Recovering Latent Time-Series from their Observed Sums: Network tomography with Particle Filters.

Edoardo Airoldi and Christos Faloutsos [1]

May 2004

CMU-CALD-04-104, CMU-CS-04-130

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[1]This material is based upon work supported by the National Science Foundation under Grants No. IIS-0083148, IIS-0113089, IIS-0209107 IIS-0205224 INT-0318547 SENSOR-0329549 EF-0331657IIS-0326322 and by the Pennsylvania Infrastructure Technology Alliance (PITA) Grant No. 22-901-0001. Additional funding was provided by donations from Intel, and by a gift from Northrop-Grumman Corporation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties.

Abstract

Hidden variables, evolving over time, appear in multiple settings, where it is valuable to recover them, typically from observed sums. Our driving application is 'network tomography', where we need to estimate the origin-destination (OD) traffic flows to determine, e.g., who is communicating with whom in a local area network. This information allows network engineers and managers to solve problems in design, routing, configuration debugging, monitoring and pricing. Unfortunately the direct measurement of the OD traffic is usually difficult, or even impossible; instead, we can easily measure the loads on every link, that is, sums of desirable OD flows.

In this paper we propose i-FILTER, a method to solve this problem. i-FILTER improves the state-of-the-art by (a) introducing explicit time dependence, and by (b) using realistic, non-Gaussian marginals in the statistical models for the traffic flows, as never attempted before. We give experiments on real data, where i-FILTER scales linearly with new observations and out-performs the best existing solutions, in a wide variety of settings. Specifically, on real network traffic measured at CMU, and at AT&T, i-FILTER reduced the estimation errors between 15% and 46% in all cases.

Keywords: Origin-Destination Traffic Flows, Link Loads, self-organizing Bayesian Dynamical System, MCMC, Particle Filter, Informative Priors, non-parametric empirical Bayes

Acknowledgments

The authors thank Dr. Srinivasan Seshan, Dr. Russel Yount, Dr. Frank Kietzke, and Dr. Jin Cao, for providing the LAN traffic data measured at Carnegie Mellon University and at AT&T, which made this work possible. Further they thank Dr. Stephen E. Fienberg for engaging discussions, and Dr. Irina Rish and Dr. Alina Beygelzimer for helpful comments.

Contents

1 Introduction

Knowledge about the origin-destination (OD) traffic matrix allows network engineers and managers to solve problems in design, routing, configuration debugging, monitoring and pricing; in fact the OD traffic matrix provides valuable information about who is communicating with whom in a network, at any given time. Unfortunately the direct measurement of the OD traffic is usually difficult, or even infeasible, in real networks. The direction of current research is to develop methods to infer the OD traffic flows from observed traffic loads on the links of the network, however the methods that have been proposed so far seem not to fully take advantage of two of the main empirically observed features of network traffic; namely its very skewed marginal distribution, and its time dependent nature.

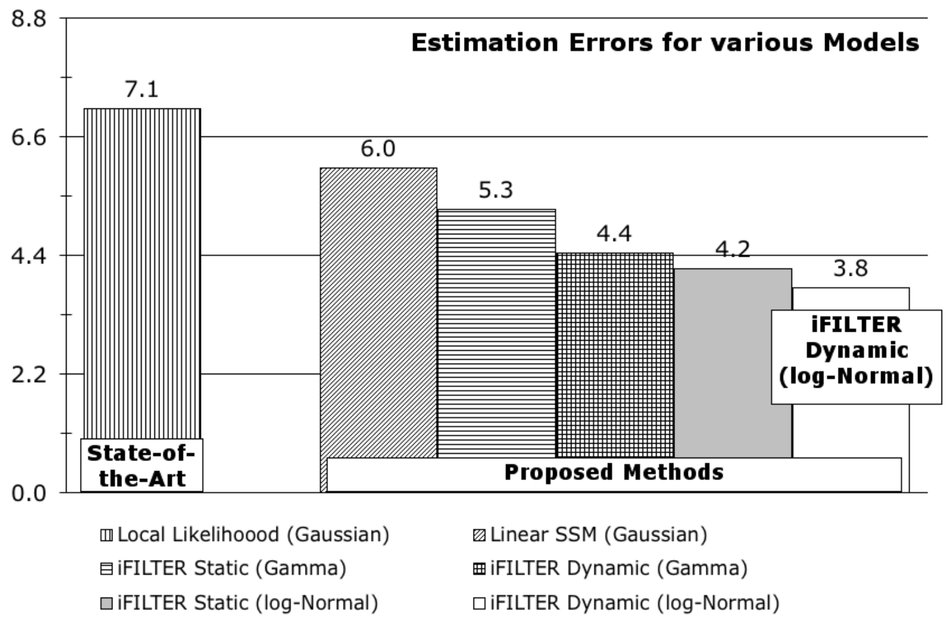

Figure 1: Estimation error in ℓ_2 distance.

In this paper we present i-FILTER, an intelligent filtering method which improves the models present in the literature by introducing two realistic assumptions:

- the successful log-Normal distribution provides a realistic model for the marginal OD traffic flows,
- time dependence between successive flows on a same OD route narrows the variability of the estimates.

i-FILTER uses a two-stage estimation procedure: estimates the pattern of time-dependencies, and eventually uses it to gain in precision.

- the learning algorithm is linear in the number of epochs times the number of OD flows that need be recovered,

- the learning algorithm (particle filter) estimates the parameters of our Bayesian dynamical system and the OD traffic flows with one pass over the data, using few high-quality particles.

In our experiments, using the best log-Normal model reduced the estimation error of previously proposed methods by 35%, and the introduction of explicit stochastic dynamical behavior reduced the estimation error up to 46%. The magnitude of the improvements entailed by the simple ideas we propose goes far beyond that of state-of-the-art re-sampling schemes that could be used to refine any given set of estimates. Further a stochastic dynamical behavior played an essential role in our models; it served as the right channel where to introduce prior information about the OD flows, gained in stage-one of our procedure.

The rest of the paper is organized as follows: Section 2 describes the problem and the relevant issues in more detail. Section 3 surveys the relevant literature. Section 4 describes our proposed methodology. Section 5 reports the results of our experiments on real data. Section 6 discusses our main findings and their relevance towards our solution to the problem, and Section 7 concludes with some final remarks and future research directions.

2 Problem - Background

In this paper we write $\boldsymbol{M}$ for matrices, $\boldsymbol{v}(t)$ for column vectors at time t, and $v(i,t)$ for their generic component i.

T	Number of time points.
ℓ	Number of observable link loads.
κ	Number of non-observable OD flows.
$\boldsymbol{A}$	$(\ell \times \kappa)$ fixed routing matrix.
$\boldsymbol{Y}$	$(\ell \times T)$ matrix of link loads.
$\boldsymbol{X}$	$(\kappa \times T)$ matrix of OD flows.
$\boldsymbol{\Lambda}$	$(\kappa \times T)$ matrix of means of $\boldsymbol{X}\|\boldsymbol{\Lambda}, \boldsymbol{\phi}$.
$\boldsymbol{\phi}$	$(1 \times T)$ vector of scale factors for $\boldsymbol{X}\|\boldsymbol{\Lambda}, \boldsymbol{\phi}$.
$\boldsymbol{\theta}$	Generic vector of hyper-parameters.
$\pi(\boldsymbol{\theta})$	Generic prior distribution.

Table 1: Summary of symbols.

We begin by defining the problem, introducing the relevant issues, and discussing related works.

2.1 The Problem and its Facets

In a formulation of the problem we want to solve there are several time series which we would like to estimate, but which we cannot observe, say, a vector of traffic flows $\boldsymbol{x}(t)$ over times $t = 1, ..., T$. However, we are able to observe linear combinations of these traffic flows, the vector of link loads $\boldsymbol{y}(t)$ over times $t = 1, ..., T$, and we know which components of $\boldsymbol{x}(t)$ mix into each of the components $y(i,t)$ at each time t trough the routing matrix $\boldsymbol{A}$, that does not change over time. This problem can be decomposed in three sub-problems, which we now define.

Problem 1. *(Network Tomography) Given the matrix of link loads* $\boldsymbol{Y}_{(\ell\times T)}$ *and a routing matrix* $\boldsymbol{A}_{(\ell\times\kappa)}$*, we want to find the matrix of non-observable OD traffic flows* $\boldsymbol{X}_{(\kappa\times T)}$ *such that* $\boldsymbol{Y} = \boldsymbol{A}\cdot\boldsymbol{X}$*. Always* $\kappa > \ell$*.*

For example, the linear equations that correspond to the routing scheme of the star network in figure 2 below are:

$$\begin{bmatrix} y(1,t) \\ y(2,t) \\ y(3,t) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x(1,t) \\ x(2,t) \\ x(3,t) \\ x(4,t) \end{bmatrix} \tag{2.1}$$

$y(1,t)$ measures the traffic load on the link from node 1 to the router and captures both the OD flow from node 1 to node 2, $x(2,t)$, and the OD flow from node 1 to itself, $x(1,t)$. $y(3,t)$ measures the traffic load on the link from the router to node 1 and captures both the OD flow from node 2 to node 1, $x(3,t)$, and the OD flow from node 1 to itself, $x(1,t)$.

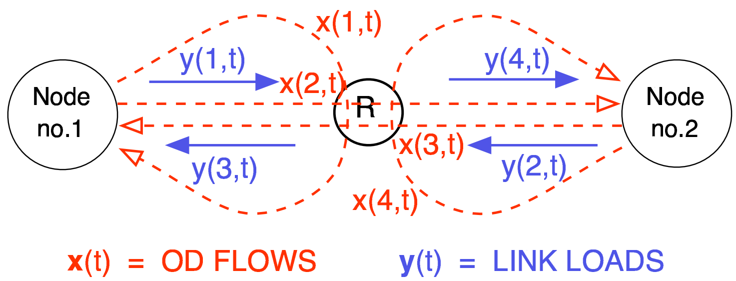

Figure 2: Two subnetworks connect to a router in a small star network. We observe the link loads (solid blue arrows), and want to infer the hidden OD flows (dashed red arrows).

In the example above we want to estimate four (κ) unobservable quantities starting from three (ℓ) independent observations[1]. The system is under-specified, $\kappa > \ell$, hence some extra information is needed in order to identify one single solution.

Problem 2. *(Prejudices) Choose a set of additional constraints to be imposed on* $\boldsymbol{X}_{(\kappa\times T)}$ *in order to make the "network tomography" problem exactly determined.*

The likelihood of the data entailed by a statistical model provides us with a natural criterion to discern "likely" solutions from unreasonable ones. Following this idea we model the unobservable quantities $\boldsymbol{x}(t)$ with a joint probability distribution; this induces a probabilistic mapping on the space of the observations $\boldsymbol{y}(t)$ via equation 2.1, so that we can compute the likelihood of the observations, and look for traffic flows that maximize the probability of particular data observations. Unfortunately in time-independent models the likelihood of $\boldsymbol{y}(t)$ is not necessarily unimodal, even as we assume independent components in $\boldsymbol{x}(t)$, and even as we use well-behaved functional forms for their distributions. More information is needed to identify a solution.

At this point there are two main ways to introduce the extra information we need. In a purely data-driven approach we would augment the data in some way, whereas in a knowledge-driven approach we would make use of informative priors in a Bayesian setting, with the complication in

[1] We assume that routers neither generate nor absorb traffic.

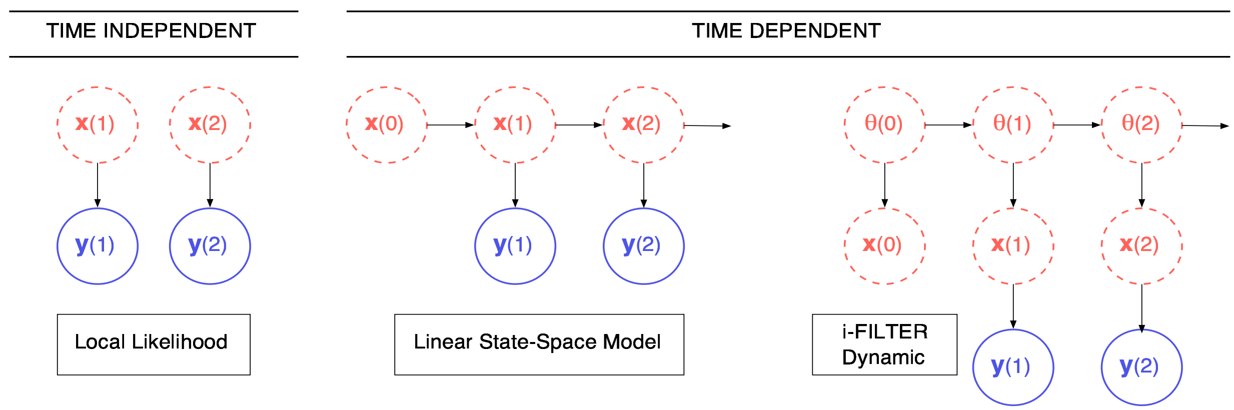

Figure 3: Graphical representations: previous works assume no explicit time dependance (left), whereas the linear SSM introduces an explicit dynamical behavior (center). Our proposed i-FILTER moves the explicit time dependence one layer up in the graphical model, thus allowing for the OD flows to be more diverse (right).

this latter case of defining what we mean by "informative". Data augmentation can be realized, for example, by raising the likelihood of the data to a power, as in simulated annealing, or by borrowing observations from epochs close in time to the current one to obtain a smoothed average solution. Alternatively, we can build "informative" priors based on partial knowledge about the magnitude of the OD flows, and update using Bayes rule and a "more accurate" data model.

Problem 3. *(Speed) Solve efficiently the "network tomography" problem, under the "prejudices" of problem 2.*

We offer a novel solution to the "prejudices" problem, which satisfies the efficiency requirement of the "speed" problem. The two-stage estimation procedure at the core of i-FILTER corresponds to a non-parametric empirical Bayes learning strategy, where the observations are used to first calibrate informative priors, and then to filter the posterior distributions of the OD flows given the data. Our solution: (1) uses realistic models for the OD flows; (2) takes advantage of the time dependence of the data while using the whole history of observations $\{\boldsymbol{y}(1), ..., \boldsymbol{y}(t)\}$ to estimate $\boldsymbol{x}(t)$ in a proper Bayesian fashion.

2.2 State-Space Models

To the best of our knowledge dynamical systems have never been used to solve the "network tomography" problem[2]. Previous works assume independent OD flows across different epochs. We use dynamical systems, which naturally extend previous approaches by assuming time dependence *explicitly*.

Definition 1. *Gaussian linear state-space models are defined by the set of equations,*

$$\begin{cases} \boldsymbol{x}(t) &= \boldsymbol{F} \cdot \boldsymbol{x}(t-1) + \boldsymbol{e}(t) \\ \boldsymbol{y}(t) &= \boldsymbol{A}_{(\ell \times \kappa)} \cdot \boldsymbol{x}(t)\,, \qquad t \geq 1 \end{cases} \tag{2.2}$$

[2]Wei et al. (2002) and others use HMMs to approach a different problem, also called "network tomography".

where $\{\boldsymbol{e}(t)\}$ *is an i.i.d. Gaussian process with variance-covariance matrix* $\boldsymbol{Q}$*, and* $\boldsymbol{F}$ *is a known matrix. Further* $\boldsymbol{x}(0) \sim Normal\,(\boldsymbol{m}, \boldsymbol{V})$ *and independent of* $\boldsymbol{e}(t)$ *for* $t \geq 1$*.*

In figure 3 above we show the graphical representations for the "network tomography" problem at subsequent time points in a static setting, and its natural extension to a dynamic setting, by means of various state-space models (SSM henceforth). Classical state-space modeling strategies a la Box and Jenkins would look for the additional constraints needed to solve problem 2 in a known dynamical behavior suggested by some physical law underlying the specific problem at hand and from known seasonal patterns in the traffic, for example the laws of motion in tracking the trajectories of moving objects, or from the presence of strong cross-correlations among the OD flows. This knowledge would translate into constraints on $\boldsymbol{F}$, and $\boldsymbol{Q}$ in the system 2.2 above, and would serve the critical role of driving the inferences towards one particular solution.

3 Related Works

3.1 Classical Literature

As we discussed above the "network tomography" problem is under-specified, and allows for infinite valid solutions. Fienberg (1968, 1970) studied the geometry of this problem and identified the $(\kappa - \ell)$-dimensional space of solutions. By assuming specific marginal probability distributions for the OD flows we induce a probability map on the the space of measurements. Vanderbei and Iannone (1994) show that assuming Poisson OD traffic does not yield a necessarily convex probability map, and Vardi (1996) further shows that solving the likelihood equations may yield local maximum whereas the actual global maximum is on the boundary, assuming Poisson traffic. Both these works assume a fixed vector of parameters for all epochs. Tebaldi and West (1998) propose a fully Bayesian framework with non-informative priors based on Poisson traffic and acknowledge the need for informative priors, but they do not specify how to obtain them. In fact, using the measurements at time t in order to estimate the OD flows at the same epoch, flat priors over the possible amount of traffic yield uniform or multi-modal posteriors. Their work is important in that they explore the extent to which it is possible to find a solution at each epoch without using the information entailed by past or future measurements.

3.2 Recent Approaches

Solutions to the problem in the non-statistical literature consider a fixed vector of parameters across all epochs as the favorite solution to overcome the lack of information entailed by the measurements. For example, a typical solution would assume traffic flows at different epochs as independent for each OD route and estimate a fixed vector of parameters by composing the constraints given by the measurements via generalized least squares. Similar solutions differ in the way they compose the constraints given by the measurements at different epochs; the alternatives include the assumptions of multivariate Gaussian or multinomial measurements and the composition is carried out via maximum likelihood or via Bayes theorem, or else using maximum entropy arguments, and so on. Economic approaches that explicitly deal with congestion had also been tried with success. As we noted above, these solutions all assume an underlying fixed vector of parameters which governs the OD traffic at all epochs, and we can express them as solutions of a common minimization problem (Airoldi 2003, Zhang et al. 2003).

Model	Time Dependence	Dynamic Estimation	Skewed Marginals
Local Likelihood	–	–	–
iFILTER Static	–	–	✓
Gaussian SSM	✓	✓	–
iFILTER Dynamic	✓	✓	✓

Figure 4: A summary of the models present in the literature in terms of their ability to model relevant features of real traffic data.

Several of these time-independent solutions have been applied locally in order to get a time-varying solution for the OD flows, otherwise constant across time. A drawback of such approaches is that they cannot obtain estimates for the OD flows corresponding to several measurements. For example, a window of 11 observations would not inform about the first and last 5 OD flows (≈ 25 mins). A recent approach due to Cao et al. (2000) assumes multivariate Gaussian OD flows, i.i.d. across time. However, they propose a clever parametrization with several desirable properties which yields reliable estimates. Further Cao et al. (2001) show how to estimate OD traffic flows in bigger networks by solving sub-problems and then composing their solutions. Medina et al. (2002) compare several methods and survey the literature. Zhang et al. (2003) propose a gravity model for the OD traffic, and use Stein's shrinkage estimator to refine their estimates.

3.3 Skewed and Bursty Traffic

The log-Normal distribution has never been used to model the OD traffic flows, however several works support this distribution for traffic in communication networks and over the Internet. Empirical studies and theoretical models that recreate realistic traffic flows from underlying processes consistently come to the conclusion that network traffic flows are skewed and bursty, see Mandelbrot (1965), Leland et al. (1993), Coutin and Carmona (1998), Sarvotham et al. (2001), Wang et al. (2002), and Cao et al. (2002). Most of the skewed distributions follow the characterization given by Zipf (1932), and its generalizations given by Mandelbrot (1953) and Bi et al. (2001). In particular Bi et al. show that the truncated log-Normal distribution is flexible enough to fit a wide variety of such observed behaviors. We show the log-log plots for the two data sets we used in figure 6 below.

3.4 Learning State-Space Models

Ghahramani and Hinton (1996) show how to learn all the parameters in the system 2.2, in our case $\boldsymbol{F}$, $\boldsymbol{Q}$, $\boldsymbol{m}$, and $\boldsymbol{V}$, by means of the EM algorithm. Higuchi (2001) shows how a self-organizing system can be built from non-linear non-Gaussian systems, so that all the relevant parameters are learned during the filtering process. Gilks and Berzuini (2001) propose a particle filter that keeps particles diverse.

4 Proposed Method: i-FILTER

Briefly, i-FILTER uses a two-stage approach to the filtering problem. In the first stage we find preliminary, smooth estimates for the OD flows, which make a good guess for the averages of the OD traffic, and in the second stage we refine these smooth estimates by looking for spikes and bursty periods with one single pass over the data. We model the OD flows as Bayesian dynamical systems, and we use a EM and particle filter as learning algorithms[3].

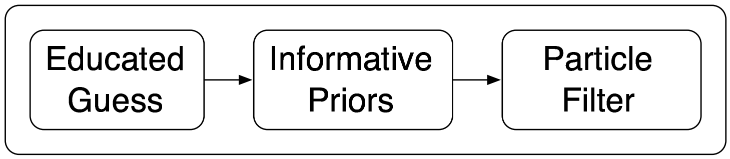

Figure 5: A non-parametric empirical Bayes approach to the filtering problem is at the core of i-FILTER.

More specifically, we use the linear Gaussian SSM and related EM steps proposed in Airoldi (2003), which includes the model in Cao et al. (2000) as a special case, to obtain smooth estimates of the OD traffic, and we then use these estimates to calibrate informative priors for the parameters underlying the dynamic of a non-Gaussian system, in non-parametric empirical Bayes fashion. Eventually the particle filter makes good use of these priors and of the skewed models, and finds a sequence of better posterior distributions for the traffic flow on each OD route; we pick their means as point estimates. Our experiments show i-FILTER achieves an error 15% to 46.5% smaller than that of state-of-the-art methods in all cases.

4.1 Models for OD Traffic

In our experiments on Carnegie Mellon origin-destination traffic we noticed that assuming a fixed relationship between $x(i,t)$ and $x(i,t+1)$ is an unrealistic constraint. Our solution is to assume a relationship between the means of the OD flows $\lambda(i,t)$ and $\lambda(i,t+1)$ instead, and to allow for some error. The SSM yields smooth estimates that capture information about this relationship, which we pass to the next estimation stage. In fact, we introduce soft constraints on the average process $\{\boldsymbol{\lambda}(t)\, t \geq 1\}$ in the form of informative priors for the parameters underlying its dynamical behavior. We reduce the number of parameters by merging dynamic and error terms into a stochastic dynamical behavior. The marginal models for the OD traffic flows are independent log-Normals[4]. The main objects of interest are then the posterior distributions $P(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(1), ..., \boldsymbol{y}(t)\,)$. In particular the point estimate for the OD traffic vector at time t is given by the mean $\hat{\boldsymbol{x}}(t) = E(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(1), ..., \boldsymbol{y}(t)\,)$.

4.1.1 i-FILTER Static

The static version of i-FILTER considers independent problems at each epoch. Briefly, we are interested in estimating $\hat{\boldsymbol{x}}(t) = E(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(t)\,) = E(\, \boldsymbol{x} \,|\, \boldsymbol{y}\,)$. To specify the full models at each time t we write:

$$\begin{cases} \boldsymbol{x} \,|\, \boldsymbol{\lambda}, \phi & \sim \quad p\left(\boldsymbol{\lambda}, \phi\right) \\ \boldsymbol{y} & = \quad \boldsymbol{A} \cdot \boldsymbol{x}, \end{cases} \tag{4.1}$$

[3]An implementation of i-FILTER described in this paper is available for the open source R environment.

[4]Airoldi (2003) also considers Gamma models.

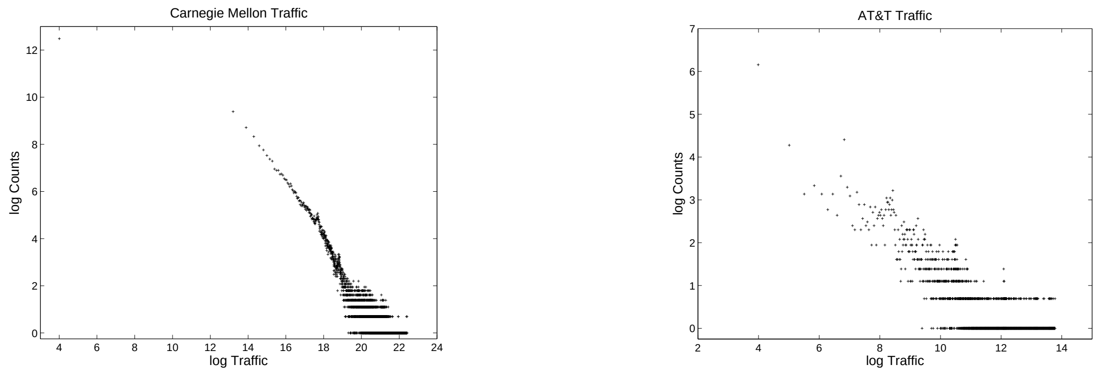

Figure 6: **Left:** a log-log plot of the 12100 traffic flows measured at Carnegie-Mellon university suggests that skewed distributions are appropriate. **Right:** a log-log plot of the 16 traffic flows measured at AT&T, which we use as a test bench for our models. i-FILTER yields the best performance when based on the successful log-Normal distribution.

where p is log-Normal, parametrized so that $E(\,x(i)\,|\boldsymbol{\lambda},\phi\,) = \lambda(i)$, $V(\,x(i)\,|\boldsymbol{\lambda},\phi\,) = \phi \cdot \lambda(i)^{\tau}$, $Cov(\,x(i),x(j)\,|\boldsymbol{\lambda},\phi\,) = 0$ for $i = 1,...,\kappa$ and $i \neq j$. Notice that ϕ is common across OD flows at each epoch, and that τ is a known scalar, which we obtain by inspection of $\boldsymbol{Y}$. The priors for the $\lambda(i)$ are *log-Normal* $(\theta_1(i),\theta_2(i))$[5], for $i = 1,...,\kappa$ and independent for $i \neq j$. The prior for ϕ is proportional to a constant, to $1/\phi$ or to $1/\phi^2$.

4.1.2 i-FILTER Dynamic

This dynamic version of i-FILTER, which yields the best results, implements the following Bayesian dynamical system:

$$\begin{cases} \lambda(i,t) &= \epsilon(i,t)\cdot\lambda(i,t-1),\; i=1,...,\kappa & (4.2)\\ \boldsymbol{x}(t)\,\big|\,\boldsymbol{\lambda}(t),\phi(t) &\sim p\left(\boldsymbol{\lambda}(t),\phi(t)\right) & (4.3)\\ \boldsymbol{y}(t) &= \boldsymbol{A}\cdot\boldsymbol{x}(t), \quad t\geq 1, & (4.4)\end{cases}$$

where p is log-Normal, parametrized so that $E(\,x(i,t)\,|\boldsymbol{\lambda}(t),\phi(t)\,) = \lambda(i,t)$, $V(\,x(i,t)\,|\boldsymbol{\lambda}(t),\phi(t)\,) = \phi\cdot\lambda(i,t)^{\tau}$, and $Cov(\,x(i,t),x(j,t)\,|\boldsymbol{\lambda},\phi\,) = 0$ for $i = 1,...,\kappa$ and $i \neq j$. Notice that $\phi(t)$ is common across OD flows at time t, and that τ is a known scalar, which we obtain by inspection of $\boldsymbol{Y}$. The priors for $\lambda(i,0)$ are *log-Normal* $(\theta(i,0),\sigma)^4$, for $i = 1,...,\kappa$ and independent for $i \neq j$, and for a big number σ that accounts for the uncertainty of the means of OD flows at time zero. The prior for $\phi(t)$ is proportional to a constant, to $1/\phi(t)$ or to $1/\phi(t)^2$. The priors for $\epsilon(i,t)$ are *log-Normal* $(\theta_1(i,t),\theta_2(i,t))^4$ for $i = 1,...,\kappa$, and independent for $i \neq j$.

4.1.3 Informative Priors for $\boldsymbol{\lambda}(t)$

The crucial question at this point is: how do we calibrate the hyper-parameters underlying the prior distributions of $\boldsymbol{\lambda}(t)$? First we obtain a preliminary set of estimates $\hat{\boldsymbol{x}}(t)$ with the Gaussian linear SSM. Then, in the case of i-FILTER static, $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ at each time are set so that mean and variance of $\boldsymbol{\lambda}$ correspond to those of $\hat{\boldsymbol{x}}(t)$. Variances can be made much larger without significant

[5] Airoldi (2003) also considers Gamma, Uniform, and truncated Gaussian priors.

loss of precision. The intuition is that the preliminary estimates indicate us where OD flows are on average. In the case of i-FILTER dynamic the intuition is the same, however it is not possible to set priors for $\boldsymbol{\lambda}(t)$ as the sequence $\{\boldsymbol{\lambda}(1), ..., \boldsymbol{\lambda}(T)\}$ is going to be determined by $\boldsymbol{\lambda}(0)$ alone. The solution is then to extract from $\{\hat{\boldsymbol{x}}(1), ..., \hat{\boldsymbol{x}}(T)\}$ information about their local dynamical behavior and use it to calibrate informative priors for $\{\boldsymbol{\epsilon}(t), t \geq 1\}$. Technically, we set $\epsilon(i,t)$ as independent log-Normals; we use the facts that product convolution of log-Normals is log-Normal (equation 4), and that $log\text{-}Normal(\theta_1(i,t), \theta_2(i,t)) = exp\{\, N(\theta_1(i,t), \theta_2(i,t))\,\}$ to solve the convolution problem exactly for $(\theta_1(i,t), \theta_2(i,t))$, for $i = 1, ..., \kappa$. In other words, values for $(\boldsymbol{\theta}_1(t), \boldsymbol{\theta}_2(t))$ are computed from $(\hat{\boldsymbol{x}}(t), \hat{\boldsymbol{x}}(t-1))$ at each time, and these parameters need not be learned. $\theta(i,0)$ is set to be the average of corresponding OD flow $\{x(i,t),\ t \geq 1\}$.

Notice that every two-stage method that finds preliminary estimates and refines them uses $\{\hat{\boldsymbol{x}}(1), ..., \hat{\boldsymbol{x}}(T)\}$ in the second stage, in some way. We prefer to translate this information into information about the means of the OD flows $\{\boldsymbol{\lambda}(1), ..., \boldsymbol{\lambda}(T)\}$, according to the intuition that preliminary estimates can identify a smooth version of the OD flows we are looking for, which make reasonable guesses for their underlying average processes.

4.2 Parameter Estimation

In order to filter the posterior distributions of the origin-destination flows and estimate the parameters of the models, i-FILTER implements a slight variation of the sample-resample-move algorithm of Gilks and Berzuini (2001), which we briefly outline below. For simplicity define $\boldsymbol{v}(t)$ to be the vector of all parameters in the model at time t, $\boldsymbol{v}(t) := \big(\boldsymbol{x}(t), \boldsymbol{\lambda}(t), \boldsymbol{\epsilon}(t), \phi(t)\big)$, and $\boldsymbol{v}(0) := \big(\boldsymbol{\lambda}(0)\big)$.

Algorithm 1. *Enhanced Particle Filter:*

At $t = 0$ generate N particles $\{\, \tilde{\boldsymbol{\lambda}}_{(i)}(0)\,\}_{i=1}^{N}$ using $\boldsymbol{\theta}(0), \sigma$.

1. *Set $t = t + 1$. Move each particle like so: (a) generate $\boldsymbol{\epsilon}_{(i)}(t)$ using $(\boldsymbol{\theta}_{1,(i)}(t), \boldsymbol{\theta}_{1,(i)}(t))$, and $\phi_{(i)}(t)$; (b) compute $\boldsymbol{\lambda}_{(i)}(t)$ using the equation 4; (c) generate $\boldsymbol{x}_{(i)}(t)$ by sampling from equation 5.*
2. *Resample N new particles from $\{\, \tilde{\boldsymbol{v}}_{(i)}(t)\,\}_{i=1}^{N}$ according to the likelihood, $P(\, \boldsymbol{y}(t) \,|\, \tilde{\boldsymbol{v}}_{(i)}(t)\,)$, they entail.*
3. *Move the new set of particles according to a MCMC for "several steps" to improve their diversity. Go to 1.*

For details about the MCMC see Airoldi (2003).

4.3 Scalability and Irreducibility

A recent result in network tomography (see Cao et al. 2001) is that it is possible to reformulate filtering problems corresponding to complex networks as a sequence of problems corresponding to small, simple networks.

Lemma 1. *The complexity of i-FILTER is $O(\, \kappa \cdot T\,)$.*

Proof. We use results by Cao et al. (2001) which imply that a tomography problem corresponding to a network with κ origin-destination flows is equivalent to $O(\kappa)$ tomography problems, which correspond to disjoint sub-sets of, say, one to four OD traffic flows in the original problem. This fact along with the fact that our solution is linear in the number of time points for which the OD traffic need be filtered, yields a total complexity of $O(\, \kappa \cdot T\,)$ for i-FILTER. □

Lemma 1 implies if we solve the "network tomography" problem for small size networks, we immediately solve it for arbitrary size networks with comparable estimation errors. Further the following result holds:

Lemma 2. *i-FILTER is based on an irreducible MCMC.*

Proof. See Appendix A. □

Lemma 2 implies that i-FILTER is able to explore the support of the whole joint posterior distribution of the OD flows. A proof of this fact is needed since the MCMC uses a Gibbs sampler with Metropolis steps.

5 Experiments

5.1 Experiment Setup

5.1.1 Data Sets Description

We had two data sets available. Both of them included validation data.

- Carnegie Mellon traffic: the first data set, which we used to choose the appropriate model, contained about 12100 origin-destination traffic flows measured every 5 minutes over slightly less than two days at Carnegie-Mellon university (CMU). We measured an average traffic of 14GB every 5 minutes.

- AT&T traffic: the second data set, which we used to test and compare the filtered traffic obtained with different methodologies, contained 16 origin-destination flows measured every 5 minutes over a one-day period at AT&T, courtesy of Dr. Jin Cao at Bell Labs.

The analysis of Carnegie Mellon origin-destination traffic flows supports the hypothesis of a very skewed distribution. In figure 6 we plotted the logarithms of the observed flows versus the logarithms of the number of times measurements of such a size appear (aka. log-log plot), after discarding the measurements smaller than a standard packet (53 bytes = 424 bytes). The log-log plot indicates a log-Normal distribution may be appropriate. A histogram of the logarithms of the flows indicated that a logarithmic transformation is actually too mild to remove all the skewness, and a double logarithmic transformation would be more appropriate. The AT&T data set is much smaller, and contains traffic flows generated on a smaller network; they are less skewed overall, and a logarithmic transformation is enough to yield a symmetric histogram for the truncated flows.

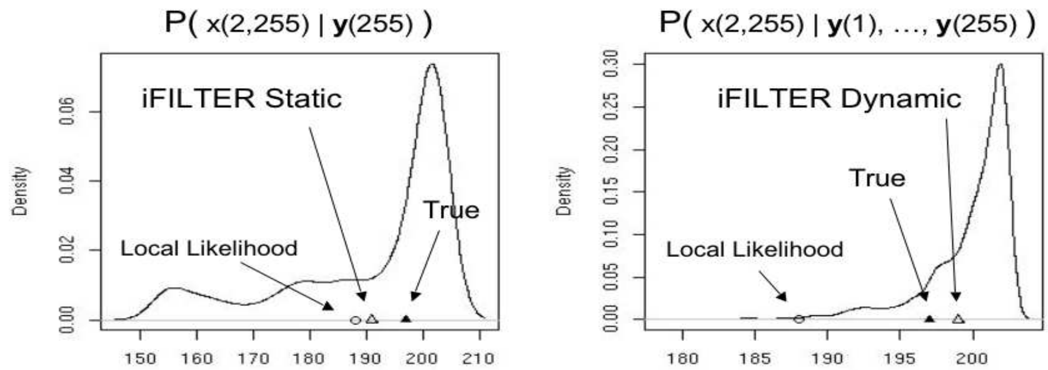

Figure 7: Example posterior distributions for the OD flow $x(2, 255)$. The traffic on the X axes is measured in Kbytes, and the figures show the posterior distribution we obtained with i-FILTER static (left panel) versus the one we obtained with i-FILTER dynamic (right panel). The solid triangles represent the true hidden OD Flow, whereas the empty triangles are our point estimates, which correspond to the means of the posterior distributions. Making use of all the observations $\{ \boldsymbol{y}(1), ..., \boldsymbol{y}(255) \}$ in computing the posterior distribution in the right panel reduced its variability — notice the different ranges — thus improving the inferences.

5.1.2 Performance and Robustness

We used the CMU data set throughout the modeling process, on star network topologies. In order to test both the performance of i-FILTER and the robustness of its underlying log-Normal model we eventually set the CMU data aside, and used the AT&T data set as an independent validation data set.

Recall, as we noted in section §4.3 above, that it is possible to break down a big filtering problem on a complex network into a sequence of smaller problems on (disjoint) simple star networks. Hence in order to test the performance of our methods, we sampled the non-observable origin-destination flows from the AT&T data set, and computed the corresponding link loads for the simple star network below, composed by one router and two nodes connected to it.

Eventually we reconstructed the origin-destination flows, starting from the available measurements about the link loads, and measured the errors of the various methods we considered.

5.2 Results

Combining realistic models and time dependence into i-FILTER reduced the error obtained with local likelihood between 15.0% and 46.5%. We performed experiments to answer the following questions:

1. *Skewed Marginals:* what is the impact of skewed model on the accuracy of the estimates? And what is the best model for the OD traffic?

2. *Time Dependence:* what is the impact of explicit time dependence on the accuracy of the estimates?

3. *Informative Priors:* what constraints should we impose to solve the "prejudice" problem? How do they impact the accuracy of our estimates?

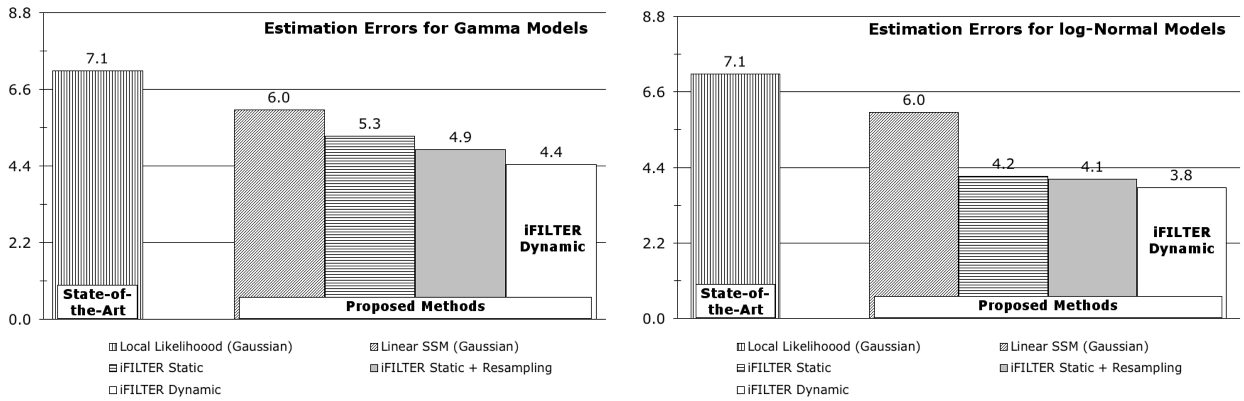

Figure 8: The bars represent the average estimation error in a validation set. Specifically we plot the ℓ_2 distance between the true OD flows and the corresponding estimates obtained with the local likelihood approach and i-FILTER in its various flavors. i-FILTER based on the Bayesian dynamical system is a clear winner! In both panels we include the estimates obtained with our Gaussian linear SSM. **Left:** error bars corresponding to i-FILTER based on Gamma models. **Right:** error bars corresponding to i-FILTER based on log-Normal models.

4. *Scalability:* does our algorithm scale well with the problem size?

We are most interested in assessing the differential impact on the estimation error of of our novel ideas, as opposed to the methods present in the literature. We compared the inferences obtained with different methods by computing the ℓ_2 distance between the true OD flows in the validation set and the estimates. The best results were obtained with log-Normal distribution for the flows and Gaussian vague priors.

5.2.1 Effect of Skewed Marginals

To isolate the effect of realistic distributions for the OD flows, we compared the estimates obtained with i-FILTER where no time dependence was assumed, for Gamma and log-Normal models, and a variety of non-conjugate priors (Uniform, Gaussian, Gamma,l log-Normal) and different parametrizations, with the estimates obtained by local likelihood. Introducing realistic model reduced the error between 25.4% and 40.8%.

5.2.2 Effect of Time Dependence

To isolate the effect of explicit time dependence, we compared the estimates we obtained with the gaussian linear SSM in Airoldi (2003) that uses independent AR(1) processes for the OD flows, with the estimates obtained with local likelihood. Introducing time dependence reduced the error by 15.5% on average; the reduction ranged between 8.5% and 31.0%.

5.2.3 Informative Priors

Using i-FILTER static in 60% of the time points uninformative priors yield flat or multi-modal posteriors, whereas in the remaining 40% of the time points flat priors yield wide uni-modal posteriors. The main effect of the data at $\boldsymbol{y}(t)$ on the posterior $P(\boldsymbol{x}(t)|\boldsymbol{y}(t))$ is on its range; impossible configurations receive zero posterior probability. Informative priors with wide variance all yield uni-modal distributions. i-FILTER dynamic with informative priors has the advantage of requiring fewer particles than the version based on flat priors; knowing where to sample may introduce bias, but the thick tails of the log-Normal distribution of both $\boldsymbol{x}(t)|\boldsymbol{\lambda}(t), \phi(t)$ and $\boldsymbol{\lambda}(t)|\boldsymbol{\theta}_1(t), \boldsymbol{\theta}_2(t)$ mitigate the problem, and i-FILTER captures several of the hidden spikes.

5.2.4 Scalability

i-FILTER reconstructs the OD flows with one single pass over the data; the first stage of the procedure and the calibration of the priors are also linear in the problem size, as we compute the posterior distributions using the gaussian linear SSM on a large sliding windows, for example. Figure 9 below shows that in less than one hour i-FILTER calibrates the informative priors and recovers the flows corresponding to one-day worth of data for the small network in figure 2, on a Powerbook G4 1.25Ghz.

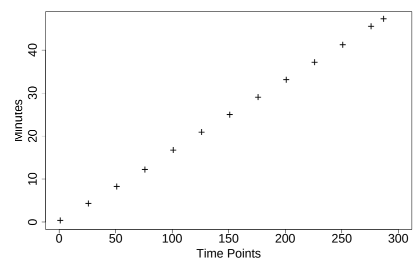

Figure 9: i-FILTER scales linearly with the problem size (number of time ticks).

6 Discussion

In this paper we propose i-FILTER, a two-stage procedure which finds its natural place in a non-parametric empirical Bayes framework. i-FILTER reduced the estimation error, when compared to state-of-the-art methods in realistic scenarios, between 15% and 46.5% in all cases. Such improvements are mainly due to three factors:

1. realistic log-Normal distributions for the OD flows,
2. dynamical systems as a natural way to deal with data with a temporal dimension,
3. informative priors,

Further, i-FILTER scales linearly with the size of the problem. Briefly, we recover a smooth version of the OD flows, we calibrate informative priors for some crucial parameters, and eventually we use a dynamical Bayesian system to refine the estimates and capture bursty traffic. This methodology allows us to combines the three simple ideas above: a realistic model for the data, the use of a filtering scheme which takes advantage of time, probabilistic constraints to overcome the under-determinacy of the problem.

In the first stage we use the Gaussian linear SMM proposed in Airoldi (2003), and we calibrate informative priors for $\boldsymbol{\lambda}(t)$ using these estimates. These priors incorporate information about the

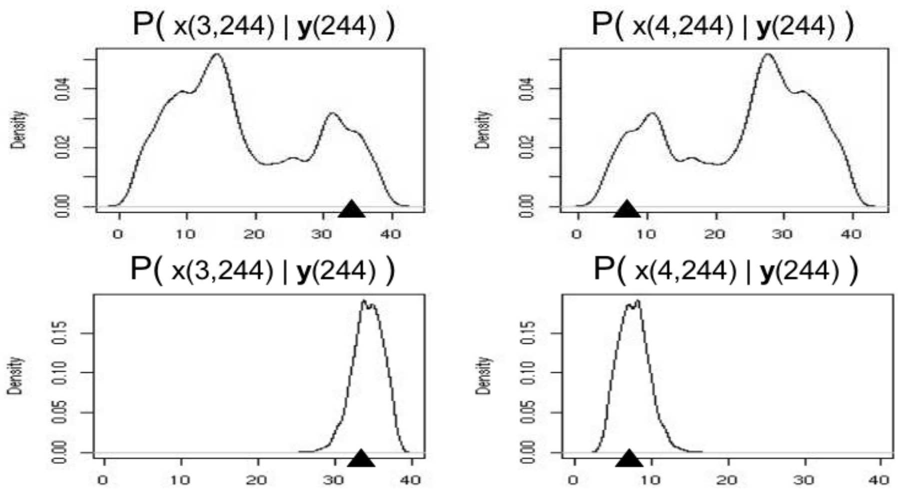

Figure 10: Example posterior distributions for the OD flows $x(3, 244)$ and $x(4, 244)$. The traffic on the X axes is measured in Kbytes, and the figures show the posterior distributions we obtained with non-informative priors (top panel) and with informative priors (bottom panel) calibrated using our Gaussian linear SSM. The solid triangles represent the true hidden OD Flows, whereas our point estimates would be the means of the posterior distributions. Making the posteriors *more unimodal* improves the estimates by reducing the bias entailed by extra modes.

magnitude and the dynamical behavior of the first stage smooth estimates, and softly constrain the location of the average processes $\{\boldsymbol{\lambda}(t),\, t \geq 1\}$. Other methods proposed in the literature make use of preliminary estimates, but they only retain the information about the magnitude of the OD flows given by the such estimates in the refining stage — see for example Zhang et al. (2003) who use shrinkage to improve the solutions given by a gravity model. In our method, the fact that we retain also the information about the local dynamical behavior yields a significant jump in the final accuracy. Another channel through which informative priors help achieve a better accuracy is by reducing the bias entailed by multiple modes in the posterior distributions. Making the posteriors more *uni-modal* improves the precision of the point estimates of the OD flows (the posterior means) as we show in figure 10 below. Informative priors do drive the inferences about the OD flows towards the preliminary guesses, however the two layers of our model and the use of soft probabilistic constraints entail enough flexibility to capture several of the spikes in many cases, for an example see figure 12 below. Further our first-stage estimates are safely based on a model which entails a one-to-one relationship between OD flows and measurements, as it includes the model by Cao et al. (2000) as a special case.

In the second stage the primary object of interest become the sequence of posterior distributions $P(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(1), ..., \boldsymbol{y}(t)\,)$. We use their means $\hat{\boldsymbol{x}(t)} = E(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(1), ..., \boldsymbol{y}(t)\,)$ as point estimates for the OD flows at time t. The Bayesian dynamical system brings further improvements, as we show in figure 7 above, due to the fact that we make use of all the observations up to time t in computing the posterior distributions $P(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(1), ..., \boldsymbol{y}(t)\,)$; conditioning on more observations yields a narrower variability. Local methods use fewer observations in a short window around t, instead. The improvement i-FILTER achieves goes beyond the contribution of state-of-the-art methods even when combined with recent resampling schemes which improve any given set of estimates.

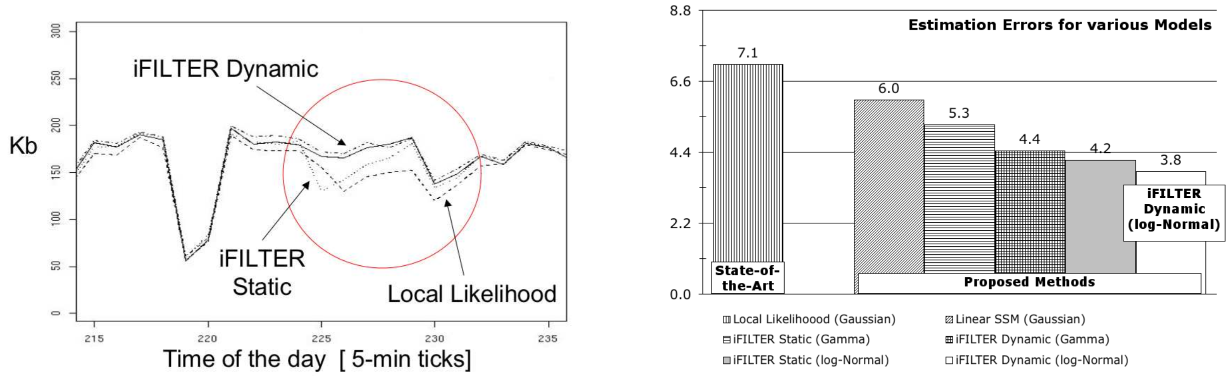

Figure 11: **Left:** the performance of the i-FILTER dynamic based on log-Normal marginals is quite amazing. The goodness of the estimates can be appreciated after sharp changes in the hidden traffic flows. In such situations local solutions tend to smooth the changes away. **Right:** The bars represent the average estimation error in a validation set. Specifically we plot the ℓ_2 distance between the true OD flows and the corresponding estimates obtained with the local likelihood approach and i-FILTER in its various flavors. We found that the i-FILTER static (Gamma and log-Normal) improved the estimates obtained with local likelihood by 15% and 30%, respectively, and that i-FILTER dynamic improved the estimates by 23% and 33%, respectively.

7 Conclusions

In this paper we describe i-FILTER, a two-stage estimation method that combines three simple ideas:

1. log-Normal models to account for skewed, and bursty unobservable OD traffic flows,
2. a novel Bayesian dynamical systems to take advantage of the time dependence of the data and narrow the variability of the estimates,
3. informative priors as soft constraints to overcome the under-determinacy of the problem.

We offer a simple heuristic to understand our modeling choices; first-stage estimates capture smooth average processes, second-stage estimates capture the spikes. We tested our methodology on star topologies using real traffic measured at Carnegie Mellon (12100 OD flows) and AT&T (16 flows) and reduced by 15% to 46.5% the estimation error obtained with state-of-the-art methods, in all cases. Finally we provided some insight in how to calibrate informative priors in Bayesian systems, where no clear guidance about the dynamic of the hidden OD flows is available. In the future we plan to explore whether periodic traffic patterns may also be discovered successfully.

References

[1] E. M. Airoldi. Advances in network tomography. Technical Report CMU-CALD-03-101, Carnegie Mellon University, 2003.

[2] Y. Bejerano, Y. Breitbart, M. Garofalakis, and R. Rastogi. Physical topology discovery for large multi-subnet networks. In *Proceedings of IEEE INFOCOM*, pages 342–352, 2003.

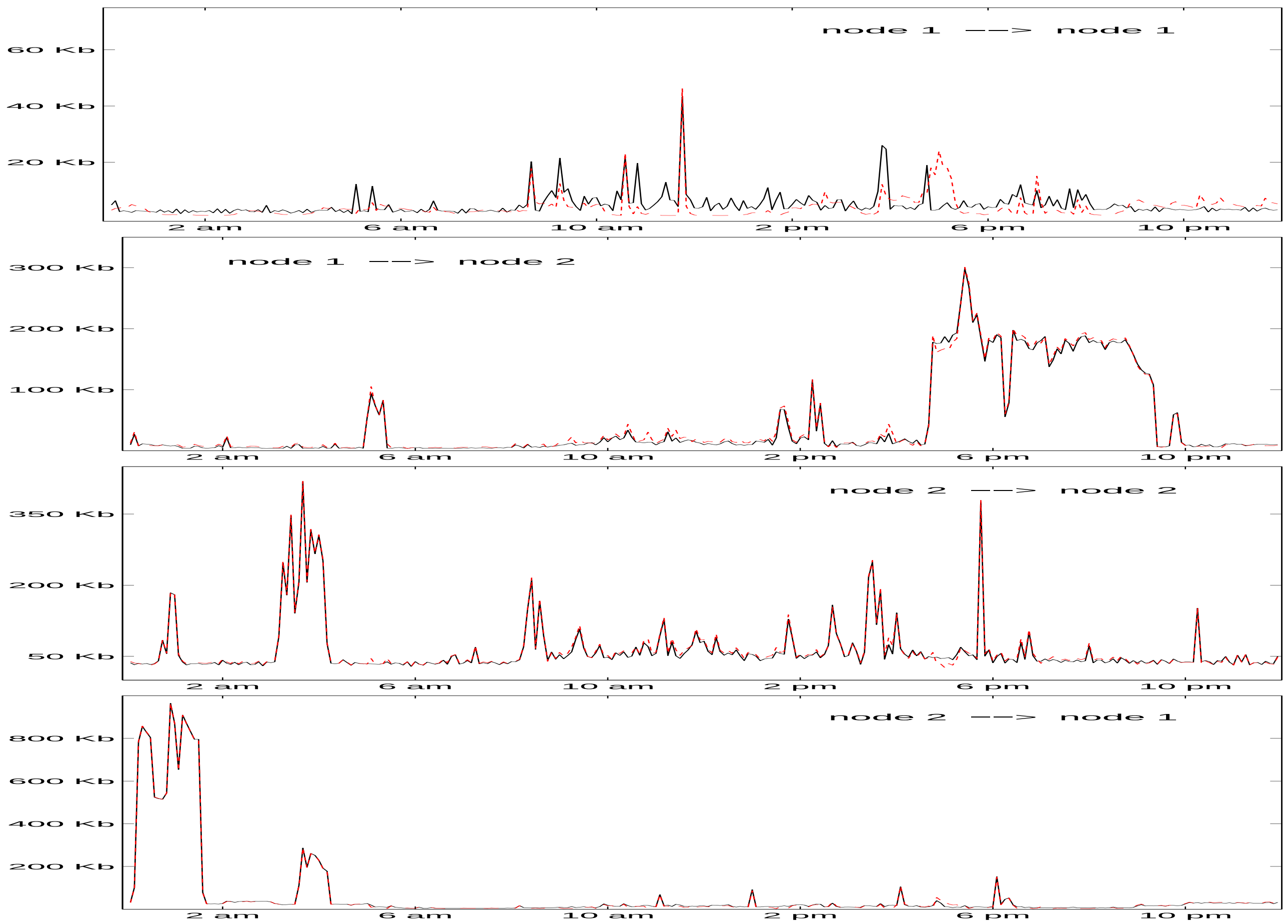

Figure 12: Example fits: actual hidden flows (solid black lines) versus reconstructed flows (dashed red lines). Notice how well we manage to reconstructs several spikes. This figure prints better in colors.

[3] Z. Bi, C. Faloutsos, and F. Korn. The DGX distribution for mining massive, skewed data. In *Knowledge Discovery and Data Mining*, pages 17–26, 2001.

[4] J. Cao, D. Davis, S. Van Der Viel, and B. Yu. Time-varying network tomography: router link data. *Journal of the American Statistical Association*, 95:1063–75, 2000.

[5] J. Cao, D. Davis, S. Van Der Viel, B. Yu, and Z. Zu. A scalable method for estimating network traffic matrices from link counts. Technical report, Bell Labs, 2001.

[6] S. E. Fienberg. The geometry of an $r \times c$ contingency table. *Annals of Mathematical Statistics*, 39:1186–90, 1968.

[7] S. E. Fienberg. The geometry of a two by two contingency table. *Journal of the American Statistical Association*, 65:694–701, 1970a.

[8] W. R. Gilks and C. Berzuini. Following a moving target — monte carlo inference for dynamic bayesian models. *Journal of the Royal Statistical Society, Series B*, 63:127–146, 2001.

[9] T. Higuchi. Self-organizing time series model. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[10] G. Liang and B. Yu. Pseudo-likelihood estimations in network tomography. In *Proceedings of IEEE INFOCOM*, 2003.

[11] J. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, 2001.

[12] B. Mandelbrot. An informational theory of the statistical structure of language. In W. Jackson, editor, *Communication Theory*. Butterworths, 1953.

[13] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic matrix estimation: existing techniques and new directions. In *Proceedings of the ACM SIGCOMM*, pages 161–174, 2002.

[14] I. Rish, M. Brodie, and S. Ma. Accuracy vs. efficiency trade-offs in probabilistic diagnosis. In *Proceedings of AAAI*, 2002.

[15] I. Rish, M. Brodie, N. Odintsova, S. Ma, and G. Grabarnik. Real-time problem determination in distributed systems using active probing. To appear in *NOMS*, 2004.

[16] C. and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 1999.

[17] C. Tebaldi and M. West. Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93:557–576, 1998.

[18] R. J. Vanderbei and J. Iannone. An em approach to od matrix estimation. Technical Report SOR 94-04, Princeton University, 1994.

[19] Y. Vardi. Network tomography: estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91:365–377, 1996.

[20] S. Vaton and A. Gravey. Iterative bayesian estimation of network traffic matrices in the case of bursty flows. In *Proceedings of the second ACM SIGCOMM Workshop on Internet measurment*, pages 89–90, 2002.

[21] W. Wei, B. Wang, D. Towsley. Continuous-time Hidden Markov Models for Network Performance Evaluation. *Performance Evaluation*, 49:129–146, 2002.

[22] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast accurate computation of large-scale IP traffic matrices from link loads. In *Proceedings of SIGMETRICS*, 2003.

[23] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An information-theoretic approach to traffic matrix estimation. In *Proceedings of SIGCOMM*, 2003.

[24] G. K. Zipf. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard Press, 1932.

Appendix A: Proof of Lemma 2

The proof is based on the following result.

Fact. *There exists a permutation ρ of the columns of $A_{(\ell\times\kappa)}$ such that $[A]_{(i,\rho(j))} = [A_1 \,|\, A_2]$, where A_1 is $(\ell \times \ell)$ and has full rank, and A_2 is $(\ell \times (\kappa - \ell))$.*

As a consequence we can permute the components of X to get $[X]'_{\rho(i)} = [X_1 \,|\, X_2]'$, and $Y = A\,X = A_1\,X_1 + A_2\,X_2$, and finally express X_1 in terms of X_2 and Y, like so:

$$X_1 = A_1^{-1} \cdot (Y - A_2\,X_2)$$

Proof. The Gibbs sampler scheme involves iterative sampling from the full conditional distributions $P(Z_i|Z_{(-i)} = z_{(-i)})$, for $i = 1, ..., N$ and Z vector. A sufficient condition to ensure the irreducibility of the chain, Besag (1974), requires that the support of the full conditional distributions is positive where that of the joint distribution of Z is positive, that is:

$$\text{if} \quad P(Z_i = z_i, Z_{(-i)} = z_{(-i)}) > 0 \quad \Rightarrow \quad P(Z_i|Z_{(-i)} = z_{(-i)}) > 0. \tag{A.1}$$

2D case: we show that condition A.1 holds. Specifically consider the situation displayed in figure 6 above, where there are $\kappa - \ell = 2$ components of X_2 that we need to sample from. The chain is at a point $X_2 > 0$ where the joint support is positive and $A_1^{-1}(Y - A_2X_2) > 0$, and it moves by $(+\epsilon, +\epsilon)'$ to the point $X_2 + (\epsilon, \epsilon)'$ where the joint support is also positive and $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, \epsilon)')\,) > 0$. We want to show that whenever both X_2 and $X_2 + (\epsilon, \epsilon)'$ are feasible, it is possible to pass from the former to the latter by means of component-wise moves, as we would with Gibbs moves; that is, the support of the full conditionals must be positive either at $A_1^{-1}(Y - A_2\,(X_2 + (0, \epsilon)')\,)$ or at $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, 0)')\,)$. In other words we want to show that

$$\{\, A_1^{-1}(Y - A_2X_2) \geq 0 \quad \wedge \quad A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, \epsilon)')\,) \geq 0 \,\} \tag{A.2}$$

implies

$$\{\, A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, 0)')\,) \geq 0 \quad \vee \quad A_1^{-1}(Y - A_2\,(X_2 + (0, \epsilon)')\,) \geq 0 \,\}. \tag{A.3}$$

Assume that A.2 holds. Notice that $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, \epsilon)')\,) = A_1^{-1}(Y - A_2X_2 - \epsilon(A_2^{11}, A_2^{21})' - \epsilon(A_2^{12}, A_2^{22})') \geq 0$. Add $A_1^{-1}(Y - A_2X_2) \geq 0$, non negative by assumption, and rearrange terms to get $A_1^{-1}(Y - A_2X_2 - \epsilon(A_2^{11}, A_2^{21})') + A_1^{-1}(Y - A_2X_2 - \epsilon(A_2^{12}, A_2^{22})') \geq 0$ which cannot be the sum of two negative quantities. QED.

Similar derivations show that whenever the joint support has positive probability at $A_1^{-1}(Y - A_2\,(X_2 - (\epsilon, \epsilon)')\,)$ then it also possible for the chain to get there either through $A_1^{-1}(Y - A_2\,(X_2 - (0, \epsilon)')\,)$ or through $A_1^{-1}(Y - A_2\,(X_2 - (\epsilon, 0)')\,)$; and that the same condition holds as we consider the moves to the points $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, -\epsilon)')\,)$ and $A_1^{-1}(Y - A_2\,(X_2 + (-\epsilon, \epsilon)')\,)$.

General case: the proof is exactly the same as in the 2D case, but more tedious. Now X_2 and $(\epsilon, ..., \epsilon)'$ are $\kappa - \ell = n$-dimensional. Assume a $A_1^{-1}(Y - A_2X_2) \geq 0$ and $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, ..., \epsilon)')\,) \geq 0$ hold true. Rewrite $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, ..., \epsilon)')\,)$ as $A_1^{-1}(Y - A_2\,X_2 - \epsilon(A_2^{11}, A_2^{21}, ..., A_2^{n1})' - ... - \epsilon(A_2^{1n}, A_2^{2n}, ..., A_2^{nn})')\,) \geq 0$. Add $(n-1) \times A_1^{-1}(Y - A_2X_2) \geq 0$, non negative by assumption, and rearrange terms to get $A_1^{-1}(Y - A_2X_2 - \epsilon(A_2^{11}, A_2^{21}, ..., A_2^{n1})') + ... + A_1^{-1}(Y - A_2X_2 - \epsilon(A_2^{1n}, A_2^{2n}, ..., A_2^{nn})') \geq 0$, which cannot be the sum of n negative terms. QED.

Again similar derivations show that condition A.1 holds as we consider moves to other points $X_2 + (\pm\epsilon, ..., \pm\epsilon)'$. □

Appendix B: Full Conditionals for the Gibbs Sampler

Say $\Theta = (\lambda_1, ..., \lambda_\kappa, \phi)'$ then $P(X, \Theta) = \prod_{i=1}^{\kappa} P(X_i|\Theta)\, P(\Theta) = \prod_{i=1}^{\kappa} P(X_i|\lambda_i, \phi)\, P(\lambda_i)\, P(\phi)$. We want $\lambda_i \in (0, \infty)$ and $\phi \in (0, \infty)$. As an example, assume priors for λ_i and $1/\phi$ proportional to a constant, and $\tau = 1$. Then, noticing that $P(\Theta|X, Y) = P(\Theta|X)\, I_{\{A^{-1}Y\}}(X)$, the following full conditional distributions can be derived.

$$
\begin{aligned}
P(\lambda_i|X, Y) &\propto \prod P(X_i|\lambda_i, \phi) \cdot P(\lambda_i) \\
&\propto \frac{1}{\lambda_i^{\frac{Ik}{2}}} e^{-\frac{1}{2\phi}\left(\frac{\log(X_i)-\lambda_i}{\lambda_i^k}\right)^2} \\
P(\phi|X, Y) &\propto P(X_i|\lambda_i, \phi) \cdot P(\phi) \\
&\propto \frac{1}{\phi^{\frac{I}{2}+2}}\, e^{-\frac{1}{2\phi}\sum_i\left(\frac{\log(X_i)-\lambda_i}{\lambda_i^k}\right)^2}.
\end{aligned}
$$

In order to compute $P(X|Y, \Theta)$ we use the fact in Appendix A to conclude that $P(X|Y, \Theta) = P(X_2|Y, \Theta) \times \times P(X_1(X_2)|Y, \Theta)$; hence for $X_i \in X_2$ and $X_j \in X_1$ it follows:

$$
\begin{aligned}
P(X_i|X_{(-i)}, Y, \Theta) &\propto P(X_i|\Theta) \cdot P(X_1|Y, \Theta) \\
&= \textit{log-Normal}_{X_i}(\lambda_i, \phi\lambda_i) \cdot \prod_j \textit{log-Normal}_{X_j}(\lambda_j, \phi\lambda_j)\, I_{\{A^{-1}Y\}}(X_j)
\end{aligned}
\qquad \text{(B.1)}
$$

In the analysis, we explored the various posterior distributions using the Gibbs sampler with Metropolis steps. In order to sample from $P(X_i|Y, \Theta)$ and $P(\lambda_i|X, Y)$, we used χ^2 and Uniform proposals, improper priors on the lambdas (all proportional to a constant), and several flavors for the improper prior on ϕ (proportional to a constant, to $\frac{1}{\phi}$, and oto $\frac{1}{\phi^2}$).

Appendix C: First Stage Linear Gaussian State-Space Model

The model we used in to obtain the preliminary estimates for the OD flows is:

$$
\begin{aligned}
&\left\{\begin{array}{lcl}
\boldsymbol{x}(t) & = & \boldsymbol{F}\cdot\boldsymbol{x}(t-1)+\boldsymbol{Q}\cdot\mathbf{1}+\boldsymbol{e}(t)\\
\boldsymbol{y}(t) & = & \boldsymbol{A}\cdot\boldsymbol{x}(t)+\boldsymbol{\epsilon}(t)
\end{array}\right.\\
=&\left\{\begin{array}{lcl}
\left[\begin{array}{c}\boldsymbol{x}(t)\\ \mathbf{1}\end{array}\right] & = & \left[\begin{array}{cc}\boldsymbol{F} & \boldsymbol{Q}\\ \mathbf{0} & \boldsymbol{I}\end{array}\right]\cdot\left[\begin{array}{c}\boldsymbol{x}(t-1)\\ \mathbf{1}\end{array}\right]+\left[\begin{array}{c}\boldsymbol{e}(t)\\ \mathbf{1}\end{array}\right]\\
\boldsymbol{y}(t) & = & [\boldsymbol{A}|\,\mathbf{0}\,]\cdot\left[\begin{array}{c}\boldsymbol{x}(t)\\ \mathbf{1}\end{array}\right]+\boldsymbol{\epsilon}(t)
\end{array}\right.\\
=&\left\{\begin{array}{lcl}
\tilde{\boldsymbol{x}}(t) & = & \tilde{\boldsymbol{F}}\cdot\tilde{\boldsymbol{x}}(t-1)+\tilde{\boldsymbol{e}}(t)\\
\boldsymbol{y}(t) & = & \tilde{\boldsymbol{A}}\cdot\tilde{\boldsymbol{x}}(t)+\boldsymbol{\epsilon}(t)
\end{array}\right.
\end{aligned}
\tag{C.1}
$$

for $t \geq 1$, where $\mathbf{1} = (1, ..., 1)'$ is a constant vector of the length κ, the parameter $\phi(t)$ enters into the variance-covariance matrix of $\boldsymbol{e}(t) \sim N(\mathbf{0}, \phi(t)\cdot\boldsymbol{Q}^{\tau})$, $\boldsymbol{x}(1) \sim N(\mathbf{0}, \boldsymbol{V}(1))$, $\boldsymbol{\epsilon}(t) \sim N(\mathbf{0}, \boldsymbol{R})$, $\boldsymbol{x}(1) \perp \boldsymbol{e}(t)$ and $\boldsymbol{x}(1) \perp \boldsymbol{\epsilon}(t)$ for all $t \geq 1$, and finally $\boldsymbol{Q}$ is a diagonal matrix with elements $(q_1, ..., q_\kappa)$, and τ is a known constant. In the model above, if we set $\boldsymbol{F} = \mathbf{0}$ there is a one-to-one mapping between $(q_1, ..., q_\kappa, \phi(t))'$ and the unique elements in $E(\boldsymbol{y}(t)), V(\boldsymbol{y}(t))$. Further it is straightforward to verify that the following lemma holds.

Lemma 3. *The linear Gaussian state-space model in equations C.1 contains the model in Cao et al. (2000, 2001) as a special case. Such a model can be obtained by simply setting $\boldsymbol{F} = \mathbf{0}$, hence imposing independence among the origin-destination flows $\boldsymbol{x}(t)$ at different epochs.*