Applied Economics Letters, 2015
Vol. 22, No. 1, 1–6, http://dx.doi.org/10.1080/13504851.2013.866199

Routledge
Taylor & Francis Group

An information theoretic approach to network tomography

Wendy K. Tam Cho[a,*] and George Judge[b]

[a]*Department of Political Science and Statistics and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[b]*Resource Economics, UC Berkeley, Berkeley, CA, USA*

In this article, we formulate an information theoretic approach to information recovery for a network flow transportation problem as an ill-posed inverse problem and use nonparametric information theoretic methods to recover the unknown adaptive-intelligent behaviour traffic flows. We indicate how, in general, information theoretic methods may provide a solution to the ill-posed inverse information flow problems, when a function must be inferred from insufficient sample information. As an application, we examine a data set which comprised traffic volumes at Bell Labs.

Keywords: inverse problem; information theoretic methods; Cressie–Read divergence; network tomography; link measurements

JEL Classification: C19; C00

I. Network Tomography

In a communication network, the efficiency of information flow in a network is predicated on designing protocols that efficiently identifies adaptive-intelligent behaviour (Wissner-Gross and Freer, 2013) and routes information. In a transportation network, the principles are no different and the emphasis is on design and efficiency in routing. Roads that are heavily used should be designed differently than less-often used roads, and traffic should be routed according to this design in infrastructure. In this article, we use information theoretic methods to analyse the problem of identifying the connection between adaptive behaviour and the entropy maximization principle and determining point-to-point traffic between subnetworks when only aggregate traffic volumes are known. This is a common problem because while point-to-point traffic information can be collected, in practice, doing so is sufficiently burdensome that such collection is not routine or typical. Instead, the usual data collection includes only aggregate information on traffic volumes.

The problem of estimating traffic volumes from aggregate link traffic measurements was first discussed in this journal by Vardi (1996). He used the term network tomography to describe a class of statistical inverse problems, and Castro *et al.* (2004) gave a useful discussion of the statistical implications of this new field. In a recent issue of this journal, Airoldi and Blocker (2013) considered this type of an ill-posed inverse problem and suggested a number of statistical models for information

*Corresponding author. E-mail: wendycho@illinois.edu

recovery. Building on this productive work, to avoid calibration, tuning parameters, regularization, pseudo likelihoods and two-stage inference methods, we propose an alternative structure of the direct formulation using information theoretic methods (see, for example, Golan *et al.*, 1996; Cho and Judge, 2008), where the information theoretic entropy solution to the inverse problem is described by the distributions of the unknown parameters.

II. The Network Inverse Problem

Point-to-point traffic volumes at any point in time, x_t, usually must be estimated from aggregate noisy traffic volumes, y_t, given information about the network routing protocol in the form of a matrix $\mathbf{A}$. The number of origin to destination routes in x is much larger than the number of aggregate traffic measures in y, and the estimation problem results in a series of noisy ill-posed linear inverse problems, $y_t = \mathbf{A}x_t + e_t$, where e_t is a noise component (see Castro *et al.*, 2004).

The transportation network problem is simple to describe. We have aggregate traffic volumes measured at T points in time, y_{it}, where $t = 1, \ldots, T$ and i denotes the subnetwork. The routing protocol matrix, $\mathbf{A}$, encompasses the routing protocol, and $\mathbf{A}_{ij} = 1$ if the point-to-point traffic at subnetwork i contributes to counter j; $\mathbf{A}_{ij} = 0$ otherwise. The network may be expressed by the relationship,

$$y_{it} = \sum_{j=1}^{n} \mathbf{A}_{ij} x_{jt} \tag{1}$$

where at any one point in time,

$$y_i = \sum_{j=1}^{n} \mathbf{A}_{ij} x_j \tag{2}$$

If we let $N = \sum_{j=1}^{n} x_j$ be the total traffic at each subnetwork, then

$$\frac{y_i}{N} = \frac{1}{N} \sum_{j=1}^{n} \mathbf{A}_{ij} x_j \tag{3}$$

If we now let $r_i = \dfrac{y_i}{N}$ and $p_j = \dfrac{x_j}{N}$, then

$$r_i = \sum_{j=1}^{n} \mathbf{A}_{ij} p_j \tag{4}$$

where $\sum_{j=1}^{n} p_j = 1$, and x_j and N are unknown. Note that we need to compute N in some way that does not entail knowing the values of x. In our application later, a peculiarity of the data will enable us to compute N from the following known information in y.

$$r_i = \frac{y_i}{N} \qquad i = 1, 2, \ldots, m \tag{5}$$

$$p_j = \frac{x_j}{N} \qquad j = 1,\ 2, \ldots, n \tag{6}$$

At any point in time, there are $m + 1$ constraints in this problem. There is an additivity constraint,

$$\sum_{j=1}^{n} p_j = 1 \tag{7}$$

as well as m other constraints,

$$r_i = \sum_{j=1}^{n} \mathbf{A}_{ij} p_j \qquad i = 1, 2, \ldots, m \tag{8}$$

This general problem captures a frequently occurring problem where a function must be inferred from insufficient information that specifies only a feasible set of functions, or solutions. The problem is fundamentally underdetermined and indeterminate because there are more unknowns than data points on which to base a solution. Thus, insufficient sample information exists to solve the problem using traditional rules of logic.

III. An Information Theoretic Estimation and Inference Base

To implement the model, we must determine how to represent the linkage to the data and how to choose the criterion or objective function. Because of the ill-posed nature of the inverse problem, traditional forward estimation methods cannot be used to recover the unknown p_j. Given the connection between adaptive behaviour and entropy maximization, to avoid adding creative assumptions and extraneous information that the researcher usually does not possess, we make use of the information theory methods that are designed to handle problems of this nature (Wissner-Gross and Freer, 2013). In this context the Cressie–Read (CR) (1984, 1988) family of

likelihoods-entropy functionals provides one basis for linking the data to the unknown model parameters and exploiting the statistical machinery of information theory.

The CR family of power divergence measures

Since we are treating this as an ill-posed inverse problem with noise and the unknown p_j are discrete random variables, we begin with the CR (Cressie and Read, 1984; Read and Cressie, 1988) multi-parametric family of goodness of fit-power divergence measures,

$$I(\boldsymbol{p}, \boldsymbol{q}, \gamma) = \frac{1}{\gamma(\gamma+1)} \sum_{i=1}^{n} p_i \left[\left(\frac{p_i}{q_i} \right)^{\gamma} - 1 \right] \tag{9}$$

In Equation 9, γ is a parameter that indexes members of the CR family, p_is represent the subject probabilities, and the q_is are interpreted as reference probabilities. The usual probability distribution characteristics, $p_i, q_i \in [0,1], \forall i$, $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} q_i = 1$, are assumed to hold. The p_js are random vectors with an unknown underlying parameterized distribution, and are assumed to be independent of each other.

The CR family of power divergences is defined through a class of additive convex functions and the CR power divergence measure encompasses a broad family of test statistics and leads to *a broad family of likelihood functions* within a moments-based estimation context. In the context of extremum metrics, maximum likelihood is embedded in the general CR (1984) family of power divergence statistics. Thus, this family represents a flexible set of pseudo-distance measures from which to derive empirical probabilities associated with the indirect data. As γ varies, the resulting CR family of estimators that minimize power divergence exhibit qualitatively different sampling behaviour. This class of estimation procedures is referred to as *Minimum Power Divergence* (MPD) estimation (Gorban *et al.*, 2010; Judge and Mittelhammer, 2011, 2012).

If in Equation 9, we let $\gamma \to 0$, and the reference distribution, q_i, be the uniform distribution, the CR distance measure yields the Shannon (1948, 1949)/ Jaynes (1957) entropy criterion,

$$\begin{aligned} -H(\mathbf{Y}) &= p_1 \log p_1 + \cdots + p_n \log p_n \\ &= -\sum_j p_j \ln (p_j) \end{aligned} \tag{10}$$

Under the Shannon and Jaynes maximum entropy estimation criterion, the pure inverse model may be formulated as

$$\arg\min_{p_j} \sum_{j=1}^{n} p_j \ln (p_j) \tag{11}$$

subject to the relevant problem constraints. In this way, the problem is stated as a constrained minimization problem that minimizes the distance between the estimated p_i and $q_i = n^{-1}$, a uniform reference distribution. Depending on the degree of external knowledge base, other fixed or random q_i may serve as the reference distribution.

The resulting Lagrangian function for the constrained maximization problem is

$$\begin{aligned} L = &-\sum_{j=1}^{n} p_j \ln p_j - (\lambda_0 - 1) \sum_{j=1}^{n} (p_j - 1) \\ &- \sum_{i=1}^{m} \lambda_i \sum_{j=1}^{n} (\mathbf{A}_{ij} p_j - r_i) \end{aligned} \tag{12}$$

where the Lagrange multipliers are $\lambda_0 - 1$, $\lambda_1, \ldots, \lambda_m$. Note that we use $\lambda_0 - 1$ instead of λ_0 for mathematical convenience. Taking the derivative yields a solution for the probabilities, p_j, in terms of the Lagrange multipliers.

$$\frac{\partial L}{\partial p_j} = -\ln p_j - \lambda_0 - \sum_{i=1}^{m} \lambda_i \mathbf{A}_{ij} = 0 \tag{13}$$

$$-\ln p_j = \lambda_0 + \sum_{i=1}^{m} \lambda_i \mathbf{A}_{ij} \tag{14}$$

$$\ln p_j = -\lambda_0 - \sum_{i=1}^{m} \lambda_i \mathbf{A}_{ij} \tag{15}$$

$$p_j = \exp\left(-\lambda_0 - \sum_{i=1}^{m} \lambda_i \mathbf{A}_{ij} \right) \qquad j = 1, 2, \ldots, n \tag{16}$$

We substitute p_j from Equation 16 into constraints (7) and (8), to determine $\lambda_0, \lambda_1, \ldots, \lambda_m$. We also substitute p_j into Equation 12 to eliminate the constraints and produce a strictly convex function to maximize

$$L = \ln\left(\sum_{j=1}^{n} \exp\left(-\sum_{i=1}^{m} \lambda_i \, \mathbf{A}_{ij}\right)\right) + \sum_{i=1}^{m} \lambda_i r_i \tag{17}$$

In general, this solution does not have a closed-form expression, and the optimal values of the unknown parameters must be numerically determined.

IV. An Empirical Application

Given the general transportation network problem formulated in Section II, as an empirical application, we examine a data set that comprised traffic volumes at Bell Labs (Cao *et al.*, 2000). The router and subnetworks set-up is depicted in Fig. 1. Aggregate traffic volumes are measured every 5 min over the course of 1 day on the Bell Labs network. In all, we have 287 sets of measurements in time.

These data include seven independent measures of aggregate adaptive behaviour traffic volume, which results in seven constraints in the form of

$$\sum_{j=1}^{7} p_j = 1 \tag{18}$$

and an additivity constraint,

$$r_i = \sum_{j=1}^{7} \mathbf{A}_{ij} p_j \qquad i = 1, 2, \ldots, 16 \tag{19}$$

where

$$r_i = \frac{y_i}{N} \qquad i = 1, 2, \ldots, 16 \tag{20}$$

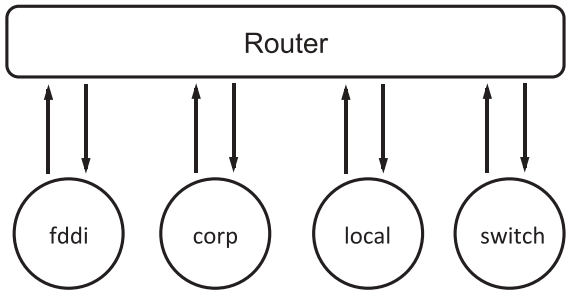

Fig. 1. Bell Labs network

$$p_j = \frac{x_j}{N} \qquad j = 1, 2, \ldots, 7 \tag{21}$$

The following 7×16 matrix encodes the routing protocol:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{22}$$

The resulting Lagrangian for this particular problem is

$$L = \ln\left(\sum_{j=1}^{16} \exp\left(-\sum_{i=1}^{7} \lambda_i \, \mathbf{A}_{ij}\right)\right) + \sum_{i=1}^{7} \lambda_i r_i \tag{23}$$

Empirical results

Using a noninformative uniform prior, the solution to this problem in each of 287 separate measurements in time has a small mean error across the 16 unknown point-to-point traffic volumes, x. The mean error across the 16 measures is essentially zero (to at least 10 decimals places of accuracy). This result is not surprising given that it is incorporated into the constraints for the problem. The individual point-to-point traffic volumes are not as close.

We obtain our estimates with a noninformative prior. However, we can see from the summary statistics of the variables that we are trying to estimate that some of the origin-destination pairs lead to adaptive behaviour that shoulders a large portion of the overall traffic burden while others do not. The first origin–destination pair, for example, has no traffic on any of the 287 measurements in time. The last origin–destination pair has almost no traffic. On the other hand, the 3rd, 6th, 7th, 8th, 9th and 14th origin–destination pairs shoulder quite a bit of traffic.

Minimum	0.0	42.74	7359.0	0.0	28.67	1636.0	3460.0	28 840.0
1st quartile	0.0	63.53	15 490.0	0.2828	60.07	2785.0	5728.0	36 670.0
Median	0.0	78.61	25 120.0	0.2828	78.07	3389.0	10 550.0	44 260.0
Mean	0.0	97.22	28 330.0	0.3057	382.80	4736.0	44 380.0	62 740.0
3rd quartile	0.0	96.45	35 550.0	0.2828	623.20	5089.0	29 490.0	58 740.0
Maximum	0.0	1132.00	188 700.0	1.9870	4528.00	43 300.0	298 200.0	416 900.0
Minimum	3730.0	463.6	4.916	1.414	0.0	3759.0	2.222	0.0
1st quartile	8892.0	985.3	5.185	3.367	0.5657	6683.0	5.131	0.0
Median	15 430.0	1627.0	10.890	6.761	0.5657	13 110.0	25.89	0.0
Mean	64 250.0	2784.0	54.860	293.0	0.5957	44 470.0	225.9	45.69
3rd quartile	26 770.0	3916.0	20.090	76.740	0.5657	25 540.0	210.9	0.0
Maximum	964 000.0	20 380.0	1591.000	36 120.0	2.6700	67 4300.0	4063.0	7984.0

The data pattern is peculiar but akin to a pattern we might surmise to characterize network/behaviour flow data. It is easy to understand why traffic might be near zero for parts of the day and higher during working hours. Traffic spikes but then are diverted, preventing spikes that overburden any part of the system. Our prior with the data was a noninformative, uniform prior, which would not be like the pattern just described. Given that we have some information on the characteristics of the system, we can improve our individual estimates with a more informative prior.

V. Conclusion

In network information flow problems, there is often only origin and destination data. The number of data points, thus, is smaller than the number of parameters that need to be estimated. To identify the unknown underlying adaptive behaviour and to measure causal influence requires one to solve a stochastic inverse problem. The resulting underdetermined inverse problem cannot be solved by traditional estimation and inference methods without imposing a large number of assumptions. A natural solution is to employ information theoretic estimation and inference methods that are designed for problems of this nature. We have demonstrated the applicability of information theoretic methods to information flow-traffic problems. As a first attempt from the CR family, we have made use of the well-known maximum likelihood entropy, $\gamma \to 0$, criterion.

One of our important contributions here is a demonstration of the connection between adaptive behaviour and likelihood or entropy maximization and the use of the CR family of entropy measures for analysing networks. Future research involves exhibiting the performance of other members of the CR family that may reflect particular characteristics of the network flow traffic data and making use of the empirical likelihood, $\gamma \to -1$ or perhaps other nonexponential criterion that makes use of convex combinations of members of the CR family.

References

Airoldi, E. M. and Blocker, A. W. (2013) Estimating latent processes on a network from indirect measurements, *Journal of the American Statistical Association*, **108**, 149–64. doi:10.1080/01621459.2012.756328

Cao, J., Davis, D., Vander Wiel, S. *et al.* (2000) Time-varying network tomography: router link data, *Journal of the American Statistical Association*, **95**, 1063–75. doi:10.1080/01621459.2000.10474303

Castro, R., Coates, M., Liang, G. *et al.* (2004) Network tomography: recent developments, *Statistical Science*, **19**, 499–517. doi:10.1214/088342304000000422

Cho, W. K. T. and Judge, G. G. (2008) Recovering vote choice from partial incomplete data, *Journal of Data Science*, **6**, 155–71.

Cressie, N. and Read, T. R. C. (1984) Multinomial goodness of fit tests, *Journal of the Royal Statistical Society, Series B*, **46**, 440–64.

Golan, A., Judge, G. G. and Miller, D. J. (1996) *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, John Wiley and Sons, New York.

Gorban, A. N., Gorban, P. A. and Judge, G. (2010) Entropy: the Markov ordering approach, *Entropy*, **12**, 1145–93. doi:10.3390/e12051145

Jaynes, E. (1957) Information theory and statistical mechanics II, *Physical Review*, **108**, 171–90. doi:10.1103/PhysRev.108.171

Judge, G. and Mittelhammer, R. (2012) Implications of the Cressie-Read family of additive divergences for information recovery, *Entropy*, **14**, 2427–38. doi:10.3390/e14122427

Judge, G. and Mittelhammer, R. (2011) *An Information Theoretic Approach to Econometrics*, Cambridge University Press, Cambridge.

Read, T. R. C. and Cressie, N. (1988) *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer-Verlag, New York.

Shannon, C. E. (1948) A mathematical theory of communication, *The Bell System Technical Journal*, **27**, 379–423, 623–56. doi:10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E. (1949) Communication in the presence of noise, *Proc IRE*, **37**, 10–21. doi:10.1109/JRPROC.1949.232969

Vardi, Y. (1996) Network tomography: estimating source-destination traffic intensities from link data, *Journal of the American Statistical Association*, **91**, 365–77. doi:10.1080/01621459.1996.10476697

Wissner-Gross, A. and Freer, C. (2013) Causal entropic forces, *Physical Review Letters*, **110**, 168702. doi:10.1103/PhysRevLett.110.168702