

Simplifying the synthesis of Internet traffic matrices

Matthew Roughan
School of Mathematical Sciences, University of Adelaide
Adelaide, 5005, Australia
matthew.roughan@adelaide.edu.au

ABSTRACT

A recent paper [8] presented methods for several steps along the road to synthesis of realistic traffic matrices. Such synthesis is needed because traffic matrices are a crucial input for testing many new networking algorithms, but traffic matrices themselves are generally kept secret by providers. Furthermore, even given traffic matrices from a real network, it is difficult to realistically adjust these to generate a range of scenarios (for instance for different network sizes). This note is concerned with the first step presented in [8]: generation of a matrix with similar statistics to that of a real traffic matrix. The method applied in [8] is based on fitting a large number of distributions, and finding that the log-normal distribution appears to fit most consistently. Best fits (without some intuitive explanation for the fit) are fraught with problems. How general are the results? How do the distribution parameters relate? This note presents a simpler approach based on a gravity model. Its simplicity provides us with a better understanding of the origins of the results of [8], and this insight is useful, particularly because it allows one to adapt the synthesis process to different scenarios in a more intuitive manner. Additionally, [8] measures the quality of its fit to the distribution's body. This note shows that the tails of the distributions are less heavy than the log-normal distribution (a counterintuitive result for Internet traffic), and that the gravity model replicates these tails more accurately.

Categories and Subject Descriptors:

C.2.5 [Computer Communications]: Local and Wide Area Networks — *Internet*; C.4 [Performance of Systems]: Modeling Techniques.

Keywords:

Traffic Characterization, Traffic Matrices, Topology.

1. INTRODUCTION

Internet Traffic Matrices (TMs), giving traffic volumes from ingress to egress in a network, are the basic input to many network engineering tasks. Much work has gone into measurement [2] or inference [15, 13, 1, 7, 16, 17] of such matrices, with a number of practical outcomes, for instance see [12].

However, as noted in a recent paper [8], there are some cases where one wishes to be able to generate synthetic traffic matrices. Synthesis gives one the control needed to generate many samples with precisely controlled parameters, in order to undertake performance analysis of, for instance, a new traffic engineering or network planning algorithm. It is critical in such synthesis that the important properties of real TMs are accurately represented in the synthetic TMs, otherwise incorrect results may arise. However, the properties of TMs have not been extensively studied as yet, and it is not currently known which are most important.

Thus there is little known on how to accurately synthesize traffic matrices, though a number of simple approaches have been used (for example see [3]). Nucci *et al.* [8] present steps along the road to providing a method. Firstly, they suggest the log-normal distribution to generate new TMs (based on comparisons of the quality of several distribution fits). Secondly, the generated TMs are then related to synthetic or real topologies to provide a complete problem (traffic and topology). The synthetic TMs fit the properties of TMs (considered in the paper), but obviously little can be said (from these results) about how the matrices might fit unobserved properties of traffic matrices. Furthermore, from such a method, it is not particularly easy to see how one should alter the parameters of the model given different underlying structural assumptions about the problems of interest. For instance, the data used in all current works on TMs was derived from North America or Europe. What approaches might be used in Asia or Australia where population demographics are rather different?

Such questions are hard to answer with complete generality, and certainly without more data than we at present possess. However, this note presents a different approach for the first step of synthesizing a TM. The method

1. provides some explanation for the results of [8], and hence an understanding of how general these results are,
2. has only one parameter that need be measured and requires generation of $O(N)$ rather than $O(N^2)$ random variables (for a network of N nodes), and so is simpler,
3. produces TM statistics with better fidelity than [8], in particular in the tails of the TM distribution, and
4. is more easily generalizable than the log-normal approach.

The method is far from new. It is based on gravity models, which have been extensively used in the social sciences, as well as telecommunications, for instance in estimation techniques for traffic matrices [16, 17]. However, in the context of estimation of Internet TMs, gravity models have been shown [16] to have accuracy limitations that might make one think twice about using them for generating synthetic TMs. This paper demonstrates that despite these limitations, the gravity model is quite reasonable for synthesis of TMs.

The method proceeds by randomly generating the edge flows of the gravity model, and then constructing the TM from these flows. The random variables that make up the TM are derived from the product of random variables, and such a distribution is known to converge, in the limit, to a log-normal distribution. Hence, it seems that this approach provides some explanation of results reported in [8]. In addition, gravity models have the advantage that they can be simply related to demographic data, allowing one to match the model to the type of problems one wishes to work, regardless of the context.

2. BACKGROUND AND RELATED WORK

An IP network can be abstractly thought of as a graph, whose nodes are routers, and whose edges are links between these. A Traffic Matrix (TM) describes the volumes of traffic traversing a network from the point at which it enters the network, to the exit point. Such a matrix is useful in capacity planning, traffic engineering, network reliability analysis, and many other network engineering tasks. It is possible to measure such a matrix using measurement technologies such as flow level traffic collection [2], but typically these are hard to implement across a large network [16]. On the other hand SNMP data is easy to collect, and almost ubiquitous. However, SNMP data only provides link load measurements, not TM measurements [16]. The link measurements \mathbf{y} are related to the TM, which is written as a column vector \mathbf{x} , by the relationship $\mathbf{y} = A\mathbf{x}$ where A is called the routing matrix [15]. The resulting problem of inferring the TM from link measurements is a classic underconstrained, linear-inverse problem.

There is extensive experience with ill-posed linear inverse problems from fields as diverse as seismology, astronomy, and medical imaging, all leading to the conclusion that some sort of side information must be brought in. Examples of side information used in the context of Internet TMs are a Poisson model [15, 13], a Gaussian model [1], a logit-choice model [7], or a gravity model [16]. Each method of estimation is sensitive to the accuracy of this side-information. The gravity model assumption was tested in [16, 17], on a large set of traffic data from a tier-1 Internet Service Provider (ISP) in North America (AT&T), where, although it resulted in fast, accurate and robust estimate of the TM, when used as a starting point, it was not found to be accurate enough for TM inference in itself. However these tests were aimed at assessing the model for estimation algorithms, not for synthesis, which is the focus of this paper.

Further effort on modeling the relationships between TM elements has been performed in [6], and successfully exploited for anomaly detection in [5]. These papers focused on Principle Components Analysis (PCA) of the traffic matrices, as times series. PCA exploits the correlations between TM elements to separate the periodic components of the traffic (see [11]), from random fluctuations, and anomalous events. It is not obvious how the structures described within [6] would lead to a simple model for use in synthesis. On the other hand, the gravity model is so simple that it has already been used as a model for network traffic, e.g. see [3].

3. GRAVITY MODELS

Gravity models, taking their name from Newton's law of gravitation, are commonly used by social scientists to model the movement of people, goods or information between geographic areas [14, 10, 9]. In Newton's law of gravitation the force is proportional to the product of the masses of the two objects divided by the distance squared. Similarly, in gravity models for interactions between cities, the relative strength of the interaction might be modeled as proportional to the product of the cities' populations. A general formulation of a gravity model is given by $X_{ij} = \frac{R_i \cdot A_j}{f_{ij}}$, where X_{ij} is the matrix element representing the force from i to j ; R_i represents the *repulsive* factors that are associated with leaving from i ; A_j represents the *attractive* factors that are associated with going to j ; and f_{ij} is a friction factor from i to j .

In network applications, gravity models have been used to model the volume of telephone calls in a network [4]. In the context of Internet TMs, we can naturally interpret X_{ij} as the traffic volume that enters the network at location i and

exits at location j , the repulsion factor R_i as the traffic volume entering the network at location i , and the attractivity factor A_j as the traffic volume exiting at location j . The friction matrix (f_{ij}) encodes the locality information specific to different source-destination pairs, however, as locality is not as large a factor in Internet traffic as in the transport of physical goods, we shall assume a common constant for the friction factors. The resulting gravity model simply states that the traffic exchanged between locations is proportional to the volumes entering and exiting at those locations.

Formally, denote the nodes by n_i , $i = 1, \dots, N$, and the volume of traffic $T(n_i, n_j)$ that enters the network at node n_i and exits at node n_j . Let $T^{\text{in}}(n_i)$ and $T^{\text{out}}(n_j)$ denote the total traffic that enters the network via node n_i , and exits the network via node n_j , respectively. The gravity model can then be computed by either of

$$T(n_i, n_j) = T \frac{T^{\text{in}}(n_i)}{\sum_k T^{\text{in}}(n_k)} \frac{T^{\text{out}}(n_j)}{\sum_k T^{\text{out}}(n_k)} = T p^{\text{in}}(n_i) p^{\text{out}}(n_j) \quad (1)$$

where T is the total traffic across the network, and $p^{\text{in}}(n_i)$ and $p^{\text{out}}(n_j)$ denote the probabilities of traffic entering and exiting the network at nodes i and j respectively. Under the (reasonable) assumption that the network is neither a source nor sink of traffic in itself, so all traffic crosses the network, then $T = \sum_k T^{\text{in}}(n_k) = \sum_k T^{\text{out}}(n_k)$ and we can also write

$$p(n_i, n_j) = p^{\text{in}}(n_i) p^{\text{out}}(n_j) \quad (2)$$

where $p(n_i, n_j)$ is the probability that a packet (or byte) enters the network at node n_i and departs at node n_j . Hence the gravity model corresponds to an assumption of independence between source and destination of the traffic. More importantly, using the above, the gravity model can be written as a matrix formed from the product of two vectors, e.g.

$$P = \mathbf{p}_{\text{in}} \mathbf{p}_{\text{out}}^T \quad (3)$$

so by characterizing these two vectors, we obtain a reasonable characterization of the matrix.

In this paper, we use data derived from the Abilene research network <http://abilene.internet2.edu/> to examine the gravity model (the data from Abilene is publically available, and therefore suitable for the type of comparison we perform here, whereas data from most providers is proprietary). Note that a gravity model *estimates* the Abilene TM (from link data) badly: the average accuracy of the simple gravity model above is $\pm 39\%$. As in previous tests of estimation techniques [16, 17] these results can be considerably improved by using a better initial gravity model (which incorporates natural routing asymmetries), or by regularization techniques. However, in this paper, the goal is not estimation, but synthesis.

In fact synthesis turns out to be quite easy. We start by taking $T^{\text{in}}(n_i)$ and $T^{\text{out}}(n_i)$ to be independent, identically-distributed exponential random variables for $i = 1, \dots, N$. The TM is then generated using (1). This method is extremely simple (an exponential distribution has only one parameter to estimate), and we need generate only $2N$ random variables (as opposed to N^2 in the log-normal approach). Figure 1 shows the distribution for the gravity model derived by simulating 1000 TMs (using the method above) and then plotting the empirical Cumulative Distribution Functions (CDFs), and Complementary Cumulative Distribution Functions (CCDFs). Note that, although the initial fit to an exponential distribution (not shown) is not perfect, the accuracy of the final fit of the gravity model to the TM is

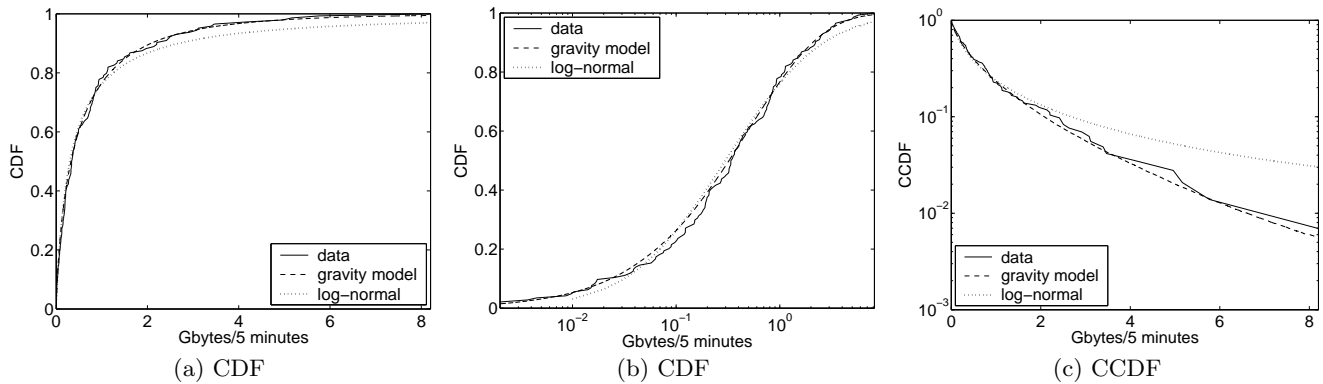


Figure 1: A comparison between the log-normal, and the gravity model fits to the empirical Abilene data (a single 5 minute PoP-PoP TM from 00:15 on the 1st of March, 2004).

excellent. Notice also that the components of the gravity model are independent, where ideally they would be correlated, but that this seems to have little impact on the quality of the results.

4. COMPARISONS AND DISCUSSION

Figure 1 shows a comparison (for one five minute TM) between the CDF and CCDF for the Abilene traffic matrix, and the log-normal and gravity-model TMs. Figure 1(a) and (b) show that all three approaches produce the same distribution with a reasonable degree of fidelity, over the body of the distribution. Nucci *et al.* [8] use two tests to demonstrate the accuracy of fit. The results here are consistent with theirs, so we show only the Kolmogorov-Smirnov (K-S) test. For Figure 1 the K-S statistic values for the log-normal, and gravity-model approaches are 0.059 and 0.047 respectively, consistent with the results of [8]. However, Figure 1(c), shows a previously unnoted feature: the tails of the TM distribution are less heavy than the log-normal distribution¹. The gravity model replicates this tail with much better accuracy than the log-normal fit. We shall provide a quantitative comparison of this feature of the distributions by comparing the relative (absolute) errors in the 99th percentiles.

Figure 2 shows a comparison of the two metrics over a week of Abilene data. As we saw before, the K-S statistic values of the two methods are close (the means over the week for the log-normal, and gravity-model approaches were 0.077 and 0.086 respectively). However, the errors in the 99th percentile were far larger for the log-normal approach (mean 59% vs 10%). Interestingly, the results were similar for aggregated data, e.g. hourly TMs.

There is an additional advantage to the approach proposed here, in that the exponential fit to the data requires only one parameter (the mean). The log-normal distribution requires a mean and variance. In the more complicated schemes of [8] the data is divided into small and large elements to be modelled separately, and this introduces additional parameters that need to be fitted, or adjusted for a particular scenarios (to be modeled). It is certainly easier to adjust one parameter, and so the gravity-model method is simpler to use than the log-normal approach.

A final additional bonus of the gravity-model approach is that it goes some way towards explaining the success of the log-normal fit. Just as the normal distribution arises naturally from the central limit theorem when we sum random

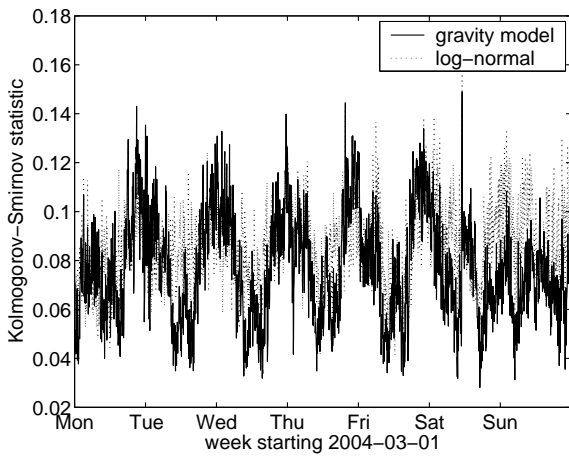
variables, the log-normal distribution arises in the limit as we take the product of random variables. Although the gravity model involves only the product of two random variables, the fact that these are simple exponentially distributed random variables allows that the body (but not tail) of the resulting distribution is reasonably approximated by a log-normal distribution. In addition to now seeing the log-normal fit as approximating the gravity model, we can now also reduce the number of parameters we need to estimate in the log-normal distribution by using the gravity model to predict a relationship between the mean and variance of the log-normal distribution. The gravity model (with exponential random variables) results in the variance of the TM elements being given by $Var[T(n_i, n_j)] = 3E[T(n_i, n_j)]^2$. Note that the results reported in [8] fit this relationship with around a 20% error for the two large/medium flow data sets². The error is much larger for the small flow data set. We do not have access to the raw data from [8] here, however, one might see the consistency in this relationship between the data from the two different networks reported in [8], in conjunction with the approximate log-normal behaviour of the gravity model (which hence matches the distributions for the data sets of [8]) as evidence that the gravity model is a reasonable approach more widely than just for Abilene data. The small flows may require separate handling (as in [8]), but given the gravity model better matches the tails of the distribution, it may also model these more accurately in a joint model.

5. CONCLUSION

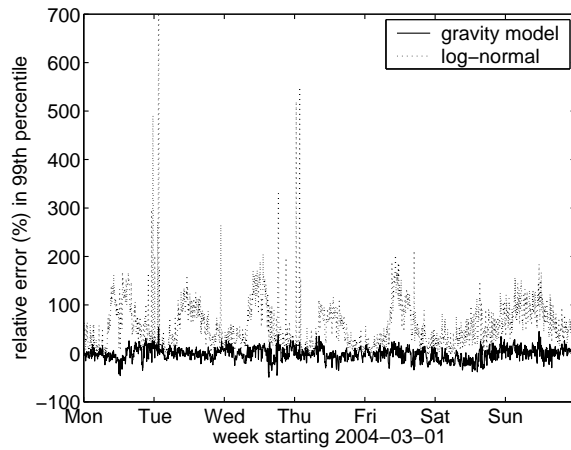
This paper has presented a very simple gravity model for simulating a TM similar to an Abilene TM data. Although the simple gravity model provides poor accuracy for estimation of traffic matrices, it appears to provide quite good approach for synthesis, in particular: it is simple, requires only one parameter, and fits the distributions of TMs well.

There is no question this approach can be improved. One advantage of the gravity model is that there are many natural ways of doing so. Future work will extend the validation of this model, and derive more refined models (for instance by including a distance related friction term). Further, [8] presents methods for producing a time-series of TMs, and matching these series to a network topology, problems not considered here, though, the simplicity of the gravity model is expected to be helpful in this regard as well.

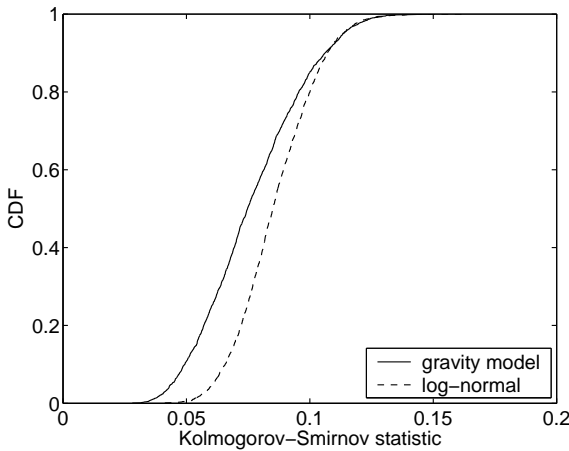
²Errors are to be expected as these results are extrapolated from the reported log-normal fit to the data (not the raw data), and the tail of the log-normal distribution, to which the variance is particularly susceptible, are incorrect.



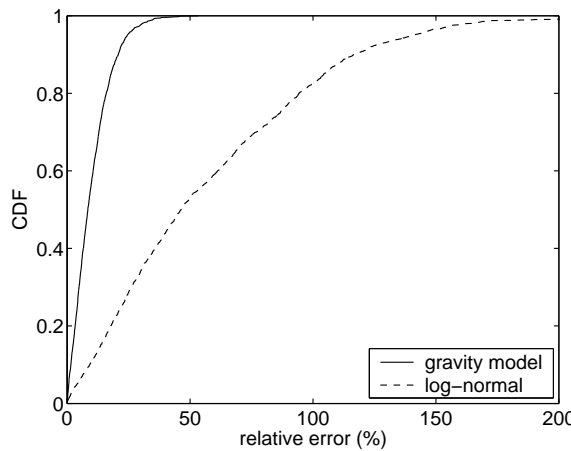
(a) Kolmogorov-Smirnov statistic.



(b) Relative errors in 99th percentile.



(c) CDF of the K-S statistic.



(d) CDF of the relative errors in 99th percentile.

Figure 2: A comparison between the log-normal fit, and the gravity model over one week (March 1st-7th, 2004) of five minute data sets from Abilene. For both metrics smaller values are better.

Acknowledgement

The Abilene data used here was generated by Yin Zhang of the University of Texas, to whom the author is grateful.

6. REFERENCES

- [1] J. Cao, D. Davis, S. V. Wiel, and B. Yu. Time-varying network tomography. *J. Amer. Statist. Assoc.*, 95(452):1063–1075, 2000.
- [2] 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 Trans. on Networking*, pp. 265–279, June 2001.
- [3] B. Fortz, J. Rexford, and M. Thorup. Traffic engineering with traditional IP routing protocols. *IEEE Communications Magazine*, 40(10):118–124, October 2002.
- [4] J. Kowalski and B. Warfield. Modeling traffic demand between nodes in a telecommunications network. In *ATNAC'95*, 1995.
- [5] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM*, 2004.
- [6] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. In *ACM SIGMETRICS / Performance*, 2004.
- [7] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic matrix estimation: Existing techniques and new directions. In *ACM SIGCOMM*, Pittsburg, USA, 2002.
- [8] A. Nucci, A. Sridharan, and N. Taft. The problem of synthetically generating IP traffic matrices: Initial recommendations. *ACM Computer Communication Review*, 35(3), 2005.
- [9] R. B. Potts and R. M. Oliver. *Flows in Transportation Networks*. Academic Press, 1972.
- [10] P. Pyhnen. A tentative model for the volume of trade between countries. *Weltwirtsch. Arch.*, 90:93–100, 1963.
- [11] M. Roughan, A. Greenberg, C. Kalmanek, M. Rumsewicz, J. Yates, and Y. Zhang. Experience in measuring Internet backbone traffic variability: Models, metrics, measurements and meaning. In *ITC-18*, pp. 379–388, Berlin, 2003.
- [12] M. Roughan, M. Thorup, and Y. Zhang. Traffic engineering with estimated traffic matrices. In *ACM SIGCOMM Internet Measurement Conference*, pp. 248–258, 2003.
- [13] C. Tebaldi and M. West. Bayesian inference on network traffic using link count data. *J. Amer. Statist. Assoc.*, 93(442):557–576, 1998.
- [14] J. Tinbergen. *Shaping the world economy: Suggestions for an international economic policy*. The Twentieth Century Fund, 1962.
- [15] Y. Vardi. Network tomography: estimating source-destination traffic intensities from link data. *J. Am. Statist. Assoc.*, 91:365–377, 1996.
- [16] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast accurate computation of large-scale IP traffic matrices from link loads. In *ACM SIGMETRICS*, pages 206–217, San Diego, California, June 2003.
- [17] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An information-theoretic approach to traffic matrix estimation. *ACM SIGCOMM*, pp. 301–312, Karlsruhe, Germany, 2003.