Rice University, ECE Department, Technical Report TR-0004
September 1, 2000

Unicast Network Tomography using EM Algorithms

Mark Coates and Robert Nowak *

Department of Electrical and Computer Engineering, Rice University

6100 South Main Street, Houston, TX 77005–1892

Email: {mcoates, nowak}@ece.rice.edu, Web: www.dsp.rice.edu

Abstract

The fundamental objective of this work is to determine the extent to which unicast, end-to-end network measurement is capable of determining internal network losses. The major contributions of this paper are two-fold: we formulate a measurement procedure for network loss inference based on end-to-end packet pair measurements, and we develop a statistical modeling and computation framework for inferring internal network loss characteristics. Simulation experiments demonstrate the potential of our new framework.

1. Introduction

In large-scale networks, end-systems cannot rely on the network itself to cooperate in characterizing its own behavior. This has prompted several groups to investigate methods for inferring internal network behavior based on end-to-end network measurements [1, 2, 3, 4, 5, 6]; the so-called *network tomography* problem. While promising, these methods require special support from the network in terms of either cooperation between hosts, internal network measurements, or multicast capability. Many networks do not currently support multicast due to its scalability limitations (routers need to maintain per group state), and lack of access control. Moreover, multicast-based methods may not provide an accurate characterization of the loss rates for the traffic of interest, because routers treat multicast packets differently than unicast packets.

*This work was supported by the National Science Foundation, grant no. MIP–9701692, the Army Research Office, grant no. DAAD19-99-1-0349, the Office of Naval Research, grant no. N00014-00-1-0390, and Texas Instruments.

In this paper, we introduce a new methodology for network tomography (specifically, inferring packet loss probabilities on internal network links) based on unicast measurement. In contrast to multicast techniques, unicast inference is easily carried out on most networks and is scalable. Our approach employs unicast, end-to-end measurement of single packet and back-to-back packet pair losses, which can be performed actively or passively. By back-to-back packet pairs we mean two packets that are sent one after the other by the source, possibly destined for different receivers, but sharing a common set of links in their paths. Throughout the remainder of the paper we work with "success" probabilities (probability of non-loss) instead of loss probabilities. This provides a more convenient mathematical parameterization of the problem, and the probability of loss is simply one minus the probability of success.

The use of back-to-back packet pair measurements is motivated by the following reasoning. If two back-to-back packets are sent across a link and the first packet is received, then it is highly likely that the second packet will also be received. We expect that the conditional success probability of the second packet (given that the first is received) may often be close to one. This observation has been verified experimentally in real networks [7] and can also be established theoretically under an M/M/1/K queue model [8]. Exploiting this correlation between back-to-back packet losses, we develop a framework for the statistical estimation of internal success probabilities based solely on unicast, end-to-end measurement. In our simulated experiments, we are able to obtain accurate loss estimates even in cases where the conditional success probabilities are significantly less than one (*e.g.,* conditional success probabilities of 0.9, which are lower than typical measurements on the Internet).

The inherent structure of networks makes this problem ideally suited to the new field of factor graph analysis. Factor graphs enable us both to visualize the relationships between statistics and network parameters and to greatly simplify the tomography problem through both probability factorization and message passing algorithms [9]. These graphical models enable very efficient and scalable estimation algorithms. In fact, the complexity of our algorithms grows linearly with the number of nodes in the network under study. A key strength of our methodology is that it can deliver not only point estimates and confidence intervals, but also probability distributions for network parameters of interest. This provides the complete characterization of the accuracy and reliability of inferred network behavior that is necessary for modeling, maintenance, and service provisioning.

The paper is organized as follows. In Section 2, we introduce the basic unicast tomography problem and the technical issues involved. In Sections 3 and 4, we formally define our loss modeling assumptions and measurement framework. Section 5 describes several basic statistical inference tasks involved in unicast tomography. In Section 6, we propose two novel inference algorithms, both of which are based on the notion of "unobserved data" and likelihood factorization. Section 7 investigates the consistency and bias of our inference algorithms. In Section 8, we examine the performance of our methods through simulation, and concluding remarks are made in Section 9.

2. Unicast Tomography

We consider a scenario in which a single source sends packets to a number of receivers (extensions to multiple sources are possible). In this case, the network topology (from the perspective of the source) is a tree-structure. Figure 2. depicts an example topology with source (node 0) and eight receivers (nodes 6 through 13). Also shown are five internal routers (nodes 1 through 5). We assume that we are able to measure network traffic only at the edge; that is, we can determine whether or not a packet sent from the source is successfully received by one of the receivers. This type of confirmation can be obtained via TCP's acknowledgment system, for example. We also assume that the routing table is fixed for the duration of the measurement process, which ensures the tree-structured topology.

The goal of this work is to estimate the loss probabilities associated with each individual link (between two routers) in the network. Here, we use the term path or subpath to refer to a connection through two or more routers and link

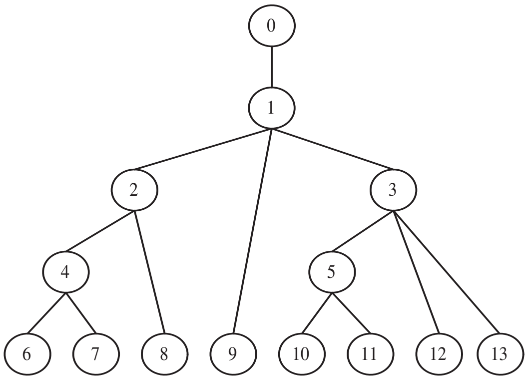

Figure 1 – Tree-structured graph representing a single-source, multiple-receiver network.

to refer to a single, direct connection between two routers. Restricting ourselves to edge-based measurement, we can measure the numbers of packets sent to and received by each receiver, providing us with a simple means of estimating the probabilities of success along each path (from source to receiver). Unfortunately, there is no unique mapping of the path success probabilities to the success probabilities on individual links (between routers) in the path. To overcome this difficulty, we propose a methodology based on measurements made using back-to-back packet pairs. These measurements provide an opportunity to collect more informative statistics that can help to resolve the links.

The basic idea employed here is quite straightforward. Suppose that we send two, closely time-spaced (back-to-back) packets from the source, with the first packet destined for receiver i and the second for receiver j. The paths traversed by the packets share some common subpath and then diverge at some point. For example, referring to Figure 2., suppose the first packet is destined for node 6 and the second for node 7. Then the two packets share a common subpath up to node 4. Now, if the first packet is received at node 6, then it is highly likely that both packets were received at node 4 (since they were closely time-spaced). Thus, if the second packet is not received at node 7, then we can deduce that it was probably dropped on the link from node 4 to 7. Repeating this packet-pair measurement numerous times and recording the number of drops of the second packet (when the first packet is received), we can isolate the loss rate on the 4-7 link.

Collecting measurements from an assortment of such back-to-back packet pairs (sent to different combinations of receivers) allows us to resolve the losses occurring on all links in the network. The key to this approach is the exploitation of the correlation between packet-pair losses on common subpaths.

In this paper, we examine several issues which in the following sections including: developing scalable estimation algorithms that are applicable to large networks; the sensitivity of the estimation procedure to cases in which the correlation between packet-pair losses on common subpaths is imperfect; and characterization of achievable estimation accuracy from limited numbers of packet measurements.

3. Loss Modeling

Here we describe our measurement method and statistical model in detail. Consider the tree-structured network associated with a single source and multiple receivers (*e.g.,* Figure 2.). A distinct path (from the source) is associated with each receiver. Each path is comprised of one or more links between routers (nodes). If isolated subpaths (subpaths consisting of two or more links with no branches) exist in the network under consideration, then these are removed and replaced by a single composite link to represent the isolated subpath. No isolated subpaths exist in the network shown in Figure 2., but if, for example, additional routers were added between nodes 1 and 2, then we would simply model this chain of links as one composite link resulting in the same tree.

For individual packet transmissions, we assume a simple Bernoulli loss model for each link. The *unconditional* success probability of link i (the link into node i) is defined as

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

where $\rho(i)$ denotes the index of the parent node of node i (the node above i-th node in the tree; *e.g.,* referring to Figure 2., $\rho(1) = 0$). A packet is successfully sent from $\rho(i)$ to i with probability α_i and is dropped with probability $1 - \alpha_i$.

We model the loss processes on separate links as mutually independent. Although spatial dependence (correlated success probabilities on neighbouring links) may be observed in networks due to common traffic, such dependence is highly circumstantial and cannot be readily incorporated in a model that is intended to be generally applicable to a variety of networks. Bolot *et al.* proposed Markovian models of packet loss in [10] based on observations of Internet traffic. Although such models do not fully account for the extended loss bursts observed in [7], we adopt a similar approach for modeling the packet loss processes on each link (the model is reminiscent of that used to explore temporal dependence in [1]).

If two, back-to-back packets are sent from node $\rho(i)$ to node i, then we define the conditional success probability as

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

where $\rho(i) \to i$ is shorthand notation denoting the successful transmission of a packet from $\rho(i)$ to i. That is, given that the first packet of the pair is received, then the second packet is received with probability β_i and dropped with probability $1 - \beta_i$. We anticipate that $\beta_i \geq \alpha_i$ for each i, since knowledge that the first packet was successfully received suggests that the queue for link i is not full. Evidence for such behavior has been provided by observations of the Internet [11, 7]. In fact, it is not unreasonable to suppose that $\beta_i \approx 1$ in many cases, as demonstrated in the next section.

4. Queueing Analysis

We now give theoretical conditions under which $\beta_i \geq \alpha_i$. Consider a modified version of the classical M/M/1/K queue model in which the arrivals are generated from an inhomogeneous Poisson process (instead of a homogeneous process). The simplest case is one where the arrivals obey a two-component mixture of homogeneous Poisson processes. That is, with probability q_0 the arrivals are drawn from a homogeneous Poisson process with rate λ_0 and with probability $q_1 = 1 - q_0$ they are drawn from a homogeneous Poisson process with rate $\lambda_1 > \lambda_0$ (*e.g.,* the so-called Markov-modulated Poisson process [12]). A binary hidden state variable s governs this selection. We assume that the state is a slowly varying process, switching between $s = 0$ and $s = 1$ at a rate $\gamma \ll \lambda_0$. This assumption means that the process is quasi-stationary, in the sense that it obeys a homogeneous Poisson process model over time intervals of significant duration. Let μ denote the (Poisson) service rate of the queue. If we take $\lambda_0 < \mu < \lambda_1$, then the two Poisson processes could be viewed as "light" traffic and "heavy" traffic models, respectively. Being in state $s = 1$ could represent a traffic "burst," for example.

Now, suppose that a pair of closely time-spaced packets arrives at the queue. Let m_0 denote the number of packets in the queue just before the first packet in the pair arrives and assume that the traffic state is $s = i$. Let $p_i(j)$, $0 \leq j \leq K$, K being the length of the queue, denote the stationary queue distribution in state i. The event that the first packet makes it into the queue is $\{m_0 \neq K\}$. This event occurs with probability $1 - p_i(K)$. If this event has occurred, then *immediately* after the first packet (before any other arrival or service event) the *conditional* distribution of the queue (probability of j packets in the queue) is given by

$$\widetilde{p}_i(j) = \begin{cases} \frac{p_i(j-1)}{1-p_i(K)} & 1 \le j \le K, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Note that there is *at least* one packet in the queue at this time.

Now we remove the conditioning on the state s. The state's true value is unknown, however we make two assumptions:

1. *the queue distribution is in a steady-state condition, p_0 or p_1, just before the first packet arrives.*

2. *the traffic remains in state $s = 0$ or state $s = 1$ over the time interval between the arrivals of the first and second packets*

Roughly speaking, these assumptions simply mean that the traffic is in one state or the other for a sufficient period of time prior to and after the arrival of the first packet. This is reasonable under the condition that rate at which the state changes is much less than the rate of the traffic in either state (*i.e.*, $\gamma \ll \lambda_0$), since then with high probability the traffic will be in one state or the other.[1] Let m_0 denote the number of packets in the queue immediately before the arrival of the first packet. The probability that the second packet *does not* make it into the queue, given that the first packet did, is

$$\widetilde{p}(K) \equiv \Pr(s=0|m_0<K)\widetilde{p}_0(K) + \Pr(s=1|m_0<K)\widetilde{p}_1(K).$$

The probability

$$\begin{aligned} \Pr(s=0|m_0<K) &= \frac{\Pr(s=0, m_0<K)}{\Pr(m_0<K)}, \\ &= \frac{q_0\,[1-p_0(K)]}{q_0\,[1-p_0(K)] + q_1\,[1-p_1(K)]} \end{aligned}$$

Thus, we have

$$\begin{aligned} \widetilde{p}(K) &= \\ &\frac{q_0\,[1-p_0(K)]\,\frac{p_0(K-1)}{1-p_0(K)} \;+\; q_1\,[1-p_1(K)]\,\frac{p_1(K-1)}{1-p_1(K)}}{q_0\,[1-p_0(K)] + q_1\,[1-p_1(K)]} \\ &= \frac{q_0 p_0(K-1) \;+\; q_1 p_1(K-1)}{q_0\,[1-p_0(K)] + q_1\,[1-p_1(K)]} \end{aligned} \tag{2}$$

[1] To be more rigorous, all subsequent statements should be qualified as "with high probability."

Theorem 1 : *Let*

$$\begin{aligned} \alpha &\equiv 1 - p(K), \\ \beta &\equiv 1 - \widetilde{p}(K). \end{aligned}$$

If $q_0 = 0$ or $q_0 = 1$, then the traffic obeys a homogeneous Poisson process of rate λ_i, $i = 0$ or $i = 1$, and we have

$$\alpha > \beta.$$

Proof: Standard queuing theory [13] tells us that $p_i(j)$ (the probability of j packets in the queue under state i) is given by

$$p_i(j) = \begin{cases} \left(\frac{1-r_i}{1-r_i^{K+1}}\right) r_i^j & 0 \le j \le K, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $r_i \equiv \frac{\lambda_i}{\mu}$. Also, observe $1 - p_i(K) = \frac{1-r_i^K}{1-r_i^{K+1}}$ so that

$$\widetilde{p}_i(j) = \left[\frac{1-r_i}{1-r_i^K}\right] r_i^{j-1}, \quad 1 \le j \le K. \tag{4}$$

Thus,

$$\frac{p_i(K)}{\widetilde{p}_i(K)} = r_i \left(\frac{1-r_i^K}{1-r_i^{K+1}}\right) < 1,$$

for both $r_i = \frac{\lambda_i}{\mu} < 1$ and $r_i > 1$. In the case where $\lambda_i = \mu$ (equivalently $r_i = 1$), the stationary queue distribution is $p_i(j) = 1/(K+1)$, $j = 0, \ldots, K$. Then

$$\widetilde{p}_i(K) = \frac{p_i(K-1)}{1-p_i(K)} = \frac{1}{K} > \frac{1}{K+1} = p_i(K).$$

□

This shows that the conditional probability of the second packet making it into the queue, conditional on the first packet making it in, is less than the unconditional probability that the first packet makes it into the queue (at least for exactly back-to-back pairs). This phenomenon was first pointed out the authors by Don Towsley.

Theorem 2 : *If $0 < q_0 < 1$, then the traffic is inhomogeneous and*

$$\begin{aligned} p(K) &= \\ &q_0 \left(\frac{1-r_0}{1-r_0^{K+1}}\right) r_0^K \;+\; q_1 \left(\frac{1-r_1}{1-r_1^{K+1}}\right) r_1^K, \end{aligned} \tag{5}$$

$$\begin{aligned} \widetilde{p}(K) &= \\ &\frac{q_0 \left(\frac{1-r_0}{1-r_0^{K+1}}\right) r_0^{K-1} \;+\; q_1 \left(\frac{1-r_1}{1-r_1^{K+1}}\right) r_1^{K-1}}{q_0 \left(\frac{1-r_0^K}{1-r_0^{K+1}}\right) \;+\; q_1 \left(\frac{1-r_1^K}{1-r_1^{K+1}}\right)}. \end{aligned} \tag{6}$$

Proof: Expression (5) is obtained by noting that $p(K) = q_0 p_0(K) + q_1 p_1(K)$ and inserting (3). Substituting expression (3) into (2) gives expression (6). $\square$

Expressions (5) and (6) can easily be evaluated. In many cases, unlike the homogeneous scenario in Theorem 1, we find that $\widetilde{p}(K) < p(K)$ (in other words, $\alpha < \beta$). A simple expression is obtained when $r_0 < 1$ and $r_1 > 1$. As $K \to \infty$ we have,

$$\frac{\widetilde{p}(K)}{p(K)} \quad \to \quad \frac{1}{q_0 r_1 + q_1},$$

showing that $\widetilde{p}(K) < p(K)$ if $q_0 r_1 + q_1 > 1$. For example, if $K = 100$, $q_0 = 0.9$, $r_0 = 0.5$, and $r_1 = 10$ (traffic with infrequent, but heavy bursts), then $\alpha = 0.91$ and $\beta = 0.99$.

Now let us consider the situation when multiple events intervene between the arrival of the first and second packets in the pair. We still assume that the traffic is either in state $s = 0$ or state $s = 1$ when both packets arrive. Define the $(K+1) \times (K+1)$ transition matrices

$$\mathbf{T}_i \equiv \begin{bmatrix} p_i^s & p_i^s & 0 & 0 & 0 & \cdots \\ p_i^a & 0 & p_i^s & 0 & 0 & \cdots \\ 0 & p_i^a & 0 & p_i^s & 0 & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & p_i^a & 0 & p_i^s \\ 0 & \cdots & 0 & 0 & p_i^a & p_i^a \end{bmatrix}, \quad i = 0, 1,$$

where $p_i^a \equiv \frac{\lambda_i}{\lambda_i + \mu}$ and $p_i^s \equiv \frac{\mu}{\lambda_i + \mu} = 1 - p_i^a$. Also define the column vectors $\mathbf{p}_i$ and $\widetilde{\mathbf{p}}_i$ as

$$\mathbf{p}_i \equiv \begin{bmatrix} p_i(0) \\ p_i(1) \\ \vdots \\ p_i(K) \end{bmatrix}, \qquad \widetilde{\mathbf{p}}_i \equiv \begin{bmatrix} \widetilde{p}_i(0) \\ \widetilde{p}_i(1) \\ \vdots \\ \widetilde{p}_i(K) \end{bmatrix}, \quad i = 0, 1.$$

Expressions (3) and (4) show that $\widetilde{p}_i(j) = C_i\, p_i(j)$, $1 \le j \le K$, where $C_i \equiv r_i^{-1} \left(\frac{1 - r_i^{K+1}}{1 - r_i^K} \right)$. Thus,

$$\widetilde{\mathbf{p}}_i = C_i \left(\mathbf{p}_i - \boldsymbol{\delta}_i \right),$$

where

$$\boldsymbol{\delta}_i = \begin{bmatrix} p_i(0) \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Suppose there are n intervening events (any combination of arrivals and services) between the arrivals of the two packets. Then, assuming the traffic is in state $s = i$, the distribution of the queue when the second packet arrives is given by

$$\begin{aligned} \widetilde{\mathbf{p}}_i^{(n)} &= \mathbf{T}_i^n \, \widetilde{\mathbf{p}}_i \\ &= C_i \left(\mathbf{p}_i - \mathbf{T}_i^n \, \boldsymbol{\delta}_i \right), \end{aligned}$$

where we use the fact that $\mathbf{p}_i$ is the stationary distribution, implying that $\mathbf{T}_i^n \, \mathbf{p}_i = \mathbf{p}_i$. Observe two key facts. First, for every n the elements of $\mathbf{T}_i^n \, \boldsymbol{\delta}_i$ are non-negative. Therefore, the $K+1$-th element of $\widetilde{\mathbf{p}}_i^{(n)}$, $\widetilde{p}_i^{(n)}(K) \le \widetilde{p}_i(K)$. Second, removing the dependence on the state, we define

$$\widetilde{\mathbf{p}}^{(n)} \equiv \Pr(s = 0 | m_0 < K) \widetilde{\mathbf{p}}_0^{(n)} + \Pr(s = 1 | m_0 < K) \widetilde{\mathbf{p}}_1^{(n)},$$

and combining this with the first fact we have

$$\widetilde{\mathbf{p}}^{(n)} \le \Pr(s = 0 | m_0 < K) \widetilde{\mathbf{p}}_0 + \Pr(s = 1 | m_0 < K) \widetilde{\mathbf{p}}_1.$$

In particular,

$$\begin{aligned} \widetilde{p}^{(n)}(K) &\le \\ & \Pr(s = 0 | m_0 < K) \widetilde{p}_0(K) + \Pr(s = 1 | m_0 < K) \widetilde{p}_1(K) \\ &= \frac{q_0 \left(\frac{1 - r_0}{1 - r_0^{K+1}} \right) r_0^{K-1} + q_1 \left(\frac{1 - r_1}{1 - r_1^{K+1}} \right) r_1^{K-1}}{q_0 \left(\frac{1 - r_0^K}{1 - r_0^{K+1}} \right) + q_1 \left(\frac{1 - r_1^K}{1 - r_1^{K+1}} \right)}. \end{aligned}$$

Thus, the probability that the queue is full when the second packet arrives is always less than or equal to that probability computed under the assumption of no intervening events. We summarize our conclusions with the following theorem.

Theorem 3 : *Let*

$$\begin{aligned} \alpha &\equiv 1 - p(K) \\ \beta^{(n)} &\equiv 1 - \widetilde{p}^{(n)}(K). \end{aligned}$$

and assume that $0 < q_0 < 1$. *Then* $p(K) > \widetilde{p}(K)$ *is a sufficient condition for*

$$\alpha < \beta^{(n)},$$

for every $n \ge 0$, *where* $p(K)$ *and* $\widetilde{p}(K)$ *are given by expressions (5) and (6), respectively.*

Corollary 1 : *If* $q_0 = 0$ *or* $q_0 = 1$, *then the traffic is homogeneous and for all* $n \ge 0$

$$\alpha > \beta^{(n)}.$$

5. Measurement Framework

Each link in the tree has two (unknown) probabilities associated with it, the unconditional and conditional success probabilities, α_i and β_i, respectively. These probabilities effect 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 can be used for active probing or TCP connections may be passively monitored, in which case back-to-back events are selected from the TCP traffic flows.

Single Packet Measurement: Suppose that n_i packets are sent to receiver i and that of these a number m_i are actually received ($n_i - m_i$ are dropped). The likelihood of m_i given n_i is binomial (since Bernoulli losses are assumed) and is given by

$$l(m_i \,|\, n_i, p_i) \quad = \quad \binom{n_i}{m_i} p_i^{m_i} \, (1-p_i)^{n_i - m_i},$$

where $p_i = \prod_{j \in \mathcal{P}(0,i)} \alpha_j$ and $\mathcal{P}(0,i)$ denotes the sequence of nodes in the path from the source 0 to receiver i. For example, in Figure 2., $\mathcal{P}(0,10) = \{1,3,5,10\}$ and so $\prod_{j \in \mathcal{P}(0,10)} \alpha_j = \alpha_1\alpha_3\alpha_5\alpha_{10}$.

Back-to-Back Packet Pair Measurement: Suppose that the source sends a large number of back-to-back packet pairs in which the first packet is destined for receiver i and the second for receiver j. We assume that the timing between pairs of packets is considerably larger than the timing between two packets in each pair. Let $n_{i,j}$ denote the number of pairs for which the first packet is successfully received at node i, and let $m_{i,j}$ denote the number of pairs for which both the first and second packets are received at their destinations. Furthermore, let $k_{i,j}$ denote the node at which the paths $\mathcal{P}(0,i)$ and $\mathcal{P}(0,j)$ diverge, so that $\mathcal{P}(0,k_{i,j})$ is their common subpath. For illustration, refer to Figure 2. and let $i = 6$ and $j = 8$, then $k_{6,8} = 2$. With this notation, the likelihood of $m_{i,j}$ given $n_{i,j}$ is binomial and is given by

$$l(m_{i,j} \,|\, n_{i,j}, p_{i,j}) \quad = \quad \binom{n_{i,j}}{m_{i,j}} p_{i,j}^{m_{i,j}} \, (1-p_{i,j})^{n_{i,j} - m_{i,j}},$$

where

$$p_{i,j} = \prod_{q \in \mathcal{P}(0,k_{i,j})} \beta_q \prod_{r \in \mathcal{P}(k_{i,j},j)} \alpha_r .$$

6. Inference Tasks

Assume that we have made an assortment of single packet and back-to-back packet measurements (sent to different receivers or combinations of receivers) as described in Section 4. Collecting all the measurements, define

$$\begin{aligned} \mathcal{M} &\equiv \{m_i\} \cup \{m_{i,j}\} \\ \mathcal{N} &\equiv \{n_i\} \cup \{n_{i,j}\}, \end{aligned}$$

where the index i alone runs over all receivers and the indices i,j run over all pairwise combinations of receivers in the network.

Let us also denote the collections of the unconditional and conditional link success probabilities as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The *joint* likelihood of all measurements is given by

$$l(\mathcal{M} \,|\, \mathcal{N}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad = \quad \prod_i l(m_i \,|\, n_i, p_i) \; \times \prod_{i,j} l(m_{i,j} \,|\, n_{i,j}, p_{i,j}).$$

Since $\mathcal{M}$ and $\mathcal{N}$ are known, we view $l(\mathcal{M} \,|\, \mathcal{N}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ as a function of the unknown probabilities $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We call $l(\mathcal{M} \,|\, \mathcal{N}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ the likelihood function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Based on the likelihood function, we wish to make inferences about the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Several options exist.

Maximum Likelihood Estimation: Maximum likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are defined as

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \quad = \quad \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \; l(\mathcal{M} \,|\, \mathcal{N}, \boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Maximum likelihood estimation enjoys many desirable properties and is widely utilized in statistical inference [14].

Maximum Integrated Likelihood Estimation: The conditional success probabilities $\boldsymbol{\beta}$ may not be of interest in many applications. In such cases, $\boldsymbol{\beta}$ are called *nuisance* parameters, and it is common to integrate the likelihood over the nuisance parameters first, then maximize the result with respect to the parameters of interest (in this case $\boldsymbol{\alpha}$). The integrated maximum likelihood estimates of $\boldsymbol{\alpha}$ are defined as

$$\widehat{\boldsymbol{\alpha}} \quad = \quad \arg\max_{\boldsymbol{\alpha}} \int l(\mathcal{M} \,|\, \mathcal{N}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \, d\boldsymbol{\beta},$$

where each conditional success probability β_i is integrated from 0 to 1. Integrated likelihood methods "automatically incorporate nuisance parameter uncertainty" [14]. As a consequence, the integrated likelihood function may provide more accurate estimates of the unconditional success probabilities than those provided by the joint likelihood function.

Marginal Likelihood Analysis: In addition to determining the success probabilities that maximize the likelihood function, it may be of interest to examine the *marginal* likelihood function of each individual probability. The marginal likelihood function of α_i is defined as

$$l(\mathcal{M} \,|\, \mathcal{N}, \alpha_i) \;\; = \;\; \int l(\mathcal{M} \,|\, \mathcal{N}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \; d\boldsymbol{\alpha}_{\bar{i}} \;\; d\boldsymbol{\beta},$$

where $\boldsymbol{\alpha}_{\bar{i}}$ is the collection of all unconditional success probabilities except α_i, and all probabilities are integrated over the interval $[0, 1]$. Similarly, the marginal likelihood function of β_i is

$$l(\mathcal{M} \,|\, \mathcal{N}, \beta_i) \;\; = \;\; \int l(\mathcal{M} \,|\, \mathcal{N}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \; d\boldsymbol{\alpha} \;\; d\boldsymbol{\beta}_{\bar{i}}.$$

The marginal likelihood functions are univariate functions of the remaining parameter. The marginals can be maximized to obtain an estimate of the parameter, or the functions can be inspected for additional information. If the marginal has a single mode (peak), then the width or spread of the likelihood function can be used to determine confidence intervals for the maximum marginal likelihood estimate. More generally, the marginal may have multiple modes (a feature completely lost when focusing only on the maximum), which may provide useful alternative explanations for the measured data.

7. Inference Algorithms

Computing maximum likelihood estimates or marginal likelihood functions can be a formidable task. Multidimensional maximizations or integrations are time-consuming and directly attempting any of the inference tasks outlined in Section 5 leads to extremely computationally demanding algorithms that are not scalable to large networks.

The basic problem is that the individual likelihood functions $l(m_i \,|\, n_i, p_i)$ or $l(m_{i,j} \,|\, n_{i,j}, p_{i,j})$ for each type of measurement involve products of the $\boldsymbol{\beta}$ and/or $\boldsymbol{\alpha}$ probabilities. Consequently, it is difficult to separate the effects of each individual success probability.

We overcome this difficulty using a common device in computational statistics known as *unobserved* data or variables. To introduce the notion of unobserved data, let us consider the likelihood

$$l(m_i \,|\, n_i, p_i) \;\; = \;\; \binom{n_i}{m_i} \, p_i^{m_i} \, (1 - p_i)^{n_i - m_i},$$

where $p_i = \prod_{j \in \mathcal{P}(0,i)} \alpha_j$. Assuming that the path consists of more than one link, the effects of the individual link success probabilities on this measurement are combined through the product over the entire path. However, suppose it were possible to measure the numbers of packets making it to each node. Let us denote these unobserved measurements by $u_{j,i}$, $j \in \mathcal{P}(0, i)$, $j \neq i$. With these measurements in hand, we can write the likelihood function of the observed *and* unobserved data as

$$l\left(\{u_{j,i}\} \,|\, n_i, p_i\right) \;\; = \;\; \prod_{j \in \mathcal{P}(0,i)} \binom{u_{\rho(j),i}}{u_{j,i}} \, \alpha_j^{u_{j,i}} \, (1 - \alpha_j)^{u_{\rho(j),i} - u_{j,i}},$$

where $\rho(j)$ again denotes the parent of node j. Also, since we are able to measure at the source and receiver, in the expression above we set $u_{0,i} = n_i$ and $u_{i,i} = m_i$. The example in Figure 7. illustrates the notion of unobserved data.

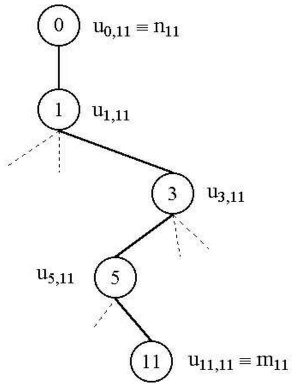

Figure 2 – Path from source to receiver $i = 11$ with unobserved data at each internal router.

Because the likelihood $l\left(\{u_{j,i}\} \,|\, n_i, p_i\right)$ depends on both the observed data and the unobserved, it is called the *complete data* likelihood. The key feature of the complete data likelihood function is that it factorizes into a product of individual binomial likelihood functions, each involving just a single success probability. Thus, the complete data likelihood function is a trivial multivariate function, and the effects of the individual link probabilities are easily separated.

In a similar fashion, we introduce unobserved data for all measured paths, and these variables allow us to factorize the joint likelihood function into a product of univariate functions. Several well-known optimization strategies take advantage of this simplification.

The Expectation-Maximization Algorithm: As the name suggests, the Expectation-Maximization (EM) Algorithm alternates between two steps; one step estimates the unobserved data and the other maximizes the complete data likelihood [15]. The EM Algorithm can be used for our problem to compute maximum likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Beginning with an initial guess for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, 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 $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The E Step can be computed in $O(N)$ operations, where N is the total number of receivers, using an upward-downward probability propagation (or message passing) algorithm [9]. The Maximization (M) Step combines the observed and expected unobserved data to form the complete data likelihood function which is then maximized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Since the complete data likelihood factorizes into a product of univariate functions, each involving just one success probability, the maximizers have closed-form, analytic expressions. Thus, the M Step can also be computed in $O(N)$ operations. Each iteration of the EM Algorithm is therefore $O(N)$ in complexity. Moreover, it can be shown that the original (observed data only) likelihood function is monotonically increased at each iteration of the algorithm, and the algorithm converges to a local maximum of the likelihood function [15]. Our experiments have shown that the algorithm typically converges in a small number of iterations.

Factor Graphs and Marginal Analysis: It may be of interest to compute maximum integrated likelihood estimates or to inspect marginal likelihood functions, as mentioned in Section 5. The EM Algorithm only delivers maximum likelihood estimates. However, using the notion of unobserved data in conjunction with probability propagation similar to that employed in the E Step above, computationally efficient algorithms do exist for computing maximum integrated likelihood estimates and marginal likelihood functions.

These algorithms are based on graphical representations of statistical models. Such representations include Bayesian networks and, more generally, factor graphs [9]. Both the parameters of interest and collected data appear as nodes in the factor graph. Each node associated with a parameter is characterised by a (potentially unknown) probability distribution. Links between the nodes indicate probabilistic dependencies. By introducing unobserved variables as additional nodes, it is possible to decouple the effects of different success probabilities in the graphical model.

Probability propagation can be used to perform exact inference, provided the graph structure is acyclic. However, this may require high-dimensional summations, leading to a heavy computational burden; thus, exact inference algorithms can scale poorly as the network size increases. To avoid the associated computational burden, we have developed an approximation to exact inference. In the approximate strategy, we first infer likelihood functions of the loss parameters at the receivers. We then use these functions to perform inference at the next level of the tree, and continue upwards to the source. Details of the algorithm appear in [8].

8. Consistency and Bias

If the conditional success probabilities $\boldsymbol{\beta}$ are all exactly one, then it can be shown that maximum likelihood estimates of the unconditional losses $\boldsymbol{\alpha}$ 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 j. The single packet measurements m_j and n_j provide an asymptotically consistent estimator of the product $p_j = \prod_{i\in\mathcal{P}(0,j)} \alpha_i$. Specifically, $\widehat{p}_j \equiv \frac{m_j}{n_j}$ converges to p_j as n_j tends to infinity. Similarly, the estimators $\widehat{p}_{i,j} \equiv \frac{m_{i,j}}{n_{i,j}}$, converge to

$$p_{i,j} = \prod_{q\in\mathcal{P}(0,k_{i,j})} \beta_q \prod_{r\in\mathcal{P}(k_{i,j},j)} \alpha_r,$$

as each $n_{i,j} \to \infty$ (recall that the node $k_{i,j}$ defines the subpath common to both receivers).

To simplify the notation, let us assume that there are L links in the path and denote them by $\mathcal{P}(0,j) = \{j_1, j_2, \ldots, j_L\}$, where $j_L \equiv j$. Define $i_1, \ldots, i_L$ so that the common subpath between $\mathcal{P}(0,i_\ell)$ and $\mathcal{P}(0,j)$ is $\mathcal{P}(0,j_\ell)$, $\ell = 1,\ldots,L$ (note that $i_L \equiv j_L = j$). Then we have

$$\begin{aligned}
\widehat{p}_j &\to \alpha_{j_1}\alpha_{j_2}\alpha_{j_3}\cdots\alpha_{j_L}, \\
\widehat{p}_{i_1,j} &\to \beta_{j_1}\alpha_{j_2}\alpha_{j_3}\cdots\alpha_{j_L}, \\
\widehat{p}_{i_2,j} &\to \beta_{j_1}\beta_{j_2}\alpha_{j_3}\cdots\alpha_{j_L}, \\
&\vdots \\
\widehat{p}_{i_L,j_L} &\to \beta_{j_1}\beta_{j_2}\beta_{j_3}\cdots\beta_{j_L}.
\end{aligned}$$

Note that if $\widehat{p}_{i_L,j_L} \to 1$, then we may deduce that $\beta_{j_\ell} = 1$, $\ell = 1,\ldots,j$. In this case, $\widehat{p}_{i_{L-1},j_L} \to \alpha_{j_L}$ and $\widehat{p}_{i_{L-\ell},j_L} \to \alpha_{j_{L-\ell+1}} \cdots \alpha_{j_L}$, for $\ell = 2,\ldots,L$. Consistent estimators of $\boldsymbol{\alpha}$ can be computed according to

$$\begin{aligned}
\widehat{\alpha}_{j_L} &\equiv \widehat{p}_{j_L,j_L}, \\
\widehat{\alpha}_{j_{L-\ell}} &\equiv \frac{\widehat{p}_{i_{L-\ell},j_L}}{\widehat{p}_{i_{L-\ell+1},j_L}}, \quad \ell = 1,\ldots,L-1.
\end{aligned} \tag{7}$$

If one or more of the β are less than one, then a systematic bias is introduced into the estimation process and the maximum likelihood estimators are not consistent. However, the severity of the bias is directly linked to the extent to which the β deviate from one; the less the deviation, the less the bias. Suppose that $\widehat{p}_{j,j} \to \gamma < 1$. Then we can deduce that

$$\gamma \le \prod_{k=1}^{\ell} \beta_{j_k} \le 1$$

for $\ell = 1 \ldots, L$. This shows that the asymptotic value of $\widehat{p}_{i_{L-i}, j_L}$ lies within the interval

$$\left[\gamma \prod_{k=L-\ell+1}^{L} \alpha_k \; , \; \prod_{k=L-\ell+1}^{L} \alpha_k \right],$$

for $\ell = 1, \ldots, L-1$. From here it follows that the the asymptotic values of the estimators $\{\widehat{\alpha}_k\}$ defined in (7) lie within the intervals

$$\left[\gamma \, \alpha_k \; , \; \frac{1}{\gamma} \, \alpha_k \right].$$

Thus, we see that the value of $\gamma = \prod_{k=1}^{L} \beta_{j_k}$ controls the asymptotic accuracy of the maximum likelihood estimators.

9. Simulation Experiments

9.1. A Simple Example

Let us now consider the simple two-receiver network shown in Figure 9.1.. Assume that we have made measurements of single packet and back-to-back packet:

$$\begin{aligned} \mathcal{M} &= \{m_i\}_{i=2,3} \cup \{m_{i,j}\}_{i,j=2,3} \\ \mathcal{N} &= \{n_i\}_{i=2,3} \cup \{n_{i,j}\}_{i,j=2,3}. \end{aligned}$$

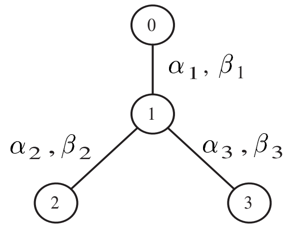

Figure 3 – A small network with two receivers. Associated with each link are an unconditional success probability, α_i, and conditional success probability, β_i.

Maximum likelihood estimates of $\alpha_1, \alpha_2, \alpha_3$ are given by

$$\begin{aligned} (\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3) &= \\ \arg \max_{\alpha_1,\alpha_2,\alpha_3} & \left[\max_{\beta_1,\beta_2,\beta_3} l(\mathcal{M} \,|\, \mathcal{N}, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3) \right]. \end{aligned}$$

Note that direct optimization requires the joint maximization of the six dimensional likelihood function; a daunting task even in this simple case. Using the EM Algorithm we can easily determine $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3)$ in $O(K)$ time, where K is the number of iterations of the algorithm. The marginal likelihood function of each α_i can also be computed using a factor graph representation of the network and a probability propagation algorithm in $O(K)$ time.

To explore the performance of these algorithms, consider three scenarios.

Scenario 1:

$$\begin{aligned} (\alpha_1, \alpha_2, \alpha_3) &= (0.80, 0.90, 0.70) \\ (\beta_1, \beta_2, \beta_3) &= (0.99, 0.99, 0.99) \end{aligned}$$

$$\mathcal{N} = \{n_i = 10000\}_{i=2,3} \cup \{n_{i,j} = 10000\}_{i,j=2,3}$$

Scenario 2:

$$\begin{aligned} (\alpha_1, \alpha_2, \alpha_3) &= (0.80, 0.90, 0.70) \\ (\beta_1, \beta_2, \beta_3) &= (0.95, 0.95, 0.95) \end{aligned}$$

$$\mathcal{N} = \{n_i = 1000\}_{i=2,3} \cup \{n_{i,j} = 1000\}_{i,j=2,3}$$

Scenario 3:

$$\begin{aligned} (\alpha_1, \alpha_2, \alpha_3) &= (0.80, 0.90, 0.70) \\ (\beta_1, \beta_2, \beta_3) &= (0.85, 0.95, 0.75) \end{aligned}$$

$$\mathcal{N} = \{n_i = 100\}_{i=2,3} \cup \{n_{i,j} = 100\}_{i,j=2,3}$$

The three scenarios were each simulated in $T = 100$ independent trials. In each trial, the maximum likelihood estimates (MLEs) and marginal likelihood functions were computed for each unconditional success probability. The maximums of the marginal likelihood functions (maximum marginal likelihood estimates - MMLEs) provide as set of alternatives to the MLEs. The mean (over all trials and links) absolute error, maximum (over all trials and links) absolute error, as well as the theoretical bound γ (as described in Section 7) for each scenario are summarized in Table 1.

Table 1. Loss estimation performance

Scenario	Absolute Error		
	MLE mean / max	MMLE mean / max	Bound γ
1	0.0106 / 0.0137	0.0053 / 0.0122	0.0199
2	0.0391 / 0.0452	0.0191 / 0.0256	0.0690
3	0.0533 / 0.1141	0.0854 / 0.1148	0.3625

In Scenario 1, we have a very large number of packet measurements (10000 of each type) and the β are almost 1. Both the MLE and marginal likelihood function produce nearly perfect inferences. In Scenarios 2 and 3, we see larger errors, but these errors are within the predicted bounds. It is also interesting to note that the maximum marginal likelihood estimator performs slightly better than the standard maximum likelihood estimator. This improvement has also been observed in many other applications [14]; marginalization over nuisance parameters tends to provide more robust estimators. Figure 9.1. displays typical results from each scenario.

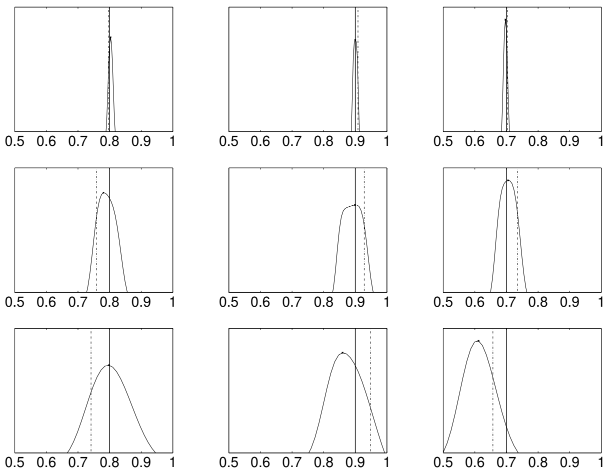

Figure 4 – Typical results from each measurement scenario. From left to right plots show results for α_1, α_2, and α_3. The true value is indicate with a solid vertical line, the MLE is indicated with a dashed vertical line. Also shown are the marginal likelihood functions for each of the α.

9.2. A Larger Network Simulation

We experimented using simulations based on the network in Figure 7.. We generated probe measurements by allowing each link in the network to assume one of two state values, 0 representing congestion, and 1 representing a light traffic burden. At time instants $t \in T$, the state of each link was updated according to a Markov process. The transition probability matrix of the process governing the state of link $(\rho(i), i)$ was determined by drawing α_i from a uniform distribution $U[0, 1]$, and then drawing β_i from $U[\alpha_i, 1]$; the matrix was designed so that if traffic were sent across the link it would experience a steady-state success probability of α_i and a conditional success probability of β_i. Packet-pair probes were sent to the various receivers in an ordered fashion designed to extract an informative subset of the possible $m_{i,j}$ and $n_{i,j}$. The times at which the first packets of these pairs were sent were determined from a Poisson process, such that inter-arrival times were well-separated. The second packet in a pair was sent one time instant later. 1600 packet pairs were sent through the network, with the destinations designed so that there was a uniform distribution across the network of divergence nodes (the node at which the paths of the individual packets in the packet-pairs separated). Such a distribution guarantees an equal (prior) exploration of all network parameters.

Figure 9.2. depicts the result of one of the experiments. The posterior distribution of success probability was calculated for each link, and plotted in the boxes; the arrows mark the true values. The confidence that can be placed on an estimate is clearly dependent on the amount of data that can be collected; estimation of the success probabilities of α_4, α_6, and α_7 is based on packet-pairs involving a packet traveling from the source to either node 6 or 7, both of which are extremely lossy paths. The maximum marginal likelihood estimators for the unconditional success probabilities resulted in a mean absolute error of 0.084, over 200 independent trials.

10. Conclusions

This work demonstrates the potential of unicast, end-to-end network measurement to determine internal network losses. We proposed a back-to-back packet pair measurement scheme that takes advantage of the correlations in losses experienced by closely time-spaced packets. We also developed two novel algorithms for likelihood analysis and estimation of internal link loss probabilities. This paper has laid the theoretical foundation for future investigations of unicast network tomography. One promising practical aspect of our framework is that it may be used in concert with various measurement tools, including active UDP probing or passive TCP monitoring. We are currently studying our framework with more sophisticated simulation tools as well

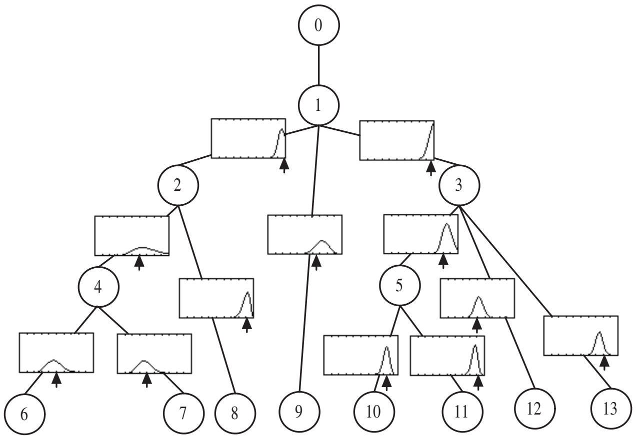

Figure 5 – An example of the results of the experiment described in Section 8.2. 1600 packet pairs were sent to various receivers in order to generate posterior probability distributions of the success rates of the links. These are plotted in the boxes on the links; the arrows mark the true values.

as with actual network measurements.

References

[1] R. Cáceres, N. Duffield, J. Horowitz, and D. Towsley, "Multicast-based inference of network-internal loss characteristics," *IEEE Trans. Info. Theory*, vol. 45, November 1999, pp. 2462–2480.

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

[3] S. Vander Wiel, J. Cao, D. Davis, and B. Yu, "Time-varying network tomography: router link data," in *Proc. Symposium on the Interface: Computing Science and Statistics*, (Schaumburg, IL), June 1999.

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

[5] "Multicast-based inference of network-internal characteristics (MINC)." See gaia.cs.umass.edu/minc.

[6] S. Ratnasamy and S. McCanne, "Inference of multicast routing trees and bottleneck bandwidths using end-to-end measurements," in *Proceedings of INFOCOM '99*, (New York, NY), March.

[7] V. Paxson, "End-to-end Internet packet dynamics," *IEEE/ACM Trans. Networking*, vol. 7, June 1999, pp. 277–292.

[8] M. Coates and R. Nowak, "Network inference from passive unicast measurement," Tech. Rep. TR0001, Rice University, Jan. 2000.

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

[10] J. Bolot and A. V. Garcia, "The case for FEC-based error control for packet audio in the internet." to appear in ACM Multimedia Systems.

[11] J.-C. Bolot, "End-to-end packet delay and loss behaviour in the Internet," in *Proc. SIGCOMM '93*, pp. 289–298, Sept. 1993.

[12] W. Fischer and K. Heier-Hellstern, "The Markov-modulated Poisson process (mmpp) cookbook," *Performance Evaluation*, vol. 18, no. 2, 1993, pp. 149–171.

[13] L. Kleinrock, *Queueing Systems. Volume I: Theory*. Wiley & Sons, New York, 1975.

[14] 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*, vol. Duke University, 1997, pp. Durham, NC.

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