Fast parameter estimation in loss tomography for networks of general topology

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:33983363

Terms of Use

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

FAST PARAMETER ESTIMATION IN LOSS TOMOGRAPHY FOR NETWORKS OF GENERAL TOPOLOGY[*]

By Ke Deng[†], Yang Li[§], Weiping Zhu[‡] and Jun S. Liu[§]

Tsinghua University[†], University of New South Wales[‡] and Harvard University[§]

As a technique to investigate link-level loss rates of a computer network with low operational cost, loss tomography has received considerable attentions in recent years. A number of parameter estimation methods have been proposed for loss tomography of networks with a tree structure as well as a general topological structure. However, these methods suffer from either high computational cost or insufficient use of information in the data. In this paper, we provide both theoretical results and practical algorithms for parameter estimation in loss tomography. By introducing a group of novel statistics and alternative parameter systems, we find that the likelihood function of the observed data from loss tomography keeps exactly the same mathematical formulation for tree and general topologies, revealing that networks with different topologies share the same mathematical nature for loss tomography. More importantly, we discover that a re-parametrization of the likelihood function belongs to the standard exponential family, which is convex and has a unique mode under regularity conditions. Based on these theoretical results, novel algorithms to find the MLE are developed. Compared to existing methods in the literature, the proposed methods enjoy great computational advantages.

1. Introduction. Network characteristics such as loss rate, delay, available bandwidth, and their distributions are critical to various network operations and important for understanding network behaviors. Although considerable attention has been given to network measurements, due to various reasons (e.g., security, commercial interest and administrative boundary) some characteristics of the network cannot be obtained directly from a large network. To overcome this difficulty, network tomography was proposed in [1], suggesting the use of end-to-end measurement and statistical inference to estimate characteristics of a large network. In an end-to-end measurement, a

1

number of sources are attached to the network of interest to send probes to receivers attached to the other side of the network, and paths from sources to receivers cover links of interest. Arrival orders and arrival times of the probes carry the information of the network, from which many network characteristics can be inferred statistically. Characteristics that have been estimated in this manner include link-level loss rates [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13], delay distributions [14, 15, 16, 17, 18, 19, 20, 21, 22, 23], origin-destination traffic [1, 16, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38], loss patterns [39], and the network topology [40]. In this paper, we focus on the problem of estimating loss rates.

Network topologies connecting sources to receivers can be divided into two classes: tree and general. A tree topology, as named, has a single source attached to the root of a multicast tree to send probes to receivers attached to the leaf nodes of the tree. A network with a general topology, however, requires a number of trees to cover all links of the network. Each tree has one source sending probes to receivers in it. Because of the use of multiple sources to send probes in a network with general topology, those nodes and receivers located at intersections of multiple trees can receive probes from multiple sources. In this case, we must consider impacts of probes sent by all sources simultaneously in order to get a good estimate. This task is much more challenging than the tree topology case.

Numerous methods have been proposed for loss tomography in a tree topology. Cáceres *et al.* [2] used a Bernoulli model to describe the loss behavior of a link, and derived the MLE for the pass rate of a path connecting the source to a node, which was expressed as the solution of a polynomial equation [2, 3, 4]. To ease the concern of using numerical methods to solve a high degree polynomial, several papers have been published to accelerate the calculation at the price of a little accuracy loss: Zhu and Geng proposed a recursively defined estimator based on a bottom-up strategy in [11, 12]; Duffield *et al.* proposed a closed-form estimator in [13], which has the same asymptotic variance as the MLE to the first order. Considering the unavailability of multicast in some networks, Harfoush *et al.* [5] and Coates *et al.* [6] independently proposed the use of the unicast-based multicast to send probes to receivers, where Coates *et al.* also suggested the use of the EM algorithm [41] to estimate link-level loss rates.

For networks beyond a tree, however, little research has been done, although a majority of networks in practice fall into this category. Conceptually, the topology of a general network can be arbitrarily complicated. However, no matter how complicated a general topology is, it can always be covered by a group of carefully selected trees, in each of which an end-

to-end experiment can be carried out independently to study the properties of the subnetwork covered by the tree. If these trees do not overlap with each other, the problem of studying the whole general network can be decomposed into a group of smaller subproblems, each for one tree. However, due to the complexity of the network topology and practical constraints, it is more than often that the selected trees overlaps significantly, i.e., some links are shared by two or more trees (see Figure 2 for example). The shared links of two selected trees induce dependence between the two trees. Simply ignoring the dependence leads to a loss of information. How to effectively integrate the information from multiple trees to achieve a joint analysis is a major challenge in network tomography of general topology.

The first effort on network tomography of general topology is due to Bu *et al.* [8], who attempted to extend the method in [2] to networks with general topology. Unfortunately, the authors failed to derive an explicit expression for the MLE like the one presented in [2] for this more general case. They then resorted to an iterative procedure (i.e., the EM algorithm) to search for the MLE. In addition, a heuristic method, called *minimum variance weighted average* (MVWA) algorithm, was also proposed in [8], which deals with each tree in a general topology separately and averages the results. The MVWA algorithm is less efficient than the EM algorithm, especially when the sample size is small. Rabbat *et al.* in [40] considered the tomography problem for networks with an unknown but general topology, mainly focusing on network topology identification, which is beyond the scope of our current paper.

In this paper, we provide a new perspective for the study of loss tomography, which is applicable to both tree and general topologies. Our theoretical contributions are: 1) introducing a set of novel statistics, which are complete and minimal sufficient; 2) deriving two alternative parameter systems and the corresponding re-parameterized likelihood functions, which benefit us both theoretically and computationally; 3) discovering that the loss tomography for a general topology shares the same mathematical formulation as that of a tree topology; and, 4) showing that the likelihood function belongs to the exponential family and has a unique mode (which is the MLE) under regularity conditions. Based on these theoretical results, we propose two new algorithms (a likelihood-equation-based algorithm called LE-ξ and an EM-based algorithm called PCEM) to find MLE. Compared to existing methods in the literature, the proposed methods are computationally much more efficient.

The rest of the paper is organized as follows. Section 2 introduces notations for tree topologies and the stochastic model for loss rate inference. Section 3 describes a set of novel statistics and two alternative parameter
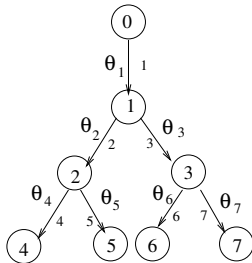
FIG 1. *A multicast tree.*

systems for loss rate analysis in tree topologies. New forms of the likelihood function are established based on the novel statistics and re-parametrization. Section 4 extends the above results to general topologies. Section 5 and 6 propose two new algorithms for finding MLE of θ, which enjoy great computational advantages over existing methods. Section 7 evaluates performances of the proposed methods by simulations. We conclude the article in Section 8.

2. Notations and Assumptions.

2.1. *Notations for Tree Topologies.* We use $T = (V, E)$ to denote the multicast tree of interest, where $V = \{v_0, v_1, ...v_m\}$ is a set of nodes representing the routers and switches in the network, and $E = \{e_1, ..., e_m\}$ is a set of directed links connecting the nodes. Two nodes connected by a directed link are called the *parent node* and the *child node*, respectively, and the direction of the arrow indicates that the parent forwards received probes to the child. Figure 1 shows a typical multicast tree. Note that the root node of a multicast tree has only one child, which is slightly different from an ordinary tree, and each non-root node has exactly one parent.

Each link is assigned a unique ID number ranging from 1 to m, based on which each node obtains its unique ID number ranging from 0 to m correspondingly so that link i connects node i with its parent node. Number 0 is reserved for the source node. In contrast to [2] and [13], whose methods are node-centric, our methods here focus on links instead. For a network with tree topology, the two reference systems are equivalent as there exists an one-to-one mapping between nodes and links: every node in the network (except for the source) has a unique parent link. For a network with general topology, however, the link-centric system is more convenient, as a node in the network may have multiple parent links.

We let f_i denote the unique parent link, B_i the brother links, and C_i the

child links, respectively, of link i. To be precise, B_i contains all links that share the same parent node with link i, including link i itself. A subtree $T_i = \{V_i, E_i\}$ is defined as the subnetwork composed of link i and all its descendant links. We let R and R_i denote the set of receivers (i.e., leaf nodes) in T and T_i, respectively. Sometimes, we also use R or R_i to denote the leaf links of T or T_i. The concrete meaning of R and R_i can be determined based on the context.

Taking the multicast tree displayed in Figure 1 as an example, we have $V = \{0, 1, \cdots, 7\}$, $E = \{1, 2, \cdots, 7\}$ and $R = \{4, 5, 6, 7\}$. For the subtree T_2, however, we have $V_2 = \{1, 2, 4, 5\}$, $E_2 = \{2, 4, 5\}$ and $R_2 = \{4, 5\}$. For link 2, its parent link is $f_2 = \{1\}$, its brother links are $B_2 = \{2, 3\}$, and its child links are $C_2 = \{4, 5\}$.

2.2. *Stochastic Model.* In a multicast experiment, n probe packages are sent from the root node 0 to all the receivers. Let $X_i^{(t)} = 1$ if the t-th probe package reached node i, and 0 otherwise. The status of probes at receivers $\{X_r^{(t)}\}_{r \in R, 1 \leq t \leq n}$ can be directly observed from the multicast experiment; but, the status of probes at internal links cannot be directly observed. In the following, we use $\mathbf{X}_R = \{X_R^{(t)}\}_{1 \leq t \leq n}$, where $X_R^{(t)} = \{X_r^{(t)}\}_{r \in R}$, to denote the data collected in a multicast experiment in tree T.

In this paper, we model the loss behavior of links by the Bernoulli distribution and assume the *spatial-temporal independence* and *temporal homogeneity* for the network, i.e. the event $\{X_i^{(t)} = 0 \mid X_{f_i}^{(t)} = 1\}$ are independent across i and t, and

$$P(X_i^{(t)} = 0 \mid X_{f_i}^{(t)} = 0) = 1, \ P(X_i^{(t)} = 0 \mid X_{f_i}^{(t)} = 1) = \theta_i.$$

We call $\theta = \{\theta_i\}_{i \in E}$ the *link-level loss rates*, and the goal of loss tomography is to estimate θ from $\mathbf{X}_R$.

2.3. *Parameter Space.* In principle, θ_i can be any value in $[0, 1]$. Thus, the natural parameter space of θ is $\Theta^* = [0, 1]^m$, an m-dimensional closed unit cube. In this paper, however, we assume that $\theta_i \in (0, 1)$ for every $i \in E$, and thus constrain the parameter space to $\Theta = (0, 1)^m$, to simplify the problem. If $\theta_i = 1$ for some $i \in E$, then the subtree T_i is actually disconnected from the other part of the network, since no probes can go through link i. In this case, the loss rate of other links in T_i are not estimable due to the lack of information. On the other hand, if $\theta_i = 0$ for some $i \in E$, the original network of interest degenerates to an equivalent network where node i is removed and all its child nodes are connected directly to node i's parent node. By constraining the parameter space of θ into Θ, we exclude these degenerate cases from consideration.

3. Statistics and Likelihood Function.

3.1. *The likelihood function and the MLE.* Given the loss model for each link, we can write down the likelihood function and use the maximum likelihood principle to determine unknown parameters. That is, we aim to find the parameter value that maximizes the log-likelihood function:

$$(3.1) \qquad \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \sum_{x \in \Omega} n(x) \log P(x; \theta),$$

where x stands for the observation at receivers (i.e., a realization of X_R), $\Omega = \{0,1\}^{|R|}$ is the space of all possible observations with $|R|$ denoting the size of set R, $n(x)$ is the number of occurrences of observation x, and $P(x; \theta)$ is the probability of observing x given parameter value θ. However, the log-likelihood function (3.1) is more symbolic than practical because of the following reasons: (1) evaluating the log-likelihood function (3.1) is an expensive operation as it needs to scan through all possible $x \in \Omega$; and, (2) the likelihood equation derived from (3.1) cannot be solved analytically, and it is often computationally expensive to pursue a numerical solution.

3.2. *Internal State and Internal View.* Instead of using the log-likelihood function (3.1) directly, we consider to rewrite it in a different form. Under the posited probabilistic model, the overall likelihood $P(\mathbf{X}_R \mid \theta)$ is the product of the likelihood from each probe, i.e.,

$$P(\mathbf{X}_R \mid \theta) = \prod_{t=1}^{n} P(X_R^{(t)} \mid \theta).$$

Thus, an alternative form of the overall likelihood can be obtained by explicating the likelihood of each single probe and accumulating them. Two concepts called *internal state* and *internal view* can be generated from this process.

3.2.1. *Internal State.* For a link $i \in E$, given the observation of probe t at R_i and R_{f_i}, we are able to partially confirm whether the probe passes link i. Formally, for observation $\{X_j^{(t)}\}_{j \in R_i}$, we define $Y_i^{(t)} = \max_{j \in R_i} X_j^{(t)}$ as the *internal state* of link i for probe t. If $Y_i^{(t)} = 1$, probe t reaches at least one receiver attached to T_i, which implies that the probe passes link i. Furthermore, by considering $Y_{f_i}^{(t)}$ and $Y_i^{(t)}$ simultaneously, we have three possible scenarios for each internal node i:

- $Y_{f_i}^{(t)} = Y_i^{(t)} = 1$, i.e., we observed that probe t passed link i; or

- $Y_{f_i}^{(t)} = 1$ and $Y_i^{(t)} = 0$, i.e., we observed that probe t reached node f_i, but we did not know whether it reached node i or not; or
- $Y_{f_i}^{(t)} = Y_i^{(t)} = 0$, i.e., we did not know whether probe t reached node f_i or not at all;

as $Y_{f_i}^{(t)} = 0$ and $Y_i^{(t)} = 1$ can never happen by definition of Y.

The three scenarios have different impacts on the likelihood function. Formally, define

$$
\begin{aligned}
E_{t,1} &= \{i \in E : Y_i^{(t)} = 1\}, \\
E_{t,2} &= \{i \in E : Y_{f_i}^{(t)} = 1, Y_i^{(t)} = 0\}, \\
E_{t,3} &= \{i \in E : Y_{f_i}^{(t)} = Y_i^{(t)} = 0\}.
\end{aligned}
$$

We have

$$
(3.2) \qquad P(X_R^{(t)} \mid \theta) = \prod_{i \in E_{t,1}} (1 - \theta_i) \prod_{i \in E_{t,2}} \xi_i(\theta),
$$

where

$$
\xi_i(\theta) = P(X_j = 0, \ \forall \ j \in R_i \mid X_{f_i} = 1; \theta)
$$

represents the probability that a probe sending out from the root node of T_i fails to reach any leaf node in T_i.

3.2.2. Internal View.

Accumulating the internal states of each link in the experiment, we have

$$
(3.3) \qquad n_i(1) = \sum_{t=1}^{n} Y_i^{(t)},
$$

which counts the number of probes whose pass through link i can be confirmed from observations. Specifically, we define $n_0(1) = n$. Moreover, define

$$
(3.4) \qquad n_i(0) = n_{f_i}(1) - n_i(1),
$$

for $\forall \ i \in E$. We call the statistics $\{n_i(1), n_i(0)\}$ the *internal view* of link i. Based on internal views, we can write the log-likelihood of $\mathbf{X}_R$ in a more convenient form:

$$
L(\theta) = \sum_{i \in E} \Big[n_i(1) \log(1 - \theta_i) + n_i(0) \log \xi_i(\theta) \Big].
$$

3.3. *Re-parametrization.* Two alternative parameter systems can be introduced to re-parameterize the above log-likelihood function. First, based on the definition of $\xi_i(\theta)$, we have

(3.5) $$\xi_i(\theta) = \theta_i + (1 - \theta_i) \prod_{j \in C_i} \xi_j(\theta), \quad i \in E.$$

Note that if $i \in R$, we have $C_i = \emptyset$, and (3.5) degenerates to $\xi_i(\theta) = \theta_i$. Let $\xi_i \triangleq \xi_i(\theta)$ for $i \in E$. Equation (3.5) defines a one-to-one mapping between two parameter systems $\theta = \{\theta_i\}_{i \in E}$ and $\xi = \{\xi_i\}_{i \in E}$, i.e.,

$$\Gamma : \Theta \mapsto \Xi, \ \xi = \Gamma(\theta) \triangleq \big(\xi_1(\theta), \cdots, \xi_m(\theta)\big),$$

where Θ and Ξ are the domain and image, respectively. The inverse mapping of Γ is

$$\Gamma^{-1} : \Xi \mapsto \Theta, \ \theta = \Gamma^{-1}(\xi) \triangleq \big(\theta_1(\xi), \cdots, \theta_m(\xi)\big), \text{ where}$$

(3.6) $$\theta_i(\xi) = \frac{\xi_i - \prod_{j \in C_i} \xi_j}{1 - \prod_{j \in C_i} \xi_j}, \quad i \in E.$$

Using ξ to replace θ in $L(\theta)$, we have the following alternative log-likelihood function with ξ as parameters:

$$L(\xi) = \sum_{i \in E} \left[n_i(1) \log\Big(\frac{1 - \xi_i}{1 - \prod_{j \in C_i} \xi_j}\Big) + n_i(0) \log \xi_i \right].$$

Second, $L(\xi)$ can be further re-organized into

$$
\begin{aligned}
L(\xi) &= \sum_{i \in E} n_i(1) \log\Big(\frac{1 - \theta_i(\xi)}{\xi_i}\Big) + \sum_{i \in E} n_{f_i}(1) \log \xi_i \\
&= n \log \xi_1 + \sum_{i \in E} n_i(1) \log \psi_i(\xi),
\end{aligned}
$$

where $\psi_i(\xi)$ is defined as

(3.7) $$\psi_i(\xi) = \begin{cases} \log \frac{1 - \theta_i(\xi)}{\xi_i}, & i \in R, \\ \log \frac{\xi_i - \theta_i(\xi)}{\xi_i}, & i \notin R. \end{cases}$$

Similarly, let $\psi_i \triangleq \psi(\xi)$ for $i \in E$. Equation (3.7) defines a one-to-one mapping between parameter system $\xi = \{\xi_i\}_{i \in E}$ and $\psi = \{\psi_i\}_{i \in E}$, i.e.,

$$\Lambda : \Xi \mapsto \Psi, \ \psi = \Lambda(\xi) \triangleq \big(\psi_1(\xi), \cdots, \psi_m(\xi)\big),$$

where $\Psi \triangleq \Lambda(\Xi)$ is the image of ψ. The inverse mapping of Λ is

$$\Lambda^{-1}: \ \Psi \mapsto \Xi, \ \xi = \Lambda^{-1}(\psi) \triangleq \big(\xi_1(\psi), \cdots, \xi_m(\psi)\big).$$

It can be shown that

$$\psi_i = \log P\big(X_i = 1 \mid X_{f_i} = 1; \ X_j = 0, \ \forall \ j \in R_i\big) \text{ for } i \notin R,$$

from which the physical meaning of ψ_i can be better understood. Using ψ to replace ξ in $L(\xi)$, we have the following log-likelihood function with ψ as parameters:

$$L(\psi) = n \log \xi_1(\psi) + \sum_{i \in E} n_i(1)\psi_i.$$

To illustrate the relations of θ, ξ and ψ, let's consider the toy network in Figure 1 with $\theta_i = 0.1$ for $i = 1, \cdots, 7$. It is easy to check that:

$$
\begin{aligned}
\xi_4 = \xi_5 = \xi_6 = \xi_7 &= 0.1, \\
\xi_2 = \xi_3 &= 0.1 + (1 - 0.1) \times 0.1^2 = 0.109, \\
\xi_1 &= 0.1 + (1 - 0.1) \times 0.109^2 \approx 0.1107; \\
\psi_4 = \psi_5 = \psi_6 = \psi_7 &= \log \frac{1 - 0.1}{0.1} \approx 2.1972, \\
\psi_2 = \psi_3 &= \log \frac{0.109 - 0.1}{0.109} \approx -2.4941, \\
\psi_1 &\approx \log \frac{0.1107 - 0.1}{0.1107} \approx -2.3366.
\end{aligned}
$$

We list these concrete values in Table 1 for comparison purpose. The notations, statistics and parameter systems are summarized into Table 2 for easy reference.

TABLE 1

Comparing different parameter systems for the toy network in Figure 1

Link	1	2	3	4	5	6	7
θ	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ξ	0.1092	0.109	0.109	0.1	0.1	0.1	0.1
ψ	-2.3366	-2.4941	-2.4941	2.1972	2.1972	2.1972	2.1972

4. Likelihood Function for General Networks. In this section, we will extend the concept of internal view and alternative parameter systems to general networks containing multiple trees. Formally, we use $\mathcal{N} = \{T_1, \cdots, T_K\}$ to denote a general network covered by K multicast trees,

TABLE 2
Collection of symbols

Symbol	Meaning
T	the multicast tree of interest
V	the node set of T
E	the link set of T
R	the set of leaf nodes (receivers) or leaf links of T
T_i	the subtree with link i as the root link
V_i	the node set of subtree T_i
R_i	the set of leaf nodes (receivers) or leaf links of T_i
f_i	the parent link of link i
B_i	the brother links of link i
C_i	the child links of link i
$X_i^{(t)}$	$X_i^{(t)} = 1$ if probe t reached node i, and 0 otherwise
$Y_i^{(t)}$	$\max_{j \in R_i} X_j^{(t)}$
$n_i(1)$	$\sum_{t=1}^n Y_i^{(t)}$, number of probes that passed link i for sure
$n_i(0)$	$n_{f_i}(1) - n_i(1)$
θ_i	$P(X_i = 0 \mid X_{f_i} = 1)$, the loss rate of link i
ξ_i	$P(X_j = 0, \ \forall \ j \in R_i \mid X_{f_i} = 1)$, the loss rate of T_i
ψ_i	$\log P(X_i = 1 \mid X_{f_i} = 1; X_j = 0, \ \forall \ j \in R_i)$

where each multicast tree T_k covers a subnetwork with S_k as the root link. For example, Figure 2 illustrates a network with $K = 2$, $S_1 = 0$ and $S_2 = 32$. Let $\mathscr{S} = \{S_1, \cdots, S_K\}$ be the set of root links in $\mathscr{N}$.

For each tree T_k, let $\{n_{k,i}(1), n_{k,i}(0)\}$ be the internal view of link $i \in E_k$ in T_k based on the n_k probes sent out from S_k. Specially, define $n_{k,i}(1) = 0$ for link $i \notin E_k$. The experiment on T_k results in the following log-likelihood functions with θ, ξ and ψ as parameters, respectively:

$$
\begin{aligned}
L_k(\theta) &= \sum_{i \in E} \left[n_{k,i}(1) \log(1 - \theta_i) + n_{k,i}(0) \log \xi_i(\theta) \right], \\
L_k(\xi) &= \sum_{i \in E} \left[n_{k,i}(1) \log\left(\frac{1 - \xi_i}{1 - \prod_{j \in C_i} \xi_j} \right) + n_{k,i}(0) \log \xi_i \right], \\
L_k(\psi) &= n_k \log \xi_{S_k}(\psi) + \sum_{i \in E} n_{k,i}(1) \psi_i.
\end{aligned}
$$

Considering that a multicast experiment in a general network $\mathscr{N}$ is a pool of K independent experiments in the K trees of $\mathscr{N}$, the log-likelihood function of the whole experiment is just the summation of K components,
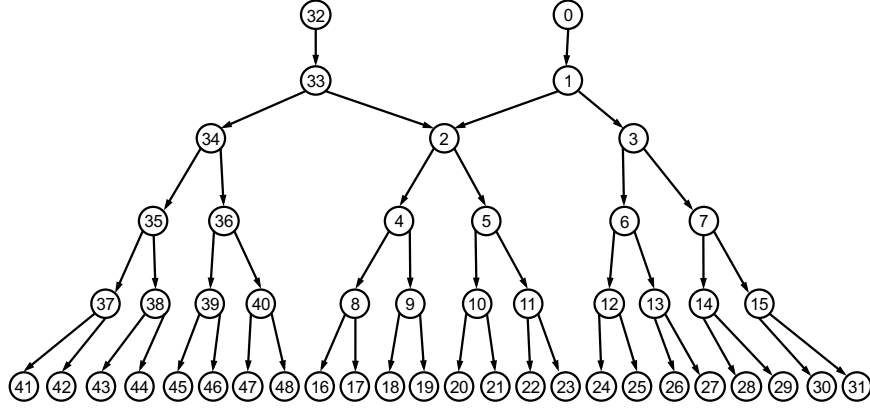
FIG 2. *A 5-layer network covered by two trees.*

i.e.,

$$L(\theta) = \sum_{k=1}^{K} L_k(\theta), \quad L(\xi) = \sum_{k=1}^{K} L_k(\xi) \quad \text{and} \quad L(\psi) = \sum_{k=1}^{K} L_k(\psi).$$

Define the internal view of link i in a general network as:

$$(4.1) \qquad n_i(1) = \sum_{k=1}^{K} n_{k,i}(1), \quad n_i(0) = \sum_{k=1}^{K} n_{k,i}(0).$$

It is straightforward to see that

$$(4.2) \qquad L(\theta) \;=\; \sum_{i \in E} \Big[n_i(1) \log(1 - \theta_i) + n_i(0) \log \xi_i(\theta) \Big],$$

$$(4.3) \qquad L(\xi) \;=\; \sum_{i \in E} \Big[n_i(1) \log(\frac{1 - \xi_i}{1 - \prod_{j \in C_i} \xi_j}) + n_i(0) \log \xi_i \Big],$$

$$(4.4) \qquad L(\psi) \;=\; \sum_{k=1}^{K} n_k \log \xi_{S_k}(\psi) + \sum_{i \in E} n_i(1) \psi_i.$$

Note that in this general case, we have:

$$\text{for } i \in \mathscr{S}, \ n_{S_k}(1) = n_k; \text{ and}$$
$$\text{for } i \notin \mathscr{S}, \ n_i(0) + n_i(1) = \sum_{j \in F_i} n_j(1).$$

Here, F_i stands for the unique or multiple parent links of link i in the general network $\mathcal{N}$.

5. Likelihood Equation. The bijections among the the three parameter systems mean that we can switch among them freely without changing the result of parameter estimation:

PROPOSITION 1. *The results of likelihood inference based on the three parameter systems are equivalent, i.e.,*

$$\Gamma\big(\arg\max_{\theta\in\Theta} L(\theta)\big) = \arg\max_{\xi\in\Xi} L(\xi) = \Lambda^{-1}\big(\arg\max_{\psi\in\Psi} L(\psi)\big);$$

and the likelihood equations under different parameters share the same solution, i.e.,

$$\frac{\partial L(\theta)}{\partial\theta}\bigg|_{\theta=\theta^*} = 0 \Leftrightarrow \frac{\partial L(\xi)}{\partial\xi}\bigg|_{\xi=\xi^*} = 0 \Leftrightarrow \frac{\partial L(\psi)}{\partial\psi}\bigg|_{\psi=\psi^*} = 0,$$

if $\theta^* \in \Theta$, or $\xi^* \in \Xi$, or $\psi^* \in \Psi$, where $\Gamma(\theta^*) = \xi^* = \Lambda^{-1}(\psi^*)$.

This flexibility provides us great theoretical and computational advantages. On the the theoretical aspect, as $\{n_k\}_{k=1}^K$ are known constants, $L(\psi)$ falls into the standard exponential family with ψ as the natural parameters. Thus, based on the properties of the exponential family [42], we have the following results immediately:

PROPOSITION 2. *The following results hold for loss tomography:*

1. *Statistics $\{n_i(1)\}_{i\in E}$ are complete and minimal sufficient;*
2. *The likelihood equation $\frac{\partial L(\psi)}{\partial\psi} = 0$ has at most one solution $\psi^* \in \Psi$;*
3. *If ψ^* exists, ψ^* (or $\theta^* = (\Lambda \circ \Gamma)^{-1}(\psi^*)$) is the MLE.*

On the computational aspect, the parameter system ξ plays a central role. Different from the likelihood equation with θ as parameters, which is intractable, the likelihood equation with ξ as parameters enjoys unique computational advantages. Let

$$r_i = \frac{n_i(1)}{n_i(1) + n_i(0)}.$$

It can be shown that the likelihood equation with ξ as parameters is:

$$(5.1) \qquad \xi_i = (1 - r_i) + r_i \cdot \prod_{j\in B_i} \xi_j \cdot I(i \notin \mathscr{S}),$$

for any link $i \in E$.

For link $i \in \mathscr{S}$, (5.1) degenerates to

$$\xi_i = 1 - r_i.$$

For link $i \notin \mathscr{S}$, define $\pi_i = \prod_{j \in B_i} \xi_j$. We note that solving ξ_i from (5.1) is equivalent to solving the following equation about π_i:

$$(5.2) \qquad \pi_i = \prod_{j \in B_i} \left[(1 - r_j) + r_j \cdot \pi_i \right],$$

which can be solved analytically when $|B_i| = 2$, and by numerical approaches [43] when $|B_i| > 2$.

From (5.1) and (5.2), it is transparent that only local statistics $\{r_j\}_{j \in B_i}$ are involved in estimating ξ_i. This observation leads to the following fact immediately: if a processor keeps the values of $\{r_j\}_{j \in B_i}$ in its memory, $\{\xi_j\}_{j \in B_i}$ can be effectively estimated by the processor independent of estimating the other parameters. Based on this unique property of the likelihood equation with ξ as parameter, we propose a parallel procedure called "LE-ξ" algorithm to estimate θ (see Algorithm 1).

Algorithm 1. The LE-ξ Algorithm

Hardwire requirement:
 $\{\mathbb{P}_i\}_i$: a collection of processors indexed by node ID $i \in V$;
Input:
 $\{r_j\}_{j \in E}$ with distributed storage where $\mathbb{P}_i$ keeps $\{r_j\}_{j \in C_i}$;
Output:
 $\hat{\theta} = \{\hat{\theta}_j\}_{j \in E}$.

Procedure:
 Operate in parallel for IDs of all non-leaf nodes $\{i \in V : i \notin R\}$
  If $i \in \mathscr{S}$,
   Get $\hat{\xi}_j = 1 - r_j$ for the unique child link j of node i with $\mathbb{P}_i$;
  If $i \notin \mathscr{S}$,
   Get $\hat{x}_i$ with $\mathbb{P}_i$ from $\{r_j\}_{j \in C_i}$ by solving equation below about x
    $x = \prod_{j \in C_i} \left[(1 - r_j) + r_j x \right]$,
   Get $\hat{\xi}_j = (1 - r_j) + r_j \hat{x}_i$ for every $j \in C_i$ with $\mathbb{P}_i$;
 End parallel operation

 Return $\hat{\theta} = \Gamma^{-1}(\hat{\xi})$.

The validity of the LE-ξ algorithm is guaranteed by the following theorem:

THEOREM 1.   *When $n_k \to \infty$ for all $1 \le k \le K$, with probability one, the LE-ξ algorithm has a unique solution in $\Theta = (0,1)^m$ that is the MLE.*

Proof of Theorem 1. First, we will show that under the following regularity conditions:

$$(5.3) \qquad n_i(1) > 0, \ n_i(0) > 0 \ \text{ and } \ \sum_{j \in F_i} n_j(1) < \sum_{j \in B_i} n_j(1) \ \text{ for } \forall \ i \in E,$$

the likelihood equation with ξ as parameters has a unique solution in $(0,1)^m$.

Note that when (5.3) holds, we have

$$0 < r_i < 1 \ \text{ and } \ \sum_{j \in B_i} r_j > 1 \text{ for } \forall \ i \in E.$$

Because the Lemma 1 in [2] has shown that: if $0 < c_j < 1$ and $\sum_j c_j > 1$, equation for x

$$x = \prod_j \left[(1 - c_j) + c_j x \right]$$

has a unique solution in $(0,1)$. It is transparent that under the regularity conditions in (5.3), equation (5.2) has a unique solution $\hat{\pi}_i \in (0,1)$ for every non-root link i. Considering that for every link $i \in E$,

$$\xi_i = (1 - r_i) + r_i \cdot \pi_i \cdot I(i \notin \mathscr{S}),$$

it is straightforward to see that the likelihood equation with ξ as parameters has a unique solution $\hat{\xi} \in (0,1)^m$.

Moreover, if

$$(5.4) \qquad\qquad\qquad\qquad \hat{\xi} \in \Xi,$$

we have $\hat{\psi} = \Lambda(\hat{\xi}) \in \Psi$. Thus, based on Proposition 1 and Proposition 2, we know that if both (5.3) and (5.4) are satisfied, $\hat{\psi}$, which is the solution of the likelihood equation with ψ as parameter, is the MLE. Considering the bijections among θ, ξ and ψ, this also means that $\hat{\theta} = \Gamma^{-1}(\hat{\xi})$ is the MLE.

Note that the probability that (5.3) or (5.4) fails goes to zero when $n_k \to \infty$ for $k = 1, \cdots, K$, we complete the proof.

The large sample properties of MLE have been well studied by [2] for tree topology, where the asymptotic normality, asymptotic variance and confidence interval are established. Considering that the likelihood function keeps exactly the same form for both tree and general topology, it is natural to extend these theoretical results for MLE established in [2] to a network of general topology.

With a finite sample, there is a chance that (5.3) or (5.4) fails. In this case, the MLE of θ falls out of Θ, and may be missed by the LE-ξ algorithm. For example,

1. If $n_i(1) = 0$ and $n_i(0) > 0$, we have $\hat{\hat{\xi}}_i = 1$;
2. If $n_i(1) > 0$ and $n_i(0) = 0$, we have $\hat{\hat{\xi}}_i = 0$;
3. If $\sum_{j \in F_i} n_j(1) = \sum_{j \in B_i} n_j(1) > 0$, we have $\hat{\theta}_i = 0$;
4. If $\hat{\xi}_i \le \hat{\pi}_i = \prod_{j \in B_i} \hat{\xi}_j$ (i.e., $\hat{\xi} \notin \Xi$), we have $\hat{\theta}_i \le 0$.

Moreover, if $n_i(1) = n_i(0) = 0$, the parameters in subtree T_i (i.e., $\{\theta_j\}_{j \in T_i}$) are not estimable due to lack of information.

In all these cases, we cannot guarantee that the estimate from the LE-ξ algorithm $\hat{\theta}$ is the global maximum in $\Theta^* = [0,1]^m$. In practice, we can increase sample size by sending additional probes to avoid this dilemma. In case that it is not realistic to send additional probes, we can simply replace $\hat{\theta}$ by the point in Θ^* that is the closest to $\hat{\theta}$, or search the boundary of Θ^* to maximize $L(\theta)$.

6. Impacts on the EM Algorithm. The above theoretical results have two major impacts to the EM algorithm widely used in loss tomography. First, as we have shown in Theorem 1, as long as regularity conditions (5.3) and (5.4) hold, likelihood function $L(\theta)$ has a unique mode in Θ. In this case, the EM algorithm always converges to the MLE for any initial value $\theta^{(0)} \in \Theta$. Considering that the chance of violating (5.3) or (5.4) goes to zero with the increase of sample size, we have the following corollary for the EM algorithm immediately:

COROLLARY 1.   *When $n_k \to \infty$ for all $1 \le k \le K$, the EM algorithm converges to the MLE with probability one for any initial value $\theta^{(0)} \in \Theta$.*

Second, the formulation of the new statistics $\{n_i(1)\}_{i \in E}$ naturally leads to a "pattern-collapsed" implementation of the EM algorithm, which is computationally much more efficient than the widely used naive implementation where the samples are processed one by one separately. In the naive implementation of the EM algorithm, one enumerates all possible configurations of the internal links compatible with each sample carrying the information on whether the corresponding probe reached the leaf nodes or not. The complexity of the naive implementation in each E-step can be $O(n2^m)$ in the worst case. With the pattern-collapsed implementation, however, the complexity of an E-step can be dramatically reduced to $O(m)$.

The first pattern collapsed EM algorithm is proposed by Deng et al. [23] in the context of delay tomography. The basic idea is to reorganize the observed

data at receivers in delay tomography into *delay patterns*, and make use of a delay pattern database to greatly reduce the computational cost in the E-step. We will show below that the spirit of this method can be naturally extended to the study of loss tomography.

Define event $\{Y_i = 0 \mid Y_{f_i} = 1\}$, or equivalently $\{X_j = 0, \ \forall\, j \in R_i \mid X_{f_i} = 1\}$, as the *loss event* in subtree T_i, denoted as $\mathcal{L}_i$. Let $L(\mathbf{X}_V; \theta)$ be the log-likelihood of the complete data where the status of probes at internal nodes are also observed, $\theta^{(r)}$ be the estimation obtained at the r-th iteration, the objective function to be maximized in the $(r+1)$-th iteration of the EM algorithm is:

$$
\begin{aligned}
Q(\theta, \theta^{(r)}) &= E_{\theta^{(r)}}\big(L(\mathbf{X}_V; \theta) \mid \mathbf{X}_R\big) = \sum_{t=1}^{n} E_{\theta^{(r)}}\big(L(X_E^{(t)}; \theta) \mid X_R^{(t)}\big) \\
(6.1) \qquad &= \sum_{i \in E} \Big[n_i(1)\log(1 - \theta_i) + n_i(0)Q(\theta_{T_i}, \theta^{(r)}) \Big],
\end{aligned}
$$

where $Q_{\mathcal{L}_i}(\theta_{T_i}, \theta^{(r)})$, which is called the *localized Q-function* of loss event $\mathcal{L}_i$, is defined as:

$$
Q_{\mathcal{L}_i}(\theta_{T_i}, \theta^{(r)}) = E_{\theta^{(r)}}\Big[logP(X_{T_i}; \theta_{T_i}) \mid \mathcal{L}_i \Big].
$$

Similar to the results for delay patterns shown in the Proposition 1 of [23], $Q_{\mathcal{L}_i}(\theta_{T_i}, \theta^{(r)})$ has the following decomposition:
(6.2)
$$
Q_{\mathcal{L}_i}(\theta_{T_i}, \theta^{(r)}) = \big(1 - \psi_i(\theta^{(r)})\big)\log(\theta_i) + \psi_i(\theta^{(r)})\Big[\log(1 - \theta_i) + \sum_{j \in C_i} Q_{\mathcal{L}_j}(\theta_{T_j}, \theta^{(r)}) \Big].
$$

Integrating (6.1) and (6.2), we have

$$
Q(\theta, \theta^{(r)}) = \sum_{i \in E} \Big[\omega_i(1)\log(1 - \theta_i) + \omega_i(0)\log(\theta_i) \Big],
$$

where $\{\omega_i(1), \omega_i(0)\}_{i \in E}$ are defined recursively as follows:
- for $i \in \{S_1, \cdots, S_K\}$, i.e., being one of the root links,

$$
\begin{aligned}
\omega_i(1) &= n_i(1) + n_i(0) \cdot \psi_i(\theta^{(r)}), \\
\omega_i(0) &= n_i(0) \cdot \big(1 - \psi_i(\theta^{(r)})\big);
\end{aligned}
$$

- for $i \notin \{S_1, \cdots, S_K\}$,

$$
\begin{aligned}
\omega_i(1) &= n_i(1) + \sum_{j \in F_i} \omega_j(0) \cdot \psi_j(\theta^{(r)}), \\
\omega_i(0) &= \sum_{j \in F_i} \omega_j(0) \cdot \big(1 - \psi_j(\theta^{(r)})\big).
\end{aligned}
$$

And, the M-step is:

$$\theta_i^{(r+1)} = \frac{\omega_i(0)}{\omega_i(0) + \omega_i(1)}, \ \forall \ i \in E.$$

In this paper, we refer to the EM with the pattern-collapsed implementation as PCEM to distinguish from the EM with the naive implementation, which is referred to as NEM. We have shown in [23] by simulation and theoretical analysis that PCEM is computationally much more efficient than the naive EM in the context of delay tomography. In context of loss tomography, it is easy to see from the above analysis that PCEM enjoys a complexity of $O(m)$ for each E-step, which is much more efficient than the naive EM, whose complexity can be $O(n2^m)$ for each E-step in the worst case.

7. Simulation Studies.
We verify the following facts via simulations:

1. PCEM obtains exactly same results as the naive EM for both tree topology and general topology, but is much faster;
2. LE-ξ obtains exactly same results as the EM algorithm when regularity conditions (5.3) and (5.4) hold, but is much faster when the speedup from parallel computation is considered.

We carried out simulations on a 5-layer network covered by two trees as showed in Figure 2. The network has 49 nodes labeled from Node 0 to Node 48, and two sources Node 0 and Node 32. We conduct two set of simulations. One is based on the ideal model, and the other is by using network simulator 2 (ns-2, [44]).

7.1. *Simulation study with the ideal model.* In the first set of simulations, the data is generated from the ideal model where each link has a pre-defined constant loss rate.

To test the performance of methods on different magnitude of loss rates, we draw the link-level loss rates from Beta distributions with different parameters. The link-level loss rates $\theta = \{\theta_i\}_i$ are randomly sampled from $\text{Beta}(1, 99)$, $\text{Beta}(5, 995)$, $\text{Beta}(2, 998)$ or $\text{Beta}(1, 999)$. For example, the mean of $\text{Beta}(1, 999)$ is 0.001, so if we draw loss rates from $\text{Beta}(1, 999)$, then on average we expect a 0.1% link-level loss rate in the network. Given the loss rate vector θ for each network, we generated 100 independent datasets with sample sizes 50, 100, 200 and 500, respectively. Thus, a total of $100 \times 4 \times 4 = 1600$ datasets were simulated. The sample size is evenly distributed into the two trees, e.g., the source of each tree will send out 100 probes when sample size $n = 200$.

TABLE 3

Performance of different methods on simulated data from ideal model on the 5-layer network in Figure 2.

		Beta(1,100)		Beta(5,1000)		Beta(2,1000)		Beta(1,1000)	
n	Method	Time (ms)	MSE (1e-5)	Time (ms)	MSE (1e-5)	Time (ms)	MSE (1e-5)	Time (ms)	MSE (1e-5)
50	NEM	5065.50	764.49	4444.10	364.88	3396.20	188.78	2778.15	91.10
	PCEM	2.25	764.49	2.55	364.88	2.55	188.78	2.60	91.10
	LE-ξ	2.15	768.14	2.25	365.44	2.15	189.12	2.20	91.22
	MVWA	2.35	772.01	2.05	368.78	2.35	189.70	2.60	91.34
100	NEM	12102.90	434.39	9109.65	245.08	7323.90	89.64	7112.55	36.69
	PCEM	3.85	434.39	3.90	245.08	3.90	89.64	4.85	36.69
	LE-ξ	3.70	436.41	3.50	247.29	3.80	90.02	4.30	36.74
	MVWA	4.70	439.50	4.40	247.36	4.90	90.08	5.00	36.80
200	NEM	31368.95	212.79	23558.50	109.03	19184.65	44.79	16249.45	18.52
	PCEM	12.10	212.79	11.05	109.03	10.20	44.79	11.20	18.52
	LE-ξ	11.65	212.87	10.65	109.69	10.70	44.91	10.75	18.54
	MVWA	13.90	213.90	11.05	109.75	11.80	44.90	10.20	18.55
500	NEM	90221.95	85.17	49625.15	43.88	45402.80	14.35	43475.15	6.98
	PCEM	25.05	85.17	22.25	43.88	22.70	14.35	22.35	6.98
	LE-ξ	28.05	85.17	21.75	43.88	21.30	14.35	23.10	6.98
	MVWA	27.00	85.45	22.50	44.02	28.10	14.38	24.25	6.98

To each of the 1600 simulated data sets, we applied NEM, PCEM, LE-ξ and MVWA, respectively. The stopping rule of the EM algorithms is $\max_i |\theta_i^{(t+1)} - \theta_i^{(t)}| \le 10^{-6}$, the initial values are $\theta_i^{(0)} = 0.03$ for all $i \in E$. We implement MVWA algorithm by obtaining MLE estimate in each tree with LE-ξ algorithm. Table 3 summarize the average running time and MSEs of these methods under different settings. From the tables, we can see that:

1. The performances of all four methods in term of MSE improve with the increase of sample size n in both networks.
2. NEM and PCEM always get exactly same results.
3. The result from LE-ξ is slightly different from these of the EM algorithms when sample size n is as small as 100 or 200 due to the violation of the regularity conditions (5.3) and (5.4). When $n = 500$, the results of LE-ξ and EM algorithms become identical.
4. LE-ξ, PCEM and MVWA algorithm implemented with LE-ξ are dramatically faster than NEM. For example, in simulation configuration with Beta$(1, 1000)$ and sample size $n = 500$, the running time of NEM is almost 2,000 times of LE-ξ and PCEM.
5. The MSE of MVWA is always larger than the MSE of the other three methods, which capture the MLE. This is consistent with the result in [8].

7.2. *Simulation study by network simulator 2.* We also conduct the simulation study using ns-2. We use the network topology shown in Figure 2, where the two sources located at Node 0 and Node 32 multicast probes to the attached receivers. Besides the network traffic created by multicast

Table 4

Performance of different methods on simulated data from network simulator 2.

Sim time (n)	Method	Time (ms)	MSE (1e-5)
1s (528)	NEM	54675.20	581.29
	PCEM	21.80	581.29
	LE-ξ	22.00	581.29
	MVWA	22.80	584.50
2s (2038)	NEM	256270.60	232.49
	PCEM	21.80	232.49
	LE-ξ	22.00	232.49
	MVWA	22.80	233.63
5s (5648)	NEM	745393.40	122.23
	PCEM	329.40	122.23
	LE-ξ	315.00	122.23
	MVWA	299.10	122.86
10s (13355)	NEM	1334478.90	75.24
	PCEM	612.90	75.24
	LE-ξ	612.10	75.24
	MVWA	657.00	76.59

probing, a number of TCP sources with various window sizes and a number of UDP sources with different burst rates and periods are added at several nodes to produce cross-traffic. The TCP and UDP cross-traffic takes about 80% of the total network traffic. We generated 100 independent datasets by running the ns-2 simulation for 1, 2, 5 or 10 simulation seconds. The longer experiments generate more multicast probes samples. We record all pass and loss events for each link during the simulation, and the actual loss rates are calculated and considered as the true loss rates for calculating MSEs. Table 4 shows the average number of packets (sample size), running time of methods and MSEs for different ns-2 simulation times.

In Table 4, we observe similar results as shown in Table 3. The performances of all four methods in term of MSE improve with the increase of sample size n in both networks. The result from LE-ξ is slightly different from these of the EM algorithms when sample size n is as small. More importantly, LE-ξ and PCEM are dramatically faster than NEM.

8. Conclusion. We proposed a set of sufficient statistics called *internal view* and two alternative parameter systems ξ and ψ for loss tomography. We found that the likelihood function keeps the exactly same formulation for both tree and general topologies under all three types of parametrization (θ, ξ and ψ), and we can switch among the three parameter systems freely without changing the result of parameter estimation. We also discovered that the parameterization of the likelihood function based on ψ falls into the standard exponential family, which has a unique mode in parameter space Ψ under regularity conditions, and the parametrization based on ξ leads to an efficient algorithm called LE-ξ to calculate the MLE, which can be carried out in a parallel fashion. These results indicate that loss tomography for general topologies enjoys the same mathematical nature

as that for tree topologies, and can be resolved effectively. The proposed statistics and alternative parameter systems also lead to a more efficient pattern-collapsed implementation of the EM algorithm for finding MLE, and a theoretical promise that the EM algorithm converges to the MLE with probability one when sample size is large enough. Simulation studies confirmed our theoretical analysis as well as the superiority of the proposed methods over existing methods.

REFERENCES

[1] Y. Vardi. Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data. *Journal of the American Statistical Association*, 91(433): 365-377, 1996.

[2] R. Cáceres, N.G. Duffield, J. Horowitz, and D. Towsley. Multicastbased inference of network-internal loss characteristics. *IEEE Transactions on Information Theory*, 45(7): 2462-2480, 1999.

[3] R. Cáceres, N.G. Duffield, S.B. Moon, and D. Towsley. Inference of Internal Loss Rates in the MBone. In *IEEE/ISOC Global Internet'99*, Vol. 3, 1853-1858, 1999.

[4] R. Cáceres, N.G. Duffield, S.B. Moon, and D. Towsley. Inferring linklevel performance from end-to-end multicast measurements. Technical report, University of Massachusetts, 1999.

[5] K. Harfoush, A. Bestavros, and J. Byers. Robust identification of shared losses using end-to-end unicast probes. In Technical Report BUCS-2000-013, Boston University, 2000.

[6] M. Coates and R. Nowak. Unicast network tomography using EM algorthms. Technical Report TR-0004, Rice University, September 2000.

[7] M. Coates and R. Nowak. Network loss inference using unicast end-to-end measurement. In the *ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, 28-1. 2000.

[8] T. Bu, N. Duffield, F.L. Presti, and D. Towsley. Network Tomography on General Topologies. In *Proc. of ACM SIGMETRICS* vol. 30, no. 1, 21-30, 2002.

[9] F. LoPresti, D.Towsley, N.G. Duffield, and J. Horowitz. Multicast topology inference from measured end-to-end loss. *IEEE Transactions on Information Theory*, 48(1): 26-45, 2002.

[10] N. Duffield, J. Horowitz, D. Towsley, W. Wei, and T. Friedman. Multicast-based loss inference with missing data. *IEEE Journal of Selected Areas of Communication*, 20(4): 700-713. 2002.

[11] W. Zhu and Z. Geng. A fast method to estimate loss rates. *Information Networking. Networking Technologies for Broadband and Mobile Networks*, 473-482, 2004.

[12] W. Zhu and Z. Geng. A bottom up inference of loss rate. *Computer Communications*, 28: 351-365, 2005.

[13] N. Duffield, J. Horowitz, F. Presti, and D. Towsley. Explicit loss inference in multicast tomography. *IEEE Transactions on Information Theory*, 52(8): 3852-3855, 2006.

[14] N.G. Duffield, J. Horowitz, F.L. Presti, and D. Towsley. Network delay tomography from end-to-end unicast measurements. *Lecture Notes in Computer Science*, 2170: 576-595, 2001.

[15] F.L. Presti, N.G. Duffield, J. Horwitz, and D. Towsley. Multicast-based inference of network-internal delay distribution. *IEEE/ACM Transactions on Networking*, 10(6): 761-775, 2002.

[16] G. Liang and B. Yu. Maximum pseudo likelihood estimation in network tomography. *IEEE Transactionson on Signal Processing*, 51(8): 2043-2053, 2003.

[17] Y. Tsang, M. Coates, and R. Nowak. Network delay tomography. *IEEE Transactionson on Signal Processing*, 51(8): 2125-2136, 2003.

[18] M.F. Shih and A.O. Hero III. Unicast-based inference of network link delay distributions with finite mixture models. *IEEE Transactionson on Signal Processing*, 51(8): 2219-2228, 2003.

[19] E. Lawrence, G. Michailidis, and V. Nair. Network delay tomography using flexicast experiments. *Journal of the Royal Statistical Society*, Series B, 68(5): 785-813, 2006.

[20] V. Arya, N.G. Duffield, and D. Veitch. Temporal delay tomography. *The 27th Conference on Computer Communications*. IEEE. 2008: 276-280, 2008.

[21] I.H. Dinwoodie and E.A. Vance. Moment estimation in delay tomography with spatial dependence. *Performance Evaluation Archive*, Volume 64, Issue 7-8, 613-628, 2007.

[22] A. Chen, J. Cao, and T. Bu. Network Tomography: Identifiability and Fourier Domain Estimation. *IEEE Transcation on Signal Processing*, 58(12): 6029-6039, 2010.

[23] K. Deng, Y. Li, W. Zhu, Z. Geng, and J.S. Liu. On Delay Tomography: Fast Algorithms and Spatially Dependent Models. *IEEE Transactions on Signal Processing*, 60(11): 5685-5697, 2012.

[24] M.G.H. Bell. The estimation of origin-destination matrices by constrained generalized least squares. *Transportation Research*, Series B 25B(1): 13-22, 1991.

[25] D.D Ortuzar and L.G. Willumsen. Modelling transport. John Wiley and Sons, 2011.

[26] E. Cascetta and S. Nguyen. A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodological*, 22(6): 437-455, 1988.

[27] H. Yang, T. Sasaki, Y. Iida and Y. Asakura. Estimation of origin-destination matrices from link traffic counts on congested networks. *Transportation Research Part B: Methodological*, 26(6): 417-434, 1992.

[28] H.P. Lo, N. Zhang and W.H.K. Lam. Estimation of an origin-destination matrix with random link choice proportions: a statistical approach. *Transportation Research Part B: Methodological*, 30(4): 309-324, 1996.

[29] L. Bianco, G. Confessore, and P. Reverberi. A network based model for traffic sensor location with implications on O/D matrix estimates. *Transportation Science*, 35(1): 50-60, 2001.

[30] C. Tebaldi and M. West. Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93(442): 557-573, 1998.

[31] J. Cao, D. Davis, S. Van Der Viel, B. Yu, and Z. Zu. A scalable method for estimating network traffic matrices from link counts. Technical report, Bell Labs, 41, 2001.

[32] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. The effect of statistical multiplexing on the long range dependence of internet packet traffic. Technical report, Bell Labs. 2002.

[33] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic matrix estimation: existing techniques and new directions. S*IGCOMM Computer Communication Review*, 32(4): 161-174, 2002.

[34] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An information-theoretic approach to traffic matrix estimation. In *Proceedings of SIGCOMM*, 301-312, 2003.

[35] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu. Network tomography: Recent developments. *Statistical Science*, 19(3): 499-517, 2004.

[36] G. Liang, N. Taft, and B. Yu. A fast lightweight approach to origin-destination IP traffic estimation using partial measurements. *IEEE Transactions on Information Theory*, 52(6): 2634-2648, 2006.

[37] J. Fang, Y. Vardi, and C. Zhang. An iterative tomogravity algorithm for the estimation of network traffic. In R. Liu, W. Strawderman, and C. Zhang (Eds.), Complex Datasets and Inverse Problems: Tomography, Networks and Beyond, Volume 54 of *Lecture Notes-Monograph Series*. IMS. 12-23, 2007.

[38] E. M. Airoldi, and A. W. Blocker, Estimating latent processes on a network from indirect measurements. *Journal of the American Statistical Association*, 108(501): 149-164, 2013.

[39] V. Arya, N.G. Duffield, and D. Veitch. Multicast inference of temporal loss characteristics. *Performance Evaluation*, 64(9): 1169-1180, 2007.

[40] M. Rabbat, R. Nowak, and M. Coates. Multiple Source, Multiple Destination Network Tomography. In *Proc. of IEEE Infocom* 04, 1628-1639, 2004.

[41] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, Volume 34, 1-38, 1977.

[42] E.L. Lehmann and G Casella. Theory of Point Estimation. Springer, 2nd edition, 1998.

[43] M.J. Todd. The computation of fixed points and applications. Springer-Verlag, New York, 1976.

[44] The network simulator 2 (ns-2). www.isi.edu/nsnam/ns2

Address of the First author
Yau Mathematical Sciences Center & Center for Statistical Science, Tsinghua University, Beijing 100084, China
E-mail: kdeng@math.tsinghua.edu.cn

Address of the Second author
Department of Statistics, Harvard University, Cambridge, MA 02138, USA
E-mail: yli01@fas.harvard.edu

Address of the Third author
Department of ITEE, ADFA@University of New South Wales, Canberra ACT 2600, Australia
E-mail: w.zhu@adfa.edu.au

Address of the Fourth author
Department of Statistics, Harvard University, Cambridge, MA 02138, USA
E-mail: jliu@stat.harvard.edu