Estimation of parameters in a network reliability model with spatial dependence.

Ian Dinwoodie

Technical Report #2003-11
October 3, 2003

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Estimation of parameters in a network reliability model with spatial dependence.

I. H. Dinwoodie
Duke University

October 3 2003

Abstract. An iterative method based on a fixed-point property is proposed for finding maximum likelihood estimators for parameters in a model of network reliability with spatial dependence. The method is shown to converge at a geometric rate under natural conditions on data.

Key words. EM-algorithm, maximum likelihood, multicast, network reliability, network tomography.

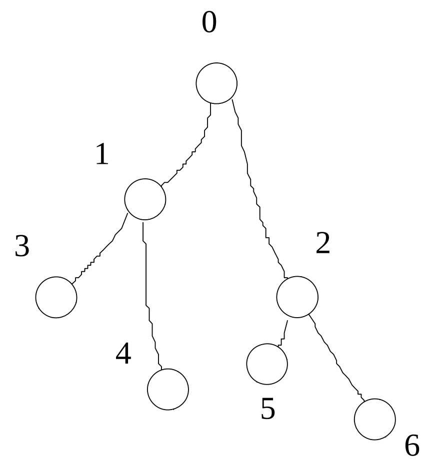
AMS Subject classifications. 62B05, 62F10

1. Introduction.

A model for network reliability with spatial dependence was formulated in [5] that generalized the Bernoulli model of [3]. An approximate maximum likelihood estimator (mle) was proposed based on a one-step relaxation method. In this paper, we describe an iterative scheme to find the numerical values of the mle based on a fixed point property.

Recent work on multicast network tomography has developed methodology for networks more general than trees [2], for missing data [6], scalability and sample size [7], and efficient probing [10], always under the assumption of independent link failures. The focus here is on the model with dependent link failures, for the simple tree topology. The iterative procedure is explained also for the traditional Bernoulli model with independent link failures. The model for dependence adds a single interaction parameter θ which corresponds to a temperature in an interaction potential over all pairs of links. The model is an analog of the Curie-Weiss model. The extra parameter complicates the identifiability and estimation because the recursive method of [3] no longer is applicable. The foundations were shown in [5], but an efficient procedure for exact maximum likelihood was not given. Rather an efficient approximate method was given that modified the estimates for the Bernoulli model. Here we give a new iterative method for maximizing the likelihood.

Let us recall the problem and introduce some notation. A tree with vertices V and edge set E has root node $0 \in V$ and "leaf" nodes $R \subset V$ (R stands for receivers). The parent nodes will be denoted $V_P := V - R$, and V_0 will be the collection of non-root nodes. All vertices in V_P will be assumed to have at least two child nodes. The multicast statistical experiment is the following. One probe is sent from the root node 0 towards the receiver nodes R, and it copies itself at each vertex onto each subsequent edge on its trip towards the receiver nodes (this is the meaning of "multicast"). The probe is lost on an edge on route to the leaf nodes with a probability that depends on the edge, say $\beta_i \in [0, 1)$ for the edge that connects vertex i to its parent $f(i)$. (We will assume that $\beta_i > 0$ in order to get parameter identifiability.) The observed data is the vector $\mathbf{y} \in \{0, 1\}^R$, where component i indicates whether the multicast signal was lost ($\mathbf{y}_i = 1$ means it was lost) on the trip from 0 to the leaf node $i \in R$. The data $\mathbf{y}$ is the image under a many-to-one linear map A of a hidden outcome $\mathbf{x}$ indicating success or failure on each edge. On the tree (1.1) below for example, $V_0 = \{1, 2, 3, 4, 5, 6\}, c = 6, V_P = \{0, 1, 2\}, R = \{3, 4, 5, 6\}$, and $f(3) = f(4) = 1$.



(1.1)

This experiment is repeated independently and identically $n \geq 1$ times, and observations $Y_1, \ldots, Y_n$ are a random sample of iid $\{0,1\}$ vectors at the receiver nodes. The goal is to estimate internal reliability parameters on edges from the incomplete receiver data. This can be done with at least two probability models, the original Bernoulli model and an interaction model with an extra parameter θ that encourages or discourages multiple losses.

2. Bernoulli Multicast Model.

In this section we describe in detail the Bernoulli multicast model of [3], for which a recursive estimation algorithm exists, but which also can solved with the proposed method.

Let $\mathcal{T}$ be the tree with vertices V numbered $0, 1, \ldots, c$. Let the descendants of node i (the set of nodes whose path back to 0 goes through i, but not including i) be denoted $d(i)$. The siblings of i would be $f^{-1} \circ f(i)$. The assumption that all parent vertices (V_P) have at least two child nodes means that for each $i \in V_P$ it holds that $f^{-1} \circ f(i) - \{i\} \neq \emptyset$.

Basic hidden outcomes are vectors of counts $\mathbf{x} = (x_i)_{i \in V_0} \in \{0,1\}^{V_0}$, where x_i specifies how many probes were lost on the edge $\{f(i), i\}$ (the edges are labelled by the outer vertex). The multicast data $\mathbf{y}$ can be written as a many-to-one function of basic hidden outcomes $\mathbf{x}$ with the help of a routing matrix A. The matrix A will have $d = |R|$ rows, one for each leaf node, and $c = |V_0|$ columns indexed by edges. The row for leaf node v will have "1" in column j if j is on the path from 0 to leaf node i. For the binary tree in figure (1.1), the matrix is given by

$$A = \begin{array}{c} \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}. \qquad (2.1)$$

The experiment is repeated $n \geq 1$ times. Then the total observed loss vector $\mathbf{y}_k$ (for experiment k out of n) at leaf nodes is given by

$$\mathbf{y}_k = A\mathbf{x}_k.$$

It is convenient to have a $|V| \times |V_0|$ routing matrix B for all nodes in V, not just the leaf nodes. The row for vertex i would have "1" in each column for vertices on the path from 0 to i. $(B\mathbf{x})_i$ for a node $i \in V_0$ would give the total number of messages lost along the path from 0 out to i. The row for vertex 0 in B is identically 0.

Now let β_i be the probability that a probe from vertex $f(i)$ will fail to cross edge $\{f(i), i\}$ to reach vertex $i \in V_0$. We will use an odds ratio parametrization of the failure probability β_i:

$$\beta_i = \frac{\lambda_i}{1+\lambda_i}, \;\; \lambda_i \geq 0.$$

For the Bernoulli model it is assumed that a probe fails to cross edge $\{f(i), i\}$ with probability $\lambda_i/(1+\lambda_i)$, and edges and probes all behave independently given failure count data on parent nodes (we generalize this below). The distribution μ_λ on $S_0 = \{\mathbf{x} = (x_i) \in Z_+^{V_0} : x_i \in \{0,1\}\}$ is

$$\begin{aligned}
\mu_\lambda(\mathbf{x}) &= \prod_{i \in V_0} \binom{1-(B\mathbf{x})_{f(i)}}{x_i} \frac{\lambda_i^{x_i}}{(1+\lambda_i)^{1-(B\mathbf{x})_{f(i)}}} \\
&= \lambda^{\mathbf{x}} \prod_{i \in V_0} \binom{1-(B\mathbf{x})_{f(i)}}{x_i} \frac{1}{(1+\lambda_i)^{1-(B\mathbf{x})_{f(i)}}} \\
&= \lambda^{\mathbf{x}} \prod_{i \in V_0} \frac{\binom{1-(B\mathbf{x})_{f(i)}}{x_i}}{(1+\lambda_i)} \prod_{i \in V_0} (1+\lambda_i)^{(B\mathbf{x})_{f(i)}} \\
&= \lambda^{\mathbf{x}} \prod_{i \in V_0} \frac{\binom{1-(B\mathbf{x})_{f(i)}}{x_i}}{(1+\lambda_i)} \prod_{i \in V_0 - R} \prod_{j \in d(i)} (1+\lambda_j)^{x_i}.
\end{aligned} \tag{2.2}$$

For $i \in V_0$, let $p_i = \prod_{j \in d(i)} (1+\lambda_j)$. Then (2.2) can be written

$$\begin{aligned}
\mu_\lambda(\{\mathbf{y}\}) &= \sum_{\{\mathbf{x} \in Z_+^c : A\mathbf{x} = \mathbf{y}\}} h(\mathbf{x}) \frac{\lambda^{\mathbf{x}} \prod_{i \in V_0} p_i(\lambda)^{x_i}}{z_\lambda} \\
&= \sum_{\{\mathbf{x} \in Z_+^c : A\mathbf{x} = \mathbf{y}\}} h(\mathbf{x}) \frac{\lambda^{\mathbf{x}} \mathbf{p}(\lambda)^{\mathbf{x}}}{z_\lambda},
\end{aligned}$$

where

$$\begin{aligned}
h(\mathbf{x}) &= \prod_{i \in V_0} \binom{1-(B\mathbf{x})_{f(i)}}{x_i} \\
z_\lambda &= \prod_{i \in V_0} (1+\lambda_i),
\end{aligned} \tag{2.3}$$

and the notation $\mathbf{p}(\lambda)^{\mathbf{x}}$ is the usual representation of $\prod_{i=1}^{c} p_i(\lambda)^{x_i}$. The vector of parameters $(\lambda_i : i \in V_0)$ is identifiable, meaning that two different vectors give rise to two different distributions $\mu_\lambda(\{\mathbf{y}\})$ on the observed (incomplete) data $\mathbf{y}$. (In the Bernoulli model, identifiability holds in fact even if vertex v_0 has only one child, and the failure probabilities are allowed to be zero.) This gives consistent maximum likelihood estimates (the estimates converge to the true parameter values) as the sample size n increases.

Now we describe some new variables that simplify the likelihood function. Consider the one-to-one reparametrization $\gamma_i = \lambda_i p_i(\lambda)$, for $i \in V_0$, from the set of positive reals in R^{V_0} to itself. Then

$$\mu_{\lambda(\gamma)}(\{\mathbf{y}\}) = \sum_{\{\mathbf{x}\in Z_+^c : A\mathbf{x}=\mathbf{y}\}} h(\mathbf{x}) \frac{\gamma^{\mathbf{x}}}{z_{\lambda(\gamma)}}.$$

For each $\mathbf{y} \in Z_+^R$, let $V^{\mathbf{y}} \subset V_0$ be the collection of edge labels that are closest to the root whose failure could lead to observation $\mathbf{y}$. For example, in the binary tree (1.1), $V^{(1,1,1,1)} = \{1,2\}$, $V^{(1,1,0,1)} = \{1,6\}$. If one defines a partial order on V_0 by $w \leq_{\mathcal{T}} v$ if and only if w is on the path from the root 0 to v, then $V^{\mathbf{y}}$ is the collection of minimal elements with respect to this order of the set $\pi^{-1}(\mathbf{y})$ where $\pi : 2^{V_0} \to \{0,1\}^R$ is defined by

$$\pi(A) = I_{\cup_{w\in A}\{v\in R \,:\, v\geq_{\mathcal{T}} w\}}.$$

Define polynomials $q_v,\ v \in V_0$ in variables $s_v,\ v \in V_0$ recursively by

$$\begin{aligned} q_r &:= s_r, r \in R \\ q_v &:= s_v + \prod_{w\in f^{-1}(v)} q_w. \end{aligned}$$

The polynomials q_v have a probabilistic meaning. For each vertex $v \in V_0$, let $\mathbf{y}^v$ be the vector in Z_+^R with 1 in each coordinate that is a descendent of v (including v if v is a receiver). Then $\mu_\lambda(\{\mathbf{y}^v\}) = q_v/z_\lambda$. The following proposition follows from the independence in the Bernoulli model. We will use the standard notation $\mathbf{q}^{\mathbf{x}}$ for $\prod_{v\in V_0} q_v^{x_v}$.

Proposition 2.1. For the Bernoulli model,

$$\begin{aligned} \mu_{\lambda(\gamma)}(\{\mathbf{y}\}) &= \frac{1}{z_{\lambda(\gamma)}} \mathbf{q}(\gamma)^{V^{\mathbf{y}}} \\ z_{\lambda(\gamma)} &= \sum_{\mathbf{y}\in\{0,1\}^R} \mathbf{q}(\gamma)^{V^{\mathbf{y}}}. \end{aligned}$$

Let $N^v, v \in V_0$ be the number of observations in the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ such that the corresponding collections $V^{\mathbf{y}_i}$ include v:

$$N^v := \#\{i : 1 \leq i \leq n, v \in V^{\mathbf{y}_i}\}.$$

Now we can represent the distribution μ_λ in a simplified form:

$$\prod_{i=1}^{n} \mu_{\lambda(\gamma)}(\{\mathbf{y}_i\}) = \frac{1}{z^n_{\lambda(\gamma)}} \prod_{i=1}^{n} \prod_{v\in V^{\mathbf{y}_i}} q_v(\gamma)$$
$$= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{v\in V_0} \prod_{i:v\in V^{\mathbf{y}_i}} q_v(\gamma)$$
$$= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{v\in V_0} q_v(\gamma)^{N^v},$$

which shows that $(N^v)_{v\in V_0}$ are sufficient statistics. This leads to a simple form of the Bernoulli log-likelihood function l_B in the parameters γ_i:

$$l_B(\gamma) = \frac{1}{n} \sum_{v\in V_0} N^v \log q_v(\gamma) - \log z_{\lambda(\gamma)}. \tag{2.4}$$

Define the polynomial $Z(q_1, \ldots, q_c)$ by

$$Z(\mathbf{q}) := \sum_{\mathbf{y}\in\{0,1\}^R} \prod_{v\in V^{\mathbf{y}}} q_v = \sum_{\mathbf{y}\in\{0,1\}^R} \mathbf{q}^{V^{\mathbf{y}}}.$$

Proposition 2.2.

$$z_{\lambda(\gamma)} = Z(\mathbf{q}(\gamma)).$$

An interior stationary point for l_B can be found as a positive solution to a system of polynomial equations in the variables q_v, by the chain rule for derivatives. Using the definition above for $Z(\mathbf{q})$ in terms of $q_v, v \in V_0$, consider l_B in the variables $q_1, \ldots, q_c$:

$$l_B(\mathbf{q}) = \frac{1}{n} \sum_{v\in V_0} N^v \log q_v - \log Z(\mathbf{q}). \tag{2.5}$$

Setting $\nabla l_B = \mathbf{0}$ (assuming an interior stationary point) leads to c polynomial equations:

$$\frac{N^v}{n} = \frac{\sum_{\mathbf{y}:v\in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}}}{\sum_{\mathbf{y}} \mathbf{q}^{V^{\mathbf{y}}}}, \quad v \in V_0$$

which is a polynomial system with a fixed point property in the vector $\mathbf{q}$:

$$q_v = \frac{N^v}{n} \frac{\sum_{\mathbf{y}} \mathbf{q}^{V^{\mathbf{y}}}}{\sum_{\mathbf{y}:v\in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}}/q_v}, \quad v \in V_0.$$

This suggests the iterative method with transformation $T : R^{V_0} \to R^{V_0}$ as below from some reasonable initial point $\mathbf{q}^0$:

$$\mathbf{q}^{k+1} = T(\mathbf{q}^k), \quad k = 1, 2, 3, \dots$$
$$T(\mathbf{q})_v = \frac{N^v}{n} \frac{\sum_{\mathbf{y}} \mathbf{q}^{V^{\mathbf{y}}}}{\sum_{\mathbf{y}: v \in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}}/q_v}, \quad v \in V_0.$$

Obviously this would not give positive values if $\mathbf{q} = 0$ or $N^v/n = 0$, so there will be some conditions on the initial point and the data in §4 for it to work.

Example 2.1. Consider the binary tree (1.1). To find an interior stationary point for the function $l_B(\mathbf{q}, \theta)$, we solve the following system for positive q_v, θ:

$$\begin{aligned}
\frac{N^1}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_1(1+q_5+q_6+q_2)\\
\frac{N^2}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_2(1+q_3+q_4+q_1)\\
\frac{N^3}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_3(1+q_5+q_6+q_2)\\
\frac{N^4}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_4(1+q_5+q_6+q_2)\\
\frac{N^5}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_5(1+q_3+q_4+q_1)\\
\frac{N^6}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_6(1+q_3+q_4+q_1).
\end{aligned}$$

3. Multicast Model with Spatial Dependence: Interaction Model

We have described the original model of [3], but the proposed method is only needed for the interaction model, which has an additional parameter $\theta \geq 0$ affecting the probability of multiple losses and breaking the Markov property. The new model reduces to the Bernoulli model when $\theta = 1$. The range of the interaction is across all edges, and values of θ greater than 1 mean that multiple losses are more likely than they would be under the Bernoulli model.

For $\mathbf{x} \in \{0,1\}^{V_0}$, let $|\mathbf{x}| := \sum_{i \in V_0} x_i$ be the number of 1's in $\mathbf{x}$. Then the notation $[|\mathbf{x}| - 1]_+$ will give $|\mathbf{x}| - 1$ if there are two or more 1's in $\mathbf{x}$, otherwise it will vanish. The new law $\nu_{\gamma,\theta}$ in parameters $\gamma_i > 0, i = 1, \dots, c, \theta \geq 0$, is specified by

$$\begin{aligned}
\nu_{\gamma,\theta}(\mathbf{x}) &:= h(\mathbf{x}) \frac{\gamma^{\mathbf{x}} \theta^{[|\mathbf{x}|-1]_+}}{w_{\gamma,\theta}}\\
w_{\gamma,\theta} &= \frac{\theta - 1 + z_{\lambda(\theta\gamma)}}{\theta}
\end{aligned} \tag{3.1}$$

where the formula for the normalizing constant $w_{\gamma,\theta}$ relates to z from the Bernoulli model as follows. Consider the one-to-one reparametrization from positive γ to positive λ with inverse given by

$$\gamma_i = \lambda_i p_i(\lambda), \quad i = 1, \ldots, c.$$

Then $\lambda(\theta\gamma)$ is the vector $(\lambda_1(\theta\gamma), \ldots, \lambda_c(\theta\gamma))$ that comes from finding the λ corresponding to $(\theta\gamma_1, \theta\gamma_2, \ldots, \theta\gamma_c)$. From the Bernoulli model, we know that

$$\prod_{i=1}^{c}(1 + \lambda_i(\gamma\theta)) = \sum_{\mathbf{x}\in Z_+^c} h(\mathbf{x})\gamma^{\mathbf{x}}\theta^{|\mathbf{x}|}$$

and separating the case $\mathbf{x} = \mathbf{0}$ gives the formula. In terms of the odds-ratio parameters λ_i, the law ν can be written

$$\nu_{\gamma(\lambda),\theta}(\mathbf{x}) := h(\mathbf{x}) \, \frac{\lambda^{\mathbf{x}}\mathbf{p}(\lambda)^{\mathbf{x}}\theta^{[|\mathbf{x}|-1]_+}}{w_{\gamma(\lambda),\theta}}. \tag{3.2}$$

The parameter pair (γ, θ) is called identifiable for positive γ and nonnegative θ if $\nu_{\gamma,\theta}(\{\mathbf{x} \geq 0 : A\mathbf{x} = \mathbf{y}\}) = \nu_{\gamma',\theta'}(\{\mathbf{x} \geq 0 : A\mathbf{x} = \mathbf{y}\})$ for all $\mathbf{y} \in Z_+^d$ implies that $\gamma = \gamma'$ and $\theta = \theta'$, where γ and γ' are assumed to be positive in each coordinate and θ and θ' are assumed to be nonnegative real numbers. If the parameters are identifiable, then different parameters will lead to different statistical patterns and consistent estimation is possible under repeated, independent experiments. Otherwise, different parameter values may be statistically indistinguishable based on repeated experimental outcomes.

In [5] it was proved that the parameters (λ, θ) are identifiable if the tree $\mathcal{T}$ has the property that all parent nodes have at least two children.

4. Estimation and Inference.

In this section, we propose a numerical method for finding maximum likelihood estimates of the unknown parameter values. With the "incomplete" data $\mathbf{y}_i = A\mathbf{x}_i$ as a many-to-one function of outcomes $\mathbf{x}_i$, it seems that the EM-algorithm [4] is appropriate. The EM-algoritm is implemented in [2], [6] and [7]. The paper [7] has some interesting approximations on the speed of convergence of the EM-algorithm, which is governed by the largest eigenvalue of a transformation matrix. The EM-algorithm is complicated when applied directly. Below we present an iterative method that has some similarities with the EM-algorithm, in that it is a fixed-point argument with geometric convergence and its justification relies on convexity. However, it does not seem to be included in the description of the EM-algorithm, rather it is a method for the "M-step" as defined in equation (2.3) p. 4, of [4]. The method is essentially a numerical way to compute a Legendre transform in a special case. The convergence to a global maximum is established under conditions. This is comparable to the results in [9] that strengthen the convergence conclusions of the EM iteration.

The basic idea can be illustrated most simply with the example of finding the value of the odds ratio parameter λ in n Bernoulli trials, say with x successes. The objective function is $\lambda^x/(1+\lambda)^n$, and the stationary point $\hat{\lambda}$ satisfies the fixed point equation $\lambda = T(\lambda) := (x/n)(1+\lambda)$. Then with $\lambda^0 = 1$ one gets a sequence of converging approximations $1, 2x/n, x/n + 2(x/n)^2, \ldots \to x/(n-x)$, as long as $x < n$.

Let the observations for an iid sample be $\mathbf{y}_1 = A\mathbf{x}_1, \ldots, \mathbf{y}_n = A\mathbf{x}_n$. Observe that the relationship between the Bernoulli model μ_λ and the interaction model $\nu_{\lambda,\theta}$ implies the following formula:

$$\nu_{\gamma,\theta}(\{\mathbf{y}\}) = \mu_{\lambda(\theta\gamma)}(\{\mathbf{y}\})(\frac{\theta I_{\mathbf{0}}(\mathbf{y}) + I_{\neq\mathbf{0}}(\mathbf{y})}{\theta - 1 + z_{\lambda(\theta\gamma)}})z_{\lambda(\theta\gamma)}$$

where $\theta\gamma := (\theta\gamma_1, \ldots, \theta\gamma_c)$.

Let $N_{\mathbf{0}}$ be the number of times the vector $\mathbf{0}$ appears in the sample. The objective function for maximum likelihood estimation is the log-likelihood function l in (γ, θ) given by

$$\begin{aligned} l(\gamma,\theta) &= \frac{1}{n}\sum_{i=1}^{n} \log \nu_{\gamma,\theta}(\{\mathbf{y}_i\}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log(\mu_{\lambda(\theta\gamma)}(\{\mathbf{y}_i\})) + \frac{N_{\mathbf{0}}}{n}\log(\theta) + \log(\frac{z_{\lambda(\theta\gamma)}}{\theta - 1 + z_{\lambda(\theta\gamma)}}). \end{aligned} \tag{4.1}$$

It follows that $l(\gamma,\theta) = l_0(\theta\gamma, \theta)$, where l_0 is defined by

$$l_0(\gamma',\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \mu_{\lambda(\gamma')}(\{\mathbf{y}_i\}) + \frac{N_{\mathbf{0}}}{n}\log(\theta) + \log(\frac{z_{\lambda(\gamma')}}{\theta - 1 + z_{\lambda(\gamma')}}) \tag{4.2}$$

The objective function l_0 (4.2) can be simplified. For each $\mathbf{y} \in Z_+^R$, let $V^{\mathbf{y}} \subset V_0$ be the collection of edge labels that are closest to the root whose failure could lead to observation $\mathbf{y}$. For example, in the binary tree (1.1), $V^{(1,1,1,1)} = \{1, 2\}$, $V^{(1,1,0,1)} = \{1, 6\}$. Define polynomials $q_v,\ v \in V_0$ in variables $s_v,\ v \in V_0$ recursively by

$$\begin{aligned} q_r &:= s_r, r \in R \\ q_v &:= s_v + \prod_{w \in f^{-1}(v)} q_w. \end{aligned}$$

By the independence in the Bernoulli model,

$$\mu_{\lambda(\gamma)}(\{\mathbf{y}\}) = \sum_{A\mathbf{x}=\mathbf{y}} h(\mathbf{x})\frac{\gamma^{\mathbf{x}}}{z_{\lambda(\gamma)}} = \frac{1}{z_{\lambda(\gamma)}} \prod_{v \in V^{\mathbf{y}}} q_v(\gamma),$$

which leads to the formula for the normalizing constant $z_{\lambda(\gamma)}$ in terms of the variables q_v:

$$z_{\lambda(\gamma)} = \sum_{\mathbf{y}\in\{0,1\}^R} \prod_{v\in V^{\mathbf{y}}} q_v(\gamma).$$

Let $N^v, v \in V_0$ be the number of observations in the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ such that the corresponding collections $V^{\mathbf{y}_i}$ include v:

$$N^v := \#\{i : 1 \le i \le n, v \in V^{\mathbf{y}_i}\}.$$

Now we can represent the distribution μ_λ in a simplified form:

$$\begin{aligned}\prod_{i=1}^n \mu_{\lambda(\gamma)}(\{\mathbf{y}_i\}) &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{i=1}^n \prod_{v\in V^{\mathbf{y}_i}} q_v(\gamma) \\ &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{v\in V_0} \prod_{i:v\in V^{\mathbf{y}_i}} q_v(\gamma) \\ &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{v\in V_0} q_v(\gamma)^{N^v}.\end{aligned}$$

This leads to a simpler form of the objective function l_0:

$$l_0(\gamma, \theta) = \frac{1}{n} \sum_{v\in V_0} N^v \log q_v(\gamma) + \frac{N_{\mathbf{0}}}{n} \log(\theta) - \log(\theta - 1 + z_{\lambda(\gamma)}). \tag{4.3}$$

The procedure to maximize l over (γ, θ) is to maximize l_0 over γ', θ and transform back:

$$\begin{aligned}(\hat\gamma', \hat\theta) &:= \texttt{arg max}_{\gamma'>0,\theta\ge 0}\, l_0(\gamma', \theta) \\ (\hat\gamma, \hat\theta) &:= (\hat\gamma'/\hat\theta, \hat\theta).\end{aligned} \tag{4.4}$$

Define the polynomial $Z(q_1, \ldots, q_c)$ by

$$Z(\mathbf{q}) := \sum_{\mathbf{y}\in\{0,1\}^R} \mathbf{q}^{V^{\mathbf{y}}}. \tag{4.5}$$

With indeterminates $q_c > q_{c-1} > \ldots > q_1$ for a ring $\mathbf{R}[q_c, \ldots, q_1]$ ordered so that the variables indexed by a node's children are greater than the one for the node itself, quantities above can be described precisely in algebraic terms. Let G be the Gröbner basis in $\mathbf{R}[q_c, \ldots, q_1]$ given by $G = \{\prod_{w\in f^{-1}(v)} q_w \; - q_v, v \in V_0 - R\}$. Then $V^{\mathbf{y}}$ is the collection of indeterminates present in the monomial $\texttt{nf}(\mathbf{q}^{\mathbf{y}}, G)$, where $\texttt{nf}$ denotes the normal form with respect to plex order of the monomial $\mathbf{q}^{\mathbf{y}} := \prod_{v\in R} q_v^{y_v}$. It can be shown that

$$\begin{aligned}z_{\lambda(\gamma)} &= Z(\mathbf{q}(\gamma)) \\ Z(\mathbf{q}) &= \texttt{nf}(\prod_{v\in R} 1 + q_v, G) = \sum_{\mathbf{y}\in\{0,1\}^R} \texttt{nf}(\mathbf{q}^{\mathbf{y}}, G).\end{aligned}$$

An interior stationary point for $l_0(\gamma, \theta)$ can be found as a positive solution to a system of polynomial equations in the variables q_v, θ, by the chain rule for derivatives. Using the definition above for $Z(\mathbf{q})$ in terms of $q_v, v \in V_0$, consider l_0 in the variables $q_1, \ldots, q_c$:

$$l_0(\mathbf{q}, \theta) = \frac{1}{n} \sum_{v \in V_0} N^v \log q_v + \frac{N_{\mathbf{0}}}{n} \log(\theta) - \log(\theta - 1 + Z(\mathbf{q})). \tag{4.6}$$

Setting $\nabla l_0 = \mathbf{0}$ leads to $c+1$ polynomial equations satisfied by the mle $(\hat{\mathbf{q}}, \hat{\theta})$:

$$\begin{aligned} \frac{N^v}{n}(Z(\mathbf{q}) + \theta - 1) &= q_v \big(\sum_{\mathbf{y}: v \in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}} / q_v \big) \\ \frac{N_{\mathbf{0}}}{n}(Z(\mathbf{q}) + \theta - 1) &= \theta. \end{aligned} \tag{4.7}$$

If we define a transformation T on $\mathbf{R}^c \times \mathbf{R}$ by

$$\begin{aligned} T(\mathbf{q}, \theta)_v &= \frac{N^v}{n} \frac{Z(\mathbf{q}) + \theta - 1}{\sum_{\mathbf{y}: v \in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}} / q_v}, \quad 1 \le v \le c \\ T(\mathbf{q}, \theta)_{c+1} &= \frac{N_{\mathbf{0}}}{n}(Z(\mathbf{q}) + \theta - 1), \end{aligned} \tag{4.8}$$

then the desired optimal interior values $(\hat{\mathbf{q}}, \hat{\theta})$ satisfy the equation

$$(\hat{\mathbf{q}}, \hat{\theta}) = T(\hat{\mathbf{q}}, \hat{\theta}).$$

This leads to the iterative method with T starting with initial values $(\mathbf{q}^0, \theta^0)$:

$$(\mathbf{q}^{k+1}, \theta^{k+1}) = T(\mathbf{q}^k, \theta^k), \quad k = 0, 1, 2, \ldots$$

Example 4.1. Consider the binary tree (1.1). To find an interior stationary point for the function $l_0(\mathbf{q}, \theta)$, we solve the following system q_v, θ:

$$\begin{aligned}
&\frac{N^1}{n}((1 + q_3 + q_4 + q_1)(1 + q_5 + q_6 + q_2) + \theta - 1) = q_1(1 + q_5 + q_6 + q_2) \\
&\frac{N^2}{n}((1 + q_3 + q_4 + q_1)(1 + q_5 + q_6 + q_2) + \theta - 1) = q_2(1 + q_3 + q_4 + q_1) \\
&\frac{N^3}{n}((1 + q_3 + q_4 + q_1)(1 + q_5 + q_6 + q_2) + \theta - 1) = q_3(1 + q_5 + q_6 + q_2) \\
&\frac{N^4}{n}((1 + q_3 + q_4 + q_1)(1 + q_5 + q_6 + q_2) + \theta - 1) = q_4(1 + q_5 + q_6 + q_2) \\
&\frac{N^5}{n}((1 + q_3 + q_4 + q_1)(1 + q_5 + q_6 + q_2) + \theta - 1) = q_5(1 + q_3 + q_4 + q_1) \\
&\frac{N^6}{n}((1 + q_3 + q_4 + q_1)(1 + q_5 + q_6 + q_2) + \theta - 1) = q_6(1 + q_3 + q_4 + q_1) \\
&\frac{N_{\mathbf{0}}}{n}((1 + q_3 + q_4 + q_1)(1 + q_5 + q_6 + q_2) + \theta - 1) = \theta.
\end{aligned}$$

The transformation T on $(\mathbf{q}, \theta)$ is

$$
\begin{aligned}
T(\mathbf{q},\theta)_1 &= \frac{N^1}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_5+q_6+q_2) \\
T(\mathbf{q},\theta)_2 &= \frac{N^2}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_3+q_4+q_1) \\
T(\mathbf{q},\theta)_3 &= \frac{N^3}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_5+q_6+q_2) \\
T(\mathbf{q},\theta)_4 &= \frac{N^4}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_5+q_6+q_2) \\
T(\mathbf{q},\theta)_5 &= \frac{N^5}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_3+q_4+q_1) \\
T(\mathbf{q},\theta)_6 &= \frac{N^6}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_3+q_4+q_1) \\
T(\mathbf{q},\theta)_7 &= \frac{N_{\mathbf{0}}}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1).
\end{aligned}
$$

The solution must be transformed to $\hat{\gamma}'$, then again transformed to $\hat{\gamma}$ using (4.4).

To prove theoretical results about the optimization procedure, introduce new parameters $\phi := (\phi_1, \phi_2, \ldots, \phi_{c+1}) = (\log(\mathbf{q}), \log(\theta))$ and statistics $\mathbf{t}' := (t_1 = N^1/n, \ldots, t_c = N^c/n, t_{c+1} = N_{\mathbf{0}}/n)$. This simplifies the objective function l_0 to a standard concave form:

$$
\begin{aligned}
l_0(\phi) &= \phi \cdot \mathbf{t} - \log(\zeta_\phi) \\
\zeta_\phi &:= \sum_{\mathbf{y} \in \{0,1\}^R} e^{\phi \cdot \mathbf{a_y}} = Z(\mathbf{q}) + \theta - 1
\end{aligned}
$$

where $\mathbf{a_y} \in \{0,1\}^{c+1}$ is a vector defined in terms of the standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_{c+1}$ by

$$
\begin{aligned}
\mathbf{a_y} &= \sum_{i \in V^{\mathbf{y}}} \mathbf{e}_i \quad \texttt{if } \mathbf{y} \neq \mathbf{0} \\
\mathbf{a_0} &= \mathbf{e}_{c+1}.
\end{aligned}
$$

Let $C \subset \mathbf{R}^{c+1}$ be the closed, convex hull of the vectors $\mathbf{a_y}, \mathbf{y} \in \{0,1\}^R$. Then we have the linear equation $\mathbf{t}' = A\mathbf{n}$ if A is the matrix with 2^R columns equal to the collection $\mathbf{a_y}, \mathbf{y} \in \{0,1\}^R$, and $\mathbf{n}$ is a vector of length 2^R that counts the number of each of the 2^R types of outcomes $\mathbf{y}$ in the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$. So if the sample includes at least one of each type of vector $\mathbf{y}$, then the data of sufficient statistics $\mathbf{t}$ will be in the interior of C, as a nontrivial convex combination of the defining vectors of C.

Theorem 4.1. If the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ includes at least one of each of the 2^R possibilities (or more generally, suppose $\mathbf{t} \in \texttt{int}\, C$) then there is a unique positive solution $(\hat{\mathbf{q}}, \hat{\theta})$ to the system (4.7) and a unique positive fixed point for T in (4.8).

Proof: This follows from Theorem 9.13 of [1] with parametrization $\phi = (\log(\mathbf{q}), \log(\theta))$.

Theorem 4.2. Suppose the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ includes at least one of each of the 2^R possibilities (or more generally suppose $\mathbf{t} \in \mathtt{int}\, C$). If $(\mathbf{q}, \theta)$ is sufficiently close to the optimal values $(\hat{\mathbf{q}}, \hat{\theta})$, then the sequence $T^n(\mathbf{q}, \theta)$ converges at a geometric rate to the fixed point $(\hat{\mathbf{q}}, \hat{\theta})$.

Proof: Let $\bar{T}(\phi) := \log(T(e^{\phi}))$ be the transformation T in the ϕ variables, which with (4.8) becomes

$$\begin{aligned}
\bar{T}(\phi)_v &= \log(N^v/n) - \log\Big(\frac{\sum_{v \in V^{\mathbf{y}}} \mathbf{q}^{V^y}/q_v}{\zeta_\phi}\Big) \\
&= \phi_v + \log(t_v) - \log \partial_{\phi_v} \log(\zeta_\phi) \\
&= \phi_v - \log\Big(\frac{\partial_{\phi_v} \log(\zeta_\phi)}{\partial_{\phi_v} \log(\zeta_{\hat{\phi}})}\Big),
\end{aligned}$$

using the optimality condition $\partial_{\phi_v} \log(\zeta_{\hat{\phi}}) = t_v$ for interior $\mathbf{t}$. The above, together with the formula for the last variable $\log(\theta) = \phi_{c+1}$ gives the final representation for the transformation T in terms of the canonical parameters ϕ :

$$\bar{T}(\phi) = \phi + \log \mathbf{t} - \log(\nabla \log \zeta_\phi) = \phi - \log\Big(\frac{\nabla \log \zeta_\phi}{\nabla \log \zeta_{\hat{\phi}}}\Big). \tag{4.9}$$

It is enough to show that the eigenvalues of the derivative $D_{\hat{\phi}}\bar{T}$ are strictly less than one in absolute value.

The derivative $D_{\hat{\phi}}\bar{T}$ is a matrix with rows $\nabla \bar{T}(\hat{\phi})_v$ indexed by edge labels $v \in V_0$. Now

$$\begin{aligned}
\partial_{\phi_w} \bar{T}(\hat{\phi})_v &= \delta_{vw} + 0 - \frac{\partial_{\phi_w}\partial_{\phi_v} \log \zeta_{\hat{\phi}}}{\partial_{\phi_v} \log \zeta_{\hat{\phi}}} \\
&= \delta_{vw} - \frac{\Sigma_{\hat{\phi}}(v,w)}{t_v},
\end{aligned}$$

where $\Sigma_{\hat{\phi}}(v,w)$ is the covariance of coordinates v, w in the vectors $\{\mathbf{a_y}\}$ with probabilities in the exponential family $p_\phi(\mathbf{a_y}) = \frac{e^{\phi \cdot \mathbf{a_y}}}{\zeta_\phi}$. Thus

$$\begin{aligned}
D_{\hat{\phi}}\bar{T} &= I - A_{\mathbf{t}}\Sigma_{\hat{\phi}} \\
A_{\mathbf{t}} &:= \begin{pmatrix} 1/t_1 & 0 & 0\ldots & \\ 0 & 1/t_2 & 0 & \ldots \\ 0 & 0 & \ldots & 1/t_{c+1} \end{pmatrix}.
\end{aligned}$$

Therefore it is sufficient to show that the eigenvalues of $A_{\mathbf{t}}\Sigma_{\hat{\phi}}$ are in $(0, 1]$, which is the same as having the eigenvalues of $\sqrt{A_{\mathbf{t}}}\Sigma_{\hat{\phi}}\sqrt{A_{\mathbf{t}}}$ in $(0, 1]$. Now $\sqrt{A_{\mathbf{t}}}\Sigma_{\hat{\phi}}\sqrt{A_{\mathbf{t}}}$ is the covariance matrix for $\sqrt{A_{\mathbf{t}}} \cdot \mathbf{a}$, or in other words $\sqrt{A_{\mathbf{t}}}\Sigma_{\hat{\phi}}\sqrt{A_{\mathbf{t}}}\,(v,w) = \mathtt{Cov}_{\hat{\phi}}(\mathbf{a}_v/\sqrt{t_v}, \mathbf{a}_w/\sqrt{t_w})$. This matrix is positive definite as a nondegenerate covariance matrix. Furthermore, the eigenvalues of any covariance matrix are bounded above by the maximum of the diagonal entries, and $\mathtt{Cov}_{\hat{\phi}}(\mathbf{a}_v/\sqrt{t_v}, \mathbf{a}_v/\sqrt{t_v}) = \mathtt{Var}_{\hat{\phi}}(\mathbf{a}_v)/t_v$. Since the vectors $\mathbf{a}_{\mathbf{y}}$ have entries in $\{0,1\}$, $\mathtt{Var}_{\hat{\phi}}(\mathbf{a}_v) = E_{\hat{\phi}}(\mathbf{a}_v - t_v)^2 \leq E_{\hat{\phi}}(\mathbf{a}_v)^2 \leq E_{\hat{\phi}}(\mathbf{a}_v) = t_v$. ∎

Theorem 4.3. Suppose the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ includes at least one of each of the 2^R possibilities (or more generally suppose $\mathbf{t} \in \mathtt{int}\, C$). If $(\mathbf{q}, \theta)$ is any positive initial value, then the sequence $T^n(\mathbf{q}, \theta)$ converges to the fixed point $(\hat{\mathbf{q}}, \hat{\theta})$.

Proof: The argument essentially compactifies the parameter space in ϕ with the moment map τ mapping into the compact, convex set C (defined to be the closed convex hull of the vectors $\{\mathbf{a}_{\mathbf{y}}\}$) . Let $\tau(\phi) = E_\phi(\mathbf{a}) = \nabla \log \zeta_\phi$. It is standard exponential-family theory that τ is continuous with continuous inverse from $\mathbf{R}^{c+1} \to \mathtt{int}\ C$. Consider the sequence of means $\mathbf{m}_n := \tau(\log T^n(\mathbf{q}, \theta)) \in \mathtt{int}\ C$. We will be done if we can show that $\mathbf{m}_n \to \mathbf{t}$.

Suppose not. Then there is a convergent subsequence $\mathbf{m}_{n_k} \to \mathbf{s} \in C$. If $\mathbf{s} \in \mathtt{int}\ C$, then it would correspond to a fixed point of positive parameter values $\phi_{\mathbf{s}}$, which is impossible by uniqueness for $\mathbf{s} \neq \mathbf{t}$. Therefore, $\mathbf{s}$ must be on the boundary of C. Let v be coordinate label where $\mathbf{s}_v \neq \mathbf{t}_v$, an let ϵ be the radius of a small neighborhood of $\mathbf{s}$ in which $\mathbf{s}_v \neq \mathbf{t}_v$. This neighborhood contains all values $\mathbf{m}_{n_k}$ for large k. The formula $\bar{T}(\phi) = \phi + \log \mathbf{t} - \log(\nabla \log(\zeta_\phi))$ implies that the v-coordinate of ϕ_{n_k} will actually move in the direction of $\mathbf{t}_v$ for all large k. This contradicts convergence to $\mathbf{s}$. Therefore, the entire sequence must converge to $\mathbf{t}$. ∎

We have proposed an iterative scheme for solving a polynomial system of equations to find the maximum likelihood estimators for the problem of network reliability, and we have proven that the method has essential convergence qualities. In practice, it seems to be stable and efficient. It is much easier to implement than a general-purpose polynomial system solver, because it finds quickly the one real, positive solution. The ϕ variables make a quasi-Newton method possible (whereas the $(\mathbf{q}, \theta)$ variables are not good with quasi-Newton because of nonnegativity conditions and a lack of convexity). But a quasi-Newton method will be relatively complex because of complicated components like ζ_ϕ, which can be handled easily with variable transformations in the proposed method.

The conditions on the data under which the method works are natural and are satisfied with probability converging geometrically to one with the sample size. On the other hand, the conditions require a sample size on the order of 2^R, where R is the number of receiver nodes. We believe such a sample size is necessary for maximum likelihood estimation. A further area of research would be Bayesian methodology, which can make likelihood functions well-behaved with smaller sample sizes.

We believe the proposed method is valuable for its simplicity and global convergence, but there are other numerical methods that may be tried for maximum likelihood estimation in this application. The EM-algorithm suggests itself at first, as a method that is designed for incomplete data. But it seems to be particularly complicated for this problem. The reparametrizations used in this paper give a much more tractable objective function, and the optimizer then satisfies a polynomial system. This means that one may try new methods of semi-definite programming, as described for example in Parrilo and Sturmfels [8] and implemented in SOS Tools. Also, there are polynomial homotopy methods that would start with a solution to the Bernoulli model, but which can track down many unwanted complex solutions. A comparative numerical study would be interesting.

Acknowledgments. This work was supported by NSF grant DMS-0200888 and SAMSI under grant DMS-0112069.

References

[1] O. Barndorff-Nielsen, *Information and Exponential Families.* Wiley, New York (1978).

[2] T. Bu, N. Duffield, F. Lo Presti, D. Towsley, Network tomography on general topologies. *Proceedings ACM Sigmetrics 2002* Marina Del Ray June 15-19 2002.

[3] R. Cáceres, N. G. Duffield, J. Horowitz, D. Towsley, T. Bu, Multicast-based inference of network internal characteristics: accuracy of packet loss estimation. *IEEE Transactions in Information Theory*, **45** (2000) 2462-2480.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39** (1997) 1-38.

[5] I. H. Dinwoodie, and E. Mosteig, Statistical inference for network reliability with spatial dependence. *SIAM Journal on Discrete Mathematics* (to appear).

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

[7] C. Ji, and A. Elwalid, Measurement-based network monitoring and inference: scalability and missing information. *IEEE Journal on Selected Areas in Communications,* **20** (2002) 714-725.

[8] P. Parrilo, and B. Sturmfels, Minimizing polynomial functions. http://www.cds.caltech.edu/ pablo/pubs.htm (2001).

[9] C. F. Jeff Wu, On the convergence of the EM algorithm. *Annals of Statistics*, **11** (1983), 95-103.

[10] B. Xi, G. Michailidis, and V. N. Nair, Estimating network internal losses using a new class of probing experiments. University of Michigan Department of Statistics Technical Report 397 (2003).

ISDS
Box 90251
Duke University
Durham, NC 27705-0251
ihd@stat.duke.edu