Multicast Topology Inference From Measured End-to-End Loss

N. G. Duffield, *Senior Member, IEEE*, Joseph Horowitz, Francesco Lo Presti, and Don Towsley, *Fellow, IEEE*

***Abstract*—The use of multicast inference on end-to-end measurement has recently been proposed as a means to infer network internal characteristics such as packet link loss rate and delay. In this paper, we propose three types of algorithm that use loss measurements to infer the underlying multicast topology: i) a grouping estimator that exploits the monotonicity of loss rates with increasing path length; ii) a maximum-likelihood (ML) estimator (MLE); and iii) a Bayesian estimator. We establish their consistency, compare their complexity and accuracy, and analyze the modes of failure and their asymptotic probabilities.**

***Index Terms*—Communication networks, end-to-end measurement, maximum-likelihood (ML) estimation, multicast, statistical inference, topology discovery.**

I. Introduction

A. Motivation

In this paper, we propose and evaluate a number of algorithms for the inference of logical multicast topologies from end-to-end network measurements. All are developed from recent work that shows how to infer per-link loss rate from measource gives rise to a copy of the packet at each receiver. Thus, a packet reaching each member of a subset of receivers encounters *identical* conditions between the source and the receivers' closest common branch point in the tree.

This approach has been used to infer the per-link packet loss probabilities for logical multicast trees with a known topology. The maximum-likelihood estimator (MLE) for the link probabilities was determined in [3] under the assumption that probe loss occurs independently across links and between probes. This estimate is somewhat robust with respect to violations of this assumption. This approach will be discussed in more detail presently.

The focus of the current paper is the extension of these methods to infer the *logical topology* when it is not known in advance. This is motivated in part by ongoing work [1] to incorporate the loss-based MLE into the National Internet Measurement Infrastructure [14]. In this case, inference is performed on end-to-end measurements arising from the exchange of multicast probes between a number of measurement hosts stationed in the Internet. The methods here can be used to infer first the logical multicast topology, and then the loss rates on the links in this topology. What we do not provide is an algorithm for identifying the physical topology of a network.

A more important motivation for this work is that knowledge of the multicast topology can be used by multicast applications. It has been shown in [9] that organizing a set of receivers in a bulk transfer application into a tree can substantially improve performance. Such an organization is central component of the widely used RMTP-II protocol [20]. The development of tree construction algorithms for the purpose of supporting reliable multicast has been identified to be of fundamental importance by the Reliable Multicast Transport Group of the Internet Engineeirng Task Force (IETF); see [7]. This motivated the work reported in [16], which was concerned with grouping multicast receivers that share the same set of network bottlenecks from the source for the purposes of congestion control. Closely related to [3], the approach of [16] is based on estimating packet loss rates for the path between the source and the common ancestor of pairs of nodes in the special case of binary trees. Since loss is a nondecreasing function of the path length, this quantity should be maximal for a sibling pair. The whole binary tree is reconstructed by iterating this procedure.

B. Contribution

This paper describes and evaluates three methods for inference of logical multicast topology from end-to-end multicast measurements. Two of these, i) and ii) below, are directly based on the MLE for link loss probabilities of [3], as recounted in Section II. In more detail, the three methods are as follows.

Manuscript received June 8, 2000; revised May 4, 2001. This work was supported in part by DARPA and the AFL under agreement P30602-98-0238.

N. G. Duffield and F. Lo Presti are with AT&T Labs–Research, Florham Park, NJ 07932 USA (e-mail: duffield@research.att.com; lopresti@research.att.com).

J. Horowitz is with the Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003 USA (e-mail: joeh@math.umass.edu).

D. Towsley is with the Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA (e-mail: towsley@cs.umass.edu).

Communicated by V. Anantharam, Associate Editor for Communication Networks.

Publisher Item Identifier S 0018-9448(02)00034-2.

We identify the physical multicast tree as comprising actual network elements (the nodes) and the communication links that

III. Deterministic Reconstruction of Loss Trees by Grouping

The use of estimates of shared loss rates at multicast receivers has been proposed recently in order to group multicast receivers that share the same set of bottlenecks on the path from the source [16]. The approach was formulated for binary trees, with shared loss rates having the direct interpretation of the loss rate on the path from the root to the (nearest) ancestor of two receivers. Since the loss rate cannot decrease as the path is extended, the pair of receivers for which shared loss rate is greatest will be siblings; if not then one of the receivers would have a sibling and the shared loss rate on the path to their ancestor would be greater. This maximizing pair is identified as a pair of siblings and replaced by a composite node that represents their parent. Iterating this procedure should then reconstruct the binary tree.

In this section and the following section, we establish theoretically the correctness of this approach, and extend it to cover

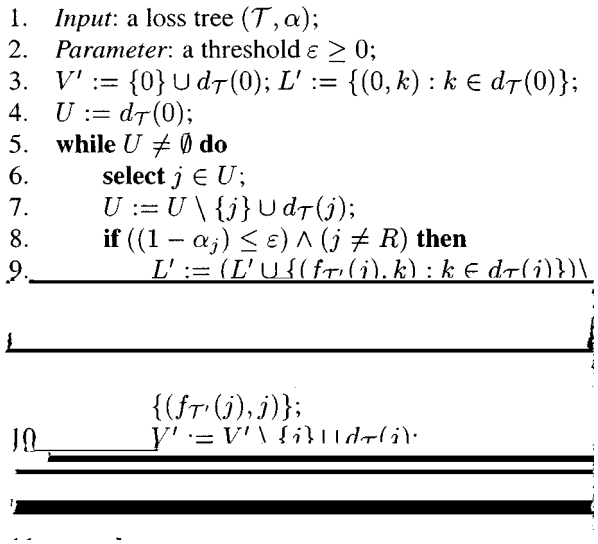


Fig. 3. Tree Pruning Algorithm $\mathsf{TP}(\varepsilon)$.

that $\min_{U \subset R'} B(U)$ is achieved when U is a sibling set. Consequently, one could replace steps 5–8 of DLT by simply finding the maximal sibling set, i.e., select a maximal $U \subset R'$ that minimizes $B(U)$. However, this approach would have worse computational properties since it requires inspecting every subset of R'.

$B(U)$ is a root of the polynomial of degree $\#U - 1$ from Proposition 1 i). For a binary subset, $B(\{j, k\})$ is written down explicitly

$$B(\{j,k\}) = \frac{\gamma(j)\gamma(k)}{\gamma(k) + \gamma(j) - \gamma(\{j,k\})}. \tag{5}$$

Calculation of B

establishes the consistency of the estimator $\hat{\mathcal{T}}_{\mathsf{BLT}}$; the proof appears in Section X.

Theorem 5: Let $(\mathcal{T}, \alpha)$

Theorem 7: Let $(\mathcal{T}, \alpha)$

expressing "maximum ignorance" about the tree topology and link probabilities. Clearly, if other prior information is available about the tree, it may be incorporated into a nonuniform prior distribution. The Bayes classifier becomes

$$\hat{\mathcal{T}}_B(x) = \arg\max_{\tau' \in \mathcal{T}(R)} \int_{A^0_{\tau'}} f(x|\tau', \alpha)\, d\alpha. \tag{13}$$

This should be compared with the ML classifier in (7).

A. Consistency of Pseudo-Bayes Classifiers

In practice, our task is to identify the specific topology giving rise to a set of measured data. When no prior distribution is specified, the concept of the Bayes classifier, as the maximizer of the probability of correct classification, does not make sense, because "the" probability of correct classification is not defined. Nonetheless, it may be convenient to construct a *pseudo-Bayes* classifier by choosing a distribution π on Θ, which plays the role of a prior, and forming the classifier in (10), which we now denote by $\hat{\mathcal{T}}_\pi$. Classifiers constructed in this way are also consistent under a mild condition.

Theorem 9: Let π be a prior distribution on Θ, and assume that $(\mathcal{T}, \alpha)$ lies in the support of π. Then $\hat{\mathcal{T}}_\pi$ is consistent in the frequentist sense, i.e., $\mathsf{P}_{\mathcal{T},\alpha}\left[\hat{\mathcal{T}}_\pi \neq \mathcal{T}\right]$

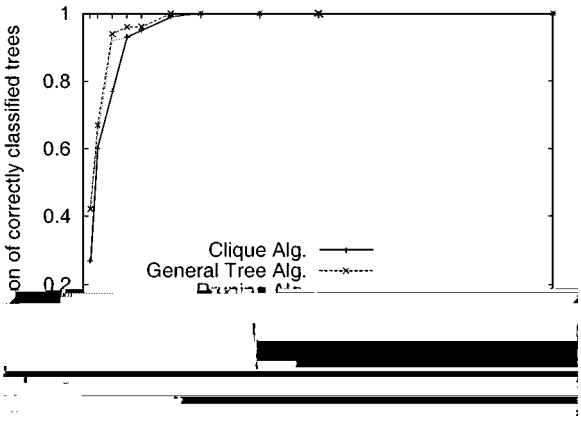


Fig. 8. $\varepsilon = 1.0\%$.

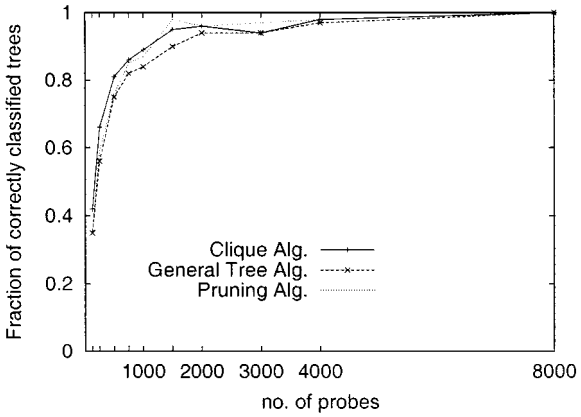


Fig. 9. $\varepsilon = 2.0\%$.

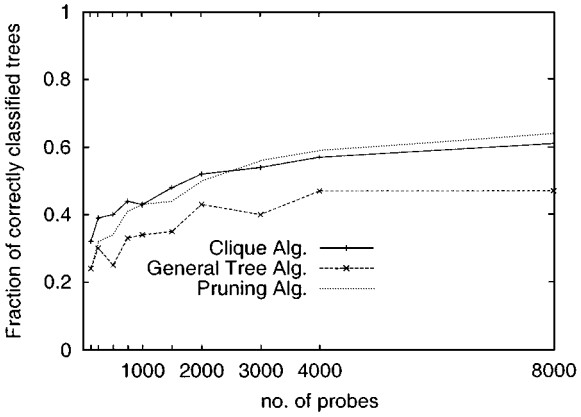


Fig. 10. $\varepsilon = 3.0\%$.

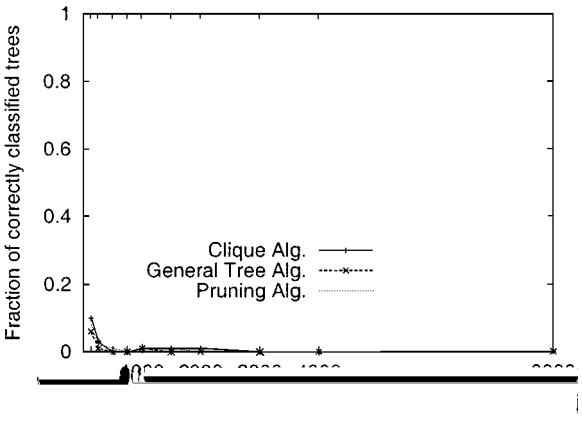


Fig. 11. $\varepsilon = 5.0\%$.

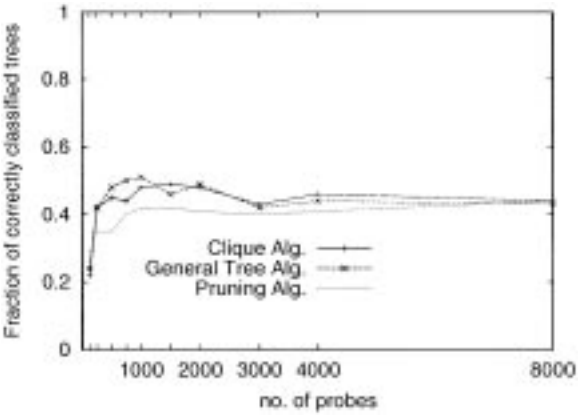


Fig. 12. $\varepsilon = 5\%$.

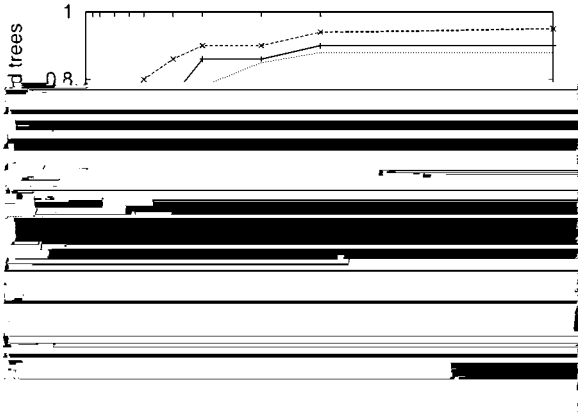
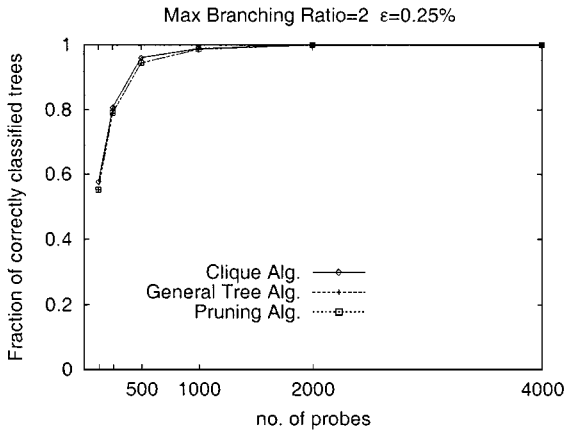
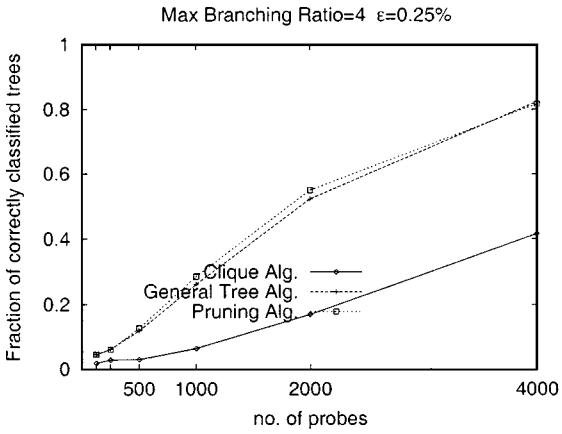


Fig. 13. $\varepsilon = 7\%$.

best for intermediate ε, decreasing for larger and smaller ε. The explanation for this behavior is that smaller values of ε lead to stricter criteria for grouping nodes. With finitely many samples, for small ε, sufficiently large fluctuations of the $\hat{B}$ cause erroneous exclusion of nodes. By increasing ε, the threshold for group formation is increased and so accuracy is initially increased. However, as ε approaches the smallest interior link loss rate, large fluctuations of the $\hat{B}$ now cause erroneous inclusion of nodes into groups. When ε is increased much beyond the smallest interior loss rate, the probability of correct classification falls to zero. The behavior is different if we ignore failures to detect links with loss rates smaller than ε. For $\varepsilon = 5\%$ and $\varepsilon = 7\%$, in Figs. 12 and 13, respectively, we plot the fraction of experiments in which the pruned topology T^{ε} was correctly identified for the three algorithms. Here, the accuracy depends on the relative values of ε and the internal link loss rates. In these experiments, the actual loss rates was often very close to 5%, so that small fluctuations results in erroneous inclusions/exclusions of nodes which accounts for the significant fraction of failures for $\varepsilon = 5\%$. In Section VIII-B, we shall analyze this


Max Branching Ratio=2 ε=0.25%
Fraction of correctly classified trees
no. of probes
Clique Alg.
General Tree Alg.
Pruning Alg.
0
0.2
0.4
0.6
0.8
1
500
1000
2000
4000


Max Branching Ratio=4 ε=0.25%
Fraction of correctly classified trees
no. of probes
Clique Alg.
General Tree Alg.
Pruning Alg.
0
0.2
0.4
0.6
0.8
1
500
1000
2000
4000

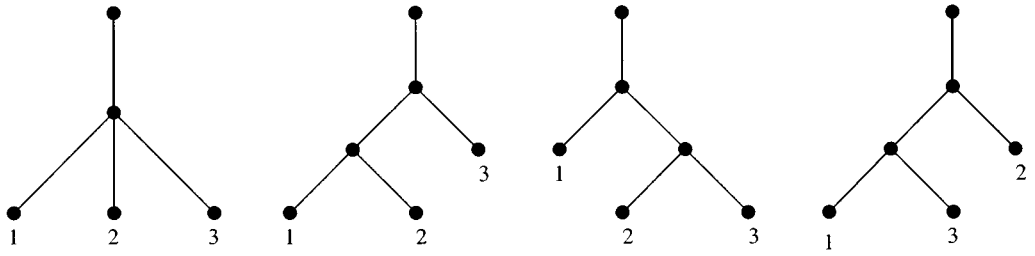


Fig. 15. ML and Bayesian classifier: The four possible topologies with three receivers.

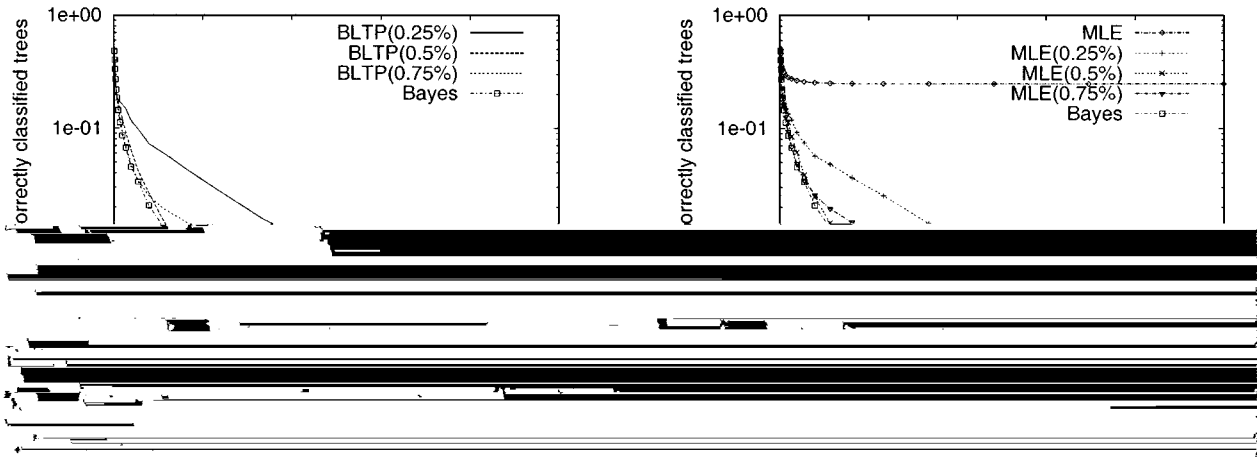


Fig. 16. Misclassification in ML, Bayesian, and BLT classifier: (τ, α) randomly drawn according to the prior distribution. (a) Bayes and BLTP(ε) classifier. (b) Bayes and ML classifiers.

the case in Fig. 16, where we plot the fraction of experiments in which the topology was incorrectly identified as function of the number of probes, for the different classifiers (for clarity, we plot separately the curves for the ML and BLTP(

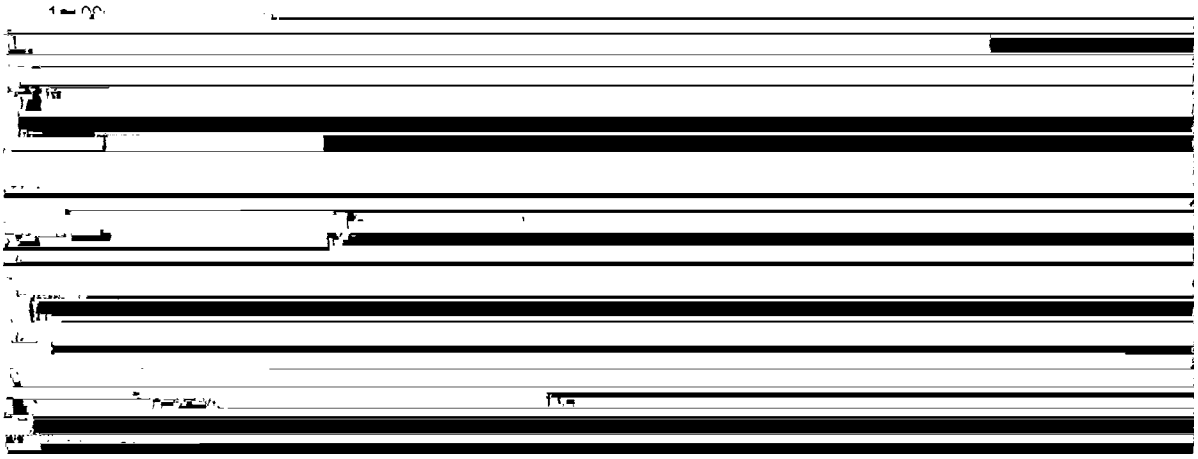
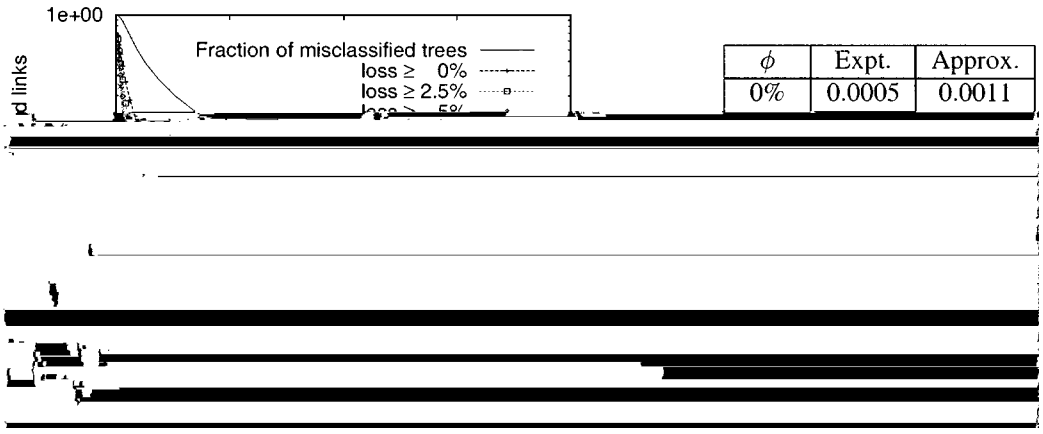


Fig. 17. Misclassification in ML, Bayesian, and BLT classifier. Fixed (

A. Misgrouping and Misclassification in BLT

We start by studying misgrouping in binary trees under BLT. Consider the event G_i that BLT correctly groups nodes in $R_{\mathcal{T}}(i)$ for some $i \in$



ϕ	Expt.	Approx.
0%	0.0005	0.0011

Fig. 18. Misclassification and misgrouping in BLT. Left: Fraction of links misclassified with loss $\geq \phi$, for $\phi = 0\%, 2.5\%, 5.0\%, 7.5\%$. Right: Comparison of experimental and approximated tail slopes.

linear slope of the fraction of misgrouped links, i.e., the one for $\phi = 0\%$.

B. Misgrouping and Misclassification in BLTP(ε)

We turn our attention to the errors in classifying general trees by the reference algorithm BLTP(ε). In the following, without loss of generality, we will study the errors in the classification of the pruned tree $(\mathcal{T}^\varepsilon, \alpha^\varepsilon) = \mathsf{TP}(\varepsilon)(\mathcal{T}, \alpha)$, with $\mathcal{T}^\varepsilon = (V^\varepsilon, L^\varepsilon)$, under the assumption that $\varepsilon \neq \alpha_k$, $k \in W$. This will include, as a special case, when ε is smaller than the internal

so that

$$\mathsf{P}\left[Q^{\text{iii)}}(\varepsilon)^c\right] \leq \sum_{(S_1, S_2, S_3)\in\mathcal{S}(\varepsilon)} Q(S_1, S_2, S_3, \varepsilon).$$

1) Misclassification Probabilities and Experiment Duration: We examine the asymptotics of the misclassification probability $P^f_{\mathsf{BLTP}(\varepsilon)}$ for large n and small $\|\overline{\alpha}\|$ by the same means as in Section VIII-A. This amounts to finding the mean $D(S_1, S_2, S_3, \varepsilon)$ and asymptotic variance $\sigma^2(S_1, S_2, S_3, \varepsilon)$ of the distribution of $\hat{D}(S_1, S_2, S_3, \varepsilon)$, then finding the dominant exponent D^2/σ^2 over the various (S_1, S_2, S_3). Let $\overline{\alpha}^f(\varepsilon) = \min_{i\in W^\varepsilon} \overline{\alpha}_i$ denote the smallest internal link loss rate of $\mathcal{T}^\varepsilon$ larger than ε and $\overline{\alpha}^p(\varepsilon) = \max_{i\in W\setminus W^\varepsilon} \overline{\alpha}_i$ the largest internal link loss rate of $\mathcal{T}$ smaller than ε or $\overline{\alpha}^p(\varepsilon) = 0$ if no such loss rate exists (which occurs when ε is smaller than all internal links loss rate). The proof of the following result is similar to that of Theorem 10 and is omitted.

Theorem 11: Let $(\mathcal{T}, \alpha)$ be a canonical loss tree. For each $0 \leq \varepsilon < 1$, $(S_1, S_2, S_3) \in \bigcup_{i\in W^\varepsilon} \mathcal{S}(i) \cup \mathcal{S}(\varepsilon)$

$$\sqrt{n}\cdot(\hat{D}(S_1, S_2, S_3, \varepsilon) - D(S_1, S_2, S_3, \varepsilon))$$

converges in distribution, as the number of probes $n \to \infty$, to a Gaussian random variable with mean 0 and variance $\sigma^2(S_1, S_2, S_3, \varepsilon)$. Furthermore, as $\|\overline{\alpha}\| = \max_{k\in V} \overline{\alpha}_k \to 0$ and $\varepsilon/\|\overline{\alpha}\| \to c \in (0, \infty)$

i) $D(S_1, S_2, S_3, \varepsilon) = s(a(S_1 \cup S_2)) - s(a(S_1 \cup S_3)) -$

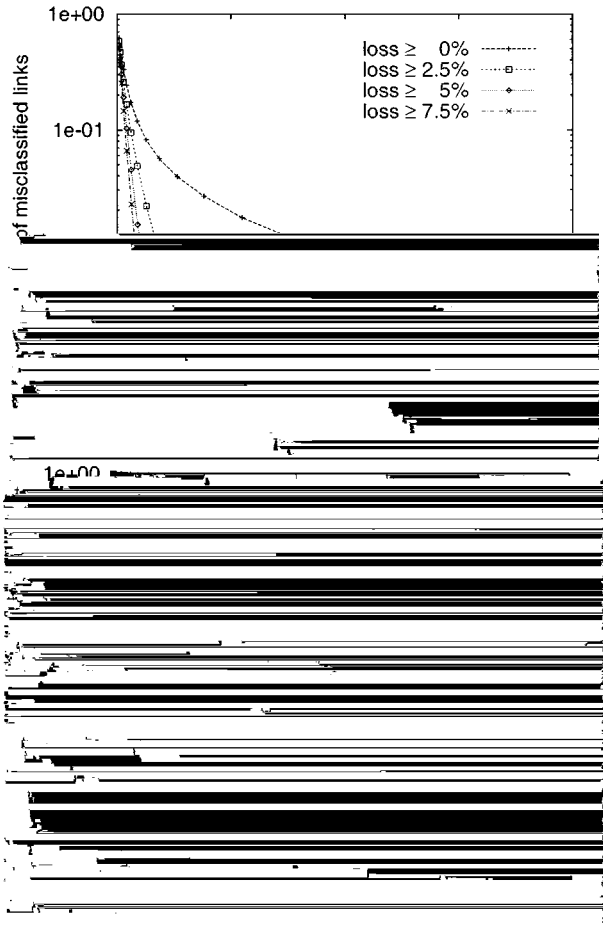


Fig. 19. Misclassification and misgrouping in

Of the algorithms presented, only the Bayesian is able to identify links with arbitrarily small loss rates. All the other classifiers require a parameter $\varepsilon > 0$ that acts as a threshold: a link with loss rate below this value will be ignored and its endpoints identified. The threshold is required in order that sibling groups not be separated due to random fluctuations of the inferred loss rates. However, we do not believe that the neces-

then before each execution of the while loop at line 4 of Fig. 2, the set R' is a stratum and the set (V', L') of nodes and links is consistent with the actual tree (V, L) in the sense that it decomposes over subtrees rooted at the stratum R', i.e., $V' = \bigcup_{k \in R'} V(k$

been grouped, the remaining pairs are still minimizers of $B(\cdot)$ among all pairs of the reduced set $(R' \setminus U^{(\ell)}) \cup \{U^{(\ell)}\}$ in line 10 of Fig. 4. Hence,