indicator matrix stemming from the multinomial model. Finally, R is defined as all the sets of complete data which yield the same result for the incomplete data, i.e. $R(k) = \{f_{ij} : y(f_{ij}) = e_k\}$, where e_k is a $(J+1)$ dimensional indicator vector with all the elements, except the kth one, being 0.

The resulting Fisher Information Matrix is symmetric with the lower triangular component given by

$$\begin{bmatrix} \frac{\sum_{j=1}^{J} p_{i(0)|j}^2 g_j}{\hat{\phi}_{i(0)}^2} & & & \\ \frac{\sum_{j=1}^{J} p_{i(0)|j} p_{i(1)|j} g_j}{\hat{\phi}_{i(0)} \hat{\phi}_{i(1)}} & \frac{\sum_{j=1}^{J} p_{i(1)|j}^2 g_j}{\hat{\phi}_{i(1)}^2} & & \\ \ldots & \ldots & & \\ \frac{\sum_{j=1}^{J} p_{i(0)|j} p_{i(J)|j} g_j}{\hat{\phi}_{i(0)} \hat{\phi}_{i(J)}} & \ldots\ldots & \frac{\sum_{j=1}^{J} p_{i(J)|j}^2 g_j}{\hat{\phi}_{i(J)}^2} \end{bmatrix} \tag{11}$$

Standard results yield that the asymptotic distribution of the estimated flow length distribution is multivariate normal; i.e.

$$\sqrt{r}(\hat{\phi} - \phi) \Rightarrow \mathrm{N}(0, I_{obs}^{-1}) \tag{12}$$

Recall that the posterior mean estimator for the an original flow length given a sampled one of length k is defined to be

$$H_{\hat{\phi}}(k) \equiv \hat{N(k)} \equiv \mathrm{E}(N(k)) = \frac{\sum_{n \in S_I} n f(k|n) \hat{\phi}(n)}{\sum_{n \in S_I} f(k|n) \hat{\phi}(n)},$$

and is continuous in ϕ. By an application of the Delta method [12], we then get that

$$\mathrm{Var}(H_{\hat{\phi}}) = \nabla H_{\hat{\phi}} I_{obs}^{-1} \nabla H_{\hat{\phi}}'. \tag{13}$$

The continuity of the above defined functional in ϕ gives that

$$\sqrt{r}(\hat{N(k)} - N(k)) \Rightarrow \mathrm{N}(0, \mathrm{Var}(H_{\hat{\phi}})). \tag{14}$$

The significance of this result is, that simultaneous confidence intervals for the original flow lengths given sampled ones, can be constructed, thus providing a measure of uncertainty about the obtained estimates. An example of such a confidence band for the estimated length of 100 Poisson flows sampled at a rate $p = 0.01$ is shown in Figure 3.

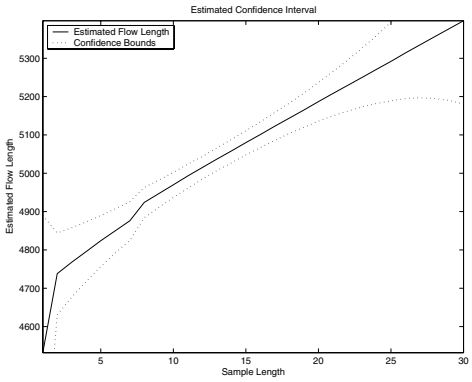

Fig. 3. Confidence band for 100 Poisson flows, sampled at a rate $p = 0.01$

III. ESTIMATING MIXTURE DISTRIBUTIONS FROM SAMPLED DATA

As discussed in the introductory section, experimental evidence strongly suggests that the proposed maximum likelihood estimator does not perform well with several real network traces, since both the packet length and consequently the byte size distributions are mixtures of two components; the first component, representing short flows, and the second one representing considerably longer flows. Figure 1 illustrated this point. In particular, the estimate of the number of active flows proves highly problematic, due to the severe nature of the biased length sampling issue in such a setting; namely, the first component of the distribution is heavily under-sampled. We propose next two procedures that deal with the problem of estimating mixture distributions. The first approach is based on the original Bernoulli sampling scheme, but employs a two-stage EM algorithm suitable for mixture distributions. The second approach looks at the problem from a different point of view and relies on an alternative sampling scheme.

It is assumed that the original flow length distribution F (and consequently the flow size one G) is a mixture of two components; i.e.

$$F = \alpha F_1 + (1 - \alpha) F_2,$$

with $\alpha \in (0, 1)$. As explained above, this would also translate to a mixture distribution for the sampled flow length distribution; namely

$$F^s = \alpha F_1^s + (1 - \alpha) F_2^s.$$

The focus on two mixtures is because this is the most common distribution for sampled flow lengths that appears in real network samples. However, the proposed algorithms can handle more complicated mixtures, at a significantly higher computational cost. In the remainder of the section, we will focus on the case where the first component is a point mass at 1; i.e. $F_1^s \equiv \delta_1$, since with a small sampling rate p at most one packet would be sampled from short flows. This assumption simplifies somewhat the derivations, but nevertheless the proposed two-stage EM algorithm works for the general case provided that the two components are adequately separated [15].

The main idea behind the proposed algorithm is as follows: in the first stage, the adaptive EM algorithm previously described is used to estimate the relative frequencies ϕ, the support of the distribution S_I and the number of active flows. In the second stage, another EM algorithm is employed that based on the current estimates of these parameters, estimates the mixing coefficient α. Therefore, the two-stage algorithm splits the parameters of interest into two subsets (blocks) and in each iteration alternates between the blocks by fixing the parameters of the other block in their current values. The theory of block relaxation algorithms [17] guarantees convergence to a local maximum. However, there are some subtle issues that need to be carefully considered. Notice that in the first stage, the sampled information on the point mass component has been included, since it is not possible to directly separate whether a sampled flow of length one comes from the first or the second component. By conditioning, we can separate $g_j(1) = P(1|j=1)g_j^1(1) + P(i \in S_I^2)g_j^2$, where S_I^2 denotes the support of the second component $S_I - \{1\}$. Then, $g_j^2(1)$ is used in the first stage EM algorithm, together with the remaining g_j's. On the other hand, all the g_j's (including g_j^1) are used in the second stage to estimate α. We discuss next the second stage of the EM algorithm that estimates the mixing coefficient α.

EM Algorithm for estimating the mixing coefficient α

The so-called *profile* likelihood function is given by

$$\begin{aligned} L(\alpha) &= \binom{M}{g_0, g_1, \ldots g_J} \prod_{j\geq 0}(f_j)^{g_j} \\ &\sim \prod_{j\geq 0}[\alpha f(j|1) + (1-\alpha) f(j|S_I^2)]^{g_j}. \end{aligned}$$

It is expressed as a mixture of two conditional densities for the two mixture components. The complete data likelihood function can be written as

$$L(\alpha) = \prod_{j\geq 0}(\alpha^{y_{j1}} f(j|1)^{y_{j1}} (1-\alpha)^{y_{j2}} f(j|S_I^2)^{y_{j2}})^{g_j},$$

where y_{jk}, is an indicator variable that identifies the mixture component ($k = 1, 2$) that a sampled flow of length j belongs to. As usual, the complete likelihood function treats the y_{jk}'s as missing data and estimates them as conditional expectations. Notice that $\mathrm{E}(y_{j1}^{(t)}|g_j, \alpha^{(t)}) = \frac{\alpha^{(t)} f(j|1)^{(t)}}{\alpha^{(t)} f(j|1)^{(t)} + (1-\alpha^{(t)}) f(j|S_I^2)^{(t)}}$, where $f(j|S_I^2)^{(t)} = f_2(j)^{(t)} = \sum_{i\in S_I^2} f(j|i)\phi_i^{(t)}$ is the estimated probability that the length of a sampled flow from the second component will be j packets. The steps of the EM algorithm are given next:

Initialize $\alpha^{(0)}$ by the empirical estimate g_1/r.

E-step

$$\begin{aligned} \mathrm{E}_{\alpha^{(t)}}(\log L(\alpha)|g_j) &= (\textstyle\sum_{j\geq 0} \mathrm{E}_{\alpha^{(t)}}(y_{j1}|g_j) g_j) \log\alpha + \\ &\quad (\textstyle\sum_{j\geq 0} \mathrm{E}_{\alpha^{(t)}}(y_{j2}|g_j) g_j) \log(1-\alpha) \end{aligned}$$

M-step

By taking the derivative of $\log L(\alpha)$ with respect to α and setting it to 0 we obtain

$$\hat{\alpha}^{(t+1)} = \frac{\sum_{j\geq 0}(\mathrm{E}_{\alpha^{(t)}}(y_{j1}|g_j) g_j)}{\hat{M}}$$

Iterate E-step and M-step until the convergence criterion is satisfied.

IV. Two-Stage Sampling Scheme

We discuss next an alternative procedure that relies on a different sampling scheme, where in the first stage flows are samples uniformly with probability p_f irrespective of their lengths, while in the second stage, packets are sampled uniformly with probability p_p from the selected during the first stage flows. Such a mechanism is feasible if different flows go through different queues on a router. It can be seen that this two-stage sampling scheme overcomes the difficulty posed by length biased sampling, since each flow has an equal probability of being selected. Notice that another difference compared to the original Bernoulli sampling scheme is that the two tasks of estimating the total number of active flow M, and the flow length/bytes distributions are well separated. A trivial unbiased estimator for the number of active flows is given by $\hat{M} = \frac{r}{p_f}$ [2]. Also note that $g_j \sim \text{Multinomial}(M, f_j)$. The likelihood function under this sampling scheme is given by

$$L(\phi, M) = \binom{M}{g_{ns}, g_0, g_1, \ldots g_J} \prod_j (f_j)^{g_j}, \qquad (15)$$

where g_{ns} is the number of unobserved flows, estimated by $\hat{g}_{ns} = \hat{n} - r$ and g_0 represents the observed number of flows with no packets selected. Another difference between the two schemes is in the f_j's, the probability of having exactly j packets sampled. Under the two-stage sampling scheme, it is given by

$$\begin{aligned} f_j &= \mathrm{P}(j \text{ packet sampled} \mid \text{this flow is selected}) \\ &\quad \times \mathrm{P}(\text{flow is selected}). \end{aligned}$$

Once again, direct maximization of the likelihood function is challenging and hence we resort to using the EM algorithm, whose steps are described next:

(1) *Initialize* $\phi_{i(j)}^{(0)}$ by the corresponding observed frequency of j, i.e,

$$\phi_{i(j)}^{(0)} = \frac{g_j}{\sum_{k=0}^{J} g_k}, \qquad (16)$$

(2) *E-step:* Given the complete set of data (f_{ij}, g_{ns}, n), f_{ij} follows a multinomial distribution with parameters $M =$

$\sum_{i,j\geq 0} f_{ij} + g_{ns}$ and p_{ij}, p_{ns}. The complete data likelihood is then given by

$$L_c(\phi, M) = \prod_{i\geq j\geq 0} (p_{ij})^{f_{ij}} \times (p_{ns})^{g_{ns}}, \quad (17)$$

where $p_{ij} = c_{ijp} p_f \phi_i$, $c_{ijp} = \binom{i}{j} p_p^j (1-p_p)^{i-j}$.
The corresponding log-likelihood is

$$l_c(\phi, M) = \sum_{i\geq j\geq 0} f_{ij} \log(\phi_i c_{ijp} p_f) + g_{ns} \log(p_{ns}). \quad (18)$$

The expectation $Q(\phi, \phi^{(k)})$ of l_c conditional on the known frequencies g_j is given by:

$$\begin{aligned} Q(\phi, \phi^{(k)}) &= \sum_{i\geq j\geq 0} E_{\phi^{(k)}}(f_{ij}|g_j) \log(\phi_i c_{ijp} p_f) \\ &\quad + E_{\phi^{(k)}}(g_{ns}|g_j) \log(1-p_f). \end{aligned} \quad (19)$$

Again, notice that

$$f_{ij}|g =^d f_{ij}|g_j \sim \text{Multinomial}(g_j, p_{i|j}),$$

where

$$p_{i|j} = \frac{\phi_i^{(k)} c_{ijp}}{\sum_{l\in S_I, l\geq j} \phi_l^{(k)} c_{ljp}} \quad (20)$$

is the probability that the flow with sample size j contains actually i packets in total. Therefore,

$$E_{\phi^{(k)}}(f_{ij}|g_j) = g_j p_{i|j}. \quad (21)$$

Meanwhile, $E_{\phi^{(k)}}(g_{ns}|g_j) = \hat{g}_{ns} = \hat{M} - r$.

(3) *M-step:* $\phi^{(k+1)} = \arg\max Q(\phi, \phi^{(k)})$, such that

$$\sum_{i\in S_I} \phi_i = 1, \text{ and } \phi_i(j) \geq 0 \text{ for } i \in S_I. \quad (22)$$

We got

$$\phi_i^{(k+1)} = \frac{\sum_{i\geq j\geq 0} g_j p_{i|j}}{\sum_{i\in S_I} (\sum_{i\geq j\geq 0} g_j p_{i|j})} \quad (23)$$

(4) *Estimate* $i|j$ *and update* S_I*:* $\hat{i}|j = \mathrm{E}(i(j)) = \sum_{k\in S_I} k f(k|j)$

Iterate steps (2)-(4) until the convergence criterion $||\phi^{(k+1)} - \phi^{(k)}|| < \delta$ is satisfied.

In order to obtain estimates of the flow sizes (in bytes) we can apply the various regression models discussed in section II.

V. Performance Assessment

In this section, we provide empirical evidence of the performance of the derived estimators for a variety of simulated and real network traffic traces. Our focus is on mixture distributions and the performance of the two-stage EM algorithm, together with maximum likelihood estimator derived for a two-stage sampling scheme. Naturally, due to lack of space we present selected results, but a comprehensive evaluation is available in [18].

We start by considering simulated flow length data from $M = 1,000$ active flows obtained from a mixture distribution with point mass at 1, while the second component following: (i) uniform with domain $[3000, 7000]$, (ii) Poisson with mean 5,000 and (iii) Pareto with shape parameter 10/9 and scale parameter 500. The parameters for these three distributions in the second component were set so as to match their expected values. The flow sizes were generated for all cases from a uniform distribution with domain $[100, 500]$ bytes. The mixing coefficient α was set to .3, .5 and .7, while the sampling rate to $p = .01$ and .05.

An example of one realization of such a mixture distribution ($\alpha = .7$ and sampling rate $p = .01$) and its estimate, where the second component is Poisson distributed, is shown in Figure 4. It can be seen that the estimate captures very well the support of the original distribution and the mixing coefficient α, as well as the second component. It should be noted that the apparent visual discrepancy between the original and the estimated distributions is mostly to the somewhat large bin size used in the histogram of the original one.

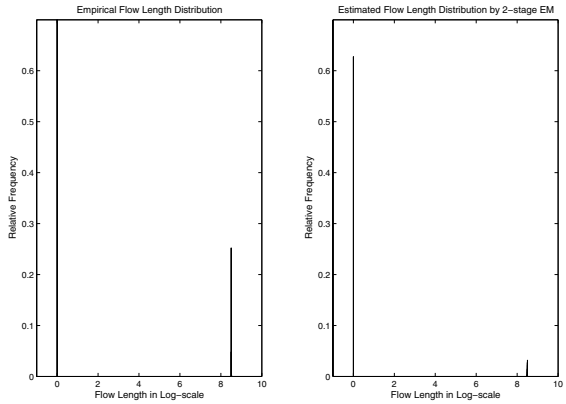

Fig. 4. Comparison of the Empirical flow length distribution and the Estimated flow length distribution through a two stage EM algorithm

In Figure 5, boxplots for the estimates of $\alpha = .7$, the mean flow length (5,000 packets) and the mean flow size (464,000 bytes) of the second component are shown. It can be seen that the estimates are very good, with a fairly narrow range of values.

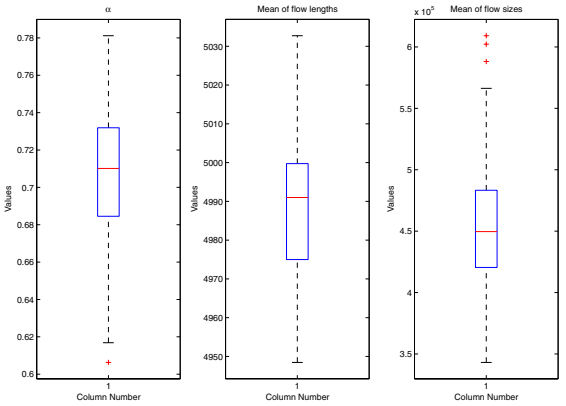

Fig. 5. Boxplots of various parameters of a mixture distribution

In Table I, the mean squared errors (MSE) obtained from 100 replications, for the estimates of the mixing coefficient α, the mean and variance of the flow length (in packets) and size (in bytes) are shown for different mixing coefficients α for the three distributions under consideration.

TABLE I

MEAN SQUARED ERRORS FOR VARIOUS PARAMETER ESTIMATES OBTAINED THROUGH THE 2-STAGE EM ALGORITHM FOR DIFFERENT DISTRIBUTIONS, WITH SAMPLING RATE $p = 0.01$.

α	$\hat{\alpha}$	Mean(P)	Var(P)	Mean(B)	Var(B)
Poisson Distribution					
0.7	0.0053	1.23E+07	5.26E+06	1.69E+10	9.48E+09
0.5	0.0162	1.23E+07	5.26E+06	1.69E+10	9.48E+09
0.3	0.0043	2.26E+06	5.26E+06	3.97E+10	7.77E+10
Uniform Distribution					
0.7	4.35E-06	1.19E+07	3.42E+06	2.19E+10	1.28E+10
0.5	6.19E-06	6.06E+06	5.06E+06	2.53E+10	3.01E+10
0.3	4.38E-06	2.10E+06	3.74E+06	3.90E+10	7.38E+10
Pareto Distribution					
0.7	0.0573	2.13E+07	4.13E+08	1.37E+11	1.22E+13
0.5	0.1801	1.04E+07	2.75E+08	3.63E+11	2.64E+13
0.3	0.4086	5.46E+06	3.47E+07	7.10E+11	1.57E+13

It can be seen that the MSE for the number of active flows is extremely small, indicating a very precise estimate for all values of α. Further, as the contribution of the second component increases (smaller α) the quality of the estimates for both the mean number of packets improves for all distributions. On the other hand, the MSE for the mean number of bytes exhibits the opposite behavior, while the MSE for the corresponding variances are of the similar order for all values of α. These results indicate the difficulty of estimating non-parametrically mixture distributions, especially in the presence of a dominant point mass, and also demonstrate that procedures that ignore the mixture structure would not fare well. Similar conclusions are reached when the sampling rate increases to $p = 0.05$ (results not shown here). The main difference is that due to the higher sampling rate the performance of the estimators improves considerably.

We examine next the performance of the maximum likelihood estimator based on the two stage sampling mechanism for same set of mixture distributions. However, in this case we consider a range of sampling rates for the flows (p_f parameter) while fixing the sampling rate for the packets.

An example of one realization of such a mixture distribution ($\alpha = .7$ and sampling rates $p_f = 0.5$ and $p_p = 0.2$, respectively) and its estimate, where the second component is uniformly distributed, is shown in Figure 6. It can be seen that the estimate captures very well the support of the original distribution and to a large extent the mixing coefficient α.

In Figure 7, boxplots for the estimates of M, the mean flow length (1,500 packets) and the mean flow size (464,000 bytes) of the entire distribution are shown. It can be seen that once again the estimates are very good, exhibiting a fairly narrow range of values.

In Table II, the mean squared errors (MSE) obtained from 100 replications, for the estimates of the number of active flows M, the mean and variance of the flow length

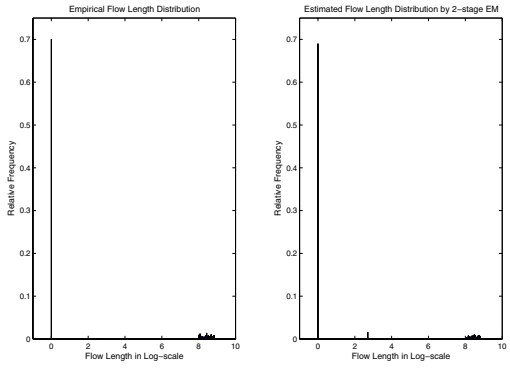

Fig. 6. Estimated flow length distribution, through a two stage sampling scheme

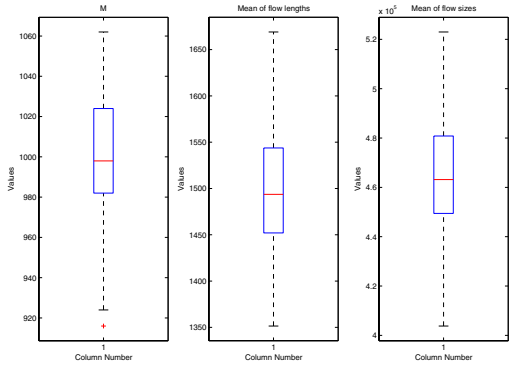

Fig. 7. Boxplots for various parameters of a mixture distribution

(in packets) and size (in bytes) are shown for different flow sampling rates p_f for the three distributions under consideration.

TABLE II

MEAN SQUARED ERRORS FOR VARIOUS PARAMETER ESTIMATES OBTAINED THROUGH THE 2-STAGE SAMPLING SCHEME FOR DIFFERENT DISTRIBUTIONS, WITH SAMPLING PACKET RATE $p_p = 0.01$.

p_f	M	Mean(P)	Var(P)	Mean(B)	Var(B)
Poisson Distribution					
0.05	17232	1.01E+05	2.67E+04	1.05E+10	1.24E+10
0.1	8871	5.12E+04	9.89E+03	6.19E+09	8.04E+09
0.3	1900	9.32E+03	2.07E+03	1.18E+09	5.17E+09
0.5	923	5.00E+03	1.33E+03	6.92E+08	4.44E+09
Uniform Distribution					
0.05	23212	8.96E+04	3.63E+04	1.09E+10	1.75E+10
0.1	8383	4.40E+04	2.31E+04	6.28E+09	1.28E+10
0.3	2110	1.70E+04	6.29E+03	1.59E+09	6.47E+09
0.5	982	6.18E+03	2.83E+03	6.55E+08	6.12E+09
Pareto Distribution					
0.05	18548	7.28E+06	4.88E+08	8.16E+11	5.57E+13
0.1	7491	4.02E+06	4.13E+08	4.54E+11	4.70E+13
0.3	2660	8.97E+05	2.44E+08	1.05E+11	2.85E+13
0.5	865	4.15E+05	1.33E+08	5.51E+10	1.63E+13

It can be seen that as the sampling rates of the flows increases, the quality of the estimates improves, as expected. The reason is that by sampling more flows, we are able

to better capture their characteristics, even when the packet sampling rate is very small. Further, the results are comparable for the estimates of M for all three distributions, but vary for the other parameters. For example, the quality of estimates for the flow lengths and sizes deteriorates for the heavy-tailed Pareto distribution. There is an interesting interplay between the two sampling rates for flows and packets within flow, but at present its full understanding is not available and is a topic of further study. Experience suggests that a flow sampling rate $p_f = .3$ performs well, even when coupled with a very small packet sampling rate, so that the overall rate $p = p_f \times p_p$ remains small.

Similar qualitatively results are obtained for a larger packet sampling rate of 0.05, although the accuracy of the estimates naturally improves. However, the improvements are not usually large enough to compensate for the increased computational complexity both in the data collection and processing.

A. Real data set application

We consider an application of the proposed methods to a real network trace obtained from the router of the Abilene network at Denver in June of 2005. The trace covers a 5-minute period and contains 65,535 active flows. The average flow length consists of 3 packets, but the variance is takes a value of 430. Similarly, the average flow size is 2.7141×10^3 bytes, while the variance is 1.9377×10^{10}. The distributions of the true flow lengths and bytes (in log-scale) for this data set are shown in Figures 8 and 9. It can easily be seen that both distributions are heavy tailed, and to a large extent comprised of two separate components.

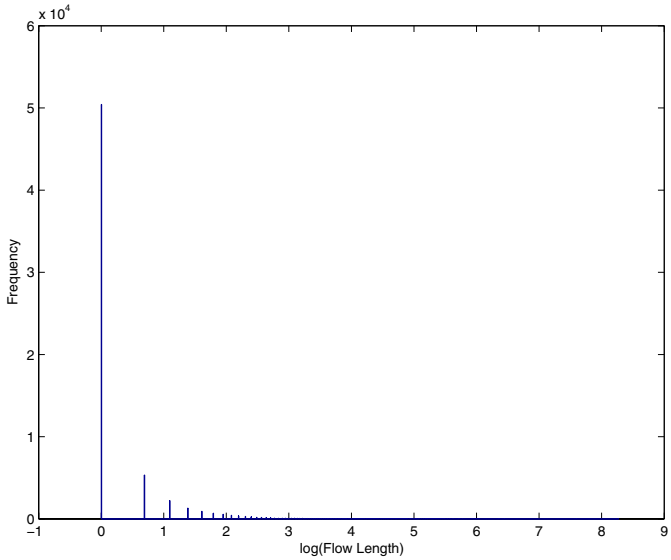

Fig. 8. Empirical flow length distribution (in log-scale) of the Abilene trace

The data were sampled by both mechanisms, and the flow length and size distributions estimated by the proposed EM algorithms. The rate for single stage Bernoulli sampling was $p = 0.01$, while the rates for the two stage mechanism were

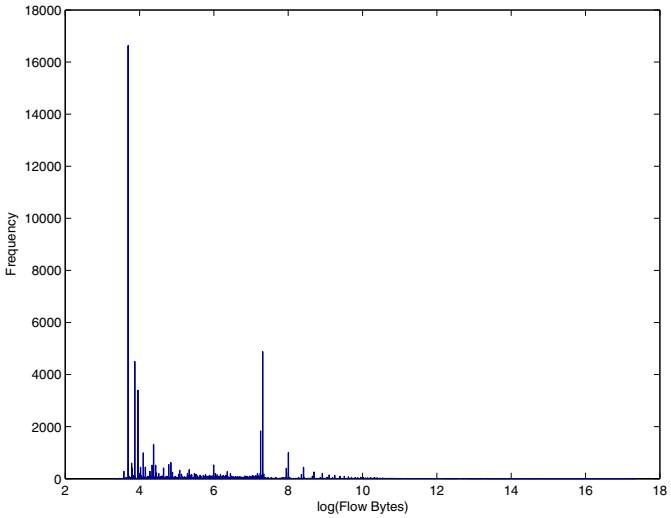

Fig. 9. Empirical flow size distribution (in log-scale) of the Abilene trace

$p_f = 0.05$ and $p_p = 0.2$. The estimate for the the number of active flows under Bernoulli sampling was about 15,000, a severe underestimation of the true value. On the other hand, the estimate under the 2-stage sampling mechanism is almost perfect, $\hat{M} = 66,208$.

In Figures 10-13, the estimates of the flow length and size distributions are shown.

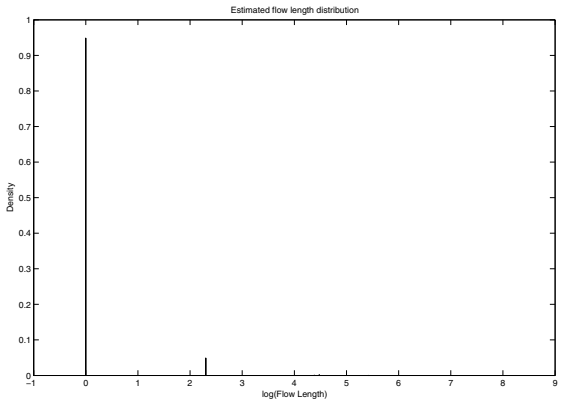

Fig. 10. Estimated flow length distribution (in log-scale) of Abilene trace using a 2-stage EM algorithm

It can be seen that the 2-stage EM algorithm captures well the first component and the support of the flow length distribution, while it focuses on the second spike in the original distribution for the flow sizes. This is mainly due to the fact that due to the severe underestimation of the number of active flows, the algorithm mainly focuses on the second component comprised of larger flows. The estimates produced by the two stage sampling scheme exhibit similar characteristics, although the spike in the flow length distribution is shifted slightly to the right. In this case, the difficulty comes from the fact that the flow sampling rate was set to a fairly small value ($p_f = 0.05$), which is as noted above does not produce

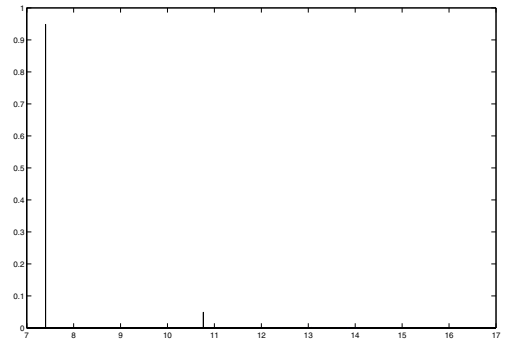

Fig. 11. Estimated flow size distribution (in log-scale) of the Abilene trace using a 2-stage EM algorithm

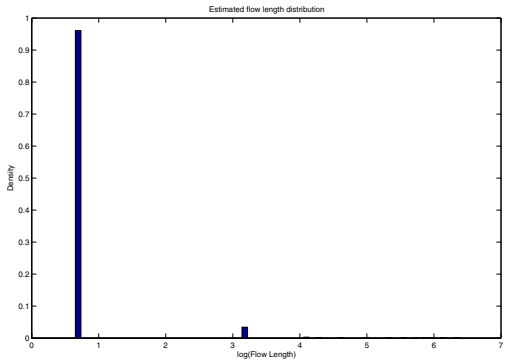

Fig. 12. Estimated flow length distribution (in log-scale) of the Abilene trace using a 2-stage sampling mechanism

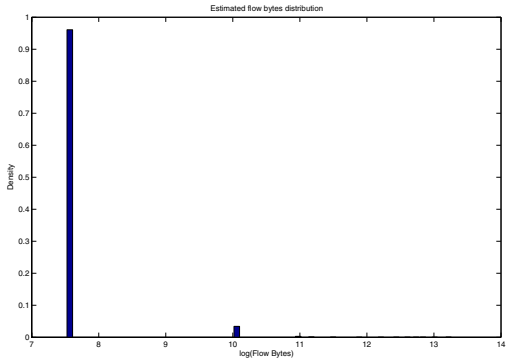

Fig. 13. Estimated flow size distribution (in log-scale) of the Abilene trace using a 2-stage sampling mechanism

particularly good estimates. This data example shows the challenging nature of the problem at hand, and its sensitivity both to the shape of the underlying distribution and the need to balance accuracy with computational efficiency.

VI. CONCLUDING REMARKS

In this paper, the problem of estimating the flow length and size distributions from sampled data is considered. A maximum likelihood non-parametric estimator for these quantities is proposed based on Bernoulli sampling and its properties briefly discussed. Subsequently, the focus shifts to mixture distributions that are prevalent in real network traffic traces. A two-stage maximum likelihood estimator is proposed based on Bernoulli sampling and an alternative estimator based on a two stage sampling mechanism. Both estimators employ the EM algorithm. Experimental evidence suggests that the quality of the estimates is very good and obviously improves for larger sampling rates. However, it should be noted that for very large data sets in terms of number of flows (M), convergence of the proposed algorithms is rather slow [9]. Therefore, a topic of current research is to develop faster alternatives. Moreover, the statistical properties of these estimators for mixture distributions from sampled data are not fully understood and require further study, as well as the choice of the sampling rates in the two-stage sampling mechanism.

Acknowledgments: This work has been partially supported by NSF grants DMS-0204247, DMS-0505535 and CCR-0325571.

REFERENCES

[1] Cisco NetFlow
[2] Cochran, W.G. (1977), *Sampling Techniques*, 3rd edition, John Wiley, New York, NY
[3] Duffield, N.G., Lund, C. and Thorup, M. (2005), Estimating flow distributions from sampled flow statistics, *IEEE/ACM Transactions on Networking*, 13, 325-336
[4] Duffield, N. (2004), Sampling for passive Internet measurement: A Review, *Statistical Science*, 19, 472-498
[5] Claffy, K.C., Polyzos, G.C. and Braun, H.W. (1993), Application of sampling methodologies to network traffic characterization, *Proceedings ACM SIGCOMM*, 13-17
[6] Hohn, M. and Veitch, D. (2003), Inverting sampled traffic", *Proceedings of Internet Measurement Conference*
[7] Internet Protocol Flow Information Export, IETF Working Group
[8] Kamiyama, T. (2005), Identifying high-rate flows with less memory, *Proceedings IEEE Infocom*, 2781-2785
[9] Keener, R.W, *Statistical Theory: A Medley of Core Topics*. in preparation
[10] Louis, T.A. (1982), Finding the observed information matrix when using the EM algorithm', *Journal of the Royal Statistical Society, Series B*, 44, 226-233
[11] Mori, T., Uchida, M. and Kawahara, R. (2004), Identifying elephant flows through periodically sampled packets, *Proceedings ACM SIGCOMM*, 115-120
[12] Zacks, S. (1991), *Theory of Statistical Inference*, John Wiley, New York, NY & Sons
[13] Faraway, J. (2005), *Extending the Linear Model with R*, CRC Press
[14] Yang, L and Michailidis, G. (2006), Estimation of Flow Lengths from Sampled Traffic, to appear in *Proceedings of Globecom*, San Francisco, CA
[15] Scott,David, *Multivariate density estimation : theory, practice, and visualization*, Wiley
[16] Diebolt, J. and Robert, C. (1994), Estimation of Finite Mixture Distributions through Bayesian Sampling, *Journal of the Royal Statistical Society, B*, 56, 363-375
[17] De Leeuw, J. (1994), Block Relaxation Algorithms in Statistics, in *Information Systems and Data Analysis*, Bock et al. (eds)., Springer, Heidelberg
[18] Yang, L and Michailidis, G. (2006), Estimating Network Traffic Characteristics based on Sampled Data, Technical Report, Department of Statistics, The University of Michigan