Estimating Point-to-Point and Point-to-Multipoint Traffic Matrices: An Information-Theoretic Approach

Yin Zhang, *Member, IEEE,* Matthew Roughan, *Member, IEEE,* Carsten Lund, *Member, IEEE,* and David Donoho

***Abstract*— Traffic matrices are required inputs for many IP network management tasks, such as capacity planning, traffic engineering and network reliability analysis. However, it is difficult to measure these matrices directly in large operational IP networks, so there has been recent interest in inferring traffic matrices from link measurements and other more easily measured data. Typically, this inference problem is ill-posed, as it involves significantly more unknowns than data. Experience in many scientific and engineering fields has shown that it is essential to approach such ill-posed problems via "regularization". This paper presents a new approach to traffic matrix estimation using a regularization based on "entropy penalization". Our solution chooses the traffic matrix consistent with the measured data that is information-theoretically closest to a model in which source/destination pairs are stochastically independent. It applies to both point-to-point and point-to-multipoint traffic matrix estimation. We use fast algorithms based on modern convex optimization theory to solve for our traffic matrices. We evaluate our algorithm with real backbone traffic and routing data, and demonstrate that it is fast, accurate, robust, and flexible.**

***Index Terms*— Traffic matrix estimation, information theory, minimum mutual information, regularization, traffic engineering, SNMP, point-to-point, point-to-multipoint, failure analysis.**

I. Introduction

Traffic matrices, which specify the amount of traffic between origin and destination in a network, are required inputs for many IP network management tasks, such as capacity planning, traffic engineering and network reliability analysis. However, it is often difficult to measure these matrices directly in large operational IP networks. So there has been a surge of interest in inferring traffic matrices from link load statistics and other more easily measured data [1], [2], [3], [4], [5].

Traffic matrices may be estimated or measured at varying levels of detail [6]: between Points-of-Presence (PoPs) [4], routers [5], links, or even IP prefixes [7]. The finer grained traffic matrices are generally more useful, for example, in the analysis of the reliability of a network under a component failure. During a failure, IP traffic is rerouted to find the new path through the network, and one wishes to test if this would cause a link overload anywhere in the network. Failure of a link within a PoP may cause traffic to reroute via alternate links within the PoP without changing the inter-PoP routing. Thus to understand failure loads on the network we must measure traffic at a router-to-router level. In general, the inference problem is more challenging at finer levels of detail, the finest so far considered being router-to-router.

Estimating traffic matrices from link loads is a non-trivial task. The challenge lies in the ill-posed nature of the problem: for a network with N ingress/egress points we need to estimate the N^2 origin/destination demands. At a PoP level N is in the tens, at a router level N may be in the hundreds, at a link level N may be tens of thousands, and at the prefix level N may be of the order of one hundred thousand. However, the number of pieces of information available, the link measurements, remains approximately constant. One can see the difficulty — for large N the problem becomes massively underconstrained.

There is extensive experience with ill-posed linear inverse problems from fields as diverse as seismology, astronomy, and medical imaging [8], [9], [10], [11], [12], all leading to the conclusion that some sort of side information must be brought in, with results that may be good or bad depending on the quality of this information. All of the previous work on IP traffic matrix estimation has incorporated prior information: for instance, Vardi [1] and Tebaldi and West [2] assume a Poisson traffic model, Cao et al. [3] assume a Gaussian traffic model, Zhang et al. [5] assume an underlying gravity model, and Medina et al. [4] assume a logit-choice model. Each method is sensitive to the accuracy of this prior: for instance, [4] showed that the methods in [1], [2], [3] were sensitive to their prior assumptions, while [5] showed that their method improved if the prior (the so called gravity model) was generalized to reflect real routing rules more accurately.

In contrast, this paper starts from a regularization formulation of the problem drawn from the field of ill-posed problems, and derives a prior distribution that is most appropriate to this problem. Our prior assumes source/destination independence, until proven otherwise by measurements. The method then blends measurements with prior information, producing the reconstruction closest to independence, but consistent with the measured data. The method proceeds by solving an optimization problem that is understandable and intuitively appealing. This approach allows a convenient implementation using modern optimization software, with the result that the algorithm is very efficient.

An advantage of the approach used in this paper is that it also provides some insight into alternative algorithms. For instance, the simple gravity model of [5] is equivalent to complete independence of source and destination, while the generalized gravity model corresponds to independence conditional on source and destination link classes. Furthermore, the algorithm of [5] is a first-order approximation of the algorithm presented here, explaining the success of that algorithm, and suggesting that it also can be extended to measure point-to-

Manuscript received March 13, 2004; revised November 11, 2004. An earlier version of this paper appeared in *Proc. ACM SIGCOMM*, August 2003.

Y. Zhang, and Carsten Lund are with AT&T Labs–Research, 180 Park Av., Florham Park, NJ, 07932. {yzhang,lund}@research.att.com.

M. Roughan is with the School of Mathematical Sciences at the University of Adelaide, Adelaide 5005, Australia, matthew.roughan@adelaide.edu.au.

D. Donoho is with Stanford University.

Digital Object Identifier: xxxxxx

multipoint demand matrices. Our method opens up further opportunities for extensions, given the better understanding of the importance of prior information about network traffic and how it can be incorporated into the process of finding traffic matrices. For instance, an appealing alternative prior generation procedure is suggested in [4]. Alternatively, the Bayesian method of [2] can be placed into the optimization framework here, with a different penalty function, as could the methods of [1], [3].

Our approach also allows us to estimate both point-to-point traffic matrices and point-to-multipoint demand matrices. Prior work on estimating traffic matrices from link data has concentrated on the point-to-point traffic, *i.e.*, the traffic from a single source to a single destination. While point-to-point traffic matrices are of great practical importance, they are not always enough for applications (as shown in [7]). Under some failures the traffic may actually change its origin and destination; its network entry and exit points. The point-to-point traffic matrix will be altered, because the point-to-point traffic matrix describes the "carried" load on the network between two points. In contrast, the *demand matrix* describes the "offered" traffic demands on the IP network and is therefore invariant under a much larger class of changes. The demand matrix is inherently point-to-multipoint in the sense that traffic coming into the network from a customer, may often depart the network via multiple egress points in order to reach its final destination. To understand this, consider a packet entering a backbone ISP through a customer link, destined for another backbone ISP's customer. Large North-American backbone providers typically are connected at multiple peering points. Our packet could reach its final destination through any of these peering links; the actual decision is made through a combination of Border Gateway Protocol (BGP) and Interior Gateway Protocol (IGP) routing protocols. If the normal exit link fails, then the routing protocols would choose a different exit point. In a more complicated scenario, the recipient of the packet might be multi-homed — that is, connected to more than one ISP. In this case the packet may exit the first ISP through multiple sets of peering links. Finally, even single homed customers may sometimes be reached through multiple inter-AS (Autonomous System) paths.

We test the estimation algorithm extensively on network traffic and topology data from an operational backbone ISP (AT&T's North American IP network). The results show that the algorithm is fast, and accurate for point-to-point traffic matrix estimation. We also test the algorithm on topologies generated through the Rocketfuel project [13], [14], [15] to resemble alternative ISPs, providing useful insight into where the algorithm will work well. One interesting side result is that there is a relationship between the network traffic and topology that is beneficial in this estimation problem. We also test the sensitivity of the algorithm to measurements errors, demonstrating that the algorithm is highly robust to errors and missing data in the traffic measurements.

We further examine some alternative measurement strategies that could benefit our estimates. We examine two possibilities: the first (suggested in [4]) is to make direct measurements of some rows of the traffic matrix, the second is to measure local traffic matrices as suggested in [16]. Both result in improvements in accuracy, however, we found in contrast to [4] that the order in which rows of the traffic matrix are included does matter — adding rows in order of the largest row sum first is better than random ordering.

Finally, the results of our evaluation of the algorithm for point-to-multipoint demand matrices are interesting in that these estimates are less accurate than the corresponding point-to-point results, for the very good reason that this estimation problem contains more ambiguity. However, we also show in this paper that the results are far more accurate (than point-to-point results) when used in real applications such as link failure analysis. In fact, the point-to-multipoint estimates produce astoundingly accurate link failure estimates. Likewise, in [17], we have also demonstrated that the resulting accuracy is well within the bounds required for another operational task, IGP route optimization.

To summarize, this paper demonstrates a specific tool that works well on large scale point-to-point and point-to-multipoint traffic matrix estimation. The results show that it is important to add appropriate prior information. Our prior information is based on independence-until-proven-otherwise, which is plausible, computationally convenient, and results in accurate estimates.

The paper begins in Section II with some background: definitions of terminology and descriptions of the types of data available. Section III describes the regularization approach used here, and our algorithm, followed by Section IV, the evaluation methodology, and Section V, which shows the algorithm's performance on a large set of measurements from an operational tier-1 ISP. Section VI examines the algorithm's robustness to errors in its inputs, and Section VII shows the flexibility of the algorithm to incorporate additional information. Section VIII shows the results for point-to-multipoint estimation, and Section IX demonstrates the utility of the point-to-multipoint results in reliability analysis. We conclude the paper in Section X.

II. Background

A. Network

An IP network is made up of routers and adjacencies between those routers, within a single AS or administrative domain. It is natural to think of the network as a set of nodes and links, associated with the routers and adjacencies, as shown in Figure 1. We refer to routers and links that are wholly internal to the network as *Backbone* Routers (BRs) and links, and refer to others as *Edge* Routers (ERs) and links.

One could compute traffic matrices with different levels of aggregation at the source and destination end-points, for instance, at the level of PoP to PoP, or router to router, or link to link [6]. In this paper, we are primarily interested in computing router to router traffic matrices, which are appropriate for a number of network and traffic engineering applications, and can be used to construct more highly aggregated traffic matrices (e.g. PoP to PoP) using topology information [6]. We may further specify the traffic matrix to be between BRs, by aggregating up to this level.

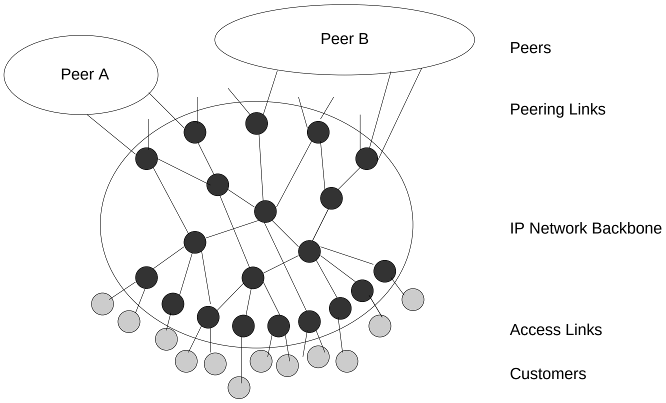


Fig. 1. IP network components and terminology

In addition, it is helpful for IP networks managed by Internet Service Providers (ISPs) to further classify the edge links. We categorize the edge links into *access* links, connecting customers, and *peering* links, which connect other (non-customer) ASs. A significant fraction of the traffic in an ISP is *inter-domain* and is exchanged between customers and peer networks. Today traffic to peer networks is largely focused on dedicated peering links, as illustrated in Figure 1. Under the typical routing policies implemented by large ISPs, very little traffic will transit the backbone from one peer to another. Transit traffic between peers may reflect a temporary step in network consolidation following an ISP merger or acquisition, but should not occur under normal operations.

In large IP networks, distributed routing protocols are used to build the forwarding tables within each router. It is possible to predict the results of these distributed computations from data gathered from router configuration files, or a route monitor such as [18]. In our investigation, we employ a routing simulator such as in [19] that makes use of this routing information to compute a routing matrix (defined in Section III-A). Note that this simulation includes load balancing across multiple shortest paths.

B. *Traffic Data*

In IP networks today, link load measurements are readily available via the Simple Network Management Protocol (SNMP). SNMP is unique in that it is supported by essentially every device in an IP network. The SNMP data that is available on a device is defined in a abstract data structure known as a Management Information Base (MIB). An SNMP *poller* periodically requests the appropriate SNMP MIB data from a router (or other device). Since every router maintains a cyclic counter of the number of bytes transmitted and received on each of its interfaces, we can obtain basic traffic statistics for the entire network with little additional infrastructure.

The properties of data gathered via SNMP are important for the implementation of a useful algorithm — SNMP data has many limitations. Data may be lost in transit (SNMP uses unreliable UDP transport; copying to our research archive may also introduce loss). Data may be incorrect (through poor router vendor implementations). The sampling interval is coarse (in our case 5 minutes). Many of the typical problems in SNMP data may be mitigated by using hourly traffic averages (of five minute data), and we shall use this approach. The problems with the finer time-scale data make time-series approaches to traffic matrix estimation more difficult.

We use flow level data in this paper for validation purposes. This data is aggregated by IP source and destination address, and port numbers at each router. This level of granularity is sufficient to obtain a real traffic matrix [7], and in the future such measurement may provide direct traffic matrix measurements, but at present limitations in vendor implementations prevent collection of this data from the entire network.

C. *Information Theory*

Information theory is of course a standard tool in communications systems [20], but a brief review will set up our terminology. We begin with basic probabilistic notation: we define $p_X(x)$ to mean the probability that a random variable X is equal to x. We shall typically abuse this notation (where it is clear) and simply write $p(x) = p_X(x)$. Suppose that X and Y are independent random variables, then

$$p(x,y) = p(x)p(y), \tag{1}$$

i.e. the joint distribution is the product of its marginals. This can be equivalently written using the conditional probability

$$p(x|y) = p(x). \tag{2}$$

In this paper we shall typically use the source S and the destination D of a packet (or bit), rather than the standard random variables X and Y. Thus $p(d|s)$ is the conditional probability of a packet (bit) exiting the network at $D = d$, given that it entered at $S = s$, and $p(d)$ is the unconditional probability of a packet (bit) going to $D = d$.

We can now define the Discrete Shannon Entropy of a discrete random variable X taking values x_i as

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i), \tag{3}$$

The entropy is a measure of the uncertainty about the value of X. For instance, if $X = x_1$ with certainty, then $H(X) = 0$, and $H(X)$ takes its maximum value when X is uniformly distributed, when the uncertainty about its value is greatest.

We can also define the conditional entropy of one random variable Y with respect to another X by

$$H(Y|X) = -\sum_j p(x_j) \sum_i p(y_i|x_j) \log_2 p(y_i|x_j), \tag{4}$$

where $p(y_i|x_j)$ is the probability that $Y = y_i$ conditional on $X = x_i$. $H(Y|X)$ can be thought of as the uncertainty remaining about Y given that we know the outcome of X. Notice that the joint entropy of X and Y can be shown to be

$$H(X,Y) = H(X) + H(Y|X). \tag{5}$$

We can also define the Shannon information

$$I(Y|X) = H(Y) - H(Y|X), \tag{6}$$

which therefore represents the decrease in uncertainty about Y from measurement of X, or the information that we gain about Y from X. The information is symmetric, $I(X|Y) = I(Y|X)$ and so we can refer to this as the *mutual information* of X and Y, and write as $I(X,Y)$. Note that $I(X,Y) \geq 0$, with equality if and only if X and Y are independent — when X and Y are independent X gives us no additional information about Y. The mutual information can be written in a number of ways, but here we write it

$$I(X,Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} = K(p_{x,y} || p_x \times p_y), \tag{7}$$

where $K(f||g) = \sum_i f_i \log(f_i/g_i)$ is the Kullback-Leibler divergence of f with respect to g, a well-known measure of distance between probability distributions.

Discrete Entropy is frequently used in coding because the entropy $H(X)$ gives a measure of the number of bits required to code the values of X. That is, if we had a large number n of randomly-generated instances $X_1, X_2, \ldots, X_n$ and needed to represent this stream as compactly as possible, we could represent this stream using only $nH(X)$ bits, using entropy coding as practiced for example in various standard commercial compression schemes.

Entropy has also been advocated as a tool in the estimation of probabilities. Simply put, the *maximum entropy principle* states that we should estimate an unknown probability distribution by enumerating all the constraints we know it must obey on ‘physical’ grounds, and searching for the probability distribution that maximizes the entropy subject to those constraints. It is well known that the probability distributions occurring in many physical situations can be obtained by the maximum entropy principle. Heuristically, if we had no prior information about a random variable X, our uncertainty about X is at its peak, and therefore we should choose a distribution for X which maximizes this uncertainty, or the entropy. In the case where we do have information about the variable, usually in the form of some set of mathematical constraints C, then the principle states that we should maximize the entropy $H(X|C)$ of X conditional on consistency with these constraints. That is, we choose the solution which maintains the most uncertainty while satisfying the constraints. The principle can also be derived directly from some simple axioms which we wish the solution to obey [21].

D. Ill-Posed Linear Inverse Problems

Many scientific and engineering problems can be posed as follows. We observe data $\mathbf{y}$ which are thought to follow a system of linear equations

$$\mathbf{y} = A\mathbf{x}, \tag{8}$$

where the n by 1 vector $\mathbf{y}$ contains the data, and the p by 1 vector $\mathbf{x}$ contains unknowns to be estimated. The matrix A is an n by p matrix. In many cases of interest $p > n$, and so there is no unique solution to the equations. Such problems are called *ill-posed linear inverse problems*. In addition, frequently the data are noisy, so that it is more accurate to write

$$\mathbf{y} = A\mathbf{x} + \mathbf{z}. \tag{9}$$

In that case any reconstruction procedure needs to remain stable under perturbations of the observations. In our case, $\mathbf{y}$ are the SNMP link measurements, $\mathbf{x}$ is the traffic matrix written as a vector, and A is the routing matrix.

There is extensive experience with ill-posed linear inverse problems from fields as diverse as seismology, astronomy, and medical imaging [8], [9], [10], [11], [12], all leading to the conclusion that some sort of side information must be brought in, producing a reconstruction which may be good or bad depending on the quality of the prior information. Many such proposals solve the minimization problem

$$\min_{\mathbf{x}} \|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda^2 J(\mathbf{x}), \tag{10}$$

where $\|\cdot\|_2$ denotes the L_2 norm, $\lambda > 0$ is a regularization parameter, and $J(\mathbf{x})$ is a penalization functional. Proposals of this kind have been used in a wide range of fields, with considerable practical and theoretical success when the data matched the assumptions leading to the method, and the regularization functional matched the properties of the estimand. These are generally called *strategies for regularization of ill-posed problems* (for a more general description of regularization see [22]).

A general approach to deriving such regularization ideas is the Bayesian approach (such as used in [2]), where we model the estimand $\mathbf{x}$ as being drawn at random from a so-called ‘prior’ probability distribution with density $\pi(\mathbf{x})$ and the noise $\mathbf{z}$ is taken as a Gaussian white noise with variance σ^2. Then the so-called posterior probability density $p(\mathbf{x}|\mathbf{y})$ has its maximum $\hat{\mathbf{x}}$ at the solution of

$$\min_{\mathbf{x}} \|\mathbf{y} - A\mathbf{x}\|_2^2 + 2 \cdot \sigma^2 \log \pi(\mathbf{x}). \tag{11}$$

Comparing this with (10) we see the penalized least-squares problems as giving the most likely reconstructions under a given model. Thus the method of regularization has a Bayesian interpretation, assuming Gaussian noise and assuming $J(\mathbf{x}) = \log \pi(\mathbf{x})$. We stress that there should be a good match between the regularization functional J and the properties of the estimand — that is, a good choice of prior distribution. The penalization in (10) may be thought of as expressing the fact that reconstructions are very implausible if they have large values of $J(\cdot)$.

Regularization can help us understand approaches such as that of Vardi [1] and Cao et al. [3], which treat this as a maximum likelihood problem where the $\mathbf{x}$ are independent random variables following a particular model. In these cases they use the model to form a penalty function which measures the distance from the model by considering higher order moments of the distributions.

III. Regularization of the Traffic Estimation Problem using Minimum Mutual Information

The problem of inference of the end-to-end traffic matrix is massively ill-posed because there are so many more routes than links in a network. In this section, we develop a regularization approach using a penalty that seems well-adapted to the structure of actual traffic matrices, and which has some appealing information-theoretic structure. Effectively, among all traffic matrices agreeing with the link measurements, we choose the one that minimizes the mutual information between the source and destination random variables.

Under this criterion, absent any information to the contrary, we assume that the conditional probability $p(d|s)$ that a source

s sends traffic to a destination d is the same as $p(d)$, the probability that the network as a whole sends packets or bytes to destination d. There are strong heuristic reasons why the largest-volume links in the network should obey this principle — they are so highly aggregated that they intuitively should behave similarly to the network as a whole.

On the other hand, as evidence accumulates in the link-level statistics, the conditional probabilities are adapted to be consistent with the link-level statistics in such a way as to minimize the mutual information between the source and destination random variables.

This Minimum Mutual Information (MMI) criterion is well-suited to efficient computation. It can be implemented as a convex optimization problem; in effect one simply adds a minimum weighted entropy term to the usual least-squares lack of fit criterion. There are several widely-available software packages for solving this optimization problem, even on very large scale problems; some of these packages can take advantages of the sparsity of routing matrices.

A. Traffic-Matrix Estimation

Let $N(s,d)$ denote the traffic volume going from source s to destination d in a unit time. Note that $N(s,d)$ is unknown to us; what can be known is the traffic $T(l)$ on link l. Let $A(s,d;l)$ denote the routing matrix, i.e. $A(s,d;l)$ gives the fraction of traffic from s to d which crosses link l (and which is zero if the traffic on this route does not use this link at all). The link-level traffic counts are

$$T(l) = \sum_{s,d} A(s,d;l)N(s,d), \quad \forall l \in L, \tag{12}$$

where L is the set of backbone links. We would like to recover the traffic matrix $N(s,d)$ from the link measurements $T(l)$, but this is the same as solving the matrix equation (8), where $\mathbf{y}$ is a vector containing the traffic counts $T(l)$, $\mathbf{x}$ is a vectorization of the traffic matrix, and A is the routing matrix. A is a matrix which is $\#L$ by ($\#S \times \#D$), where there are $\#L$ link measurements, $\#S$ sources, and $\#D$ destinations.

B. The Independence Model

We propose thinking about $N(s,d)$ in probabilistic terms, so that if a network carries N end-to-end packets (or bits) total within a unit time then the number of packets sent from source s to destination d, $N(s,d)$ say, is a random variable with mean $N \cdot p(s,d)$, with $p(s,d)$ the joint probability that a randomly chosen one of the N packets (or bits) goes from s to d. We consider the marginal probabilities

$$p_S(s) = \sum_d p(s,d), \qquad p_D(d) = \sum_s p(s,d), \tag{13}$$

the chance that a randomly-chosen packet (bit) enters the network at s, and the chance that a randomly chosen packet (bit) departs at d, respectively. We can expand this notation to measure sets:

$$p_{S,D}(Q_s, Q_d) \quad = \quad \sum_{s\in Q_s}\sum_{d \in Q_d} p(s,d), \tag{14}$$

for all sets of source and destination links Q_s, Q_d, and similarly for the marginal probabilities p_S and p_D.

We let S be the random variable obtained looking at the source of a random packet (or bit), and let D denote the destination. Suppose for sake of discussion that S and D are independent random variables. Then (2) means that, given that a packet (bit) originates at $S = s$, it is no more likely to go to $D = d$ than would a randomly-chosen packet (bit) originating anywhere in the network. For networks containing a few extremely high volume links carrying very large fractions of the packets, the assumption (2) should work well for the very largest circuits, since they have been so highly aggregated that their behavior may be very similar to the network as a whole.

Note that the independence of source and destination is equivalent to the simple *gravity model* [4], [5], with the form

$$N(s,d) \approx \text{Const } N(s)N(d) \tag{15}$$

where $N(s)$ is the traffic entering at s, and $N(d)$ is the traffic exiting at d. While there is experience with the gravity model above and some success in its application, it is also known that it gives results that are not as accurate as may be obtained using additional information [4], [5].

Section II suggests that regularization is a way of using prior information in conjunction with link measurements to help decide which traffic matrices from the set satisfying (8) are more plausible. We propose using a regularization functional that uses the independence/gravity model as a point of departure, but which considers other models as well. Recall from our discussion of information theory that independence of source and destination is tantamount to the statement that the mutual information vanishes: $I(S,D) = 0$. Recall also that $I(S,D) \geq 0$. It follows that the penalty functional on traffic matrices $p(s,d)$, is given by $J(p) \equiv I(S,D) \geq 0$, with equality if and only if S and D are independent.

This functional $J(\cdot)$ has an interpretation in terms of the compressibility of addresses in IP headers. Suppose we have a large number of IP headers — abstracted to be simply source/destination address pairs (s_i, d_i), $i = 1, \ldots, N$. We want to know: what is the minimal number of bits required (per header) to represent the source destination pair. It turns out that this is just $H(S) + H(D) - I(S,D)$. Now if we simply applied entropy compression to the S_i and D_i streams separately, we would pay $H(S) + H(D)$ bits per header to represent headers. Hence the functional $I(S,D)$ measures the number of bits of additional compression possible beyond the separate compression of source and destination based on traditional entropy compression. This extra compression is possible because of special dependencies that make it more likely to have traffic between certain source/destination pairs than we would have expected by independence. In fact measurements of $H(S)$ and $H(D)$ (on real datasets described below) are typically around 5, while $I(S,D)$ is very small, typically around 0.1. This suggests that the independence assumption is a reasonable fit to the real data, at least on average. There may be some links for which it is not, but the MMI method specifically allows for correction to these (see below).

Suppose we adopt a Bayesian viewpoint, assigning an *a priori* probability $\pi(p)$ to the traffic matrix p that is proportional to $2^{-J(p)}$. Then we are saying we regard as *a* *priori* implausible those traffic matrices where much higher compression is possible based on joint source-destination pairs as compared to compression of sources and destinations separately. Each bit saved reduces our *a priori* likelihood by about a factor $1/2$.

C. Regularization Method

We propose now to reconstruct traffic matrices by adopting the regularization prescription (10) with the regularization functional $J(p) = I(S, D)$. Translating (10) into traffic-matrix notation, we seek to solve

$$\text{minimize } \sum_l \left(T(l) - N \sum_{s,d} A(s,d;l) p(s,d) \right)^2 + \lambda^2 I(S,D), \tag{16}$$

Recalling the Bayesian interpretation of regularization, we are saying that we want a traffic matrix which is a tradeoff between matching the observed link traffic counts and having *a priori* plausibility, where our measure of plausibility, as just explained, involves the 'anomalous compressibility' of source-destination pairs. The traffic matrix obtained as the solution to this optimization will be a compromise between two terms based on the size of λ, which is a proxy for the noise level in our measurements. Note that

$$I(S,D) = \sum_{d,s} p(s,d) \log \frac{p(s,d)}{p(s)p(d)} = K(p(s,d)||p(s)p(d)), \tag{17}$$

where $K(\cdot||\cdot)$ again denotes the Kullback-Leibler divergence. Here $p(s)p(d)$ represents the gravity model, and $K(\cdot||\cdot)$ can be seen as a distance between probability distributions, so that we can see (16) as having an explicit tradeoff between fidelity to the data and deviation from the independence/gravity model. Note also that the Kullback-Leibler divergence is the negative of the relative entropy of $p(s,d)$ with respect to $p(s)p(d)$, and so this method also has an interpretation as a maximum entropy algorithm.

Both terms in the above tradeoff are convex functionals of the traffic matrix p. Hence, for each given λ, they can be rewritten in constrained optimization form:

$$\begin{array}{l} \text{minimize } K(p(s,d)||p(s)p(d)) \text{ subject to} \\ \sum_l (T(l) - N \sum_{s,d} A(s,d;l) p(s,d))^2 \le \chi^2. \end{array} \tag{18}$$

Here $\chi^2 = \chi^2(\lambda)$ is chosen appropriately so that the solution of this problem and the previous one are the same, at the given value of λ. The problem is saying: among all traffic matrices adequately accounting for the observed link counts, find the one closest to the gravity model.

Note that in all these optimization problems, there are additional constraints (as on any probability distribution): non-negativity, normalization, and (13). We leave these implicit.

D. Algorithm

The problem we attack in this paper is the BR-to-BR traffic matrix. While this problem is an order of magnitude more complex than a PoP-to-PoP traffic matrix, a router-to-router traffic matrix is absolutely necessary for many network engineering tasks. A PoP-to-PoP traffic matrix is useful when designing a network from scratch, but typically, in a real network changes are incremental, and so we need to see how these changes affect traffic at the router level. We use techniques from [5] to reduce the size of the problem initially, by removing redundant information, and a large number of traffic matrix elements that we know to be zero from routing information. This processing does not improve accuracy, but does speed up later computations.

To make the exact formulation explicit, we define

$$\begin{aligned} x_i &= N(s_i, d_i), & (19) \\ y_j &= \text{ traffic counts } = T(l_j), & (20) \\ g_i &= N(s_i)N(d_i), & (21) \end{aligned}$$

where

$$\begin{aligned} N &= \text{total traffic in network} & (22) \\ N(s_i) &= \text{total traffic originating at } s_i & (23) \\ N(d_i) &= \text{total traffic departing at } d_i & (24) \end{aligned}$$

and we define the column vectors $\mathbf{x}$, and $\mathbf{y}$ with elements x_i and y_i, respectively. Note that if $N(s_i) = 0$ or $N(d_i) = 0$, then both $g_i = 0$ and $x_i = 0$, so we exclude these i from the penalty function. The problem formulation is then given by

$$\begin{array}{l} \text{argmin}_x \left\{ ||\mathbf{y} - A\mathbf{x}||^2 + \lambda^2 \sum_{i:\ g_i > 0} \frac{x_i}{N} \log \left(\frac{x_i}{g_i} \right) \right\} \\ \text{subject to } x_i \ge 0. \end{array} \tag{25}$$

The additional constraints (normalization, etc.) on the marginal distributions are satisfied by supplementing the routing matrix and measurements to ensure that they include these constraints.

This penalized least-squares formulation has been used in solving many other ill-posed problems, and so there exist publicly available software in Matlab (such as routine MaxEnt in Per Christian Hansen's Inverse Problems Toolbox [23], [24]) to solve small-scale variants of such problems. Our problems are, however, large in scale and not suited to such basic implementations. The problem of solving such large-scale traffic matrices is only possible if we can exploit one of the main properties of routing matrices: they are very sparse — the proportion of exact zero entries in each column and row is overwhelming. Accordingly, we use PDSCO [25], a MATLAB package developed by Michael Saunders of Stanford University, which has been highly optimized to solve problems with sparse matrices. PDSCO has been used (see e.g. [25]) to solve problems of the order 16,000 by 256,000 efficiently. We have found that its performance is very good (taking no more than a few seconds) on the largest problems we consider here.

In principle, the choice of λ depends on the noise level in the measurements, but we show later that the method's performance is highly insensitive to this parameter.

An interesting point is that if one were to have additional information such as used in the choice model of [4] then this could also be incorporated by conditioning the initial model $P_{S,D}(s,d)$ on this information (for an example of this type see Section III-E). Alternatively, such information could be included in the constraints underlying the optimization (as shown in Section VII).

E. Inter-domain Routing

1) Zero Transit Traffic: The above algorithm assumes that independence of source and destination is a reasonable starting model. However, there are good reasons we may want to modify this starting model. In real backbone ISPs, routing is typically asymmetric due to hot-potato routing — traffic from the customer edge to peers will be sent to the "nearest" exit point, while traffic in peer networks will do likewise resulting in a different pattern for traffic from peering to customers. Also there should be no traffic transiting the network from peer to peer [5]. Both these factors demand departures from the gravity/independence model.

Suppose we assume there is zero transit traffic. We suggest that *conditional independence* of source and destination, *given appropriate side information*, will be more accurate than pure independence. More specifically, suppose we have available as side information, the source and destination class (access or peering). We would then model the probabilities of a packet (bit) arriving at s and departing at d as conditionally independent *given the class of arrival and destination link*. In Appendix I we prove that this results in the following model. Define A and P to be the sets of access and peering links, respectively, then the conditionally-independent model is

$$p_{S,D}(s,d) = \begin{cases} \frac{p_S(s)}{p_S(A)}\frac{p_D(d)}{p_D(A)}(1-p_S(P)-p_D(P)), & \text{for } s \in A, d \in A, \\ p_S(s)\frac{p_D(d)}{p_D(A)}, & \text{for } s \in P, d \in A, \\ \frac{p_S(s)}{p_S(A)}p_D(d), & \text{for } s \in A, d \in P, \\ 0, & \text{for } s \in P, d \in P. \end{cases} \tag{26}$$

to which we can naturally adapt the algorithm above (by modifying g_i).

2) Point to Multipoint: As noted in the introduction a point-to-point traffic matrix is not suitable for all applications. Sometimes we need a point-to-multipoint demand matrix, for instance, when we want to answer questions about the impact of link failures outside the backbone, e.g. "would a peering link failure cause an overload on any backbone links?" In this case, traffic would reroute to an alternate exit point, changing the point-to-point traffic matrix in an unknown way. However, the point-to-multipoint demand matrix would remain constant.

Ideally such a matrix would be at the prefix level, but a number of operational realities make an approximation to router level useful for many engineering tasks. The first such reality is that backbone networks that exchange large traffic volumes are connected by private peering links as opposed to Internet Exchange Points. This allows us to see the proportion of traffic going to each individual peer using only SNMP link measurements, so we can partition traffic per peer. The second such reality is that the BGP policies across a set of peering links to a single peer are typically the same. Therefore, the decision as to which peering link to use as the exit point is made on the basis of shortest IGP distance. This distance is computed at the link level, as opposed to BGP policies, which can act at the prefix level. While we cannot test that this property is true for all large ISPs (and in general it is not always true even on the network from which we have measurements), the methodology above does not need this, because the algorithm above only uses this as a prior, to be corrected through the use of link (and other) information.

The step required to generate a point-to-multipoint demand matrix requires consideration of the control ISPs have over interdomain routing. Interdomain routing gives an ISP little control over where traffic enters its network, so we shall not make any changes to (26) for access-to-access, and peering-to-access traffic. However, a provider has considerable control over where traffic will leave their network across the peering edge. Traffic destined for a particular peer may be sent on any of the links to that peer.

The result is that we must modify (26) for access-to-peer traffic. We do so by not specifying which link d in the set of links to peer i (i.e. P_i) is used for traffic leaving the network to peer i. We can do this formally by not specifying $p_{S,D}(s,d)$ for $s \in A, d \in P$ but rather $p_{S,D}(s,P_i)$ for all peers i. This simple point-to-multipoint model can then be used in the estimation through using

$$p_{S,D}(s,P_i) = \frac{p_S(s)}{p_S(A)}p_D(P_i), \tag{27}$$

for $s \in A$, in place of the access-to-peering equation from (26). We do not determine the exit point in the estimates. The algorithm can then proceed by minimizing the mutual information of the final distribution with respect to (26) and (27). The exit points are implicit in the routing matrix used in the optimization (25), but are left undetermined in the estimate, and can therefore be fixed when applied to a particular case.

We should also note that this is a quite general extension. We use it here on sets of peering links P_i, but in a network with different policies, we can partition the peering links in some different fashion (even through a non-disjoint partition) to reflect some particular idiosyncrasies in routing policy.

F. Relationship to Previous Algorithms

The work in this paper presents a general framework, within which we can place a number of alternative methods for estimating IP traffic matrices. For instance, by taking a linear approximation to the log function in the Kullback-Leibler information distance information and exploiting the fact that $\sum_x [f(x) - g(x)] = 0$ we get

$$\begin{aligned} K(f||g) &\approx \sum_x f(x)\left[\frac{f(x)-g(x)}{g(x)}\right] - \sum_x [f(x)-g(x)] \\ &= \sum_x \left[\frac{f(x)-g(x)}{\sqrt{g(x)}}\right]^2. \end{aligned} \tag{28}$$

From this we can see that the MMI solution may be approximated by using a quadratic distance metric with square root weights. This explains the success of the approach in [5], as well as why square root weights give the best performance (which was unknown in [5]). The conditional independence of Section III-E explains the use of the generalized gravity model as an initial condition in [5].

The quadratic optimization is convenient, because it can be simply solved using the Singular Value Decomposition (SVD) [5], with non-negativity enforced by a second step using Iterative Proportional Fitting (IPF) [3]. In this paper we will compare the performance of the pure MMI approach, its quadratic approximation, and the method of [5] (referred to here as SVD-IPF), and we see that the approximation works well in the cases considered. We defer the comparison with maximum likelihood approaches [1], [3], [4] to future work, because scaling these methods to the size of problem described here requires additional techniques (for instance see [26], [27]) that have only recently been developed.

The point of interest here is that the MMI principle above produces (an approximation of) the algorithm previously derived from an initial gravity model solution. However in the case of the MMI solution, the principle precedes practice — that is, the decision to regularize with respect to a prior is not an arbitrary decision, but a standard step in ill-posed estimation problems. The close approximation has a practical impact in that we can use the fact that [5] already demonstrated that the conditional independence of Section III-E to be a better prior than complete independence. We use this fact here by using (26) and (27) in the remainder of the paper.

IV. Evaluation Methodology

In this paper, we apply the traffic matrix benchmarking methodology developed in [5], [28] to real Internet data to validate different algorithms. One major advantage of the methodology is that it can provide a *consistent* data set that is as realistic as practically possible. Below we provide an overview of this methodology, followed by a summary of the performance metrics we use.

A. Validation Methodology

In [5] sampled flow-level data were used, as well as topology and routing information as derived in [19]. Flow level data contains details of numbers of packets and bytes transferred between source and destination IP addresses, and also gives information such as the interface at which the traffic entered our network. Combining these datasets one may derive a traffic matrix [7].

The resulting traffic matrix in our experiments covers around 80% of the real network traffic (including all the peering traffic) on the real topology of a large operational tier-1 ISP. Following [5], we compute the traffic matrices on one hour time scales to deal with the limitations of the measurements. Given these traffic matrices and the network topology and routing information, we only need a consistent set of link load measurements to proceed.

[5] solves the problem of providing a consistent set of traffic, topology and link measurement data as follows. Simulate the network routing using the available topology and routing information. From this we may compute a routing matrix $\mathbf{A}$, and then derive a set of link measurements $\mathbf{y}$ from (8). Thus the traffic matrix $\mathbf{x}$, the routing matrix $\mathbf{A}$ and the measured link loads $\mathbf{y}$ are all consistent. We can then perform the estimation procedure to compute $\hat{\mathbf{x}}$, the traffic matrix estimate.

The validation approach allows us to work with a problem for which we know the "ground truth" — the real traffic matrix. It can also be extended in several different ways. For example, it allows one to take a traffic matrix and apply it on an arbitrary topology, for instance a simulated network such as a star, or a measured topology such as those produced by Rocketfuel [13], [14]. Thus we can gain insight into the effect of different topologies on the performance of the algorithm. We may also introduce controlled measurement errors to assess the algorithm's robustness, or simulate alternative measurements to see their impact in a rigorous manner.

B. Performance Metrics

In this paper we use two basic methods for assessing and comparing the results. The first method is to estimate the relative error (that is, the average of the absolute value of the errors, relative to the average traffic matrix element). The second method is to plot the Cumulative Distribution Function (CDF) of the errors relative to the average traffic matrix element. However, many elements of a router to router traffic matrix are zero due to routing constraints, and these constrained elements are easy to estimate. This results in a large number of entries to the traffic matrix with near zero error. To more accurately indicate the errors on the positive elements we separate the zero and non-zero elements and compute their errors separately. The errors on the zero elements are very small (99% of the errors are below 1%), and so we shall not display these separately here. We shall report the relative errors of the positive elements.

V. Performance

A. Sensitivity to the Choice of λ

The choice of the parameter λ determines how much weight is given to independence, versus the routing constraint equations. One typically wants to find a λ such that $\|A\mathbf{x} - \mathbf{y}\| \leq \epsilon\|\mathbf{y}\|$, where ϵ specifies the desired level of accuracy to which the linear constraints $A\mathbf{x} = \mathbf{y}$ should be satisfied. This can be done by applying a line search process exploiting the fact that we are optimizing with respect to a unimodal function.

In our experiments, however, we find that the algorithm's performance is not sensitive to the choice of λ. Figure 2 shows the relative error in the estimates for varying λ. Figure 2 (a) and (b) show the results for the quadratic and MMI algorithms respectively, for a single-hour data set given different levels of error in the input measurements (see below for details of the introduced measurement errors). Figure 2 (c) and (d) show the average results over a month of data.

Most notably, in each graph there is a distinct region where the curves are all quite flat, and that this region is largely the same regardless of the error level. Thus the choice of λ is insensitive to the level of noise in the measurements, and it is easy to choose a good value. We choose a fixed value from the middle of the insensitive range, $\lambda = 0.01$ throughout the rest of the paper, with a result that is at worst only a few percent off that for the optimal choice of λ.

B. Comparison of Algorithms

We now compare the three algorithms described above (MMI, quadratic optimization, and SVD-IPF) applied to the problem of computing a BR-to-BR traffic matrix. The results below are based on 506 data sets from AT&T's North American IP network, representing the majority of June 2002, and

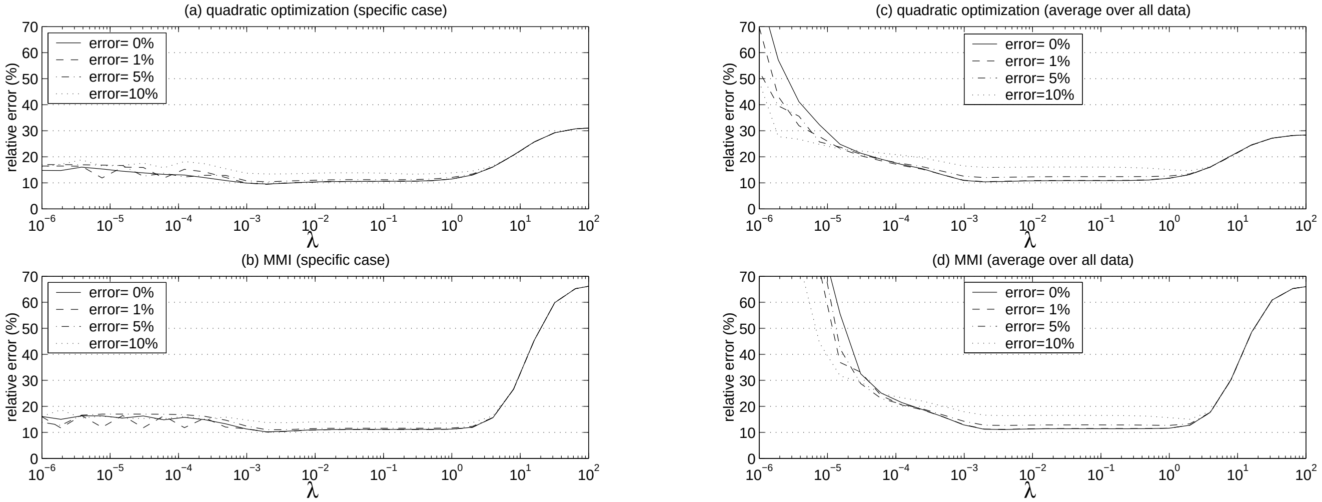


Fig. 2. The relative errors for the quadratic and MMI algorithms for a given value of λ.

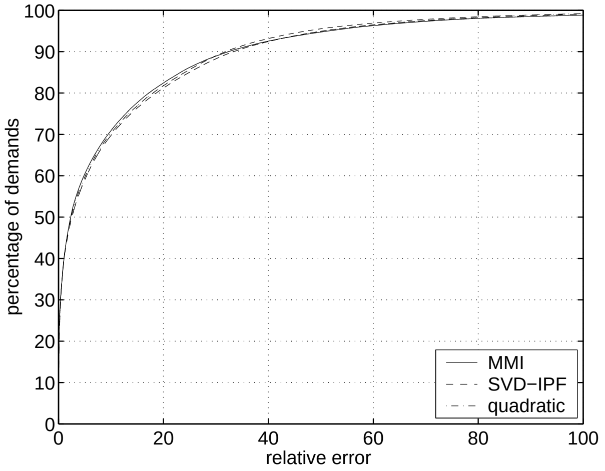


Fig. 3. A comparison of the relative errors for the methods.

covering all days of the week, and times of day. Figure 3 shows the CDF of the relative errors for the three methods. We can see that their performance is almost identical. The mean relative error is 11.3%. Furthermore, note that more than 80% of the traffic matrix elements have errors less than 20%. The CDFs for individual data sets are very similar, but generally less smooth. All three algorithms are remarkably fast, delivering the traffic matrix in under six seconds. The fastest algorithm is SVD-IPF, which is about twice as fast as MMI, the slowest one. We also compare the three algorithms for robustness. The results are very similar, and are omitted here in the interest of brevity.

Note also that [5] showed a number of additional performance metrics for the SVD-IPF algorithm (which we can see has very similar performance to the MMI algorithm). Those results indicated that not only are the errors on the flows reasonable, but also that the errors on the largest flows are small, and that the errors are stable over time, which is important if the results are to be used to detect network events.

C. Topological Impact

While traffic data is generally considered highly proprietary, and is therefore hard to obtain from network operators, there has been an effort underway recently to measure ISP topologies via a tool called Rocketfuel [13], [14], [15]. Using the topological information provides us a means of examining the impact of other topologies on our algorithm. In this section, we investigate the impact of different topologies on the performance of the algorithm using Rocketfuel and simulated topologies. Since we also need IGP weights, we use the maps for three North American networks (Sprint, Abovenet, and Exodus), for which the IGP weights have been estimated by Rocketfuel. Note that these are not real weights from the networks of interest, but a set consistent with observed routing. Nor are these the real networks, as Rocketfuel maps are unlikely to be perfectly accurate. Furthermore, the Rocketfuel data do not contain the peering relationships of a network, and so we are limited to using the same initial conditional independence assumptions in our exploration of topology. These issues are not a big problem here because we are primarily concerned with the impact of varying the internal network topology on the estimates, and as such we only need realistic networks, rather than exact maps of other networks. The results should, however, *not* be used in ISP comparisons.

The approach for testing the impact of topology is as follows. We map locations (origins and destination in the original network) to locations (in the Rocketfuel network) at the PoP level, and map (26) and (27) to this new network, assuming the same peering relationships, thus removing dependence on data we don't have access to. More specifically, let $\mathcal{M} : A \to B$ denote a mapping from the original set of locations $i \in A$ to a set of Rocketfuel locations $j \in B$. Then the mapping of demands from one network to another is accomplished by

$$x_j^B = \sum_{i:\mathcal{M}(i)=j} x_i^A, \quad \forall j \in B, \qquad (29)$$

and we map the g_i from (21) similarly. We consider two mappings, the first based on geographical location, which is provided in the Rocketfuel dataset. Geographical information does not provide any way of mapping from router to router in the new network, so we perform our mapping at the PoP level, and therefore also perform the estimation at this level, and compare to AT&T data likewise aggregated to PoP level.

The second mapping is a random permutation that destroys the dependency between the traffic and the network topology.

Figure 4 shows a summary of the results (detailed results can be found in [28]). The figure shows (as squares), the results for the Rocketfuel networks where the mapping from location to location is done on the basis of nearest geographical equivalent, i.e.,

$$\mathcal{M}(i) = j, \text{ where } d(i,j) \leq d(i,k) \; \forall k \in B, \qquad (30)$$

where $d(i,j)$ is the geographic distance between PoPs i and j. The figure also shows PoP level results for AT&T, and two simple simulated networks (a star and a clique with 20 nodes).

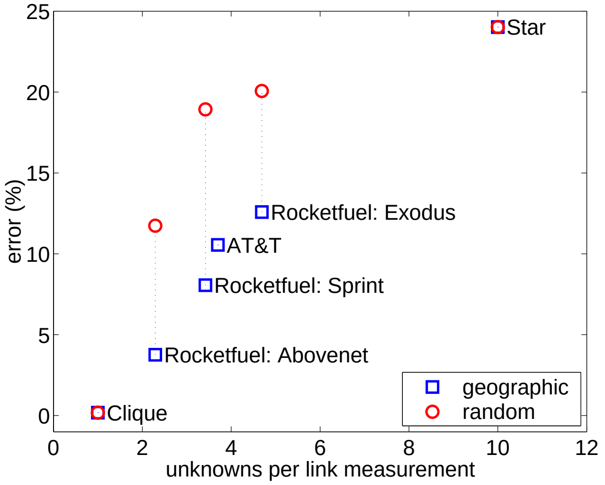


Fig. 4. Results on Rocketfuel, and simulated topologies.

The most obvious thing to note in Figure 4 is that there is a direct correlation between the ratio of number of unknowns to number of measurements, and the accuracy of the results. The star and clique form extreme examples where we either have complete data (in the clique we measure each origin-destination demand directly) and thus almost no measurement error; or almost no additional data (in the star the link measurements tell us no more than the total volumes entering and exiting at a location) and therefore the most inaccurate results. This illustrates the basis for the MMI method. It will work best where either the conditionally independent estimate is good to start with, or the topology has sufficiently diverse links to allow for the results to be accurately refined. The networks measured by Rocketfuel appear to have such diversity, as their results are of similar or better quality than those for AT&T.

However, there is more to the problem than this. In fact it appears that there is a relationship between the network traffic, and the network topology that benefits the performance of the algorithm. Figure 4 also shows the result of mapping the locations in AT&T's network to the Rocketfuel ISPs using a random permutation (the figure is based on 100 random permutations of 24 data sets drawn from one day in June). The performance under a random mapping is worse than under a geographical mapping.

This is interesting because, typically in large networks, regions of the network with higher demand tend to have more connections to the other PoPs (in the measured network the correlation coefficient between node degree and traffic volume was 0.7). A higher degree at a node results in more information about the corresponding row of the traffic matrix, and thence a better estimate of this row. Good estimates of the larger elements make it easier to estimate other elements elsewhere in the network, and so we get a better overall result. This naturally leads to better estimates when the traffic is correlated to the network degree, but when we perform the random mapping, the correlation no longer holds. We shall see later that this property has an impact on the design of network measurement infrastructure to further improve traffic matrix estimates: it is better to put measurement infrastructure in the nodes with the largest traffic volume.

VI. Robustness

A critical requirement for any algorithm that will be applied to real network data is robustness. In general this refers to the sensitivity of an algorithm to violations of the algorithm's assumptions (implicit and explicit). In the MMI method, the assumptions are that the MMI criteria constitute a reasonable approach (verified above) and that the input data are correct. Network data are often error prone, and there can be missing data, and so we must consider how robust the algorithm is to such errors. In the following sections we consider the impact of incorrect or missing link data, and incorrect routing data on the MMI algorithm. Only the latter form of incorrect input data has an important impact on the results of the algorithm.

A. *Incorrect Link Data*

All measurements, including network data, contain errors. Therefore, we shall introduce a range of errors, and study their impact. Comparisons with flow level data have shown that errors in either source are not generally large, and the sources of such errors lead one to believe that they will not be strongly correlated. Hence we shall introduce independent Gaussian errors to the measurements $\mathbf{y}$ and compare with the zero error case. More specifically, take the relative error in the traffic of link i to be $\epsilon_i \sim N(0,\sigma)$, where $N(0,\sigma)$ is the normal distribution with mean 0 and standard deviation σ. We vary σ from 0 to 0.1, with the latter corresponding to quite large (around $\pm$ 20%) relative errors in the measurements (remember the 95th percentiles of the normal distribution lie at $\pm 1.96\sigma$.)

Also note that errors on access and peering links will have minimal impact on a BR to BR traffic matrix because the data from access links is aggregated across many links (to form the traffic volumes entering and exiting the network at a router) and so we only consider here errors in the backbone-link measurements.

Figure 5 shows the CDF of the results given different noise levels. Clearly noise impacts the results, but note that the additional errors in the measurements are actually smaller (for the most part) than the introduced errors in the measurements. This is likely due to the redundant link constraints, which provide an averaging effect to reduce the impact of individual errors. The first row of Table I presents a summary.

B. *Missing Link Data*

We next consider the impact of missing data, for instance missing because a link was not polled over an extended

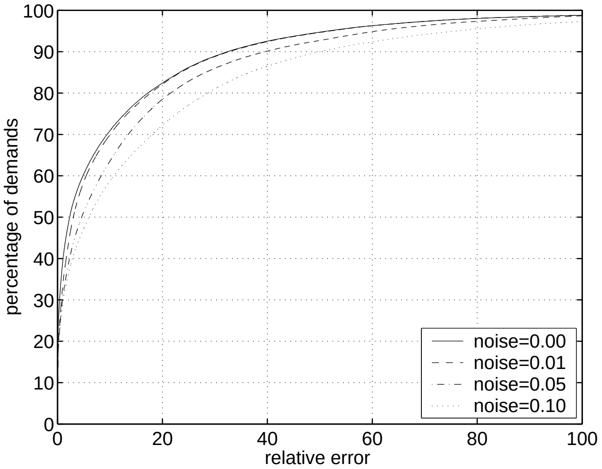


Fig. 5. Relative errors for MMI given measurement noise.

interval. A few missing data points can be replaced using interpolation; trading missing data for data with some error. Furthermore, ERs are typically connected very simply to the backbone (typically by sets of redundant links), and almost all ($> 99\%$) of ER traffic is between the backbone and the edge. Thus if data are missing from a single edge link we may estimate the corresponding traffic using measurements of the traffic between the ER and the backbone. Thus, except in the rare case where we miss multiple edge links, we need only consider missing backbone link data.

Figure 6 shows the effect of missing the top N backbone links (rated in terms of traffic on those links). The results are shown for the 24 data sets from each of three days in June. The results show that despite loosing the links with the largest traffic, the results are hardly impacted at all (except in one case). This suggests that there is often enough redundant information in the network to compensate for the missing data.

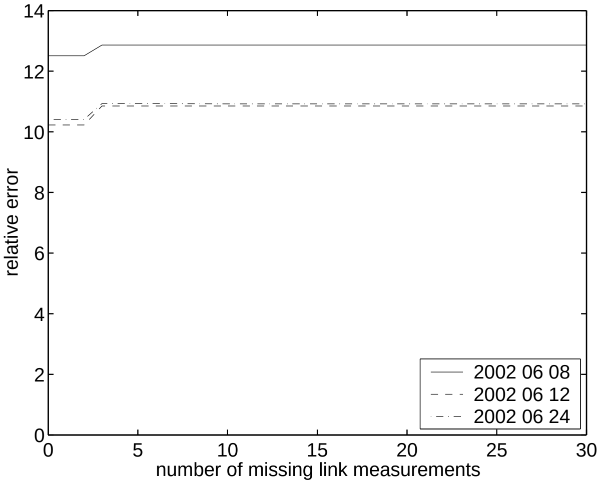


Fig. 6. The impact of missing data on the relative errors for three days (each comprising 24 data sets).

C. Incorrect Routing Data

A third source of data in which we may find errors is the routing matrix. Errors in this matrix can have a large impact on the performance of estimation methods, because if we have errors in a significant number of routes, this corresponds to changing many elements of the matrix from 1 (in the absence of load sharing) to zero and visa versa. However, as in all other reports on traffic matrix estimation, we assume the routing matrix input is accurate. This assumption is reasonable because there are good methods for reliably obtaining routing information (for instance see [18]).

VII. Additional information

One major benefit of adopting the information theoretic approach describe here is that it provides a natural framework for including additional information. In this section, we examine the impact of two sources of information: (i) flow level data at some locations, and (ii) the local traffic matrix at a router [16].

A. Flow Level Data

In this section we consider the impact of having flow level data at some locations, which gives the rows of the traffic matrix for those locations. This inclusion was explored in [4] in a simulation. They showed that the methods of [2], [3] provided improvements to traffic matrix estimates roughly in proportion to the number of rows measured, but that it did not matter whether one selected the rows to be measured randomly, or in order of largest row sum.

Flow level information can be included in our algorithm by simply including additional constraint equations. Results are presented for three separate days of data, each consisting of twenty four, one-hour data sets. We compare the error in the estimates as we include a variable number of known rows of the traffic matrix, both in row sum order, and randomly. Figure 7 shows the results. In the random-ordering case, we see an approximately linear improvement as additional information is included, but in contrast to the results of [4] row sum order is significantly better. In fact, once 10 rows are included, the error for the row sum case is about half that of the random ordered case, and this ratio improves until we have included around half of the rows, when the error for the row sum ordered case becomes negligible. One possible reason why these results do not agree with [4] is that the traffic matrices used in the simulation were not as skewed as those observed in real networks.

The result is a clear win for measuring flow, or packet level data. Such data on a fraction of the network may provide a disproportionate improvement in the estimates. The results were similar even when errors were added to the flow level measurements, and so sampled flows may also provide practical improvements.

B. Local Traffic Matrices

Another appealing alternative to collect additional information with minimal cost is to collect local router traffic matrices. That is, for the router to keep a table of traffic from in-interface to out-interface. As shown in [16], the collection of local traffic matrices only requires minimal changes to router hardware, and can be included in our algorithm as constraints. Figure 8 shows the CDF including local traffic matrices, and Table I shows a summary of the results in comparison to those without local traffic information. Notice that the results with a local

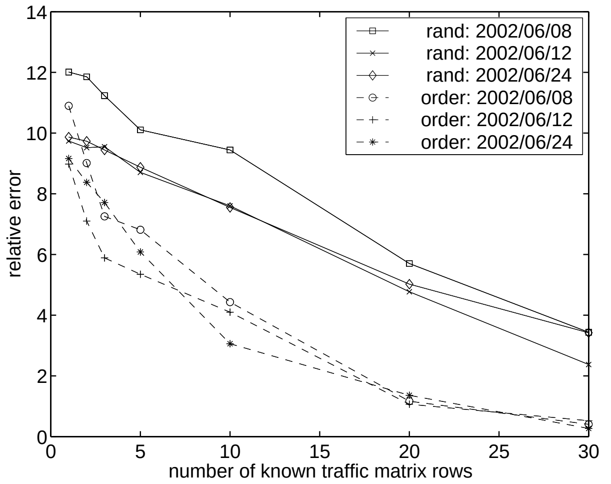


Fig. 7. Effect of addition of known traffic matrix rows. Dashed lines show largest row sum ordering, and solid show random order. There are over 60 rows in the traffic matrix.

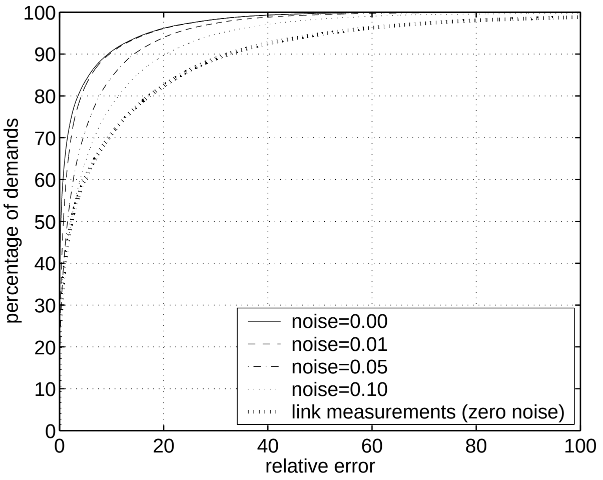


Fig. 8. The result of including local traffic matrices, for varying error levels. Also included as a baseline is the zero noise case from Figure 3.

noise level (σ)	0	0.01	0.05	0.10
w/o local TM	11.26%	11.63%	14.00%	18.01%
with local TM	3.06%	3.40%	5.04%	7.3%

TABLE I

THE RELATIVE ERRORS GIVEN A PARTICULAR NOISE LEVEL, WITH AND WITHOUT LOCAL TRAFFIC MATRIX DATA.

traffic matrix, are not only better, but also less sensitive to measurement errors.

The star topology illustrates why a local traffic matrix helps. In that case, a local traffic matrix at the hub router provides the traffic matrix directly. In reality the network is not a star, so a large amount of additional information is redundant. In our problem, the number of constraints is of the order of a factor of 20 times the simple link measurement constraints, but the number of independent constraints is only roughly doubled. However, this redundant information is still useful because it makes the algorithm more robust to noise in the measurements, as seen in Table I.

These results show that it is quite practical to improve the estimates above by incorporating additional information.

VIII. POINT-TO-MULTIPOINT

Up to this point we have only considered the performance of our algorithm for estimating PTP traffic matrices. We now test the algorithms performance on Point-To-MultiPoint (PTMP) traffic demands. Figure 9 shows an example of PTMP estimation results. In each of the figures we compare the estimated results with their true value – if the estimates were perfectly accurate the points would all line up along the solid diagonal lines. The dashed lines show $\pm 20\%$. Figure 9 (a) shows the results for Point-To-Point (PTP) estimates. One can see that although there are errors, the results are clustered reasonably closely around the diagonal, particularly for the important larger traffic matrix elements. See [5], [28] for much more extensive assessment of the quality of the PTP estimates – for instance, on a large data set the average errors in the PTP estimates were 11.26%, which is well within the bounds for operational usefulness

Figure 9 (b) shows the results of the PTMP estimates, which are still reasonably accurate, but not as good as the PTP estimates. The results are more spread, and there are noticeable outliers well outside the $\pm 20\%$ bounds.

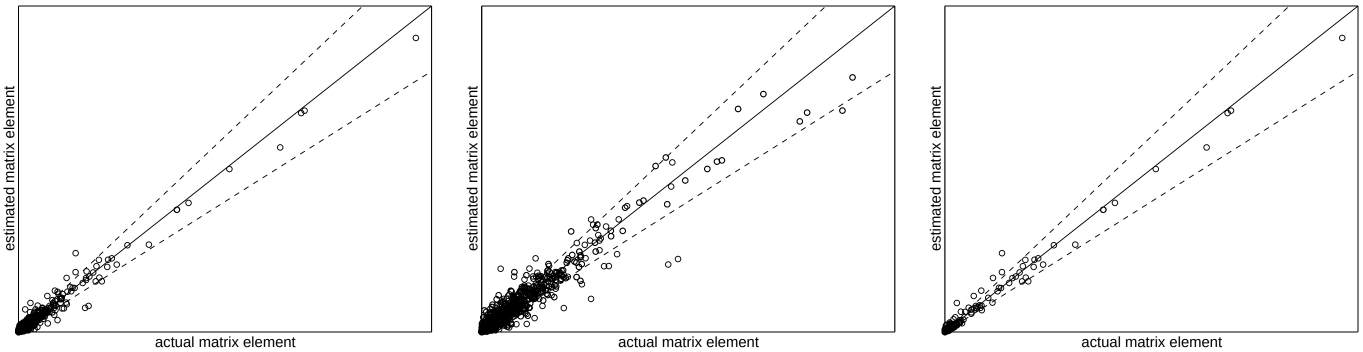
We have tested both PTP and PTMP estimates on considerably larger data sets and these results appear to be consistent. For instance, Figure 10 shows average errors of the results over the course of one day (June 6th 2003) for the PTMP estimates, and compares them to the previously estimated errors for PTP estimates. One sees immediately that the PTMP estimates have errors around the 25% mark – more than twice those of the PTP estimates. We examine the reasons for this worse performance below.

A. Why PTMP estimates are less accurate

In order to understand why PTMP estimates are less accurate, consider that given two different point-to-point demands, the underlying routes are guaranteed to be different, because either the source or the destination is different. For point-to-multipoint demands, however, this is no longer the case. That is, it is possible for different point-to-multipoint demands to use exactly the same route.

Specifically, let $D_1 = A \rightarrow \mathcal{R}_1$ and $D_2 = A \rightarrow \mathcal{R}_2$ be the demands from A to peer P_1 and P_2, respectively, where $\mathcal{R}_i$ are the sets of possible egress routers for peer i $(i = 1, 2)$. Assume that the BGP routing policies are the same for P_1 and P_2. If the router closest to A in $\mathcal{R}_1 \bigcup \mathcal{R}_2$ also belongs to $\mathcal{R}_1 \bigcap \mathcal{R}_2$, then D_1 and D_2 will use the same egress point and therefore the same route. Figure 11 gives such an example where $\mathcal{R}_1 = \{R_1, R_2\}$ and $\mathcal{R}_2 = \{R_2, R_3, R_4\}$, and R_2 is the closest egress point from A within both $\mathcal{R}_1$ and $\mathcal{R}_2$. As a result, both D_1 and D_2 use the route that exits the network at R_2.

Demands with the same routes manifest themselves as equal columns in the routing matrix. This means the constraints on link loads (8) impose restrictions on the sum of such demands instead of each individual demand. As a result, we expect the estimated sum to be accurate. However, the quality of the splitting among different demands largely depends on the conditional independence assumption, which is only an approximation to reality. Therefore, we expect the estimated



(a) PTP.

(b) PTMP.

(c) PTMP aggregated (customer to peering traffic).

Fig. 9. Comparison of the quality of the estimated traffic matrices.

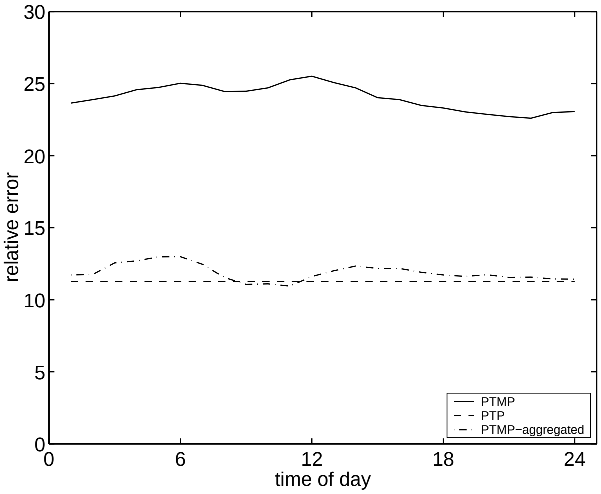


Fig. 10. A comparison of the different methods for estimation over the course of 24 hours. Each data point represents the average errors for a one hour data set. The solid line shows the errors for PTMP estimates, the dashed line those for PTP estimates and the dot-dash line shows those for aggregated PTMP estimates (for customer to peer traffic)).

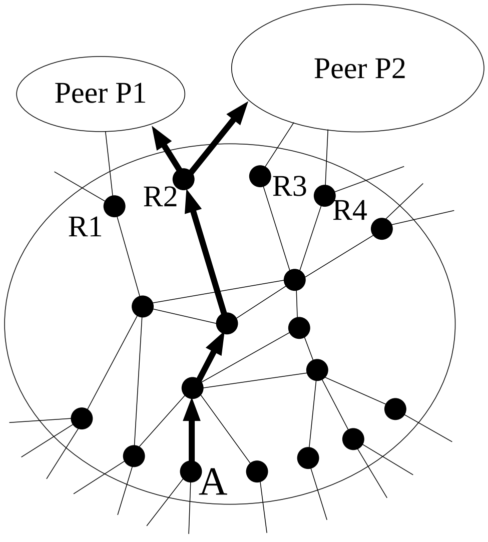


Fig. 11. An example of different point-to-multipoint demands using the same route. There are two point-to-multipoint demands here: the demand from A to peer P_1: $D_1 = A \to \{R_1, R_2\}$, and the demand from A to peer P_2: $D_2 = A \to \{R_2, R_3, R_4\}$. Among $\{R_1, R_2, R_3, R_4\}$, R_2 is the closest egress point from A. As a result, both D_1 and D_2 use the route that exits the network at R_2.

individual demands to be less accurate. Note that PTP estimation problem is also ill-posed and so the problem of ambiguity exists there also, but it is compounded in PTMP case, with a corresponding impact on the performance.

For some applications, it may be possible to avoid such inaccuracy by merging D_1 and D_2 into a single demand and only estimate the sum D_1+D_2. Figure 9 (c) shows an example of such aggregation. The figure shows the point-to-multipoint traffic elements (those from customers to peers) aggregated by summing where we have identical columns in the routing matrix. We can see that the results are once again good — in fact they are very similar to (though not exactly the same as) the PTP results. Figure 10 also shows the errors over the course of one day, and we can see that these errors are now very close to those of the PTP estimation technique.

However, for many applications like reliability analysis, we cannot simply merge them. The two demands D_1 and D_2 may use the same route during normal operations, but not under failure conditions. For example, if router R_2 in Figure 11 fails, D_1 will exit the network at R_1, whereas D_2 will now exit at either R_3 or R_4 depending on which one is closer to A.

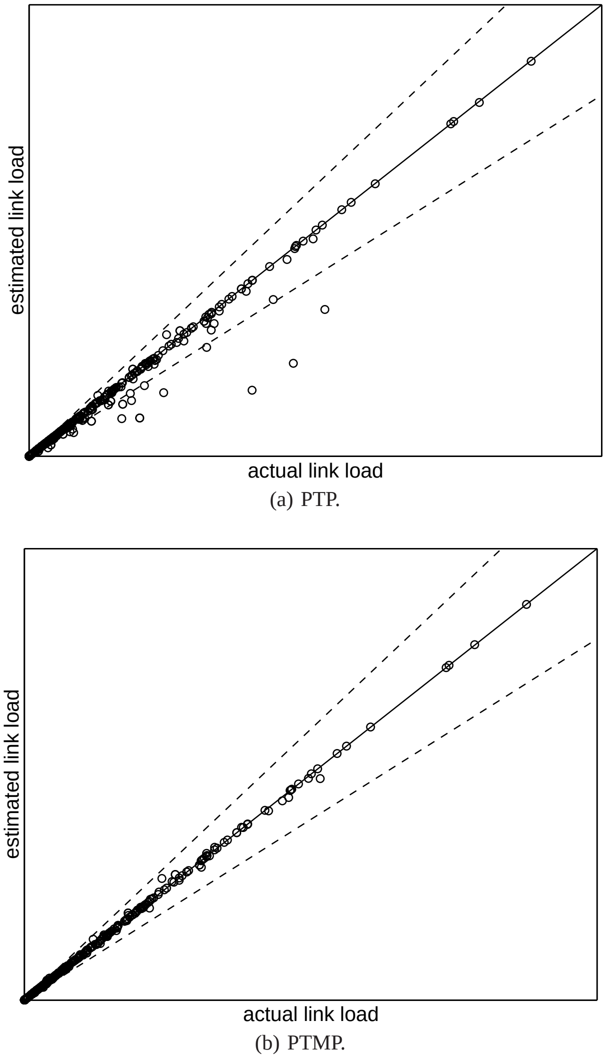
Given that the estimates for individual PTMP demands are less accurate it is tempting to say "the MMI method is intrinsically limited in that it can not accurately distinguish different demands using the same route; we need better information (for instance from NetFlow [7]) in order to perform accurate reliability analysis." Surprisingly, we find in Section IX that the PTMP traffic matrices obtained using MMI work remarkably well for reliability analysis. In addition, they work considerably better than the estimated PTP traffic matrices, which highlights the importance of using PTMP matrices.

IX. Reliability Analysis Results

In [17] the authors found that there is not a simple relationship between error statistics such as those considered above, and the operational usefulness of a set of estimates. In fact the performance of an estimate in route optimization was not even monotonic in the average magnitude of its errors. So here, we shall consider how well the PTMP estimates perform in the task for which we required them in the first place: reliability analysis, in particular of the peering edge of the network.

To do this, we simulate the failure of edge nodes of the AT&T backbone, and consider the resulting loads on links in the network. Under such failures traffic is rerouted, and may result in an overload on the network unless it is carefully planned. We could also simulate single peering link failures, but such failures are a subset of the node failures, and so are less demanding. Under the failure, the PTMP traffic demands will reroute to alternative exit points, as determined by the new shortest IGP distance. However, the PTP traffic to the node in question is simply lost, as we have no way of rerouting this traffic accurately.

In Figure 12 we compare the link loads on the simulated network as produced by the estimated, and true traffic matrix. Figure 12 (a) shows the results for the PTP traffic matrices. The results are reasonable for many links, because many are unaffected, or affected only in a minor way by the routing changes. However, as might be expected, there are a significant number of links for which the loads are underestimated, because the traffic to the failed edge node has been dropped, rather than rerouted. In comparison, Figure 12 (b), the same picture for the PTMP demands, shows remarkably accurate results. This is in direct contrast to the larger value of the average errors in the elements.



(a) PTP.

(b) PTMP.

Fig. 12. Comparison of the reliability analysis results.

To extend these results we consider the failures of 14 different edge nodes: those with the largest number of peering connections. Figure 13 shows these results, and we can see that the errors from the PTMP estimates remain negligible (in practice one typically uses such estimates to predict future traffic, and the prediction errors are generally greater than the estimation errors found here), while the PTP estimates result in average errors up to 8%, but note that in the worst case the PTP estimates result in substantial underestimates of the link loads under the failures (under-estimation is a more serious problem here than over-estimation).

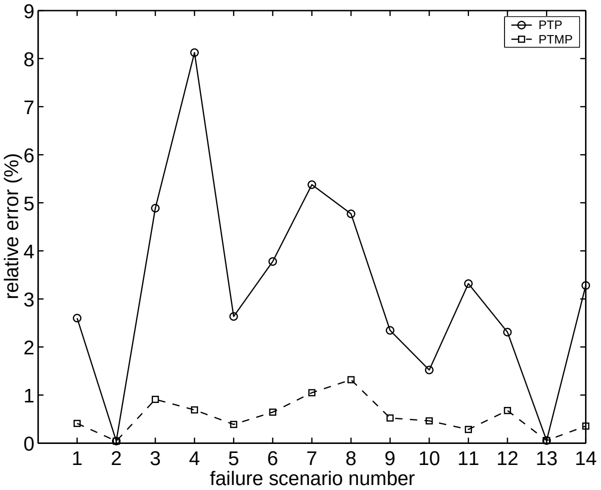


Fig. 13. A comparison between the average errors in the reliability analysis using PTP (solid line), and PTMP estimates (dashed line).

One natural explanation for the quality of the results lies in the insight above that sums of the traffic that use the same routes will be accurately estimated. If such aggregates are simultaneously moved to the same alternate route, then the resulting link loads will be accurately estimated. The results above suggests that it is often the case that such aggregates are shifted to new routes as aggregates, or at least that this is a reasonable approximation. In effect the errors are such that they mostly cancel, when the estimates are used in this way. While such a property is not necessarily guaranteed in general IP networks, it does appear to be the case on AT&T's North American backbone network, and intuition supplied by [29] supports the idea that this is also the case on other networks.

X. Conclusion

To summarize, we present a new approach to traffic matrix estimation for IP networks. We demonstrate on real data that the method has excellent properties: it is fast, accurate, flexible, and robust. In addition, this paper provides some insight into the problem of traffic matrix estimation itself. In particular, by testing the method on Rocketfuel topologies we provide some measure of what aspects of a network make the problem easier or harder: estimates on more highly meshed networks were more accurate. Further, we found that the relationship between the traffic volumes and the topology played a significant role in the accuracy of the estimates. Apart from this, the method also provides additional insight into a broad range of approaches to traffic matrix estimation.

There is still considerable work to do in this area: for instance, the choice of priors is interesting. It is known that a better prior results in a better estimate. While the prior used here seems adequate, one may be able to do better (for instance by using [4]). Other areas of future work include, understanding why the methods are so insensitive to the value of λ, and performing further validations of the method, on alternate data sets (including different traffic patterns).

ACKNOWLEDGMENTS

We would like to thank Michael Saunders for the PDSCO code used here, George Varghese for suggesting the collection of local traffic matrices, anonymous reviewers for their helpful comments, and the people who helped collect some of the data used here, in particular Joel Gottlieb, and Fred True.

REFERENCES

[1] Y. Vardi, "Network tomography: estimating source-destination traffic intensities from link data," *J. Am. Statist. Assoc.*, vol. 91, pp. 365–377, 1996.

[2] C. Tebaldi and M. West, "Bayesian inference on network traffic using link count data," *J. Amer. Statist. Assoc*, vol. 93, no. 442, pp. 557–576, 1998.

[3] J. Cao, D. Davis, S. V. Wiel, and B. Yu, "Time-varying network tomography," *J. Amer. Statist. Assoc*, vol. 95, no. 452, pp. 1063–1075, 2000.

[4] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, "Traffic matrix estimation: Existing techniques and new directions," in *ACM SIGCOMM*, (Pittsburg, USA), August 2002.

[5] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, "Fast accurate computation of large-scale IP traffic matrices from link loads," in *ACM SIGMETRICS*, (San Diego, California), pp. 206–217, June 2003.

[6] A. Medina, C. Fraleigh, N. Taft, S. Bhattacharyya, and C. Diot, "A taxonomy of IP traffic matrices," in *SPIE ITCOM: Scalability and Traffic Control in IP Networks II*, (Boston, USA), August 2002.

[7] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: Methodology and experience," *IEEE/ACM Transactions on Networking*, pp. 265–279, June 2001.

[8] M. Bertero, T. Poggio, and V. Torre., "Ill-posed problems in early vision.," *Proc. of the IEEE*, vol. 76, no. 8, pp. 869–889, 1988.

[9] I. J. Craig and J. C. Brown, *Inverse Problems in Astronomy: A Guide to Inversion Strategies for Remotely Sensed Data*. Adam Hilger, Boston, 1986.

[10] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization.," *SIAM Review*, vol. 40, no. 3, 1998.

[11] R. Parker, *Geophysical Inverse Theory*. Princeton University Press, Princeton, NJ, 1994.

[12] G. Wahba, *Statistical Decision Theory and Related Topics III*, vol. 2, ch. Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine., pp. 383–418. Academic Press, 1982.

[13] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP topologies with Rocketfuel," in *ACM SIGCOMM*, (Pittsburg, USA), August 2002.

[14] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson, "Inferring link weights using end-to-end measurements," in *ACM SIGCOMM Internet Measurement Workshop*, (Marseilles, France), November 2002.

[15] N. Spring, R. Mahajan, D. Wetherall, and H. Hagerstrom, "Rocketfuel: An ISP topology mapping engine." `http://www.cs.washigton.edu/research/networking/rocketfuel/`.

[16] G. Varghese and C. Estan, "The measurement manifesto," in *2nd Workshop on Hot Topics in Networks (HotNets-II)*, 2003.

[17] M. Roughan, M. Thorup, and Y. Zhang, "Traffic engineering with estimated traffic matrices," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, (Miami Beach, FL, USA), pp. 248–258, 2003.

[18] A. Shaikh, C. Isett, A. Greenberg, M. Roughan, and J. Gottlieb, "A case study of OSPF behavior in a large enterprise network," in *ACM SIGCOMM Internet Measurement Workshop*, (Marseille, France), 2002.

[19] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, and J. Rexford, "Netscope: Traffic engineering for IP networks," *IEEE Network Magazine*, pp. 11–19, March/April 2000.

[20] G. Jumarie, *Relative Information*. Springer-Verlag, 1990.

[21] J. Skilling, "The axioms of maximum entropy," in *Maximum-Entropy and Bayesian Methods in Science and Engineering* (G. J. Erickson and C. R. Smith, eds.), vol. Volume 1: Foundations, pp. 173–187, Kluwer Academic Publishers, 1988.

[22] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, 1997.

[23] P. C. Hansen, "Regularization tools (for Matlab)." `http://www.imm.dtu.dk/~pch/Regutools/index.html`.

[24] P. C. Hansen, "Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems," *Numerical Algorithms*, vol. 6, pp. 1–35, 1994.

[25] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[26] J. Cao, S. V. Wiel, B. Yu, and Z. Zhu, "A scalable method for estimating network traffic matrices from link counts." Preprint. Available at `http://stat-www.berkeley.edu/~binyu/publications.html`.

[27] B. Yu, "Maximum pseudo likelihood estimation in network tomography," in *NISS Internet Tomography Technical Day*, (Research Triangle Park), March 28 2003.

[28] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," in *ACM SIGCOMM*, (Karlsruhe, Germany), pp. 301–312, August 2003.

[29] A. Maghbouleh, "How well do I need to know my traffic matrix? applications in capacity planning and traffic engineering," in *INternet TraffIc MATrices Estimation (INTIMATE)*, (Paris, France), June 2003.

APPENDIX I
CONDITIONAL INDEPENDENCE

In Section III-E we present a result based on conditional independence, rather than simple independence. Zero transit traffic makes it is more reasonable to adopt a conditionally independent model in which the probabilities of a packet (bit) arriving at s and departing at d given the class of arrival and destination link (peering or access) are independent:

$$\begin{aligned} &p_{S,D}(s,d|S \in C_s, D \in C_d) = \\ &\qquad p_S(s|S \in C_s, D \in C_d)\ p_D(d|S \in C_s, D \in C_d), \end{aligned} \tag{31}$$

where C_s, and C_d are the source the destination's link class, respectively. We know

$$p_{S,D}(s,d) = p_{S,D}(s,d|S \in C_s, D \in C_d)\ p_{S,D}(C_s, C_d) \tag{32}$$

The source and destination links only depend on the class of the source and destination respectively, so

$$p_S(s|S \in C_s, D \in C_d) = p_S(s|S \in C_s), \tag{33}$$

$$p_D(d|S \in C_s, D \in C_d) = p_D(d|D \in C_d). \tag{34}$$

Furthermore, from the definition of conditional probability

$$p_S(s|S \in C_s) = p_S(s)\ /\ p_S(C_s), \tag{35}$$

$$p_D(d|D \in C_d) = p_D(d)\ /\ p_D(C_d), \tag{36}$$

with the result

$$p_{S,D}(s,d) = \frac{p_S(s)}{p_S(C_s)} \frac{p_D(d)}{p_D(C_d)} p_{S,D}(C_s, C_d) \tag{37}$$

If the class of source and destination were independent, then (37) would result in the independent model $p_{S,D}(s,d) = p_S(s)p_D(d)$. However, noting that all traffic from peering must go to access, and likewise, all traffic to peering links comes from access, and further that the four probabilities must add to one, we get.

$$\begin{aligned} p_{S,D}(P,P) &= 0 \\ p_{S,D}(P,A) &= p(d \in A|s \in P)\ p_S(P) = p_S(P) \\ p_{S,D}(A,P) &= p(s \in A|d \in P)\ p_D(P) = p_D(P) \\ p_{S,D}(A,A) &= 1 - p_S(P) - p_D(P). \end{aligned}$$

Substituting into (37) we get (26).