Maximum Entropy Traffic Matrix Synthesis

Paul Tune and Matthew Roughan
School of Mathematical Sciences
The University of Adelaide, Australia
{paul.tune,matthew.roughan}@adelaide.edu.au

ABSTRACT

The traffic matrix (TM) is an important input in traffic engineering and network design. However, the design of current synthesis models of TMs has been rather *ad hoc*, and does not necessarily conform to observed traffic constraints. We apply the *principle of maximum entropy* to develop fast TM synthesis models, with the future goal of developing realistic spatio-temporal TMs.

Categories and Subject Descriptors

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

Keywords

Gravity model; maximum entropy; traffic matrix synthesis

1. INTRODUCTION

The Internet traffic matrix (TM) is an important input in network traffic engineering. Each entry of the matrix contains the traffic volume (typically measured in bytes) from an ingress router to an egress router in a network [6].

Despite their usefulness, there is a lack of work on synthesising these matrices for practical network design optimisation and protocol testing. Often in TM synthesis, there exists observations about the traffic, for *e.g.*, the total ingress and egress traffic [5, 7].

Ideally, a TM synthesis method should satisfy the following criteria:

- *speed*: generation of elements of the model must be fast, scaling well with the dimensions of the matrix,
- *aggregation*: an aggregate of the models should preserve the structure of the original model,
- *model complexity*: the number of parameters controlling the method should be minimal, and
- *conformance*: artificial TMs should conform to observed traffic constraints.

To systematise work on TM synthesis, we propose using the principle of *maximum entropy* [1] as a starting point in deriving models based on the traffic observation. The models we derive satisfy the above four criteria and have a strong theoretical basis.

2. SYNTHESIS MODELS

We first develop spatial models *i.e.,* model of a TM in one measurement bin (about 5 minutes to an hour). We consider the modeling of the Ingress/Egress (IE) TM, as traffic information regarding this TM class is more readily available compared to the Origin/Destination (OD) TM.

Our motivation for adopting the principle of maximum entropy to derive the models is related to the concept of *parsimony*: the model fits the data with the least number of assumptions. In such a case, differential entropy turns out to be a very good measure of complexity. Let $h(f) = -\int_0^\infty \cdots \int_0^\infty f(\mathbf{X}) \log f(\mathbf{X})\ d\mathbf{X}$ denote the differential entropy of a random TM $\boldsymbol{X}$ distributed with density $f(\mathbf{X})$ (note that TMs are nonnegative). Let $\mathcal{C} = \{\mathbf{X}\,|\,\{\phi_\ell(\mathbf{X}) = a_\ell\}_{\ell=0}^{L}\}$ be the convex set of $L+1$ constraints on $\mathbf{X}$, with

$$\phi_0(\mathbf{X}) = \int_0^\infty \cdots \int_0^\infty f(\mathbf{X})\ d\mathbf{X} = 1,$$

since $f(\mathbf{X})$ is a density function. The principle of maximum entropy says that the best model fitting the constraints is the unique solution to the optimization problem

$$\begin{array}{cc} \max\limits_{f(\mathbf{X})} & h(f) \\ \text{s.t.} & f(\mathbf{X}) \geq 0, \mathbf{X} \in \mathcal{C}. \end{array} \tag{1}$$

The constraints we consider here are the first and second order statistics of the following random variables:

- $\boldsymbol{R}$: $N \times 1$ total ingress traffic, *i.e.,* $R_i = \sum_j X_{i,j}$,
- $\boldsymbol{C}$: $N \times 1$ total egress traffic, *i.e.,* $C_j = \sum_i X_{i,j}$, and
- S: total traffic, *i.e.,* $S = \sum_{i,j} X_{i,j}$.

These constraints provide a natural starting point because the traffic can be measured fairly well from SNMP data [6]. Other constraints may also be used, so as long as the constraints are convex to ensure (1) has a global solution.

Table 1 lists the purely spatial maximum entropy models in an increasing number of constraints, labeled from PS1 to PS4. Let $\bar{\mathbf{X}} = \mathbb{E}[\boldsymbol{X}]$. The covariance matrix is defined as

$$\text{Cov}(\boldsymbol{X}) = \mathbb{E}[\text{vec}(\boldsymbol{X} - \bar{\mathbf{X}})\text{vec}(\boldsymbol{X} - \bar{\mathbf{X}})^{\text{T}}], \tag{2}$$

where $\text{vec}(\boldsymbol{X})$ is the vectorization of $\boldsymbol{X}$, *i.e.,* stacking the columns of $\boldsymbol{X}$ on top of another, beginning from the first column. The operation $\mathbf{A} \otimes \mathbf{B}$ defines the Kronecker product

Label	Constraints	Model	Mean	Covariance
PS1	$\mathbb{E}[S] = T$	$X_{i,j} \sim \text{Exp}\left(\frac{N^2}{T}\right), \forall i, j$	$\frac{T}{N^2}\mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}$	$\frac{T^2}{N^4}\mathbf{I}_{N^2}$
PS2	$\mathbb{E}[S] = T,$ $\mathbb{E}[\boldsymbol{R}] = \mathbf{r},$ $\mathbb{E}[\boldsymbol{C}] = \mathbf{c}$	$\boldsymbol{X} = T\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}},$ $U_i \sim \text{Exp}\left(\frac{T}{r_i}\right),$ $V_j \sim \text{Exp}\left(\frac{T}{c_j}\right)$	$\frac{1}{T}\mathbf{r}\mathbf{c}^{\mathrm{T}}$	$\frac{1}{T^2}([\text{diag}(\mathbf{c})]^2 + \mathbf{c}\mathbf{c}^{\mathrm{T}}) \otimes ([\text{diag}(\mathbf{r})]^2 + \mathbf{r}\mathbf{r}^{\mathrm{T}})$ $-\frac{1}{T^2}(\mathbf{c}\mathbf{c}^{\mathrm{T}} \otimes \mathbf{r}\mathbf{r}^{\mathrm{T}})$
PS3	$\mathbb{E}[S] = T,$ $\mathbb{E}[\boldsymbol{R}] = \mathbf{r},$ $\mathbb{E}[\boldsymbol{C}] = \mathbf{c},$ $\mathbb{E}[(R_i - r_i)^2] = \sigma_{r_i}^2, \forall i,$ $\mathbb{E}[(C_j - c_j)^2] = \sigma_{c_j}^2, \forall j$	$\boldsymbol{X} = T\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}},$ $U_i \sim \text{TNorm}\left(\frac{r_i}{T}, \frac{\sigma_{r_i}^2}{T^2}\right),$ $V_j \sim \text{TNorm}\left(\frac{c_j}{T}, \frac{\sigma_{c_j}^2}{T^2}\right)$	$\frac{1}{T}\mathbf{r}\mathbf{c}^{\mathrm{T}}$	$\frac{1}{T^2}(\text{diag}(\boldsymbol{\sigma}_{\mathbf{c}}) + \mathbf{c}\mathbf{c}^{\mathrm{T}}) \otimes (\text{diag}(\boldsymbol{\sigma}_{\mathbf{r}}) + \mathbf{r}\mathbf{r}^{\mathrm{T}})$ $-\frac{1}{T^2}(\mathbf{c}\mathbf{c}^{\mathrm{T}} \otimes \mathbf{r}\mathbf{r}^{\mathrm{T}})$
PS4	$\mathbb{E}[S] = T,$ $\mathbb{E}[\boldsymbol{R}] = \mathbf{r},$ $\mathbb{E}[\boldsymbol{C}] = \mathbf{c},$ $\mathbb{E}[(\boldsymbol{R} - \mathbf{r})(\boldsymbol{R} - \mathbf{r})^{\mathrm{T}}] = \boldsymbol{\Sigma}_{\mathbf{r}},$ $\mathbb{E}[(\boldsymbol{C} - \mathbf{c})(\boldsymbol{C} - \mathbf{c})^{\mathrm{T}}] = \boldsymbol{\Sigma}_{\mathbf{c}}$	$\boldsymbol{X} = T\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}},$ $\boldsymbol{U} \sim \text{TNorm}\left(\frac{\mathbf{r}}{T}, \frac{1}{T^2}\boldsymbol{\Sigma}_{\mathbf{r}}\right),$ $\boldsymbol{V} \sim \text{TNorm}\left(\frac{\mathbf{c}}{T}, \frac{1}{T^2}\boldsymbol{\Sigma}_{\mathbf{c}}\right)$	$\frac{1}{T}\mathbf{r}\mathbf{c}^{\mathrm{T}}$	$\frac{1}{T^2}(\boldsymbol{\Sigma}_{\mathbf{c}} + \mathbf{c}\mathbf{c}^{\mathrm{T}}) \otimes (\boldsymbol{\Sigma}_{\mathbf{r}} + \mathbf{r}\mathbf{r}^{\mathrm{T}})$ $-\frac{1}{T^2}(\mathbf{c}\mathbf{c}^{\mathrm{T}} \otimes \mathbf{r}\mathbf{r}^{\mathrm{T}})$

Table 1: Purely spatial maximum entropy models under various constraints, with corresponding mean and covariance. The models are listed in the order of an increasing number of constraints. Note that the mean is an N by N matrix, while the covariance matrix is N^2 by N^2.

between matrices $\mathbf{A}$ and $\mathbf{B}$. The notation $\text{Exp}(\lambda)$ denotes the exponential distribution with rate λ and $\text{TNorm}(\mu, \Sigma)$ denotes the truncated normal distribution (nonnegative support) with mean μ and covariance Σ.

The mean of these models correspond to their deterministic gravity model counterparts. For instance, for PS1,

$$\mathbb{E}[\boldsymbol{X}] = \frac{T}{N^2}\mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}$$

is precisely the gravity model when there is only a constraint on the total traffic, and for PS2, PS3 and PS4,

$$\mathbb{E}[\boldsymbol{X}] = T\mathbb{E}[\boldsymbol{U}]\mathbb{E}[\boldsymbol{V}^{\mathrm{T}}] = \frac{\mathbf{r}\mathbf{c}^{\mathrm{T}}}{T}, \tag{3}$$

which is precisely the deterministic gravity model under row and column sum constraints [8]. Thus, the models are *stochastic* extensions of the classic gravity model. These models also obey the *independence* property, where the source and destination are independent to each other, and the *aggregation* property, just like the gravity model [6].

Computationally, each of these models require $2N$ random variables to be generated, which is as simple as the method outlined in [5]. Moreover, the number of random variables required scales linearly with N.

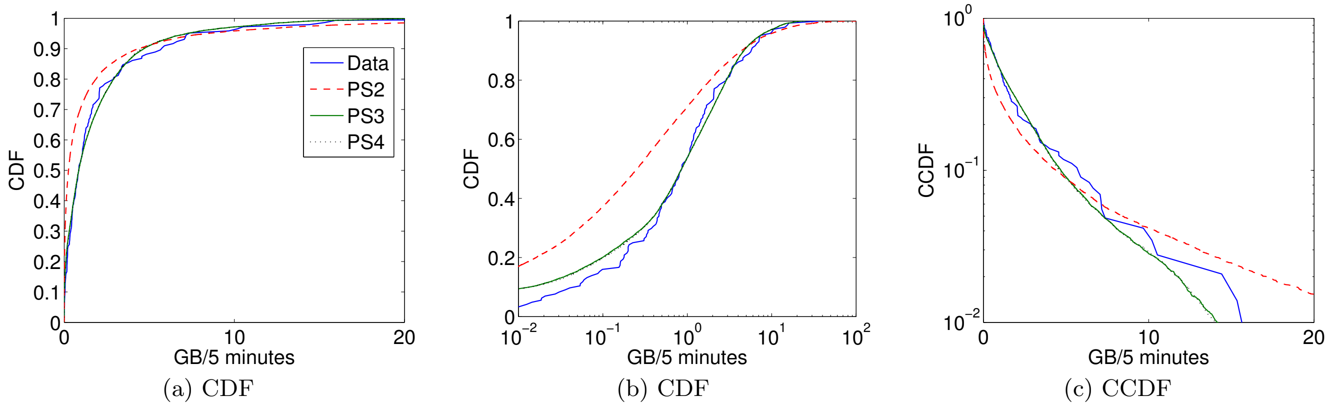
Our models use classical distributions, resulting in efficient generation of these random variables as there already exist efficient algorithms for this task. The truncated normal distribution can be generated via accept-reject sampling *i.e.,* one only needs to first generate normally distributed random variables and choosing only values of U_i and V_j that are nonnegative, or via Gibbs sampling [4]. Similarly, in the generalized case, the covariance matrices $\boldsymbol{\Sigma}_{\mathbf{r}}$ and $\boldsymbol{\Sigma}_{\mathbf{c}}$ can be easily incorporated, simply by generating spatially correlated normally distributed random vectors and using accept-reject sampling to select samples. The sample covariance matrices $\hat{\boldsymbol{\Sigma}}_{\mathbf{r}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{c}}$ can be estimated from data (though this potentially conflicts with our stance, since there is a degree of inaccuracy in any estimate).

2.1 Data fitting

Figure 1 shows an example of three purely spatial maximum entropy models (PS2, PS3 and PS4 from Table 1) fitted on a single 5 minute PoP–PoP TM taken from Abilene at 0140 to 0145 on March 1st, 2004 [2]. The plots present the cumulative and complementary cumulative distribution functions (CDF and CCDF) of the flow volume distributions of the test TM.

We assume that we have measurements of the row, column and total sums. The row and column sums correspond to measurement of the total incoming traffic of ingress nodes and total outgoing traffic of the egress nodes. These measurements are not perfect (due to sampling errors), and spatial correlations between measurements may be present.

All models provide a fairly good fit of the distribution of the empirical Abilene traffic matrix, as seen in Figure 1(a). However, we see here that assuming the measurements are perfect is not enough. PS2's fit is less accurate compared to the other two models. Figure 1(b) and (c) highlight the discrepancy between the fit and the underlying flow distri-

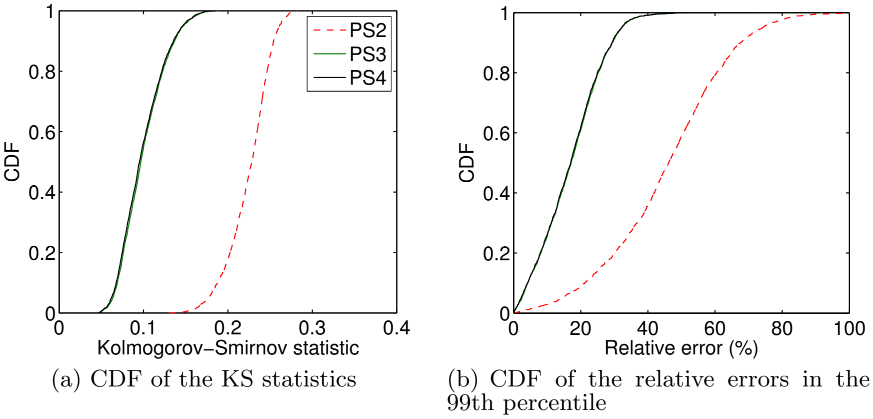


(a) CDF

(b) CDF

(c) CCDF

Figure 1: *Comparison between several purely spatial maximum entropy models fits to empirical Abilene data: the exponential, truncated normal and generalized truncated normal. The real TM is a single 5 minute PoP–PoP TM from 0140 on March 1st, 2004. Note the logarithm scale on the x-axis of (b) and the y-axis of (c).*



(a) CDF of the KS statistics

(b) CDF of the relative errors in the 99th percentile

Figure 2: *Comparison of the CDFs of the KS statistics against the relative errors in the 99th percentile of the models over a week of data on the Abilene network, beginning March 1st, 2004.*

bution for the smaller and larger flows respectively.

All three models do not fit small flows well. The same observation was reported by [3, 5], where small flows were handled differently from medium and large sized flows.

Figure 2 presents the CDF of the Kolmogorov-Smirnov (KS) statistics and the relative errors in the 99th percentile of all three models computed over one week of data on Abilene, beginning from March 1st, 2004. For each traffic matrix lasting 5 minutes (duration of a single measurement bin) over the entire measurement interval, we generated 1000 instances for each of the three models and then compared their traffic flow size CDFs to actual traffic flow size CDFs of Abilene over the period. Using these CDFs, we compute each of their KS statistics and relative error. Interestingly, we varied the measurement bin size from 5 minutes to an hour but observed little change in the results. A similar observation was noted in [5]. We found that although both PS3 and PS4 clearly outperformed PS2, PS4 only marginally outperformed PS3. This is almost indistinguishable in the figure, with a zoom-in only providing the required resolution to verify the outperformance.

Future work will extend maximum entropy to spatio-temporal ensembles of TMs.

3. REFERENCES

[1] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[2] NLANR. Abilene Trace Data. http://pma.nlanr.net/Special/ipls3.html.

[3] A. Nucci, A. Sridharan, and N. Taft. The problem of synthetically generating IP traffic matrices: Initial recommendations. *SIGCOMM Comput. Commun. Rev.*, 35:19–32, July 2005.

[4] C. P. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125, 1995.

[5] M. Roughan. Simplifying the synthesis of Internet traffic matrices. *SIGCOMM Comput. Commun. Rev.*, 35(5):93–96, 2005.

[6] P. Tune and M. Roughan. Internet traffic matrices: A primer. In H. Haddadi and O. Bonaventure, editors, *Recent Advances in Networking, Vol. 1*. ACM SIGCOMM, August 2013.

[7] P. Tune and M. Roughan. Network design sensitivity analysis. In *ACM SIGMETRICS 2014*, June 2014. To appear.

[8] 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*, pages 206–217, 2003.