This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2007 proceedings.

Sensitivity of Traffic Matrix Estimation Techniques to Their Underlying Assumptions

Ilmari Juva
Helsinki University of Technology, Finland
ilmari.juva@tkk.fi

Abstract— **In this paper we study the traffic matrix estimation problem. Based on the nature of additional information that is used to make the problem solvable, we identify two major groups among the estimation methods proposed in literature: The gravity model based methods and the second moment methods. All methods make some assumptions in order to deploy the extra information in the estimation process. We study the sensitivity of the estimation accuracy to these underlying assumptions. The gravity model methods are found to be more accurate when the assumptions hold, but on the other hand their accuracy declines faster than that of the second moment methods when the assumption is not exactly true. In addition, we propose a novel estimation technique which incorporates both sources of extra information. This method is shown in many cases to outperform the current estimation methods relying only on one or the other.**

Keywords: Traffic Matrix Estimation, Gravity model, mean-variance relation

I. Introduction

The traffic matrix gives the volumes of traffic between each origin/destination (OD) pair in the network and is a required input in, for instance, traffic engineering tasks, where the underlying traffic volumes are typically assumed to be known. However, in reality, the traffic matrix is rarely readily available in current IP networks.

Thus, the traffic matrix has to be estimated using information that is available, typically link count measurements $\boldsymbol{y}$ obtained from SNMP measurements and the routing matrix $\boldsymbol{A}$. The basic relationship between $\boldsymbol{y}$ and origin-destination traffic volumes $\boldsymbol{x}$ can be written as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{1}$$

The above equation holds exactly in any given moment in time, and also the equation where we take the expectations of $\boldsymbol{y}$ and $\boldsymbol{x}$ holds. The expected value of $\boldsymbol{x}$ is the traffic matrix $\boldsymbol{\lambda}$, and we get the first moment equation of traffic matrix estimation

$$\overline{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{\lambda}, \tag{2}$$

where $\overline{\boldsymbol{y}}$ denotes the sample mean of the link counts.

Since in any realistic network there are many more OD pairs than links, the problem of solving $\boldsymbol{\lambda}$ from $\boldsymbol{A}$ and $\boldsymbol{y}$ is strongly underdetermined. Thus, explicit solutions cannot be found as there is an infinite number of solutions for $\boldsymbol{\lambda}$ that satisfy equation (2). To overcome this ill-posedness, some additional extra information is needed to solve the problem. Reviews of the proposed methods can be found e.g. in [1] and [2].

The proposed methods use typically either the gravity model or the mean-variance relation to bring the required extra information to make the system identifiable. In this paper we study how sensitive the methods are to the aforementioned underlying assumptions, as well as the number of measurements available. It is found that the gravity model methods are rather sensitive to the gravity assumptions but not at all sensitive to the size of the measurement sample. The methods relying on the mean-variance relation on the other hand are not as sensitive to how well the mean-variance relation holds, but are very sensitive to the sample size.

We propose a novel estimation method combining two sources of extra information and show that in many situations this performs better than methods using only one or the other, thus representing an improvement over current estimation methods.

The rest of the paper is organized as follows. In section II we give a brief overview of the methods proposed in literature and introduce the two methods used later in the paper in the simulation study and propose a novel estimation method combining two sources of extra information. In section III we study the performance of each of these methods when their underlying assumptions do not hold, or hold only to some degree. Section IV concludes the paper.

II. Estimation methods

Based on a comprehensive literature review we found that while several different estimation methods have been proposed, an overwhelming majority of these fall into two main categories with regard to the nature of the extra information utilized to yield an estimate.

1) Gravity model based methods [1], [3], [4], [5].
2) Methods using second moment statistics through a mean-variance relation [6], [7], [8], [9].

The accuracy of these methods depend strongly on the validity of the assumptions, which obviously never hold exactly in real data sets. The first group of methods make only the gravity model assumptions. This is found to be accurate for some networks, but inaccurate for others in [10]. The second group makes the assumption of a functional relation between the mean and the variance of an OD pairs traffic volume. In addition, a traffic distribution, typically Gaussian distribution, has to be assumed to formulate the maximum likelihood equation.

1-4244-0353-7/07/$25.00 ©2007 IEEE

A fundamental difference between the groups is also that gravity methods only require a snapshot of the link counts. If there are several measurements available the link count averages can be used. On the other hand, second moment methods need several measurements of a locally stationary process.

A. Gravity Model Methods

The first group of estimation methods uses the gravity model assumption [3] to gain the extra information. The model is named after Newton's law of gravitation. In the law of gravitation the force between two objects is proportional to the masses of the objects and the inverse of the squared distance between them

$$F \propto \frac{m_1 m_2}{r^2}.$$

In gravity models the quantity to be estimated is proportional to the product of some readily available quantities. Gravity models have been used in social science to model the movement of people or goods between two cities, as well as in telephone networks.

In gravity modelling for data networks the idea is that if we have no knowledge of where a bit is coming or where it is going, the best guess is to make the estimate proportional to traffic volumes sent and received by each node in the network. The traffic between a source node s and a destination node d is assumed to be directly proportional to the product of the total traffic sent by s and the total traffic received by d. This estimate is not, however, always unbiased. (see the Appendix.)

Based on the gravity model it is possible to form a prior estimate. This information is then combined with the link count information to yield the final estimate [4].

1) Information Theoretic Approach: The gravity model estimate must be incorporated with the link count measurements to yield the final estimate. In [5] an information theoretic approach is used. The gravity model is based on independence between origin and destination of the traffic. In information theoretic terms this can be expressed by the mutual information $I(S,D)$ between source and destination addresses, where S and D are random variables with values s and d for a specific source s and destination d.

The mutual information can be expressed in many different ways, but the most useful interpretation for this problem is

$$I(S,D) = K(p(s,d)||p(s)p(d)),$$

where

$$K(f||g) = \sum_i f_i \log\left(\frac{f_i}{g_i}\right)$$

is the Kullback-Leibler divergence, which measures the distance between distributions f and g. So we can write

$$I(S,D) = \sum_{s,d} p(s,d) \log_2 \left(\frac{p(s,d)}{p(s)p(d)}\right)$$

The authors note that a typical way of solving ill-posed linear inverse problems is to solve the regularized minimization problem with a penalty function. In this case

$$\min_{\boldsymbol{x}} ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||_2^2 + \lambda^2 J(\boldsymbol{x}), \tag{3}$$

where λ is a regularization parameter and J is a penalization functional.

The authors use probabilistic terms in their notation of the problem. Total traffic in the network is denoted by N, and the traffic sent from source s to destination d is denoted by $N(s,d)$, where

$$N(s,d) = Np(s,d)$$

and $p(s,d)$ is the probability that a random bit in the network goes from node s to node d. The OD pairs are indexed by i, and the origin and destination of the ith OD pair are denoted by s_i and d_i, respectively. The gravity estimate g_i for the OD pair's traffic is defined as the product of all traffic originating from s_i and all traffic terminating at d_i. Thus the gravity estimate

$$g_i \sim N(s_i)N(d_i),$$

and the traffic matrix $\boldsymbol{x}$ is the actual traffic going from source node s_i to destination node d_i

$$x_i = N(s_i, d_i).$$

In information theoretic terms the independence between source and destination, implied by the gravity model, is equivalent to the mutual information being zero. As the mutual information $I(s_i, d_i)$ is also always positive, it is thus an appropriate penalty function.

$$J(\boldsymbol{x}) = I(s_i, d_i) = \sum_i f_i \log\left(\frac{f_i}{g_i}\right) = \sum_i \frac{x_i}{N} \log\left(\frac{x_i}{g_i}\right).$$

Now equation 3 can be written as

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}} & ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||^2 + \lambda^2 \sum\limits_{i:\ g_i>0} \frac{x_i}{N} \log\left(\frac{x_i}{g_i}\right) \\ \text{subject to} & x_i \geq 0 \end{array} \tag{4}$$

That is, we want a solution that is a tradeoff between satisfying the link count relation and having an a priori plausibility, which here means that the mutual information is small and the solution is thus close to the gravity model. The final solution depends on the selection of λ. The authors use value $\lambda = 0.01$, but demonstrate that the accuracy of the method is not very sensitive to the choice of λ.

B. Second moment methods

The second moment methods use the link count sample covariance matrix as the source of the extra information. When a functional relation

$$\boldsymbol{\Sigma} = \phi \cdot \text{diag}\{\boldsymbol{\lambda}^c\} \tag{5}$$

is assumed between the mean and the variance of the traffic matrix, it is possible to formulate a maximum likelihood problem that becomes identifiable through the use of the sample covariance.

In maximum likelihood estimation method [7] the assumed mean-variance relation is used to write the covariance matrix in the likelihood equation as a function of the mean. Then the system is identifiable and can be solved numerically by the EM algorithm.

But, in fact, the second order statistics for OD-pairs are identifiable based solely on the second order statistics of the link counts. As long as we assume independence among OD-pairs and a sensible routing scheme, we can solve analytically the variances of the OD-pairs by least squares method.

$\boldsymbol{S}^{(x)}$ is the unknown covariance matrix of the OD pair traffic in vector form and $\boldsymbol{S}^{(y)}$ is the covariance matrix of the link counts in vector form. Now we can solve for $\boldsymbol{S}^{(x)}$ using the sample covariance matrix of the link counts.

$$\boldsymbol{S}^{(x)} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{S}^{(y)}, \tag{6}$$

where the matrix $\boldsymbol{B}$ is a function of the routing matrix $\boldsymbol{A}$. (See [9] for further details.)

As the power-law relation between variance and mean is assumed, we can solve the traffic matrix from our variance estimate. This technique was deployed in our Quick method [9].

1) Regularization Quick method: In the Quick method a starting point $\boldsymbol{q}$ is obtained from the variance estimates through the mean variance relation.

$$\boldsymbol{q} = (\phi^{-1}\boldsymbol{S}^{(x)})^{\frac{1}{c}}. \tag{7}$$

Then, the final estimate is obtained taking into account both the starting point and the link count measurements.

The projection from the starting point to the feasible subspace defined by the link counts was done by an analytical expression,

$$\boldsymbol{\lambda} = \boldsymbol{q} + \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{q}). \tag{8}$$

In this paper we use the regularization method in order to bring both methods on the same footing and allow to identify solely the effects that the nature of the extra information has on estimation accuracy.

The optimization problem is now

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||^2 + \lambda^2 \sum_i \frac{x_i}{N} \log\left(\frac{x_i}{q_i}\right), \quad (9) \\ \text{subject to} \quad & x_i \geq 0. \end{aligned}$$

The efficiency of the methods can be compared to the optimal maximum likelihood method by simulation with synthetic data. The sample variances of the estimators can be compared to the Cramér Rao lower bound. The variance/covariance matrix of any unbiased estimator cannot be lower than the inverse of the Fisher information matrix. This is the Cramér-Rao lower bound for the variance of an estimator. Since the maximum likelihood method is asymptotically efficient [11], it has the lowest variance of any estimator. Thus its variance coincides with the bound and we can evaluate any estimator against it by comparing the sample variance of the estimator in question to the bound. (see [12] for a full derivation and discussion of the Cramér-Rao bounds.)

The sample standard deviations of each OD pair is compared to the bound, and the average of these is calculated. The projection Quick method has a standard deviation which is on average 2.02 times larger than the bound, while the regularization Quick method has 2.07 times larger deviation. Thus, they are approximately equally efficient and yield results with about two times larger errors than the MLE. This differs somewhat from the results in [9], where the errors of the projection Quick method were 1.4 to 1.7 times larger than those of the MLE. This is due to the fact that in this simulation study we consider the access links. That is, we have the knowledge of the traffic that enters and leaves the network through each node. The MLE uses this information more effectively than the Quick methods, and thus performs better.

C. Combining both sources of extra information

As stated before, the traffic matrix estimation problem is underconstrained and some extra information has to be brought into the situation to get an unique estimate. The accuracy of this estimate depends on the relevance of the extra information viz. the validity of the assumptions made in order to use the information in the estimation. Above we have reviewed the two common sources of extra information and methods based on them. Current methods utilize one or the other of these information sources. However, as both are relevant information to the problem, it might be a good idea to incorporate both into the estimate.

There are two ways to do this. We can write both starting points into the regularization equation, and optimize them simultaneously. The objective function becomes

$$\min\left\{||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||^2 + \lambda^2 \sum \frac{x_i}{N}\log\left(\frac{x_i}{g_i}\right) + \mu^2 \sum \frac{x_i}{N}\log\left(\frac{x_i}{q_i}\right)\right\}, \tag{10}$$

where $\boldsymbol{g}$ is the gravity model prior and $\boldsymbol{q}$ is the quick method prior.

Another possibility is to just take a componentwise average of the two priors and insert the resulting combined prior into the regularization function. This yields

$$\min\left\{||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||^2 + \lambda^2 \sum \frac{x_i}{N}\log\left(\frac{x_i}{wg_i + (1-w)q_i}\right)\right\} \tag{11}$$

as the objective function.

It turns out that the latter method which is computationally simpler also outperforms the first method. Thus we concentrate on that approach. For the moment we use weight $w = 0.5$. It is left as further work to see whether a weighted average would be a better approach.

III. Simulation Study

A. Simulation methodology

We create synthetic data sets in which the assumptions are true to various degrees, starting from a perfect fit and then making it gradually worse. For any given situation, the goodness of fit value of the mean-variance relation is placed on the vertical axis and the goodness of fit of the gravity model on the horizontal axis of a diagram. Then, at each point of this grid we can make a simulation study to find out which method is more accurate with these particular goodness of fit values for the assumptions.

We use the fictional backbone topology of Figure 1. The generated traffic volumes of the OD pairs follow the gravity model and are denoted by $\boldsymbol{\lambda}_g$. Each OD pair traffic volume is proportional to an assigned dummy variable, or *mass*, of the origin node and the *mass* of the destination node, with the mass variables in this case approximately follow the population of the cities. This is somewhat different from the way the gravity model approach is used in estimation, where the link counts are used directly as the masses. In this case, however, this would lead to an inconsistency making it impossible to create a dataset with different size OD pair traffic, as discussed in the Appendix.

For each simulation we draw identically and independently distributed samples from a Gaussian distribution with parameter vector $\boldsymbol{\lambda}$ for the means of the OD pairs and covariance matrix $\boldsymbol{\Sigma}$ defining the variances of the OD pairs.

To achieve data sets which follow the gravity model assumption and the mean-variance relation only to some degree we add a random component to the parameters $\boldsymbol{\lambda}, \boldsymbol{\Sigma}$.

The deterministic component for the mean is denoted by $\boldsymbol{\lambda}_g$, and follows the gravity model exactly. We add to this a random component

$$\boldsymbol{\epsilon}_\lambda \in (-\lambda_g, \lambda_g).$$

As the traffic matrix $\boldsymbol{\lambda}_g$ follows the gravity model exactly and the error term is random, we can produce synthetic data with a desired goodness of fit value with regard to the gravity model by changing the coefficient $w \in [0, 1]$ which determines the weight of the error term to the final parameter.

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}_g + w\boldsymbol{\epsilon}_\lambda.$$

The variance of the ith OD pair is the element $\boldsymbol{\Sigma}(i, i)$ of the covariance matrix, denoted by $\sigma^2(i)$, or for shorter notation just by σ^2. Again, we have a deterministic component

$$\sigma_m^2 = \phi\lambda^c,$$

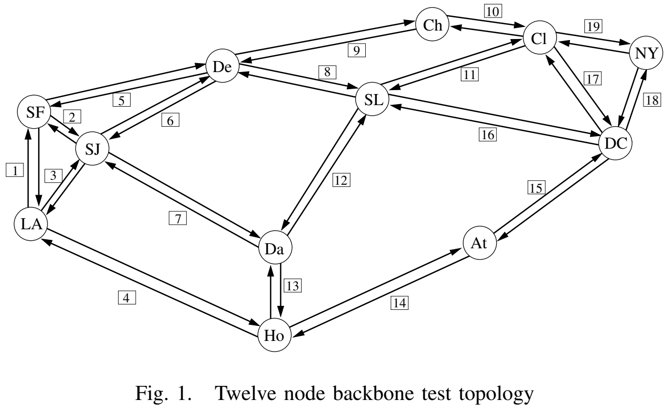

Fig. 1. Twelve node backbone test topology

which follows the mean variance relation by definition, and a random component

$$\epsilon_\sigma \in (-\sigma_m^2, \sigma_m^2).$$

The final parameter value is taken as

$$\sigma^2 = \sigma_m^2 + v\epsilon_\sigma,$$

where $v \in [0, 1]$.

To produce the synthetic data set we draw T samples of traffic counts $\boldsymbol{x}_t$ from a Gaussian distribution with the parameter values obtained above.

$$\boldsymbol{x}_t \sim \mathrm{N}(\boldsymbol{\lambda}, \boldsymbol{\Sigma}).$$

For each parameter value the methods are used to obtain an estimate for each sample. As the most important thing is to estimate accurately the larger OD pairs, we assess the accuracy by considering the mean relative error for the largest OD pairs, comprising 80% of total traffic in the network. This is repeated for one hundred times and the error of the method for the given parameters is taken to be the average of the mean relative errors over the 100 simulations.

1) Calculating the goodness of fit: We need some measure to describe how close the synthetic OD pair means are to the values the gravity model would suggest and how close the synthetic OD pair variances are to the variance suggested by the mean-variance relation.

The distance between the gravity model traffic matrix $\boldsymbol{\lambda}_g$ and the actual traffic matrix $\boldsymbol{\lambda}$ is measured as the error sum of squares over each OD pair i

$$\mathrm{ESS} = \sum_i (\lambda(i) - \lambda_g(i))^2.$$

To make the quantity scale invariant we use the goodness of fit value

$$R^2 = 1 - \frac{\mathrm{ESS}}{\mathrm{TSS}},$$

where TSS is the total sum of squares of the OD pairs

$$\mathrm{TSS} = \sum_i (\lambda_g(i) - \overline{\lambda}_g)^2.$$

If the fit is perfect, then ESS is zero and $R^2 = 1$. If the traffic matrix is totally random, then ESS is approximately same as TSS and $R^2 \approx 0$.

The R^2-value for the mean-variance relation is calculated similarly with

$$\mathrm{ESS} = \sum_i (\sigma^2(i) - \sigma_m^2(i))^2.$$

B. Simulation results

1) Gravity model methods vs. Second moment methods: We use the mean relative error of the largest OD pairs, that comprise 80% of total traffic, as the performance metric. In the sequel we refer to this as *error* for short.

On the left hand side of Figure 2 the mean relative errors of the estimate of equation (4), which uses the gravity model assumption as the extra information, are displayed as

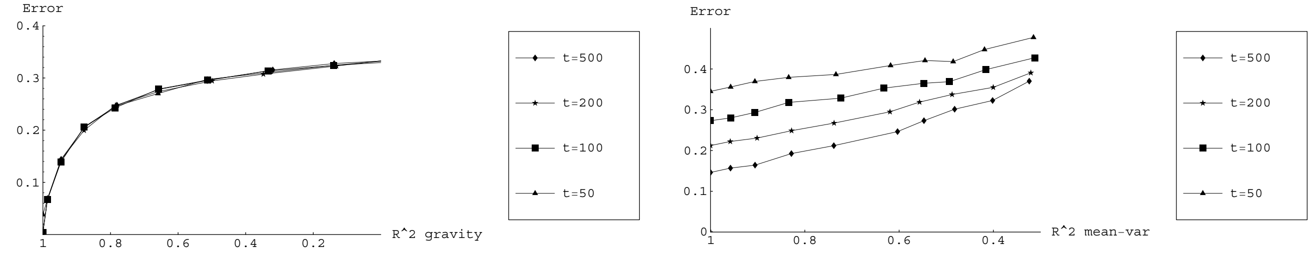

Fig. 2. Effect on estimation accuracy when assumptions goodness of fit deteriorates. Left: Gravity model assumption. Right: Mean-variance relation.

a function of the R^2 value of the goodness of fit of the gravity assumption. The estimates are very accurate when the assumption holds but quickly grow worse when the fit of the gravity model becomes less exact. We can also note that there is no significant difference between the different sample sizes used. This is due to the fact that even a smaller data set is enough to yield accurate estimate for the mean of the link counts, which is the only required input for the model.

On the right hand side of Figure 2 a similar situation is shown regarding the method of equation (9) using the second moment extra information. This time the goodness of fit value on the horizontal axis is for the mean-variance relation. It can be seen that the accuracy of the method depends strongly on the sample size, since it is difficult to get accurate estimates of the sample variance with small sample sizes. The accuracy of the estimator does not deteriorate as quickly as that of the gravity model estimator when the underlying assumption does not hold. On the other hand, even with an exact fit, there are significant errors. As the fit of the mean-variance relation becomes worse, the sample size does not matter as much. This is intuitively clear: if the extra information is not relevant to the problem, having more of it does not help that much.

It is to be expected that if the gravity model holds well, while the mean-variance relation does not, the gravity based methods are more accurate, and vice versa. At some point the methods are equally accurate. If we consider the diagram where the fits of the mean-variance relation and gravity model are set on the axis, we can find a equivalence curve through the area comprised of the points where the comparison between two methods yields a tie.

Figure 3 displays curves that show when the two estimators are equally accurate. On the left side of the figure the gravity assumption holds exactly and on the bottom of the figure the mean-variance relation holds exactly. Thus, on the area from the equivalence curve to the top left corner the gravity model method is more effective and from the curve to the bottom right corner the second moment methods yield better results. As the latter method was dependent on the sample size, also the equivalence curves depend on the sample size. For example, if the mean-variance relation holds, the gravity model assumption needs to have a R^2-value of over 0.83 to be more accurate if there are 500 measurements available. If there are only 100 measurements available a 0.50 goodness of fit for the gravity model makes it the better choice of extra information.

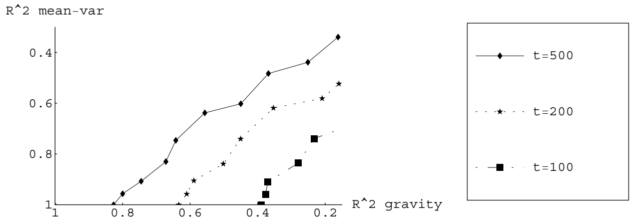

Fig. 3. Equivalence curves showing which R^2 values of gravity model and mean-variance relation yield similar estimation errors.

TABLE I

R^2 VALUES CALCULATED FROM REAL DATA SETS

	Mean-var relation	Gravity model
Abilene	0.76	0.84
Funet	0.83	N/A
Lucent	0.76	0.96

Typical R^2 values for real data sets (Funet[13], Lucent[7] and Abilene[1]) available are listed in Table I. The traffic in Abilene and Lucent networks is considerably more bursty than in internet backbone links and thus the values for the mean-variance relation might not be representative.

We notice that these values fall onto the area in Figure 3 which is to the left of the equivalence curves. Thus, even with sample size of 500 the Gravity model seems to be the better choice for estimating the traffic matrix in these networks. Furthermore, the sample size of 500 measurements might be unrealistically large, considering that we need the measured sample to be locally stationary. Thus the result of our simulations seems to strongly indicate that the gravity model is the better of the two approaches.

2) Accuracy of the Combined method: Figure 4 shows a comparison of three methods. The gravity model method, the second moment method using mean-variance relation to obtain

[1] http://www.cs.utexas.edu/users/yzhang/

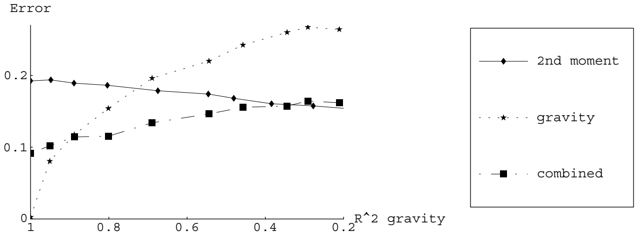

Fig. 4. Comparisons of using the combined prior vs. using one or the other.

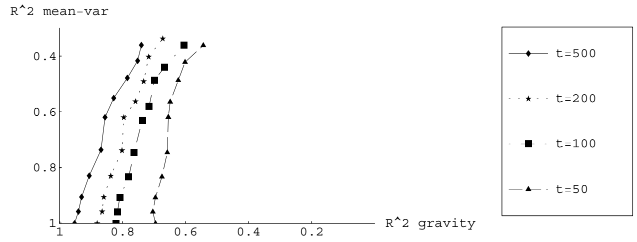

Fig. 5. The equivalence lines for gravity prior vs combined prior

a prior and the combined prior method. We have set sample size to $t = 500$ and the mean-variance relation is set to a realistic level based on Table I, so that $R^2_{meanvar} \approx 0.8$. The method using the gravity model gives the best estimates when the gravity model assumptions hold, but then deteriorates quickly and the combined method becomes the best estimator already when $R^2_{gravity} < 0.9$. The second moment method surpasses the gravity model when $R^2_{gravity} < 0.7$. However, it outperforms the combined method only when the gravity model fit is exceptionally bad, and even then the two are pretty much equally accurate. The second moment method becomes slightly more accurate as the gravity model fit becomes worse. This is most likely a byproduct of the simulation setup: as we create data sets far from the gravity model, the error terms have to be large. Thus, some of the OD pairs are close to zero, and a smaller portion of the OD pairs comprise the 80% of total traffic, and are included in the estimation error.

Calculating results similar to Figure 4 for different values of $R^2_{meanvar}$ we can draw equivalence curves similar to the ones of Figure 3. This time comparing the combined method and the gravity method. The results are depicted in Figure 5. Again, the gravity model is the best one to the left of the equivalence curves and the combined method is the best on the right side of the curves. We notice that the combined method is rather accurate, and the fit of the gravity model assumption needs to be very good, or the fit of the mean-variance relation needs to be bad, to justify using the gravity model as the lone source of extra information.

IV. Conclusion

In this paper we studied the traffic matrix estimation problem, and especially the sensitivity of the two most common methods to their underlying crucial assumptions. We found that the gravity method's accuracy deteriorates more quickly as a function of the gravity model fit, than the second moment method does as a function of the fit of the mean-variance relation. However, when the assumptions hold, the gravity method is significantly more accurate. Also it needs only small sample sizes to achieve good accuracy, while the second moment method needs to estimate the sample covariance matrix, and thus is not very accurate with smaller samples. It would appear that based on our study the gravity based methods are superior to the second moment methods with most realistic sample sizes.

We proposed a novel estimation method, which combines these two competing sources of extra information. Since typically both of them are relevant at least to some extent, as shown by the study of the three real data sets in Table I, it usually makes sense to use both. We showed that in many cases the combined method is the most accurate estimation technique.

Appendix

Gravity modeling is well known in social sciences. It is used, for instance, to estimate the amount of people moving from one city to another. This is done by assuming that the amount is proportional to the populations of the two cities, as well as the distance between them. In economy the trade between nations can be estimated by gravity model by considering it proportional to the GDP of the nations.

The use of the gravity modeling in communication networks is slightly different. There are no easily obtainable natural candidates to play the role of the masses, so the total traffic volumes entering (leaving) the network through a node are appointed to serve as the mass variables. If the traffic from the node to itself is negligible, as in our simulation examples, this leads to an inconsistency, which makes it next to impossible to find a situation where the gravity estimate would be unbiased. Let $N_{in}(s)$ denote the traffic entering the network at node s and $N_{out}(d)$ the traffic going out of the network at node d. The traffic of an OD pair sd is given by

$$x_{sd} = N_{in}(s)\frac{N_{out}(d)}{\sum_{i\neq s} N_{out}(i)}. \tag{12}$$

and also by

$$x_{sd} = \frac{N_{in}(s)}{\sum_{i\neq s} N_{in}(i)} N_{out}(d) \tag{13}$$

The only non-zero solution to these equations is one where each OD pair has the same traffic volume. This is of course extremely restricting, and thus hardly appropriate for real traffic situations. It is also next to impossible to create synthetic traffic in a simulation to be consistent with this gravity model.

We can revise the gravity model to be consistent with gravity modeling in other fields by attaching to each node a *mass* variable so that the OD pair traffic between any two nodes s and d is proportional to the product of these variables instead of the total traffic entering/leaving the network through that node.

$$x_{sd} \propto m_s m_d \tag{14}$$

When solving the traffic matrix, we can then first infer values for the variables m_i using measurements of $N_{in/out}$, and then calculate the traffic volumes according to (14). This estimator would be unbiased. However, the advantage gained in the accuracy due to this unbiasedness concerns only the cases where the gravity model holds in the data set almost perfectly. In any realistic case the difference in the result of equation (4) is negligible.

However, this approach allows us to create synthetic data sets consistent with it, by assigning the mass variables to the nodes, and then calculating the traffic. This is the approach we have used in creating the data sets in our simulation study.

ACKNOWLEDGMENTS

The author would like to thank CSC - the Finnish IT center for science - for providing access to Funet network, and the Statistics Research Department at Bell Labs for providing the Lucent data set.

The work for this paper was carried out as a part of project Fancy funded by Academy of Finland

REFERENCES

[1] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, “Traffic matrix estimation: Existing techniques and new solutions,” in *SIGCOMM’02*, Pittsburg, USA, 2002.

[2] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and C. Diot, “Traffic matrices: Balancing measurements, inference and modeling,” in *SIGMETRICS’05*, Banff, Canada, 2005.

[3] J. Kowalski and B. Warfield, “Modeling traffic demand between nodes in a telecommunications network,” in *ATNAC 95*, 1995.

[4] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, “Fast accurate computation of large-scale ip traffic matrices from link loads,” in *ACM Sigmetrics*, 2003.

[5] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, “An information-theoretic approach to traffic matrix estimation,” in *ACM SIGCOMM*, 2003.

[6] Y. Vardi, “Network tomography: estimating source-destination traffic intensities from link data,” *Journal of the American Statistical Association*, pp. 365–377, 1996.

[7] J. Cao, D. Davis, S. V. Wiel, and B. Yu, “Time-varying network tomography,” *Journal of the American Statistical Association*, vol. 95, pp. 1063–1075, 2000.

[8] G. Liang and B. Yu, “Pseudo likelihood estimation in network tomography,” in *IEEE Infocom*, 2003.

[9] I. Juva, S. Vaton, and J. Virtamo, “Quick traffic matrix estimation based on link count covariances,” in *ICC 2006*, Istanbul, Turkey, 2006.

[10] A. Gunnar, M. Johansson, and T. Telkamp, “Traffic matrix estimation on a large ip backbone – a comparison on real data,” in *IMC’04*, Taormina, Italy, 2004.

[11] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley and Sons, Inc, 1997.

[12] P. Bermolen, S. Vaton, and I. Juva, “Search for optimality in traffic matrix estimation: A rational approach by cramér-rao lower bounds,” in *NGI 2006*, Valencia, Spain, 2006.

[13] R. Susitaival, I. Juva, M. Peuhkuri, and S. Aalto, “Characteristics of od pair traffic in funet,” in *ICN’06*, Mauritius, 2006.