Estimating arterial traffic conditions using sparse probe data

Ryan Herring
Industrial Engineering and Operations Research
University of California, Berkeley
ryanherring@berkeley.edu

Aude Hofleitner
Electrical Engineering and Computer Science
University of California, Berkeley
aude@eecs.berkeley.edu

Pieter Abbeel
Electrical Engineering and Computer Science
University of California, Berkeley
pabbeel@cs.berkeley.edu

Alexandre Bayen
Civil and Environmental Engineering
Systems Engineering
University of California, Berkeley
bayen@berkeley.edu

Abstract

Estimating and predicting traffic conditions in arterial networks using probe data has proven to be a substantial challenge. In the United States, sparse probe data represents the vast majority of the data available on arterial roads in most major urban environments. This article proposes a probabilistic modeling framework for estimating and predicting arterial travel time distributions using sparsely observed probe vehicles.

We evaluate our model using data from a fleet of 500 taxis in San Francisco, CA, which send GPS data to our server every minute. The sampling rate does not provide detailed information about where vehicles encountered delay or the reason for any delay (i.e. signal delay, congestion delay, etc.). Our model provides an increase in estimation accuracy of 35% when compared to a baseline approach for processing probe vehicle data.

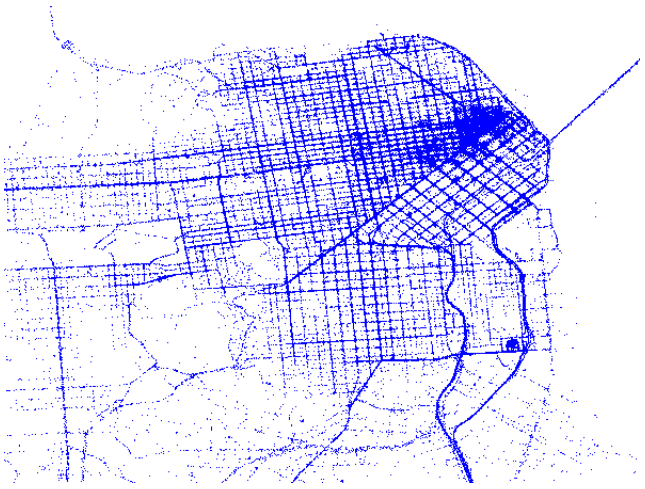

Figure 1. San Francisco taxi measurement locations for a single day, observed at a rate of once per minute.

1. Introduction

Traffic congestion has a significant impact on economic activity throughout much of the world. An essential step towards active congestion control is the creation of accurate, reliable traffic monitoring systems.

Historically, traffic monitoring systems have been mostly limited to highways and have relied on public or private data feeds from a dedicated sensing infrastructure, which often includes loop detectors, radars, video cameras. For highway networks covered by such an infrastructure, it has become common practice to perform both system identification of highway parameters (free flow speed, traffic jam density and flow capacity) and estimation of traffic state (flow, density, length of queues, bulk speed and shockwave location) at a very fine spatio-temporal scale [22, 17]. These highway traffic monitoring approaches heavily rely upon both the ubiquity of data and highway traffic flow models developed over the last half century [15, 4, 20].

For arterials (the secondary network) and highways not covered by dedicated sensing infrastructure, traffic monitoring is substantially more difficult: probe vehicle data is the only significant data source available today with the prospect of global coverage in the future. The features of probe vehicle data today, including the lack of ubiquity and reliability, the variety of data types and specifications, and the randomness of its spatio-temporal coverage, make it insufficient for fully characterizing macroscopic traffic model parameters and doing state estimation with these models for large transportation networks. Figure 1 shows probe measurements from San Francisco taxis for one day, which illustrates the breadth of coverage when aggregating data over longer periods of time. However, this data does not provide enough information to directly infer the macroscopic state of traffic at a fine spatio-temporal scale. Traffic models and data assimilation algorithms must be developed to efficiently transform this data into reliable traffic information. See, e.g., [22, 21, 10, 13] for a discussion on the use of cell phone data for highway traffic monitoring.

Aside from less abundant sensing compared to existing highway traffic monitoring systems, the arterial network presents additional modeling and estimation challenges as

the underlying flow physics which governs them is more complex because of traffic lights (often with unknown cycles), intersections, stop signs, parallel queues, and others. Collecting the detailed parameters of the arterial road network into an accessible electronic database would require the cooperation of numerous government agencies, making this information unreliable and tedious to obtain. Moreover, at the low penetration rate typical for arterial traffic, even small changes in the road network can greatly affect the estimation. This makes the detailed spatio-temporal modeling and estimation approaches developed for highway traffic impractical for arterials—at least until the data volume significantly increases [6, 18].

We propose a statistical approach for arterial traffic estimation from probe vehicle data by modeling the evolution of traffic *states* as a *Coupled Hidden Markov Model* (CHMM), which is a particular form of a probabilistic graphical model. Our approach starts from well established first principle traffic flow models for arterial traffic [15, 7]. We then show how these traffic flow models can be leveraged to estimate historical travel time probability distributions as well as predict the short-term evolution of travel times.

CHMMs have been used for predicting the evolution of sensor readings on highways [14]. The approach in [14] relies on the fact that fixed infrastructure sensors (loop detectors) provide exactly one measurement every 30 seconds at a fixed location. Probe data on arterials is available at random times and random locations, making this model not applicable for our study. Statistical approaches have been proposed that rely on either a single measurement per time interval or aggregated measurements per time interval [9, 6], neither of which is appropriate in our setting. Another probabilistic graphical model approach based on the statistical physics Ising model was proposed in [8]. This model relies on measuring a binary quantity stating whether traffic is congested or uncongested. Transforming probe data into binary congested/uncongested values is a very difficult process by itself and has not been specifically addressed to our knowledge.

Some researchers have examined the case of how to process high-frequency probe data (one measurement approximately every 20 seconds or less) [21]. High-frequency data allows for reliable calculation of short distance speeds and travel times. In this paper, we specifically address the processing of sparse probe data where this level of granularity is not available. Finally, other approaches based on regression [16], optimization [2], neural networks and pattern matching [5] have all been proposed. None of these approaches addresses the issue of processing sparse probe data on a dense arterial network.

The contribution of our work specifically addresses the case of noisy, sparse probe data. In particular, we propose a model and algorithm to do traffic estimation with measurements received at *random locations* and *random times*. We define the travel time distribution of each observation (two consecutive measurements of a vehicle) as a function of (i) the travel time distributions of the links traversed and (ii) the spatial distribution of vehicle locations on each traversed link. The key insight is that, on average, vehicles spend more time traveling through the part of a link just before an intersection than they spend on the part of a link just after an intersection (section 2). Based on traffic modeling assumptions, we construct a graphical model (section 3) representing the travel time distribution of each link at each time interval and their spatio-temporal evolution. Leveraging the results from section 2, the graphical model also represents travel time distributions on any portion of the links of the network (*partial links*) and estimates the probability of an observation given travel time distribution parameters. We develop an expectation maximization (EM) algorithm (section 4) for learning the parameters of our CHMM. We estimate the current state of the network using a particle filter, which is used for predicting the link travel time distributions in the short-term future. Finally, we present the results of a case study (section 5) in San Francisco, for which a fleet of 500 taxis provides sparse location measurements as part of the *Mobile Millennium* project [11]. The initial results indicate that travel time distributions can be accurately estimated using only sparse GPS data.

2. Traffic modeling framework

We present the assumptions and notations of a model of traffic through a signalized intersection. Given these assumptions, we derive travel time distributions between any two points of the network based on the spatial distribution of vehicles along each link of the path. This provides a framework for computing the likelihood of a probe vehicle observation given the parameters of the network.

2.1. Traffic modeling assumptions

To model traffic dynamics, we use the following set of model parameters: maximum density $\rho_{\max}$, critical density ρ_c, free flow speed v_f, cycle time C and red time R. The demand in traffic is represented by the arrival density on a link, ρ_a. These parameters are defined for each link l and each day d. For simplicity of notations, we omit to write the dependency on the link l and the day d.

We denote $\mathbf{N}_n^l$ the spatial neighbors of link l of order n, where first order neighbors ($\mathbf{N}_1^l$) are the links sharing an intersection with link l (including link l). The higher order spatial neighbors are defined by the following recursive formula:

$$\mathbf{N}_{n+1}^l = \bigcup_{j \in \mathbf{N}_n^l} \mathbf{N}_1^j \tag{1}$$

To formulate our model, we make the following assumptions:

1. *Triangular fundamental diagram.*

2. *Stationarity of traffic*: during each estimation interval, the parameters of the light cycles (red and cycle time) do not change. The arrival density ρ_a is constant. The

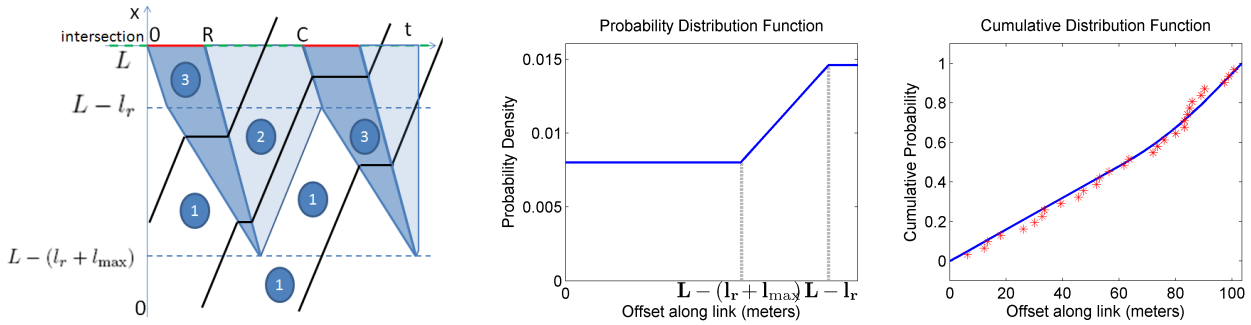

Figure 2. The estimation of the spatial distribution of vehicles on a link as derived from the model of traffic. The space-time plane is used to represent the density of vehicles (left). Using a maximum likelihood estimation, we derive the parameters of the model. The probability density at location x (center) and the estimated and empirical cumulative distribution of the vehicle locations (right) demonstrate that the data fit the model assumptions well. The data used are on one link of the San Francisco network from a single time interval (5:00pm-5:30pm) aggregated over 20 days.

traffic dynamics are stationary: they evolve periodically with period C (length of the light cycle). In particular, there is no consistent increase or decrease in the length of the queue, or instability.

3. *First In First Out* (FIFO) model: overtaking on the road network is neglected.

4. *Discrete congestion states*: for each day d and each time interval t, the traffic conditions on link l are represented by a *discrete* value, $s^l_{d,t}$, which indicates the level of congestion. There are S discrete levels of congestion.

5. *Conditional independence of link travel times*: conditioned on the state $s^l_{d,t}$ of a link l, the travel time distribution of that link is independent from all other traffic variables.

6. *Conditional independence of state transitions*: conditioned on the states of the spatial neighbors of link l of order n (denoted $\mathbf{N}^l_n$) at time t, the state of link l at time $t+1$ is independent from all other current link states, all past link states and all past travel time observations.

2.2. Path travel time probability distribution

As the location measurements are taken uniformly over time, more densely populated areas of the link will have more location measurements. We estimate the probability distribution $\mathscr{P}_X$ of vehicle locations within a link using a statistical model derived from traffic theory and modeling assumptions 1, 2, and 3. For a vehicle traveling from location x_1 to location x_2 on an arterial link l, we assume that the partial travel time Y_{x_1,x_2} is distributed as $\alpha_{x_1,x_2}Y_l$, where Y_l is the random variable of travel time on link l—with realizations denoted y_l—and

$$\alpha_{x_1,x_2} = \int_{x_1}^{x_2} \mathscr{P}_X(x)\,dx. \qquad (2)$$

Note that a baseline approach would assume that α_{x_1,x_2} is the ratio between the distance $|x_2 - x_1|$ and the length of the link, assuming that the travel time on a link is uniformly distributed on the link. Our model takes into account the non spatial uniformity of travel time along a link of the network.

Probability distribution of vehicle locations. On a road segment, at a location x and a time t, the density takes one of the following values: (1) arrival density ρ_a for the vehicles that are upstream of the queue, (2) maximum density $\rho_{\max}$ for the vehicles stopped at time t and location x and (3) critical density ρ_c for the vehicles downstream of the queue—vehicles that have already stopped on their trajectory on link l. These different values of the density at location x and time t are represented in the space-time diagram of trajectories (Figure 2, left). In a stationary regime, we define the triangular queue (from its triangular shape on the space-time diagram of trajectories) as the spatio-temporal region where vehicles stop for the first time on the link. Its length is called the maximum queue length, denoted $l_{\max}$. Under congested conditions, the part of the queue downstream of the triangular queue, called the *remaining queue* with length l_r, corresponds to vehicles which have to stop more than once before going through the intersection. In a stationary regime, we can define the temporal average density at location x, denoted $d(x)$. It is constant upstream of the maximum queue length—equal to ρ_a. It increases linearly until the beginning of the remaining queue l_r. In the remaining queue, the density is constant, equal to ρ_b. The density ρ_b is computed as a convex combination of the maximum density $\rho_{\max}$ and the critical density ρ_c—the density at which the queue discharges. The weights are given by the proportion of the cycle experiencing each of the density, thus the expression for ρ_b, $\rho_b = \frac{R}{C}\rho_{\max} + (1 - \frac{R}{C})\rho_c$.

Remark: The model presented is not restrictive with respect to the presence of a remaining queue, which is only present under congested conditions. During undersaturated conditions, we have $l_r = 0$. In the triangular queue the density increases linearly. The value of the density at the intersection is computed as a convex combination of $\rho_{\max}$, ρ_c and ρ_a where the weights represent the fraction of the cycle dur-

ing which each of the density is observed. The average density at the intersection is $\rho_b = \frac{R}{C}\rho_{\max} + \frac{\tau}{C}\rho_c + (1 - \frac{R+\tau}{C})\rho_a$, where τ is the *clearing time*—time during which the light is green and the queue is dissipating.

The probability distribution of vehicle locations $\mathscr{P}_X$ is proportional to the average density. The normalizing constant Z is the average number of vehicles on the link

$$\mathscr{P}_X(x) = \frac{1}{Z} d(x), \tag{3}$$

with $Z = \int_0^L d(u)\,du$.

After normalization the probability distribution $\mathscr{P}_X(x)$ is equal to

$$\begin{cases} \tilde{\rho}_a, & x \in [0, L-(l_r+l_{\max})] \\ \tilde{\rho}_a + \tilde{\rho}_b \frac{x-(l_r+l_{\max})}{l_{\max}}, & x \in [L-(l_r+l_{\max}), L-l_r] \\ \tilde{\rho}_a + \tilde{\rho}_b, & x \in [L-l_r, L], \end{cases} \tag{4}$$

where $\tilde{\rho}_a = \frac{\rho_a}{Z}$, $Z = \rho_a L + \frac{1}{2} l_{\max}\rho_b + l_r\rho_b$ and $\tilde{\rho}_b = \rho_b/Z = 2\frac{1-\tilde{\rho}_a L}{l_{\max}+2l_r}$. The distribution is fully determined with the three parameters $\tilde{\rho}_a$, l_r and $l_{\max}$.

We estimate the parameters of the distribution of vehicle location on a link $\mathscr{P}_X$ by maximizing the likelihood of the set of location observations (denoted $(x_o)_{o\in O}$). This optimization problem is written in equation (5).

$$\underset{\tilde{\rho}_a, l_r, l_{\max}}{\textbf{argmax}} \sum_{o\in O} \ln(\mathscr{P}_X(x_o)) \quad \textbf{s.t.} \begin{cases} 0 \le \tilde{\rho}_a \le \frac{1}{L} \\ l_r + l_{\max} \le L \\ 0 \le l_r, 0 \le l_{\max} \end{cases} \tag{5}$$

The constraints come from the physics of the problem. The first constraint can be rewritten as $\rho_a \le \frac{Z}{L}$, where $\frac{Z}{L}$ represents the average density on the whole link. It illustrates the fact that the arrival density is inferior to the average density on the link. The other constraints illustrate that the total queue cannot extend the length of the link and that both the triangular queue and the remaining queue must be non-negative.

Experimental results have shown that the model learns the parameters with a relatively small amount of data. An example of the learned and experimental cumulative distributions of vehicle location for a link of the network are represented in figure 2 (right).

3. Modeling framework

Arterial traffic conditions vary over space and time. Given assumptions 4, 5, and 6 in section 2.1, we model the spatio-temporal conditional dependencies of arterial traffic using a probabilistic graphical model known as a *Coupled Hidden Markov Model* (CHMM) [3]. A *Hidden Markov Model* (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved states. CHMMs model systems of multiple interacting processes. In this article, the multiple processes evolving over time are the discrete *states* (assumption 4) of each link in the arterial network. Since we do not observe the state of each link for all times, these processes are considered *hidden*. The travel time distribution on each link is

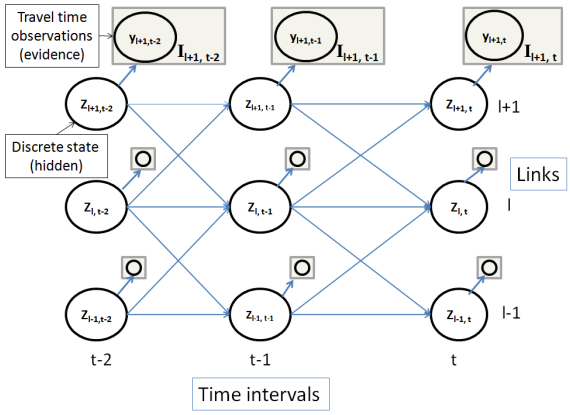

Figure 3. Spatio-temporal model of arterial traffic evolution represented as a coupled hidden Markov model. The circular nodes represent the (hidden) discrete state of traffic for each link at each time interval. The square nodes represent travel time observations from the distribution defined by the traffic state.

conditioned on its hidden state (assumption 5) from which we have sparse observations from probe vehicles traveling through the arterial network. Assumption 6 gives the *coupled* structure to the HMM by specifying local dependencies between adjacent links of the road network. Figure 3 illustrates our model representation of link states and probe vehicle observations. Each circular node in the graph represents the state of a link in the road network. The forward arrows indicate the local spatial dependency of links from one time period to the next. Each square node in the graph represent probe vehicle observations on the link to which it is attached.

The observations are successive GPS measurements of vehicle trajectories (approximately one per minute). The issues of filtering the noise of the GPS to estimate the most likely location of the measurements and inferring the path taken by the vehicle are not addressed in this article. There are multiple approaches to solving this problem including using statistical filtering [12]. In the remainder of this article, we assume that we are given the most likely measurement locations on the road network as well as the most likely path of the vehicle.

To completely specify the CHMM-based model, we have to estimate (i) the initial state probabilities for each link, denoted $\pi_{l,s}$, (ii) the discrete transition probability distribution functions (assumption 6), denoted $A_{l,t}$, and (iii) the distribution of travel time on a link given the state of that link (assumption 5), denoted $g_{l,s,t}$.

For each link l and each time interval t, the probability of link l to be in state s at time $t+1$ given the state of its neighbors at time t is given by the *discrete transition probability distribution* function of link l. It is fully characterized by a matrix of size $S^{\mathbf{N}_n^l} \times S$, denoted $A_{l,t}$. The element of line r and column s, $A_{l,t}(r,s)$, represents the probability of link l to be in state s at time $t+1$ given that the neighbors of l are in state r at time t. Note that the index of the lines

of the matrix represent the different state configurations of the links in $\mathbf{N}_n^l$, where r is the decimal representation of the state configuration (expressed in base S). For each link l, each state s, and each time interval t, the travel time distribution is denoted $g_{l,s,t}$.

A simplifying assumption for computational tractability is to assume that for each link l, the state transition matrix $A_{l,t}$ and the conditional travel time distribution function $g_{l,s,t}$ do not depend on time. They are denoted respectively by A_l and $g_{l,s}$ in the reminder of this article. To relax this assumption, one can assume that these functions are piecewise constant in time and estimate them for each period of time during which the stationarity assumption is satisfied. We also assume that, given the state of a link, the travel time distribution on that link is independent from all the other random variables. In general, travel time distributions across links are not independent (due to light synchronization, platoons, and other factors), although it is a reasonable approximation in many cases. Future work will specifically address the challenge of using correlated distributions, which have the potential to capture more complex dynamics in the arterial road network.

4. Parameter estimation

In this section, we describe how the traffic modeling assumptions lead to an estimation of the parameters of the model and the state variables using the path observations. Given the parameters of the model, we can estimate the most likely state of the links given observations and their evolution over time. Similarly, given the state of the links of the network over a period of time, we can estimate the parameters of the model (state transition matrix, and conditional travel time probability distributions). This well known type of problem is solved using an *Expectation Maximization* (EM) algorithm which iterates between finding the probability of each state for each link of the network and each time interval given some values of the model parameters (E step). Then, the probabilities of each state for each link and each time interval are used to update the value of the parameters by maximizing the log likelihood (M step).

One challenge of our graphical model approach is that we do not observe link travel times directly since the probe observations we receive can span up to 12 links of the network between two consecutive measurements. This difficulty is addressed by computing the most likely link travel times that make up the path of the probe vehicle (*travel time allocation*), which is described in section 4.1. It is possible to have a graphical model representation that does not have this decomposition approach, but it leads to a difficult nonlinear parameter optimization (M-step) problem, for which the number of variables increase quadratically in the number of links. This optimization problem would require an approximation technique to solve, which is why we propose a more intuitive decomposition scheme called *travel time allocation*.

A high-level description of the parameter estimation step is presented in Algorithm 4.3.

4.1. Travel time allocation

An observation consists of a travel time over a path consisting of multiple (partial) links. In order to use the graphical model presented in section 3, the total travel time must be decomposed into a travel time for each (partial) link on the path. This can be achieved by maximizing the log-likelihood of the link travel times for each observation given the model parameters. This optimization problem for a single observation is

$$\underset{y}{\mathbf{argmax}} \left\{ \sum_{l \in P} \ln\Big(\sum_{s=1}^{S} z_l^s g_{l,s}(y_l)\Big) \; : \sum_{l \in P} \alpha_{x_1^l, x_2^l} y_l = \tilde{y} \right\}, \quad (6)$$

where P is the set of links on the path, y_l is the travel time assigned to link l, x_1^l and x_2^l are the start and end location on link l, $\tilde{y}$ is the observed travel time between the GPS measurements, and z_l^s is the probability of link l to be in state s. The values of x_1^l and x_2^l will be equal to the start and end of the link for all intermediate links and will only have non-trivial values for the first and last link of the path (where the actual GPS observations are). The values of z_l^s are obtained from the E-step of the EM algorithm, except in the first iteration where they have been initialized with reasonable values (see Algorithm 4.3). The optimization problem in equation (6) has a number of variables equal to the number of links of the path between consecutive GPS measurements, which is always a relatively small number. This makes the optimization problem easy to solve using numerical methods.

As a reminder, we use the density model of section 2 to compute α_{x_1,x_2}, the proportion of the full link travel time to use.

4.2. E step: Particle filtering

On small networks, it is possible to do exact inference in the CHMM by converting the model to an HMM with a state of dimension number of links. However, the transition matrix is a S^N by S^N matrix (N is the number of links in the network), which is intractable for any reasonable traffic network. Instead, we use an approximation based on particle filtering. Each particle represents an instantiation of the time evolution of the network. Each particle has a weight proportional to the probability of having this instantiation of the state evolution of the network given the available data. We simulate a high number of particles that evolve through the graphical model. These particles are used to estimate the probabilities of the state of each link and each time interval and the probabilities of transition between the state of the neighbors of link l at time $t-1$ and the state of link l at time t. For more information on particle filtering, see, for example [19].

4.3. M step: Update of the parameters

For each link and each state, we assume that the travel time distribution $g_{l,s}$ is parameterized by a set of parameters

$p_{l,s}$ and we note the set of all parameters $\mathbf{P} = (p_{l,s})_{l,s}$. To update these parameters, we maximize the expected complete log-likelihood given the expected values of the probabilities that each link l is in state s at time t and day d ($z^s_{d,t,l}$) and the expected values of link l to be in state s given that the neighbors of link l are in state r at time $t-1$ and day d ($q^{s,r}_{d,t,l}$). We also update the transition matrices A_l and the initial state probabilities π_l for each link of the network, which corresponds to optimizing on the set of parameters $\mathbf{A} = (A_l)_l$ and $\pi = (\pi_{l,s})_{l,s}$.

The expected complete log likelihood reads

$$\begin{aligned}
&\Lambda(Y|\mathbf{z},\mathbf{q},\mathbf{P},\mathbf{A},\pi) = \\
&\sum_{l=1}^{N}\sum_{s=1}^{S}\sum_{d=1}^{D}\sum_{t=1}^{T_d} z^s_{d,t,l}\left(\sum_{i=1}^{I_{d,t,l}} \ln(g_{l,s}(y_i))\right) + \\
&\sum_{l=1}^{N}\sum_{d=1}^{D}\sum_{t=2}^{T_d}\sum_{s=1}^{S}\sum_{r=1}^{S^{N^l_n}} q^{s,r}_{d,t,l}\ln(A_l(r,s)) + \\
&\sum_{l=1}^{N}\sum_{d=1}^{D}\sum_{s=1}^{S} z^s_{d,0,l}\ln(\pi_{l,s}),
\end{aligned} \tag{7}$$

where $I_{d,t,l}$ is the set of travel time observations for day d, time interval t, and link l as provided by the travel time allocation method presented in section 4.1.

The usual optimization problem is modified to take into account the varying number of observations for each link and each time interval. The optimization problem is stated as

$$\max_{\mathbf{P},\mathbf{A}} \quad \Lambda(Y|\mathbf{z},\mathbf{q},\mathbf{P},\mathbf{A},\pi) \ : \begin{cases} \sum_{s=1}^{S} A_l(r,s) = 1, \forall l, r \\ A_l(r,s) \in [0,1], \forall l, r, s \\ \sum_{s=1}^{S} \pi_{l,s} = 1, \forall l \\ \pi_{l,s} \in [0,1], \forall l, s \end{cases} \tag{8}$$

The updates of the transition probabilities A_l and of the initial state probabilities π_l are straightforward. The update of the travel time distributions depends on the type of distribution used in the model. Due to the travel time allocation, the optimization problem on all the parameters $\mathbf{P}$ of the network decouples in $S \times N$ smaller optimization problems, one for each state and link of the network. For state s and link l, the optimization problem is

$$\max_{p_{l,s}} \quad \sum_{d=1}^{D}\sum_{t=1}^{T_d} z^s_{d,t,l}\left(\sum_{i=1}^{I_{d,t,l}} \ln(g_{l,s}(y_i))\right), \tag{9}$$

where $p_{l,s}$ represents the parameters of the travel time distribution $g_{l,s}$. Decoupling the optimization problem makes it highly scalable as each of the optimization subproblems can be performed in parallel. If the travel time allocation method is not used, then the resulting optimization problem is coupled across the whole network resulting in a large non-linear optimization problem that does not scale well.

Algorithm 1 Estimation of the historical distribution of travel time and state transition probability matrices.

Estimate the link parameters for the density model (section 2.2)

Initialize the parameters $P_{l,s}$ of the distributions, the state transition probability matrices A_l, the initial state probabilities $\pi_{l,s}$, and the state probabilities $z^s_{d,t,l}$

EM-algorithm with travel time allocation:
while The algorithm has not converged **do**
  Travel time allocation (section 4.1)
  $y_l \leftarrow$ Allocated travel times given the parameters $P_{l,s}$ and the state probabilities $z^s_{d,t,l}$

  E Step (section 4.2): compute the expected state probabilities $z^s_{d,t,l}$ and transition probabilities $q^{r,s}_{d,t,l}$ given $(y_l)_l$, $(P_{l,s})_{l,s}$ and $(A_l)_l$
  $z^s_{d,t,l} \leftarrow E(z^s_{d,t,l}|y_l, P_{l,s}, A_l)$
  $q^{r,s}_{d,t,l} \leftarrow E(q^{r,s}_{d,t,l}|y_l, P_{l,s}, A_l)$

  M Step (section 4.3): maximize the expected complete log-likelihood, given the state probabilities $z^s_{d,t,l}$ and the transition probabilities $q^{r,s}_{d,t,l}$.
  $(P_{l,s}, A_l, \pi_l) \leftarrow \mathbf{argmax}_{\mathbf{P},\mathbf{A},\pi}\Lambda(Y|\mathbf{z},\mathbf{q},\mathbf{P},\mathbf{A},\pi)$
end while

4.4. Real-time estimation and forecast

Estimating and forecasting traffic conditions in real-time can be achieved after the travel time distributions and transition probabilities have been learned. We use the graphical model with its learned parameters to perform inference using data up to the time the estimate or forecast is produced. This is done by running the particle filter (E-step only) to determine which state of traffic is most likely for each link and time interval. Forecast is done by propagating the particle filter forward from the current time interval (with no additional data).

5. Experiments

We tested our arterial traffic forecasting method using probe data from a fleet of about 500 taxis in San Francisco as provided to us by the Cabspotting project [1]. Each taxi provides a measurement of its location approximately once every minute (generally between 50 and 70 seconds). In addition to its location, the taxi also reports whether or not it is carrying a customer or not. This information allows us to filter out the points when a taxi is loading or unloading a passenger. This data is sent to the *Mobile Millennium* traffic server, where it is processed and visualized in real-time.

In our case study, we used data from November 25, 2009 through February 27, 2010, focusing on weekdays from 3pm-8pm in the subnetwork of San Francisco depicted in figure 4. This subnetwork contains 322 links (where a link

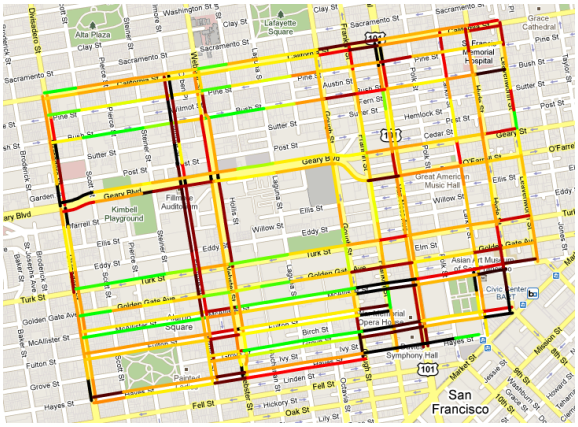

Figure 4. Real-time traffic estimation for a subnetwork of San Francisco. The color scale represents the estimated travel time divided by the speed limit travel time. Green is for values close to 1 (travel time is about the same as speed limit) and black indicates values around 5.

is defined as the road between two signals) and has an average of 600 observations per half hour time interval. We use 30 minutes (half an hour) as the time interval in the graphical model presented in section 3. We assume that the observation probability distribution functions g (section 3) are independent Gaussians. In general, the choice of a Gaussian distribution restricts the flexibility of the model to capture unique traffic characteristics, but it is also far more tractable to solve in practice. Finding tractable approximation methods for solving the problem using traffic theory inspired travel time distributions will be the subject of future work.

Our approach requires a training period (section 4) before it can be used to make predictions in real-time. We used data from November 25, 2009 through February 19, 2010 as our training period. We only used Tuesdays, Wednesdays and Thursdays to train our model, which totalled 18 training days (after removing holidays and days with system malfunctions that prevented data collection). We then tested the model by running it over all Tuesdays, Wednesdays and Thursdays between February 20, 2010 and February 27, 2010, which totalled 3 days.

We first learn the traffic density parameters (section 2) for each hour of the day from 3pm to 8pm, where each hour period is assumed to have its own characteristics in terms of the average density on a link. We then run the EM algorithm (section 4) over all the training data, with the assumption that the transition matrix $\mathbf{A}$ and the Gaussian distributions for each link are stationary over the study period. Once the parameters have been learned through the EM algorithm, we use a particle filter to compute the most likely state of each link given real-time data on a test day. Figure 4 shows a map of the subnetwork of San Francisco with each link colored according to its level of congestion, defined as the mean travel time divided by a reference free flow travel time. The free flow travel time is computed as the travel time experienced when traveling at the speed limit and accounting for an expected delay (due to traffic signals) under light traffic conditions.

Model	RMSE (sec)	MPE
Graphical (with density)	46	30.1%
Graphical (without density)	50	34.3%
Baseline	63	44.4%

Table 1. Experimental results comparison between the proposed graphical model and the baseline model.

To quantify the validity of our estimates, we compare the actual travel time of an observed path to the estimate obtained by summing over the mean travel time for all links of the path. Table 1 shows the root mean squared error (RMSE) and mean percentage error (MPE) of our travel time predictions as compared to a baseline approach. The baseline approach computes the average speed for each observation and assigns it to each link along its path. Then all of the speeds on each link are averaged to give a historical average speed for each link. The real-time version of this approach does the same thing and then takes a weighted average between the historical and the real time speed to give a speed estimate for each link of the network, which can be used to estimate travel times. The two versions of the graphical model show the effect of using the density model of section 2.2 to compute partial link travel times instead of simply using a travel time proportional to the partial link distance.

These results were computed on the data obtained between February 20 and February 27, 2010. The data was split into two sets, one for computing the real-time traffic estimates and one for computing the error metrics. This was done to ensure an unbiased comparison of the proposed graphical model and the baseline model. Approximately 70% of the data was used for computing the real-time traffic estimates with the other 30% used for computing the error metrics.

6. Conclusion and discussion

In this article, we proposed a new probabilistic modeling framework for estimating arterial traffic conditions from sparse probe data. Our initial results suggest that this approach outperforms the baseline approach in predicting short-distance arterial travel times by 36.9% in terms of the root mean squared error metric. We believe that the proposed modeling approach provides a fundamental basis for estimating arterial traffic conditions. The key features that our model possesses are:

1. Each link has a discrete traffic state that cannot be directly observed.
2. Traffic states of nearby links are correlated and evolve over time in a Markov manner (i.e. the future is independent of the past given the present).
3. Expectation maximization provides the right framework for learning the transition and observation model

parameters.

There are numerous ways in which our model can be extended to take into account a wider variety of traffic features. These enhancements include:

1. Traffic-specific travel time distributions instead of independent Gaussians.
2. Traffic-specific meanings for the discrete states of each link instead of just undersaturated/congested.
3. Direct calculation of the E-step and M-step in the EM algorithm using the path travel times instead of relying on the travel time allocation step.
4. Relate link travel time distributions to route travel time distributions as estimating short-distance travel times are of less interest than longer trips through city network.

Each of the listed items are part of ongoing research and we expect that these enhancements to the basic model will result in a much richer model capable of giving precise route travel time distributions. The ability to reliably estimate route travel time distributions will be a valuable tool for commuters, fleets, and public agencies.

Acknowledgement

Special thanks to Timothy Hunter for help with data preprocessing and path inference of the taxi data. Additional thanks to the Mobile Millennium team for their technical support.

References

[1] Cabspotting. `http://www.cabspotting.org`.

[2] S.J. Agbolosu-Amison, B. Park, and Ilsoo Yun. Comparative evaluation of heuristic optimization methods in urban arterial network optimization. In *Intelligent Transportation Systems, 2009. ITSC '09. 12th International IEEE Conference on*, pages 1–6, 2009.

[3] Matthew Brand. Coupled hidden Markov models for modeling interacting processes. Technical report, The Media Lab, Massachusetts Institute of Technology, 1997.

[4] C. Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research B*, 28(4):269–287, 1994.

[5] C. de Fabritiis, R. Ragona, and G. Valenti. Traffic estimation and prediction based on real time floating car data. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 197–203, 2008.

[6] C. Furtlehner, J. M Lasgouttes, and A. De La Fortelle. A belief propagation approach to traffic prediction using probe vehicles. In *Proc. IEEE 10th Int. Conf. Intel. Trans. Sys*, pages 1022–1027, 2007.

[7] N. Geroliminis and C.F. Daganzo. Macroscopic modeling of traffic in cities. *86th Annual Meeting Transportation Research Board, Washington D.C.*, 2007.

[8] Bruce Hellinga, Pedram Izadpanah, Hiroyuki Takada, and Liping Fu. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C: Emerging Technologies*, 16(6):768 – 782, 2008.

[9] R. Herring, A. Hofleitner, S. Amin, T. Abou Nasr, A. Abdel Khalek, P. Abbeel, and A. Bayen. Using mobile phones to forecast arterial traffic through statistical learning. In *89th Annual Meeting Transportation Research Board*, Washington D.C, 2010.

[10] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005.

[11] The *Mobile Millennium* Project. `http://traffic.berkeley.edu`.

[12] T. Hunter, R. Herring, A. Hofleitner, A. Bayen, and P. Abbeel. Trajectory reconstruction of noisy gps probe vehicles in arterial traffic. *IEEE Transactions on Inteligent Transportation Systems*. To be submitted.

[13] A. Krause, E. Horvitz, A. Kansal, and F. Zhao. Toward community sensing. In *Proceedings of ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, St. Louis, MO, April 2008.

[14] Jaimyoung Kwon and Kevin Murphy. Modeling freeway traffic with coupled hmms. Technical report, University of California, Berkeley, 2000.

[15] M. J. Lighthill and G. B. Whitham. On kinematic waves. II. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 229(1178):317–345, May 1955.

[16] Xinyu Min, Jianming Hu, Qi Chen, Tongshuai Zhang, and Yi Zhang. Short-term traffic flow forecasting of urban network based on dynamic STARIMA model. In *Intelligent Transportation Systems, 2009. ITSC '09. 12th International IEEE Conference on*, pages 1–6, 2009.

[17] J. Kwon J. Rice E. Van Zwet P. J. Bickel, C. Chen and P. Varaiya. Measuring traffic. *Statistical Science*, 22(4):581–597, 2007.

[18] T. Park and S. Lee. A Bayesian approach for estimating link travel time on urban arterial road network. In *Computational Science and Its Applications ICCSA 2004*, pages 1017–1025. Perugia, Italy, May 2004.

[19] S. Russell and P. Norvig. *Artificial Intelligence - A Modern Approach*. Prentice-Hall, Inc, Englewood Cliffs, NJ, 1995.

[20] X. Sun, L. Munoz, and R. Horowitz. Mixture Kalman filter based highway congestion mode and vehicle density estimator and its application. In *Proceedings of the 2004 American Control Conference*, pages 2098–2103, Boston, MA, 2004.

[21] A. Thiagarajan, L. R. Sivalingam, K. LaCurts, S. Toledo, J. Eriksson, S. Madden, and H. Balakrishnan. VTrack: Accurate, Energy-Aware Traffic Delay Estimation Using Mobile Phones. In *7th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Berkeley, CA, November 2009.

[22] D. Work, S. Blandin, O.-P. Tossavainen, B. Piccoli, and A. Bayen. A distributed highway velocity model for traffic state reconstruction. *In press, Applied Research Mathematics eXpress (ARMX)*, 2010.