

# Estimation of Flow Lengths from Sampled Traffic

Lili Yang

George Michailidis

Department of Statistics  
University of Michigan  
Ann Arbor, MI 48109

Department of Statistics  
University of Michigan  
Ann Arbor, MI 48109

**Abstract**— In this paper, we consider the problem of nonparametric estimation of the original length of traffic flows, based on sampled flow data. The proposed approach is a two-step one. In the first stage, the flow length distribution is estimated by an expectation-maximization (EM) algorithm that in addition provides an estimate for the number of active flows in the link. In the second stage, two estimators are derived for the original flow length, using information from the posterior distribution previously obtained. The proposed approach is illustrated on a number of synthetic and real data sets.

## I. INTRODUCTION

Network traffic measurement and monitoring constitute an important component for several network management and engineering tasks, such as quality of service provisioning, usage based accounting, traffic profiling and control, just to name a few [3], [4]. However, collecting the necessary information on every packet that goes through a high speed link becomes prohibitive in terms of processing capacity, cache memory and required bandwidth. Packet sampling has emerged as an attractive option that offers a scalable alternative to address this problem, as can be seen by recent recommendations in the Internet Engineering Task Force working groups [6] and its implementation in high-speed routers [1]. An overview of applications where sampling proves useful for passive Internet measurements is given in [3].

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 distribution has attracted some attention in the literature (see for example [2], [10], [7], [5] and references therein). In this paper, we address this problem in a rigorous statistical manner. Specifically, we adopt a non-parametric model and obtain the maximum likelihood estimator for the number of active flows and the flow length distribution, based on sampled packet data. Further, the Fisher information matrix of the estimator is derived that allows one to construct confidence intervals. Further, the length of an original flow is estimated from that of a sampled flow through the posterior mean and the maximum a posteriori estimate. The proposed method is validated on a number of simulated and real network traces. The paper is organized as follows: in section II, the non-parametric model is formulated and the maximum likelihood

estimator based on the Expectation Maximization Algorithm is derived (section II(A)), together with posterior estimates of flow lengths (section II(C)). Further, asymptotic properties of the estimator -consistency, Fisher information matrix- are obtained in section II(D). In section III, the performance of the non-parametric estimator is assessed on several synthetic and real network traffic data sets. Some concluding remarks are drawn in section IV.

## II. PROBLEM FORMULATION

Suppose that there are  $M$  active flows on a link, comprised of  $N_m, m = 1, \dots, M$  packets each. The number of packets in each flow is referred to as the flow length. Further, assume that packets are sampled according to a Bernoulli sampling scheme; i.e. each packet is selected with probability  $p$ , independent of any other characteristic. Each sampled packet can be assigned to a particular flow, by observing its flow key obtained from information available in the packet header [3]. Therefore, the available data are sampled flow lengths  $n_1, n_2, \dots, n_r$ , where  $r$  is the number of sampled flows. For flows whose packets have not been sampled, there is no available information. 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. The objective of this study is threefold: (i) estimate non-parametrically the flow length distribution of the link, (ii) estimate the original length of a sampled flow and (iii) estimate the number of flows in the link. In addition, we would like to provide uncertainty assessments in the form of confidence intervals for these quantities of interest. The model employed is described next.

Let  $\phi_i$  denote the probability that a flow contains  $i$  packets and let  $\phi = \{\phi_i\}$ . Further, let  $g_j, j = 0, 1, \dots, J$  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  $i$  and 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:

$$L(\phi, M) = \binom{M}{g_0, g_1, \dots, g_J} \prod_{j \geq 0} \left( \sum_{i \geq j} \phi_i c_{ij} \right)^{g_j} \quad (1)$$

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  $i \in S_I = \{i(0), i(1), \dots, i(J)\}$ , 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 are restricted to integer values closest to  $j/p$ .

*Remark:* A fairly similar formulation of the problem appeared in [2]. However, two different likelihood functions were considered, one for  $\phi$  (the flow length distribution) and a separate one for the total number of flows  $n$ . Further, some algebra shows that the proposed estimate of  $M$  in [2] given by the number of observed flows divided by the sampling probability (i.e.  $r/p$ ) is not the maximum likelihood estimate. Instead, we propose an integrated framework for this joint estimation problem, and treat  $M$  as a nuisance parameter [8].

#### A. Maximum Likelihood Estimation

Maximizing the likelihood function given in (1) is a hard task, due to the presence of the constraints. However, the Expectation-Maximization algorithm [8] achieves the goal. The main idea is that if the frequencies of all the original flows whose length is  $i$  packets were observed, then the estimation of  $\phi$  and  $M$  would be rather trivial; hence, the EM algorithm imputes such values (E-step) and then maximizes the likelihood over the parameters of interest (M-step). Its main steps are described next:

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

$$\phi_{i(j)}^{(0)} = \frac{g_j}{\sum_{k=1}^J g_k + \hat{g}_0^{(0)}}, \quad (2)$$

where  $\hat{g}_0^{(0)}$  is estimated by the odds ratio

$$\hat{g}_0^{(0)} = \frac{\sum_{j=1}^J \frac{g_j}{\sum_{k=1}^J g_k} c_{i0}}{1 - \sum_{j=1}^J \frac{g_j}{\sum_{k=1}^J g_k} c_{i0}} r. \quad (3)$$

(2) *E-step:* Given the complete set of data  $(f_{ij}, g_j, n)$ ,  $f_{ij}$  follows a multinomial distribution with parameters  $M = \sum_{i,j} f_{ij}$  and  $p_{ij}$ . The data complete likelihood is then given by  $L_c(\phi, M) = \prod_{i \geq j \geq 0} (\phi_i c_{ij})^{f_{ij}}$ , with the corresponding log-likelihood being

$$l_c(\phi, M) = \sum_{i \geq j \geq 0} f_{ij} \log(\phi_i c_{ij}). \quad (4)$$

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

$$Q(\phi, \phi^{(k)}) = \sum_{i \geq j \geq 0} E_{\phi^{(k)}}(f_{ij} | g_j, j = 1, 2, \dots, J) \log(\phi_i c_{ij}). \quad (5)$$

For  $j \geq 1$ , notice that  $f_{ij} | g_j$  follows a Multinomial( $g_j, p_{i|j}$ ) distribution, where

$$p_{i|j} = \frac{\phi_i^{(k)} c_{ij}}{\sum_{l \in S_I, l \geq j} \phi_l^{(k)} c_{lj}} \quad (6)$$

is the probability that a sampled flow of length  $j$  contains actually  $i$  packets in total. Therefore,

$$E_{\phi^{(k)}}(f_{ij} | g_j, j = 1, 2, \dots, J) = g_j p_{i|j}. \quad (7)$$

However, when  $j = 0$  a specific form for  $E_{\phi^{(k)}}(f_{i0} | g_j, j = 1, 2, \dots, J)$  can not be obtained, since the quantity  $g_0$  is not observed. Further, there is no specific distributional assumption connecting  $g_0$  and the  $g_j$ 's,  $j \geq 1$ . In order to overcome this difficulty, the *nuisance parameter*  $\hat{g}_0^{(k)}$  is introduced and updated by

$$\hat{g}_0^{(k)} = \frac{\sum_{i \in S_I} \phi_i^{(k)} c_{i0}}{1 - \sum_{i \in S_I} \phi_i^{(k)} c_{i0}} r, \quad (8)$$

the updated odds ratio defined by the probability of observing a sampled flow of length zero over the probability of observing non-zero flow lengths. Accordingly, we obtain  $\hat{M}^{(k)} = r + \hat{g}_0^{(k)}$ . We then get that

$$E_{\phi^{(k)}}(f_{i0} | g_j, j = 1, 2, \dots, J) = \hat{g}_0^{(k)} p_{i|0}. \quad (9)$$

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

$$\sum_{i \in S_I} \phi_i = 1, \text{ and } \phi_{i(j)} \geq 0 \text{ for } i \in S_I. \quad (10)$$

The method of Lagrange multipliers gives

$$\phi_i^{(k+1)} = \frac{\sum_{i \geq j \geq 1} g_j p_{i|j} + \hat{g}_0^{(k)} p_{i|0}}{\sum_{i \in S_I} (\sum_{i \geq j \geq 1} g_j p_{i|j} + \hat{g}_0^{(k)} p_{i|0})}. \quad (11)$$

*Iterate* steps (2) and (3) until the convergence criterion is satisfied; i.e.  $\|\phi^{(k+1)} - \phi^{(k)}\| < \delta$ .

#### B. Estimation of Original Flow Lengths

The next objective is to come up with an estimate of the *actual* length of sampled flows. Given the estimated flow length distribution  $\phi$ , we can, through a straightforward application of Bayes formula, obtain the posterior probability distribution of a flow being of length  $i$  given that  $k$  of its packets have been sampled. Specifically, let  $f(i|k)$ ,  $k = 1, 2, \dots, J$  denote this probability distribution,  $f(i, k)$  the joint distribution of original and sampled flow lengths and  $f(k)$  the distribution of observing a sampled flow of length  $k$ . We then have that

$$f(n|k) = \frac{f(n, k)}{f(k)} = \frac{c_{nk} \phi_n}{\sum_{n \in S_I} c_{nk} \phi_n}. \quad (12)$$

For any given sampled flow of length  $k$ , we provide next two estimators of the original flow length  $N(k)$ .

(1) *Average*.  $\hat{N}(k) = \mathbb{E}(N(k)) = \sum_{n \in S_I} n f(n|k)$ . This estimator (the posterior mean) is the weighted average of all possible flow lengths.

(2) *Maximum a posteriori estimator*.  $\tilde{N}(k) = \operatorname{argmax}_{n \in S_I} f(n|k)$ . The estimated length corresponds to the value that maximizes the probability of observing a sampled flow of length  $k$ .

The attraction of the second estimator ( $\tilde{N}(k)$ ) is that it minimizes the total risk, where the risk function is given by

$$R(\hat{N}(k), N) = \mathbb{E}[L(N, \hat{N}(k)) | \text{True flow length is } N],$$

with the loss function  $L(A, B) = 1$  if  $A \neq B$ , and 0 otherwise.

On the other hand, the first estimator ( $\hat{N}(k)$ ) gives smoother estimates as seen in the experimental results section III. Due to space considerations, our performance evaluation focuses on the first estimator.

### C. Smoothing the Flow Length Distribution

In the previous section, a non-parametric density estimator was obtained for the flow length distribution, calculated at integer values  $i(j)$  closest to  $\frac{j}{p}$ . However in practice, some of the  $i(j)$ 's may be beyond the range of the original flow lengths. This comes from the fact that  $\operatorname{Var}(j) = p(1-p)\mathbb{E}(i) + p^2\operatorname{Var}(i)$ . We then have that  $\operatorname{Var}[i(j)] = \frac{1-p}{p}\mathbb{E}(i) + \operatorname{Var}(i)$ . Hence, when  $p$  is small, the variance of the selected flow lengths  $i$  is rather large. In order to obtain an estimator that exhibits reduced variability, we implement a smoothing technique [8].

Notice that because flow lengths are integer valued, only smoothing kernels defined on the integers are appropriate choices. An example of such a kernel used in the numerical work in Section III is shown next:

$$K(x, y) = \begin{cases} \frac{1}{2} & \text{if } x = y \\ \frac{c}{|x-y|} & \text{if } x \neq y \end{cases} \quad (13)$$

where  $c$  is a constant that ensures that  $\sum_i K(x_i - x) = 1$  for all values of  $i$  observed. In principle, we can extend the values of  $i$  to all the integers in the domain of the flow length distribution. It is shown in section III that smoothing the flow length distribution yields better results for small sampling rates (low  $p$ ), small sample sizes and heavy tailed distributions for the original flow lengths.

### D. Statistical Inference

We briefly discuss next some of the asymptotic properties of the derived maximum likelihood estimator. Some algebra (omitted for space considerations) shows that the usual regularity conditions [8] are satisfied by the posited multinomial model and hence it can be established that the maximum likelihood estimator converges to the true parameters; i.e.  $(\hat{\phi}, \hat{M}) \xrightarrow{p} (\phi, M)$ , where  $\xrightarrow{p}$  denotes convergence in probability.

Given the above consistency result, we then need to calculate the Fisher Information Matrix for the parameters

of interest  $(\phi, M)$ , in order to obtain their asymptotic distribution, and subsequently calculate the variance of  $\hat{N}(k)$ . We implemented Louis' [9] procedure for obtaining the observed information matrix when using the EM algorithm. Gradient vectors of the complete log-likelihood take the form of  $S = (\frac{f_{i(0)}}{\hat{\phi}_{i(0)}}, \dots, \frac{f_{i(J)}}{\hat{\phi}_{i(J)}})'$ . Then, the observed Information Matrix is given by  $I_{obs} = \sum_{i=1}^r S(\hat{f}_i, \hat{\phi}, \hat{M})S(\hat{f}_i, \hat{\phi}, \hat{M})'$ , 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  $k$ th one, being 0.

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

$$\begin{bmatrix} \frac{\sum_{j=1}^J p_{i(0)|j}^2 g_j}{\hat{\phi}_{i(0)}^2} & & & & \\ \frac{\sum_{j=1}^J p_{i(0)|j} p_{i(1)|j} g_j}{\hat{\phi}_{i(0)} \hat{\phi}_{i(1)}} & \frac{\sum_{j=1}^J p_{i(1)|j}^2 g_j}{\hat{\phi}_{i(1)}^2} & & & \\ \dots & \dots & \dots & \dots & \\ \frac{\sum_{j=1}^J p_{i(0)|j} p_{i(J)|j} g_j}{\hat{\phi}_{i(0)} \hat{\phi}_{i(J)}} & \dots & \dots & \frac{\sum_{j=1}^J p_{i(J)|j}^2 g_j}{\hat{\phi}_{i(J)}^2} & \end{bmatrix} \quad (14)$$

Standard results yield that the asymptotic distribution of the estimated flow length distribution is multivariate normal; i.e.  $\sqrt{r}(\hat{\phi} - \phi) \Rightarrow N(0, I_{obs}^{-1})$

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

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

and is continuous in  $\phi$ . By an application of the Delta method [8], we then get that  $\operatorname{Var}(G_{\hat{\phi}}) = \nabla G_{\hat{\phi}}' I_{obs}^{-1} \nabla G_{\hat{\phi}}$ . The continuity of the above defined functional in  $\phi$  gives that  $\sqrt{r}(\hat{N}(k) - N(k)) \Rightarrow N(0, \operatorname{Var}(G_{\hat{\phi}}))$ . 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.

## III. EXPERIMENTAL EVALUATION

In this section, we provide empirical evidence of the performance of the derived estimators for a variety of simulated and real network traffic traces. The simulated flow length data were obtained from the following distributions: (i) uniform with domain  $[0, 10000]$ , (ii) Poisson with mean 5,000 and (iii) Pareto with shape parameter 10/9 and scale parameter 500 (Pareto-I) and also with scale parameter 50 (Pareto-II). The parameters for the first three distributions were set so as to match their expected values, while Pareto-II has a heavier tail. Another set of simulated data were obtained from the ns2 network simulation package [12]. Two networking scenarios were considered. In the first scenario lasting 2 minutes, 100 constant bit rate sources generated traffic, whose duration followed a Pareto distribution with

shape parameter 1.5 and scale parameter 100/3. Packet sizes were identical within the same flow (source), but different across flows, following a normal distribution with mean 800 and standard deviation 100. In the second scenario 100 ftp transmissions were generated on a link, whose duration follows the previously defined Pareto distribution. Finally, two real data sets were considered. The first one contains 34,514 network flows collected over a 2 hour period at the router of a small local area network. The average flow length consists of 29 packets, but the variance is  $4.5 \times 10^5$ . The second data set contains 256,835 flows passing through the gateway link of the UNC campus network. The average flow length consists of 39 packets with a variance of  $3.22 \times 10^5$ . The distribution of the true flow lengths (in log-scale) for these data sets are shown in Figures 1 and 2, respectively. It can easily be seen that both data sets have heavy right-tailed flow length distributions.

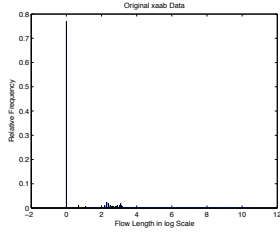


Fig. 1. Empirical distribution (in log-scale) of the LAN flow data

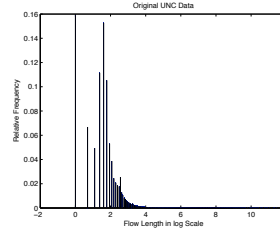


Fig. 2. Empirical distribution (in log-scale) of the UNC flow data

We first look at the estimates ( $\hat{M}$ ) of the true number of flows, plotted against the true ones ( $M$ ), where the flow length distribution is Poisson with mean 5,000 for varying values of  $M$  with a sampling rate of  $p = .01$  (see Figure 7). It can be seen that the estimated number of flows is extremely accurate, a result also obtained for many other distributions (not shown here).

We show next quantile-quantile plots of the empirical distribution of the true flow lengths, vs the non-parametrically estimated distribution for 1000 flows, whose lengths follow a mean 5,000 uniform distribution with  $p = .01$  and  $.05$  sampling rates (see Figures 3 and 4). A high degree of agreement for both sampling rates, between the true and estimated distributions, can be observed.

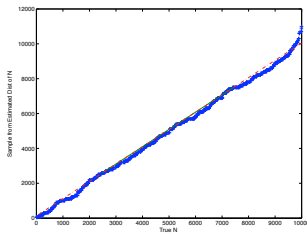


Fig. 3. Quantile-quantile plot of the true vs the estimated flow length distribution for 1000 uniformly distributed flows, sampling rate 0.01

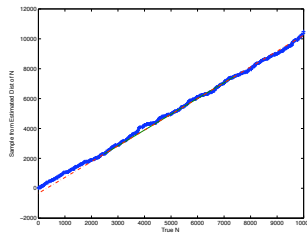


Fig. 4. Quantile-quantile plot of the true vs the estimated flow length distribution for 1000 uniformly distributed flows, sampling rate 0.05

In order to obtain a better perspective about the effect of the sampling rate, the estimated flow length distribution for 1,000 Poisson flows of mean length 5,000 are given in Figures 5 and 6, corresponding to sampling rates of  $p = .01$  and  $.05$ . It can be seen that with a higher sampling rate the number of possible sampled flow lengths ( $J$  in our notation) increases significantly, thus allowing a more accurate estimate of the underlying distribution. Finally, the confidence intervals obtained for the original flow length estimates  $\hat{N}(k)$ , when a sampled flow of length  $k$  was observed for 100 Poisson flows of mean length 5,000, are shown in Figure 9. It can be seen that the confidence intervals are fairly tight except at the two ends, which reflects the higher uncertainty of our estimates for the corresponding  $\phi_i$  parameters.

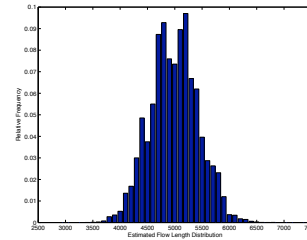


Fig. 5. Estimated flow length distribution of 1,000 Poisson flows with .01 sampling rate

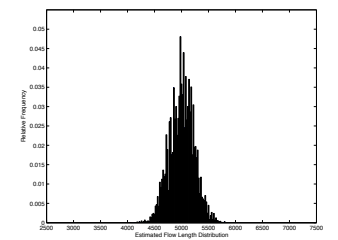


Fig. 6. Estimated flow length distribution of 1,000 Poisson flows with .05 sampling rate

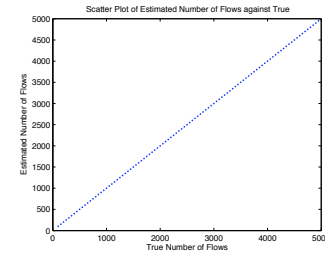


Fig. 7. Scatter plot of true vs estimated number of active flows in the link

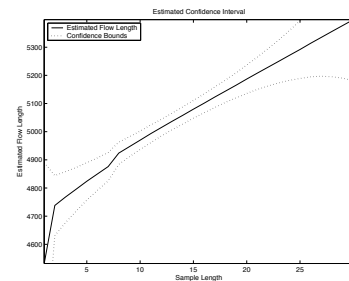


Fig. 8. Confidence intervals for 100 Poisson Flows with sampling rate 0.01

In the following tables, we give  $\chi^2$ -distances between the true and estimated flow length distributions, together with mean squared errors (MSE) of the  $\hat{N}(k)$  estimates for a variety of model and ns simulated data sets. As expected, it can be seen that as the sampling rate increases both

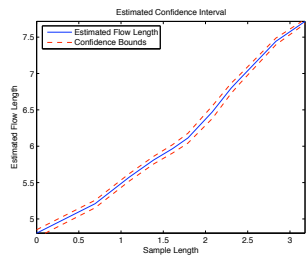


Fig. 9. Confidence intervals for 100 Pareto Flows (in log-scale) with sampling rate 0.01

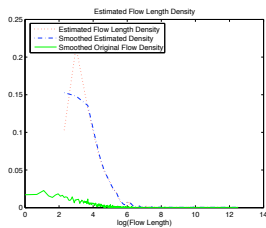


Fig. 10. Estimated flow length distribution of 2500 Pareto flows with .05 sampling rate

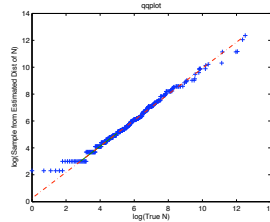


Fig. 11. Quantile-quantile plot of the true vs the estimated flow length distribution for 2500 Pareto distributed flows, sampling rate 0.05

performance measures decrease. Further, the MSE criterion shows rather comparable performance for the relatively large amount of data with 2500 flows.

Figures 12 and 13 show the estimated flow length distribution for the LAN and UNC flow data sets, respectively. Comparing the estimated distributions to the original ones given in Figures 1 and 2, it can be seen that that even with low sampling rate -1% for the LAN flow data, and 0.1% for the UNC flow data- their match is fairly good. Although there are some minor gaps, they both capture most patterns of the distributions, especially considering the extremely heavy-tailed nature of the original data.

TABLE I  
STATISTICS FROM SIMULATED DATA

$\chi^2$ distance	M	rate p			
		Uniform	Poisson	Pareto-I	Pareto-II
100	0.01	225.66	437.29	648.92	1.24E+03
	0.05	121.83	285.07	106.91	88.12
1000	0.01	4.74e+04	1.24e+05	1.45e+04	2.92E+06
	0.05	4.99e+03	3.82e+04	3.22e+03	4.45E+05
2500	0.01	3.29e+05	2.35e+05	1.29e+07	6.23E+07
	0.05	3.50e+04	2.27e+05	6.84e+05	6.61E+06
MSE					
		Uniform	Poisson	Pareto-I	Pareto-II
100	0.01	3.94e+005	2.97e+004	4.15e+005	3.73E+04
	0.05	7.95e+004	1.05e+004	1.97e+005	3.48E+03
1000	0.01	4.06e+05	3.72e+04	1.14e+05	8.30E+04
	0.05	8.85e+04	1.31e+04	7.14e+04	1.88E+04
2500	0.01	4.52e+05	1.79e+04	5.12e+05	2.45E+05
	0.05	9.04e+04	1.65e+04	1.02e+05	1.51E+04

TABLE II  
STATISTICS FROM NS2 SIMULATOR

rate	$\chi$ -distance		MSE	
	UDP/CBR	TCP	UDP/CBR	TCP
0.01	122.5285	765.3391	5.2948e+003	7.4811e+05
0.05	17.0740	242.1123	1.5526e+003	7.8458e+04

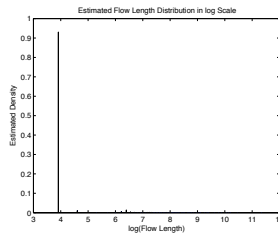


Fig. 12. Estimated LAN flow length distribution

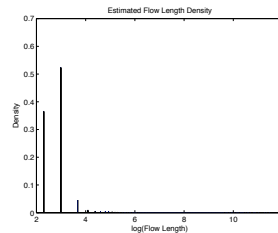


Fig. 13. Estimated UNC flow length distribution

#### IV. CONCLUSIONS

In this paper, a maximum likelihood non-parametric estimator for the flow length distribution and the number of active flows in a link was developed based on the expectation-maximization algorithm. In addition, the information matrix of the estimates was derived, that allows one to calculate confidence intervals for the parameters of interest. 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 algorithm is rather slow [8]. Therefore, a topic of current research is to develop faster alternatives that can be used in an online manner.

#### REFERENCES

- [1] Cisco NetFlow.
- [2] Duffield, N.G., Lund, C. and Thorup, M. (2005), Estimating flow distributions from sampled flow statistics, *IEEE/ACM Transactions on Networking*, 13, 325-336
- [3] Duffield, N. (2004), Sampling for passive Internet measurement: A Review, *Statistical Science*, 19, 472-498
- [4] Claffy, K.C., Polyzos, G.C. and Braun, H.W. (1993), Application of sampling methodologies to network traffic characterization, *Proceedings ACM SIGCOMM*, 13-17
- [5] Hohn, M. and Veitch, D., Inverting sampled traffic (2003), *Proceedings of Internet Measurement Conference*, Miami Beach, FL
- [6] Internet Protocol Flow Information Export, IETF Working Group
- [7] Kamiyama, T. (2005), Identifying high-rate flows with less memory, *Proceedings IEEE Infocom*, 2781-2785
- [8] Keener, R.W. *Statistical Theory: A Medley of Core Topics*, Springer, New York, NY, to appear
- [9] 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
- [10] Mori, T., Uchida, M. and Kawahara, R. (2004), Identifying elephant flows through periodically sampled packets, *Proceedings ACM SIGCOMM*, 115-120
- [11] Duffield, N.G., Lund, C. and Thorup, M. (2002) Properties and prediction of flow statistics from sampled packet streams, *Proc. ACM SIGCOMM Internet Measurement Workshop 159-171*.
- [12] The ns-simulator, Information Sciences Institute, available at [www.isi.edu/nsnam/ns](http://www.isi.edu/nsnam/ns)