Unicast Network Tomography Using the EM Algorithm

Robert D. Nowak and Mark J. Coates

Department of Electrical and Computer Engineering

Rice University, MS-380, P.O. Box 1892

Houston, Texas 77251-1892 USA

Fax: (+1 713) 737-6196

Email: nowak,mcoates@rice.edu

Web: www.ece.rice.edu

Submitted to the IEEE TRANSACTIONS ON INFORMATION THEORY

November, 2001

Part of this work was presented at the

ITC Conference on IP Traffic, Modelling and Management, Monterey, CA, September 2000

Abstract

One of the predominant schools of thought in networking today is that monitoring and control of large scale networks is only practical at the edge. With intelligent and adaptive elements at the edge of the network, core devices can function as simple, robust routers. However, the effectiveness of edge-based control can be significantly enhanced by information about the internal network state. A fundamental component of the state is the loss rates of internal links in the network. The task of estimating these loss rates solely from host-to-host measurements is an example of "network tomography". This paper investigates a new network tomography procedure based on unicast packet-pair measurement, in contrast to previously proposed multicast probing strategies. We develop a likelihood formulation for unicast loss rate network tomography and devise an EM algorithm for computing the MLE. We conduct a theoretical analysis of the algorithm and report the results of simulations and network testbed experiments designed to explore performance under realistic conditions.

I. INTRODUCTION

A. Background and Motivation

Network tomography involves estimating network performance parameters such as packet loss rates or traffic intensities from traffic measurements at a limited set of measurement locations. Y. Vardi was one of the first to rigorously study this sort of problem and coined the term *network tomography* [31] due to the

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

similarity between network inference and medical tomography. Two forms of network tomography have been addressed in the recent literature: (i) link-level parameter estimation based on end-to-end, path-level traffic measurements [5, 12, 13, 18, 19, 21, 24, 25, 27, 33] and (ii) sender-receiver path-level traffic intensity estimation based on link-level traffic measurements [7, 8, 28, 30, 31]. This paper is concerned with the former problem, and more specifically it deals with a new technique for inferring packet loss rates on internal network links based solely on end-to-end measurements [5, 12, 18, 19]. Direct measurement of such properties is impractical in many cases because of the necessary hardware/software overhead or non-cooperative internal routers, and thus end-to-end measurement and inference methodologies are of great practical interest.

This paper investigates the use of back-to-back (closely time-spaced) unicast packets for network tomography. Network probing using back-to-back packets has been proposed in a number of measurement schemes [1, 4, 9, 23, 26]. The basic idea behind our estimation procedure is quite straightforward. Suppose two closely back-to-back packets are sent to two different receivers. The paths to these receivers share a common set of links from the sender but later diverge. If one of the packets is dropped and the other successfully received, then (assuming the packets experience the same fates on shared links) one can infer that the packet must have been dropped on one of the unshared links. This enables the resolution of losses and delay on individual links. We first proposed the basic elements of our approach in [12].

The methodology and analysis presented in this paper are important extensions of the work presented in [12], but the basic idea of exploiting correlations of closely-spaced packets remains the same. The technique we propose is unique in that it accounts for potential imperfections in the correlations. We develop a probabilistic model to describe packet losses and devise a maximum likelihood estimator (MLE) for the internal loss rate parameters. The maximization cannot be solved analytical, and we propose a novel Expectation-Maximization (EM) algorithm to compute the MLE. The EM algorithm presented in this paper differs in key respects from the method we proposed in earlier work [12], resulting in a much more computationally efficient procedure.

We also study the convergence behavior of the EM algorithm, proving that it converges to the set of global maxima. We characterize this set and show that in certain cases it contains a single, global maximum. We also provide a theoretical analysis of the correlation between losses of packet-pairs under an M/M/1/K queue model. This analysis demonstrates that the correlations are generally quite strong under this theoretical model, corroborating experimental observations in real networks [18, 23] and `ns` experiments. We further assess the performance of our approach through `ns` simulation experiments and more realistic experiments in a network testbed comprised of eight freeBSD routers.

We emphasize that the main contributions in this paper are a likelihood criterion for unicast network tomography, an EM algorithm for computing the result, and theoretical analysis of the algorithm. We do not discuss the validity of the assumptions underlying the likelihood criterion in great detail. Many of the issues surrounding our modeling assumptions are investigated elsewhere [5, 11]. Suffice it to say, there is compelling theoretical and experimental evidence to suggest that the key assumptions are reasonable approximations in many pratical situations.

B. Related Work

The problem of estimating internal loss rates was first considered in the MINC (Multicast Inference of Network Characteristics) [21] Project. An interesting strategy was proposed for estimating loss rates based on multicast packet probes [5]. The technique exploits the correlation between the losses/delays observed by multicast receivers. The performance of these algorithms is impressive [6], but there are two serious deficiencies in the methodology. Firstly, multicast protocols are not supported by significant portions of the Internet. Secondly, the internal performance measured by active multicast probes often differs significantly from that encountered by unicast packets, which comprise by far the most substantial component of Internet traffic [18].

Related tomography schemes have been developed or proposed independently by other researchers. Harfoush *et al.* developed a similar unicast tomography technique in [19], building on earlier work by Rubenstein *et al.* who devised a technique for detecting shared congestion in traffic flows [26]. The estimation procedure in [19] is based on the assumption of perfect correlations between losses of packet pairs on shared links. The authors in [18] proposed an alternative strategy based on sending multiple-packet probes to improve the observed correlations (compared to the correlations of packet-pairs), and then applied the multicast-based algorithms of the MINC project for loss estimation (thus also assuming perfect correlation). Our framework is distinct from these other methods in that we do not assume perfect correlations, since that assumption (when erroneous) produces biased estimates. Moreover, our framework procedure allows us to assess the severity of imperfect correlations and the impact this has on the accuracy of the loss rate estimates.

C. Organization

The paper is organized as follows. In Section 2, we review the basic unicast network tomography problem and describe the loss modeling and measurement framework. In Section 3, we pose the MLE problem and develop the EM algorithm. In Section 4, we study the convergence behavior of the EM algorithm. In Section 5, we investigate the correlations between back-to-back packet losses under an M/M/1/K queuing model. In Section 6, we report on `ns` simulations and network testbed experiments that explore the efficacy of our estimation procedures. In Section 7, we discuss possible alternatives to MLEs, describe extensions of the proposed framework to delay distribution estimation, and make concluding remarks.

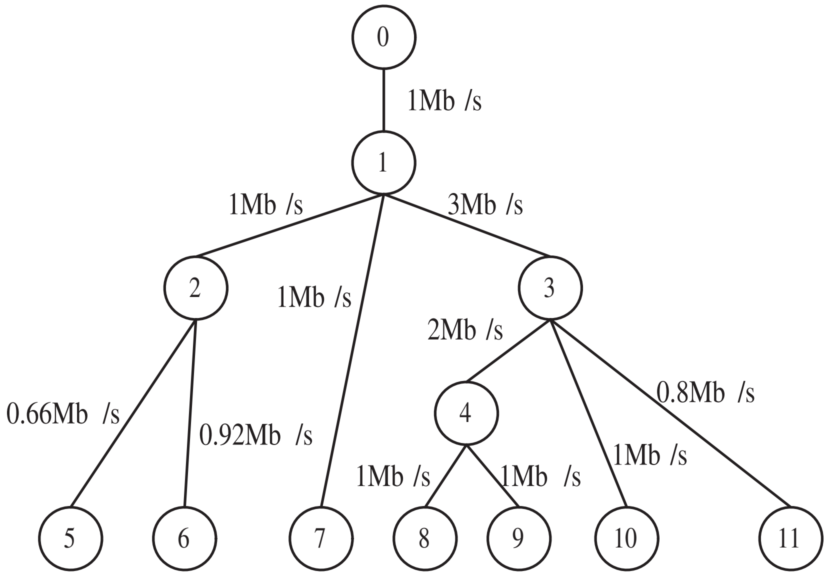


Fig. 1. Tree-structured graph representing a single-source, multiple-receiver network. Vertex 0 is the source, vertices 1-4 internal routers, and vertices 5-11 receivers. Beside each edge we indicate the capacity in megabits per second.

II. Problem Statement, Loss Models, and Measurement Framework

A. Problem Setup

In this paper we focus on networks in which a sender transmits packets to multiple receivers. The network tomography problem and methodology are can be extended to the multiple sender cases. Figure 1 depicts an example of this form of topology; the network appears to the sender as a tree. The vertices of the tree correspond to the sender (vertex 0), internal routers (vertices 1–4) and receivers (vertices 5–11). We define a *edge* as the connection between any two adjacent vertices in the tree, deem the set of edges connecting a sender and any receiver a *path*, and a subset of connected edges in a path is referred to as a *subpath*. We enumerate the edges by associating each edge with the vertex it connects to below; e.g., the edge connecting vertices 1 and 3 is called edge 3. The tree of Figure 1 does not necessarily depict all routers encountered by packets travelling from the sender to receivers. It is possible that a number of routers are passed as a packet travels along each edge; e.g., from between vertices 1 and 3. In the context of our framework, each vertex simply corresponds to a junction (or branching point) in the paths of different receivers. The routing is assumed to be known and constant throughout the measurement period. Although the routing tables in the Internet are periodically updated, these changes occur at intervals of several minutes. Our measurement periods are typically of much shorter durations.

The network tomography problem we consider in this paper deals with the estimation of loss rate on each edge in the network based on end-to-end (sender-to-receiver) measurements of packet losses. That is, we can observe whether or not each sent packet is received or not at the end receiver, but no other measurements are available. Such information is readily available through the TCP (Transmission Control

Protocol) acknowledgement system, for example.

B. Loss Models

Two types of packet measurements are utilized in our network tomography methodology:

(1) end-to-end losses of individual packets

(2) end-to-end losses of closely time-spaced (back-to-back) packet pairs

In both cases we assume that the packet measurements are well separated in time (i.e., the time periods between successive single packet or packet pair measurments is much larger than the spacing between back-to-back pairs).

We model loss/success of an individual packet with a Bernoulli distribution. The success probability of an individual measurement packet on edge i (the edge into vertex i) is defined as

$$p_{i,1} \equiv \Pr(\text{packet successfully transmitted from } \rho(i) \text{ to } i),$$

where $\rho(i)$ denotes the index of the parent vertex of vertex i (the vertex above i-th vertex in the tree; *e.g.,* referring to Figure 1, $\rho(1) = 0$). A packet is successfully sent from $\rho(i)$ to i with probability $p_{i,1}$ and is dropped with probability $1 - p_{i,1}$.

If a back-to-back packet pair is sent from vertex $\rho(i)$ to vertex i, then we define the conditional success probabilities:

$$p_{i,2} \equiv \Pr(\text{1st packet } \rho(i) \to i \mid \text{2nd packet } \rho(i) \to i\ ),$$

where 1st and 2nd refer to the temporal order of the two packets as they exit $\rho(i)$, and $\rho(i) \to i$ is shorthand notation denoting the successful transmission of a packet from $\rho(i)$ to i.

Although the loss models are merely simple approximations to the behavior of real networks, our `ns` simulations and network testbed experiments in Section VI demonstrate that these models appear to be reasonable for tomography purposes. Specifically, above models produce loss rate estimates that agree quite well with direct counts of actual packet losses.

C. Measurement Framework

Each edge in the tree has two (unknown) probabilities associated with it, the unconditional and conditional success probabilities, $p_{i,1}$ and $p_{i,2}$, respectively. These probabilities are related to the the single packet and back-to-back packet measurements that we will make, as described below. The measured data can be collected in a number of possible ways. For example, UDP (User Datagram Protocol) can be used for active probing or existing TCP connections may be passively sampled, in which case back-to-back events are selected from the TCP traffic flows.

Single Packet Measurement: Suppose that n_a packets are sent to receiver a and that of these a number $m_{a,1}$ are actually received ($n_a - m_{a,1}$ are dropped). The likelihood of $m_{a,1}$ given n_a is binomial (since Bernoulli losses are assumed) and is given by

$$l(m_{a,1} \,|\, n_a, q_a) \quad = \quad \binom{n_a}{m_{a,1}} q_a^{m_{a,1}} \, (1 - q_a)^{n_a - m_{a,1}},$$

where $q_a = \prod_{i \in \mathcal{P}(0,a)} p_{i,1}$ and $\mathcal{P}(0,a)$ denotes the sequence of vertices in the path from the sender 0 to receiver a.

Back-to-Back Packet Pair Measurement: Suppose that the sender transmits a large number of back-to-back packet pairs in which the first packet is destined for receiver a and the second for receiver b. We assume that the timing between pairs of packets is considerably larger than the timing between two packets in each pair. Let $n_{a,b}$ denote the number of pairs for which the second packet is successfully received at vertex b, and let $m_{a,b,1}$ denote the number of pairs for which both the first and second packets are received at their destinations. Furthermore, let $s_{a,b}$ denote the vertex at which the paths $\mathcal{P}(0,a)$ and $\mathcal{P}(0,b)$ diverge, so that $\mathcal{P}(0, s_{a,b})$ is their shared subpath. With this notation, the likelihood of $m_{a,b,1}$ given $n_{a,b}$ is binomial and is given by

$$l(m_{a,b,1} \,|\, n_{a,b}, q_{a,b}) \quad = \quad \binom{n_{a,b}}{m_{a,b,1}} q_{a,b}^{m_{a,b,1}} \, (1 - q_{a,b})^{n_{a,b} - m_{a,b,1}}, \tag{1}$$

where

$$q_{a,b} = \prod_{i \in \mathcal{P}(0, s_{a,b})} p_{i,2} \prod_{i \in \mathcal{P}(s_{a,b}, a)} p_{i,1}. \tag{2}$$

D. Plausibility of Modeling Assumptions

Here we attempt to shed some light on the physical plausibility of our loss models. The following assumptions regarding network behavior, partially support the loss models described above.

A1. The routing matrix is assumed to be known and constant throughout the measurement period. Although the routing tables in the Internet are periodically updated, these changes occur at intervals of several minutes. Our measurement periods are typically of much shorter durations.

A1. Packet losses (drops) are due to solely queue buffer overflow.

A2. The queuing behavior on all edges is stochastic and stationary over the observation period.

A3. Spatial Independence. The losses on each edge are assumed statistically independent of losses on all other edges.

A4. Temporal Independence. All packet and packet-pair measurements are statistically independent of each other (which is reasonable if the time separation between measurements is sufficiently large).

A5. The measurements do not effect that stationarity of the network. This assumption is reasonable if the measurement packets are well separated in time, and if the total number of measurement packets is negligible compared to the total traffic.

Although many of the simplifying assumptions do not strictly hold, our `ns` simulations and network testbed experiments demonstrate that these approximations appear to be reasonable for tomography purposes. The possibility of long-range temporal dependencies in network traffic due to common cross-traffic on different edges could presumably lead to temporal and spatial correlations in the losses experience on those edges. However, theoretical analysis of spatial and temporal dependencies in multicast trees shows that the dependencies may not strongly affect the MLEs in the multicast setting. [5]. A similar analysis carries over to the unicast case considered here and suggests that our MLE estimator (derived in Section III-A) can yield accurate results in the presence of moderate levels of dependence. Practically speaking, network routers usually have many inputs and many outputs; in many cases the proportion of shared traffic on edges is relatively small compared to the total traffic, in which case the losses on different links should be at worst weakly dependent.

III. MLE and EM Algorithm

We wish to estimate the network success rates $\boldsymbol{p} = \{p_{i,1}, p_{i,2}\}$. Notice that $\boldsymbol{p}$ contains the single-packet, unconditional success probabilities as well as the packet-pair, conditional success probabilities. Estimates of the latter probabilities provide an accuracy measure for the unconditional success probabilities $\{p_{i,1}\}$, as shown below. This is an important issue that does not arise in the multicast case, because the multicast probes are perfectly correlated (in effect, the probabilities $\{p_{i,2}\}$ are all exactly one).

We will derive a maximum likelihood estimator of $\boldsymbol{p}$ given the entire set of single packet and back-to-back packet measurements. For convenience, define $m_{a,0} \equiv n_a - m_{a,1}$ and $m_{a,b,0} \equiv n_{a,b} - m_{a,b,1}$. Collecting all the measurements, define

$$\boldsymbol{y} \equiv \{m_{a,k}\} \cup \{m_{a,b,k}\} \tag{3}$$

where the index a alone runs over all receivers and the indices a, b run over all pairwise combinations of receivers in the network. The index k is a binary variable that indicates failure (0) or success (1).

As before denote the collection of the unconditional and conditional edge success probabilities as $\boldsymbol{p}$. The *joint* likelihood of all measurements is given by

$$l(\boldsymbol{y}|\boldsymbol{p}) \propto \prod_a l(m_{a,0}, m_{a,1}\,|\boldsymbol{p}) \times \prod_{a,b} l(m_{a,b,0}, m_{a,b,1}|\boldsymbol{p}). \tag{4}$$

Since $\boldsymbol{y}$ is known, we view $l(\boldsymbol{y}\,|\,\boldsymbol{p})$ as a function of the unknown probabilities $\boldsymbol{p}$. We call $l(\boldsymbol{y}\,|\,\boldsymbol{p})$ the likelihood function of $\boldsymbol{p}$. The maximum likelihood estimate of $\boldsymbol{p}$ is defined as

$$\widehat{\boldsymbol{p}} = \arg\max_{\boldsymbol{p}}\ l(\boldsymbol{y}\,|\,\boldsymbol{p}). \tag{5}$$

Maximum likelihood estimation enjoys many desirable properties and is widely utilized in statistical inference [2].

Computing the maximum likelihood estimates is a formidable task. Directly attempting the maximization inherent in (5) leads to extremely computationally demanding algorithms that are not scalable to large networks. The basic problem is that the individual likelihood functions $l(m_{a,0}, m_{a,1} \,|, \boldsymbol{p})$ or $l(m_{a,b,0}, m_{a,b,1} \,|\, \boldsymbol{p})$ for each type of measurement involve products of subsets of the $\boldsymbol{p}$ parameters. Consequently, it is difficult to separate the effects of each individual success probability. As a result, numerical optimization strategies are required.

Before describing such an optimization technique, let us comment on the identifiability of the success probabilities. It is not hard to see that in general a unique maximum likelihood solution does not exist (e.g., consider even a simple two receiver network). However, if both packets in all back-to-back pairs experience the same fate on each shared edge (either both are successful or both are dropped), then a unique maximum likelihood solution exists. In such a setting, the conditional success probabilities $\{p_{i,2}\}$ are all equal to 1 and, consequently, it is easy to verify that the unconditional success probabilities $\{p_{i,1}\}$ are identifiable. In practice the conditional success probabilities are less than perfect, and the potential non-uniqueness of a maximum likelihood solution can lead to biased estimates. It is possible, however, to quantify the potential severity of this bias in terms of observable quantities.

If the conditional success probabilities $\{p_{i,2}\}$ are all exactly one, then it can be shown that maximum likelihood estimates of the unconditional losses $\{p_{i,1}\}$ will tend to their true values as the number of packet measurements increases. This can be understood by considering a single path from the source to receiver a. The single packet measurements $m_{a,1}$ and $n_a = m_{a,0} + m_{a,1}$ provide an asymptotically consistent estimator of the product $q_a = \prod_{i \in \mathcal{P}(0,a)} p_{i,1}$. Specifically, $\widehat{q}_a \equiv \frac{m_{a,1}}{n_a}$ converges to q_a as n_a tends to infinity. Similarly, the estimators $\widehat{q}_{a,b} \equiv \frac{m_{a,b,1}}{n_{a,b}}$, converge to

$$q_{a,b} = \prod_{i \in \mathcal{P}(0,s_{a,b})} p_{i,2} \prod_{i \in \mathcal{P}(s_{a,b},a)} p_{i,1},$$

as each $n_{a,b} \to \infty$ (recall that the vertex $s_{a,b}$ defines the subpath common to both receivers). If $p_{i,2} = 1$ for all i, then there exists a one-to-one mapping between path success probabilities $\{q_a, q_{a,b}\}$ and the edge success probabilities $\{p_{i,1}\}$.

If the $p_{i,2}$ are close to but not exactly one, then the relationship between $\boldsymbol{q} = \{q_a, q_{a,b}\}$ and $\boldsymbol{p} = \{p_{i,1}, p_{i,2}\}$ is one-to-many (i.e., there may be more than one $\boldsymbol{p}$ corresponding to each value of $\boldsymbol{q}$). However, the inverse image of $\boldsymbol{q}$ in $\boldsymbol{p}$-space is shown to be well-concentrated about the "true" $\boldsymbol{p}$ value so long as the $\{p_{i,2}\}$ are close to one (see Section IV-B). Thus, so long as the back-to-back success probabilities are sufficiently close to one (as theory and experiments strongly suggest), any member of the inverse image set will provide a fairly accurate result.

A. EM Algorithm

We overcome the difficulty in maximizing the joint likelihood function by using a common device in computational statistics known as *unobserved* data or variables. Suppose it were possible to measure how many packets successfully traversed each internal edge and how many were dropped. We will use $z_{i,1,1}$ to denote the number of single packets that successfully traversed edge i and $z_{i,1,0}$ to denote the number that were dropped. Similarly, we use $z_{i,2,1}$ to denote the number of packet-pairs that successfully traversed edge i and $z_{i,2,0}$ to denote the number of times that the second packet in a pair was successful but the first packet was dropped. Let $z_i = \{z_{i,j,k}\}_{j=1,2;k=0,1}$ and $\boldsymbol{z} = \{z_i\}$. These measurements are not observed, so $\boldsymbol{z}$ is called the *unobserved data*. Define the complete data $\boldsymbol{x} = \{\boldsymbol{y}, \boldsymbol{z}\}$.

Associated with the complete data is the *complete data likelihood function*. To simplify the notation in this section, we let $p_{i,j,1}$ denote the success probability associated with $z_{i,j,1}$ (note that throughout the other sections in the paper this probability is denoted $p_{i,j}$). The probability of loss associated with $z_{i,j,0}$ is denoted by $p_{i,j,0}$ (this probability is simply $1 - p_{i,j}$). The key feature of the complete data likelihood function is that it is a product of factors, each involving just a single success probability $p_{i,j,1}$ or loss probability $p_{i,j,0}$. We can write

$$l(\boldsymbol{x} \,|\, \boldsymbol{p}) \propto \prod_{i,j,k} p_{i,j,k}^{z_{i,j,k}}. \tag{6}$$

Thus, the complete data likelihood function is a trivial multivariate function, and the effects of the individual edge probabilities are easily separated.

The EM algorithm [20] uses the complete data likelihood function to perform the maximization in (5). Beginning with an initial value for $\boldsymbol{p}$, denoted $\boldsymbol{p}^{(0)}$, the algorithm is iterative and alternates between two steps until convergence. The Expectation (E) Step computes the conditional expected value of the unobserved data given the observed data, under the probability law induced by the current estimates of $\boldsymbol{p}$. At the $r+1$-st iteration of the EM algorithm the E-Step computes

$$Q(\boldsymbol{p}, \boldsymbol{p}^{(r)}) = \mathrm{E}_{\boldsymbol{p}^{(r)}}[\log l(\boldsymbol{x}|\boldsymbol{p}) \,|\, \boldsymbol{y}], \tag{7}$$

where $\boldsymbol{p}^{(r)}$ is the iterate from the previous iteration. The Maximization (M) Step maximizes this $Q(\boldsymbol{p}, \boldsymbol{p}^{(r)})$ with respect to $\boldsymbol{p}$, thus updating the current estimate. That is,

$$\boldsymbol{p}^{(r+1)} = \arg\max_{\boldsymbol{p}} Q(\boldsymbol{p}, \boldsymbol{p}^{(r)}). \tag{8}$$

Evaluations of the original likelihood function at the iterates produced by the EM algorithm form a non-decreasing sequence; i.e., $l(\boldsymbol{y}|\boldsymbol{p}^{(0)}) \leq l(\boldsymbol{y}|\boldsymbol{p}^{(1)}) \leq l(\boldsymbol{y}|\boldsymbol{p}^{(2)}) \leq \ldots$, and thus the EM algorithm tends to increase the original likelihood objective.

Notice that the complete data log likelihood is linear in $\boldsymbol{z}$:

$$\log l(\boldsymbol{x}|\boldsymbol{p}) \propto \sum_{i,j,k} z_{i,j,k} \log p_{i,j,k}.$$

Thus, in the E-Step we need only compute the expectation of $\boldsymbol{z} = \{z_{i,j,k}\}$. Consider the expectation of $z_{i,1,1}$, the parameter counting the success of individual packets traversing edge i. Let $\boldsymbol{p}^{(r)}$ denote the estimate of $\boldsymbol{p}$ after the r-th iteration. For each successful measurement $m_{a,1}$ such that the path $\mathcal{P}(0,a)$ involves edge i, we know that the packet successfully traversed edge i. We can make a similar observation for all the packet-pair measurements $m_{a,b,1}$ such that $i \in \mathcal{P}(s_{a,b}, a)$. The case of the unsuccessful measurements $m_{a,0}$ is somewhat more complicated. For each of these measurements, the probability that edge i was successfully traversed is equal to the the probability that the drop occurred on some edge on the path from vertex i to the receiver. Based on these considerations, we can write the expectation of $z_{i,1,1}$ as:

$$\begin{aligned} \mathrm{E}_{\boldsymbol{p}^{(r)}}[z_{i,1,1}|\boldsymbol{y}] &= \sum_{a\,:\,i\in\mathcal{P}(0,a)} m_{a,1} + m_{a,0}\left[\prod_{t\in\mathcal{P}(0,i)} p^{(r)}_{t,1,1} \sum_{u\in\mathcal{P}(i,a)} p^{(r)}_{u,1,0} \prod_{v\in\mathcal{P}(i,\rho(u))} p^{(r)}_{v}, 1, 1\right] + \\ &\sum_{(a,b)\,:\,i\in\mathcal{P}(s_{a,b},a)} m_{a,b,1} + m_{a,b,0}\left[\prod_{t\in\mathcal{P}(0,s_{a,b})} p^{(r)}_{t,2,1} \prod_{u\in\mathcal{P}(s_{a,b},i)} p^{(r)}_{u,1,1} \sum_{v\in\mathcal{P}(i,a)} p^{(r)}_{v,1,0} \prod_{w\in\mathcal{P}(i,\rho(v))} p^{(r)}_{w,1,1}\right] \end{aligned} \tag{9}$$

Similarly, we can determine the expectation of $z_{i,2,1}$:

$$\begin{aligned} \mathrm{E}_{\boldsymbol{p}^{(r)}}[z_{i,2,1}|\boldsymbol{y}] &= \sum_{(a,b)\,:\,i\in\mathcal{P}(0,s_{a,b})} m_{a,b,1} + m_{a,b,0}\left[\prod_{t\in\mathcal{P}(0,i)} p^{(r)}_{t,2,1} \sum_{u\in\mathcal{P}(i,s_{a,b})} p^{(r)}_{u,2,0} \prod_{v\in\mathcal{P}(i,\rho(u))} p^{(r)}_{v,2,1}\right] + \\ &m_{a,b,0}\left[\prod_{t\in\mathcal{P}(0,s_{a,b})} p^{(r)}_{t,2,1} \sum_{u\in\mathcal{P}(s_{a,b},a)} p^{(r)}_{u,1,0} \prod_{v\in\mathcal{P}(s_{a,b},\rho(u))} p^{(r)}_{v,1,1}\right] \end{aligned} \tag{10}$$

The expectations of the failure counts ($z_{i,1,0}$ and $z_{i,2,0}$) can be calculated directly from those of the success counts.

With the expectation expressions in hand, the EM algorithm takes the following form.

The computational complexity of the EM algorithm is related to the number of edges L as follows. The M-Step requires $O(L)$ operations. In general, the E-Step poses the majority of the computational burden. In the E-Step, we evaluate (9) and (10) for each internal vertex in the network. The computational complexity of the calculation of all the necessary conditional expectations is dependent on the total number of edges and the network topology. It ranges from $O(L)$ to $O(L^2)$ operations, where L is the total number of edges in the network. The two extreme cases are depicted in Figures 2. The computational complexity in the case of a perfectly balanced binary tree (all subtrees have the same depth) is $O(L \log_2 L)$. The Appendix contains the complexity analysis leading to these results. Thus, the overall complexity of each iteration of the EM algorithm lies between $O(L)$ to $O(L^2)$ operations.

EM Algorithm

Initialize:	Initialize the estimates $\boldsymbol{p}^{(0)}$
E-Step (iteration r):	Calculate the conditional expectation of $\boldsymbol{z}$ given $\boldsymbol{p}^{(r)}$ and $\boldsymbol{y}$ using (9) and (10). Label the vector $\boldsymbol{z}^{(r)}$ and plug $\boldsymbol{z}^{(r)}$ into the expression for the complete data log likelihood to obtain $Q(\boldsymbol{p}, \boldsymbol{p}^{(r)})$
M-Step (iteration r):	Computer $\boldsymbol{p}^{(r+1)}$ according to (8). This corresponds to evaluating $$p_{i,1,1}^{(r+1)} = \frac{z_{i,1,1}^{(r)}}{z_{i,1,1}^{(r)} + z_{i,1,0}^{(r)}}$$ $$p_{i,2,1}^{(r+1)} = \frac{z_{i,2,1}^{(r)}}{z_{i,2,1}^{(r)} + z_{i,2,0}^{(r)}}$$ The loss probability estimates $p_{i,1,0}^{(r+1)}$ and $p_{i,2,0}^{(r+1)}$ are simply one minus the success probability estimates.

IV. Analysis of EM Algorithm

A. Convergence

The EM algorithm proposed above is guaranteed to converge to a global maximum point of the likelihood function. This is established by noting the following properties.

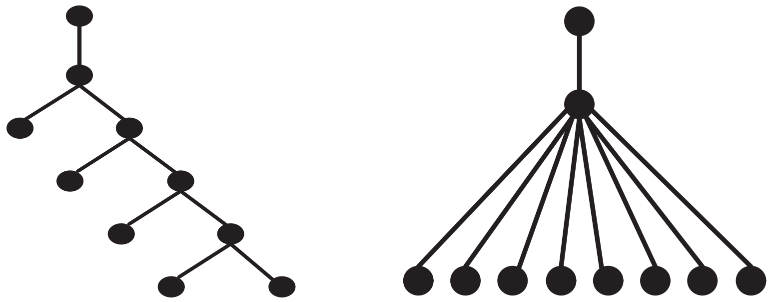
1. The EM algorithm generates a monotonic increasing sequence of likelihood values.

2. $Q(\boldsymbol{p}, \boldsymbol{p}^{(r)})$ defined in (7) is continuous in both arguments (in the interior of the parameter space $[0,1]^{2L}$).

3. The log-likelihood is a concave function in $\log \boldsymbol{p}$.

The first and second properties are easy to verify for the problem at hand. The second property and results of [32] guarantee that the EM algorithm converges to a stationary point of the likelihood function (assuming all stationary points are in the interior of the parameter space, e.g., $(0,1)^{2L}$. The third property (concavity) guarantees that interior stationary points are points where the global maximum likelihood value is achieved.

The concavity property is established as follows. The likelihood function can be reparameterized in terms of $\{\log p_{i,1}, \log p_{i,2}\}$. Note that there is a one-to-one mapping between this and the original parameterization, and the MLEs for $\{\log p_{i,1}, \log p_{i,2}\}$ are simply the logarithms of the MLEs for $\{p_{i,1}, p_{i,2}\}$. Hence, it suffices to show that the log-likelihood is concave in $\{\log p_{i,1}, \log p_{i,2}\}$.



(a) (b)

Fig. 2. The two topologies that lead to the extremes of computational complexity for the EM algorithm. Topology (a) leads to the worst case complexity of $O(L^2)$ operations, where L is the number of edges in the tree. Topology (b) leads to the best case complexity of $O(L)$ operations.

Recall that the conditional *path* success probability for a packet-pair measurement is given by

$$q_{a,b} = \prod_{i \in \mathcal{P}(0,s_{a,b})} p_{i,2} \prod_{i \in \mathcal{P}(s_{a,b},a)} p_{i,1},$$

where the conditional edge success probabilities $\{p_{i,2}\}$ appear on shared edges and the unconditional edge success probabilities $\{p_{i,1}\}$ appear on unshared edges. The likelihood function, in terms of the path success probabilities, is a product of binomials parameterized by $\{q_{a,b}\}$ and hence concave in $\{q_{a,b}\}$. The log conditional *path* success probability is

$$\log q_{a,b} = \sum_{i \in \mathcal{P}(0,s_{a,b})} \log p_{i,2} \sum_{i \in \mathcal{P}(s_{a,b},a)} \log p_{i,1},$$

and it is easy to check that the log-likelihood function is also a concave function in $\{\log q_{a,b}\}$. Single-packet measurements have a similar form involving only unconditional edge success probabilities. Note that the log path success probabilities are linear combinations of the log edge success probabilities.

Let $\mathbf{q}$ denote the collection of all path success probabilities (conditional and unconditional), and let ℓ' denote the log-likelihood function for all the measurements (i.e., a sum of log binomial likelihood functions). It follows from the discussion above that the log-likelihood function $\ell'(\log \mathbf{q})$ is concave in $\log \mathbf{q}$. Furthermore, the log path success probabilities are linearly related to the log edge success probabilities: $\mathbf{q} = \mathbf{T} \log \boldsymbol{p}$, where $\mathbf{T}$ is a matrix of 1's and 0's whose rows correspond to the combinations of log edge success probabilities required for each path. This allows us to express the log-likelihood as a function of $\{\log \boldsymbol{p}\}$ with the equivalence $\ell'(\log \mathbf{q}) = \ell'(\mathbf{T} \log \boldsymbol{p}) \equiv \ell(\log \boldsymbol{p})$, where ℓ is parameterized in terms of the log edge success probabilities. ¿From here it is easy to verify that ℓ is concave in $\log \boldsymbol{p}$. Let $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ be any two sets of edge success probabilities. Then for $\lambda \in [0,1]$

$$\begin{aligned}\ell(\lambda \log \boldsymbol{p}_1 + (1-\lambda)\log \boldsymbol{p}_2) &= \ell'(\lambda \mathbf{T} \log \boldsymbol{p}_1 + (1-\lambda)\mathbf{T}\log \boldsymbol{p}_2), \\ &\geq \lambda \ell'(\mathbf{T}\log \boldsymbol{p}_1) + (1-\lambda)\ell'(\mathbf{T}\log \boldsymbol{p}_2), \\ &= \lambda \ell(\log \boldsymbol{p}_1) + (1-\lambda)\ell(\log \boldsymbol{p}_2),\end{aligned}$$

where the inequality above follows from the concavity of ℓ'. This establishes the concavity of ℓ in $\log \boldsymbol{p}$. We summarize this result in the following theorem.

Theorem 1 *The log of the likelihood function given by (4) is concave in* $\log \boldsymbol{p}$*. Furthermore, if all stationary points are in the interior of* $[0,1]^{2L}$*, then the EM algorithm converges to the set of global maxima.*

It is also interesting to consider cases in which the conditional success probabilities are fixed (possibly incorrectly) to be one for all edges. In this case, the likelihood function has the same form as before, except that each product of conditional success probabilities in (2) is equal to one; e.g., $\prod_{i\in\mathcal{P}(0,s_{a,b})} p_{i,2} = 1$. As pointed out earlier, in this case there is a one-to-one mapping between $\mathbf{q}$, the path success probabilities, and $\boldsymbol{p} = \{p_{i,1}\}$, the unconditional success probabilities. Since the MLE of $\boldsymbol{q}$ is unique, it follows that the MLE of $\boldsymbol{p}$ is also unique in this case. Thus, the likelihood has a single stationary point — the global maximum. The convergence results above guarantee that the EM algorithm converges to the global maximum. Again, we summarize the results with a theorem.

Theorem 2 *If the conditional success probabilities are set to one* ($p_{i,2} = 1,\ \forall i$)*, then the log of the likelihood function given by (4), with* $\mathbf{p} = \{p_{i,1}\}$*, is concave in* $\log \boldsymbol{p}$ *and the EM algorithm converges to the the global maximum.*

B. Characterizing the Set of Global Maxima

In the general case ($p_{i,2}$ not fixed to one), the EM algorithm converges to the set of global maxima. We study the structure of this set as the number of measurements tend to infinity. We derive coordinate-wise bounds on this limit set and show that it is highly concentrated about the "true" values if the conditional success probabilities $\{p_{i,2}\}$ are close to one.

First consider single packet measurements on the path from the sender to receiver a. The measurements $m_{a,1}$ and n_a provide an asymptotically consistent estimator of the product $q_a = \prod_{i\in\mathcal{P}(0,a)} p_{i,1}$. Specifically, $\widehat{q}_a \equiv \frac{m_{a,1}}{n_a}$ converges to q_a as n_a tends to infinity. Similarly, for packet-pair measurements, the estimators $\widehat{q}_{a,b} \equiv \frac{m_{a,b,1}}{n_{a,b}}$, converge to

$$q_{a,b} = \prod_{i\in\mathcal{P}(0,s_{a,b})} p_{i,2} \prod_{j\in\mathcal{P}(s_{a,b},a)} p_{j,1},$$

as each $n_{a,b} \to \infty$ (recall that the vertex $s_{a,b}$ defines the subpath common to both receivers).

Now consider packet-pair measurements along the same path. Denote the sequence of vertices in the path $\mathcal{P}(0,a)$ by $\{i_1, i_2, \ldots, i_T\}$, where $i_T = a$. Let $r_1, \ldots, r_T$ also denote a set of receiver vertices chosen so that the shared subpath between $\mathcal{P}(0, r_t)$ and $\mathcal{P}(0,a)$ is $\mathcal{P}(0, i_t)$, $t = 1, \ldots, T$ (note that $r_T \equiv i_T = a$). Then we have

$$\begin{aligned}
\widehat{q}_a &\rightarrow p_{i_1,1}\, p_{i_2,1}\, p_{i_3,1} \cdots p_{i_T,1}, \\
\widehat{q}_{a,r_1} &\rightarrow p_{i_1,2}\, p_{i_2,1}\, p_{i_3,1} \cdots p_{i_T,1}, \\
\widehat{q}_{a,r_2} &\rightarrow p_{i_1,2}\, p_{i_2,2}\, p_{i_3,1} \cdots p_{i_T,1}, \\
&\vdots \\
\widehat{q}_{a,r_T} &\rightarrow p_{i_1,2}\, p_{i_2,2}\, p_{i_3,2} \cdots p_{i_T,2}.
\end{aligned}$$

If $\widehat{q}_{a,a} \rightarrow \gamma < 1$, then we can deduce that

$$\gamma \ \leq \ \prod_{k=1}^{t} p_{i_k,2} \ \leq \ 1$$

for $t = 1 \ldots, T$. This shows that the asymptotic value of $\widehat{q}_{a,r_{T-i}}$ lies within the interval

$$\left[\gamma \prod_{k=T-t+1}^{T} p_{i_k,1} \ , \ \prod_{k=T-t+1}^{T} p_{i_k,1} \right],$$

for $t = 1, \ldots, T-1$. From here it follows that for any global maximum point $\widehat{\boldsymbol{p}}$ the $p_{i_t,1}$-coordinate must lie within the interval

$$\left[\gamma\, p_{i_t,1} \ , \ \frac{1}{\gamma}\, p_{i_t,1} \right];$$

if not, then the vector $\widehat{\boldsymbol{p}}$ cannot map to the MLE $\widehat{\mathbf{q}}$, contradicting the fact that it is one of the global maxima. We summarize our conclusions with the following theorem.

Theorem 3 *Suppose that the number of all measurements tends to infinity ($n_a \rightarrow \infty$ and $n_{a,b} \rightarrow \infty$ for all receivers a, b). Define*

$$\gamma_i \equiv \max_{a \,:\, i \in \mathcal{P}(0,a)} \ \lim_{n_{a,a} \rightarrow \infty} \frac{m_{a,a,1}}{n_{a,a}}.$$

The $p_{i,1}$-coordinates of the limit set of global maxima of the likelihood function lie within the interval

$$\left[\gamma_i\, p_{i,1} \ , \ \frac{1}{\gamma_i}\, p_{i,1} \right], \quad \forall\ i,$$

where $p_{i,1}$ is the true unconditional success probability.

Theorem 3 shows that the $p_{i,1}$-coordinates of the EM iterates $\boldsymbol{p}^{(r)}$ tend to values very close to their corresponding "true" values, if the conditional success probabilities are close to one; the closer to one they are, the tighter the limit set is about the true unconditional success probabilities. Note that the values γ_i can be obtained directly from the observable measurements, giving one a computable estimate of the "accuracy" of the

MLE of $p_{i,1}$. A queuing-theoretic argument in the next section shows that the conditional success probabilities $\{p_{i,2}\}$ are indeed very close to unity. This conclusion is supported by our experimental results. Also note that the estimate obtained by setting the $\{p_{i,2}\}$ to one generally leads to an underestimation of the unconditional success probabilities $\{p_{i,1}\}$; i.e., in that case the MLE attributes losses due to imperfect conditional success probabilities to the unconditional losses. The joint MLE of both types of success probabilities can mitigate this deficiency.

V. M/M/1/K Queuing Behavior of Back-to-Back Packet Pairs

If both packets in a back-to-back pair share the same fate on each edge (either both are successful or both are dropped), then the unicast tomography problem is somewhat similar to the multicast tomography problem in that, like multicast probes, the losses of packet pairs are perfectly correlated. In such a setting, the conditional success probabilities $\{p_{i,2}\}$ are all equal to one and, consequently, it is easy to check that the unconditional success probabilities $\{p_{i,1}\}$ are identifiable. In practice the conditional success probabilities are less than perfect, and it is again easy to verify that the collection of success probabilities $\{p_{i,1}, p_{i,2}\}$ is not identifiable from the measurements described in the previous section. There is reason to believe that the conditional success probabilities may be very close to 1, in which case the success probabilities are "almost" identifiable. We will examine the issue of identifiability in more detail in Section IV-A. Internet measurements [4, 18, 23] have shown that the conditional success probabilities $\{p_{i,2}\}$ are typically very close to one, but to the best of our knowledge there are no previous theoretical studies that corroborate these findings.

To investigate this phenomenom further, here we explore the queuing behavior of back-to-back packets under the classical M/M/1/K queue model. We then examine the results of `ns` simulations of scenarios that closely mirror the traffic arrival patterns measured at queues in the Internet and demonstrate a close correspondence to our theoretical results.

Consider an M/M/1/K queue with arrival rate λ_a and service rate λ_s. The queue is obeys a $K+1$-state Markov chain with transition probabilities p_a and p_s, corresponding to moves up or down the chain, respectively. Let $\{q_j\}_{j=0}^{K}$ denote the stationary queue distribution.

Suppose that two closely time-spaced packets reach the queue at nearly the same time. Specifically, assume that there are r intervening events (arrivals and services) between the arrivals of the two packets. We are interested in the probability that the first packet makes it into the queue, conditioned on the event that the second packet also successfully enters the queue. In other words, we will examine the probability that the queue is not full when the first packet arrives, conditional on the fact that it is not full when the second packet arrives.

There are four possible outcomes for the two packets: $\{0,0\}, \{0,1\}, \{1,0\}$, and $\{1,1\}$, where a 0 or 1 in the first position indicates the loss and success, respectively, of the first packet, and the second position denotes the outcome of the second packet. The probability we are interested in is given by

$$p^* \equiv \frac{\Pr(\{1,1\})}{\Pr(\{1,1\}) + \Pr(\{0,1\})},$$

the joint probability that both packets successfully enter the queue divided by the marginal probability of success for the second packet alone.

First, consider the probability $\Pr(\{1,1\})$. We can write this probability as

$$\Pr(\{1,1\}) = \sum_{j=0}^{K-1} q_j p_{j+1},$$

where p_j denotes the probability that the queue is not full after r steps of the chain beginning at vertex j (the dependence on r is suppressed for notational convenience). Explicit expressions for the probabilities $\{p_j\}$ can be obtained, but it is not necessary for the purposes of our analysis. The expression above can be interpreted as follows. In order that the first packet is successful, the queue must not be full when it arrives (corresponding to q_j in the above expression). The second packet will only make it into the queue if, after r intervening events, the queue is again not full (corresponding to the p_j in the above expression).

Second, consider the probability $\Pr(\{0,1\})$. We can write this probability simply as

$$\Pr(\{0,1\}) = q_K p_K,$$

since the queue must be full when the first packet arrives and not full when the second arrives. Combining the expressions for $\Pr(\{1,1\})$ and $\Pr(\{0,1\})$ we obtain an expression for the desired conditional probability

$$\begin{aligned} p^* &= \frac{\sum_{j=0}^{K-2} q_j p_{j+1} + q_{K-1} p_K}{\sum_{j=0}^{K-2} q_j p_{j+1} + q_{K-1} p_K + q_K p_K} \\ &= 1 - q_K \frac{p_K}{\sum_{j=0}^{K-2} q_j p_{j+1} + (q_{K-1} + q_K) p_K}. \end{aligned}$$

Note that the *unconditional* probability that the first packet successfully enters the queue is simply $1 - q_K$. Thus, from the above expression, we see that the conditional success probability will be greater than or equal to the unconditional success probability if the following condition holds:

$$\frac{p_K}{\sum_{j=0}^{K-2} q_j p_{j+1} + (q_{K-1} + q_K) p_K} \leq 1.$$

The following argument shows that in fact the condition is true. Invert the inequality above to obtain an equivalent condition

$$\sum_{j=0}^{K-2} q_j \frac{p_{j+1}}{p_K} + q_{K-1} + q_K \geq 1.$$

This inequality holds since $\frac{p_{j+1}}{p_K} \geq 1$, $j = 0, \ldots, K-2$, and $\sum_{j=0}^{K} q_j = 1$. The condition $p_K \leq p_{j+1}$ is a consequence of the fact that every sequence of r steps that leads to the full state K starting from state j also leads to a full state starting from state K. However, the converse is not necessarily true. A sequence of r steps that leads to state K starting from K may not lead to state K starting from j.

Note that if $r = 0$, then $p_K = 0$ and we have $p^* = 1$ (*i.e.*, if there are no intervening events, then the conditional success probability is perfect). Also, observe that as $r \to \infty$ the effect of the initial state of the queue diminishes and $p_j \to 1 - q_K$, $j = 0, \dots, K$. Thus, as $r \to \infty$ the conditional probability $p^* \to 1 - q_K$, the unconditional success probability. The results are summarized in the next theorem.

Theorem 4 *Under an M/M/1/K queue model, the conditional success probability p^* is greater than or equal to the unconditional success probability. Moreover, if r denotes the number of intervening events between the two packets, then $p^*(0) = 1$ and $\lim_{r\to\infty} p^*(r) = 1 - q_K$, the unconditional success probability.*

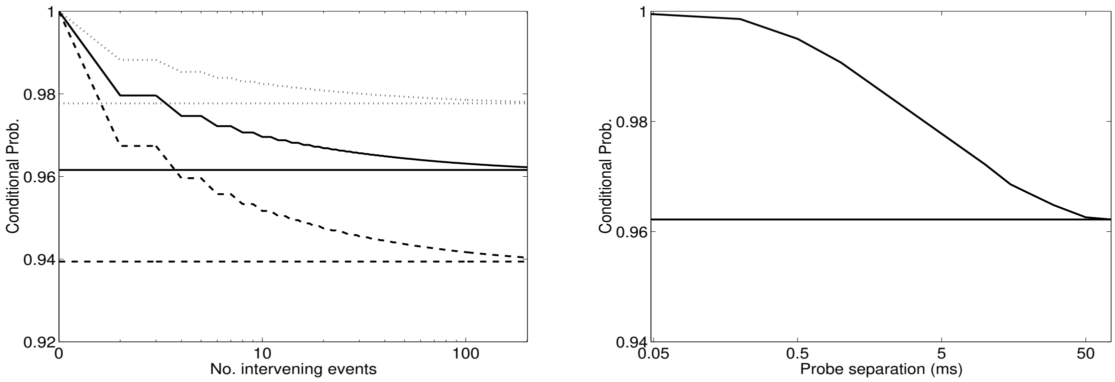
This theorem describes the behavior of the conditional success probability as a function of intervening queueing events under the unrealistic model of an M/M/1/K queue. Figure 3(a) plots the conditional success probabilities as functions of the number of intervening events r for several values of unconditional success probability. To examine the behavior in a more realistic environment, we simulated a network using the `ns-2` simulation environment. Competing traffic was generated at a queue by multiple TCP and UDP connections from 40 different links entering the queue. We calculated the conditional success rate by sending several thousand packet pairs into the queue. The experiment was repeated as the spacing between the probe packets within a pair was varied. Figure 3(b) shows the variation in conditional probability as the spacing between the probe packets changes. We observe a similar behavior as in the theoretical M/M/1/K result: the conditional probability is very close to one for very small packet spacing and decays to the unconditional probability as the spacing is increased.

VI. ns Simulations and Testbed Experiments

A. *ns Simulation*

Using the 12-vertex network topology of Figure 1, we evaluated the performance of the EM loss inference algorithm in the `ns-2` simulation environment [22]. The topology is intended to reflect the heterogeneous nature of many networks – a slower entry edge from the source, a faster internal backbone, and then slower exit edges to the receivers. This chosen topology gives us the flexibility to explore the effects of having receivers at different distances from the source (number of edges in path), and to examine the effect of varying fan-outs. We fix the queue size at each router to be 35 packets, and drops (losses) occur when a queue overflows.

Our experiments investigated a variety of network traffic conditions, comprised of TCP connections from the source to receivers as well as background cross-traffic flows. Single packet and packet pair statistics were collected by monitoring the TCP connections. Within these connections, we identify two packets as a "pair" if the time-spacing between them is less than 2 msec. Details of the scheme for packet pair identification



(a) (b)

Fig. 3. (a) The conditional success probability (M/M/1/K queue) as a function of the number of queueing events intervening between the arrivals of the two probe packets comprising a packet pair. The behaviors for queues with three different unconditional success probabilities are displayed. The horizontal lines indicate the unconditional success probabilities (dashed 0.9395, solid 0.961, dotted 0.9785). The decaying lines are the corresponding conditional success probabilities. (b) The results of `ns-2` experiments simulating a queue with competing traffic generated from multiple TCP and UDP connections. The horizontal line shows the unconditional success probability of probes entering the queue with a spacing of 100 ms. The decaying line displays the conditional success probability observed as the spacing between the probes in a packet-pair is increased from 0.05 ms to 75 ms.

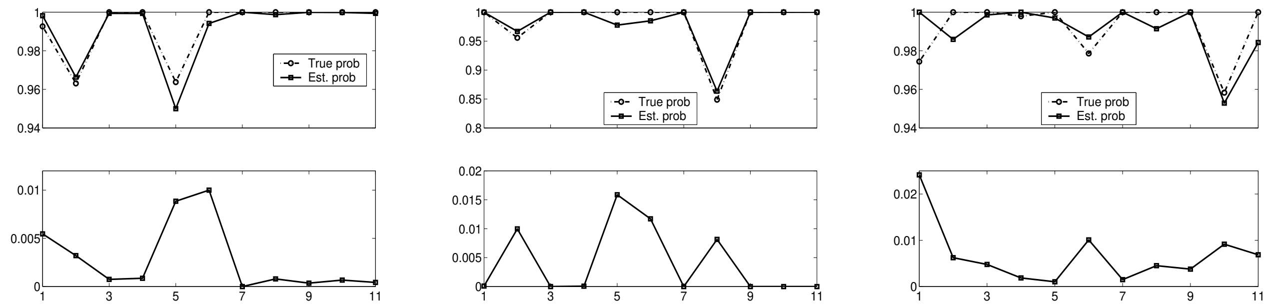
appear in [29].

In this paper, we report the results from measurements collected over a 300 second interval in three different traffic scenarios. The first two scenarios investigate cases in which traffic and losses are heaviest on two edges. The scenarios test the ability of the algorithm to resolve cascaded losses (edges 2 and 5, Scenario (a) in Figure 3) or identify isolated lossy edges in the network (edges 2 and 8, Scenario (b) in Figure 3). In the third scenario, more evenly distributed traffic introduces medium losses at several edges, exploring performance in more benign conditions (Scenario (c) in Figure 3).

In each case, we conducted ten independent simulations. Figure 3 displays the results. The top panel illustrates an example of the estimated and true success rate for each edge, chosen arbitarily from the ten realizations. We see that the estimated success rates are in good agreement with the true TCP success rates. The bottom panel shows the mean absolute error for each edge over the 10 trials. In all three scenarios, we see that the worst-case mean absolute error is roughly 2%.

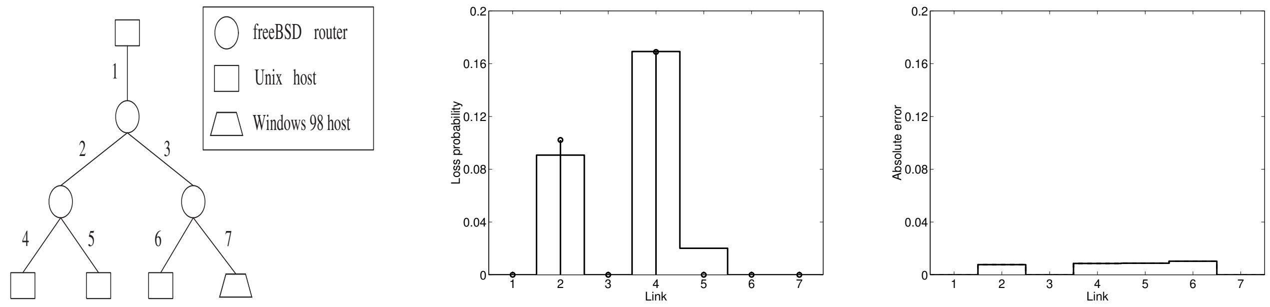
B. Testbed Experiment

We have constructed a testing framework on a testbed network of prototype freeBSD v3.2 routers, with topology as depicted in Figure 3. Maximum edge speeds are 10 Mb/s for routers and 100 Mb/s for hosts, as manually configured by multiport Ethernet cards. The buffer size of the routers is 250 packets. Competing traffic is generated at all hosts and routers according to a Pareto on/off model. We apply the lost estimation



(a) (b) (c)

Fig. 4. Simulation Results. True and estimated edge-level success rates of TCP flows from source to receivers for several traffic scenarios: (a) Heavy losses on edges 2 and 5, (b) Heavy losses on edges 2 and 8, and (c) Traffic mixture - medium losses. In each subfigure, the two panels display for each edge 1-11 (horizontal axis): (top) an example of true and estimated success rates and (bottom) mean absolute error between estimated and true success rates over 10 trials for each edge.



(a) (b) (c)

Fig. 5. Testbed experimental results. (a) Network testbed architecture. Heavy cross-traffic was generated on links 2 and 4. (b) The results of a typical experiment. Circles show true loss rates on links 1-7. Solid line shows estimated loss rates. (c) The absolute error for each link loss rate estimate, averaged over ten experiments.

algorithm at the source vertex, using measurements collected by sending packet-pair probes to the receivers. There was 75 ms spacing between probes, and the experiment was conducted for 2 minutes, generating 1600 measurements. UDP probing and measurement on individual edges was employed to obtain "true" unconditional loss probabilities. Our experiments show a general agreement between the tomographic estimates of the edge loss probabilities derived from packet pair measurements and the direct "true" estimates obtained by UDP probing along individual edges.

VII. Extensions and Conclusions

This paper developed a maximum likelihood estimation approach to the problem of unicast network tomography. We jointly estimate the unicast success probability and the conditional (back-to-back packet pair) success probability on each edge of the network graph. This allows us to account for imperfect conditional success probabilities (less than one) and can help to avoid underestimation of the unicast success rates. We devised a fast, EM algorithm for computing MLEs. We also examined the convergence behavior of the EM algorithm, proving that it converges to the set of global maxima. A theoretical analysis of the correlation between successes of closely spaced packet-pairs under an M/M/1/K queue model demonstrated that the correlations are generally quite strong (implying that the conditional success probabilities tend to be very close to one), corroborating experimental observations in real networks [18,23] and our `ns` and network testbed experiments. We assessed the performance of our unicast network tomography algorithm through `ns` simulation experiments and more realistic experiments in a network testbed comprised of eight freeBSD routers.

The ideas in this paper can be extended in several directions. First, because our approach is likelihood-based Bayesian estimation methods can easily be developed. We proposed a Bayesian approach that incorporates the prior assumption that the conditional success probabilities are greater than or equal to the unconditional success probabilities in earlier work [12]. Other types of prior information or regularization can be easily applied within our EM algorithm by simply modifying the M-Step. A second extension is to the maximum likelihood estimation of queuing delays. This can also be formulated as a maximum likelihood estimation problem and solved using an EM algorithm related to the one derived in this paper. We developed and investigated an MLE/EM approach to the unicast network delay tomography problem in other papers [13,14].

There are also several related issues that we are currently investigating. The network tomography problem studied in this paper and others mentioned above assume knowledge of the network topology. While this information may be readily available in many situations, in others it is not. Several researchers, including us, have investigated measurement-based techniques for estimating the network topology [3, 10, 15–17, 25]. We were the first to propose a MLE approach to the topology identification problem [10]. This perspective on the problem demonstrates the enormous computational challenge associated with topology identification (the only way we know to compute the exact MLE topology estimate is by testing every possible tree topology connecting the sender to the receivers). Finding compuational efficient, optimal or near-optimal methods for topology identification is an important open problem. Another key issue is the possibility of non-negligible temporal and spatial dependencies, which could arise due to long-range temporal dependencies in network traffic and the common cross-traffic flows. Assessing the impact of such dependencies on network tomography algorithms and developing new algorithms that mitigate their effects or even new measurement methods immune to such dependencies are important directions for future work.

Appendix

I. EM Algorithm Computational Complexity

In this appendix, we examine the computational complexity of the EM algorithm, leading to the results stated in Section III-A. The E-step involves almost all the computation in the algorithm; it involves the calculation of the conditional expectations of the unobserved data according to (9) and (10). Both of these equations involve the summation of terms, each term being a product of success/loss probabilities and a measurement count.

We begin by analysing the total number of multiplications necessary to construct all these terms. A term in the conditional expectation calculation is the product of a specific combination of the conditional and unconditional success probabilities associated with a certain subpath of the tree (one probability term being chosen from each edge in the subpath). Each subpath starts from the sender and ends at a vertex lower in the tree. The nature of the packet pair measurement process places a restriction on the way in which conditional and unconditional probabilities can be grouped within each term. Specifically, the sequence of edge probabilities, ordered in the traversal order from sender to end vertex, cannot involve an unconditional edge success probability followed by a conditional edge success probability; such a sequence would correspond to a measurement not possible in our framework. This restriction limits the number of unique product terms that can be formed and must be computed in the E-step calcultation. We can relate the number of distinct subpath probabilities to the number of edges (or vertices) in a tree. For each vertex i, there are $l(i)+1$ subpaths to that vertex, where $l(i)$ is the number of edges in the path $\mathcal{P}(0,i)$. If the tree has L edges in total, let us enumerate the vertices $0,\ldots,L$ and assume that there is a single edge emerging from the source. Then the number of subpaths is equal to $L+\sum_{i=1}^{L} l(i)$.

The number of terms that must be calculated is therefore equal to $L(\bar{l}+1)$, where $\bar{l}=\frac{1}{L}\sum_{i=0}^{L} l(i)$ is the average depth of the tree. If the tree is binary and complete, the average depth grows as $\log_2 L$. If the tree is complete and has constant fanout equal to f, then the average depth grows as $log_f L$. When the fanout is set to its maximum possible value, $L-1$ (see Figure 2(b)), the average depth grows as $\log L/\log(L-1)$. This approaches one as L grows large (clearly the average depth approaches 2 for a tree of the form in Figure 2(b)). The worst-case tree is depicted in Figure 2(a). In this case, the average depth grows proportionally to L. We can now state that the growth of number of unique terms in all the summations relative to the number of edges L lies between $O(L)$ and $O(L^2)$. Because of the repetition of the combinations of probabilistic weights in the each term, the formation of all terms requires less than two multiplications per term.

We now briefly consider the number of additions involved in the evaluation of the conditional expectations. The expressions (9) and (10) involve a summation of measurements, but this need only be performed once for the entire algorithm, so can be disregarded when determining the computational requirements per iteration. For each subpath terminating at vertex i, there are $l(i)+1$ terms to sum. Once these summations have been performed, the resulting L terms must be summed. The total number of summations is then $L-1+\sum_{i=1}^{L} l(i)$. The growth of the number of summations required as a function of the number of edges is therefore the same as the growth of the number of multiplications.

References

[1] M. Allman and V. Paxson. On estimating end-to-end network path properties. In *Proc. ACM SIGCOMM '99*, Cambridge, MA, Aug. 1999.

[2] J. O. Berger, B. Lisco, and R. L. Wolpert. Integrated likelihood methods for eliminating nuisance parameters. *Work Paper 97-01, Institute of Statistics & Decision Sciences*, Duke University:Durham, NC, 1997.

[3] A. Bestavros, J. Byers, and K. Harfoush. Inference and labeling of metric-induced network topologies. Technical Report BUCS-2001-010, Computer Science Department, Boston University, June 2001.

[4] J-C. Bolot. End-to-end packet delay and loss behaviour in the Internet. In *Proc. ACM SIGCOMM 1993*, pages 289–298, Sept. 1993.

[5] R. Cáceres, N. Duffield, J. Horowitz, and D. Towsley. Multicast-based inference of network-internal loss characteristics. *IEEE Trans. Info. Theory*, 45(7):2462–2480, November 1999.

[6] R. Cáceres, N. Duffield, J. Horowitz, D. Towsley, and T. Bu. Multicast-based inference of network-internal characteristics: Accuracy of packet loss estimation. In *Proceedings of IEEE INFOCOM 1999*, March 1999.

[7] J. Cao, D. Davis, S. Vander Wiel, and B. Yu. Time-varying network tomography: router link data. *J. Amer. Statist. Assoc*, 95:1063–1075, 2000.

[8] J. Cao, S. Vander Wiel, B. Yu, and Z. Zhu. A scalable method for estimating network traffic matrices from link counts. URL: http://www.stat.berkeley.edu/∼binyu/publications.html, 2000.

[9] R. Carter and M. Crovella. Measuring bottleneck link speed in packet-switched networks. Technical Report BU-CS-96-006, Computer Science Dept., Boston University, Mar. 1996.

[10] R. Castro, M. Coates, M. Gadhiok, R. King, R. Nowak, E. Rombokas, and Y. Tsang. Maximum likelihood network topology identification from edge-based unicast measurements. Technical Report TREE-0107, Rice University, Nov. 2001.

[11] M. Coates and R. Nowak. Network inference from passive unicast measurement. Technical Report TREE-0002, Rice University, Jan. 2000.

[12] M. Coates and R. Nowak. Network loss inference using unicast end-to-end measurement. In *ITC Seminar on IP Traffic, Measurement and Modelling*, Monterey, CA, Sep. 2000.

[13] 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.

[14] M. Coates, R. Nowak, and Y. Tsang. Nonparametric estimation of internal delay distributions from unicast end-to-end measurement. Technical Report TREE-0106, Rice University, Aug. 2001.

[15] N.G. Duffield, J. Horowitz, and F. Lo Presti. Adaptive multicast topology inference. In *Proceedings of IEEE INFOCOM 2001*, Anchorage, Alaska, April 2001.

[16] N.G. Duffield, J. Horowitz, F. Lo Presti, and D. Towsley. Multicast topology inference from end-to-end measurements. In *ITC Seminar on IP Traffic, Measurement and Modelling*, Monterey, CA, Sep. 2000.

[17] N.G. Duffield, J. Horowitz, F. Lo Presti, and D. Towsley. Multicast topology inference from measured end-to-end loss. to appear in *IEEE Trans. Information Theory*, 2002.

[18] 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.

[19] K. Harfoush, A. Bestavros, 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.

[20] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley, New York, 1997.

[21] Multicast-based inference of network-internal characteristics (MINC). http://gaia.cs.umass.edu/minc.

[22] UCB/LBNL/VINT network simulator ns (version 2). URL: http://www.isi.edu/nsnam/ns/.

[23] V. Paxson. End-to-end Internet packet dynamics. *IEEE/ACM Trans. Networking*, 7(3):277–292, June 1999.

[24] F. Lo Presti, N.G. Duffield, J. Horowitz, and D. Towsley. Multicast-based inference of network-internal delay distributions. Technical report, University of Massachusetts, 1999.

[25] 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.

[26] 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.

[27] M.F. Shih and A.O. Hero. Unicast inference of network link delay distributions from edge measurements. Technical report, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, May 2001.

[28] C. Tebaldi and M. West. Bayesian inference on network traffic using link count data (with discussion). *J. Amer. Stat. Assoc.*, pages 557–576, June 1998.

[29] Y. Tsang, M. Coates, and R. Nowak. Passive network tomography using EM algorithms. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, May 2001.

[30] R.J. Vanderbei and J. Iannone. An EM approach to OD matrix estimation. Technical Report SOR 94-04, Princeton University, 1994.

[31] Y. Vardi. Network tomography: estimating source-destination traffic intensities from link data. *J. Amer. Stat. Assoc.*, pages 365–377, 1996.

[32] C.F.J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[33] A.-G. Ziotopolous, A.O. Hero, and K. Wasserman. Estimation of network link loss rates via chaining in multicast trees. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, May 2001.