Multicast-Based Inference of Network-Internal Loss Characteristics

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

February 9, 1998

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, in particular packet loss rates. We develop statistically rigorous techniques for estimating loss rates on internal links based on losses observed by multicast receivers. These techniques exploit 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 validate these techniques through simulation and discuss possible extensions and applications of this work. The bandwidth efficiency of multicast traffic makes these techniques suitable for large-scale measurements of both end-to-end and internal network dynamics.

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,

*AT&T Labs–Research, Rm. A173, 180 Park Avenue, Florham Park, NJ 07932, USA; E-mail: ramon@research.att.com

†AT&T Labs–Research, Rm. A175, 180 Park Avenue, Florham Park, NJ 07932, USA; E-mail: duffield@research.att.com

‡Department of Mathematics & Statistics, Lederle Graduate Research Tower, Box 34515 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, 13, 18, 20, 21]). A number of ongoing measurement infrastructure projects (Felix [5], IPMA [7], NIMI [12] and Surveyor [26]) 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. Furthermore, these efforts are all unicast-based.

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.

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 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) and derive an expression for their rate of convergence.

We have evaluated our approach for two- and four-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 [17] 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) [23]. In this case we also 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 correlations (i.e., correlations between links), which violates the Bernoulli assumption. We believe that large and long-lasting spatial correlations are unlikely in a real network such as the Internet because of its traffic and link diversity. Furthermore, we believe that the introduction of Random Early Detection (RED) [6] policies in Internet routers will help break such correlations. In any case, our analysis shows that the introduction of correlations introduces inference errors in a continuous manner: if the correlations are small, the errors in the estimates are also small. Also, the analysis shows how prior knowledge of the likely magnitude of correlations could be used to correct the Bernoulli-based estimates.

We note that interference from TCP sources introduces temporal correlations (i.e., correlations between packets) that also violate the Bernoulli assumption. These correlations are apparent in our simulated net-

work, where probe losses often occur back-to-back due to burstiness in the competing TCP streams. Such correlations have also been measured in the Internet, but they rarely involve more than a few consecutive packets [1]. Our MLEs are still asymptotically accurate for large numbers of probes when losses have temporal correlations. Our MLEs only require ergodicity, which will hold, e.g., when the correlations between losses have sufficiently short range. However, the rate of convergence of the estimates to their true values will be slower. In our experiments, inferred loss rates closely tracked actual losses rates despite the presence of temporal correlations.

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 which grows at worst N per link. The exchange of unicast probes can lead to local loads which grow as N^2, depending on the topology.

The work presented in this paper assumes that the multicast tree is known in advance. We are presently developing algorithms to infer the multicast tree from the probe measurements themselves. We propose that such an approach could be used in combination with tools such as `mtrace` [15] in order to determine the topology.

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 [26], 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, with an emphasis on network survivability. A fourth, NIMI (National Internet Measurement Infrastructure) [12], 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.

Turning 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 [31] 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. Some analytic methods for inference of traffic intensities have been proposed quite recently [28, 29]. 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 other papers address a substantially different problem, namely, the determination of traffic matrices. Moreover, the inference problems generally do not have a unique or easily identifiable solution, sometimes needing ad hoc methods to identify a candidate

solution. This was a consequence of a combination of the coarseness of the data (average data rates) 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 along with their rate of convergence and tests for data consistency. The specific example of the simple two-leaf tree is worked out explicitly. In Section 4 we present an algorithm for computing packet loss estimates. Section 5 presents the results of simulation experiments that validate our approach. In Section 6 we analyze the effects of spatial and temporal correlations on our estimators. The paper ends with a summary of our contributions and proposals for further work.

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 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 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 1. A source multicasts a sequence of probes to two receivers, R_1 and R_2. The probes traverse the multicast tree illustrated in Figure 1(a). Figure 1(b) illustrates the logical multicast tree, where each path between branch points in the tree depicted in Figure 1(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

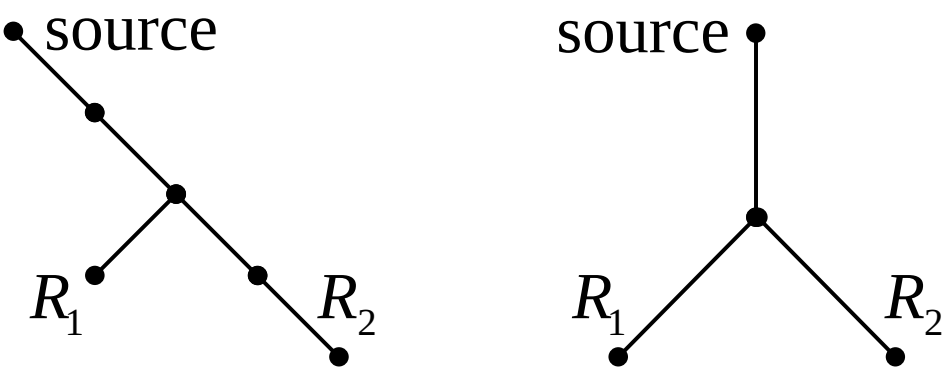

Figure 1: (a) A multicast tree with two receivers. (b) The corresponding logical multicast tree.

these assumptions; in Section 6 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 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 2 and 3 examples of two- and four-leaf logical multicast trees which we shall use for analysis and experiments.

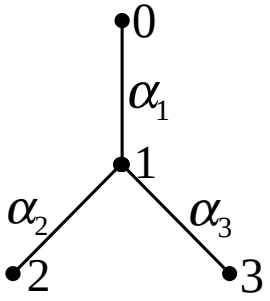

Figure 2: A two-leaf logical multicast tree

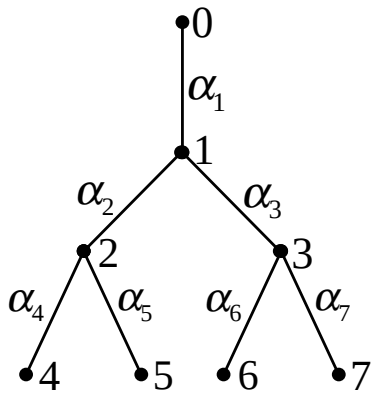

Figure 3: A four-leaf logical multicast tree

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 $\mathbf{P}_\alpha$. The probability mass function for a single outcome $x \in \Omega$ is $p(x;\alpha) = \mathbf{P}_\alpha(X_{(R)} = x)$.

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 (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 the same or 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 [18], 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. We determine the rate at which the estimate converges to its 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 2.

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 $k \leq k'$ for $k, k' \in V$ whenever k is a descendant of k' or $k = k'$ and $k < k'$ whenever $k \leq k'$ 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 $k < f(k) < f^2(k) \ldots < f^\ell(k) = 0$ 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 $\gamma_k = \gamma_k(\alpha) = \mathbf{P}_\alpha[\Omega(k)]$. 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 $\beta_k = \mathbf{P}[\Omega(k) \mid X_{f(k)} = 1]$. 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 $\mathcal{G}$.*

The proof of Theorem 1 relies of the following Lemma whose proof is in the Appendix.

Lemma 1 *Let $c_i \in (0,1)$, $i = 1, 2, \ldots, i_{\max}$, with $\sum_i c_i > 1$. The equation $(1 - x) = \prod_i (1 - c_i x)$ has a unique solution $x \in (0,1)$.*

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

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

By 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 $A_0 = \alpha_0 = 1$):

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

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 the Appendix. That done, we must complete the argument by showing that the stationary point does have maximum likelihood. For this we must impose additional conditions. Although the set $\mathcal{A}$ contains all positive α_k, let us now restrict our attention to link probabilities $\alpha \in \mathcal{B} = (0,1)^{\#R} \subset \mathcal{A}$. (As we explain in more detail in Section 4, there is no loss of generality in excluding the probabilities 0 and 1). Stationarity does not preclude $\widehat{\alpha}$ from being either a minimum or a saddle for the likelihood function, with the maximum falling on the boundary of $\mathcal{B}$. 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 $\breve{\alpha}$. that $\mathcal{L}(\alpha)$ is concave in α at the $\alpha = \widehat{\alpha}$ which implies that $\widehat{\alpha}$ is 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 $\widehat{\alpha}$. The proof is given in the Appendix.

Theorem 3

(i) The model is identifiable *in $\mathcal{B}$, i.e., $\alpha, \alpha' \in \mathcal{B}$ and $\mathbf{P}_\alpha = \mathbf{P}_{\alpha'}$ implies $\alpha = \alpha'$. As a consequence, distinct link probabilities α produce distinct statistical behavior of the $\widehat{\gamma}$ as $n \to \infty$.*

(ii) As $n \to \infty$, $\breve{\alpha} \to \alpha$, $\mathbf{P}_\alpha$ almost surely.

(iii) With probability 1, for sufficiently large n, $\breve{\alpha} = \widehat{\alpha}$.

3.2 Rates of Convergence of the MLEs

In this section we examine in more detail the rate of convergence of the MLE $\widehat{\alpha}$ to the true value α. Specifically, we can apply some general results on the asymptotic properties of MLEs in order to show that $\sqrt{n}(\widehat{\alpha} - \alpha)$ is asymptotically normally distributed as $n \to \infty$. 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 desired accuracy. Secondly, we can estimate the likely accuracy of $\widehat{\alpha}$ from actual probe data and associate confidence intervals to the estimates.

Large Sample Behavior of the MLEs. The *Fisher Information Matrix* at α based on $X_{(R)}$ is a mapping from $\mathcal{B}$ to the $\#U$-dimensional real matrices defined by $\mathcal{I}_{jk}(\alpha) := \mathrm{Cov}\left(\frac{\partial \mathcal{L}}{\partial \alpha_j}(\alpha), \frac{\partial \mathcal{L}}{\partial \alpha_j}(\alpha)\right)$. It is straightforward to verify that $\mathcal{L}$ satisfies conditions (see Section 2.3.1 of [25]) under which $\mathcal{I}$ is equal to the following more convenient expression which we will use in the sequel:

$$\mathcal{I}_{jk}(\alpha) = -\mathbf{E}\frac{\partial^2 \mathcal{L}}{\partial \alpha_j \partial \alpha_k}(\alpha) \tag{13}$$

Theorem 4 *When $\mathcal{I}(\alpha)$ is non-singular, then as $n \to \infty$, under $\mathbf{P}_\alpha$, $\sqrt{n}(\widehat{\alpha} - \alpha)$ converges in distribution to a $\#U$-dimensional Gaussian random variable with mean 0 and covariance matrix $\mathcal{I}^{-1}(\alpha)$.*

This enables us to determine, for example, that asymptotically for large for large n, the estimator $\widehat{\alpha}_k$ will lie between the points

$$\alpha_k \pm z_{\delta/2}\sqrt{\frac{\mathcal{I}^{-1}_{kk}(\alpha)}{n}}, \tag{14}$$

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*:

$$\widehat{\mathcal{I}}_{jk}(\widehat{\alpha}) = -\frac{\partial^2 \mathcal{L}}{\partial \alpha_j \partial \alpha_k}(\widehat{\alpha}), \quad \text{where} \quad \widehat{\alpha} = \Gamma^{-1}(\widehat{\gamma}). \tag{15}$$

Now, the proof of Theorem 2 (see particularly (37)) 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 $\mathbf{P}_{\widehat{\alpha}}[\Omega(k)] = \Gamma(\Gamma^{-1}(\widehat{\gamma}))_k = \widehat{\gamma}_k$, we have $\widehat{\mathcal{I}}(\widehat{\alpha}) = \mathcal{I}(\widehat{\alpha})$.

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

$$\widehat{\alpha}_k \pm z_{\delta/2}\sqrt{\frac{\mathcal{I}^{-1}_{kk}(\widehat{\alpha})}{n}}. \tag{16}$$

An issue for further study is the development of simultaneous confidence intervals for all of the link probabilities, and especially how they grow with the number of links.

3.3 Example: the Two-leaf Tree

In this section we illustrate the application of the results of Sections 3.1 and 3.2 to the two-leaf tree of Figure 2.

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

$$\widehat{\gamma}_1 = \widehat{p}(11) + \widehat{p}(10) + \widehat{p}(01), \quad \widehat{\gamma}_2 = \widehat{p}(11) + \widehat{p}(10), \quad \widehat{\gamma}_3 = \widehat{p}(11) + \widehat{p}(01). \tag{17}$$

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

$$\widehat{\alpha}_1 = \frac{\widehat{\gamma}_2\widehat{\gamma}_3}{\widehat{\gamma}_2 + \widehat{\gamma}_3 - \widehat{\gamma}_1} = \frac{(\widehat{p}(01) + \widehat{p}(11))(\widehat{p}(10) + \widehat{p}(11))}{\widehat{p}(11)} \tag{18}$$

$$\widehat{\alpha}_2 = \frac{\widehat{\gamma}_2 + \widehat{\gamma}_3 - \widehat{\gamma}_1}{\widehat{\gamma}_3} = \frac{\widehat{p}(11)}{\widehat{p}(01) + \widehat{p}(11)} \tag{19}$$

$$\widehat{\alpha}_3 = \frac{\widehat{\gamma}_2 + \widehat{\gamma}_3 - \widehat{\gamma}_1}{\widehat{\gamma}_2} = \frac{\widehat{p}(11)}{\widehat{p}(10) + \widehat{p}(11)} \tag{20}$$

Confidence Intervals. An elementary calculation shows that the inverse of the Fisher information matrix governing the confidence intervals for models in (14) is

$$\mathcal{I}^{-1}(\alpha) = \begin{pmatrix} \frac{\alpha_1(\overline{\alpha}_3 - \alpha_2(1+\alpha_3(\alpha_1-2)))}{\alpha_2\alpha_3} & \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_3} & \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_2} \\ \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_3} & \frac{\overline{\alpha}_2\alpha_2}{\alpha_1\alpha_3} & \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_1} \\ \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_2} & \frac{-\overline{\alpha}_2\overline{\alpha}_3}{\alpha_1} & \frac{\overline{\alpha}_3\alpha_3}{\alpha_1\alpha_2} \end{pmatrix}. \tag{21}$$

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 (16) is obtained by inserting (18)–(20) into (21)

4 Data Consistency and Parameter Computation

In this section we address computational issues around 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.

4.1 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 $\widehat{\gamma} \in \Gamma(\mathcal{B})$. 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 is 0 then 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) If $\widehat{\alpha}_k > 1$ for any $k \in U$ then we reject consistency of the data with the model class since the link probabilities are required to lie in $[0, 1]$ (subject to (i) and (ii) above).

(iv) If $\widehat{\gamma}_k = \sum_{j \in d(k)} \widehat{\gamma}_j$ for any $k \in U \setminus R$, we reject consistency of the data with the model class. This will occur only if the observed losses satisfy the strong dependence property that each packet reaching a receiver in $R(k)$ reaches no other receiver in $R(k)$. The possibility $\widehat{\gamma}_k > \sum_{j \in d(k)} \widehat{\gamma}_k$. is precluded by the relations (23) and (24) below.

4.2 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 $\widehat{\gamma}_k$. These can be calculated essentially by reconstruction of a sample path of the full process $(X_k)_{k \in V}$ from the measured data $(X_k)_{k \in R}$. From the data $X^1_{(R)}, \ldots, X^n_{(R)}$ from n probes, we define the n-element binary vector $(\widehat{X}_k)_{k \in V}$ recursively by

$$\widehat{X}_k = X_k, \quad k \in R \tag{22}$$

$$\widehat{X}_k(i) = \bigvee_{j \in d(k)} \widehat{X}_j(i), \quad k \in V \setminus R \tag{23}$$

so that

$$\widehat{\gamma}_k = n^{-1} \sum_{i=1}^{n} \widehat{X}_k(i). \tag{24}$$

For simplicity we assume now that $\widehat{\gamma} \in \Gamma(\mathcal{B})$, so that, if necessary, steps (i) and (ii) of Section 4.1 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 pseudo-code below. (They could be combined). The procedure $\mathrm{find\_x}$ calculates the $\widehat{X}_k$ and $\widehat{\gamma}_k$, assuming $\widehat{X}_k$ initializes to X_k for $k \in R$ and 0 otherwise. The procedure infer calculates the $\widehat{\alpha}_k$.

```
procedure main {
      find_x ( 0 ) ;
      infer ( 0, 1 ) ;
}

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

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

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. For node k with 2 or 3 children, A_k can be found explicitly as the solutions of a linear and quadratic equation respectively.

4.3 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 $(\widehat{X}_j)_{j \in d(k)}$. 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 $\widehat{X}_j$. 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 which is unlucky enough to be the representative for all distinct subtrees to which it belongs).

5 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

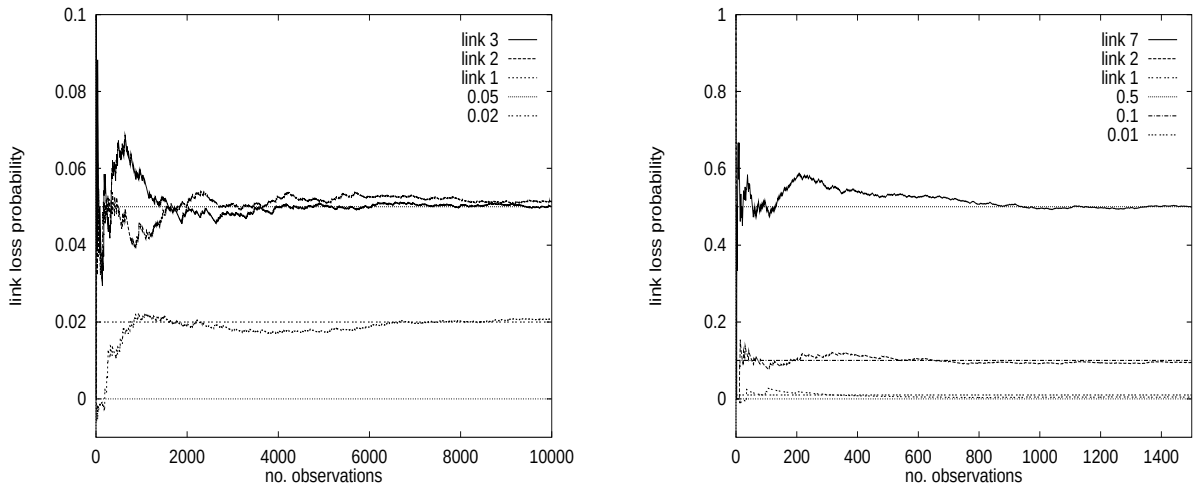

Figure 4: CONVERGENCE OF INFERRED LOSS PROBABILITIES TO ACTUAL LOSS PROBABILITIES IN MODEL SIMULATIONS. Left: Two-leaf tree of Figure 2 with parameters $\overline{\alpha}_1 = 0.02$; $\overline{\alpha}_2 = \overline{\alpha}_3 = 0.05$. Right: Selected links from four-leaf tree of Figure 3, with parameters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

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 [27]. The following two subsections describe our results from these two simulation efforts.

5.1 Model Simulations

Topology. For the model simulations, we used ad hoc software written in C++. We simulated the two tree topologies shown in Figures 2 and 3. 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.

Convergence. Figure 4 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 5 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

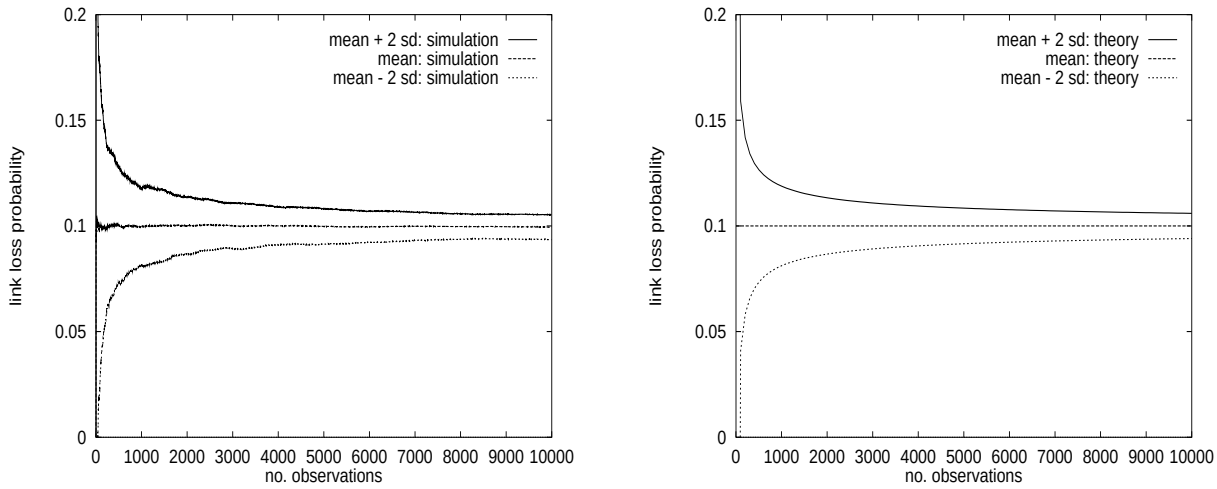

Figure 5: AGREEMENT BETWEEN SIMULATED AND THEORETICAL CONFIDENCE INTERVALS. Left: Results from 100 model simulations. Right: Predictions from (14). The graphs show two-sided confidence estimates at 2 standard deviations for link 2 of the four-leaf tree of Figure 3. Parameters were $\overline{\alpha}_1 = 0.01$; $\overline{\alpha}_2 = 0.1$; $\overline{\alpha}_3 = \overline{\alpha}_4 = \overline{\alpha}_5 = \overline{\alpha}_6 = 0.01$; $\overline{\alpha}_7 = 0.5$. Simulation matches theory extremely well – the two sets of curves are indistinguishable when plotted in the same graph.

intervals are as predicted by (14). As shown, simulation matches theory extremely well – we show the two graphs separately because the two sets of curves are indistinguishable when plotted together. In addition, the 95% confidence intervals converge to within 0.02 of the mean probabilities after 2,000 observations.

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 shorter 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.

5.2 TCP Simulations

Topology. For the TCP simulations, we used the ns network simulator [17]. We configured ns to simulate the two tree topologies shown in Figures 2 and 3. 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.

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

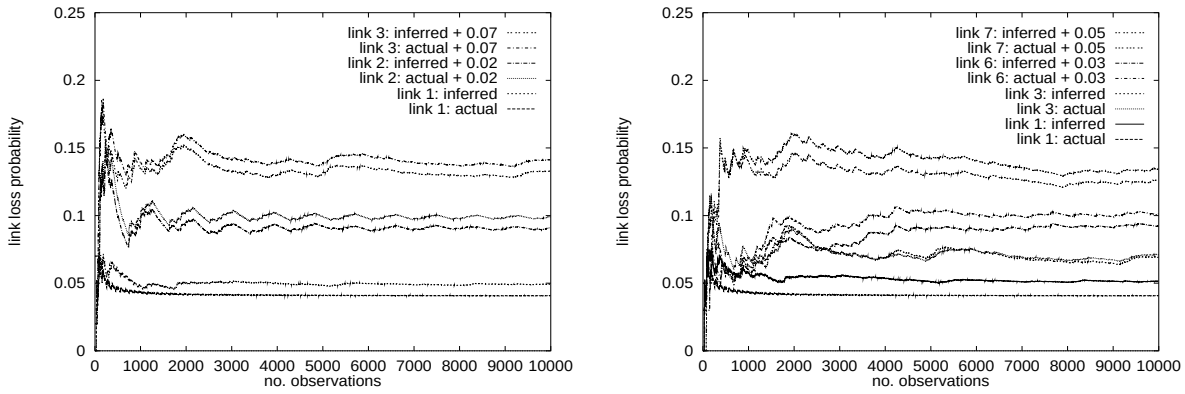

Figure 6: CONVERGENCE OF INFERRED LOSS RATES TO ACTUAL LOSS RATES IN TCP SIMULATIONS. Left: Two-leaf tree of Figure 2. Right: Selected links from four-leaf tree of Figure 3 (some pairs of probabilities are offset for clarity). The graphs show that the inferred loss rates closely track the actual loss rates over 10,000 observations.

also carried one TCP stream.

Node 0 sent multicast probe packets generated by a Constant Bit Rate (CBR) 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.

Convergence. Figure 6 compares inferred loss rates to actual loss rates on selected links of our two- and four-leaf topologies. As shown, the inferred rates closely track the actual rates over 10,000 observations.

We note that the inferred values were 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 [16]. Inspection of the autocorrelation function of the time series of packet losses for a series of experiments predominantly showed correlations indistinguishable from zero beyond a lag of 1 (i.e. greater than back-to-back losses). As we explain in more detail in Section 6.2, the estimator $\widehat{\alpha}$ is still asymptotically accurate for large numbers of probes when losses have temporal correlations of sufficiently short range. However, the rate of convergence of the estimates to their true values will be slower.

Figure 7 shows the Root Mean Square (RMS) differences between the inferred and actual loss rates on our standard 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

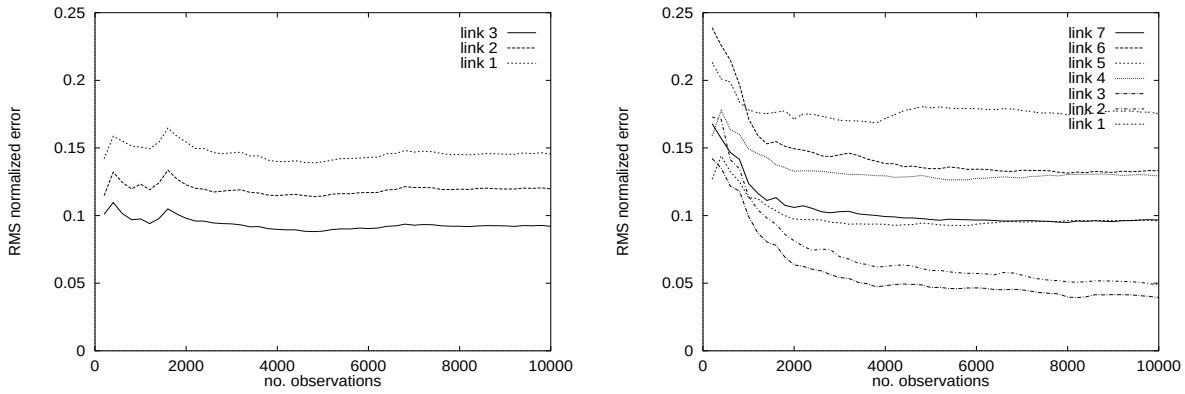

Figure 7: ACCURACY OF INFERENCE IN TCP SIMULATIONS. Left: Two-leaf tree of Figure 2. Right: Four-leaf tree of Figure 3. The graphs show normalized root mean square differences between actual and inferred loss rates, computed across 100 simulations. After an initial transient, inferred loss rates settle down to within 4 to 18% of actual loss rates, depending on the link. These errors can be reduced to approximately 1% by modifying the MLEs to correct for spatial loss correlations.

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 correlation. In our simulations, the same multicast probe is lost on sibling links more often than the spatial independence assumption dictates. These correlated 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 correlation present in our simulations. We can also calculate the effect of such correlation on the inferred loss probabilities by extending our previous analysis. Thus a prior estimate of the degree of correlation could be used to obtain corrections to the Bernoulli inference. We discuss this in more detail in Section 6.1 and give an example of how to apply the correction. Applied to the inferences on the two-leaf tree summarized in Figure 6, they reduce an RMS error of between 8 and 15% to one of around 1%. The key observation behind these analyses is that the error in the inferred values varies smoothly with the degree of spatial correlation. The more correlation in the network, the larger the error. We can arrange for perfectly correlated losses in a simulated network, for example by creating perfectly synchronized interference streams on sibling links. However, we believe that large and long-lasting spatial loss correlation is unlikely in real networks like the Internet because of their traffic and link diversity.

6 The Analysis of Correlations

6.1 Analysis and Correction for Spatial Correlations

When spatial correlations are present in packet losses, the Bernoulli model assumption is violated. But even with such correlations, we can still ask what are the *marginal* loss probabilities for each link independently.

In this section we assess and quantify the effect of correlations between losses on links that share a common parent. The comparison of inferred and actual loss rates in Figure 6 shows that the inferred loss is greater than the actual loss at the leaf-links, while the effect is reversed at the root link. This effect can be understood qualitatively by observing that in the Bernoulli model, a predominance of common losses amongst the descendants of a node k will push the estimator $\widehat{\alpha}$ towards ascribing loss as having a common origin in the path from the root 0 to k.

However, as we now show, the effect of violating the Bernoulli property is continuous: if the correlations are small, the errors in the estimates are also small. We can investigate this quantitatively by extending the model class. For simplicity we analyze the two-leaf tree of Figure 2, although the method extends to arbitrary trees. We introduce correlations into the model while maintaining the same marginal link probabilities. All models with these properties can be described by the addition of a single parameter. Consider then the family of models in which for each ν the probabilities of the outcomes $ij \in (11, 10, 01, 00)$ are $p(ij; \alpha, \nu)$, where

$$p(11;\alpha) = \alpha_1\alpha_2\alpha_3 \quad \longrightarrow \quad p(11;\alpha,\nu) = \alpha_1(\alpha_2\alpha_3 + \nu\alpha_2\alpha_3) \tag{25}$$

$$p(10;\alpha) = \alpha_1\alpha_2\overline{\alpha}_3 \quad \longrightarrow \quad p(10;\alpha,\nu) = \alpha_1(\alpha_2\overline{\alpha}_3 - \nu\alpha_2\alpha_3) \tag{26}$$

$$p(01;\alpha) = \alpha_1\overline{\alpha}_2\alpha_3 \quad \longrightarrow \quad p(01;\alpha,\nu) = \alpha_1(\overline{\alpha}_2\alpha_3 - \nu\alpha_2\alpha_3) \tag{27}$$

$$p(00;\alpha) = \overline{\alpha}_1 + \alpha_1\overline{\alpha}_2\overline{\alpha}_3 \quad \longrightarrow \quad p(00;\alpha,\nu) = \overline{\alpha}_1 + \alpha_1(\overline{\alpha}_2\overline{\alpha}_3 + \nu\alpha_2\alpha_3) \tag{28}$$

When $\nu > 0$, the correlations between the losses on links 2 and 3 positive; e.g. then $\mathbf{P}[X_2 = 1 \mid X_3 = 1, X_1 = 1] = \mathbf{P}[X_2 = X_3 = 1 \mid X_1 = 1]/\mathbf{P}[X_3 = 1 \mid X_1 = 1] = (1+\nu)\alpha_2 > \alpha_2 = \mathbf{P}[X_2 = 1 \mid X_1 = 1]$. Now consider sending probes through a network with losses described by these probabilities. As the number of probes $n \to \infty$, the proportion $\widehat{p}(x)$ of measurements in outcome x converges almost surely to $p(x; \alpha, \nu)$. To see how the estimator $\widehat{\alpha}$ interprets this data, we insert these limit values $p(x; \alpha, \nu)$ in place of $\widehat{p}(x; \alpha)$ in (18)–(20) for the Bernoulli model. The resulting estimates of the link probabilities would then be $\alpha(\nu)$, as given by

$$\alpha_1(\nu) = \frac{\alpha_1}{1+\nu}, \quad \alpha_2(\nu) = (1+\nu)\alpha_2, \quad \alpha_3(\nu) = (1+\nu)\alpha_3. \tag{29}$$

This confirms the qualitative statements above: as ν increases from 0, the estimate $1 - \alpha_1(\nu)$ of the probability of loss on link 1 increases. As $\nu \to 0$ we recover the result of the Bernoulli model.

If some knowledge of the correlations 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 these correlations. We intend to undertake these experiments in the future.

For the moment, we illustrate with our data from the ns simulation of the two-leaf tree of Figure 2. During 50 simulations, each with 10,000 probes, the correlation κ between losses on links 2 and 3 was found to be about 0.1. Let Y_2 defined on those trials for which $X_1 = 1$, taking the value X_2, with Y_3 defined similarly.

Then using the measured losses, the average estimated value $\widehat{\kappa}$ of $\kappa := \mathrm{Cov}(Y_2, Y_3)/\sqrt{\mathrm{Var}(Y_2)\mathrm{Var}(Y_3)}$ over the first 50 simulations was found to be 0.1.

We now show how to use $\widehat{\kappa}$ to adjust subsequent inferences to take account of spatial correlations. An explicit calculation of κ using the model (25)–(28) shows that

$$\nu = \kappa \frac{\overline{\alpha}_2 \overline{\alpha}_3}{\alpha_2 \alpha_3}. \tag{30}$$

Now consider a new set of probes. We suppose that the spatial correlations are characterized by $\widehat{\kappa}$. Applying the Bernoulli estimator will yield estimates which we denote by $\widehat{\alpha}(\widehat{\nu})$, where $\widehat{\nu} = \widehat{\kappa}(1-\widehat{\alpha}_2)(1-\widehat{\alpha}_3)/\widehat{\alpha}_2\widehat{\alpha}_3$, following (30). Inserting $\widehat{\alpha}(\widehat{\nu})$ and $\widehat{\nu}$ into (29) in place of $\alpha(\nu)$ and ν yields an estimate of the true marginal link probabilities, the α of (29). We conducted 50 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.

6.2 Consequences of Temporal Correlations

The key property possessed by the estimator $\widehat{\alpha}$ is that it is asymptotically accurate for a large number of observations, even in the presence of sufficiently short-range temporal correlations, provided that the underlying processes are stationary and ergodic. (This happens, for example, when recurrence conditions are satisfied; see e.g. [14]). Then, the observed probabilities $\widehat{\gamma}$ converge to the true values γ with probability 1 as the number of observations $n \to \infty$. A simple argument involving the Inverse Function Theorem (e.g., see [24]) shows that Γ^{-1} is continuous on $\Gamma(\mathcal{B})$, and hence $\widehat{\alpha} \to \alpha$ with probability 1. As a specific example, when the observations $X^1, X^2, \ldots$ form a Markov chain (and also under more general mixing conditions) then the observed frequencies $\widehat{p}(x)$ will converge exponentially fast to their mean values $p(x;\alpha)$ (e.g., see Chapter 6 of [3]).

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. The price of correlations is, however, that the rate of convergence is slower than for the Bernoulli case. One can understand this qualitatively from the fact that burstiness in the packet loss processes means that the average takes longer to approach. Quantitatively, the rate of convergence can be determined from the model, and so it should be possible to determine the impact of temporal correlations on convergence rates from knowledge of the ambient correlations of loss in the network, in much the same manner as the impact of spatial correlations was determined in Section 6.1. Another approach to avoiding temporal correlations 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.

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. [30]). This approach has been proposed for network measurements [22] 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.

7 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 correlations.

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 correlations. 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.

Third, we plan to experiment with multicast-based inference on the Internet. As a preliminary step, we plan to measure ambient correlations 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.

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] J. Mahdavi, V. Paxson, A. Adams, M. Mathis, "Creating a Scalable Architecture for Internet Measurement," *to appear in Proc. INET '98*.

[13] M. Mathis and J. Mahdavi, "Diagnosing Internet Congestion with a Transport Layer Performance Tool," *Proc. INET '96*, Montreal, June 1996.

[14] S.P. Meyn and R.L. Tweedie, "Markov chains and stochastic stability", Springer, New York, 1993.

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

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

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

[18] V. Paxson, "End-to-End Routing Behavior in the Internet," *Proc. SIGCOMM '96*, Stanford, Aug. 1996.

[19] V. Paxson, "Towards a Framework for Defining Internet Performance Metrics," *Proc. INET '96*, Montreal, 1996.

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

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

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

[23] J. Postel, "Transmission Control Protocol," RFC 793, September 1981.

[24] W. Rudin, "Functional Analysis", McGraw-Hill, New York, 1973.

[25] M.J. Schervish, "Theory of Statistics", Springer, New York, 1995.

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

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

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

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

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

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

Appendix: Proofs of Theorems

Proof of Lemma 1: Let $h_1(x) = (1-x)$, $h_2(x) = \prod_i (1 - c_i x)$. Let $q_i = c_i/(1-c_i x)$. Then for $x \in [0,1]$ $h_1''(x) = 0$, $h_2''(x) = h_2(x)\left\{\left(\sum_i q_i\right)^2 - \sum_i q_i^2\right\} > 0$. 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 $x \in (0,1)$. ∎

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 $x \in \Omega(f^i(k)) \setminus \Omega(f^{i-1}(k))$,

$$\frac{\partial \log p(x)}{\partial \alpha_k} = \frac{1}{\overline{\beta}_{f^{i-1}(k)}} \frac{\partial \overline{\beta}_{f^{i-1}(k)}}{\partial \alpha_k} \tag{31}$$

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

$$\frac{\partial \log p(x)}{\partial \alpha_k} = \frac{1}{\overline{\beta}_0} \frac{\partial \overline{\beta}_0}{\partial \alpha_k} \tag{32}$$

On combining these:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \alpha_k} &= \frac{1}{\alpha_k} \sum_{x \in \Omega(k)} n(x) + \frac{1}{\overline{\beta}_0} \frac{\partial \overline{\beta}_0}{\partial \alpha_k} \sum_{x \in \Omega(0)^c} n(x) \\ &\quad + \sum_{i=1}^{\ell(k)} \left\{ \frac{1}{\overline{\beta}_{f^{i-1}(k)}} \frac{\partial \overline{\beta}_{f^{i-1}(k)}}{\partial \alpha_k} \sum_{x \in \Omega(f^i(k)) \setminus \Omega(f^{i-1}(k))} n(x) \right\} \end{aligned} \tag{33}$$

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

$$\frac{\partial \overline{\beta}_k}{\partial \alpha_k} = -\beta_k/\alpha_k, \qquad \text{and} \tag{34}$$

$$\frac{\partial \overline{\beta}_{f^i(k)}}{\partial \alpha_k} = \frac{\alpha_{f^i(k)} - \beta_{f^i(k)}}{\overline{\beta}_{f^{i-1}(k)}} \frac{\partial \overline{\beta}_{f^{i-1}(k)}}{\partial \alpha_k} = -\frac{\beta_k}{\alpha_k} \prod_{m=1}^{i} \frac{\alpha_{f^m(k)} - \beta_{f^m(k)}}{\overline{\beta}_{f^{m-1}(k)}}. \tag{35}$$

The right hand term in equation (35) follows by iterating the middle term. Observe that

$$\sum_{x \in \Omega(f^i(k)) \setminus \Omega(f^{i-1}(k))} \frac{n(x)}{n} = \widehat{\gamma}_{f^i(k)} - \widehat{\gamma}_{f^{i-1}(k)} \quad \text{and} \quad \sum_{x \in \Omega(0)^c} \frac{n(x)}{n} = 1 - \widehat{\gamma}_0. \tag{36}$$

Combining (33), (34), (35) and (36) we get

$$\frac{\alpha_k}{n} \frac{\partial \mathcal{L}}{\partial \alpha_k} = \widehat{\gamma}_k - \beta_k \sum_{i=1}^{1+\ell(k)} \frac{\widehat{\gamma}_{f^i(k)} - \widehat{\gamma}_{f^{i-1}(k)}}{\overline{\beta}_{f^{i-1}(k)}} \prod_{m=1}^{i-1} \frac{\alpha_{f^m(k)} - \beta_{f^m(k)}}{\overline{\beta}_{f^{m-1}(k)}}. \tag{37}$$

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 $k = 0$, (37) yields $0 = \widehat{\gamma}_0 - \beta_0(1 - \widehat{\gamma}_0)/\overline{\beta}_0$, whence

$$\widehat{\gamma}_0 = \beta_0 = \gamma_0. \tag{38}$$

For any other k, combining (37) for k and $j = f(k)$ yields

$$\widehat{\gamma}_k = \frac{\beta_k}{\overline{\beta}_k}\left(\widehat{\gamma}_j - \widehat{\gamma}_k + \frac{(\alpha_j - \beta_j)\widehat{\gamma}_j}{\beta_j}\right), \quad \text{whence} \quad \frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} = \frac{\beta_k \alpha_j}{\beta_j} = \frac{\gamma_k}{\gamma_j}. \tag{39}$$

Together with (38) this gives $\widehat{\gamma}_k = \gamma_k$ for all $k \in V$. ∎

Proof of Theorem 3: (i) By the strong law of large numbers, $\widehat{\gamma} \to \Gamma(\alpha)$, $\mathbf{P}_\alpha$ 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) Fix some $\alpha^0 \in \mathcal{B}$, $M \subset \mathcal{B}$, $x \in \Omega$ and define

$$Z(M, x) = \inf_{\alpha' \in M} \log \frac{p(x; \alpha^0)}{p(x; \alpha')} = \log p(x; \alpha^0) - \sup_{\alpha' \in M} \log p(x; \alpha'). \tag{40}$$

Observe that $p(x; \alpha)$ is polynomial in the α_k, and hence continuous. According to Lemma 7.54 in [25], it suffices to show that, for each $\alpha' \neq \alpha^0$, there is an open set $N_{\alpha'}$ containing α', such that $\mathbf{E}_{\alpha^0} Z(N_{\alpha'}, X) > -\infty$. (Here $\mathbf{E}_{\alpha^0}$ is the expectation w.r.t. $\mathbf{P}_{\alpha^0}$).

Look at the two terms in $\mathbf{E}_{\alpha^0} Z(M, X)$ for any $M \subset \mathcal{B}$. The first is $\mathbf{E}_{\alpha^0} \log p(X; \alpha^0) = \sum_{x \in \Omega} p(x; \alpha^0) \log p(x; \alpha^0)$. This is finite since $p \log p$ is bounded for $p \in [0, 1]$ and Ω is finite. For the second term, note that $p(x; \alpha') \leq 1 \Rightarrow \log p(x; \alpha') \leq 0 \Rightarrow \sup_{\alpha' \in M} \log p(x; \alpha') \leq 0 \Rightarrow -\sup_{\alpha' \in M} \log p(x; \alpha') \geq 0 \Rightarrow \mathbf{E}_{\alpha^0} Z(M, X) \geq \mathbf{E}_{\alpha^0} \log p(X; \alpha^0) > -\infty$. Finally, we note that although it is not mentioned there, Lemma 7.54 in [25] requires identifiability, which we proved in (i) above.

(iii) Now let $\alpha \in \mathcal{B}$ be the true set of link probabilities. From part (ii), with $\mathbf{P}_\alpha$ 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 $\mathcal{B}$. 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: We refer to Theorem 7.63 of [25]. Clearly $\mathcal{L}$ is 3-times continuously differentiable on $\mathcal{B}$, and has bounded expectation in some neighborhood of α. This establishes the relation (7.64) in [25]. $\frac{\partial \log p(x,\alpha)}{\partial \alpha_j \alpha_k}(\alpha)$ is clearly finite on $\mathcal{B}$ and $n(x)$ is bounded above. Hence $\mathcal{I}$ is finite in $\mathcal{B}$, so together with Theorem 3 and the assumption of non-singularity, we are able to conclude the result. ∎