Multicast-Based Inference of Network-Internal Loss Characteristics*

R. Cáceres† N.G. Duffield‡ J. Horowitz§ D. Towsley¶

Abstract

Robust measurements of network dynamics are increasingly important to the design and operation of large internetworks like the Internet. However, administrative diversity makes it impractical to monitor every link on an end-to-end path. At the same time, it is difficult to determine the performance characteristics of individual links from end-to-end measurements of unicast traffic. In this paper, we introduce the use of end-to-end measurements of *multicast* traffic to infer network-internal characteristics. The bandwidth efficiency of multicast traffic makes it suitable for large-scale measurements of both end-to-end and internal network dynamics.

We develop a Maximum Likelihood Estimator for loss rates on internal links based on losses observed by multicast receivers. It exploits the inherent correlation between such observations to infer the performance of paths between branch points in the tree spanning a multicast source and its receivers. We derive its rate of convergence as the number of measurements increases, and we establish robustness with respect to certain generalizations of the underlying model. We validate these techniques through simulation and discuss possible extensions and applications of this work.

1 Introduction

Background and Motivation. Fundamental ingredients in the successful design, control and management of networks are mechanisms for accurately measuring their performance. Two approaches to evaluating network performance have been:

(i) Collecting statistics at internal nodes and using network management packages to generate link-level performance reports; and

(ii) Characterizing network performance based on end-to-end behavior of point-to-point traffic such as that generated by TCP or UDP.

A significant drawback of the first approach is that gaining access to a wide range of routers in an administratively diverse network can be difficult. Introducing new measurement mechanisms into the routers themselves is likewise difficult because it requires persuading large companies to alter their products. Also,

*This work was sponsored in part by the DARPA and the Air Force Research Laboratory under agreement F30602-98-2-0238.

†AT&T Labs–Research, Rm. B125, 180 Park Avenue, Florham Park, NJ 07932, USA; E-mail: ramon@research.att.com

‡AT&T Labs–Research, Rm. B139, 180 Park Avenue, Florham Park, NJ 07932, USA; E-mail: duffield@research.att.com

§Dept. of Math. & Statistics, University of Massachusetts Amherst, MA 01003-4515, USA; E-mail: joeh@math.umass.edu

¶Dept. of Computer Science, University of Massachusetts, Amherst, MA 01003-4610, USA; E-mail: towsley@cs.umass.edu

the composition of many such small measurements to form a picture of end-to-end performance is not completely understood.

Regarding the second approach, there has been much recent experimental work to understand the phenomenology of end-to-end performance (e.g., see [1, 2, 14, 19, 21, 22]). A number of ongoing measurement infrastructure projects (Felix [5], IPMA [7], NIMI [13] and Surveyor [28]) aim to collect and analyze end-to-end measurements across a mesh of paths between a number of hosts. **pathchar** [10] is under evaluation as a tool for inferring link-level statistics from end-to-end point-to-point measurements. However, much work remains to be done in this area.

Contribution. In this paper, we consider the problem of characterizing link-level loss behavior within a network through end-to-end measurements. We present a new approach based on the measurement and analysis of the loss behavior of *multicast* probe traffic. The key to this approach is that multicast traffic introduces correlation in the end-to-end losses measured by receivers. This correlation can, in turn, be used to infer the loss behavior of the links within the multicast routing tree spanning the sender and receivers. This enables the identification of links with higher loss rates as candidates for the origin of the degradation of end-to-end performance.

Using this approach, we develop *maximum likelihood estimators* (MLEs) of the link loss rates within a multicast tree connecting the sender of the probes to a set of receivers. These estimates are, initially, derived under the assumption that link losses are described by independent Bernoulli losses, in which case the problem is that of estimating the link loss rates given the end-to-end losses for a series of n probes. We show that these estimates are strongly consistent (converge almost surely to the true loss rates). Moreover, the asymptotic normality property of MLEs allows us to derive an expression for their rate of convergence to the true rates as n increases.

We evaluate our approach for two-, four-, and eight-receiver populations through simulation in two settings. In the first type of experiment, link losses are described by time-invariant Bernoulli processes. Here we find rapid convergence of the estimates to their actual values as the number of probes increases. The second type of experiment is based on **ns** [18] simulations where losses are due to queue overflows as probe traffic competes with other traffic generated by infinite data sources that use the Transmission Control Protocol (TCP) [24]. In the two- and four- receiver topologies with few background connections we find fast convergence although there are persistent, if small, differences between the inferred and actual loss rates.

The cause of these differences is that losses in our simulated network display spatial dependence (i.e., dependence between links), which violates the Bernoulli assumption. We believe that large and long-lasting spatial dependence is unlikely in a real network such as the Internet because of its traffic and link diversity. This is supported by experiments with an eight-receiver topology with diverse background traffic in which we found closer agreement between inferred and actual loss rates. Furthermore, we believe that the intro-

duction of Random Early Detection (RED) [6] policies in Internet routers will help break such dependence.

The potential for both spatial and temporal dependence of loss motivates investigation into their effect. Our analysis shows that dependence introduces inference errors in a continuous manner: if the dependence small, the errors in the estimates are also small. Furthermore, the errors are a second order effect: in the special case of a binary tree with statistically identical dependent loss on sibling links, the Bernoulli MLE of the marginal loss rates are actually unaffected for interior links of the tree. More generally, the MLE will be insensitive to spatial dependence of loss within regions of similar loss characteristics. Furthermore, the analysis shows how prior knowledge of the likely magnitude of dependence–e.g. from independent network measurements–could be used to correct the Bernoulli MLE.

We note that interference from TCP sources introduces temporal dependence (i.e., dependence between different packets) that also violates the Bernoulli assumption. This dependence is apparent in our simulated network, where probe losses often occur back-to-back due to burstiness in the competing TCP streams. Such dependence has also been measured in the Internet, but rarely involves more than a few consecutive packets [1]. The consistency of the estimator does not require independence between probes; it is sufficient that the loss process be ergodic. This property holds, e.g., when the dependence between losses has sufficiently short range. However, the rate of convergence of the estimates to their true values will be slower. We quantify this for Markovian losses by applying the Central Limit Theorem for the occupation times of Markov processes. We use this approach to compare the efficacy of two sampling strategies in the presence of Markovian losses. In our experiments, inferred loss rates closely tracked actual losses rates despite the presence of temporal dependence.

The work presented in this paper assumes that the topology of the multicast tree is known in advance. We are presently developing algorithms to infer the multicast tree from the probe measurements themselves.

We envisage deploying inference engines as part of a measurement infrastructure comprising hosts exchanging probes in a WAN. Each host will act as the source of probes down a multicast tree to the others. A strong advantage of using multicast rather than unicast traffic is efficiency. N multicast servers produce a network load that grows at worst linearly as a function of N. On the other hand, the exchange of unicast probes can lead to local loads which grow as N^2, depending on the topology. We illustrate this in Figure 1. In this example, $2N$ servers exchange probes. For unicast probes, the load on central link grows as N^2; for multicast probes it grows only as $2N$.

Related Work. There are a number of measurement infrastructure projects in progress, all based on the exchange of unicast probes between hosts in the current Internet. Two of these, IPMA (Internet Performance Measurement and Analysis) [7] and Surveyor [28], focus on measuring loss and delay statistics; in the former between public Internet exchange points, in the latter between hosts deployed at sites participating in Internet 2. A third, Felix [5], is developing linear decomposition techniques to discover network topology,

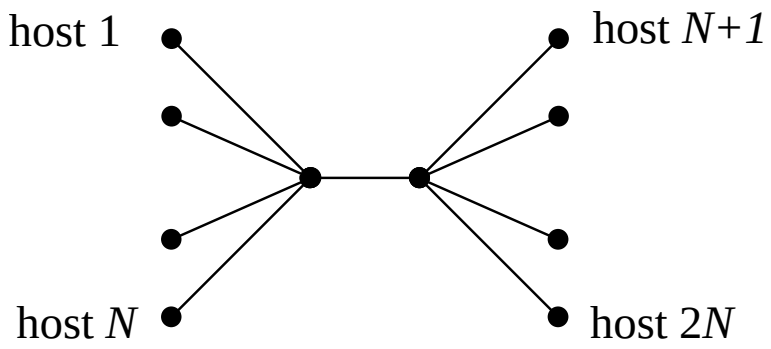

Figure 1: PROBE METHOD, LOAD AND TOPOLOGY: $2N$ servers exchange probes. For unicast probes, load on central link grows as N^2; for multicast probes it grows only as $2N$.

with an emphasis on network survivability. A fourth, NIMI (National Internet Measurement Infrastructure) [13], concentrates on building a general-purpose platform on which a variety of measurements can be carried out. These infrastructure efforts emphasize the growing importance of network measurements and help motivate our work. We believe our multicast-based techniques would be a valuable addition to these measurement platforms.

There is a multicast-based measurement tool, **mtrace** [16], already in use in the Internet. **mtrace** reports the route from a multicast source to a receiver, along with other information about that path such as per-hop loss and delay statistics. Topology discovery through **mtrace** is performed as part of the **tracer** tool [12].

However, **mtrace** suffers from performance and applicability problems in the context of large-scale measurements. First, **mtrace** traces the path from the source to a single receiver by working back through the multicast tree starting at that receiver. In order to cover the complete multicast tree, **mtrace** would need to run once for each receiver, which does not scale well to large numbers of receivers. In contrast, the inference techniques described in this paper cover the complete tree in a single pass. Second, **mtrace** relies on multicast routers to respond to explicit measurement queries. Current routers support these queries. However, Internet service providers may choose to disable this feature since it gives anyone access to detailed delay and loss information about paths in their part of the network. In contrast, our inference techniques do not rely on cooperation from any network-internal elements.

We now turn our attention to related theoretical work on inference methodologies. There has been some ad hoc, statistically non-rigorous work on deriving link-level loss behavior from end-to-end multicast measurements. An estimator proposed in [33] attributes the absence of a packet at a set of receivers to loss on the common path from the source. However, this is biased, even as the number of probes n goes to infinity.

For a different problem, some analytic methods for inference of traffic matrices have been proposed quite recently [30, 31]. The focus of these studies was to determine the intensities of individual source-destination flows from measurements of aggregate flows taken at a number of points in a network. Although

there are formal similarities in the inference problems with those of the present paper, the problem addressed by the other papers was substantially different. The solutions are not always unique or easily identifiable, sometimes needing supplementary methods to identify a candidate solution. This was a consequence of a combination of the coarseness of the data (average data rates in the class of Poissonian traffic processes) and the generality of the network topology considered.

Structure of the Paper. The remainder of the paper is structured as follows. In Section 2 we present a loss model for multicast trees and describe the framework within which analysis will occur. Section 3 contains the derivation of the estimators themselves; the specific example of the two-leaf tree is worked out explicitly. Section 4 analyzes the rates of convergence of estimators as the number of probes is increased. In particular, we obtain a simple approximation for estimator variance in the regime of small loss probabilities. In Section 5 we present an algorithm for computing packet loss estimates, and tests for consistency of the data with the model. Section 6 presents the results of simulation experiments that validate our approach. Motivated in part by the experimental results, we continue by examining the effects of violation of the Bernoulli assumption. In Section 7 we analyze the effects of spatial dependence on our estimators. We show how to correct for them on the basis of some a priori knowledge of their magnitude; we show that in any case they deform the estimates based on the Bernoulli assumption only to second order. In Section 8 we analyze the effect of temporal dependence on the loss process. We show that the asymptotic accuracy of the Bernoulli-based estimator is unaffected, although it may converge more slowly. We conclude in Section 9 with a summary of our contributions and proposals for further work. Some of the proofs are deferred to Section 10.

2 Model & Framework

2.1 Description of Logical Multicast Trees

Let $\mathcal{T} = (V, L)$ denote the logical multicast tree from a given source, consisting of the set of nodes V, including the source and receivers, and the set of links L. A link is ordered pair $(j, k) \in V \times V$ denoting a link from node j to node k. The set of children of a node j is denoted by $d(k)$ (i.e. $d(j) = \{k \in V : (j, k) \in L\}$). For each node $j \in V$ apart from the root 0, there is a unique node $k = f(j)$, the parent of j, such that $(j, k) \in L$. We shall define $f^n(k)$ recursively by $f^n(k) = f(f^{n-1}(k))$. We say that j is a descendant of k if $k = f^n(j)$ for some integer $n > 0$.

The root $0 \in V$ will represent the source of the probes. The set of leaf nodes $R \subset V$ (those with no children) will represent the set of receivers. The logical multicast tree has the property that every node has at least two descendants, apart from the root node (which has one) and the leaf-nodes (which have none). On the other hand, nodes in the full (as opposed to logical) multicast tree can have only one descendant. The logical multicast tree is obtained from the full multicast tree by deleting all nodes which have a single child (apart from the root 0) and adjusting the links accordingly. More precisely, if $i = f(j) = f^2(k)$ are

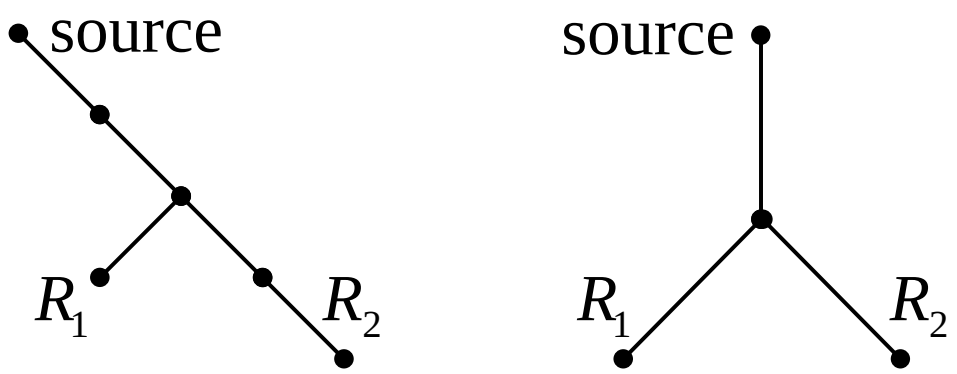

Figure 2: (a) A multicast tree with two receivers. (b) The corresponding logical multicast tree.

nodes in the full tree and $\#d(j) = 1$, then we assign to the logical tree only the nodes i, k and the link (i, k). Applying this rule to all such i, j and k in the full multicast tree yields the logical multicast tree.

A two receiver example is illustrated in Figure 2. A source multicasts a sequence of probes to two receivers, R_1 and R_2. The probes traverse the multicast tree illustrated in Figure 2(a). Figure 2(b) illustrates the logical multicast tree, where each path between branch points in the tree depicted in Figure 2(a) has been replaced by a single logical link.

2.2 Modeling the Loss of Probe Packets

We model the loss of probe packets on the logical multicast tree by a set of mutually independent Bernoulli processes, each operating on a different link. Losses are therefore independent for different links and different packets. In the introduction we discussed the reasons why network traffic can be expected to violate these assumptions; in Sections 7 and 8 we discuss the extent to which they affect the estimators described below, and how these effects can be corrected for.

We now describe the loss model in more detail. With each node $k \in V$ we associate a probability $\alpha_k \in [0, 1]$ that a given probe packet is not lost on the link terminating at k. We model the passage of probes down the tree by a stochastic process $X = (X_k)_{k \in V}$ where each X_k takes a value in $\{0, 1\}$; $X_k = 1$ signifies that a probe packet reaches node k, and 0 that it does not. The packets are generated at the source, so $X_0 = 1$. For all other $k \in V$, the value of X_k is determined as follows. If $X_k = 0$ then $X_j = 0$ for the children j of k (and hence for all descendants of k). If $X_k = 1$, then for j a child of k, $X_j = 1$ with independent probability α_j, and $X_j = 0$ with probability $\overline{\alpha}_j = 1 - \alpha_j$. (We shall write $1 - a$ as $\overline{a}$ in general). Although there is no link terminating at 0, we shall adopt the convention that $\alpha_0 = 1$, in order to avoid excluding the root link from expressions concerning the α_k. We display in Figures 3 and 4 examples of two- and four-leaf logical multicast trees which we shall use for analysis and experiments.

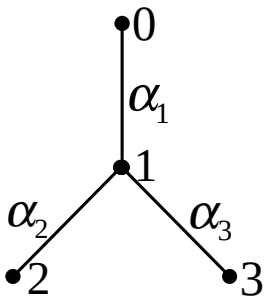

Figure 3: A two-leaf logical multicast tree

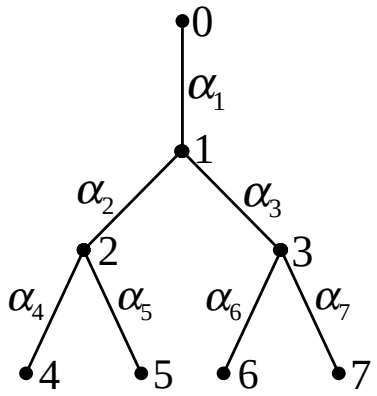

Figure 4: A four-leaf logical multicast tree

2.3 Data, Likelihood, and Inference

In an experiment, a set of probes is dispatched from the source. We can think of each probe as a trial, the outcome of which is a record of whether or not the probe was received at each receiver. Expressed in terms of the random process X, each such outcome is the set of values of X_k for k in the set of leaf nodes R, i.e. the random quantity $X_{(R)} = (X_k)_{k \in R}$, an element of the space $\Omega = \{0,1\}^R$ of all such outcomes. For a given set of link probabilities $\alpha = (\alpha_k)_{k \in V}$, the distribution of the outcomes $(X_k)_{k \in R}$ will be denoted by P_α. The probability mass function for a single outcome $x \in \Omega$ is $p(x;\alpha) = \mathsf{P}_\alpha(X_{(R)} = x)$.

Let us dispatch n probes, and, for each possible outcome $x \in \Omega$, let $n(x)$ denote the number of probes for which the outcome x obtained. The probability of n independent observations $x^1, \ldots, x^n$ (with each $x^m = (x_k^m)_{k \in R}$) is then

$$p(x^1, \ldots, x^n; \alpha) = \prod_{m=1}^{n} p(x^m; \alpha) = \prod_{x \in \Omega} p(x;\alpha)^{n(x)} \tag{1}$$

Our task is to estimate the value of α from a set of experimental data $(n(x))_{x \in \Omega}$. We focus on the class of *maximum likelihood estimators* (MLEs): i.e. we estimate α by the value $\breve{\alpha}$ which maximizes $p(x^1, \ldots, x^n; \alpha)$ for the data $x^1, \ldots, x^n$. Under very mild conditions, which are satisfied in the present situation, MLEs exhibit many desirable properties, including *strong consistency, asymptotic normality, asymptotic unbiasedness, and asymptotic efficiency* (see [11]). Strong consistency means that MLEs converge almost surely (i.e., with probability 1) to their target parameters as the sample size increases. The last three properties mean that, if the sample size is large, we can compute confidence intervals for the parameters at a given confidence level, the estimators are approximately unbiased, and there is no other estimator that would give the same level of precision with a smaller sample size.

Because of these properties, when a parametric model is available, MLEs are usually the estimators of choice. Moreover, the confidence intervals allow us to estimate the accuracy of the estimates of α, and in

particular their rate of convergence to the true parameter α as the number of samples n becomes large. This is important for understanding the number of probes which must be sent in order to obtain an estimate of α with some desired accuracy. Furthermore, in view of the possibility of large time-scale fluctuation in WANs, e.g. Internet routing instabilities as reported by Paxson [19], the period over which probes are sent should not be unnecessarily long.

3 The Analysis of the Maximum Likelihood Estimator

In this section we establish the form of the MLE and determine the rate at which it converges to the true value as the number of probes increases; this can be used to make prediction for given models, and also to estimate the likely accuracy of estimates derived from actual data. We work this out completely for the two-leaf tree of Figure 3.

3.1 The Likelihood Equation and its Solution

It is convenient to work with the log-likelihood function

$$\mathcal{L}(\alpha) = \log p(x^1, \ldots, x^n; \alpha) = \sum_{x \in \Omega} n(x) \log p(x; \alpha), \tag{2}$$

In the notation we suppress the dependence of $\mathcal{L}$ on n and $x^1, \ldots, x^n$. Since $\log$ is increasing, maximizing $p(x^1, \ldots, x^n; \alpha)$ is equivalent to maximizing $\mathcal{L}(\alpha)$.

We introduce the notation that $k \preceq k'$ for $k, k' \in V$ whenever k is a descendant of k' or $k = k'$ and $k \prec k'$ whenever $k \preceq k'$ but $k \neq k'$. We shall say that a link k is at level $\ell = \ell(k)$ if there is a chain of ℓ ancestors $k = f^0(k) \prec f(k) \prec f^2(k) \ldots \prec f^\ell(k) = 0$ leading back to the root 0 of $\mathcal{T}$. Levels 0 and 1 have only one node. We will occasionally use U to denote $V \setminus \{0\}$. Let $\mathcal{T}(k) = (V(k), L(k))$ denote the subtree within $\mathcal{T}$ rooted at node k. $R(k) = R \cap V(k)$ will be the set of receivers which are descended from k. Let $\Omega(k)$ be the set of outcomes x in which at least one receiver in $R(k)$ receives a packet, i.e.,

$$\Omega(k) = \{x \in \Omega : \bigvee_{j \in R(k)} x_j = 1\}. \tag{3}$$

Set $\gamma_k = \gamma_k(\alpha) = \mathsf{P}_\alpha[\Omega(k)]$. An estimate of γ_k is

$$\widehat{\gamma}_k = \sum_{x \in \Omega(k)} \widehat{p}(x), \quad \text{where} \quad \widehat{p}(x) := \frac{n(x)}{n}, \tag{4}$$

is the observed proportion of trials with outcome x. We will show that α can be calculated from $\gamma = (\gamma_k)_{k \in V}$, and that the MLE

$$\breve{\alpha} = \arg\max_{\alpha \in [0,1]^{\#R}} \mathcal{L}(\alpha) \tag{5}$$

can be calculated in the same manner from the estimates $\widehat{\gamma}$. The relation between α and γ is as follows. Define $\beta_k = \mathsf{P}[\Omega(k) \mid X_{f(k)} = 1]$. The β_k obey the recursion

$$\overline{\beta}_k = \overline{\alpha}_k + \alpha_k \prod_{j \in d(k)} \overline{\beta}_j, \quad k \in V \setminus R, \tag{6}$$

$$\beta_k = \alpha_k, \quad k \in R. \tag{7}$$

Then

$$\gamma_k = \beta_k \prod_{i=1}^{\ell(k)} \alpha_{f^i(k)}. \tag{8}$$

Theorem 1 *Let $\mathcal{A} = \{(\alpha_k)_{k \in U} : \alpha_k > 0\}$, and $\mathcal{G} = \{(\gamma_k)_{k \in U} : \gamma_k > 0 \, \forall k;\ \gamma_k < \sum_{j \in d(k)} \gamma_j \, \forall k \in U \setminus R\}$. There is a bijection Γ from $\mathcal{A}$ to $\mathcal{G}$. Moreover, Γ and Γ^{-1} are continuously differentiable.*

The proof of Theorem 1 relies of the following Lemma whose proof is given in Section 10.

Lemma 1 *Let C be the set of $c = (c_i)_{i=1,2,\ldots,i_{\max}}$ with $c_i \in (0,1)$ and $\sum_i c_i > 1$. The equation $(1-x) = \prod_i (1 - c_i x)$ has a unique solution $x(c) \in (0,1)$. Moreover, $x(c)$ is continuously differentiable on C.*

Proof of Theorem 1: The γ_k have been expressed as a function of the α_k, and clearly $\alpha_k > 0 \, \forall k \in U$ implies the conditions for $\mathcal{G}$. Thus it remains to show that the mapping from $\mathcal{A}$ to $\mathcal{G}$ is injective. Let $A_k = \prod_{i=0}^{\ell(k)} \alpha_{f^i(k)}$. From (8) we have

$$\gamma_k = A_k, \quad k \in R, \tag{9}$$

while combining (6) and (8) we find

$$H_k(A_k, \gamma) := (1 - \gamma_k/A_k) - \prod_{j \in d(k)} (1 - \gamma_j/A_k) = 0, \quad k \in U \setminus R. \tag{10}$$

Since $H_k(A_k, \gamma) = h(\gamma_k/A_k, \{\gamma_j/\gamma_k : j \in d(k)\})$ from Lemma 1, there is a unique $A_k > \gamma_k$ which solves (10). We recover the α_k uniquely from the A_k by taking the appropriate quotients (and setting $A_0 = \alpha_0 = 1$):

$$\alpha_k = A_k / A_{f(k)}, \quad k \in U. \tag{11}$$

Clearly Γ is continuously differentiable; that Γ^{-1} is also follows from the corresponding statement for $x(c)$ in Lemma 1. ■

Candidates for the MLE are solutions of the *likelihood equation* for the stationary points α of $\mathcal{L}$:

$$\frac{\partial \mathcal{L}}{\partial \alpha_k}(\alpha) = 0, \quad k \in U. \tag{12}$$

Theorem 2 *When $\widehat{\gamma} \in \mathcal{G}$, the likelihood equation has the unique solution $\widehat{\alpha} := \Gamma^{-1}(\widehat{\gamma})$.*

Note that in the notation we have suppressed the dependence of $\breve{\alpha}$ and $\widehat{\alpha}$ on n and $x^1, \ldots, x^n$. We defer the proof of Theorem 2 to Section 10. That done, we must complete the argument by showing that the stationary point does have maximum likelihood. For this we must impose additional conditions. $\widehat{\alpha}$ is not precluded from being either a minimum or a saddle for the likelihood function, the maximum falling on the boundary of $[0,1]^{\#U}$. For some simple topologies we are able to establish directly that $\mathcal{L}(\alpha)$ is (jointly) concave in the parameters at $\alpha = \widehat{\alpha}$, which is hence the MLE $\breve{\alpha}$. For more general topologies we use an argument which establishes that $\widehat{\alpha} = \breve{\alpha}$ for all sufficiently large n, and whose proof also establishes some useful asymptotic properties of $\widehat{\alpha}$.

If $\alpha_k = 0$ for some link k, then $X_k = 0$ for all $j \in R(k)$, regardless of the values of α_j for j descended from k, and hence these cannot be determined. For this reason we now restrict attention to the case that all $\alpha_k > 0$, by passing to a subtree if necessary; see Section 5.

Theorem 3 *Assume $\alpha_k \in (0,1], k \in U$.*

(i) The model is identifiable*, i.e., $\alpha, \alpha' \in (0,1]^{\#R}$ and $\mathsf{P}_\alpha = \mathsf{P}_{\alpha'}$ implies $\alpha = \alpha'$.*

(ii) As $n \to \infty$, $\breve{\alpha} \to \alpha$ and $\widehat{\alpha} \to \alpha$, P_α almost surely.

(iii) Assume also $\alpha_k < 1,\ k \in U$. With probability 1, for sufficiently large n, $\breve{\alpha} = \widehat{\alpha}$.

Maximum Likelihood Estimator for the Two-leaf Tree Denote the 4 points of $\Omega = \{0,1\}^2$ by $\{00, 01, 10, 11\}$. Then

$$\widehat{\gamma}_1 = \widehat{p}(11) + \widehat{p}(10) + \widehat{p}(01), \quad \widehat{\gamma}_2 = \widehat{p}(11) + \widehat{p}(10), \quad \widehat{\gamma}_3 = \widehat{p}(11) + \widehat{p}(01). \tag{13}$$

The equations (10) for $\widehat{A}_k$ in terms of the $\widehat{\gamma}_k$ can be solved explicitly; combining with (11) we obtain the estimates

$$\widehat{\alpha}_1 = \frac{\widehat{\gamma}_2\widehat{\gamma}_3}{\widehat{\gamma}_2 + \widehat{\gamma}_3 - \widehat{\gamma}_1} = \frac{(\widehat{p}(01) + \widehat{p}(11))(\widehat{p}(10) + \widehat{p}(11))}{\widehat{p}(11)} \tag{14}$$

$$\widehat{\alpha}_2 = \frac{\widehat{\gamma}_2 + \widehat{\gamma}_3 - \widehat{\gamma}_1}{\widehat{\gamma}_3} = \frac{\widehat{p}(11)}{\widehat{p}(01) + \widehat{p}(11)} \tag{15}$$

$$\widehat{\alpha}_3 = \frac{\widehat{\gamma}_2 + \widehat{\gamma}_3 - \widehat{\gamma}_1}{\widehat{\gamma}_2} = \frac{\widehat{p}(11)}{\widehat{p}(10) + \widehat{p}(11)} \tag{16}$$

Note that although it is possible that $\widehat{\alpha}_1 > 1$ for some finite n, this will not happen when n is sufficiently large, due to Theorem 3(ii).

There is a simple interpretation of the estimates in (15) and (16). With the $\widehat{p}$'s replaced by their corresponding true probabilities p, (15) would give the probability of receiving a probe at node 1, given that it known to be received at node 2. For independent losses, this is just the marginal probability that the probe is received at node 1. We have found, however, the corresponding formulas when there are more than 2 sibling nodes do not allow such a direct interpretation.

4 Rates of Convergence of Loss Estimator

4.1 Large Sample Behavior of the Loss Estimator

In this section we examine in more detail the rate of convergence of $\widehat{\alpha}$ and the MLE $\breve{\alpha}$ to the true value α. We can apply some general results on the asymptotic properties of MLEs in order to show that $\sqrt{n}(\breve{\alpha} - \alpha)$ is asymptotically normally distributed as $n \to \infty$; some general properties of MLEs ensure that the same hold for $\sqrt{n}(\widehat{\alpha} - \alpha)$, and with the same asymptotic variance. We can use these results in two ways. First, for models of loss processes with typical parameters, we can estimate the number of probes required to obtain an estimate with a given accuracy. Secondly, we can estimate the likely accuracy of $\widehat{\alpha}$ from actual probe data and associate confidence intervals to the estimates.

The fundamental object controlling convergence rates of the MLE $\breve{\alpha}$ is the *Fisher Information Matrix* at α. This is defined for each $\alpha \in (0,1)^{\#U}$ as the $\#U$-dimensional real matrix $\mathcal{I}_{jk}(\alpha) := \mathsf{Cov}\left(\frac{\partial \mathcal{L}}{\partial \alpha_j}(\alpha), \frac{\partial \mathcal{L}}{\partial \alpha_k}(\alpha)\right)$. It is straightforward to verify that $\mathcal{L}$ satisfies conditions (see Section 2.3.1 of [27]) under which $\mathcal{I}$ is equal to the following more convenient expression which we will use in the sequel:

$$\mathcal{I}_{jk}(\alpha) = -\mathsf{E}\frac{\partial^2 \mathcal{L}}{\partial \alpha_j \partial \alpha_k}(\alpha) \tag{17}$$

On the other hand, a direct calculation of the asymptotic variance of $\widehat{\alpha}$ follows from the Central Limit Theorem. The random variables $\widehat{\gamma}$ are asymptotically Gaussian as $n \to \infty$ with

$$\sqrt{n}\,(\widehat{\gamma} - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma), \tag{18}$$

where $\sigma_{jk} = \lim_{n\to\infty} n\,\mathsf{Cov}(\widehat{\gamma}_j, \widehat{\gamma}_k)$, for $j,k \in U$. Here $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. Since by Theorem 1, Γ^{-1} is continuously differentiable on $\mathcal{G}$, then by the Delta method (see Chapter 7 of [27]) $\widehat{\alpha} = \Gamma^{-1}(\widehat{\gamma})$ is also asymptotically Gaussian, so establishing the first part of the following theorem. We note that the matrices ν and $\mathcal{I}^{-1}(\alpha)$ agree on the interior of the parameter space, but, as we shall see below, $\mathcal{I}(\alpha)$ may be singular on the boundary. Let $D_{ij}(\alpha) = \frac{\partial \Gamma_i^{-1}}{\partial \gamma_j}(\Gamma(\alpha))$ and D^T denotes the transpose.

Theorem 4 *(i) When $\alpha_k \in (0,1]$, $k \in U$, then as $n \to \infty$,*

$$\sqrt{n}\,(\widehat{\alpha} - \alpha) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \nu), \quad \text{where} \quad \nu = D(\alpha) \cdot \sigma \cdot D^T(\alpha). \tag{19}$$

(ii) When $\alpha_k \in (0,1)$, $k \in U$ then $\mathcal{I}(\alpha)$ is non-singular and $\mathcal{I}^{-1}(\alpha) = \nu$.

(iii) When $\alpha_k \in (0,1)$, $k \in U$, $\sqrt{n}\ (\breve{\alpha} - \alpha)$ converges in distribution as $n \to \infty$ to a $\#U$-dimensional Gaussian random variable with mean 0 and covariance matrix $\mathcal{I}^{-1}(\alpha)$.

Theorem 4 enables us to determine, for example, that asymptotically for large n, with probability $1 - \delta$, the $\widehat{\alpha}$ will lie between the points

$$\alpha_k \ \pm\ z_{\delta/2}\sqrt{\frac{\mathcal{I}_{kk}^{-1}(\alpha)}{n}}, \tag{20}$$

where $z_{\delta/2}$ denotes the number that cuts off an area $\delta/2$ in the right tail of the standard normal distribution. This is used for a confidence interval of level $1-\delta$. As we are interested in a 95% confidence interval for single link measurements, we take $z_{\delta/2} \approx 2$.

Confidence Intervals for Parameters. With slight modification, the same methodology can be used to obtain confidence intervals for the parameters $\widehat{\alpha}$ derived from measured data from n probes. Following [4] we use the *observed Fisher Information*:

$$\widehat{\mathcal{I}}_{jk}(\widehat{\alpha}) = -\frac{\partial^2 \mathcal{L}}{\partial \alpha_j \partial \alpha_k}(\widehat{\alpha}), \quad \text{where} \quad \widehat{\alpha} = \Gamma^{-1}(\widehat{\gamma}). \tag{21}$$

Now, the proof of Theorem 2 (see particularly (57)) shows that the $\partial \mathcal{L}/\partial \alpha_k$ depend on the $n(x)$ only through the combinations $n\widehat{\gamma}_k$. Hence the same is true for the $\partial^2 \mathcal{L}/\partial \alpha_j \partial \alpha_k$. Since $\mathsf{P}_{\widehat{\alpha}}[\Omega(k)] = \Gamma(\Gamma^{-1}(\widehat{\gamma}))_k = \widehat{\gamma}_k$, we have $\widehat{\mathcal{I}}(\widehat{\alpha}) = \mathcal{I}(\widehat{\alpha})$.

We then use confidence intervals for $\widehat{\alpha}_k$ of the form

$$\widehat{\alpha}_k \ \pm \ z_{\delta/2} \sqrt{\frac{\mathcal{I}^{-1}_{kk}(\widehat{\alpha})}{n}}. \tag{22}$$

This allows us to find simultaneous confidence regions from the asymptotic distribution for α for a given tree. An issue for further study is to understand how the confidence intervals change as the tree grows.

Example: Confidence Intervals for the Two-leaf Tree An elementary calculation shows that the inverse of the Fisher information matrix governing the confidence intervals for models in (20) is

$$\mathcal{I}^{-1}(\alpha) = \begin{pmatrix} \frac{\alpha_1(\overline{\alpha}_3 - \alpha_2(1+\alpha_3(\alpha_1-2)))}{\alpha_2\alpha_3} & \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_3} & \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_2} \\ \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_3} & \frac{\overline{\alpha}_2\alpha_2}{\alpha_1\alpha_3} & \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_1} \\ \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_2} & \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_1} & \frac{\overline{\alpha}_3\alpha_3}{\alpha_1\alpha_2} \end{pmatrix}. \tag{23}$$

Here, the order of the coordinates is $\alpha_1, \alpha_2, \alpha_3$. The inverse of the observed Fisher information governing the confidence intervals for data in (22) is obtained by inserting (14)–(16) into (23). We note that in this case $\mathcal{I}$ is singular at the boundaries $\alpha_2 = 1$ and $\alpha_3 = 1$.

4.2 Dependence of Loss Estimator Variance on Topology

The variance of $\widehat{\alpha}$ determines the number of probes which must be used in order to obtain an estimate of a given desired accuracy. Thus it is important to understand how the variance depends on the underlying topology. Growth of the variance with the size of the tree might preclude application of the estimator to large internetworks. Long timescale instability has been observed in the Internet [19]; if the timescale required for accurate measurements approaches that at which variability occurs, the estimator's requirement of stationarity would be violated. In this section we show that the asymptotic variance ν of $\widehat{\alpha}$ is independent of topology for loss ratios approaching zero.

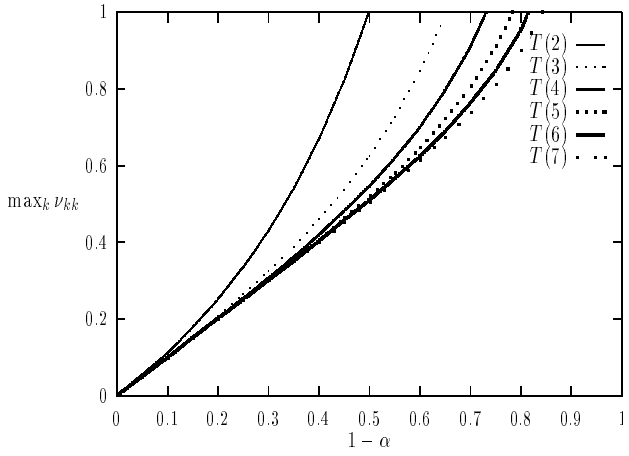

Figure 5: ASYMPTOTIC ESTIMATOR VARIANCE AND BRANCHING RATIO Depth 2 tree, 2 to 7 leaves. Variance decreases towards linear approximation $1-\alpha$ as branching ratio increases.

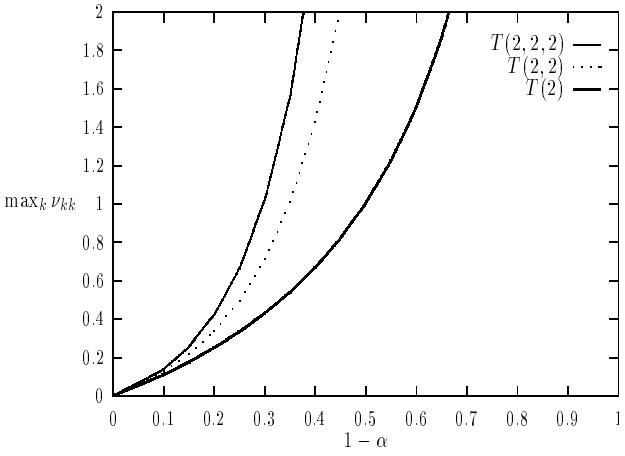

Figure 6: ASYMPTOTIC ESTIMATOR VARIANCE AND TREE DEPTH Binary Tree of depth 2, 3 or 4. Variance increases with tree depth.

The following theorem characterizes the behavior of ν for small loss ratio, independently of the topology of the logical tree. Set $\|\overline{\alpha}\| = \max_{k \in U} \overline{\alpha}_k$. Set $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.

Theorem 5 $\nu_{jk} = \overline{\alpha}_k \delta_{jk} + O(\|\overline{\alpha}\|^2) \quad$ as $\|\overline{\alpha}\| \to 0$.

Theorem 5 says that the variance of $\widehat{\alpha}$ is, to first order in $\overline{\alpha}$, independent of topology. However, nothing is said about higher order dependence, and in particular whether the difference between ν_{jk} and $\overline{\alpha}_k \delta_{jk}$ converges to zero uniformly for all topologies as $\overline{\alpha} \to 0$. For a section of trees we used computer algebra to calculate the maximum asymptotic variance over links $\max_k \nu_{kk}$ for a selection of trees, as a function of the uniform Bernoulli probability $\alpha_k = \alpha$. We use the notation $T(r_1, r_2, \ldots, r_n)$ denote the tree of depth $n+1$ (depth = maximum level ℓ of any leaf) with successive branching ratios $1, r_1, r_2, \ldots, r_n$, i.e. the root node 0 has the single descendent node 1 which has r_1 descendents, each of which has r_2 descendents, and so on. We show the dependence on branching ratio in Figure 5 for trees of depth 2. In these examples, increasing the branching ratio decreases the variance. In Figure 6, we show the dependence on tree depth for binary trees of depth 2, 3 and 4. In this example, estimator variance increases with tree depth, roughly linearly. In all examples, estimator variance is approximately linear for $\overline{\alpha}$ less than about 0.1, and independent of topology, in keeping with Theorem 5. For larger $\overline{\alpha}$ it appears from these examples that the change in estimator variance of moving from simple topologies to more complex ones is governed by two opposing effects; variance reduction with increasing branching ratio, and variance growth with increasing tree depth. The reason for this appears to be that increasing the branching ratio increases the size of $R(k)$ (the set of leaf-nodes descended from k) so providing more data points for estimation, while increasing the tree depth increases cumulative error per link in estimation.

5 Data Consistency and Parameter Computation

In this section we address computational issues associated with the estimator $\widehat{\alpha}$. We specify consistency checks which must be applied to the data before $\widehat{\alpha}$ is computed. We describe an algorithm for computation of $\widehat{\alpha}$ and discuss its suitability for implementation in a network, in particular the extent to which it is distributable.

5.1 Data Consistency

In this section we describe tests for consistency of the empirical probabilities $\widehat{\gamma}$ with the model. The validations of the methodology carried out in this paper are all within controlled simulations. So we do not address here the additional consistency checks which would be required for applications to real network data, such as tests for stationarity.

The rest of this section focuses on range checking and tree surgery. An arbitrary data set $(n(x))_{x\in\Omega}$ may not give rise to $\widehat{\gamma} \in \Gamma((0,1)^{\#U})$. If this is because some of the $\widehat{\gamma}_k$ take values 0 or 1, then it can be dealt with by reducing the tree. In particular, when one of the $\widehat{\gamma}_k$ is 0, not all of the α_k can be inferred from the data. Those which cannot must be removed from consideration. In other cases, the data is not consistent with the assumptions that loss occurs independently on different links. We discuss these now.

(i) If $\widehat{\gamma}_k = 0$ for any $k \in V$, we construct a new tree by deleting node k and all its descendants, and perform the analysis on this pruned tree instead. We are unable to distinguish between the various ways in which γ_k may be zero, e.g. $\alpha_k = 0$, or $\alpha_k > 0$ but $\alpha_j = 0$ for children $j \in d(k)$.

(ii) If $\widehat{\alpha}_k = 1$ for any $k \in U$ then we can assign probability 1 to α_k. Then, for the purposes of calculation only, we consider a reduced tree obtained by excising node k in the same manner as nodes with a single descendant are excised from the physical multicast tree to generate the logical multicast tree; see Section 2.1.

(iii) Any $\widehat{\alpha}_k > 1$ is a nonphysical value, since the link probabilities are required to lie in $[0,1]$ (subject to (i) and (ii) above). Theorem 3 tells us this will not occur for sufficiently large n. Thus in implementations of the inference algorithm, this event may be used to trigger the dispatch of further probes.

(iv) The condition $\widehat{\gamma}_k = \sum_{j\in d(k)} \widehat{\gamma}_j$ for any $k \in U \setminus R$ prevents the calculation of $\widehat{A}_k$ and hence also link probabilities for links that include k as a vertex, namely $\widehat{\alpha}_k = \widehat{A}_k/\widehat{A}_{f(k)}$ and $\widehat{\alpha}_j = \widehat{A}_j/\widehat{A}_k$ for $j \in d(k)$. Instead, we estimate only the probabilities $\{\alpha_k\alpha_j : j \in d(k)\}$ on the composite links from $f(k)$ to the elements of $d(k)$, estimating $\widehat{\alpha_k\alpha_j} = A_j/A_{f(k)}, \quad j \in d(k)$. The possibility $\widehat{\gamma}_k > \sum_{j\in d(k)} \widehat{\gamma}_k$. is precluded by the relations (25) and (26) below. Equality occurs only if the observed losses satisfy the strong dependence property that each packet reaching a receiver in $R(k)$ reaches no other receiver in $R(k)$.

5.2 Computation of the Estimator on a General Tree

In this section we describe the algorithm for computing $\widehat{\alpha}$ on a general tree. An important feature of the calculation is that it can be performed recursively on trees. First we show how to calculate the $\widehat{\gamma}_k$. Denote by $(\widehat{X}_k(i))_{k\in R, i=1,2,..,n}$ the measured values at the leaf nodes of process X for n. Define the binary quantities $(\widehat{Y}_k(i))_{k\in V, i=1,2,..,n}$ recursively by

$$\widehat{Y}_k(i) = \widehat{X}_k(i), \quad k \in R \tag{24}$$

$$\widehat{Y}_k(i) = \bigvee_{j\in d(k)} \widehat{Y}_j(i), \quad k \in V \setminus R \tag{25}$$

so that

$$\widehat{\gamma}_k = n^{-1} \sum_{i=1}^{n} \widehat{Y}_k(i). \tag{26}$$

For simplicity we assume now that $\widehat{\gamma} \in \Gamma((0,1)^{\#U})$, so that, if necessary, steps (i) and (ii) of Section 5.1 have been performed on the data and/or the logical multicast tree in order to bring it to this form. The calculation of $\widehat{\alpha}$ can be done by another recursion. We formulate both recursions in pseudocode in Figure 7. The procedure **find_gamma** calculates the $\widehat{Y}_k$ and $\widehat{\gamma}_k$, assuming $\widehat{Y}_k$ initializes to $\widehat{X}_k$ for $k \in R$ and 0 otherwise. The procedure **infer** calculates the $\widehat{\alpha}_k$. The procedures could be combined. The full set of link probabilities is estimated by executing **main(**1**)** where node 1 is the single descendant of the root node 0.

Here, an empty product (which occurs when the first argument of **infer** is a leaf node) is understood to be zero. We assume the existence of a routine **solvefor** that returns the value of the first symbolic argument which solves the equation specified in its second argument. We know from Theorem 1 that under the conditions for $\widehat{\gamma}$ a unique such value exists.

5.3 Implementation of Inference in a Network

The recursive nature of the algorithm has important consequences for its implementation in a network setting. Observe that the calculation of $\widehat{\gamma}_k$ and A_k depends on X only through the $(\widehat{Y}_j)_{j\in d(k)}$. Put another way, if j is a child of k, the contribution to the calculation of $\widehat{\alpha}_k$ of all data measured at the set of receivers $R(j)$ descended from j, is summarized through $\widehat{Y}_j$. In a networked implementation this would enable the calculation to be localized in subtrees at a representative node, the computational effort at each node being at worst proportional to the depth of the tree (for the node that is the representative for all distinct subtrees to which it belongs).

Moreover, estimates from measurements at receivers $R(k)$ descended from a node k are consistent with those from the full set of receivers in the following sense. Executing **main(**k**)** yields the A_k calculated by **main(**1**)** as the value for $\widehat{\alpha}_k$. Thus is the effective probability that a probe traverse a (fictitious) link from the root 0 directly to k. But when the full inference **main(**1**)** is performed, it is not hard to see that the $\widehat{\alpha}$ obey $A_k = \prod_{i=0}^{\ell(k)} \widehat{\alpha}_{f^i(k)}$, i.e the probability of traversing the path from 0 to k without loss.

```
procedure main (k ) {
        find_gamma ( k ) ;
        infer ( k, 1 ) ;
}

procedure find_gamma ( k ) {
        foreach ( j ∈ d(k) ) {
                Ŷ_j = find_gamma ( j ) ;
                foreach ( i ∈ {1, ..., n} ) {
                        Ŷ_k[i] = Ŷ_k[i] ∨ Ŷ_j[i] ;
                }
        }
        γ̂_k = n^{-1} Σ_{i=1}^{n} Ŷ_k[i] ;
        return Ŷ_k ;
}

procedure infer ( k, A ) ;
        A_k = solvefor( A_k , (1 − γ̂_k/A_k) == ∏_{j∈d(k)} (1 − γ̂_j/A_k) );
        α̂_k = A_k/A ;
        foreach ( j ∈ d(k) ) {
                infer ( j , A_k ) ;
        }
}
```

Figure 7: PSEUDOCODE FOR INFERENCE OF LINK PROBABILITIES

6 Simulation Results

We evaluated our inference techniques through simulation and verified that they performed as expected. This work had two parts: *model simulations* and *TCP simulations*. In the model simulations, losses were determined by time-invariant Bernoulli processes. These losses follow the model on which we based our earlier analysis. In the TCP simulations, losses were due to queue overflows as multicast probes competed with other traffic generated by infinite TCP sources. We used TCP because it is the dominant transport protocol in the Internet [29]. The following two subsections describe our results from these two simulation efforts.

6.1 Model Simulations

Topology. For the model simulations, we used ad hoc software written in C++. We simulated the two tree topologies shown in Figures 3 and 4. Node 0 sent a sequence of multicast probes to the leaves. Each link exhibited packet losses with temporal and spatial independence. We could configure each link with a different loss probability that held constant for the duration of a simulation run. We fed the losses observed by the leaves to a separate Perl script that implements the inference calculation described earlier.

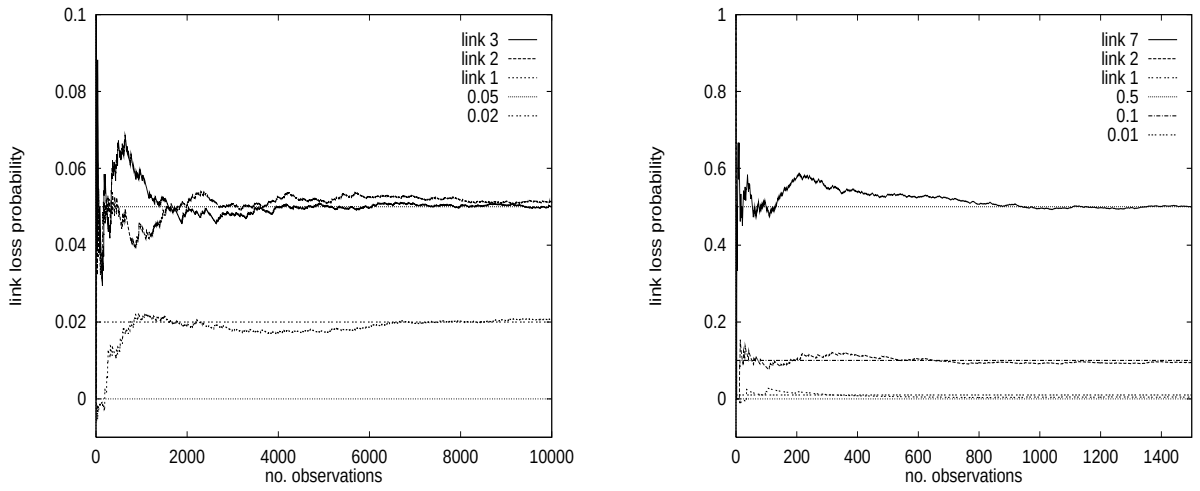

Figure 8: CONVERGENCE OF INFERRED LOSS PROBABILITIES TO ACTUAL LOSS PROBABILITIES IN MODEL SIMULATIONS. Left: Two-leaf tree of Figure 3 with parameters $\overline{\alpha}_1 = 0.02$; $\overline{\alpha}_2 = \overline{\alpha}_3 = 0.05$. Right: Selected links from four-leaf tree of Figure 4, with parameters $\overline{\alpha}_1 = 0.01$; $\overline{\alpha}_2 = 0.1$; $\overline{\alpha}_3 = \overline{\alpha}_4 = \overline{\alpha}_5 = \overline{\alpha}_6 = 0.01$; $\overline{\alpha}_7 = 0.5$. The graphs show that inferred probabilities converge to within 0.01 of the actual probabilities after 2,000 or fewer observations.

Convergence. Figure 8 compares inferred packet loss probabilities to actual loss probabilities. The left graph shows results for all three links in our two-leaf topology, while the right graph shows results for selected links in the four-leaf topology. In all cases, the inferred probabilities converge to within 0.01 of the actual probabilities after 2,000 observations.

Figure 9 compares the empirical and theoretical 95% confidence intervals of the inferred loss probabilities for the two-leaf topology. The empirical intervals were calculated over 100 simulation runs using 100 different seeds for the random number generator that underlies the Bernoulli processes. The theoretical intervals are as predicted by (20). As shown, simulation matches theory extremely well – we show the two graphs separately because the two sets of curves are indistinguishable when plotted together. For 2,000 observations, the confidence intervals lie with within 20% of the true probabilities.

It may seem that thousands of probes constitute too many network resources to expend and too long to wait for a measurement. However, it is important to note that a stream of 200-byte packets every 20 ms represents only 10 Kbps, equivalent to a single compressed audio transfer. Furthermore, a measurement using 5,000 such packets lasts less than two minutes. There already exist a number of MBone "radio" stations that send long-lived streams of sequenced multicast packets. In some cases we can use these existing multicast streams as measurement probes without additional cost. Overall, we feel that multicast-based inference is a practical and robust way to measure network dynamics.

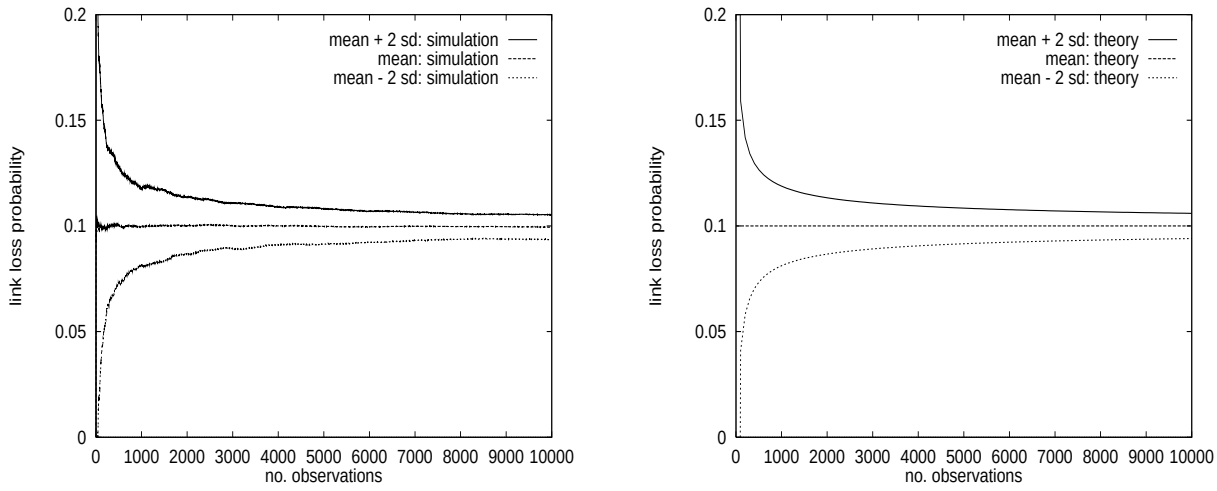

Figure 9: AGREEMENT BETWEEN SIMULATED AND THEORETICAL CONFIDENCE INTERVALS. Left: Results from 100 model simulations. Right: Predictions from (20). The graphs show two-sided confidence estimates at 2 standard deviations for link 2 of the four-leaf tree of Figure 4. Parameters were $\overline{\alpha}_1 = 0.01$; $\overline{\alpha}_2 = 0.1$; $\overline{\alpha}_3 = \overline{\alpha}_4 = \overline{\alpha}_5 = \overline{\alpha}_6 = 0.01$; $\overline{\alpha}_7 = 0.5$. Simulation matches theory extremely well – the two sets of curves are indistinguishable when plotted in the same graph.

6.2 TCP Simulations

Topology. For the TCP simulations, we used the **ns** network simulator [18]. We configured **ns** to simulate tree topologies shown in Figures 3, 4 and 11. All links had 1.5 Mbps of bandwidth, 10 ms of propagation delay, and were served by a FIFO queue with a 4-packet limit. Thus, a packet arriving at a link was dropped when it found four packets already queued at the link.

In each topology, node 0 sent multicast probe packets generated by a source with 200-byte packets and interpacket times chosen randomly between 2.5 and 7.5 msec. The leaf nodes received the multicast packets and monitored losses by looking for gaps in the sequence numbers of arriving probes. We fed the losses observed by the multicast receivers to the same inference implementation used for the model simulations described above. We also had **ns** report losses on individual links in order to compare inferred losses with actual losses.

In the two- and four-receiver topologies, each node maintained TCP connections to its child nodes. These connections used the Tahoe variant of TCP, sent 1,000-byte packets, and were driven by an infinite data source. Links to left children carried one such TCP stream, while links to right children carried two TCP streams. The link between nodes 0 and 1 also carried one TCP stream.

In the eight-receiver topology, the traffic more more diverse, with 52 TCP connections between different pairs of nodes, giving rise to approximately 8 connections per link on average.

Convergence. Figure 10 compares inferred loss rates to actual loss rates on selected links of our two- and

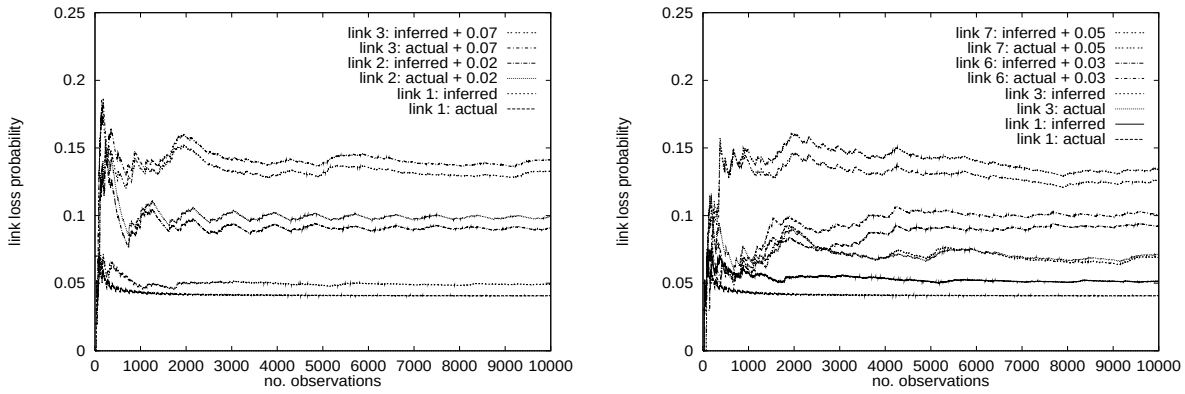

Figure 10: TRACKING OF ACTUAL LOSS RATES BY INFERRED LOSS RATES IN TCP SIMULATIONS. Left: Two-leaf tree of Figure 3. Right: Selected links from four-leaf tree of Figure 4 (some pairs of probabilities are offset for clarity). The graphs show that the inferred loss rates closely track the actual loss rates over 10,000 observations.

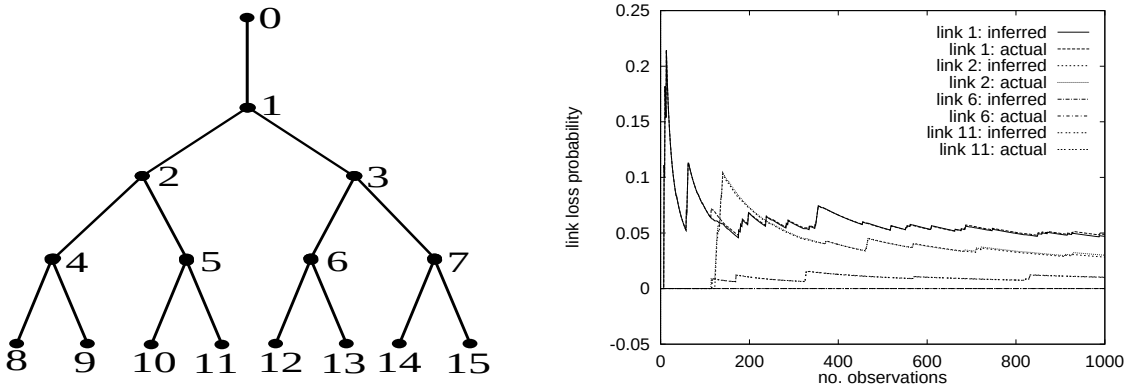

Figure 11: TRACKING OF ACTUAL LOSS RATES BY INFERRED LOSS RATES IN TCP SIMULATIONS WITH DIVERSE BACKGROUND TRAFFIC. LEFT: Eight-leaf binary tree. RIGHT: Close tracking of actual loss rates by estimated loss rates as number of observations is increased up to 1,000.

four-leaf topologies. As shown, the inferred rates closely track the actual rates over 10,000 observations. Figure 11 compares inferred and actual loss rates in the eight-receiver topology with diverse background traffic; in this case the tracking is even closer.

We note that the inferred values are accurate even though queue overflows due to TCP interference do not obey our temporal independence assumption. TCP is a bursty packet source, particularly in the region of exponential window growth during a slow start [9]. In our simulations, multicast probes are often lost in groups as they compete for queue space with TCP bursts. This phenomenon is readily apparent when watching animations of our simulations with the **nam** tool [17]. Inspection of the autocorrelation function of the time series of packet losses for a series of experiments predominantly showed correlation indistinguishable from zero beyond a lag of 1 (i.e. greater than back-to-back losses). As we explain in more

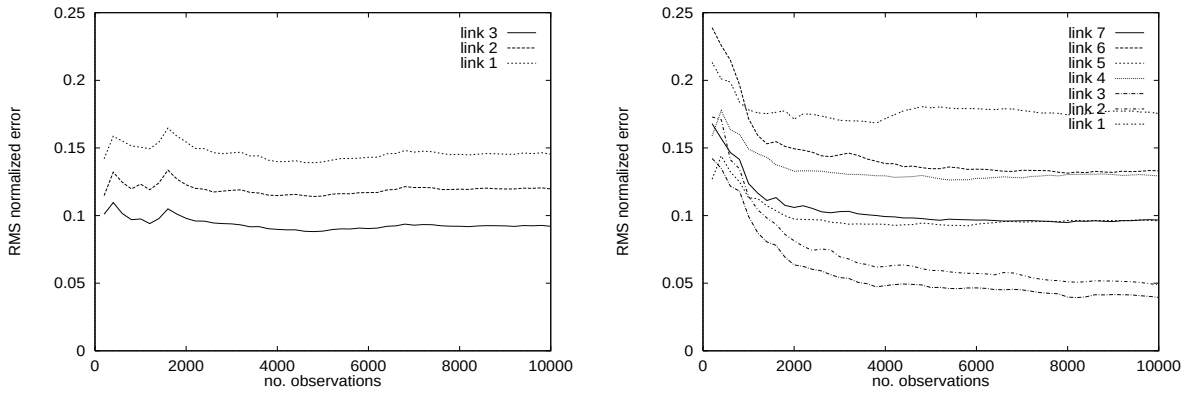

Figure 12: ACCURACY OF INFERENCE IN TCP SIMULATIONS. Left: Two-leaf tree of Figure 3. Right: Four-leaf tree of Figure 4. The graphs show normalized root mean square differences between actual and inferred loss rates, computed across 100 simulations. After an initial transient, inferred loss rates settle down to within 8 to 15% (in the two-leaf tree) and 4 to 18% (in the four-leaf tree) of actual loss rates, depending on the link. The RMS error was reduced to approximately 1% by modifying the MLEs to correct for spatial loss dependence.

detail in Section 8, the estimator $\widehat{\alpha}$ is still asymptotically accurate for large numbers of probes when losses have temporal dependence of sufficiently short range. However, the rate of convergence of the estimates to their true values will be slower.

Figure 12 shows the Root Mean Square (RMS) differences between the inferred and actual loss rates in the two- and four-leaf topologies. These differences were calculated over 100 simulation runs using 100 different seeds for the random number generator that governs the time between probe packets. As shown, the differences can drop significantly during the first 2,000 observations. However, at some point they level off and do not drop much further, if at all. This persistence reveals a systematic, although small, error in the inferred values because of spatial loss dependence. In our simulations, the same multicast probe is lost on sibling links more often than the spatial independence assumption dictates. These dependent losses lead the inference calculation to underestimate losses on the sibling links and to overestimate losses on the parent link.

We can quantify the spatial loss dependence present in the simulations. We can also calculate the effect of such dependence on the inferred loss probabilities by extending our previous analysis. Thus a prior estimate of the degree of dependence could be used to obtain corrections to the Bernoulli inference. We discuss this in more detail for spatial dependence in Section 7 and give an example of how to apply the correction. Applied to the inferences on the two-leaf tree summarized in Figure 10, they reduce an RMS error of between 8 and 15% to one of around 1%. The key observation behind these analyzes is that the error in the inferred values varies smoothly with the degree of spatial dependence. The greater the dependence in the network, the larger the error. We can arrange for correlated losses in a simulated network,

for example by creating synchronized interference streams on sibling links. However, the results for the eight-receiver topology with diverse background traffic support our belief that large and long-lasting spatial loss dependence is unlikely in real networks like the Internet because of their traffic and link diversity.

7 The Analysis and Correction of Spatial Dependence

7.1 Analysis of Spatial Dependence

When spatial dependence present in packet losses, the Bernoulli model assumption is violated. But even with such dependence, we can still ask what are the *marginal* loss probabilities for each link separately. In this section we quantify the effects of this dependence and show how they may be corrected for on the basis of a priori knowledge of them. We propose that this knowledge should be obtained by independent measurements on instrumented networks. Moreover, we establish that dependence deforms the Bernoulli estimates *continuously* in the sense that small divergences from independence of the losses lead to small divergence of the estimates of the marginal loss probabilities from their true values. For binary trees we find that the effect of such dependence on the estimates of marginal loss probabilities for links in the interior of the network is second order, and become negligible in regions of the network across which loss and dependence change little.

One motivation for considering dependent losses comes from the well-known example of synchronization between TCP flows which can occur as a result of the slow-start after packet loss; see [9]. Flows which have experienced common loss on a link k will then have some degree of dependence. Viewed as background traffic against which the probe packets compete, they can be expected to give rise to dependent losses of probe packets on links on the subtree descended from k. However, the dependence of probe loss can be expected to decrease on progressing down the tree from k. This happens if we assume that flows which became dependent though losses a given node k typically have a spread of destination address; then their paths through the network will subsequently diverge. Then the fraction of the total traffic contributed on links descended from k will decrease on progressing down the tree from k; hence the dependent influence of such flows on probe loss will decrease likewise.

The foregoing discussion motivates us to capture such dependence to first order by considering, within the class of dependent loss processes, those for which dependence only occurs between losses on sibling links, i.e., between those X_j and $X_{j'}$ for which $f(j) = f(j')$. Let $\Delta = \{\{j_1, \ldots j_n\} \subset d(k), k \in V \setminus R\}$ denote the set of subsets of sibling links. We characterize the joint distribution of the $(X_k)_{k\in V}$ through the family of joint conditional probabilities $(\alpha_{k_1,\ldots,k_n})_{\{j_1,\ldots j_n\}\in\Delta}$ where for $k = f(j_1) = \ldots = f(j_n)$,

$$\alpha_{j_1,\ldots,j_n} := \mathsf{P}[X_{j_1} = 1, \ldots, X_{j_n} = 1 | X_k = 1] \tag{27}$$

(For Bernoulli loss, $\alpha_{j_1,\ldots,j_n} = \prod_{m=1}^{n} \alpha_{j_m}$). We now derive analogous relations to (6) in this case. It is

convenient to work initially with the quantities

$$\xi_k := \mathsf{P}[\Omega(k) \mid X_k = 1] = \mathsf{P}[\Omega(k) \mid X_{f(k)} = 1]/\mathsf{P}[X_k = 1 \mid X_{f(k)} = 1] = \beta_k/\alpha_k \tag{28}$$

For $n \leq \#d(k)$ let $d_n(k)$ denote the set of subsets of $d(k)$ of cardinality n. By the Inclusion-Exclusion Principle (see e.g. Chapter 5.2 of [25])

$$\mathsf{P}[\Omega(k)] = \mathsf{P}\left[\cup_{j\in d(k)}\Omega(j)\right] = \sum_{n=1}^{\#d(k)} (-1)^{n+1} \sum_{\{j_1,\ldots,j_n\}\subset d_n(k)} \mathsf{P}\left[\Omega(j_1)\cap\ldots\cap\Omega(j_n)\right], \tag{29}$$

from which we find using (27) and (28) that

$$\xi_k = \sum_{n=1}^{\#d(k)} (-1)^{n+1} \sum_{\{j_1,\ldots,j_n\}\subset d_n(k)} \alpha_{j_1,\ldots,j_n}\xi_{j_1}\ldots\xi_{j_n} \tag{30}$$

Reexpressed in term of the γ_k we obtain the following analog of (10) for $k \in U \setminus R$:

$$H_k(A_k,\gamma,\psi) := \gamma_k/A_k - \sum_{n=1}^{\#d(k)} (-1)^{n+1} \sum_{\{j_1,\ldots,j_n\}\subset d_n(k)} \psi_{j_1,\ldots,j_n} \frac{\gamma_{j_1}\cdots\gamma_{j_n}}{A_k^n} = 0 \tag{31}$$

where $\psi_{j_1,\ldots,j_n} = \alpha_{j_1,\ldots,j_n}/(\alpha_{j_1}\ldots\alpha_{j_n})$ and we write $\psi = (\psi_{j_1,\ldots,j_n})_{\{j_1,\ldots,j_n\}\in\Delta}$. For a given loss model one can in principle compute ψ and compute A_k from γ_k. Rather than do this, however, we establish some structural results.

We can compare the actual values $A_k(\psi)$ which solve (31) for A_k, with those obtained from (10) with the Bernoulli assumption, which we can write as $A_k(1)$. The following theorem shows that the deformation from $A_k(\psi)$ to $A_k(1)$ is continuous in the neighborhood of the Bernoulli values $\psi = 1$ (i.e. $\psi_{j_1,\ldots,j_n} = 1$ for all $\{j_1,\ldots,j_n\} \in \Delta$).

Theorem 6 *Let $\alpha_k > 0$. There exists a neighborhood of $\psi = 1$ in $\mathbb{R}^{\#\Delta}$ on which $\psi \mapsto A_k(\psi)$ is continuous.*

Proof of Theorem 6: The result then follows from the Implicit Function Theorem (see [26]) provided that $\partial_{A_k} H_k(A_k(1),\gamma,1) \neq 0$. But $H_k(A_k,\gamma,1) = H_k(A_k,\gamma) = h(\gamma_k/A_k, \{\gamma_j/\gamma_k : j \in d(k)\})$ appearing in (10) and Lemma 1, and so the result follows from $\partial_x h(x(c),c) < 0$ as established during the proof of Lemma 1. ■

7.2 Spatially Dependent Losses in Binary Trees

When $\mathcal{T}$ is a binary tree we can obtain explicit results. For $k \in U\setminus R$ write $\psi^{(k)} = \psi_{jj'}$ where $d(k) = \{j,j'\}$ Then from (31) we have

$$\gamma_k = \begin{cases} A_k, & k \in R \\ \gamma_j + \gamma_{j'} + \psi^{(k)}\gamma_j\gamma_{j'}/A_k, & k \in U \setminus R \end{cases} \tag{32}$$

Let $\alpha(\psi)$ be the true value of α, i.e. that obtained by combining (32) with (11). $\alpha(1)$ is then the value previously obtained using the Bernoulli assumption. Let $k = 1$ denote the single descendent of the root node 0.

Theorem 7 *Let $\mathcal{T}$ be a binary tree.*

(i) There is a bijection Γ_ψ from $\mathcal{A}$ to $\mathcal{G}$ such that $\Gamma_\psi^{-1}(\gamma) = \alpha(\psi)$, with $\Gamma_1 = \Gamma$ from Theorem 1.

(ii)

$$\alpha_k(\psi) = \begin{cases} \alpha_1(1)/\psi^{(1)}, & k = 1 \\ \alpha_k(1)\psi^{(f(k))}, & k \in R \\ \alpha_k(1)\psi^{(f(k))}/\psi^{(k)}, & \text{otherwise} \end{cases} \tag{33}$$

Proof of Theorem 7: From (32), $A_k(\psi) = (\gamma_j + \gamma_{j'} - \gamma_k)/(\gamma_j\gamma_{j'}\psi^{(k)}) = A_k(1)/\psi^{(k)}$. The form of (ii) then follows from (11); this is used as the definition of Γ_ψ^{-1} for (i). ∎

Theorem 7(ii) has the interesting interpretation that in the interior of the network (i.e. except for node 1 and the leaf-nodes) the error in using $\alpha_k(\psi)$ in place of $\alpha_k(1)$ is a second order effect. For the error depends only on the on the relative magnitude of correlations at adjacent nodes through the quotient $\psi^{(f(k))}/\psi^{(k)}$. If the link probabilities and dependencies are (approximately) equal at each node of the tree, then this quotient will be (approximately) one, and so the Bernoulli estimate $\widehat{\alpha}_k(1) := \Gamma_1^{-1}(\widehat{\gamma})$ will be (approximately) equal to $\Gamma_\psi^{-1}(\widehat{\gamma})$, for interior k. Thus we see that the presence of dependent losses in binary trees perturbs the Bernoulli-based estimator little for links within the interior of regions across which the degree of dependence is similar. On the other hand, at the boundaries between such regions, a priori knowledge of the degree of dependence can help make the estimates more accurate. This motivates future work both in simulation studies and instrumentation of heterogeneous networks in order to establish the degree of dependence is influenced by dynamic factors such as utilization, and (comparatively) static factors such link technology and relative link speeds.

It is interesting to see that the TCP Simulations of the 4-leaf tree display some of the features one might expect from the above discussion. Observe in the RHS of Figure 10 that for the leaf-links (6 and 7) the inferred loss rate underestimates the actual loss rate, while for link 1 it overestimates it. For the interior link 3, the inferred and actual values are almost identical. This is consistent with the above discussion if $\psi_k > 1$ and $\psi_3 \approx \psi_{f(3)} = \psi_1$. Note that for $d(k) = \{j, j'\}$,

$$\psi_k > 1 \iff \alpha_{jj'} > \alpha_j\alpha_{j'} \iff \mathsf{E}[X_jX_{j'} \mid X_k = 1] > \mathsf{E}[X_j \mid X_k = 1]\mathsf{E}[X_{j'} \mid X_k = 1]. \tag{34}$$

In other words, $\psi_k > 1$ iff X_j and $X_{j'}$ are (conditional on $X_k = 1$) positively correlated. We expect this to be the case when synchronized losses occur as described at the start of this section.

	RMS difference from actual loss	
	adjusted	original
link 1	0.012	0.142
link 2	0.009	0.114
link 3	0.007	0.089

Table 1: CORRECTING FOR SPATIAL DEPENDENCE: RMS proportional difference of inferred from actual losses in **ns** simulation of two-leaf tree in Figure 3, after 10,000 probes. Adjustment of inference to account for dependence (left column) shows order of magnitude improvement over original inference (right column)

7.3 Correction for Spatial Dependence in Binary Trees

If some knowledge of the degree of dependence in the traffic is available, then this can be used to adjust the inferred loss probabilities accordingly. This motivates experimental studies of real networks with instrumented links in order to ascertain the magnitude of the dependence. We intend to undertake these experiments in the future. Here we show how knowledge of dependence can be used to correct the Bernoulli-based estimates of link probabilities for non-interior nodes. We consider the set of leaf-nodes $\{j, j'\} \in d(k)$. Let Y_j have the the distribution of X_j conditioned on $X_k = 1$. Suppose we know a priori an estimate $\widehat{\kappa}$ for the correlation of Y_j and $Y_{j'}$. Now the theoretical value of the correlation is

$$\kappa = \frac{\mathsf{Cov}(Y_j, Y_{j'})}{\sqrt{\mathsf{Var}(Y_j)\mathsf{Var}(Y_j)}} = \frac{\alpha_{jj'} - \alpha_j \alpha_{j'}}{\sqrt{\alpha_j \overline{\alpha}_j \alpha_{j'} \overline{\alpha}_{j'}}} = \psi^{(k)} \left(1 - \sqrt{\frac{\alpha_j \alpha_{j'}}{\overline{\alpha}_j \overline{\alpha}_{j'}}}\right) \tag{35}$$

Thus we expect to improve our estimates $\widehat{\alpha}_j(1)$ by using $\widehat{\alpha}_j(1)\widehat{\psi}^{(k)}$ instead where $\psi^{(k)}$ is obtained from (35) by using $\widehat{\kappa}$ and $\widehat{\alpha}(1)$ in place of κ and α.

To test this approach, we measured the loss dependence in an **ns** simulation of 10,000 probes in the two-leaf tree, then conducted 100 further **ns** simulations of 10,000 probes, and adjusted the inferred link probabilities in this manner. Comparing the actual, adjusted, and originally inferred loss ratios we see this provides improvement: the root mean square error goes down from between 8 and 15% (depending on the link) to about 1% in this case; see Table 1.

8 Temporal Dependence and Convergence Rates

8.1 Ergodicity and Asymptotic Accuracy

In this section we investigate the impact of temporal dependence on the estimator $\widehat{\alpha}$. Denote by $X(n) = (X_k(n))_{k \in V}$ the (spatial) process of the n^{th} probe. The first observation is that, if we replace the assumption of independence between probes to merely assuming that the (temporal) process $(X(n))_{n \in \mathbb{N}}$ is stationary and ergodic, then $\widehat{\alpha}$ still converges to α almost surely as the number of observations grows to ∞. This is because, by definition, the observed probabilities $\widehat{\gamma}$ of the ergodic process converge almost surely to the long

term averages. By stationarity, these are just the $\gamma = \Gamma(\alpha)$ as before, where the α are the (time)-marginal distributions of the link probabilities. A simple argument involving the Inverse Function Theorem (e.g., see [26]) shows that Γ^{-1} is continuous on $\Gamma((0,1)^{\#U})$, and hence $\widehat{\alpha} \to \alpha$ almost surely. Note we do not rely on $\widehat{\alpha}$ being the maximum likelihood estimator, with respect to some parameter space, for the marginal probabilities α of the general process. Rather, we have shown that the Bernoulli estimator is asymptotically accurate for stationary ergodic processes.

In the remainder of this section we examine the rate of convergence when X possesses temporal dependence. In an application of the method to measurement on real networks however, inherent variability (due do large scale events such as routing changes) may impose limits on the durations over which we can expect the loss process to be stationary. For this reason it is important to understand in more detail the impact of time-dependent packet loss on convergence rates. We propose to examine this through models. Markovian models of packet loss have been proposed on the basis of observations of the Internet (e.g., see [1]), although some longer bursts of losses were also found. We shall see that the price of temporal dependence is slower convergence than for the Bernoulli case. One can understand this qualitatively from the fact that burstiness in the packet loss processes means that the long-term average of $\widehat{\gamma}$ takes longer to approach.

8.2 Convergence Rates for Markovian Congestion

The main tool in understanding convergence rates is the following. Let Γ_k^{-1} denote the node k component of Γ^{-1}, so that $\widehat{\alpha}_k = \Gamma_k^{-1}(\widehat{\gamma})$. Suppose now that the random variables $\widehat{\gamma}$ are asymptotically Gaussian as $n \to \infty$ with

$$\sqrt{n}\,(\widehat{\gamma} - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma), \tag{36}$$

where $\sigma_{jk} = \lim_{n\to\infty} n\,\mathsf{Cov}(\widehat{\gamma}_j, \widehat{\gamma}_k)$, for $j, k \in U$. Here $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. Then by the Delta method (see Chapter 7 of [27]), since Γ_k^{-1} is continuously differentiable on $\mathcal{G}$ (see Theorem 1), $\Gamma_k^{-1}(\widehat{\gamma})$ is also asymptotically Gaussian:

$$\sqrt{n}\,(\Gamma_k^{-1}(\widehat{\gamma}) - \alpha_k) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \nu_k), \quad \text{where} \quad \nu_k = \nabla\Gamma_k^{-1}(\gamma) \cdot \sigma \cdot \nabla\Gamma_k^{-1}(\gamma). \tag{37}$$

In the remainder of this section we establish (36) within the context of Markov loss processes, and perform some explicit calculations for the 2-leaf tree.

We expand the class of loss processes as follows. We will define a Markov process $(Y(n))_{n\in\mathbb{N}}$, where $Y(n)$ will describe the state of the network encountered by the n^{th} probe; this description is used whether, for example, the interprobe times are constant, variable or random. Y is constructed as follows. For each $k \in U$ let $(Y_k(n))_{n\in\mathbb{N}}$ be an independent Markov process on the state space $\{0, 1\}$. We think of $Y_k(n)$ as representing the state of link k at time n, taking the value 0 if the link is congested, 1 if it is not. A probe that encounters a congested link is lost. We represent this by the process $X = (X_k(n))_{k\in U, n\in\mathbb{N}}$ defined by

letting $X_k(n)$ be conditionally independent of $(X_j(m), Y_j(m))_{k \prec j, m<n}$ given $(X_{f(k)}(n), Y_k(n))$, with

$$(X_k(n) \mid X_{f(k)}(n), Y_k(n)) = \begin{cases} 0, & X_{f(k)} = 0 \\ Y_k(n), & X_{f(k)} = 1 \end{cases} \tag{38}$$

When $Y_k(\cdot)$ is Bernoulli with probability α_k to be in the state 1, then the $X(n)$ are independent for each n, with the $X_k(n)$ distributed as described in Section 2.2. X is not a Markov process, but rather is a function of the Markov process Y. Moreover, $X(n)$ is a some function of $Y(n)$ alone, which we denote by χ. For each $k \in U$, let $\Phi(k)$ be the set of configurations y of Y such that $\chi(y)$ has outcome $\chi(y)_{(R)}$ in $\Omega(k)$, i.e.,

$$\Phi(k) = \{y \in \{0,1\}^{\#U} : \chi(y)_{(R)} \in \Omega(k)\}. \tag{39}$$

Let Q denote the transition matrix for Y, i.e., $Q = \otimes_{k \in U} Q(k)$ is the Kronecker product of the transition matrices of the individual Y_k. Let $q(k) = \{1 - \alpha_k, \alpha_k\}$ and let $q = \otimes_{k \in U} q(k)$ be the corresponding product distribution.

Theorem 8 *With the above notation, assume $\alpha_k \in (0,1)$ for all $k \in U$. Then (37) holds with*

$$\sigma_{jk} = \sum_{y \in \Phi(j)} \sum_{z \in \Phi(k)} \left[q_y(\delta_{yz} - q_z) + 2 \sum_{m=1}^{\infty} (Q^m_{yz} - q_y) q_z \right], \tag{40}$$

where Q^m denotes the m-step transition matrix.

Observe that in the Bernoulli case, the second term in (79) vanishes, while the first depends only on the marginal probabilities α. This means that the first term in (79) gives rise to the diagonal elements of (23); in what follows we can thus restrict our attention to the increase in the asymptotic variance as specified by the second term.

We parameterize the transition matrix of Y_k as

$$Q(k) = \begin{pmatrix} 1 - \alpha_k \overline{\omega}_k & \overline{\alpha}_k \overline{\omega}_k \\ \alpha_k \overline{\omega}_k & 1 - \overline{\alpha}_k \overline{\omega}_k \end{pmatrix}, \tag{41}$$

where $\overline{\omega}_k \in (0, 1/\max\{\alpha_k, \overline{\alpha}_k\}]$. ω_k parameterizes the burstiness of Y_k without changing its marginal probabilities. $Y_k(m)$ and $Y_k(m+1)$ are positively (or negatively) correlated when $\omega_k > 0$ (or $\omega_k < 0$). When $\omega_k = 0$, Y_k is Bernoulli. By calculation of the matrix powers of $Q(k)$ through its spectral decomposition, we find that $Q^n(k)_{yz} q_z(k)$ is given by the matrix

$$[Q^n(k)_{yz}] = \omega_k^n F(k) + G(k), \quad \text{where} \quad F(k) = \alpha_k \overline{\alpha}_k \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad G(k) = q(k) \otimes q(k). \tag{42}$$

Expanding $Q^n = \otimes_{k \in U} Q^n(k)$ and summing over n we find

$$\sum_{m=1}^{\infty} \left[(Q^m_{yz} - q_y) q_z \right] = \sum_{W \subset U} g(W) \left(\otimes_{k \in W} F(k) \right) \otimes \left(\otimes_{k \in U \setminus W} G(k) \right), \tag{43}$$

where $g(\emptyset) = 0$ and otherwise $g(W) = (\prod_{k \in W} \omega_k)/(1 - \prod_{k \in W} \omega_k)$.

8.3 Example: the Two-leaf Tree

Taking gradients in (14)–(16) and reexpressing them in terms of α we find

$$\nabla\Gamma_1^{-1}(\Gamma(\alpha)) = \frac{(1,-\overline{\alpha}_3,-\overline{\alpha}_2)}{\alpha_2\alpha_3}, \quad \nabla\Gamma_2^{-1}(\Gamma(\alpha)) = \frac{(-1,1,\overline{\alpha}_2)}{\alpha_1\alpha_3}, \quad \nabla\Gamma_3^{-1}(\Gamma(\alpha)) = \frac{(-1,\overline{\alpha}_3,1)}{\alpha_1\alpha_2}, \quad (44)$$

Using the notation (abc), with $a,b,c \in \{0,1\}$, to denote a value of $Y(n)$, we have from (13):

$$\Phi(1) = \{(111),(110),(101)\}, \quad \Phi(2) = \{(111),(110)\}, \quad \Phi(3) = \{(111),(101)\}. \quad (45)$$

For simplicity we set the α_k and ω_k equal to α,ω. Then (43) becomes

$$\begin{aligned}\sum_{m=1}^{\infty}\left[(Q_{yz}^m - q_y)q_z\right] &= \frac{\omega^3}{1-\omega^3}F(1)\otimes F(2)\otimes F(3) \quad (46)\\ &+\frac{\omega^2}{1-\omega^2}\left(F(1)\otimes F(2)\otimes G(3)+F(1)\otimes G(2)\otimes F(3)+G(1)\otimes F(2)\otimes F(3)\right)\\ &+\frac{\omega}{1-\omega}\left(F(1)\otimes G(2)\otimes G(3)+G(1)\otimes G(2)\otimes F(3)+G(1)\otimes F(2)\otimes G(3)\right).\end{aligned}$$

Combining (44), (45) and (46) in (37) in (46) with Theorem 8

$$\mathcal{I}_{11}^{-1} = \frac{\overline{\alpha}-\alpha(1+\alpha(\alpha-2))}{\alpha}, \quad (47)$$

$$\mathcal{I}_{22}^{-1} = \mathcal{I}_{33}^{-1} = \frac{\overline{\alpha}}{\alpha} \quad (48)$$

$$\nu_1 = \mathcal{I}_{11}^{-1} + \frac{\overline{\alpha}\omega(\alpha^2+\alpha\omega+\alpha^2\omega+\omega^2-\alpha\omega^2+2\alpha^2\omega^2+\omega^3-\alpha\omega^3+\alpha^2\omega^3)}{\alpha(1+\omega)(1-\omega^3)} \quad (49)$$

$$\nu_2 = \nu_3 = \mathcal{I}_{22}^{-1} + \frac{\overline{\alpha}\omega((\alpha+\omega)^2+\omega^2(\alpha^2+\omega))}{\alpha(1+\omega)(1-\omega^3)} \quad (50)$$

From (42), ω is the geometric decay rate of correlations. We can interpret $\tau = 1/(1-\omega)$ as the mean correlation time of the losses; $\tau = 1$ for Bernoulli losses. In Figure 13 we display the increase in asymptotic variance by plotting the ratio $\nu_1/\mathcal{I}_{11}^{-1}$ of the asymptotic variance with Markovian correlations to that without. We do this for $\alpha \in [.5,1]$ and $\tau \in [1,10]$. $\nu_2/\mathcal{I}_{22}^{-1}$ displayed very similar behavior. The ratio is increasing in correlation time τ, and in the link transmission probability α.

8.4 Temporal Dependence and Probing Methodology

An approach to avoiding the effect of temporal dependence would be to time probes at intervals larger than the typical correlation time of losses. Although this will reduce the number of probes required for a given level of convergence, the absolute time of convergence may increase due to the increased time between probes. Increasing the probes spacing by a factor τ', but with all probes lying within a given measurement period would increase the variance of the estimates by a factor τ' for independent losses. With Markovian losses, the effect of dependence between probes could be ameliorated by taking $\tau' > \tau$, the correlation

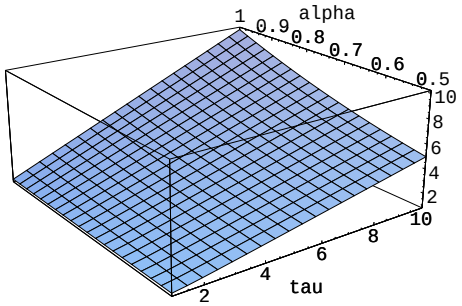

Figure 13: IMPACT OF TEMPORAL DEPENDENCE ON CONVERGENCE OF ESTIMATES: The ratio $\nu_1/\mathcal{I}_{11}^{-1}$ of the asymptotic variances of $\widehat{\alpha}_1$ with and without temporal dependence. Ratio is increasing in correlation time τ, and in link transmission probability α.

time. But for the two-leaf tree we see from (47) that when $\alpha \to 1$, then $\nu_k/\mathcal{I}_{kk}^{-1} \to 1/(1-\omega) = \tau$ for $k = 1, 2, 3$. Thus for small loss probabilities, the slow-down in the rate of convergence of $\widehat{\alpha}$ is no worse than that obtained by spacing probes to be approximately independent. In this example then, one may as well use all probes irrespective of their mutual dependence, rather than try to space them out to avoid dependence.

We envisage that direct measurement of the correlation time of received probes could be used, in combination with calculations of the previous section, to determine the number of probes, in an ongoing measurement, that are required in order to infer the link probabilities for a given accuracy. In the example considered we have seen that in order to estimate the increase in the asymptotic variation due to dependence between losses of small probability, it is sufficient to determine the correlation time of observed losses. When losses are heterogeneous, this will be conservative, since the autocorrelation will be dominated by the component with slowest decay.

A related issue is the randomization of interprobe times in order to avoid bias in the selection of network states which are observed via the probes. Probes with exponentially distributed spacings will see time averages; this is the PASTA property (Poisson Arrivals See Time Averages; see e.g. [32]). This approach has been proposed for network measurements [23] and is under consideration in the IP Performance Metrics working group of the IETF [8]. In the context of the above discussion, lengthening the interprobe time is to be understood as increasing the mean of the exponential distribution.

9 Summary and Future Work

In this paper, we introduced the use of end-to-end measurements of multicast traffic to infer network-internal characteristics. We developed statistically rigorous techniques for estimating packet loss rates on internal links, and validated these techniques through simulation. We showed that the inferred values quickly converged to within a small error of the actual values. We also presented evidence that our techniques yield accurate results even in the presence of moderate levels of temporal and spatial loss dependence.

We are extending our work in several directions. First, we are applying multicast-based inference to metrics other than packet loss. In particular, we have developed estimators for link delay. We are also investigating ways to infer link bandwidth and network topology using multicast probes. The ability to determine topology would free our measurements from the assumption of a priori knowledge of topology or of a separate topology-discovery tool.

Second, we plan to do more extensive simulations. We plan to substitute RED queueing for FIFO queueing to study the effect of RED on loss dependence. We also plan to substitute Poisson probes for CBR probes to avoid inadvertent synchronization of the probe traffic with periodic network processes. At the same time, we plan to simulate more complex topologies than the simple examples used throughout this paper. Topologies other than complete binary trees would stress our MLE for general trees, while larger topologies would test the convergence properties of our techniques on larger problem instances. This will be complemented by a theoretical analysis of the dependence of convergence rates on topology. Furthermore, we would like to explore how closely loss rates experienced by our probes agree with loss rates experienced by other network applications and protocols, for example TCP. We expect that our multicast-based measurements will yield ambient loss rates that are meaningful in a broad context.

Third, we plan to experiment with multicast-based inference on the Internet. As a preliminary step, we plan to measure ambient dependence in the real network, and determine the extent to which we need to adapt our estimates to their presence. We also plan to deploy our inference tools in multicast-enabled portions of the Internet, including the MBone, to test our techniques on a real network.

Finally, we would like to integrate our inference tools with one or more of the large-scale measurement infrastructures under construction. NIMI seems particularly suited because of its intended role as a general framework where many types of measurement can be carried out. The challenge will be to adapt a unicast-based infrastructure to perform multicast-based measurements, and in particular to schedule measurements, collect results, and perform inference calculations when large numbers of receivers are involved.

In conclusion, we feel that multicast-based inference is a powerful approach to measuring Internet dynamics. The rigorous statistical analysis behind our techniques gives them a firm theoretical footing, while the bandwidth efficiency of multicast traffic gives them much desired scalability. Robust and efficient measurements are increasingly important as the Internet continues to grow in size and diversity.

10 Proofs of Theorems

Proof of Lemma 1: Let $h_1(x) = (1-x)$, $h_2(x,c) = h_2(x) = \prod_i (1-c_i x)$. Let $q_i = c_i/(1-c_i x)$. Then for $x \in [0,1]$ $h_1''(x) = 0$, $h_2''(x) = h_2(x)\left\{(\sum_i q_i)^2 - \sum_i q_i^2\right\} > 0$. Hence $h(x) = h_1(x) - h_2(x)$ is strictly concave on $[0,1]$. Now $h(0) = 0$, $h(1) < 0$ and $h'(0) = -1 + \sum_i c_i > 0$. So since h is concave and continuous on $[0,1]$ there must be exactly one solution to $h(x) = 0$ for $x \in (0,1)$. Now set write

$h(x,c) = h_1(x) - h_2(x,c)$. Let $x(c)$ be the unique solution to $h(x(c),c) = 0$. The above derivation implies that $h'(x(c)) = (\partial h(x,c)/\partial x)|_{x=x(c)} < 0$, so in particular, is different from 0. Since h is continuously differentiable, then by the Implicit Function Theorem [26], so is $c \mapsto x(c)$. ∎

Proof of Theorem 2: The idea is to split up the sum (2) into portions on which $\frac{\partial \log p(x)}{\partial \alpha_k}$ is constant. These will be $\Omega(k)$, the $\Omega(f^i(k)) \setminus \Omega(f^{i-1}(k))$ for $i = 1, 2, \ldots, \ell(k)$, and $\Omega(0)^c$.

Consider first the case that $x \in \Omega(k)$. Then α_k occurs in $p(x)$ as a factor, and hence $\frac{\partial \log p(x)}{\partial \alpha_k} = 1/\alpha_k$. When $x \in \Omega(f^i(k)) \setminus \Omega(f^{i-1}(k))$ for $i = 1, 2, \ldots, \ell(k)$, then $p(x) = \overline{\beta}_{f^{i-1}(k)} R_k(x)$ where $R_k(x)$ does not depend on α_k (or indeed on any α_j for $j \leq f^{i-1}(k)$. Hence for $x \in \Omega(f^i(k)) \setminus \Omega(f^{i-1}(k))$,

$$\frac{\partial \log p(x)}{\partial \alpha_k} = \frac{1}{\overline{\beta}_{f^{i-1}(k)}} \frac{\partial \overline{\beta}_{f^{i-1}(k)}}{\partial \alpha_k} \tag{51}$$

Similarly, when $x \in \Omega(0)^c$,

$$\frac{\partial \log p(x)}{\partial \alpha_k} = \frac{1}{\overline{\beta}_0} \frac{\partial \overline{\beta}_0}{\partial \alpha_k} \tag{52}$$

On combining these:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \alpha_k} &= \frac{1}{\alpha_k} \sum_{x \in \Omega(k)} n(x) + \frac{1}{\overline{\beta}_0} \frac{\partial \overline{\beta}_0}{\partial \alpha_k} \sum_{x \in \Omega(0)^c} n(x) \\ &\quad + \sum_{i=1}^{\ell(k)} \left\{ \frac{1}{\overline{\beta}_{f^{i-1}(k)}} \frac{\partial \overline{\beta}_{f^{i-1}(k)}}{\partial \alpha_k} \sum_{x \in \Omega(f^i(k)) \setminus \Omega(f^{i-1}(k))} n(x) \right\} \end{aligned} \tag{53}$$

For the derivatives, some algebra with (7) shows that

$$\frac{\partial \overline{\beta}_k}{\partial \alpha_k} = -\beta_k/\alpha_k, \qquad \text{and} \tag{54}$$

$$\frac{\partial \overline{\beta}_{f^i(k)}}{\partial \alpha_k} = \frac{\alpha_{f^i(k)} - \beta_{f^i(k)}}{\overline{\beta}_{f^{i-1}(k)}} \frac{\partial \overline{\beta}_{f^{i-1}(k)}}{\partial \alpha_k} = -\frac{\beta_k}{\alpha_k} \prod_{m=1}^{i} \frac{\alpha_{f^m(k)} - \beta_{f^m(k)}}{\overline{\beta}_{f^{m-1}(k)}}. \tag{55}$$

The right hand term in equation (55) follows by iterating the middle term. Observe that

$$\sum_{x \in \Omega(f^i(k)) \setminus \Omega(f^{i-1}(k))} \frac{n(x)}{n} = \widehat{\gamma}_{f^i(k)} - \widehat{\gamma}_{f^{i-1}(k)} \quad \text{and} \quad \sum_{x \in \Omega(0)^c} \frac{n(x)}{n} = 1 - \widehat{\gamma}_0. \tag{56}$$

Combining (53), (54), (55) and (56) we get

$$\frac{\alpha_k}{n} \frac{\partial \mathcal{L}}{\partial \alpha_k} = \widehat{\gamma}_k - \beta_k \sum_{i=1}^{1+\ell(k)} \frac{\widehat{\gamma}_{f^i(k)} - \widehat{\gamma}_{f^{i-1}(k)}}{\overline{\beta}_{f^{i-1}(k)}} \prod_{m=1}^{i-1} \frac{\alpha_{f^m(k)} - \beta_{f^m(k)}}{\overline{\beta}_{f^{m-1}(k)}}. \tag{57}$$

Here we adopt the convention that the empty product for $i = 1$ means 1, and that the symbol $\widehat{\gamma}_{f(0)}$ that occurs when $i = 1 + \ell(k)$ means 1.

Set $\frac{\partial \mathcal{L}}{\partial \alpha_k}$ for all $k \in V$. For $k = 0$, (57) yields $0 = \widehat{\gamma}_0 - \beta_0(1 - \widehat{\gamma}_0)/\overline{\beta}_0$, whence

$$\widehat{\gamma}_0 = \beta_0 = \gamma_0. \tag{58}$$

For any other k, combining (57) for k and $j = f(k)$ yields

$$\widehat{\gamma}_k = \frac{\beta_k}{\overline{\beta}_k}\left(\widehat{\gamma}_j - \widehat{\gamma}_k + \frac{(\alpha_j - \beta_j)\widehat{\gamma}_j}{\beta_j}\right), \quad \text{whence} \quad \frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} = \frac{\beta_k \alpha_j}{\beta_j} = \frac{\gamma_k}{\gamma_j}. \tag{59}$$

Together with (58) this gives $\widehat{\gamma}_k = \gamma_k$ for all $k \in V$. ∎

Proof of Theorem 3: (i) By the strong law of large numbers, $\widehat{\gamma} \to \Gamma(\alpha)$, P_α almost surely, as $n \to \infty$. Since Γ is, in particular, bijective, then the model is identifiable, since $\Gamma(\alpha) = \Gamma(\alpha')$ implies $\alpha = \alpha'$.

(ii) Convergence of $\widehat{\gamma}$ to γ (from (i)) and continuity of Γ^{-1} (from Theorem 1) yield convergence of $\widehat{\alpha} = \Gamma^{-1}(\widehat{\gamma})$ to $\alpha = \Gamma^{-1}(\gamma)$ as $n \to \infty$. We now establish convergence of $\breve{\alpha}$. Fix some $\alpha^0 \in (0,1)^{\#U}$, $M \subset (0,1)^{\#U}$, $x \in \Omega$ and define

$$Z(M, x) = \inf_{\alpha' \in M} \log \frac{p(x;\alpha^0)}{p(x;\alpha')} = \log p(x;\alpha^0) - \sup_{\alpha' \in M} \log p(x;\alpha'). \tag{60}$$

Observe that $p(x;\alpha)$ is polynomial in the α_k, and hence continuous. According to Lemma 7.54 in [27], it suffices to show that, for each $\alpha' \neq \alpha^0$, there is an open set $N_{\alpha'}$ containing α', such that $\mathsf{E}_{\alpha^0} Z(N_{\alpha'}, X) > -\infty$. (Here E_{α^0} is the expectation w.r.t. P_{α^0}).

Look at the two terms in $\mathsf{E}_{\alpha^0} Z(M, X)$ for any $M \subset (0,1)^{\#U}$. The first is $\mathsf{E}_{\alpha^0} \log p(X;\alpha^0) = \sum_{x\in\Omega} p(x;\alpha^0) \log p(x;\alpha^0)$. This is finite since $p \log p$ is bounded for $p \in [0,1]$ and Ω is finite. For the second term, note that $p(x;\alpha') \leq 1 \Rightarrow \log p(x;\alpha') \leq 0 \Rightarrow \sup_{\alpha'\in M} \log p(x;\alpha') \leq 0 \Rightarrow -\sup_{\alpha'\in M} \log p(x;\alpha') \geq 0 \Rightarrow \mathsf{E}_{\alpha^0} Z(M, X) \geq \mathsf{E}_{\alpha^0} \log p(X;\alpha^0) > -\infty$. Finally, we note that although it is not mentioned there, Lemma 7.54 in [27] requires identifiability, which we proved in (i) above.

(iii) Now let $\alpha \in (0,1)^{\#U}$ be the true set of link probabilities. From part (ii), with P_α probability 1, the MLE $\breve{\alpha} \to \alpha$ as $n \to \infty$. Hence, for each sequence of probes we have that for n sufficiently large, $\breve{\alpha}$ lies in the interior of $(0,1)^{\#U}$. For such n, $\breve{\alpha}$ must then solve the likelihood equation (12). We know from Theorem 2, that solutions of the likelihood equation are unique, and hence this $\breve{\alpha} = \widehat{\alpha}$. ∎

Proof of Theorem 4: (ii) Recall $V(k) = \{j \in V : j \preceq k\}$, $R(k) = V(k) \cap R$ and $U = V \setminus \{0\}$. Set $S(\alpha) = (S_k(\alpha))_{k\in U}$ with $S_k(\alpha) = \frac{\partial \mathcal{L}}{\partial \alpha_k}(\alpha)$ (the score vector). Then $\mathcal{I}_{jk}(\alpha) = \mathsf{Cov}(S_j(\alpha), S_k(\alpha)) = \mathsf{E}_\alpha(S_j(\alpha) S_k(\alpha))$ since $\mathsf{E}_\alpha(S_\alpha) = \sum_{x\in\Omega} p(x,\alpha) \frac{\partial}{\partial \alpha_k} \log p(x,\alpha) = \sum_{x\in\Omega} \frac{\partial}{\partial \alpha_k} p(x,\alpha) = 0$.

Suppose that $\mathcal{I}(\alpha)$ is singular for some $\alpha = (\alpha_k)_{k\in U} \in (0,1)^{\#U}$. Then there exists some nonzero vector $c = (c_k)_{k\in U}$ for which $c \cdot \mathcal{I} \cdot c = 0$. But $c \cdot \mathcal{I} \cdot c$ is the variance of the mean-zero random variable $c \cdot S(\alpha)$, so then we would have that $c \cdot \mathcal{S}(\alpha) = 0$, P_α almost surely, or equivalently

$$\sum_{k\in U} c_k \frac{\partial \log p(x,\alpha)}{\partial \alpha_k} = 0 \quad \forall x \in \Omega \tag{61}$$

since $\mathsf{P}_\alpha(\{x\}) > 0$ for all $x \in \Omega$. We show that, in fact, (61) implies $c_k = 0$, first for $k \in R$, then for all $k \in U$.

Let $x^{(0)} \in \Omega$ be such that $x_j^{(0)} = 1$ for all $j \in R$, and for some $k \in R$ let $x_j^{(1)} = 1$ for $j \neq k$ and 0 for $j = k$. Then

$$p(x^{(0)}, \alpha) = \prod_{j \in U} \alpha_j \quad \text{while} \quad p(x^{(1)}, \alpha) = \overline{\alpha}_k \prod_{j \in U \setminus \{k\}} \alpha_j \tag{62}$$

and so from (61)

$$\sum_{j \in U} \frac{c_j}{\alpha_j} = 0 \quad \text{while} \quad -\frac{c_k}{\overline{\alpha}_k} + \sum_{j \in U \setminus \{k\}} \frac{c_j}{\alpha_j} = 0. \tag{63}$$

Combining the last two equations we find $c_k = 0$.

We now proceed by induction. For $k \in U$ assume that $c_j = 0$ for all $j \prec k$. We now prove that $c_k = 0$. Let $x^{(0)}$ be as before, and set

$$x_j^{(3)} = \begin{cases} 1 & j \in R \setminus R(k) \\ 0 & j \in R(k). \end{cases} \tag{64}$$

Then

$$p(x^{(3)}, \alpha) = (\overline{\alpha}_k + \alpha_k \phi_k) \prod_{j \in V \setminus V(k)} \alpha_j \tag{65}$$

where $\phi_k = \prod_{j \in d(k)} \overline{\beta}_j = \mathsf{P}_\alpha[X_j = 0 \, \forall j \in R(k) \mid X_k = 1]$. Hence from (61)

$$\frac{c_k(\phi_k - 1)}{\overline{\alpha}_k + \phi_k \alpha_k} + \sum_{j \in V \setminus V(k)} \frac{c_j}{\alpha_j} = 0, \tag{66}$$

recalling the assumption that $c_j = 0$ for all $j \prec k$. For the same reason (61) reads

$$\frac{c_k}{\alpha_k} + \sum_{j \in V \setminus V(k)} \frac{c_j}{\alpha_j} = 0. \tag{67}$$

Combining (66) and (67), then we find $c_k = 0$. The equality of ν with $\mathcal{I}^{-1}$ in the interior of the space of parameters α is standard under the conditions established during the proof of Theorem 3; see, e.g., Chapter 6.4 of [11].

(iii) We refer to Theorem 7.63 of [27]. Clearly $\mathcal{L}$ is 3-times continuously differentiable on $(0,1)^{\#U}$, and has bounded expectation in some neighborhood of α. This establishes the relation (7.64) in [27]. $\frac{\partial \log p(x,\alpha)}{\partial \alpha_j \alpha_k}(\alpha)$ is clearly finite on $(0,1)^{\#U}$. Hence $\mathcal{I}$ is finite in $(0,1)^{\#U}$, so together with Theorem 3 and the non-singularity of $\mathcal{I}$ established in (ii) above, we are able to conclude the result. ∎

Proof of Theorem 5: Let $j \vee k$ denote the nearest common ancestor of j and k, i.e. $j \vee k$ is the $\prec$-least common upper bound of j and k. The proof proceeds by a number of subsidiary results. Since probes are assumed independent, it suffices to evaluate all random quantities for $n = 1$ probes.

(i) As $\|\overline{\alpha}\| \to 0$,

$$\text{(a) } 1 - A_k = s(k) + O(\|\overline{\alpha}\|^2); \quad \text{(b) } \overline{\beta}_k = O(\|\overline{\alpha}\|), \quad \text{(c) } 1 - \gamma_k = s(k) + O(\|\overline{\alpha}\|^2), \tag{68}$$

where

$$s(k) = \sum_{j \succeq k} \overline{\alpha}_j. \tag{69}$$

The relation (a) is clear by expanding $A_k = \prod_{j \succeq k}(1 - \overline{\alpha}_j)$. (b) follows by an inductive argument. Observe from (6) that if (b) holds for all $k \in d(j)$, it also holds for j. But since $\beta_k = \alpha_k$ for leaf-nodes $k \in R$, (b) holds for all k. (c) then follows from the relation $\gamma_k = A_k(1 - \prod_{j \in d(k)} \overline{\beta}_j)$.

(ii) As $\|\overline{\alpha}\| \to 0$,

$$\mathsf{Cov}(\widehat{\gamma}_j, \widehat{\gamma}_k) = s(j \vee k) + O(\|\overline{\alpha}\|^2) \tag{70}$$

To see this, we write $\mathsf{Cov}(\widehat{\gamma}_j, \widehat{\gamma}_k) = \mathsf{E}[\widehat{\gamma}_j\widehat{\gamma}_k] - \mathsf{E}[\widehat{\gamma}_j]\mathsf{E}[\widehat{\gamma}_k]$, and $\mathsf{E}[\widehat{\gamma}_j] = \gamma_j$ by definition. If k is an ancestor of j then $\widehat{\gamma}_j = 1 \Rightarrow \widehat{\gamma}_k = 1$ and so $\mathsf{E}[\widehat{\gamma}_j\widehat{\gamma}_k] = \gamma_j$. Similarly, if j is an ancestor of k, then $\mathsf{E}[\widehat{\gamma}_j\widehat{\gamma}_k] = \gamma_k$. Otherwise $\widehat{\gamma}_j = 1, \widehat{\gamma}_k = 1 \Rightarrow \widehat{\gamma}_{j\vee k} = 1$, and so we write $\mathsf{E}[\widehat{\gamma}_j\widehat{\gamma}_k] = \mathsf{P}[\widehat{\gamma}_j = 1 \mid X_{j\vee k} = 1]\mathsf{P}[\widehat{\gamma}_k = 1 \mid X_{j\vee k} = 1]\mathsf{P}[X_{j\vee k} = 1] = \mathsf{P}[\widehat{\gamma}_j = 1]\mathsf{P}[\widehat{\gamma}_k = 1]/\mathsf{P}[X_{j\vee k} = 1] = \gamma_j\gamma_k/A_{j\vee k}$. Thus,

$$\mathsf{Cov}(\widehat{\gamma}_j, \widehat{\gamma}_k) = \begin{cases} \gamma_k(1-\gamma_j) & j \succeq k \\ \gamma_j(1-\gamma_k) & k \succeq j \\ \gamma_j\gamma_k(1/A_{j\vee k} - 1) & \text{otherwise} \end{cases} \tag{71}$$

(70) then follows from (68) and the fact that $j \vee k = j$ when $j \succeq k$.

(iii) As $\|\overline{\alpha}\| \to 0$,

$$D(\alpha) = D + O(\|\overline{\alpha}\|) \quad \text{where} \quad D_{jk} := \begin{cases} 1 & k = j \\ -1 & k = f(j) \\ 0 & \text{otherwise} \end{cases} \tag{72}$$

To establish this, note first that $D(\alpha)$ has inverse $D^{-1}(\alpha)$ whose elements are $(D(\alpha)^{-1})_{ij} = \partial\gamma_i/\partial\alpha_j$. Now $\partial\gamma_i/\partial\alpha_j = \gamma_i/\alpha_j$ when $j \succeq i$. When $j \prec i$, then from the proof of Theorem 2

$$\frac{\partial\gamma_i}{\partial\alpha_j} = A_i\frac{\partial\beta_i}{\partial\alpha_j} = A_i\frac{\beta_i}{\alpha_j}\prod_{m=1}^{\ell(j)-\ell(i)}\prod_{k\in d(f^m(j))\setminus f^{m-1}(j)}\overline{\beta}_k \tag{73}$$

From (68) (b), this goes to 0 as $\|\overline{\alpha}\| \to 0$. Finally, for all other j, γ_i does not depend on α_i, and so the derivative is 0. Summarizing, as $\|\overline{\alpha}\| \to 0$,

$$D(\alpha)^{-1} = \widetilde{D} + O(\|\overline{\alpha}\|) \quad \text{where} \quad \widetilde{D}_{ij} := \begin{cases} 1 & j \succeq i \\ 0 & \text{otherwise} \end{cases} \tag{74}$$

Since matrix inversion is continuous in an open neighborhood of the non-singular matrices, then (72) follows if we can show that $\widetilde{D}_{ij}$ and D_{ij} are inverses. First $\sum_i D_{ki}\widetilde{D}_{ij} = \widetilde{D}_{kj} - \widetilde{D}_{f(k)j} = \delta_{kj}$ as required. Second $\sum_i \widetilde{D}_{ij}D_{jk} = \widetilde{D}_{ik} - \sum_{j\in d(k)}\widetilde{D}_{ik}$. The second term is only potentially non-zero when $k \succ i$. In this case the only term that contributes to the sum is when $j \succeq i$, giving -1. Hence $\sum_i \widetilde{D}_{ij}D_{jk} = \delta_{ik}$ as required.

(iv) By (iii), and continuity of finite dimensional matrix products, we have as $\|\overline{\alpha}\| \to 0$ that

$$\nu_{ik} = \sum_{j,j'} D_{ij}s(j \vee j')D_{kj'} + O(\|\overline{\alpha}\|^2). \tag{75}$$

It remains to evaluate

$$\sum_{j,j'} D_{ij} s(j \vee j') D_{kj'} = s(i \vee k) - s(i \vee f(k)) - s(f(i) \vee k) + s(f(i) \vee f(k)). \tag{76}$$

When $i = k$, then $i \vee k = i$, $i \vee f(k) = f(i) \vee f(k) = f(i)$ and so (76) yields $s(i) - s(f(i)) = \overline{\alpha}_i$, All other possible i and k yield zero, as we now show. If $i \prec k$ then $i \vee k = f(i) \vee k = k$, while $i \vee f(k) = f(i) \vee f(k) = f(k)$, and hence (76) is zero. The case $k \prec i$ is similar. In all other cases $i, k \prec i \vee k$ and so $i \vee k = i \vee f(k) = f(i) \vee k = f(i) \vee f(k)$. ∎

Proof of Theorem 8: Since $\alpha_k \in (0,1)$, each $Y_k(\cdot)$ is irreducible, and hence so is $Y(\cdot)$, and so q is the unique stationary distribution for Q, i.e. $\sum_z Q_{yz} q_z = q_y$. For n probes, $\widehat{\gamma}_j = \sum_{y \in \Phi(j)} \widehat{q}_y$ where $\widehat{q}_y = n^{-1} \sum_m \delta_{y\, Y(m)}$. By the Central Limit Theorem for Markov processes, see e.g. Chapter 17 of [15], $\widehat{q}$ is asymptotically Gaussian as $n \to \infty$ with

$$\sqrt{n}\,(\widehat{q} - q) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \xi) \tag{77}$$

where

$$\xi_{yz} = \lim_{n\to\infty} n\, \mathsf{Cov}(\widehat{q}_y, \widehat{q}_z) = \lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} \sum_{m'=1}^{n} \mathsf{Cov}(\delta_{y\, Y(m)}, \delta_{z\, Y(m')}) \tag{78}$$

$$= q_y(\delta_{yz} - q_z) + 2 \sum_{m=1}^{\infty} (Q^m_{yz} - q_y) q_z . \tag{79}$$

∎

References

[1] J-C. Bolot and A. Vega Garcia "The case for FEC-based error control for packet audio in the Internet" ACM Multimedia Systems, to appear.

[2] R. L. Carter and M. E. Crovella, "Measuring Bottleneck Link Speed in Packet-Switched Networks," *PERFORMANCE '96*, October 1996.

[3] A. Dembo and O. Zeitouni, "Large deviations techniques and applications", Jones and Bartlett, Boston, 1993.

[4] B. Effron and D.V. Hinkley, "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information", Biometrika, 65, 457–487, 1978.

[5] Felix: Independent Monitoring for Network Survivability. For more information see **ftp://ftp.bellcore.com/pub/mwg/felix/index.html**

[6] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," IEEE/ACM Transactions on Networking, 1(4), August 1993.

[7] IPMA: Internet Performance Measurement and Analysis. For more information see **http://www.merit.edu/ipma**

[8] IP Performance Metrics Working Group. For more information see **http://www.ietf.org/html.charters/ippm-charter.html**

[9] V. Jacobson, "Congestion Avoidance and Control", Proceedings of ACM SIGCOMM '88, August 1988, pp. 314–329.

[10] V. Jacobson, Pathchar - A Tool to Infer Characteristics of Internet paths. For more information see **ftp://ftp.ee.lbl.gov/pathchar**

[11] E.L. Lehmann. "Theory of Point Estimation". Wiley-Interscience, 1983.

[12] B.N. Levine, S. Paul, J.J. Garcia-Luna-Aceves, “Organizing multicast receivers deterministically according to packet-loss correlation”, Preprint, University of California, Santa Cruz.

[13] J. Mahdavi, V. Paxson, A. Adams, M. Mathis, “Creating a Scalable Architecture for Internet Measurement,” *to appear in Proc. INET ’98*.

[14] M. Mathis and J. Mahdavi, “Diagnosing Internet Congestion with a Transport Layer Performance Tool,” *Proc. INET ’96*, Montreal, June 1996.

[15] S.P. Meyn and R.L. Tweedie, “Markov Chains and Stochastic Stability”, Springer, New York, 1993.

[16] **mtrace** – Print multicast path from a source to a receiver. For more information see **ftp://ftp.parc.xerox.com/pub/net-research/ipmulti**

[17] **nam** – Network Animator. For more information see **http://www-mash.cs.berkeley.edu/ns/nam.html**

[18] **ns** – Network Simulator. For more information see **http://www-mash.cs.berkeley.edu/ns/ns.html**

[19] V. Paxson, “End-to-End Routing Behavior in the Internet,” *Proc. SIGCOMM ’96*, Stanford, Aug. 1996.

[20] V. Paxson, “Towards a Framework for Defining Internet Performance Metrics,” *Proc. INET ’96*, Montreal, 1996.

[21] V. Paxson, “End-to-End Internet Packet Dynamics,” *Proc. SIGCOMM 1997*, Cannes, France, 139–152, September 1997.

[22] V. Paxson, “Automated Packet Trace Analysis of TCP Implementations,” *Proc. SIGCOMM 1997*, Cannes, France, 167–179, September 1997.

[23] V. Paxson, “Measurements and Analysis of End-to-End Internet Dynamics,” Ph.D. Dissertation, University of California, Berkeley, April 1997.

[24] J. Postel, “Transmission Control Protocol,” RFC 793, September 1981.

[25] K. Ross & C. Wright, “Discrete Mathematics”, Prentice Hall, Englewood Cliffs, NJ, 1985.

[26] W. Rudin, “Functional Analysis”, McGraw-Hill, New York, 1973.

[27] M.J. Schervish, “Theory of Statistics”, Springer, New York, 1995.

[28] Surveyor. For more information see **http://io.advanced.org/surveyor/**

[29] K. Thompson, G.J. Miller and R. Wilder, “Wide-Area Internet Traffic Patterns and Characteristics,” *IEEE Network*, 11(6), November/December 1997.

[30] R.J. Vanderbei and J. Iannone, “An EM approach to OD matrix estimation,” Technical Report, Princeton University, 1994

[31] Y. Vardi, “Network Tomography: estimating source-destination traffic intensities from link data,” J. Am. Statist. Assoc., 91: 365–377, 1996.

[32] R.R. Wolff “Poisson Arrivals See Time Averages”, Operations Research, 30: 223–231, 1982

[33] M. Yajnik, J. Kurose, D. Towsley, “Packet Loss Correlation in the MBone Multicast Network,” *Proc. IEEE Global Internet*, Nov. 1996