# NONPARAMETRIC INTERNET TOMOGRAPHY

*Yolanda Tsang, Mark Coates, Robert Nowak*

Department of Electrical and Computer Engineering, Rice University
6100 South Main Street, Houston, TX 77005-1892
*Email: {ytsang, mcoates, nowak}@ece.rice.edu, Web: www.dsp.rice.edu*

## ABSTRACT

The substantial overhead of performing global Internet monitoring motivates techniques for inferring spatially localized information about performance using only host-based, end-to-end measurements. In this paper, we present a novel methodology for inferring queuing delay distributions across internal links in the network based solely on unicast, end-to-end measurements. A key feature of our new approach is that it is nonparametric, meaning that no *a priori* limit is placed on the number of unknown parameters used to model the delay distributions. The nonparametric approach is required in order to accurately estimate the wide variety of internal delay distributions. The methodology is formulated according to a recently proposed nonparametric, wavelet-based density estimation method in combination with an expectation-maximization optimization algorithm that employs a novel fast Fourier transform implementation. We perform network level ns simulations to verify the accuracy of the estimation procedure.

## 1. INTRODUCTION

Spatially localized information about network performance plays an important role in isolation of network congestion and detection of performance degradation. Routing algorithms, servicing strategies, security programs and performance verification can benefit from monitoring techniques that report such information. Monitoring can be performed internally, but it is impractical to directly measure traffic characteristics at all internal devices for a number of reasons [1]. This has prompted several groups to investigate methods for inferring internal network behavior based on “external” end-to-end network measurements [1, 2, 3, 4, 5, 6, 7, 8]. This problem is often referred to as *network tomography*.

Queuing delays are one of the most critical performance characteristics. Optimizing communication network routing and service strategies requires knowledge of the queueing delay at different points in the network. Measuring end-to-end (source to receiver) delays using timestamps [6, 9, 10] is relatively easy and inexpensive in comparison to internal measurement, although there are of course measurement issues that must be addressed. It is natural to consider the following problem: from end-to-end measurements can we resolve the queueing delay experienced at internal points in the network? More precisely, the goal of the network tomography problem considered in this paper is to estimate the probability distribution of the queueing delay on each link based on end-to-end packet pair measurements.

In this paper, we describe a *nonparametric* framework for the inference of internal delay distributions based on unicast end-to-end measurement. By nonparametric we mean that no *a priori* limit is placed on the number of parameters or degrees of freedom used to describe the observed delay measurements. Most work to date in network tomography is based on *parametric* models. A nonparametric approach based on cumulant generating functions was proposed in [11], but this approach, unlike ours, requires internal measurements. Parametric models assume that the measured traffic data depends on a finite number of parameters. For example, earlier work in delay distribution estimation, including our own, has been based on discretized (or quantized) delay measurements, with internal delay distributions modeled as discrete probability mass functions (pmfs) [1, 2, 3]. In this context, the parameters are simply the probabilities associated with each pmf. It has been our experience, however, that no sufficiently simple parametric model is capable of portraying the wide variety of internal delay distributions observed in practice, thus motivating the consideration of nonparametric or continuous models. We compare parametric, pmf-based delay distribution estimators with the new nonparametric approach in network simulation experiments, and find that the nonparametric method offers significantly superior performance.

The remainder of the paper is structured in the following manner. In Section 2 we describe the measurement framework, modeling assumptions and implementation requirements. In Section 3 we describe the inference methodology, detailing the nonparameteric estimation procedure and an expectation-maximization (EM) algorithm used to compute the estimators. In Section 4 we describe the results of ns experiments designed to explore the performance of the methodology. In Section 5, we make some concluding remarks and indicate avenues of future research.

## 2. MEASUREMENT FRAMEWORK

Throughout this paper, we concentrate on networks comprised of a single sender transmitting measurement probes to multiple receivers. There is no difficulty extending the approach to measurements made at multiple sources, although care must be taken that measurements are sufficiently separated for independence assumptions to hold. We assume that the topology is fixed throughout the measurement period, but straightforward extensions can account for changes in topology over coarse time scales.

For the networks we consider, standard network routing protocols produce a tree-structured topology, with the *source* at the root and the *receivers* at the leaves. A network with six receivers is de-

This work is supported by the National Science Foundation, grant nos. ANI–0099148 and MIP–9701692, the Office of Naval Research, grant no. N00014-00-1-0390, and the Army Research Office, grant no. DAAD19-99-1-0290.

picted in Fig. 1. The nodes between the source and receivers represent internal *routers*. Connections between the source, routers, and receivers are called *links*. Each link between routers may be a direct connection, or there may be "hidden" routers (where no branching occurs) along the link that are not explicit in our representation. We adopt the notation that link $i$ connects node $i$ to its parent node $\omega(i)$.

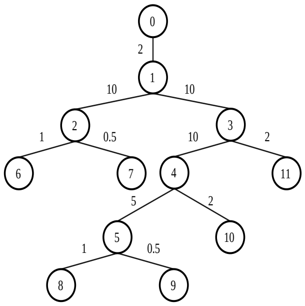

**Fig. 1**. Tree-structured network topology used for ns simulation experiments. Source (node 0) transmits to 6 receivers (nodes 6-11). Link speeds in Mb/s are shown next to the links. Link $i$ connects node $i$ to its parent node $\omega(i)$, e.g. link 9 connects nodes 5 and 9.

The basic measurement and inference idea is quite straightforward. Suppose two closely time-spaced (back-to-back) packets are sent from the source to two different receivers. The paths to these receivers traverse a common set of links, but at some point the two paths diverge (as the tree branches). The two packets should experience approximately the same delay on each shared link in their path. This facilitates the resolution of the delays on each link. We collect measurements of the end-to-end delays from source to receivers, and we index the packet pair measurements by $k = 1, \ldots, N$. For the $k$-th packet pair measurement, let $y_1(k)$ and $y_2(k)$ denote the two end-to-end delays measured. The ordering 1 and 2 is *arbitrary*; the indices are randomly selected with no dependence the order in which the packets were sent from the source. Since we are interested in inferring queuing delay, our first step is to extract the minimum delay (propagation + transmission) on each measurement path. This is estimated as the smallest delay measurement we acquire on the path during the measurement period.

To illustrate the basic ideas behind our inference methodology in its simplest form, suppose that we send many packet pairs to receivers 6 and 7 in Fig. 1 and measure the delays experienced by each packet. Each measurement consists of a pair of delays, one being the delay to receiver 6 and the other the delay to receiver 7. From these measurements, collect events where '0' delay (a delay in bin zero) is measured at receiver 6. Now, assuming that the delay is the same for both packets on the common links (1 and 2 in this case), any "additional" delay observed to the receiver at 7 can be attributed to link 7 alone. We can then build a histogram estimate of the delay pmf for link 7. We describe the complete, large-scale, EM algorithm inference procedure in Section 3.

There are several assumptions in the framework that are worthy of discussion. Firstly, we assume spatial independence of delay. Delay on neighboring links is generally correlated to a greater or lesser extent depending on the amount of shared traffic. In the presence of weak correlation, our framework is able to derive good estimates of the delay distributions. As the correlation grows stronger, we see a gradual increase of bias in the estimates. We also assume temporal independence (successive probes across the same link experience independent delays). Temporal dependence was observed in [1] and in our experiments. As in [1], the maximum likelihood estimator we employ remains consistent in the presence of temporal dependence, but the convergence rate slows. Finally, our framework hinges on an assumption that packets in a pair experience a common delay on shared links. In actual Internet experiments we have found that any discrepancies between delays are very slight and thus will not significantly effect the performance of our methodology [12].

## 3. DELAY DISTRIBUTION INFERENCE

We commence the description of our inference framework by formalizing our measurement and modeling notation. We assume that these measurements are independent and identically distributed according to a continuous delay density $p_i(t)$, where without loss of generality we assume that $t \in [0,1]$ (for convenience of exposition we take the maximum delay to be unity). Define a discrete pmf via $p_{i,j} = \int_{(j)/K}^{(j+1)/K} p_i(t)dt$, $j = 0, \ldots, K-1$, where $K$ is the smallest power of two greater than or equal to $N_i$. Let $p_i = \{p_{i,0} \ldots, p_{i,K-1}\}$ denote the probabilities of a delay of $0, 1, \ldots, K-1$ time units, respectively, on link $i$. We denote the packet pair measurements $\boldsymbol{y} \equiv \{y_1(k), y_2(k)\}_{k=1}^{N}$. In general, only a relatively small amount of data can be collected over the period when delay distributions can be assumed approximately stationary. A natural estimate would be the maximum likelihood estimates (MLEs) of $\boldsymbol{p} \equiv \{p_i\}$, the collection of all delay pmfs. Under the assumption of spatial independence, the likelihood of each delay measurement $\{y_1(k), y_2(k)\}$, denoted $l(\boldsymbol{y}|\boldsymbol{p})$, is parameterized by a convolution of the pmfs in the path from the source to receiver. We have developed an EM algorithm [2] to compute MLEs for the parametric instance of this problem (i.e., when $K \ll N$).

In this paper, our goal is nonparametric delay density estimation on each link. This is accomplished by choosing the number of bins to be equal to or larger than the number of measurements, and thus the problem is ill-posed and the MLE tends to over-fit to the probe data. This results highly variable estimates that do not accurately reflect the delay distribution of the traffic at large. One way to reduce this irregularity is to maximize a penalized likelihood. We replace the maximum (log) likelihood objective function $L(\boldsymbol{p}) = \log l(\boldsymbol{y}|\boldsymbol{p})$ with an objective function of the form:

$$L(\boldsymbol{p}) - pen(\boldsymbol{p}), \tag{1}$$

where

$$\text{pen}(\{p_{i,j}\}) \equiv \sum_i \frac{1}{2}\log(N_i) \times \#_i, \tag{2}$$

where $N_i$ denotes the number probe packets passing through link $i$ and $\#_i$ denotes the number of non-zero Haar wavelet coefficients in the delay pmf of link $i$. This wavelet-based scheme is called a Multiscale Maximum Penalized Likelihood Estimator (MMPLE), and it was proposed in [13] for conventional (non-tomographic) probability density estimation. It penalizes the irregularity and complexity of the pmf — the larger the value of $\#_i$, the more "bumps" in the pmf. There are two important features of the MMPLE: (1) the global maximizer can be computed in $O(K)$ operations; (2) the MMPLE is nearly minimax optimal in the rate of

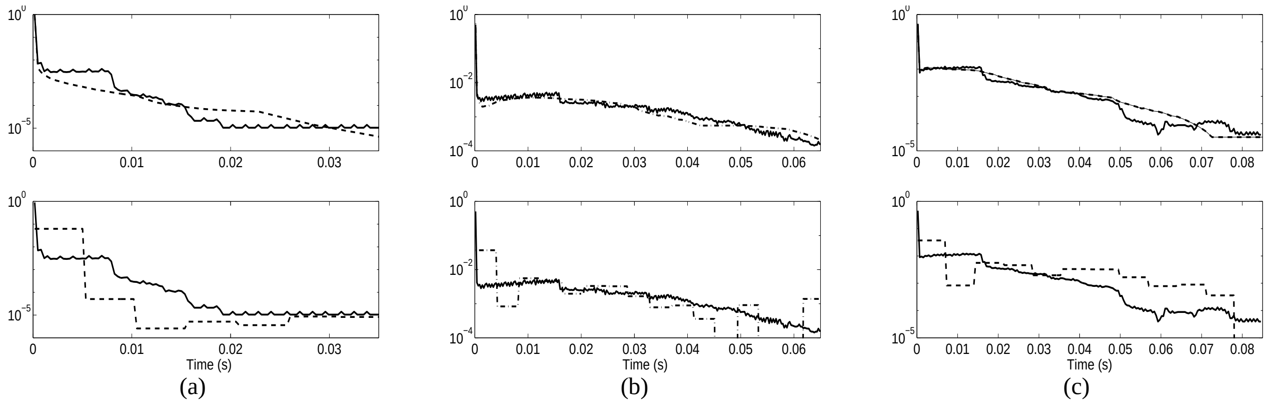

**Fig. 2**. Comparison between true pmfs (solid) and estimated pmfs (dashed). Top panel shows true pmf and MMPLE (calculated using 512 bins); bottom panel shows true pmf and MLE (calculated using 16 bins). 16 bins is determined as the bin size at which the MLE obtains the best fit. (a) Link 5. (b) Link 7. (c) Link 9.

convergence over a broad class of function spaces. In all results in this paper we employ a *translation-invariant* version of the MMPLE, in which multiple MMPLEs are computed with $K$ different shifted versions of the Haar wavelet basis and the resulting estimates are averaged. This produces a slight improvement over the basic MMPLE and can be efficiently computed in $O(K \log K)$ operations.

The penalized likelihood function in (1) cannot be maximized analytically due to the convolutional relationship between link delay pmfs and end-to-end measurements $\boldsymbol{y}$. The EM algorithm is an iterative procedure designed to maximize (1) that takes advantage of the $O(K)$ computational simplicity of the MMPLE technique.

The first step in developing an EM algorithm is to propose a suitable *complete data* quantity that simplifies the likelihood function. Let $z_i(k)$ denote the delay on link $i$ for the packets in the $k$-th pair. Let $\boldsymbol{z}_i = \{z_i(k)\}$ and $\boldsymbol{z} = \{\boldsymbol{z}_i\}$. The link delays $\boldsymbol{z}$ are not observed, and hence $\boldsymbol{z}$ is called the *unobserved data*. Define the *complete data* $\boldsymbol{x} \equiv \{\boldsymbol{y}, \boldsymbol{z}\}$. Note that the complete data likelihood may be factorized as follows $l(\boldsymbol{x}|\boldsymbol{p}) = f(\boldsymbol{y}|\boldsymbol{z})g(\boldsymbol{z}|\boldsymbol{p})$, where $f$ is the conditional pmf of $\boldsymbol{y}$ given $\boldsymbol{z}$ (which is a point mass function since $\boldsymbol{z}$ determines $\boldsymbol{y}$), and $g$ is the likelihood of $\boldsymbol{z}$. The factorization shows that $l(\boldsymbol{x}|\boldsymbol{p}) \propto g(\boldsymbol{z}|\boldsymbol{p})$, since $f(\boldsymbol{y}|\boldsymbol{z})$ does not depend on the parameters $\boldsymbol{p}$. Next note that the likelihood $g(\boldsymbol{z}|\boldsymbol{p}) = \prod_{i,j} p_{i,j}^{m_{i,j}}$, where $m_{i,j} \equiv \sum_{k=1}^{N} \mathbf{1}_{z_i(k)=j}$ is the number of packets (out of all the packet pair measurements) that experienced a delay of $j$ on link $i$; here $\mathbf{1}_A$ denotes the *indicator function* of the event $A$. Therefore, we have

$$L(\boldsymbol{p}) \propto \sum_{i,j} m_{i,j} \log p_{i,j}. \qquad (3)$$

If the $m_{i,j}$ were available, then we could directly apply the MMPLE described above.

The EM algorithm is an iterative method that uses the complete data likelihood function to maximize the penalized log-likelihood function. Specifically, the EM algorithm alternates between computing the conditional expectation of the complete data log likelihood given the observations $\boldsymbol{y}$ and maximizing the sum of this expectation and the imposed complexity penalty $(-\text{pen}(\boldsymbol{p}))$ with respect to $\boldsymbol{p}$. Notice that the complete data log likelihood (3) is linear in $\boldsymbol{m}$. Thus, in the E-Step we need only compute the expectation of $\boldsymbol{m} = \{m_{i,j}\}$. This can be efficiently performed using a message passing (or upward-downward) procedure [14]. Unfortunately, a straightforward implementation of the message passing procedure, as proposed in [2, 3], has a computational complexity which is $O(MK^3)$, where $M$ is the number of links in the tree and $K$ is the number of bins. This is impractical in our nonparametric setting since $K$ is not fixed, but rather increases in proportion to the number of measurements. We use a novel, fast Fourier transform based implementation which is $O(MK^2 \log K)$, a tremendous reduction in complexity when $K$ is large. For more details see [12]. In the M-Step, we apply the MMPLE algorithm of [13] to the conditional expectation $\boldsymbol{m}$. The EM algorithm typically converges in 10 to 15 iterations.

## 4. SIMULATION EXPERIMENTS

In order to verify the performance of our estimation methodology, we conducted ns simulation experiments using the network depicted in Fig. 1. The network was simulated for multiple two minute measurement periods. This time duration corresponds to 500 packet-pairs. Throughout the measurement period, queue lengths in the network were determined at a fine time scale by monitoring the arrivals of every packet at each queue. A "true" pmf for each link was formed by calculating delays from queue lengths and link capacities, quantizing and forming a histogram. When generating this true pmf, so much data is available that the quantization can be very fine (constructing an excellent estimate of the delay density) without affecting estimation stability.

In Fig. 2, we show the results of one experiment, comparing the true pmfs to the nonparametric MMPLE estimator and the MLE estimator of [2] (the estimation technique in [1] also returns the MLE, although it is devised in a different setting). We display results for the lower bandwidth links because for our experimental set-up, queueing delay was concentrated in these links. There is substantial mass in the tails of these pmfs and we can evaluate how well the pmf estimates generated by our proposed methodology estimates match the tails. In the higher bandwidth links, there is much less mass in the pmf tails. For these links, both the MLE and MMPLE estimates match the true pmf where probability mass is concentrated, but there is insufficient information to closely match the tails. We calculated the MLE for a variety of bin sizes, but show the bin size that achieved the best fit to the true pmf (in this case 16 bins). The nonparametric estimator was calculated from 512 bins.

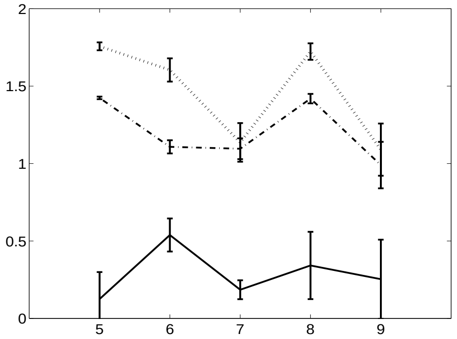

**Fig. 3**. $L_1$ error criterion averaged over 25 simulations (means and standard deviation). Solid line is MMPLE, dashed line is MLE (16 bins), dotted line MLE (64 bins).

In Fig. 3, we plot the magnitude of the $L_1$ error norm between the true pmf and the MMPLE for the links in the network, as averaged over 25 simulations. Also shown are the results for the MLE for medium (64 bins) and large (16 bins) bin sizes. The $L_1$ error norm is simply the sum of the absolute difference between the estimated pmf and the true pmf over the K bins (MLE estimates made with fewer bins are appropriately converted). As discussed in [15], the $L_1$ error criterion is a common measure of the performance of a density estimate. The advantage of such a measure as opposed to a mean-squared error criterion is that more attention is paid to the tails of the distributions. It also enjoys several theoretical advantages over other measures [15].

As is evident from the two figures, the MMPLE technique generates estimates which are smooth, close fits to the true pmfs. In order to introduce some degree of smoothness, MLE estimates must be calculated using a large bin size, resulting in an inability to capture the finer details of a pmf.

When the amount of probing that can be performed is limited, we believe that the most substantial source of error is the intrinsic variability in probe measurements. Another potential source of error is the discrepancy between the delays experienced by the two packets in each pair on their common path. We therefore examined the extent and effect of the delay discrepancy; with 512 bins, the overwhelming majority of the discrepancy was concentrated in 0-3 bins, with a maximum value of 16 bins. The effect of these discrepancies on the quality of the estimates is relatively minor when such a small amount of data is available for inference. If we use our queue monitoring to construct an artificial set of measurements (thereby providing ideal packet-pair probe measurements to our algorithms), the estimates we obtain are very similar to those we report here.

## 5. CONCLUSIONS

We have described a *nonparametric* framework for the inference of internal delay distributions based on unicast end-to-end measurement. The key features of the framework are its flexibility (the ability to capture fine details and smooth regions) and the introduction of a complexity penalization that allows smooth, accurate estimates to be generated even when the amount of data is very small. The basic MMPLE framework developed here could be extended to the multicast approach suggested in [4] and may also be applicable in time-varying contexts like those considered in [2, 3].

## 6. REFERENCES

[1] F. Lo Presti, N.G. Duffield, J. Horowitz, and D. Towsley, “Multicast-based inference of network-internal delay distributions,” Tech. Rep., University of Massachusetts, 1999.

[2] M. Coates and R. Nowak, “Network delay distribution inference from end-to-end unicast measurement,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, May 2001.

[3] M. Coates and R. Nowak, “Sequential Monte Carlo inference of internal delays in nonstationary communication networks,” to appear in *IEEE Trans. Signal Processing, Special Issue on Monte Carlo Methods for Statistical Signal Processing*, 2001.

[4] N.G. Duffield, F. Lo Presti, V. Paxson, and D. Towsley, “Inferring link loss using striped unicast probes,” in *Proceedings of IEEE INFOCOM 2001*, Anchorage, Alaska, April 2001.

[5] A. Bestavros K. Harfoush and J. Byers, “Robust identification of shared losses using end-to-end unicast probes,” in *Proc. IEEE Int. Conf. Network Protocols*, Osaka, Japan, Nov. 2000, *Errata* available as Boston University CS Technical Report 2001-001.

[6] K. Lai and M. Baker, “Measuring link bandwidths using a deterministic model of packet delay,” in *Proc. ACM SIGCOMM 2000*, Stockholm, Sweden, Aug. 2000.

[7] S. Ratnasamy and S. McCanne, “Inference of multicast routing trees and bottleneck bandwidths using end-to-end measurements,” in *Proceedings of IEEE INFOCOM 1999*, New York, NY, March 1999.

[8] D. Rubenstein, J. Kurose, and D. Towsley, “Detecting shared congestion of flows via end-to-end measurement,” in *Proc. ACM SIGMETRICS 2000*, Santa Clara, CA, June 2000.

[9] J. Kurose and K. Ross, *Computer Networking*, Addison Wesley, 2001.

[10] *Netdyn*, www.cs.umd.edu/projects/netcalliper/NetDyn.html.

[11] M.F. Shih and A.O. Hero, “Unicast inference of network link delay distributions from edge measurements,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, May 2001.

[12] M. Coates, R. Nowak, and Y. Tsang, “Nonparametric estimation of internal delay densities from unicast end-to-end measurement,” Tech. Rep. TR0106, Rice University, Aug. 2001.

[13] E. Kolaczyk and R. Nowak, “A multiresolution analysis for likelihoods: Theory and methods,” submitted to *A*nnals of Statistics, 2000.

[14] B. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, 1998.

[15] D.W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley, New York, 1992.