

Estimating Network Loss Rates Using Active Tomography

Bowei XI, George MICHAILIDIS, and Vijayan N. NAIR

Active network tomography refers to an interesting class of large-scale inverse problems that arise in estimating the quality of service parameters of computer and communications networks. This article focuses on estimation of loss rates of the internal links of a network using end-to-end measurements of nodes located on the periphery. A class of flexible experiments for actively probing the network is introduced, and conditions under which all of the link-level information is estimable are obtained. Maximum likelihood estimation using the EM algorithm, the structure of the algorithm, and the properties of the maximum likelihood estimators are investigated. This includes simulation studies using the ns (network simulator) to obtain realistic network traffic. The optimal design of probing experiments is also studied. Finally, application of the results to network monitoring is briefly illustrated.

KEY WORDS: EM algorithm; Inference on graphs; Network modeling; Network monitoring; Network tomography; Probing experiments.

1. INTRODUCTION

The term “network tomography” was introduced by Vardi (1996) to characterize a certain class of inverse problems in computer and communication networks. The goal here, as in medical tomography problems, is to recover higher-dimensional network information from lower-dimensional data. Early work dealt with estimation of the *origin–destination matrix* using passive monitoring (Vardi 1996; Zhang, Roughan, Lund, and Donoho 2003); that is, monitoring agents are placed at internal routers/switches in the network, and aggregate-level data are collected on total traffic flowing into and out of the monitored nodes. Because of the high volume of traffic, origin–destination information cannot be collected at the individual “packet” level. The inverse problem is to recover, from the aggregate data, origin–destination information of the traffic patterns in the network.

Active network tomography refers to a different class of large-scale inverse problems that arise with networks, that is, estimation of quality of service (QoS) parameters such as loss rates and delay distributions at the routers and link-level bandwidths (Castro, Coates, Liang, Nowak, and Yu 2004). Characterizing these parameters is critical for detecting congestion, faults, and other anomalous behavior, ensuring compliance of service-level agreements with users (Coates, Hero, Nowak, and Yu 2003), and management of overlay networks (Chen, Bindel, and Katz 2003). New applications with stringent QoS requirements, such as video conferencing, Internet telephony, and online games, point to an even greater need for fast and efficient algorithms for assessing and responding to changes in network performance.

The traditional approach for characterizing network performance is based on detailed queueing models at the individual router level. But this has become inadequate for capturing overall network performance, because of the complexity and size of modern networks. More importantly, estimating link-level QoS

parameters requires access to the internal links and routers. But the lack of centralized control of modern networks means that Internet service providers typically do not have access to all the nodes of interest, making collection of detailed QoS information at the individual router/link level difficult. Active tomography provides an alternative approach through the use of active “probing,” i.e., sending “probe packets” from a source to one or more receiver nodes located on the periphery of the network. The active tomography problem involves “reconstructing” link-level information from the end-to-end path-level measurements.

This article deals with two aspects of the active tomography problem: design of probing experiments, and estimation of link-level loss rates from end-to-end measurements using these experiments. (A loss occurs when the packet is lost at an internal router, typically due to buffer overflow.) The first part of the article introduces a flexible class of experiments for probing a large network and studies its properties. The second part focuses on estimation of loss rates and related issues.

The article is organized as follows. Section 2 introduces the main elements of the active network tomography problem. Section 3 describes the class of flexible experiments and addresses associated issues of identifiability. Section 4 deals with various aspects of maximum likelihood estimation using the EM algorithm, and Section 5 addresses optimal design of probing experiments in terms of allocation of probes. Section 6 describes a simulation study using the ns-2 (network simulator) to assess the bicast schemes in more realistic environments. The article concludes with an application of the results to network monitoring.

2. FRAMEWORK AND BACKGROUND INFORMATION

2.1 Background

Some relevant facts about networking are briefly summarized here. (See, e.g., Marchette 2001 for more details and a very accessible introduction.) Suppose that one wants to transfer a file from a remote location to the local workstation. The file’s content is broken into pieces, and additional information on origin–destination, reassembly instructions (such as sequence numbers), and error-correcting features are added. These pieces, called *packets*, are transmitted over the network. All of the packets associated with a particular file are referred to

Bowei Xi is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907 (E-mail: sbw@stat.purdue.edu). George Michailidis is Associate Professor, Department of Statistics (E-mail: gmichail@umich.edu), and Vijayan N. Nair is Professor, Department of Statistics and Industrial and Operations Engineering (E-mail: vnn@umich.edu), University of Michigan, Ann Arbor, MI 48109. The authors thank the anonymous reviewers for many useful suggestions and Xiaodong Yang for help with the ARL calculations. They are especially grateful to Professor F. J. Samaniego, the editor, for an extraordinarily helpful review and constructive comments that have substantially changed the manuscript. This research has been supported by National Science Foundation grants DMS-02-04247, CCR-0325571, and DMS-05-05535.

as a “flow.” The origin–destination information is used by the network elements (routers and switches) in conjunction with the Internet protocol (IP), which is primarily responsible for routing packets to their destination, to deliver the packets to the intended recipient. The sequence numbers are crucial to the operation of the transport protocol (e.g., TCP), which also is responsible for regulating the transmission rate within a flow and thus alleviating network congestion. The routers/switches located at the core of the network play a role similar to that of traffic intersections in road networks; namely, they queue up incoming packets and forward them toward their destination along the most effective route. The forwarding of packets at routers follows some scheduling discipline, such as first-come, first-served. Because a queue consists of a physical block of computer memory (finite buffer size), if there are too many incoming packets, then the router may be unable to accommodate some of them and will drop them. Depending on the transmission protocol, senders of dropped packets may be notified to arrange for retransmission. This queuing mechanism is responsible for observed packet losses and, to a large extent, for packet delays. Estimation of link-level delay distributions is another important problem, but we do not consider it here.

Computer and communications networks can be represented by graphs with the nodes corresponding to various computing devices such as workstations, routers, and switches and the edges corresponding to physical links (e.g., fiberoptic cables) connecting the devices. In active tomography, the network is “probed” by actually sending packets from one or more source nodes to a set of receiver nodes. The end-to-end measurements on packet losses, delays, and other attributes are then used to recover the information about performance at the individual links. To make things concrete, consider the graph in Figure 1(a), which shows a small network comprised of workstations located on its periphery and routers at its core and their links. This graph actually depicts an active probing scenario in which packets are sent from the “source” node “0” to the “receiver” nodes on the periphery: 5, 6, 7, 10, 11, 12, 13, 14, and 15.

For the purpose of the probing experiment, the physical topology of the network in Figure 1(a) can be transformed to the *logical topology* in Figure 1(b). Note that this has the structure of a tree: an acyclic graph with one vertex designated as the root. We describe logical topologies in more detail in the next section. Note, however, that the router located between nodes

1 and 3 in Figure 1(a) has disappeared in Figure 1(b). This is because it forms a “chain” (a combination of two links without a split), and the information for the two links cannot be estimated separately. Thus any chains will be collapsed into a single link in the logical topology.

Suppose now that packets are sent from source node 0 in Figure 1 to destination nodes 6, 7, 10, and 11 and the corresponding path-level information on losses is obtained; that is, the losses for the paths 0–6, 0–7, 0–10, and 0–11 are recorded. The goal is to estimate from these end-to-end measurements the link-level parameters of interest for the individual links 0–1, 1–2, 1–4, 2–6, and so on. This is the active network tomography problem.

2.2 Logical Topology and Trees

Most of the literature deals with logical topologies that can be described by trees: acyclic graphs with one vertex designated as the root [see Fig. 2(a)]. We also restrict our attention to single-source topologies in the present article. Some of the same issues also arise with multiple-sources, but multiple-source topologies do present new challenges that we do not deal with here. Note that the logical topology for any active tomography problem with a single source can be represented as a tree. We will also assume, as is commonly done in the literature, that the logical topology is known and fixed during the probing experiment.

We use the following notation. Let $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ denote a tree with root $0 \in \mathcal{V}$, a set of nodes \mathcal{V} , and a set of edges/links \mathcal{E} . A link between nodes i and j is an ordered pair, $(i, j) \in \mathcal{V} \times \mathcal{V}$. The root node 0 represents the source (sender) of the transmitted packets. Let $d(i)$ be a *direct descendant* (child) of node i , and let $\mathcal{D}(i) = \{j \in \mathcal{V} : j = d(i)\}$ denote the set of all direct descendants (children) of node i . [In Fig. 2(a), $\mathcal{D}(1) = \{2, 3, 4, 5\}$.] The set of receiver nodes, denoted by $\mathcal{R} \subset \mathcal{V}$, consists of all nodes without children, that is, $\mathcal{R} = \{i \in \mathcal{V} : \mathcal{D}(i) = \emptyset\}$. [Again, for Fig. 2(a), $\mathcal{R} = \{2, 3, 6, 8, 9, 10, 11, 12, 13, 14, 15\}$.] The set of *internal nodes* \mathcal{I} comprises the nodes that are neither the root nor the receivers (i.e., $\mathcal{I} = \{s \in \mathcal{V} - \{\mathcal{R} \cup \{0\}\}$). We assume throughout that each internal node has at least two children; otherwise, the internal link characteristics (losses) associated with the node and its child cannot be estimated separately.

For each node $i \in \mathcal{V} - \{0\}$, there is a unique node j such that $d(j) = i$. We refer to this as the parent node of i and denote it

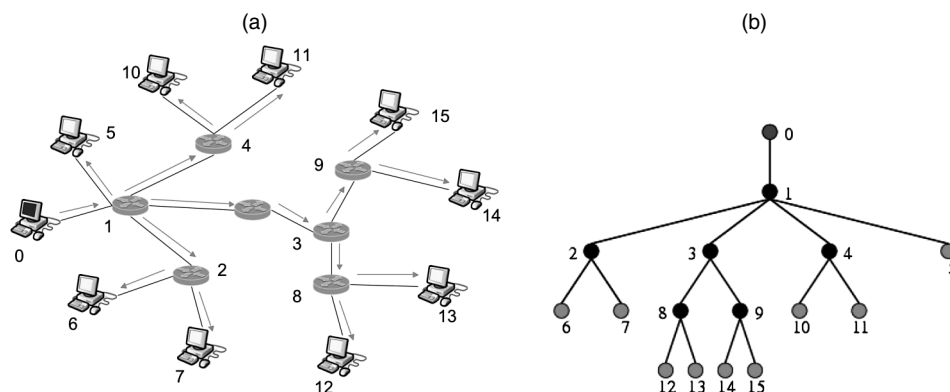


Figure 1. A Layout of a Small Computer Network (a) and the Corresponding Logical Topology of the Network for the Probing Experiment (b).

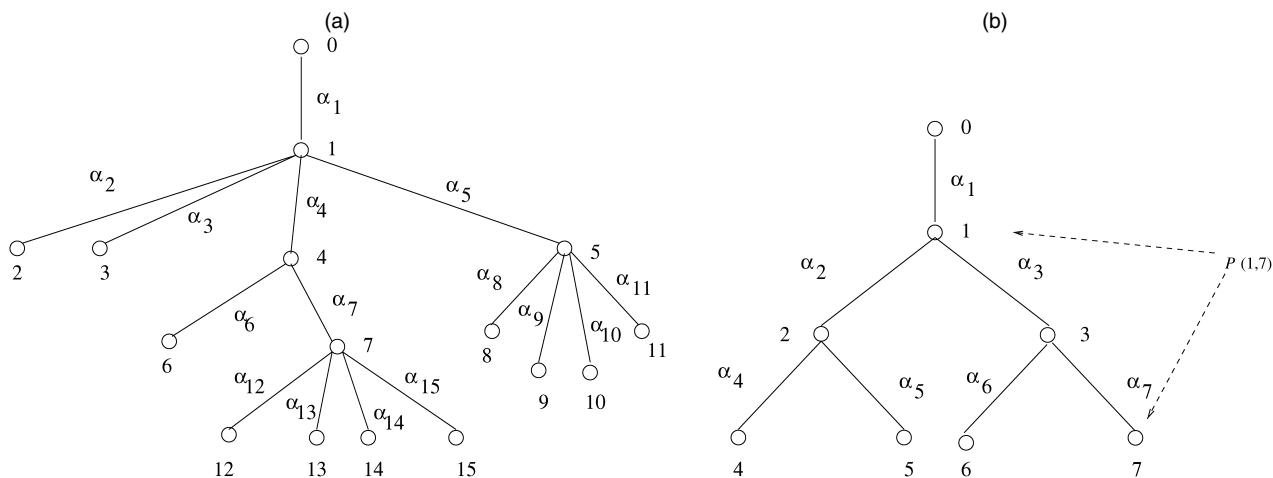


Figure 2. A General Tree Topology (a) and a Three-Layer Symmetric Binary Tree (b).

as $f(i)$. Defining $f^n(i)$ recursively by $f^n(i) = f(f^{n-1}(i))$, we say that i is a descendant of j if $j = f^n(i)$ for some integer $n > 0$. [In Fig. 2, $f(6) = 4$, $f^2(6) = 1$, and $f^3(6) = 0$.] Let $\mathcal{L}_j, j = 1, 2, \dots$, denote the j th layer of a tree, defined as the set of all nodes whose shortest path from the root node 0 has j links; that is, $\mathcal{L}_j = \{i \in \mathcal{V} : 0 = f^j(i)\}$. [In Fig. 2(a), $\mathcal{L}_3 = \{6, 7, 8, 9, 10, 11\}$.] Finally, we let $\mathcal{P}(i, j)$ denote a path between nodes i and j that comprises a set of connected links [see Fig. 2(b)].

We consider *binary trees* extensively in the numerical and simulation sections of this article, because of their simplicity. A binary tree is one in which each internal node has exactly two children, that is, $|\mathcal{D}(i)| = 2$ for all $i \in \mathcal{V} - (\mathcal{R} \cup \{0\})$. For a *symmetric binary tree*, the j th layer has 2^{j-1} nodes, for $j = 1, 2, \dots$. Figure 2(b) shows an example of a three-layer symmetric binary tree.

The size of the networks being studied can vary from local area networks (e.g., a university campus network) involving a few dozen receivers to wide-area networks with several hundred nodes and 10–20 layers. However, the size of the logical topology depends on the resolution that investigators want to achieve. For a coarser look at network performance, several links may be aggregated, whereas for detailed capacity planning, a finer resolution is required.

2.3 Transmission Protocols

There are two types of protocols for transmitting a probe packet from a source node to a specified set of receiver nodes. The most common type is the unicast scheme, which sends a packet from the source to one receiver at a time (Walrand and Varaiya 1999). At the other extreme, the multicast scheme sends a packet to a collection of prespecified receivers simultaneously. For example, consider Figure 1(b), and suppose that the packet needs to be sent to receivers 6, 7, and 12. One packet is sent by the source node to node 1. At this node, the packet is duplicated, and one copy is placed on each of the links going to nodes 2 and 3. At node 2, the packet is further duplicated and sent along to each child node, whereas node 3 sends it on to node 8, which transmits the packet to 12. In the literature, the case in which all of the receiver nodes in a network are probed using a single multicast transmission scheme is called a

multicast experiment. In this article we refer to it instead as an *omnicast* probing experiment, to distinguish the multicast transmission protocol from a multicast experiment. (This distinction is made clear in Sec. 3.) The class of flexible experiments in Section 3 is also based on the multicast protocol.

Some networks have disabled the multicast protocol for security reasons; in these situations the unicast protocol must be relied on. It is known that all of the link-level information cannot be recovered from end-to-end data using just independent unicast probing experiments (Coates and Nowak 2000). The higher-order correlation information present in multicast schemes) is critical for recovering link-level information. This has led to the proposed back-to-back unicast protocol, which seeks to mimic the multicast scheme by sending unicast probes spaced very close together in time to several receivers (Coates and Nowak 2000; Nowak 2001; Castro et al. 2004). Usually, this involves just one pair of receivers at a time. If the pair of probes are sent back-to-back within nanoseconds of each other, then the probes likely will experience identical network conditions on the common links. In this case, back-to-back unicast will mimic a multicast (specifically, a bicast) scheme.

In this article we consider this idealized back-to-back scheme to be interchangeable with the multicast protocol. However, it is important to keep in mind that the back-to-back probes may not always experience the same losses in shared links, especially if the shared path has many links. Moreover, if the back-to-back probes are sent to all of the receivers in a large network (mimicking a multicast scheme to all receivers), then there can again be differences in the performance of the shared links. Lo Presti, Paxon, and Towsley (2001) proposed using striped probes to improve the correlation among back-to-back unicasts. In ongoing work, we are formally investigating the properties of such back-to-back schemes using a latent-variable temporal model.

2.4 Stochastic Model

Let $Z_r(m) = 1$ if the m th probe packet sent from the root node 0 reached receiver node $r \in \mathcal{R}$, and 0 otherwise. For a one-cast (unicast) scheme, the root node transmits packets $m = 1, 2, \dots$ to one receiver at a time, so we observe only $Z_r(m)$

1 for a single receiver r for each probe packet. For an omnicast
2 scheme, where the sender transmits each packet simultaneously
3 to all receiver nodes, the observed outcome for the m th probe
4 packet consists of $Z_r(m)$ for all $r \in \mathcal{R}$.

5 Define hypothetical random variables $X_i(m)$ associated with
6 all of the links in the network as the outcome of the probe
7 sent to node i from its parent $f(i)$, with $X_i(m) = 1$ if the
8 packet traverses link $i \in \mathcal{E}$ successfully and 0 otherwise. We
9 analyze the data under the following independence model,
10 which also has been commonly used in the literature (Caceres,
11 Duffield, Horowitz, and Towsley 1999). We assume through-
12 out that the $X_i(m)$'s are independent across i and m . Let
13 $\alpha_i(m) = P(X_i(m) = 1)$. We also assume temporal homogene-
14 nity, that is, $\alpha_i(m) \equiv \alpha_i$ for all probes m . Then $P(Z_r(m) = 1) =$
15 $\prod_{s \in \mathcal{P}(0,r)} \alpha_s$. Further, $P(X_j(m) = 1 \forall j \in \mathcal{D}(i)) = \prod_{s \in \mathcal{P}(0,i)} \alpha_s \times$
16 $\prod_{j \in \mathcal{D}(i)} \alpha_j$.

17 These assumptions have also been used in the network en-
18 gineering literature (Caceres et al. 1999; Castro et al. 2004;
19 Coates et al. 2002). The temporal homogeneity assumption is
20 not critical, because the time frame for the probing experiment
21 is on the order of minutes, but the effect of spatial dependence
22 merits further study. Extensions to situations with spatiotem-
23 poral dependence will be considered in future work. We do, how-
24 ever, consider a limited assessment of the assumptions using the
25 ns simulator in Section 6.

26 Work has been done on the estimation of link-level param-
27 eters from active probing schemes. Caceres et al. (1999) con-
28 sidered multicast experiments (omnicast experiments in our
29 terminology here) and developed estimation methods that are
30 asymptotically equivalent to the maximum likelihood estimator
31 (MLE) for loss rates. Unfortunately, this method does not ex-
32 tend to the flexible experiments considered herein. Moreover,
33 these estimators can fall outside the range of $(0, 1)$ in finite
34 samples (see Sec. 4.2). Coates and Nowak (2000) considered
35 maximum likelihood estimation using the EM algorithm for
36 link losses but under back-to-back unicast probing. The prob-
37 lem of estimating link-level delay distributions has also been
38 studied (see, e.g., Lo Presti, Duffield, Horowitz, and Towsley
39 2002; Liang and Yu 2003; Tsang, Coates, and Nowak 2003).

3. A CLASS OF FLEXIBLE PROBING EXPERIMENTS

40 Although the omnicast experiment is conceptually simple, it
41 has several drawbacks. First, the number of possible outcomes
42 in the experiment increases exponentially with the number of
43 layers in the tree topology. For example, consider a symmet-
44 ric binary tree with L layers with $R = 2^{L-1}$ receiver nodes.
45 The omnicast scheme corresponds to a multinomial experiment
46 of dimension R , so there are $2^R = 2^{2^{L-1}}$ possible outcomes.
47 Thus data complexity will be a major problem with large net-
48 works. More importantly, network service providers rarely want
49 to probe the entire network with the same degree of intensity. It
50 is more common to allocate different levels of probing effort to
51 different regions of the network at different times. In network
52 monitoring, for example, the goal is to monitor the network reg-
53 ularly and study regions of the network in which problems oc-
54 cur. This calls for a more flexible class of probing experiments
55 that allows for studying different regions of the network with
56 varying intensities. Such experiments raise interesting ques-
57 tions about how to design them, when the experiments will lead
58 to identifiability of all of the link-level parameters, how to com-
59 bine the data to estimate all of the parameters, and so on.

3.1 Flexible Experiments

60 We begin with a description of a k -cast scheme. A k -cast
61 scheme sends a probe simultaneously to a given subset k of the
62 receivers in \mathcal{R} and is completely specified by the k -tuple of re-
63 ceiver nodes, $\langle r_1, r_2, \dots, r_k \rangle$, $r_j \in \mathcal{R}$, $j = 1, \dots, k$. For example,
64 two possible four-cast schemes for the general tree topology in
65 Figure 2(a) are $\langle 12, 13, 14, 15 \rangle$ and $\langle 2, 3, 6, 12 \rangle$.
66

67 The class of flexible probing experiments, denoted generi-
68 cally as \mathcal{C} , is given by a collection of independent schemes
69 $\{C_h, N_h\}$, where C_h is a k_h -cast scheme for $1 < k_h < R$, with R
70 the total number of receiver nodes, N_h the number of probes al-
71 located to C_h , and $h = 1, \dots, H$ the number of k_h -cast schemes
72 used. In practice, k_h can (and often will) be different for dif-
73 ferent C_h . Throughout, let $N = \sum_h N_h$ denote the total number
74 of probes for the experiment. This class of experiments allows
75 us to allocate different numbers of probes N_h to schemes C_h .
76 Thus different parts of the network can be probed with different
77 intensities and possibly at different times. We can combine the
78 end-to-end measurements from all of the schemes to estimate
79 the link-level parameters as well as continuously update the es-
80 timates of the QoS parameters based on the data over time.

81 If $k = 1$ for all C_h , then \mathcal{C} is composed of a collection of
82 unicast schemes. If $k = R$, then \mathcal{C} corresponds to a single omni-
83 cast experiment (Caceres et al. 1999). As we discuss later, an
84 efficient experiment from a computational standpoint is a “min-
85 imal” experiment based on a collection of *bicast* ($k = 2$) and
86 unicast ($k = 1$) probing schemes.

87 Note that a probe packet for a k -cast scheme has 2^k possible
88 outcomes, each of dimension k . These correspond to whether
89 the outcome for the receiver node is 1 or 0 (whether or not the
90 node receives the transmission). For example, for the four-cast
91 scheme $\langle 12, 13, 14, 15 \rangle$ in Figure 2, there are 16 possible out-
92 comes, with the outcome $(Z_{12} = 0, Z_{13} = 1, Z_{14} = 0, Z_{15} = 1)$
93 indicating that the packet was successfully received by receivers
94 13 and 15 but not by receivers 12 and 14. If we send N_h probes
95 using this k -cast scheme, then, under the posited stochastic
96 model, it leads to a multinomial experiment with 2^k outcomes.
97 The “success” probability for each outcome is a complex func-
98 tion of the underlying link success rates α 's. For example, the
99 probability of the event $(Z_{12} = 0, Z_{13} = 1, Z_{14} = 0, Z_{15} = 1)$ is
100 given by a sum of products of α_i 's and $(1 - \alpha_i)$'s. We discuss
101 this in more detail in Section 5.1.

102 The experiment \mathcal{C} is then just a collection of these indepen-
103 dent multinomial experiments. The data complexity of the ex-
104 periment \mathcal{C} is dictated by that of the largest k_h -cast scheme C_h
105 in \mathcal{C} . Typically, this will be much smaller than that of the omni-
106 cast experiment corresponding to the entire network ($k_h = R$).
107

3.2 Identifiability

108 A natural question now is whether for a given experiment \mathcal{C}
109 all of the internal link-level parameters can be identified. We
110 already know that the answer is negative for the collection of
111 unicast experiments. In this section we characterize necessary
112 and sufficient conditions for identifiability of all of the link-
113 level parameters.
114

115 We need the notion of a splitting node. First consider a two-
116 cast (or bicast) scheme with receiver nodes $\langle r_1, r_2 \rangle$. Then the
117 internal node s is a splitting node if $\mathcal{P}(0, s)$ is the longest com-
118 mon path that $\{r_1, r_2\}$ share on the tree. For a k -cast scheme, the

splitting nodes can be defined in terms of the splitting nodes of pairs of receiver nodes. Consider the k -cast scheme with receiver nodes $\langle r_1, r_2, \dots, r_k \rangle$, and let $\{r_i, r_j\}$ be any subset of them. Then the internal node $s \equiv s(r_i, r_j)$ is a *splitting node* for this particular pair if $\mathcal{P}(0, s)$ is the longest common path that $\{r_i, r_j\}$ share on the tree.

Note that the number of splitting nodes for a k -cast scheme can range from 1 to $(k - 1)$. For example, for the four-cast scheme $\langle 12, 13, 14, 15 \rangle$ in Figure 2(a), there is only one splitting node, 7. But there are three splitting nodes (1, 4, and 7) for the scheme $\langle 2, 3, 6, 12 \rangle$. We are interested mainly in k casts with a single split.

Proposition 1. We assume that $\alpha_j > 0$ for all links in the logical topology. Let \mathcal{C} be a probing experiment comprising a collection of schemes $\{C_h, N_h\}$ with $N_h \geq 1$. The experiment \mathcal{C} identifies the parameters $\vec{\alpha}$ if and only if the following conditions are satisfied: (I) every internal node s in the tree topology corresponds to a splitting node for some scheme $C_h \in \mathcal{C}$ with $k_h \geq 2$, and (II) all of the receiver nodes in the tree are covered by \mathcal{C} .

The proof is deferred to the Appendix. It proceeds by mapping any given experiment to an equivalent one involving a collection of bicast and unicast schemes and proving the result for this case.

To understand the implications of Proposition 1, consider the tree topology in Figure 2(a). Suppose that we use the experiment \mathcal{C} comprising the schemes $C_1 = \langle 2, 3 \rangle$, $C_2 = \langle 6 \rangle$, $C_3 = \langle 12, 13, 14, 15 \rangle$, and $C_4 = \langle 8, 9, 10, 11 \rangle$ with $N_h \geq 1$ for all C_h . The internal nodes 1, 7, and 5 are splitting nodes for $\langle 2, 3 \rangle$, $\langle 12, 13, 14, 15 \rangle$, and $\langle 8, 9, 10, 11 \rangle$; however, 4 is not a splitting node, and so this experiment will not identify all of the parameters. In particular, the unicast experiment $C_2 = \langle 6 \rangle$ can estimate the entire path parameter $\pi(0, 6)$ but cannot separate the individual link parameters α_4 and α_6 . The problem can be rectified by replacing, for example, C_2 and C_3 with $C'_2 = \langle 6, 12 \rangle$ and $C'_3 = \langle 13, 14, 15 \rangle$. Of course, there are many ways of modifying this to identify all of the parameters.

This example also illustrates the advantage of this class of schemes. We can probe the different subnetworks C_1 , C_2 , C'_3 , and C'_4 separately with different intensities, even at different times, and combine the results to characterize the overall network behavior. The subnetworks are much smaller and can be studied more easily. Note, however, that the individual subnetworks by themselves do not allow for estimation of all of the internal link-level parameters within each, so this cannot be viewed as four separate problems.

An experiment comprising collection bicast and unicast schemes has the least data complexity, because the complexity is no more than that of a bicast scheme that yields a multinomial experiment with four possible outcomes. For such a collection, *minimal* experiments (i.e., smallest collections) can be found that lead to identifiability of all of the internal link parameters as follows:

1. For each internal node s , use exactly *one* bicast pair b , whose splitting node is s .
2. Choose these bicast pairs to maximize the number of receiver nodes covered.

3. Choose unicast schemes to cover the remaining receiver nodes, $r \in \mathcal{R}$ not covered by the bicast pairs.

As an illustration, consider the binary tree in Figure 2(b). A minimal experiment consists of the bicast pairs $C_1 = \langle 4, 5 \rangle$, $C_2 = \langle 6, 7 \rangle$, and $C_3 = \langle 5, 6 \rangle$. This is not unique, however, because we could replace C_3 with $C'_3 = \langle 4, 7 \rangle$.

Note that Proposition 1 provides a simple and elegant characterization of the identifiability condition. It can also be used to construct an experiment \mathcal{C} that satisfies the identifiability condition by formulating it as a set-covering problem (Chvatal 1979).

Finally, the identifiability result in Proposition 1 is also useful for the back-to-back unicast transmission protocols used in the literature (Nowak 2001; Castro et al. 2004). (The use of back-to-back unicast schemes in the literature has been limited to pairs of receiver nodes, because this is the most reasonable scenario. Back-to-back transmissions to many receivers at a time is unlikely to mimic the multicast protocol well, because the probes may not see the same environment on the common links due to the time delay between many probes.) There is no discussion in the literature on the design of back-to-back unicast experiments. Questions of interest include whether they should be sent to all possible pairs and whether there is a subset of the pairs that will be sufficient to ensure identifiability of all the internal link parameters and, if so, how these should be chosen. Proposition 1 and the ensuing discussion about minimal bicast/unicast experiments answer all of these questions. In particular, we see that send back-to-back probes need to be sent to only subset of all possible pairs to cover all of the internal nodes, and any remaining nodes can be covered by just unicasts. Thus the results in this section are also useful for designing back-to-back unicast experiments.

4. MAXIMUM LIKELIHOOD ESTIMATION

4.1 The Likelihood Function

As noted earlier, the experiment $\mathcal{C} = \{C_h, N_h\}$ comprises a collection of independent schemes C_h with number of probes N_h . In the remainder, we assume that \mathcal{C} satisfies the identifiability condition of Proposition 1. Recall that a k -cast scheme can be viewed as a k -dimensional multinomial experiment with parameters that depend on the link transmission rates, α_i . To see this more clearly, denote the probability of a successful transmission of a packet over the path $\mathcal{P}(s, u)$ by $\pi(s, u)$; therefore,

$$\pi(s, u) = \prod_{\ell \in \mathcal{P}(s, u)} \alpha_\ell. \quad (1)$$

Consider the simple case of bicast probes, and suppose that a bicast probe b is sent to the pair of receiver nodes $\langle i_b, j_b \rangle$, $i_b, j_b \in \mathcal{R}$, with splitting node s_b . The observed outcome can take one of the following four values: $(Z_{i_b}(t), Z_{j_b}(t)) = (0, 0)$, $(0, 1)$, $(1, 0)$, or $(1, 1)$, depending on whether the packet was received by none, one, or both of the intended receivers. Let γ_{ij} denote the corresponding probability of any of these events. Then

$$\gamma_{11} = \pi(0, s_b)\pi(s_b, i_b)\pi(s_b, j_b), \quad (2)$$

$$\gamma_{10} = \pi(0, s_b)[1 - \pi(s_b, i_b)]\pi(s_b, j_b), \quad (3)$$

$$\gamma_{01} = \pi(0, s_b)\pi(s_b, i_b)[1 - \pi(s_b, j_b)], \quad (4)$$

and

$$\gamma_{00} = [1 - \pi(0, s_b)] + \pi(0, s_b)[1 - \pi(s_b, i_b)][1 - \pi(s_b, j_b)]. \quad (5)$$

The corresponding bicast experiment has $2^2 = 4$ possible outcomes, $N_{1,1}$, $N_{1,0}$, $N_{0,1}$, and $N_{0,0}$, with probabilities given by the foregoing corresponding γ 's. Because $\gamma_{1,1} + \gamma_{1,0} + \gamma_{0,1} + \gamma_{0,0} = 1$, there are only three free probabilities in the bicast scheme b . Also note that $\pi(0, i_b) = \gamma_{1,1} + \gamma_{1,0}$ and $\pi(0, j_b) = \gamma_{1,1} + \gamma_{0,1}$. Analogous expressions can be derived for k -cast schemes.

For a general scheme C_h with k receiver nodes $\langle r_{1,h}, \dots, r_{k,h} \rangle$, let $N_{(i_1, \dots, i_k), h}$ denote the number of outcomes corresponding to the event $\{i_1, \dots, i_k\}$, where $i_j = 1$ means that receiver $r_{j,h}$ received the packet and 0 means that it did not. Let $\gamma_{(i_1, \dots, i_k), h}$ denote the corresponding probability of this event. Then the overall log-likelihood of the experiment \mathcal{C} is just the sum of the individual likelihoods of the C_h 's and is given (up to additive constants) by

$$\log(\Lambda(\mathbf{N}|\vec{\alpha})) = \sum_{C_h \in \mathcal{C}} \sum_{i_1, \dots, i_k} N_{(i_1, \dots, i_k), h} \log(\gamma_{(i_1, \dots, i_k), h}). \quad (6)$$

The parameters $\gamma_{(i_1, \dots, i_k), h}$ are complicated functions of the underlying α 's. To understand the complexity, consider the score functions for a k -cast scheme C_h with a single split. As before, let $C_h = \langle r_{1,h}, \dots, r_{k,h} \rangle$, with s_h as the splitting node. There are two cases to consider.

Case 1. Links in the path above the split. For node $k \in \mathcal{P}(0, s_h)$,

$$\frac{\partial \log \Lambda_h}{\partial \alpha_k} = \frac{1}{\alpha_k} \left[(N_h - N_{(0, \dots, 0), h} + N_{(0, \dots, 0), h} \frac{\gamma_{(0, \dots, 0)h} - 1}{\gamma_{(0, \dots, 0)h}}) \right].$$

Case 2. Links in the path above the split. For node $k \in \mathcal{P}(s_h, r_j)$,

$$\begin{aligned} \frac{\partial \log \Lambda_h}{\partial \alpha_k} &= \frac{1}{\alpha_k} \left[N_{1+, r_j, h} - N_{0, r_j, h} \frac{\pi(s_h, r_j)}{1 - \pi(s_h, r_j)} \right. \\ &\quad \left. - N_{0, \dots, 0} \frac{\pi(0, s_h) \prod_{i \neq j} (1 - \pi(s_h, r_i)) \pi(s_h, r_j)}{\gamma_{0, \dots, 0}} \right], \end{aligned}$$

where $N_{1+, r_j, h}$ is the sum of all outcomes where r_j has a 1 and $N_{0, r_j, h}$ is the sum of all outcomes where r_j has a 0 and at least one of the remaining receivers has a 1.

We see that the likelihood equations are involved and cannot be solved analytically to get explicit expressions for the MLE in general. We resort to iterative optimization methods for maximizing the likelihood. The EM algorithm has been found to be useful in the literature (Coates and Nowak 2000; Coates et al. 2002; Castro et al. 2004; Liang and Yu 2003), because this falls naturally into the class of missing-data problems.

4.2 The EM Algorithm

According to the posited model, we have $Z_r(m) = \prod_{i \in \mathcal{P}(0, r)} X_i(m)$, with $\alpha_i = P(X_i(m) = 1)$ for all m . The end-to-end measurements, Z_r 's, are observed, but the internal link-level data $X_i(m)$ are not. Thus the collection $\{X_i(m); i \in \mathcal{T}, m = 1, 2, \dots\}$ can be treated as the unobserved complete data, and the EM algorithm (Dempster, Laird, and Rubin 1977) can be used to compute the MLEs.

Let $V_i = \sum_{m=1}^N X_i(m)$, the total number of "successes" at node i under the hypothetical experiment. Then the complete-data likelihood is given by

$$\Lambda(\vec{\alpha} | V_1, \dots, V_E) \propto \prod_{i \in \mathcal{V}} \alpha_i^{V_i} (1 - \alpha_i)^{N - V_i}. \quad (7)$$

This complete-data likelihood function is based on multinomial experiments that involve the α_i 's directly. It can be maximized easily to obtain the MLEs. It belongs to the exponential family, so the E-step involves computing the conditional expectation of the V_i 's, the complete data-sufficient statistics, given the observed data and current values of the parameters. The general expression for $(t + 1)$ st iteration of the algorithm is given as follows:

E-step. For every scheme $C_h \in \mathcal{C}$ and node $i \in \mathcal{V}_h$, compute the conditional expectations given the observed end-to-end data \mathcal{N}_h ,

$$\begin{aligned} V_{i,h}^{(t+1)} &= E_{\vec{\alpha}^{(t)}} \left(\sum_m I\{X_{i,h}(m) = 1\} \mid \mathcal{N}_h \right) \\ &= N_h - E_{\vec{\alpha}^{(t)}} \left(\sum_m I\{X_{i,h}(m) = 0\} \mid \mathcal{N}_h \right). \end{aligned}$$

M-step.

$$\alpha_i^{(t+1)} = \frac{\sum_{C_h \in \mathcal{C}} V_{i,h}^{(t+1)}}{\sum_{C_h \in \mathcal{C}} N_h}.$$

It clearly would be useful to obtain explicit expressions for the E- and M-steps. We develop these here for the important special case where the k -cast schemes have a single splitting node (which is the most interesting case for our flexible experiments). The situation is conceptually analogous for schemes with multiple splitting nodes, but the notation becomes messy because the form of the E-step depends on the exact form of the tree topology.

Let s_h be the splitting node for scheme $C_h = \langle r_{1,h}, r_{2,h}, \dots, r_{k,h} \rangle$. The $k + 1$ path probabilities for this scheme, $\pi(0, s_h)$, $\pi(s_h, r_{1,h}), \dots, \pi(s_h, r_{k,h})$, can be obtained from (1). Further, let $N_{(0, \dots, 0), h}$ denote the number of probes corresponding to the outcome of 0 for all of the receiver nodes $r_{1,h}, r_{2,h}, \dots, r_{k,h}$, and let $\gamma_{(0, 0, \dots, 0), h}$ be the corresponding probability. Finally, let $N_{0, r_j, h}$ denote the number of probes corresponding to the event that 0 is observed at receiver node $r_{j,h}$ and at least one of the remaining receiver nodes receives a 1. Starting with an initial value $\vec{\alpha}^{(0)}$, let $\vec{\alpha}^{(t)}$ be the value after the t st iteration. Then we can write the $(t + 1)$ st iteration of the E-steps as follows.

For each scheme $C_h \in \mathcal{C}$, proceed as follows:

1. Use $\vec{\alpha}^{(t)}$ and (1) to get the updated path probabilities.

2. Compute $V_{\ell,h}^{(t+1)} \equiv E_{\tilde{\alpha}^{(t)}}[V_\ell | \mathcal{N}_h]$ as follows:

For link $\ell \in \mathcal{P}(0, s_h)$,

$$V_{\ell,h}^{(t+1)} = N_h - N_{(0,\dots,0),h} \frac{1 - \alpha_\ell^{(t)}}{\gamma_{(0,\dots,0),h}^{(t)}}.$$

For link $\ell \in \mathcal{P}(s_h, r_{j,h})$,

$$\begin{aligned} V_{\ell,h}^{(t+1)} &= N_h - N_{0,r_j,h} \frac{1 - \alpha_\ell^{(t)}}{1 - \pi^{(t)}(s_h, r_{j,h})} \\ &\quad - N_{(0,\dots,0),h} \\ &\quad \times \left((1 - \alpha_\ell^{(t)}) \left[1 - \pi^{(t)}(0, s_h) \right. \right. \\ &\quad \left. \left. + \pi^{(t)}(0, s_h) \prod_{\{i: i \neq j\}} (1 - \pi^{(t)}(s_h, r_{i,h})) \right] \right) \\ &\quad \times (\gamma_{(0,\dots,0),h}^{(t)})^{-1}. \end{aligned}$$

In a bicast scheme $b = \langle i_b, j_b \rangle$, if the path $\mathcal{P}(0, s_b)$ consists of only the single link ℓ under consideration, then $\alpha_\ell = \pi(0, s_b)$. The same holds for $\mathcal{P}(s_b, i_b)$ and $\mathcal{P}(s_b, j_b)$. In these cases, some of the foregoing calculations will simplify. For example, consider the binary symmetric three-layer tree given in Figure 2(b) together with a minimal bicast experiment consisting of the three pairs $\langle 4, 5 \rangle$, $\langle 6, 7 \rangle$, and $\langle 5, 6 \rangle$. Then, $V_{4,(4,5)}^{(t+1)}$ simplifies to

$$V_{4,(4,5)}^{(t+1)} = N^{4,5} - N_{0,1}^{4,5} - N_{(00),(4,5)} \frac{(1 - \alpha_4^{(t)})(1 - \pi^{(t)}(0, 5))}{\gamma_{(00),(4,5)}^{(t)}}$$

and

$$\gamma_{(00),(4,5)}^{(t)} = [1 - \alpha_1^{(t)} \alpha_2^{(t)}] + \alpha_1^{(t)} \alpha_2^{(t)} (1 - \alpha_4^{(t)})(1 - \alpha_5^{(t)}).$$

Further,

$$\alpha_4^{(t+1)} = \frac{V_{4,(4,5)}^{(t+1)}}{N_{(4,5)}},$$

because this is the only pair in the minimal bicast that includes node 4.

Caceres et al. (1999) developed a clever algorithm for computing approximate MLEs for loss rates for an omnicastr experiment. The basic idea is to reduce the data to sufficient statistics and obtain explicit expressions for solving the likelihood equations. If a node has k children, then the equation involves solving a polynomial of order $(k - 1)$. For symmetric binary trees, this reduces to linear equations. These estimates solve the likelihood equations and hence are asymptotically equivalent to the MLEs.

It does not appear that this algorithm can be generalized to the class of flexible experiments considered in this article. Moreover, there are situations in which the approximate estimator can behave poorly, leading to estimates outside the range of $(0, 1)$. This seems to occur when there is considerable variability in the link loss rates, with some loss rates being very small. This point was already noted by Caceres et al. (1999). To see this, consider a three-layer tree with $\alpha_1 = \alpha_2 = \alpha_4 = \alpha_5 = .8$, $\alpha_3 = .05$, and $\alpha_6 = \alpha_7 = .2$. The first two rows of Table 1 shows the results from an omnicastr experiment with 400

Table 1. Comparison of Omnicastr and Bicastr MLEs

	α_1	α_2	α_3	α_4	α_5	α_6	α_7
Omnicastr $\hat{\alpha}_{\text{approx MLE}}$	1.3250	.5223	.0226	.8056	.7803	.2500	.3333
Omnicastr $\hat{\alpha}_{\text{MLE}}$	1.0000	.6921	.0300	.8056	.7803	.2500	.3333
Bicastr $\hat{\alpha}_{\text{MLE}}$.8310	.7576	.0521	.7796	.7727	.2142	.2266

probes. The first row shows the approximate MLEs obtained using the algorithm of Caceres et al. (1999). In this case the approximate MLE does a poor job of estimating α_1 and α_2 . The second row shows the omnicastr MLEs obtained through the EM algorithm. This was computationally expensive, taking more than 1,200 iterations to compute these MLEs. Although the MLE for α_1 lies inside the range $(0, 1)$, it also does poorly in estimating α_1 and α_2 . Recall that the true value is .8.

The third row gives the results from an experiment with four bicastrs: $\langle 4, 5 \rangle$, $\langle 6, 7 \rangle$, $\langle 5, 6 \rangle$, and $\langle 4, 7 \rangle$. This experiment allocated more probes to links in which the loss rate is high, to estimate them more precisely. Specifically, 600 probes were sent to pair $\langle 6, 7 \rangle$, 160 probes were sent each to pair $\langle 5, 6 \rangle$ and pair $\langle 4, 7 \rangle$, and only 40 probes were sent to pair $\langle 4, 5 \rangle$. The expected amount of total traffic under this scheme is 1,212, only slightly larger than that under the omnicastr experiment. The bicastr experiment did a much better job estimating the link loss rates. This situation provides another example of the flexibility and advantages of the experiments proposed in this article.

4.3 Convergence and Computational Complexity of the Algorithm

General convergence properties of the EM algorithm are well known (see, e.g., Tanner 1996; Wu 1983). It does not appear that the log-likelihood is strictly concave in our case, and so the uniqueness of the MLE is not easy to establish. However, we have studied this problem numerically for many datasets and encountered no problems with multiple maxima. Proposition 2 shows that the Fisher information matrix is positive definite. This establishes that, with probability tending to 1, there will be a unique maximum at least in local neighborhoods around the true value $\tilde{\alpha}_0$.

Denote by $I_N(\mathcal{C}, \tilde{\alpha}_0)$ the Fisher information matrix at the true value $\tilde{\alpha}_0$. The following result is proved in the Appendix.

Proposition 2. $I_N(\mathcal{C}, \tilde{\alpha}_0)$ is a finite and positive-definite matrix.

Let $\tau_h = N_h/N$, the proportion of probe size allocated to the scheme C_h . Then we can write $I_N(\mathcal{C}, \tilde{\alpha}_0) = N \sum_{C_h \in \mathcal{C}} \tau_h \times I(C_h, \tilde{\alpha}_0)$, where $I(C_h, \tilde{\alpha}_0)$ is the (normalized) information for the scheme C_h (i.e., corresponding to a probe of size $N_h = 1$). The individual elements of $I(C_h, \tilde{\alpha}_0)$ can be computed as the variance-covariance matrix of the score functions (given in Sec. 4.1) or the expectation of the negative second derivatives of the observed data log-likelihood. A scheme C_h will typically involve only a few of the link-level parameters, so many of the entries in $I(C_h, \tilde{\alpha}_0)$ will be 0, leading to a sparse matrix.

Figure 3 shows the number of iterations needed for convergence of the likelihood function and convergence of selected $\hat{\alpha}$'s for a symmetric binary three-layer tree with all elements of $\tilde{\alpha} > .6$. About 50 iterations are needed for a convergence criterion of 10^{-4} for the log-likelihood. Using a reasonable initial

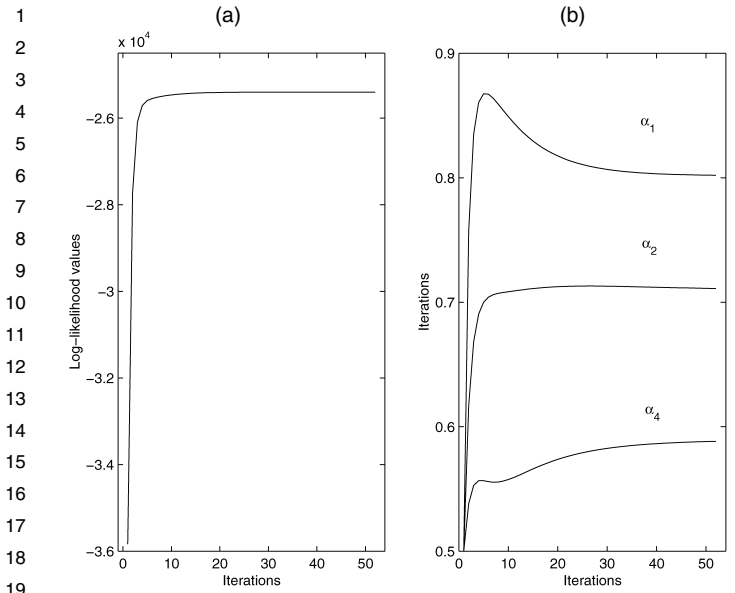


Figure 3. Convergence of the Log-Likelihood Function (a) and of Selected $\hat{\alpha}$'s (b) for a Three-Layer Tree.

estimate of $\bar{\alpha}$ significantly reduces the number of iterations, especially for fairly small values of α 's. As we show in the next section, the variability of $\hat{\alpha}_i$ increases as α_i gets smaller. This affects the number of iterations, which increases as the loss rate $(1 - \alpha_i)$ increases (Fig. 4).

The computational complexity of the algorithm depends in general on the structure of the individual schemes and the underlying tree topology and hence is difficult to assess. However, in the special case of a *minimal* experiment comprising bicast and unicast schemes, we can establish a lower bound for any given tree topology \mathcal{T} .

Let $L = |\mathcal{L}|$, denote the number of layers and let $E = |\mathcal{E}|$ denote the number of links in \mathcal{T} . Note that most of the computations stem from the E-step. Consider the complexity of one iteration of the EM algorithm for an arbitrary bicast pair $b = \langle i_b, j_b \rangle$ with splitting node s_b . The path probabili-

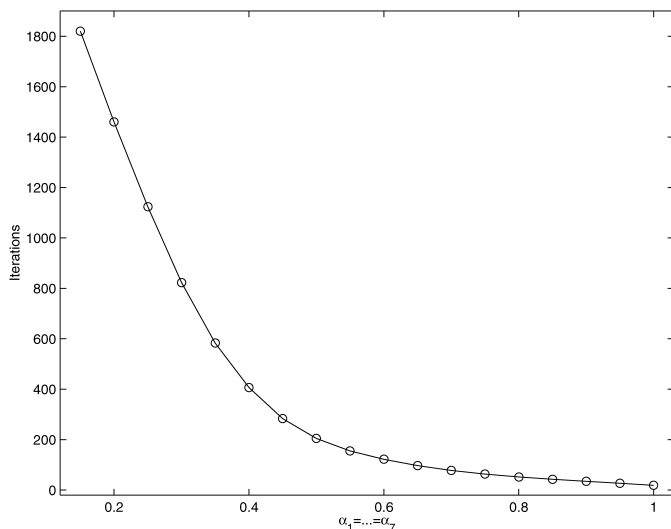


Figure 4. Number of EM Iterations for a Three-Layer Tree as a Function of the Success Probability.

ties given in (1)–(5) need to be computed at the beginning of the k th iteration. This involves $\mathcal{O}(Q_b)$ multiplications and a fixed number of additions/subtractions, where Q_b denotes the maximum length of paths $\mathcal{P}(0, i_b)$ and $\mathcal{P}(0, j_b)$, that is, $Q_b = \max\{|\mathcal{P}(0, i_b)|, |\mathcal{P}(0, j_b)|\}$. Note that $Q_b < L$ for any bicast scheme. At the second stage, the updates of $V_{\ell, b}$ need to be computed, which involves a large but constant number of operations. Therefore, $\mathcal{O}(L)$ operations are required in the E-step for the bicast schemes. For an arbitrary unicast scheme with receiver node u , only the path probability $\pi^{(k)}(0, u)$ needs to be calculated; this also requires at most $\mathcal{O}(L)$ operations. The second stage of updating $V_{\ell, u}^{(k+1)}$ involves a constant number of operations. Finally, the M-step involves a single division for each α_i for the whole experiment.

Therefore, the complexity of the minimal experiment is given by $\mathcal{O}(|\mathcal{B}| + |\mathcal{U}|L)$. Minimal experiments require $|\mathcal{I}|$ bicast pairs, whereas the number of unicast schemes is bounded by $|\mathcal{R}|$. Therefore, the lowest possible complexity is $\mathcal{O}(E \times L)$. The relationship between E and L depends on the structure of the topologies. For the special case of symmetric binary trees, $L = \log(E)$.

4.4 Behavior of the Variances

This section studies how the behavior of the variance of the MLEs varies with the true loss rates and the layer of the links in the tree. We consider just the three-layer symmetric binary tree Figure 2(b) with equal loss rates for all links, that is, $\alpha_1 = \dots = \alpha_7$. Figure 5 shows the variances of the MLEs for a bicast experiment with an equal allocation of 25% to the four bicast: $C_1 = \langle 4, 5 \rangle$, $C_2 = \langle 6, 7 \rangle$, $C_3 = \langle 5, 6 \rangle$, and $C_4 = \langle 4, 7 \rangle$.

Unlike a binomial experiment in which the variance is proportional to $\alpha(1 - \alpha)$, the variance here increases as α gets smaller. Thus there is a higher level of uncertainty when a link has high loss rate (small α). Further, the variance of the MLE at the first layer or link 1 [Fig. 5(a)] is uniformly lower than that at the second layer corresponding to nodes 2 and 3 [Fig. 5(b)]. This is due to the larger number of probes that go through the nodes at higher layers of the tree. Similarly, the variance at nodes 2 and 3 [Fig. 5(b)] is lower than that of the receiver nodes [Fig. 5(c)], although the differences now are much smaller. This is because there is much more information about the receiver nodes from bicast pairs that split at the lowest layer (e.g., $\langle 4, 5 \rangle$ and $\langle 6, 7 \rangle$). This offsets the loss due to the fewer number of probes. Our investigations suggest that similar conclusions hold for four-layer and larger binary trees.

4.5 Large-Sample Properties

Recall that $N = \sum_h N_h$, the total number of probes in the experiment. Let $\hat{\alpha}_{\text{MLE}}$ denote the MLE. Further, let $I(\mathcal{C}, \vec{\alpha}_0)$ be the normalized information matrix given by $\sum_{C_h \in \mathcal{C}} \tau_h I(C_h, \vec{\alpha}_0)$, where $I(C_h, \vec{\alpha}_0)$ is the per-unit information for the scheme C_h .

Proposition 3. Assume that $\lim_{N \rightarrow \infty} N_h/N = \tau_h$, with $0 < \tau_h < 1$. Then (a) $\hat{\alpha}_{\text{MLE}} \rightarrow \vec{\alpha}_0$ a.s., and (b) $\sqrt{N}(\hat{\alpha}_{\text{MLE}} - \vec{\alpha}_0) \xrightarrow{\mathcal{L}} \text{MVN}(\mathbf{0}, \Sigma)$, where $\Sigma^{-1}(\vec{\alpha}) = I(\mathcal{C}, \vec{\alpha}_0)$, the normalized information matrix and MVN stands for multivariate normal.

Proof. Because the end-to-end data $N_{(i_1, \dots, i_k), h}$ have asymptotic normal distributions, the results follow in a straightforward manner if we can establish that the mapping $N_{(i_1, \dots, i_k), h} \rightarrow$

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59

60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118

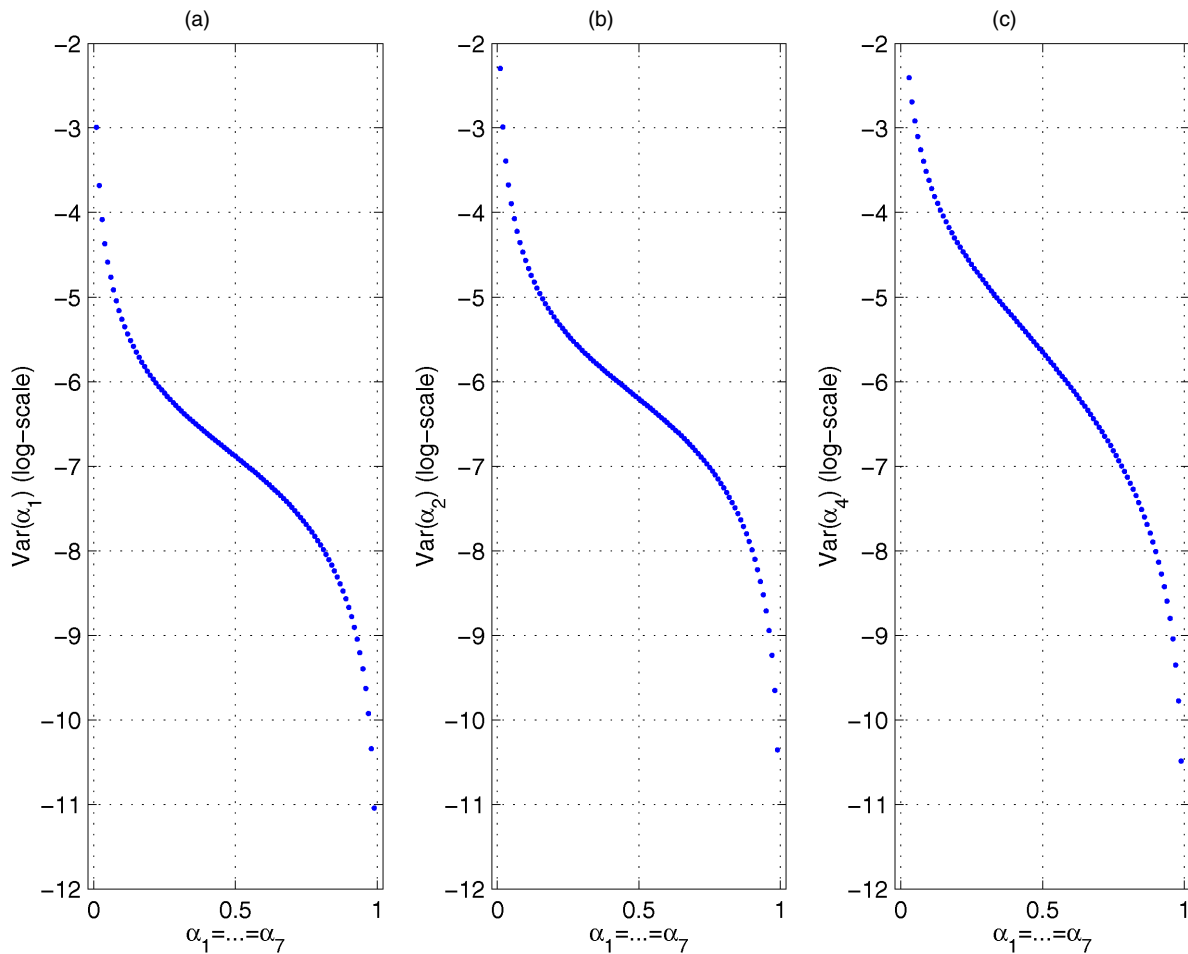


Figure 5. Variances of the MLEs for Selected Links in a Three-Layer Tree With Equal Loss Rates ($=P$).

$\hat{\alpha}_{MLE}$ is a continuously differentiable function. It can be shown that this is in fact true in local neighborhoods of the true values of the parameters, using the positive-definiteness of the Fisher information matrix and an argument based on the implicit function theorem. The details are omitted here.

We can use the asymptotic normality to construct confidence regions and hypothesis tests. We can also use likelihood-ratio methods for inference. These require computing the Hessian and using the observed information matrix to estimate the asymptotic variance-covariance matrix. Also note that the additive structure of the log-likelihood function over the individual schemes C_h simplifies the calculations considerably. The structure for k -cast scheme with a single split simplifies things, with the computations depending only on whether the node of interest is above or below the splitting node.

5. OPTIMAL DESIGN ISSUES RELATED TO PROBE ALLOCATION

There are two design issues associated with the flexible experiments $\mathcal{C} = \{C_h, N_h\}$: selection of appropriate schemes C_h , and allocation of the total number of probes to specific schemes N_h . We have already discussed the first problem. Here we consider the second problem, optimal allocation $\{N_h\}$ of a fixed budget of N probes to a given set of schemes $\{C_h\}$. Our goal

here is to develop a general formulation of the optimal allocation problem and to investigate the results for special cases to gain some insight.

The problem can be formulated as an optimal design of experiments problem. Given total probe size N , let τ_h denote the proportion of probes to be allocated to C_h . The optimal design problem is to choose $\{\tau_h\}$ to minimize an appropriate measure of variance of the MLEs of the link-level loss rates. The two most common criteria used in the optimal design literature are D-optimality and A-optimality (Pukelsheim 1993). D-optimality minimizes the determinant of the variance-covariance matrix (or maximizes that of the Fisher information matrix), whereas A-optimality minimizes the trace, that is, the sum of the variances. D-optimal designs are more common, because A-optimality ignores the covariances; thus we restrict our attention to the former criterion.

Let the experiment be denoted by $\{C_h, \tau_h\}$, with fixed total probe size N . The Fisher information matrix $I(N, \alpha)$ can be written as a weighted sum $N \sum_h \tau_h I(C_h, \alpha)$, where $I(C_h, \alpha)$ is the normalized information matrix corresponding to the scheme C_h . The D-optimal allocations of the τ 's are those that maximize the determinant of the Fisher information matrix. We see that $\det(\sum_h \tau_h I(C_h, \alpha))$ can be expressed as a polynomial in $\vec{\tau}$. The optimal value of $\vec{\tau}$ that maximizes this must be determined numerically. The more difficult issue is that the optimal allocations depend on the unknown values of the link-level

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59

60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118

1 parameters. This issue, called local optimality in the literature
 2 (Chernoff 1953; Pukelsheim 1993); is a common problem in
 3 most nonlinear (and nonnormal) design situations.

4 There are several ways to address this problem in prac-
 5 tice. The first, most common approach is to use any available
 6 preliminary information about the loss rates to determine the
 7 optimal allocations and assess their sensitivity to the inputs
 8 (sometimes called planning values). In our setup, the prelimi-
 9 nary information can come from historical data, specifica-
 10 tions for service-level agreements, and so on. If the results
 11 are very sensitive to the planning information, then one typi-
 12 cally will decide against using optimal allocations. (See, e.g.,
 13 Meeker and Escobar 1998 for a detailed discussion of this
 14 approach in the context of accelerated test planning.) A sec-
 15 ond approach that provides a formal framework for incorpo-
 16 rating prior information is Bayesian-optimal design theory (see
 17 Chaloner and Verdinelli 1995 for an excellent review). Let $p(\vec{\alpha})$
 18 be the prior distribution on the link probabilities. Then we can
 19 get the Bayes-optimal allocations for our problem by minimiz-
 20 ing the criterion $\phi(\tau) = \int \log(\det[N \sum_h \tau_h I(C_h, \alpha)]) p(\vec{\alpha}) d\vec{\alpha}$
 21 [Chaloner and Verdinelli 1995, p. 286, eq. (15)]. A third, non-
 22 Bayesian, alternative is to use a two-stage approach in which
 23 initial estimates are obtained from a first-stage experiment and
 24 the estimates are used to decide on the (approximately) opti-
 25 mal allocation in the second stage. We discuss the application
 26 of these approaches for a specific example.

27 First, we investigate the behavior of the optimal allocations
 28 (assuming that the true α 's are known) for some special cases
 29 to develop insights. Again, for simplicity we restrict attention
 30 to three- and four-layer symmetric binary trees with bicast exper-
 31 iments. We have conducted extensive investigations, but here we
 32 provide only selected results due to space limitations.

33 Figures 7–9 show the results for symmetric three-layer and
 34 four-layer (Fig. 6) trees. For the three-layer case, we used a bi-
 35 cast experiment with four schemes: $C_1 = \langle 4, 5 \rangle$, $C_2 = \langle 6, 7 \rangle$,
 36 $C_3 = \langle 5, 6 \rangle$, and $C_4 = \langle 4, 7 \rangle$. This includes one more bicast
 37 pair than a minimal experiment, so that all of the receiver nodes
 38 are treated symmetrically. For the four-layer tree, we used a bi-
 39 cast experiment with eight pairs: $C_1 = \langle 8, 9 \rangle$, $C_2 = \langle 10, 11 \rangle$,
 40 $C_3 = \langle 12, 13 \rangle$, $C_4 = \langle 14, 15 \rangle$, $C_5 = \langle 9, 10 \rangle$, $C_6 = \langle 11, 12 \rangle$,
 41 $C_7 = \langle 13, 14 \rangle$, and $C_8 = \langle 8, 15 \rangle$.

42 Figure 7(a) shows τ , the total D-optimal allocation for the
 43 two pairs that split at the second layer ($C_1 = \langle 4, 5 \rangle$ and $C_2 =$
 44 $\langle 6, 7 \rangle$). This is for the three-layer tree with equal α 's for all of
 45 the links. We see that τ varies in a small range around $2/3$, so
 46 each bicast gets around $1/3$ of the allocation and the remaining
 47 two pairs $C_3 = \langle 5, 6 \rangle$ and $C_4 = \langle 4, 7 \rangle$ each get about $1/6$. Note
 48 that the schemes that split at the lower level get more probes,
 49 and that the optimal allocations are remarkably stable across a
 50 broad range, $\alpha \in (.5, .99)$.

51 Figure 7(b) shows the corresponding results for the four-layer
 52 tree with equal α 's. The total D-optimal allocation for the four
 53 pairs that split at the lowest layer, τ_1 , is around $.60$. Recall that
 54 the total was $2/3$ in the three-layer case. The total allocation
 55 for three pairs that split at the middle layer, τ_2 , is around $.24$,
 56 and that for the single pair that splits at the top is $.14$. These
 57 values are again remarkably stable for α in the range $(.5, .99)$.
 58 It is also interesting that the schemes that split at the lowest and
 59 highest levels receive greater allocation than those that split at

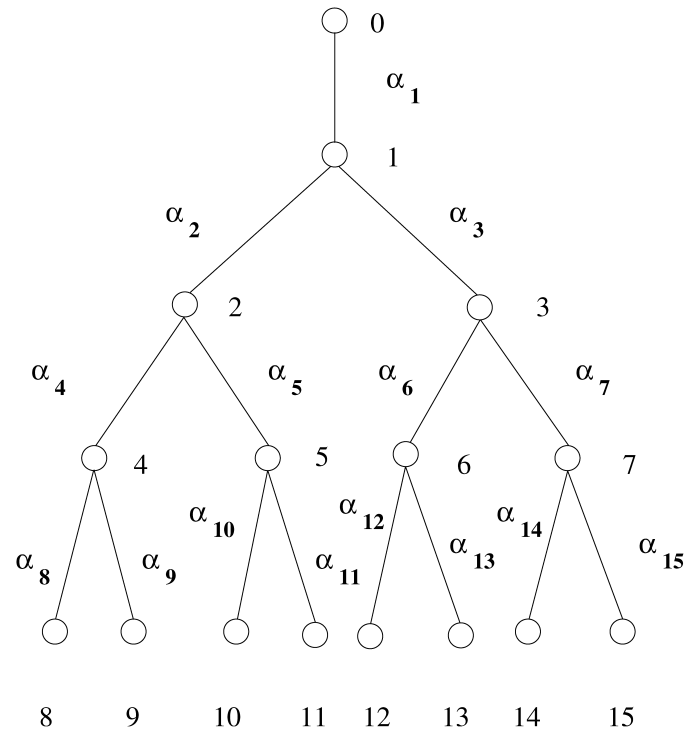


Figure 6. A Four-Layer Symmetric Binary Tree.

the middle. This is due to a combination of factors. Schemes
 that split near the top provide less information about links near
 the bottom, implying a need to increase the allocation. On the
 other hand, more probes traverse the links near the top than at
 the bottom, suggesting a need to increase the allocation to lower
 links. For example, all probes will traverse the 0–1 link. These
 effects trade off against one another to yield higher allocations
 for the top and bottom links and lower allocations for links in
 the middle.

Consider now the optimal allocations when the loss rates are
 unequal, that is, rates for links in the top and bottom layers
 are equal to P_1 and those in the middle layer(s) are equal to P_2 .
 Figure 8 deals with the three-layer tree. The y-axis shows the
 total allocation τ for the pairs that split at the second layer of
 the topology ($\langle 4, 5 \rangle$, $\langle 6, 7 \rangle$) for three cases: $P_1 = .8, .9,$
 and $.99$. The x-axis corresponds to values of $P_2 \in (.5, .99)$.
 We see that τ again varies in a small range, from $.68$ to $.73$,
 and is only slightly higher than the value of $2/3$ obtained pre-
 viously. Figure 9 shows the results for the four-layer tree,
 with again τ_1 represent the total allocation for the four pairs
 that split at the lowest layer, τ_2 representing that for the
 middle layer, and τ_3 representing that for the top layer. Again,
 τ_1 varies in a small range around $.6$ – $.65$ (close to the values
 for the case with equal α 's); τ_2 and τ_3 display similar
 behavior.

We have also investigated other bicast experiments for the
 three- and four-layer trees. It can be shown analytically, us-
 ing symmetry arguments, that for the three-layer tree and the
 foregoing choice of α 's, the experiment with all possible bi-
 cast pairs (six pairs) has exactly the same optimal allocations
 as the one that we considered earlier. For the four-layer tree,
 on the other hand, the case with all possible bicasts (28 pairs)
 exhibits slightly different behavior. We found only very small
 differences in the optimal allocations very small, however.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59

60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118

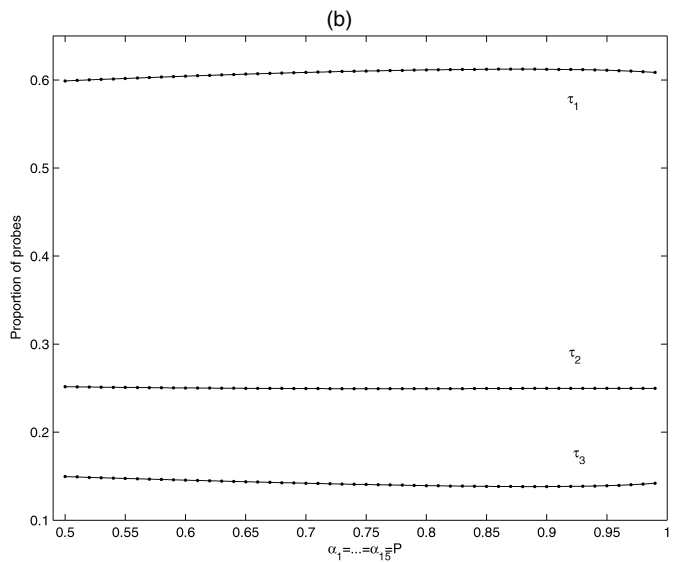
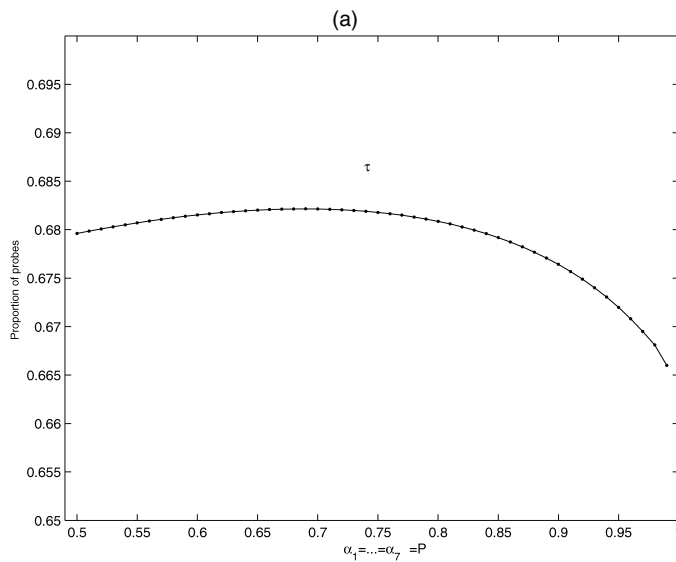


Figure 7. D-Optimal Allocations When the True Link-Loss Rates Are All Equal to P . Optimal allocations for (a) a three-layer symmetric binary tree and (b) a four-layer symmetric binary tree.

Let us now return to the practical problem where the loss rates are unknown. First, we see from Figures 7–9 that the optimal allocations are fairly stable for the region of interest, that is, the interval $(.90, .99)$. Specifically, Figure 7(b) shows that the allocations are remarkably constant for the four-layer tree with equal loss rates, suggesting that the results are robust to misspecification of the prior information. A similar conclusion holds for Figure 9, which shows the results for a four-layer tree with unequal loss rates. For the three-layer tree (Figs. 7 and 8), there is some change in the allocations as the probabilities get close to 1, but this is still very small [0.665–0.685 in Fig. 7(a)]. Thus we can conclude that in these cases, the optimal allocations are reasonably robust to uncertainty in the preliminary information. Because of this stability, the Bayesian D-optimal allocations will also be close to the locally optimal ones.

We also investigated the usefulness of the two-stage approach (discussed earlier) on a symmetric three-layer tree using a collection of four bicast schemes: $\langle 4, 5 \rangle$, $\langle 6, 7 \rangle$, $\langle 5, 6 \rangle$, and $\langle 4, 7 \rangle$. Given a total budget of $N = 1,000$ probes, a proportion q was allocated equally to all four bicast schemes in stage I. The data from the initial sample were used to estimate the success probabilities α . Based on the estimates $\hat{\alpha}$, the remainder of the $(1 - q)N$ probes were allocated using the optimal allocation scheme. The final estimate of α is a weighted combination of stage I and stage II estimates given by $q\hat{\alpha}^1 + (1 - q)\hat{\alpha}^2$. This procedure was repeated for $M = 1,000$ simulations. A number of scenarios were considered for the true values of α and values of q , but only selected results are reported here. For equal loss rates of $\alpha = .99$ and $q = .3$, the optimal allocations using the two-stage approach ranged from about .65 to .75 with about 70% in the interval .66–.72. Note from Figure 7(b) that the locally optimal allocation is around .67.

6. NETWORK SIMULATION STUDIES

So far we have studied the behavior of the estimators under the assumption of spatial and temporal stationarity. In this section we do a small simulation using the network simulator (ns) package to study the performance in a more realistic environment. Details about the ns simulator are available at <http://www.isi.edu/nsnam/ns>.

For simplicity, we consider a three-layer binary symmetric tree [see Fig. 2(b)]. In the simulation, all links had 1.5 Mb/sec of bandwidth and 10 ms of propagation delay and were served by a FIFO queue with a finite buffer of size 10. Thus a packet arriving at a node will be dropped if it encounters 10 packets already queued up. We considered two different scenarios: constant-bit-rate (CBR) (see Walrand and Varaiya 1999) traffic traversing the network and background traffic consisting of TCP background traffic and CBR probing traffic. The reason for investigating these two scenarios is that CBR traffic would lead to a stationary environment, as the posited model

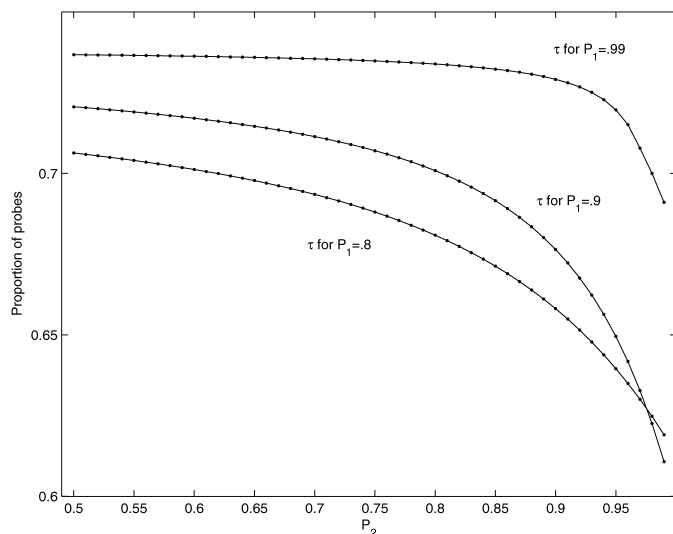


Figure 8. D-Optimal Allocations for the Three-Layer Tree With Unequal Loss Rates. Loss rates are equal to P_1 for links at the top and bottom layers and equal to P_2 for links at middle layer.

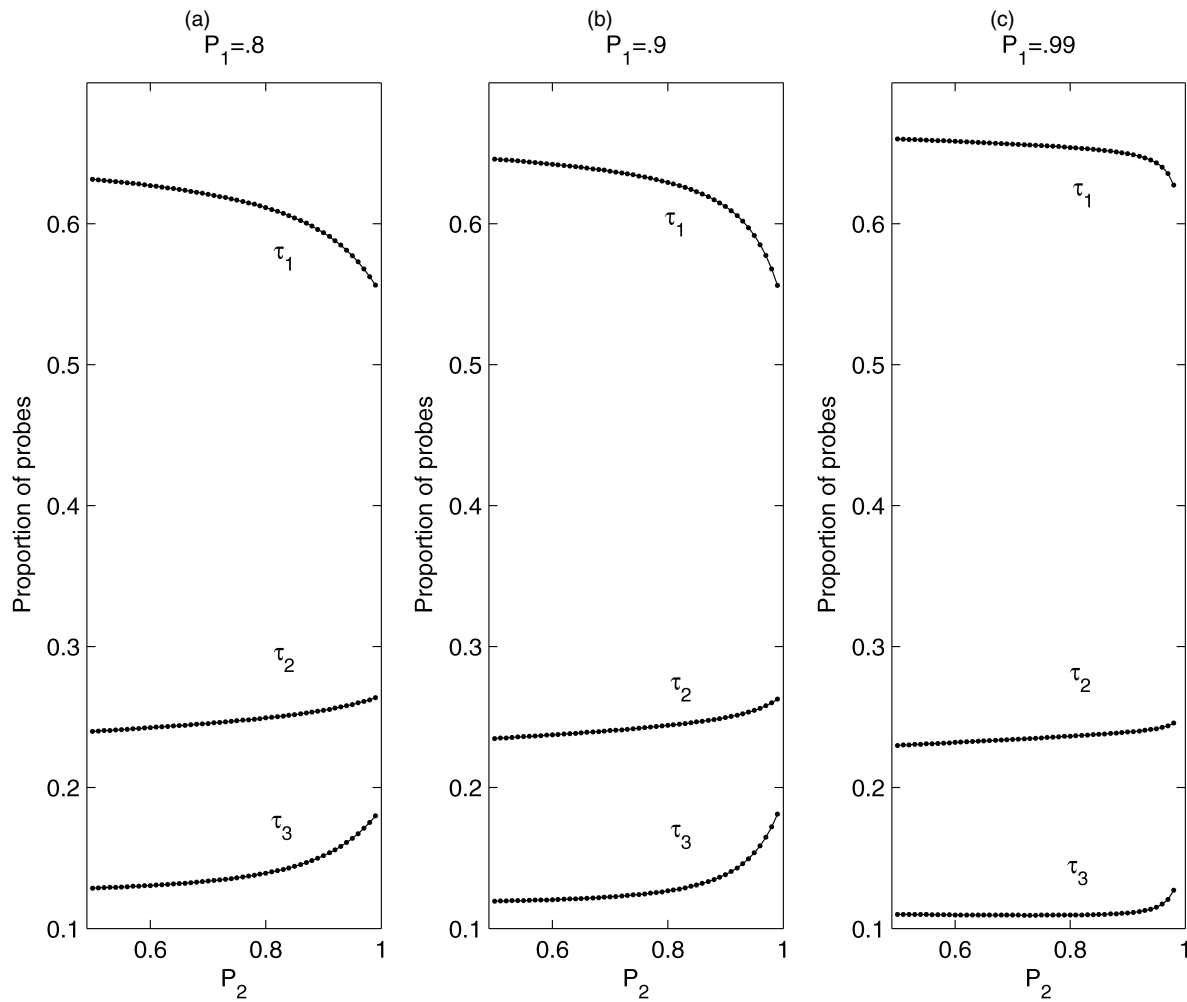


Figure 9. D-Optimal Allocation for the Four-Layer Tree With Unequal Loss Rates: $=P_1$ for the Top and Bottom Links and $=P_2$ for the Middle Links. (a) $P_1 = .80$; (b) $P_1 = .90$; (c) $P_1 = .99$.

requires. In contrast the TCP protocol, which is the predominant protocol in real networks, is a bursty packet source due to its “linear increase–exponential backoff” rate of transmission nature (Walrand and Varaiya 1999).

For the all CBR traffic scenario, the root link and the receiver links carried a single flow, whereas the middle links (1–2 and 1–3) had two flows. The background traffic was generated by infinite data sources that sent 500-byte packets with a uniform interpacket distribution in (1, 3) ms. The probing experiment for the three-layer tree consisted of the three bicast schemes: $\langle 4, 5 \rangle$, $\langle 6, 7 \rangle$, and $\langle 5, 6 \rangle$. Forty-byte packets were transmitted with a uniform interpacket distribution in (2.5, 7.5) ms. Hence, probing traffic was a small fraction ($<5\%$) of the total traffic in the network.

Figure 10 shows the inferred and the actual (tracked by the ns simulator) loss rates on selected links over 5,000 observations. Although we are dealing with a highly congested network, we see that the estimates track the actual loss rates extremely well. In the second simulation scenario there were 52 TCP connections on the various links, resulting in about seven or eight flows per link. The TCP connections sent 1,000-byte packets, and the FIFO queue buffer was 4 packets. The characteristics of the probing traffic were the same as

before. Figure 11 shows the actual and inferred loss rates for selected links. Note the higher loss rate in the 3–7 link due to the presence of eight connections compared with the seven present in the 2–4 link. Although the tracking of the actual link losses is very consistent, there exists a small systematic bias in the estimates (a fact also observed in Caceres et al. 1999). This is likely due to nonstationarity caused by persistent losses on neighboring links. This issue merits further study.

7. APPLICATION TO NETWORK MONITORING

A major goal in network engineering is to monitor the network over time for anomalous behavior and to diagnose where the problems occur, that is, identify the affected nodes or sub-networks. In this section we demonstrate the usefulness of the results for network monitoring in an idealized setting. A comprehensive methodology for the monitoring problem is the subject of ongoing work. This will involve taking into account the considerable variation in network parameters due to diurnal, day-of-week, and other effects. These can be ignored in getting (local-in-time) estimates of the QoS parameters, because the probing experiments are conducted within a span of minutes. But they become important in the context of monitoring done over a longer period. In ongoing work, we are studying methods

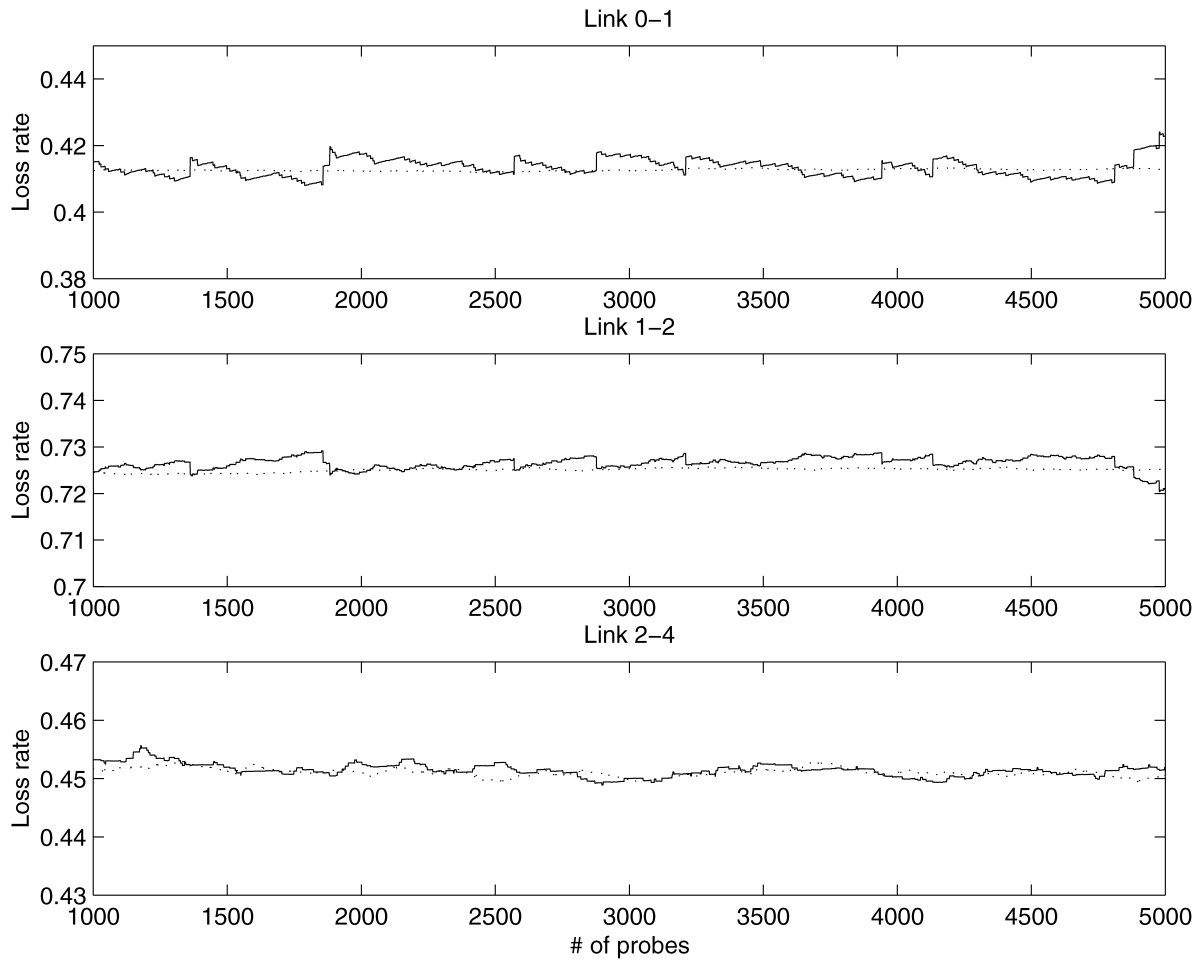


Figure 10. Tracking the Actual Loss Rates (.....) by Inferred Loss Rates (—) in a CBR Simulation for Selected Links of a Three-Layer Tree Topology.

for estimating the systematic effects and removing them to detect changes in the presence of nonstationarity. Our goal in this section is more limited: to demonstrate the potential usefulness of the methods developed in the article for network monitoring. The flexibility of the new class of schemes makes it particularly well suited for the monitoring application.

We consider the following idealized framework for monitoring and detecting changes: $\vec{\alpha} = \vec{\alpha}_0$ for all times $t \leq T$ and some of the α_j 's change to $\alpha_{1j} < \alpha_{0j}$ at some random point in time $T > 0$. Our goal is to detect the change as quickly as possible and to identify the link or collection of links where the problem has occurred.

We first estimate the individual link-level loss rates to establish baselines as follows. Time is divided in $\Delta > 0$ time intervals, and within every Δ interval a number of N_Δ probes are used for the probing experiment. (The total number of probes N_Δ is appropriately allocated among the k -cast schemes used in the probing experiment.) That is, $t = k\Delta$, $k = 0, 1, 2, \dots$. Using the data obtained from the N_Δ probes, an estimate of $\hat{\alpha}(t)$ using the EM algorithm is obtained. There are various ways to monitor for changes in the values of $\vec{\alpha}(t)$. One method that is suitable for detecting both small and medium changes is the exponentially weighted moving average procedure (EWMA) (Basseville and Benvensite 1986). The EWMA statistic can be

expressed as

$$Z_j(t) = \lambda \hat{\alpha}_j(t) + (1 - \lambda)Z_j(t - 1),$$

where $\hat{\alpha}_j(t)$ is the local estimate of α_j at time t , $Z_j(1) = \hat{\alpha}_j(1)$, λ is an appropriate weight, and $Z_j(t)$ is obtained iteratively from the foregoing.

We illustrate the methods on the four-layer binary symmetric tree in Figure 6 as the logical topology of the network being monitored. We consider two different scenarios to capture different types of changes. In the first scenario, $\alpha_i = .99$, $i = 1, \dots, 5$; that is, the network is in its normal state for the first five time periods. Then there is small increase in the loss rate for link 1–3 from .99 to .95; all other links remain the same. Figure 12 shows the EWMA chart, which gives the EWMA statistic and the lower and upper control limits for the links in the path 0–15, that is, $\alpha_1, \alpha_3, \alpha_7$, and α_{15} . The control limits were calculated using the “null” values of the success probabilities, that is, taking the mean level equal to .99. The probing experiment consisted of 8 bicast schemes with 250 probe packets allocated to each bicast pair. A value of $\lambda = .6$ was used, a common choice in the process control literature (see, e.g., Basseville and Benvensite 1986). We see from Figure 12 that the change in α_3 was clearly detected, even though it was relatively small. The EWMA statistic for the other links in the path are within the control limits, except for α_7 , which had just

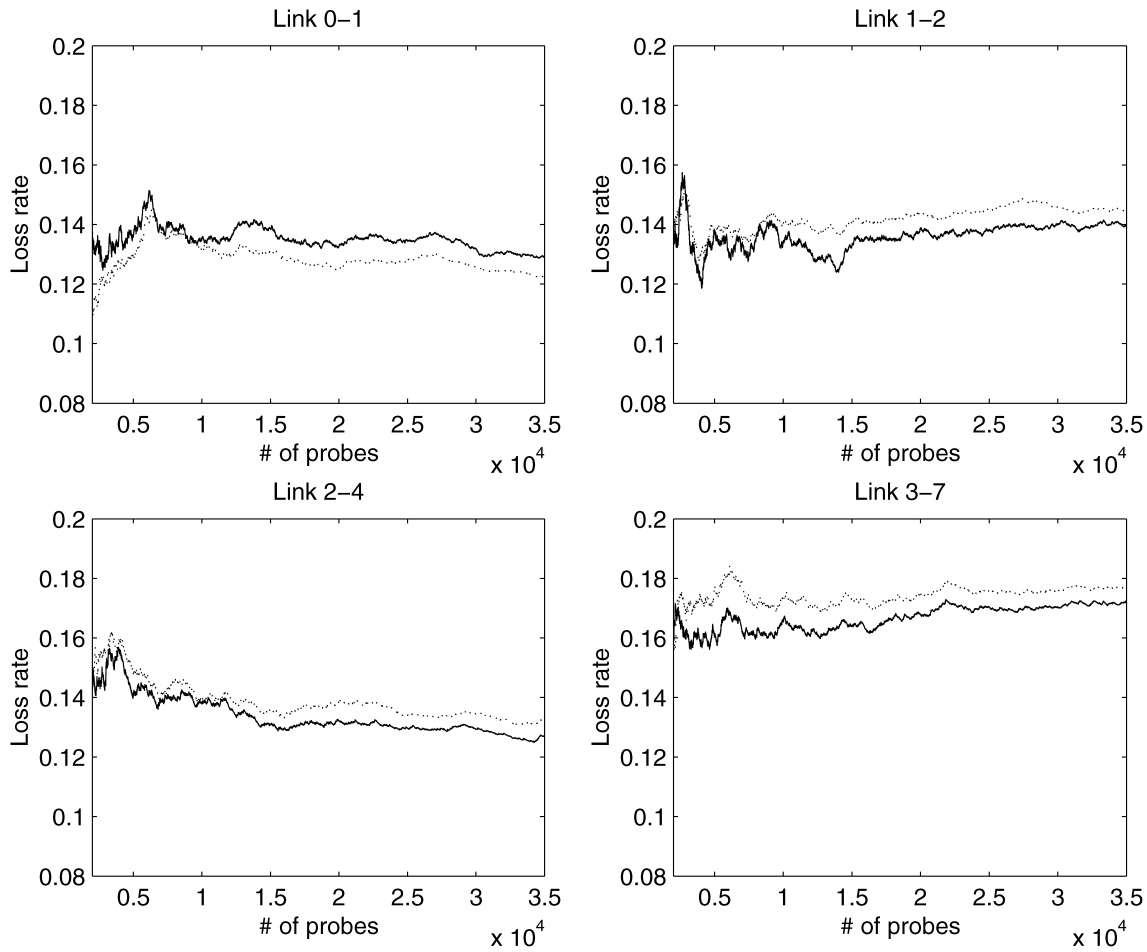


Figure 11. Tracking the Actual Loss Rates (.....) by Inferred Loss Rates (—) in a TCP Simulation for Selected Links of a Three-Layer Tree Topology.

one point outside the control limits. The figure also shows the variability of the moving average process because of the rather small number of probing packets used. This variability can, of course, be reduced by increasing the probe size. The design of monitoring schemes, including the choice of monitoring statistic, probe sizes, and average run lengths, are being studied in ongoing work.

To understand how well the procedure works, one must study the run-length distribution of the monitoring procedure. Here run length (RL) is defined as the number of periods before a change is detected; that is, the statistic falls outside the control limits (Basseville and Benvensite 1986). Although one can investigate the RL distribution in general, it is common to focus on the expected or average RL (ARL). It is desirable to have a large ARL under the null hypothesis of no change (ARL_0) and a small ARL when there is a change (ARL_1). The RL can be viewed as the first-passage time of the underlying process across the control limits (one- or two-sided boundaries). The most common method for computing ARLs (aside from simulation) uses a Markov chain approximation (Brook and Evans 1972; Ringer and Prabhu 1996) by discretizing the state space. Crowder (1987) developed a better, integral-equation approach for EWMA-based statistics. Numerical routines are available in SAS for computing the ARLs when the underlying process is normal. We used these routines for our problem, using a normal

approximation for $\hat{\alpha}_j(t)$ s. The normal approximation is reasonable when the probe size $n \geq 100$ but is not as good for $n = 50$. We did some simulations to calibrate the numerical results in this small-sample case and found that the ARL values from simulation were slightly smaller than those reported in Tables 2 and 3. Our setup is also a bit more complicated than the usual normal case in which the mean shift is not related to the variance. We used the integral equation with control limits under the null but the variance of the process under the alternative.

Table 2 gives the ARLs for the situation of interest: α_3 changes from .99 to .95 and all other α_k 's remain unchanged at .99. ARL values for different probe sizes n and different values of the weights λ are given. The value of L refers to the width of the control limits ($\pm L\sigma$) and was chosen so that the in-control ARL is about 250 in all cases. The ARL values displayed in the table are the expected number of time intervals before a change is detected. We see that even with a small sample of 50 probes, the change is detected within three time periods; this reduces to about two periods with probe size of 100. For $n = 250$, there is almost immediate detection (one time period). The optimal weighting parameter (corresponding to the smallest ARL in each row) changes with changing sample size (because a larger sample size implies a larger shift size in terms of the noncentrality parameter).

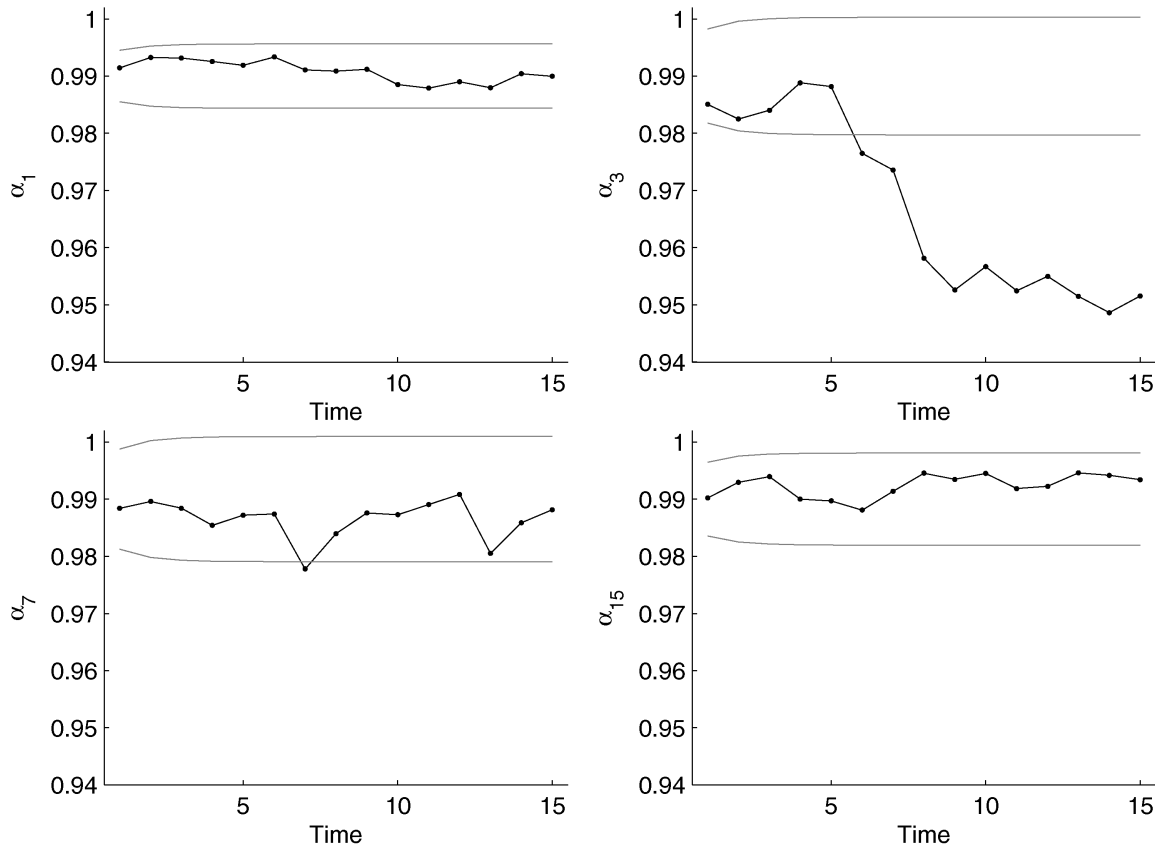


Figure 12. Link Success Probabilities Monitoring of a Sudden Change in a Single Link Using an EWMA Chart.

To put these numbers into perspective, note that probes can be sent approximately 15–20 milliseconds apart without interfering with the operation of the network. Therefore, one (monitoring) period ranges from about 1 second (for 50 probes) to 5 seconds (250 probes). So we see that a small-magnitude change can be safely detected in 3–5 seconds.

The second scenario is similar to the first scenario but now involves deterioration in two links, α_3 and α_7 , along the path $\mathcal{P}(1, 8)$; that is, α_7 also changes from 0.99 to .95 for $t = 6, \dots, 10$. Once again, 250 probes per bicast pair were used. The results, shown in Figure 13, indicate that changes in both links can be successfully detected while having no false alarms on the remaining two links on the path $(0, 15)$.

Table 3 gives the ARLs for α_7 . The ARLs for α_3 were qualitatively very similar to those in Table 2 under Scenario 1 and thus are omitted due to space limitations. The conclusions from Table 3 are very similar to those under Scenario 1 with the single-link change problem.

As noted earlier, in practice, network monitoring is done over a period during which the QoS parameters will vary. We will have to accommodate for systematic variation due to time-of-day, day-of-the-week, and other effects. Furthermore, in the

foregoing illustration, we were solving the inverse problem to estimate the α 's at each time point. However, for the purpose of detection, we can just monitor the end-to-end path estimates $\hat{\pi}(0, r_h)$ for all of the receiver nodes. Once a change in performance is detected, we can solve the inverse problem to estimate the α 's and identify the regions in which performance has degraded. A comparison of this alternative approach to the one that we illustrated earlier merits further study. Finally, network monitoring and intrusion detection is a very important area, and network engineers use a wide array of tools and data sources to address this problem. The results from active tomography must be effectively combined with other sources of information and tools for effective monitoring.

8. CONCLUDING REMARKS

There are a number of interesting directions for further work in the context of computer and communication networks. These include design issues for multisource topologies, incorporation of temporal and spatial dependence, and the network monitoring problems discussed in the preceding section.

We have formulated and presented the results in terms of the application to network tomography, because this is an interest-

Table 2. ARLs for Scenario 1 and Link 3

Probe size n	$L =$	2.439	2.532	2.582	2.611	2.629	2.646
	$\lambda =$.2	.3	.4	.5	.6	.8
$n = 50$		4.59	4.50	4.62	4.90	5.34	6.81
$n = 100$		3.02	2.82	2.74	2.73	2.79	3.13
$n = 250$		1.91	1.73	1.61	1.52	1.46	1.41

Table 3. ARLs for Scenario 2 and Link 7

Probe size n	$L =$	2.439	2.532	2.582	2.611	2.629	2.646
	$\lambda =$.2	.3	.4	.5	.6	.8
$n = 50$		3.32	3.13	3.08	3.11	3.23	3.74
$n = 100$		2.31	2.10	1.99	1.92	1.88	1.91
$n = 250$		1.54	1.37	1.26	1.18	1.14	1.09

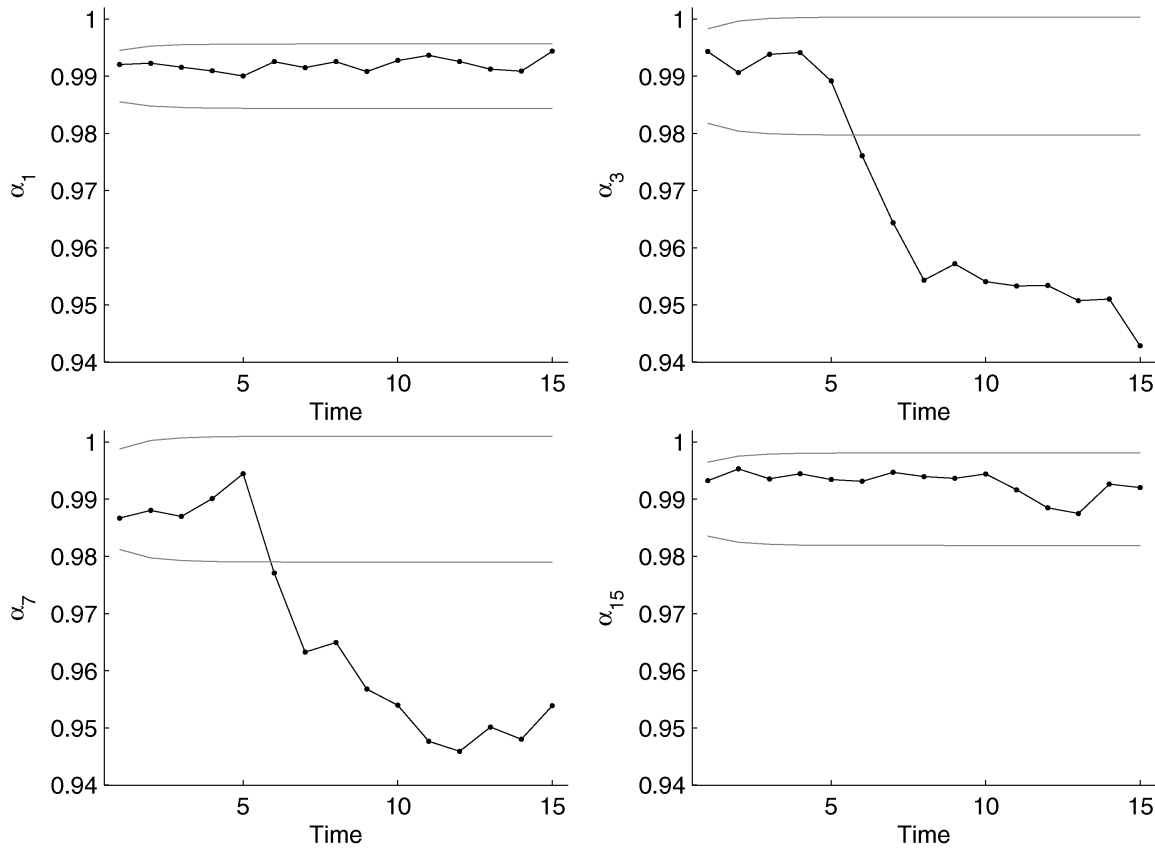


Figure 13. Link Success Probabilities Monitoring of a Sudden Change Along a Path Using an EWMA Chart.

ing class of inverse problems. However, the results can be also viewed more generally as inference for tree-structured graphs. There are other applications, such as manufacturing assembly processes and distribution networks, where these results can also be applied with appropriate modification.

APPENDIX: PROOFS

A.1 Proof of Proposition 1

If every internal node s in \mathcal{T} is a splitting node for some scheme C_h , then it is automatically a splitting node for a two-cast subset of the k -cast scheme. Thus we can study an equivalent problem involving an experiment $\tilde{\mathcal{C}}$ comprising bicast and unicast schemes with the following characteristics: (I) for each internal node s in \mathcal{T} , there is at least one bicast pair $b \in \mathcal{B}$ whose splitting node is s , and (II) the unicast schemes in \mathcal{U} are chosen to cover the remaining receiver nodes $r \in \mathcal{R}$ that are not covered by the bicast pairs in \mathcal{B} .

It suffices to establish the existence of a bijection between $\vec{\alpha}$ and the parameters $\Gamma \cup \Delta$, where Γ and Δ are defined as follows. Define \mathcal{B} to denote the collection of all bicast pairs used in the experiment and let \mathcal{U} denote the collection of unicast schemes. Let $\Gamma^b = \{\gamma_{1,1}^b, \gamma_{1,0}^b, \gamma_{0,1}^b\}$ denote the set of free probabilities from a pair of receiver nodes $b = \langle i, j \rangle$ and let $\Gamma = \{\Gamma^b : b \in \mathcal{B}\}$ denote the probabilities generated by all bicast pairs in \mathcal{B} . Let $\Delta^u = \{\delta_1^u, \delta_0^u\}$ denote the probabilities of the two outcomes for unicast scheme u and let $\Delta = \{\Delta^u : u \in \mathcal{U}\}$.

Sufficiency. It is easy to see that $\Gamma = \{\Gamma^b; b \in \mathcal{B}\}$ and $\Delta = \{\Delta^u; u \in \mathcal{U}\}$ are uniquely determined by $\vec{\alpha}$. We next show that the elements of $\vec{\alpha}$ are also uniquely determined by $\Gamma \cup \Delta$.

Recall that a node $i \in \mathcal{V} - \{0\}$ belongs to the k th layer \mathcal{L}_k of \mathcal{T} if its shortest path from the root node has k links. We need to consider the following three cases: (1) the splitting node for bicast pair b is node 1,

that is, belongs to first layer \mathcal{L}_1 ; (2) the splitting node is any internal node, that is, $s \in \mathcal{I}$; and (3) the case of receiver nodes $r \in \mathcal{R}$.

Case 1. For bicast pair $b_0 = \langle i_b, j_b \rangle$ with splitting node 1, we have

$$\alpha_1 = \pi^{b_0}(0, 1) = \frac{(\gamma_{11}^{b_0} + \gamma_{10}^{b_0})(\gamma_{11}^{b_0} + \gamma_{01}^{b_0})}{\gamma_{11}^{b_0}}.$$

Therefore, it is determined by the elements of Γ .

Case 2. We proceed by induction. Suppose that for all internal nodes s such that $s \in \mathcal{L}_1 \cup \mathcal{L}_2 \cup \dots \cup \mathcal{L}_{k-1}$, the α_s 's are determined by Γ . We need to show that $\alpha_t, t \in \mathcal{L}_k$ is also determined by Γ . Because t is an internal node, there exists a bicast scheme $b_0 \in \mathcal{C}$ with splitting node corresponding to t . We have that $\pi^{b_0}(0, t) = \pi^{b_0}(0, f(t))\alpha_t$, with all members of $\pi^{b_0}(0, f(t))$ already determined. As before, we have that $\pi^{b_0}(0, t) = (\gamma_{11}^{b_0} + \gamma_{10}^{b_0})(\gamma_{11}^{b_0} + \gamma_{01}^{b_0})/\gamma_{11}^{b_0}$, which, combined with the previous observation, establishes the identifiability of α_t from the elements of Γ .

Case 3. We now deal with the receiver nodes $r \in \mathcal{R}$. Note that due to the induction hypothesis in the previous step, all α_s 's, with $s \in \mathcal{I}$, have been identified. A receiver node can be covered by either a unicast scheme or a bicast scheme. For the unicast case, $\pi^{u_0}(0, r) = \pi^{u_0}(0, f(r))\alpha_r$, with all elements of $\pi^{u_0}(0, f(r))$ already identified by the induction. We also have that $\delta_1^{u_0} = \pi^{u_0}(0, r)$, which, combined with the previous observation, establishes the identifiability of α_r from elements of Δ . For the bicast scheme $b_0 = \langle i_{b_0}, j_{b_0} \rangle$ with splitting node s_{b_0} , we have that $\pi(s_{b_0}, i_{b_0}) = \gamma_{11}^{b_0}/(\gamma_{11}^{b_0} + \gamma_{01}^{b_0})$ and $\pi(s_{b_0}, j_{b_0}) = \gamma_{10}^{b_0}/(\gamma_{11}^{b_0} + \gamma_{10}^{b_0})$. But $\pi^{b_0}(s_{b_0}, r) = \pi^{b_0}(0, f(r))\alpha_r$, with r being either i_{b_0} or j_{b_0} , and the result follows as before.

This establishes that there is a bijection between $\vec{\alpha}$ and $\Gamma \cup \Delta$.

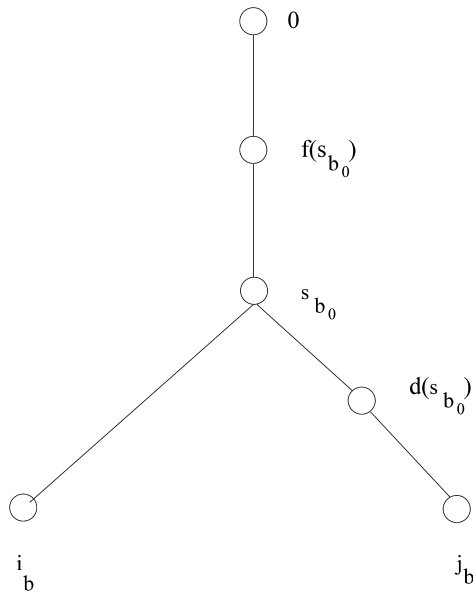


Figure A.1. Demonstration of (A.1) and (A.2).

Necessity. We argue by contradiction. Suppose that there is a combination of bicast schemes that includes *all* possible pairs, except the collection of pairs $B_0 = \{b \in \mathcal{C} : \text{splitting node is } s_{b_0} \in \mathcal{T}\}$ that have as their splitting node s_{b_0} an internal node of \mathcal{T} . We show that this fails to identify all of the elements of $\vec{\alpha}$.

Let $f(s_{b_0})$ and $d(s_{b_0})$ denote the parent and any child node of node s_{b_0} (see Fig. A.1). From the previous derivations, it is easy to see that the following relationships hold:

$$\pi^b(0, d(s_{b_0})) = \pi^b(0, f(s_{b_0})) \times \alpha_{s_{b_0}} \times \alpha_{d(s_{b_0})} \quad (\text{A.1})$$

and

$$\pi^b(f(s_{b_0}), \ell) = \alpha_{s_{b_0}} \times \alpha_{d(s_{b_0})} \times \pi^b(d(s_{b_0}), \ell), \quad (\text{A.2})$$

for any $b \in \mathcal{C}$, where ℓ corresponds to some receiver node for pair b . Note that, as argued in the sufficiency part of the proof, only these π 's are uniquely determined by their Γ^b 's. Further note, that both (A.1) and (A.2) correspond to two actual equations, because node s_{b_0} has two children nodes for bicast schemes in B_0 . A straightforward calculation shows that only the values of the products $\alpha_{s_{b_0}} \times \alpha_{d(s_{b_0})}$ can be calculated uniquely from the elements of Γ by taking the appropriate ratios, but the individual parameters cannot be disentangled. Hence \mathcal{C} fails to identify all of the elements of $\vec{\alpha}$. This completes the proof of the proposition.

A.2 Proof of Proposition 2

In this section we denote the information matrix by Σ . Suppose that there exists a vector $\vec{c} \in \mathcal{R}^E$ such that $\vec{c}'\Sigma\vec{c} = 0$. We show that every element of \vec{c} must be 0, which establishes the result.

Suppose that $\text{var}(\vec{c}'S(\vec{\alpha})) = \vec{c}'\Sigma\vec{c} = 0$ and $E(\vec{c}'S(\vec{\alpha})) = 0$. We must have that $\vec{c}'S(\vec{\alpha}) = 0$, a.s. Equivalently,

$$\sum_{e=1}^E c_e \frac{\partial \Lambda(\vec{\alpha}|\mathbf{N})}{\partial \alpha_e} = 0 \quad (\text{A.3})$$

for all possible elements of \mathbf{N} .

We demonstrate the result for a collection of schemes \mathcal{C}_h comprising bicast and unicast transmissions, and then indicate how it generalizes for an arbitrary collection. Recall from our construction of minimal experiments that unicast schemes may uniquely cover receiver links only, whereas *all* links between internal nodes are covered by bicast schemes. We show that $\vec{c} = 0$.

We next examine the three cases.

Case 1. Consider an arbitrary bicast scheme, $b = \langle i, j \rangle$, that covers receivers i and j with splitting node s . Without loss of generality, assume that every bicast and unicast scheme used in the collection \mathcal{C} receives a single probe packet. Furthermore, because the result must be true for all \mathbf{N} , assume that in all of the other bicast schemes in the collection \mathcal{C} , the observed outcomes are also (1, 1), and in all of the unicast schemes, the observed outcome is 1.

If the observed outcome is also (1, 1) for the bicast pair b , then we have

$$\begin{aligned} & \log \Lambda(\vec{\alpha}|\mathbf{N}) \\ &= \sum_{\ell \in \mathcal{P}(0,s)} \log(\alpha_\ell) + \sum_{\ell \in \mathcal{P}(s,i)} \log(\alpha_\ell) + \sum_{\ell \in \mathcal{P}(s,j)} \log(\alpha_\ell) \\ & \quad + \sum_{b \in \mathcal{C}, b \neq b} \log(\gamma_{1,1}^b) + \sum_{u \in \mathcal{C}} \log(\delta_1^u), \end{aligned} \quad (\text{A.4})$$

which implies that (A.3) becomes

$$\sum_{\ell \in \mathcal{P}(0,s)} \frac{c_\ell}{\alpha_\ell} + \sum_{\ell \in \mathcal{P}(s,i)} \frac{c_\ell}{\alpha_\ell} + \sum_{\ell \in \mathcal{P}(s,j)} \frac{c_\ell}{\alpha_\ell} + g(\vec{c}) = 0, \quad (\text{A.5})$$

with $g(\vec{c})$ capturing the terms in the sum over all bicast and unicast schemes, but bicast pair b .

Suppose now that the observed outcome for pair b is (1, 0) whereas for all the other bicast schemes in \mathcal{C} were still (1, 1) and at all unicast schemes 1; then, (A.3) becomes

$$\begin{aligned} & \sum_{\ell \in \mathcal{P}(0,s)} \frac{c_\ell}{\alpha_\ell} + \sum_{\ell \in \mathcal{P}(s,i)} \frac{c_\ell}{\alpha_\ell} \\ & \quad + \sum_{\ell \in \mathcal{P}(s,j)} \frac{c_\ell \times \pi(s,j)}{\alpha_\ell \times (\pi(s,j) - 1)} + g(\vec{c}) = 0. \end{aligned} \quad (\text{A.6})$$

Furthermore, assume that the observed outcome at pair b is (0, 1), whereas that for all of the other bicast schemes in \mathcal{C} is still (1, 1); then (A.3) becomes

$$\begin{aligned} & \sum_{\ell \in \mathcal{P}(0,s)} \frac{c_\ell}{\alpha_\ell} + \sum_{\ell \in \mathcal{P}(s,i)} \frac{c_\ell \times \pi(s,i)}{\alpha_\ell \times (\pi(s,i) - 1)} \\ & \quad + \sum_{\ell \in \mathcal{P}(s,j)} \frac{c_\ell}{\alpha_\ell} + g(\vec{c}) = 0. \end{aligned} \quad (\text{A.7})$$

Finally, assume that the observed outcome for pair b is (0, 0), whereas that for all of the other bicast schemes in \mathcal{C} is still (1, 1) and that for the unicast schemes is 1; then (A.3) becomes

$$\begin{aligned} & \sum_{\ell \in \mathcal{P}(0,s)} \frac{c_\ell}{\alpha_\ell} \left[\frac{\pi(0,s)[(1 - \pi(s,i)) \times (1 - \pi(s,j)) - 1]}{\gamma_{00}^b} \right] \\ & \quad + \sum_{\ell \in \mathcal{P}(s,i)} \frac{c_\ell}{\alpha_\ell} \left[\frac{\pi(s,i)\pi(0,s)(\pi(s,j) - 1)}{\gamma_{00}^b} \right] \\ & \quad + \sum_{\ell \in \mathcal{P}(s,j)} \frac{c_\ell}{\alpha_\ell} \left[\frac{\pi(s,j)\pi(0,s)(\pi(s,i) - 1)}{\gamma_{00}^b} \right] + g(\vec{c}) \\ &= 0. \end{aligned} \quad (\text{A.8})$$

Subtracting (A.6) from (A.5) gives

$$\begin{aligned} & \sum_{\ell \in \mathcal{P}(s,j)} \frac{c_\ell}{\alpha_\ell} - \sum_{\ell \in \mathcal{P}(s,j)} \frac{c_\ell \times \pi(s,j)}{\alpha_\ell \times (\pi(s,j) - 1)} = 0 \\ & \Rightarrow \sum_{\ell \in \mathcal{P}(s,j)} \frac{c_\ell}{\alpha_\ell} \times \frac{-1}{(\pi(s,j) - 1)} = 0 \\ & \Rightarrow \sum_{\ell \in \mathcal{P}(s,j)} \frac{c_\ell}{\alpha_\ell} = 0. \end{aligned} \quad (\text{A.9})$$

Subtracting (A.7) from (A.5) and going through similar steps gives that

$$\sum_{\ell \in \mathcal{P}(s,i)} \frac{c_\ell}{\alpha_\ell} = 0. \tag{A.10}$$

From (A.8), and using the results from (A.9) and (A.10), we get

$$\sum_{\ell \in \mathcal{P}(0,s)} \frac{c_\ell}{\alpha_\ell} \left[\frac{\pi(0,s)[(1-\pi(s,i)) \times (1-\pi(s,j)) - 1]}{\gamma_{00}^b} \right] + g(\vec{c}) = 0. \tag{A.11}$$

From (A.5), together with the results obtained in (A.9) and (A.10), we get

$$\sum_{\ell \in \mathcal{P}(0,s)} \frac{c_\ell}{\alpha_\ell} + g(\vec{c}) = 0. \tag{A.12}$$

Subtracting (A.11) from (A.12), and after some algebra, we finally get

$$\sum_{\ell \in \mathcal{P}(0,s)} \frac{c_\ell}{\alpha_\ell} = 0. \tag{A.13}$$

Furthermore, by adding (A.10) to (A.13), we have

$$\sum_{\ell \in \mathcal{P}(0,i)} \frac{c_\ell}{\alpha_\ell} = 0, \tag{A.14}$$

with node i a receiver.

Case 2. Now consider an arbitrary unicast u that covers receiver r . Suppose that for all bicast schemes in \mathcal{C} , the observed outcome is $(1, 1)$, whereas for all unicast schemes, the observed outcome is 1 . Then (A.3) becomes

$$\sum_{\ell \in \mathcal{P}(0,r)} \frac{c_\ell}{\alpha_\ell} + g'(\vec{c}) = 0. \tag{A.15}$$

Assume now that the observed outcome for unicast scheme u is 0 instead, whereas all of the observed outcomes for all other unicast and bicast schemes in \mathcal{C} remain as before. Then, (A.3) becomes

$$\sum_{\ell \in \mathcal{P}(0,r)} \frac{c_\ell}{\alpha_\ell} \left[\frac{-\pi(0,r)}{1-\pi(0,r)} \right] + g'(\vec{c}) = 0. \tag{A.16}$$

Subtracting (A.16) from (A.15), after some algebra, we get

$$\sum_{\ell \in \mathcal{P}(0,r)} \frac{c_\ell}{\alpha_\ell} = 0. \tag{A.17}$$

Due to the construction of the collection \mathcal{C} , every internal node must be a splitting node for one bicast scheme, which in turn implies that (A.13) holds for all internal nodes s . Furthermore, because collection \mathcal{C} covers all links, (A.17) holds for all receiver nodes r . Therefore, by taking successive differences along every path $\mathcal{P}(0, j), j \in \mathcal{V} - \{0\}$, of the form

$$\sum_{\ell \in \mathcal{P}(0,j)} \frac{c_\ell}{\alpha_\ell} - \sum_{\ell \in \mathcal{P}(0,f(j))} \frac{c_\ell}{\alpha_\ell} = \frac{c_{(f(j),j)}}{\alpha_{(f(j),j)}} = 0,$$

we can easily establish that $c_\ell = 0, \forall \ell \in \mathcal{E}$. Therefore, $\vec{c} = 0$, and hence the Fisher information matrix $\mathcal{I}_N(\mathcal{C}, \vec{\alpha})$ is positive definite in the interior of $(0, 1)^E$.

For a general collection of flexicast schemes \mathcal{C} , we can proceed along similar lines as follows. Consider the h th k -cast scheme with splitting nodes denoted by $s^h_1, s^h_2, \dots, s^h_d$. A similar strategy of using all of the possible 2^k outcomes for the h th scheme and assuming that for all remaining schemes in the collection, only 1's are observed, we

can establish the following relationships:

$$\sum_{\ell \in \mathcal{P}(0,s^h_j)} \frac{c_\ell}{\alpha_\ell} = 0,$$

$$\sum_{\ell \in \mathcal{P}(s^h_j, s^h_j)} \frac{c_\ell}{\alpha_\ell} = 0, \quad j = 1, \dots, d-1, \quad \text{and}$$

$$\sum_{\ell \in \mathcal{P}(s^h_j, r)} \frac{c_\ell}{\alpha_\ell} = 0.$$

Then, taking differences as before, we establish the result that $c_\ell = 0$ for all $\ell \in \mathcal{E}$, which in turn proves the nonsingularity of the Fisher information matrix.

[Received July 2003. Revised December 2003.]

REFERENCES

Basseville, M., and Benveniste, A. (1986), *Detection of Abrupt Changes in Signals and Dynamical Systems*, New York: Springer-Verlag.

Brook, D., and Evans, D. A. (1972), "An Approach to the Probability Distribution of CUSUM Run Lengths," *Biometrika*, 59, 539–549.

Caceres, R., Duffield, N. G., Horowitz, J., and Towsley, D. (1999), "Multicast-Based Inference of Network Internal Loss Characteristics," *IEEE Transactions on Information Theory*, 45, 2462–2480.

Cao, J., Davis, D., Wiel, S. V., and Yu, B. (2000), "Time-Varying Network Tomography: Router Link Data," *Journal of the American Statistical Association*, 95, 1063–1075.

Castro, R., Coates, M. J., Liang, G., Nowak, R., and Yu, B. (2004), "Internet Tomography: Recent Developments," *Statistical Science*, 19, 499–517.

Chaloner, K., and Verdinelli, I. (1995), "Bayesian Experimental Design: A Review," *Statistical Science*, 10, 273–304.

Chernoff, H. (1953), "Locally Optimal Designs for Estimating Parameters," *The Annals of Mathematical Statistics*, 24, 586–602.

Chen, Y., Bindel, D., and Katz, R. H. (2003), "Tomography-Based Overlay Network Monitoring," in *Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement 2003*, pp. 216–231.

Chvatal, V. (1979), "A Greedy Heuristic for the Set-Covering Problem," *Mathematics of Operations Research*, 4, 233–235.

Coates, M. J., Hero, A., Nowak, R. M., and Yu, B. (2002), "Internet Tomography," *IEEE Signal Processing Magazine*, 19, 47–65.

_____ (2003), ???.

Coates, M. J., and Nowak, R. (2000), "Network Loss Inference Using Unicast End-to-End Measurement," in *Proceedings of the ITC Conference on IP Traffic, Modelling and Management*, Monterey, CA.

Crowder, S. V. (1987), "A Simple Method for Studying Run-Length Distributions of Exponentially Weighted Moving Average Charts," *Technometrics*, 29, 401–407.

Dempster, A. P., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Liang, G., and Yu, B. (2003), "Maximum Pseudo-Likelihood Estimation in Network Tomography," *IEEE Transactions on Signal Processing*, ??, ???–???

Lo Presti, F., Duffield, N. G., Horowitz, J., and Towsley, D. F. (2002), "Multicast-Based Inference of Network-Internal Delay Distributions," *IEEE/ACM Transactions on Networking*, 10, 761–775.

Lo Presti, F., Paxson, V., and Towsley, D. F. (2001), "Inferring Link Loss Using Striped Unicast Probes," in *Proceedings of IEEE Infocom 2001*, Anchorage, Alaska, April 22–26, 2001.

Marchette (2001), ???.

Meeke, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: Wiley.

Multicast-Based Inference of Network Internal Characteristics (MINC), available at: <http://www.research.att.com/projects/minc/>.

Network simulator, available at <http://www.isi.edu/nsnam/ns>.

Nowak (2001), ???.

Pukelsheim, F. (1993), *Optimal Design of Experiments*, New York: Wiley.

Ringer, G. C., and Prabhu, S. S. (1996), "A Markov Chain Model for the Multivariate Exponentially Weighted Moving Averages Control Chart," *Journal of the American Statistical Association*, 91, 1701–1706.

Tanner, M. A. (1996), *Tools for Statistical Inference*, New York: Springer-Verlag.

Tsang, Y., Coates, M. J., and Nowak, R. (2003), "Network Delay Tomography," *IEEE Transactions on Signal Processing*, 51, 2125–2136.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59

60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118

1	Vardi, Y. (1996), "Network Tomography: Estimating Source-Destination Traffic Intensities From Link Data," <i>Journal of the American Statistical Association</i> , 91, 365–377.	60
2		61
3	Walrand, J., and Varaiya, P. (1999), <i>High-Performance Communications Networks</i> , New York: Morgan Kaufmann.	62
4		63
5		64
6		65
7		66
8		67
9		68
10		69
11		70
12		71
13		72
14		73
15		74
16		75
17		76
18		77
19		78
20		79
21		80
22		81
23		82
24		83
25		84
26		85
27		86
28		87
29		88
30		89
31		90
32		91
33		92
34		93
35		94
36		95
37		96
38		97
39		98
40		99
41		100
42		101
43		102
44		103
45		104
46		105
47		106
48		107
49		108
50		109
51		110
52		111
53		112
54		113
55		114
56		115
57		116
58		117
59		118