

Learning the Dynamics of Arterial Traffic From Probe Data Using a Dynamic Bayesian Network

Aude Hofleitner, Ryan Herring, Pieter Abbeel, and Alexandre Bayen

Abstract—Estimating and predicting traffic conditions in arterial networks using probe data has proven to be a substantial challenge. Sparse probe data represent the vast majority of the data available on arterial roads. This paper proposes a probabilistic modeling framework for estimating and predicting arterial travel-time distributions using sparsely observed probe vehicles. We introduce a model based on hydrodynamic traffic theory to learn the density of vehicles on arterial road segments, illustrating the distribution of delay within a road segment. The characterization of this distribution is essentially to use probe vehicles for traffic estimation: Probe vehicles report their location at random locations, and the travel times between location reports must be properly scaled to match the map discretization. A dynamic Bayesian network represents the spatiotemporal dependence on the network and provides a flexible framework to learn traffic dynamics from historical data and to perform real-time estimation with streaming data. The model is evaluated using data from a fleet of 500 probe vehicles in San Francisco, CA, which send Global Positioning System (GPS) data to our server every minute. The numerical experiments analyze the learning and estimation capabilities on a subnetwork with more than 800 links. The sampling rate of the probe vehicles does not provide detailed information about the location where vehicles encountered delay or the reason for any delay (i.e., signal delay, congestion delay, etc.). The model provides an increase in estimation accuracy of 35% when compared with a baseline approach to process probe-vehicle data.

Index Terms—Expectation-maximization algorithms, probes, queuing analysis, real-time systems, statistical learning.

I. INTRODUCTION AND BACKGROUND

TRAFFIC congestion has a significant impact on economic activity. An essential step toward active congestion control is the creation of accurate reliable traffic monitoring systems, leveraging the latest advances in technology and research. Historically, traffic monitoring systems have been mostly limited



Fig. 1. Probe measurements in San Francisco, CA. The small (resp. large) dots represent the measurement of the location of a taxi, received between midnight and 7 A.M. (resp. at 7 A.M.).

to highways and have relied on data feeds from a dedicated sensing infrastructure (loop detectors, radars, etc.). For highway networks covered by such an infrastructure, it is common practice to perform both system identification (free-flow speed, traffic jam density, and flow capacity) and estimation of traffic state (flow, density, bulk speed, and shockwave location) at a very fine spatiotemporal scale [5], [46]. These highway traffic monitoring approaches rely upon both the ubiquity of data and highway traffic flow models developed over the last half century [11], [38].

For arterials, traffic monitoring is substantially more difficult: Probe-vehicle data are the only significant data source available today with the prospect of global coverage in the future. It comes from various sources such as fleet vehicles periodically reporting their location (a rate of 1 min is currently the standard), smartphones, aftermarket devices, or radio-frequency identification tags. The features of probe-vehicle data today, including the lack of ubiquity and a uniform penetration rate (the percentage of vehicle reporting their location varies across the network and throughout the day), the variety of data types and specifications, and the randomness of its spatiotemporal coverage, make it insufficient for fully characterizing macroscopic traffic model parameters and doing state estimation for large transportation networks. Fig. 1 shows the probe measurements collected on a day from midnight to 7 A.M. by one of the feeds of the *Mobile Millennium* system. It also shows a snapshot of the location of probes at 7 A.M. The figure shows both the breadth of coverage when aggregating data over long periods of time and the limited information available at a given

Manuscript received September 21, 2011; revised February 9, 2012; accepted May 11, 2012. This work was supported in part by the Federal and California Departments of Transportation, by Nokia, by NAVTEQ, and by Cabspotting. The Associate Editor for this paper was S. Sun.

A. Hofleitner is with Partners for Advanced Transportation Technology, University of California, Berkeley, Berkeley, CA 94720-3830 USA (e-mail: aude@eecs.berkeley.edu).

R. Herring was with the California Center for Innovative Transportation, University of California, Berkeley, Berkeley, CA 94720-3830 USA. He is now with Apple Inc., Cupertino CA 95014 USA (e-mail: ryanherring@engineeralum.berkeley.edu).

P. Abbeel and A. Bayen are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720-3830 USA (e-mail: pabbeel@cs.berkeley.edu; bayen@berkeley.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TITS.2012.2200474

point in time, limiting the direct estimation capabilities at a fine spatiotemporal scale. Traffic models and data assimilation algorithms must be developed to efficiently transform these data into reliable traffic information (see, e.g., [28], [35], [44], and [46] for a discussion on the use of cell phone data for highway traffic monitoring).

Aside from less abundant sensing compared with existing highway traffic monitoring systems, the arterial network presents additional modeling and estimation challenges: The underlying flow physics that governs them is more complex because of traffic lights (often with unknown cycles), intersections, and others [8], [37]. 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. This makes the detailed spatiotemporal modeling and estimation approaches developed for highway traffic impractical for arterials, i.e., at least until the data volume significantly increases [3], [42], [45].

We are interested in a statistical approach to arterial traffic estimation based on *Dynamic Bayesian Networks* (DBNs). The model characterizes the variability of travel times among vehicles traveling on the network and the stochasticity of congestion dynamics on arterial networks. Our approach is adapted to the only significant data source available today for arterial estimation: sparse measurements from probe-vehicle data. We present a brief overview of existing research focused on statistical approaches for traffic estimation and underline the contributions of this paper. An extended review of the literature is available in [20]. Previous research has studied the estimation and short-term prediction of sensor readings using DBNs [36], [43] and regression models [40]. These articles assume that sensors (such as loop detectors) provide measurements with a fixed frequency at fixed locations. Probe data on arterials are available at *random times* and *random locations*, making this assumption not applicable for this paper. Other approaches [16], [22] assume that either a single measurement per time interval or aggregated measurements per time interval are available for each road segment of the network (according to the map discretization). This assumption limits the capacity to represent the variability of travel times among the vehicles traveling on the network. Moreover, such approaches are not adapted to missing data, when no information is available on some parts of the network. An approach inspired from the Ising model was developed in [16]. It relies on binary measurements stating whether traffic is congested or uncongested that are not directly available from traffic sensors. Transforming traffic data into binary congested/uncongested values is a difficult process by itself and has not been specifically addressed in the literature to our knowledge. Our model offers such a binary quantization from probe-vehicle travel times. Neural networks and pattern matching [12] have been used to estimate traffic from Global Positioning System (GPS) data under the critical assumption that the velocity is spatially homogeneous and similar among drivers. This assumption does not take into account the variability of travel times due to the frequent stops at traffic signals. High-frequency probe data (one measurement approximately every 20 s or less) [29] allow 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. Numerous estimation algorithms from probe-vehicle data rely on the decomposition of path travel times to individual road segments, also referred to as links [19], [23]. However, when a vehicle travels more than one link, the location of its delay is unknown, and this decomposition can lead to inaccuracies. In this paper, we use probe-vehicle data without the use of a travel-time decomposition algorithm.

The contribution of this paper specifically addresses the estimation and short-term forecast of the *probability distribution function* (pdf) of travel times in the case of noisy sparse probe data. In particular, we propose a model and an algorithm for traffic estimation with measurements received at *random locations* and *random times*, which are based on the learning of the dynamics of congestion on the network. Each observation, which is defined as two consecutive GPS measurements including the travel time between these measurements, has a probability density that depends on 1) the pdf of travel times of the links traversed and 2) the spatial distribution of vehicles on each traversed link. The key insight is that, on average, vehicles are more likely to experience delay close to intersections because of the presence of traffic signals. We assume that the pdf of travel times on each link of the network depends on the level of congestion (*congestion state*) of this link, and we model and learn the dynamics of congestion on the network using a DBN. We define a link as the road segment between signalized intersections; however, this choice of discretization can be more fine if desired.

This paper is organized as follows. Section II presents a graphical model representing the dependence between the travel-time observations and the congestion state of each link at each time interval and their spatiotemporal evolution. Section III uses queuing theory to formalize the intuition that vehicles are more likely to experience delays close to intersections. We discuss how this information can be used to compute the pdf of travel times on an arbitrary path from the pdf of travel times of the links traversed. Leveraging the modeling assumptions of Section II and the results from Section III, the DBN represents the probabilistic dynamics of traffic congestion and the probabilistic observation model of the congestion states from probe data. We develop an *expectation-maximization* (EM) algorithm (see Section IV) for learning the parameters of the DBN. We perform the *expectation step* (E step) using particle filtering and solve a large convex optimization problem using an interior point method for the *maximization step* (M step). After the historical learning of the parameters of the system's dynamics, we estimate the current state of the network and predict the probability of congestion and the pdf of link travel times from the probe data available in real time. Finally, we present the results of a case study (see Section V) in San Francisco, CA, for which a fleet of 500 probe vehicles provides sparse location measurements [1]. These data are one of the feeds available in the Mobile Millennium system [4]. The system provides real-time streaming data and a history of the data collected since October 2009. The initial results indicate that travel-time distributions can be accurately estimated using sparse GPS data only.

II. TRAFFIC MODELING ASSUMPTIONS

A. Dynamical Model

Arterial traffic can be viewed as a dynamic stochastic process. Our model represents the main characteristics of traffic dynamics while making assumptions necessary for the tractability of the estimation process. The validity and limitation of the model are further discussed in Section VI, where we also analyze how the model can be refined or generalized.

- 1) *Time discretization*: We model traffic as a discrete-time dynamical system. We call Δ the time discretization, which is chosen depending on the data available and the desired temporal scale of the estimation. This paper is focused on estimating travel-time distributions when measurements are sparse. We choose Δ to be equal to 5 min in the numerical experiments as we are interested in estimating trends rather than fluctuations. For $t \in \mathcal{T} = \{0, \dots, (T-1)\}$, time interval t is given by $[t_0 + t\Delta, t_0 + (t+1)\Delta]$.
- 2) *Characterization of the state of traffic*: For each link $i \in I$ (I is the set of link indices), traffic conditions are characterized by a discrete random variable (RV) $\xi^{i,t}$. We denote by $s^{i,t} \in \{0, \dots, S-1\}$ the realization of the RV $\xi^{i,t}$, representing a discrete congestion state. We choose a binary representation of traffic states ($S=2$), characterizing an *undersaturated* and a *congested* state. The derivations are easily generalized to a finer discretization of the number of states.
- 3) *Dynamical model*: Transitions between time intervals model information propagation on the road network by taking into account the spatiotemporal dependence of the state of the links. We denote by $\xi^{I,t}$ (with realization $s^{I,t} \in \{0, \dots, S-1\}^{|I|}$) the state of the network at time interval t . We call π_i the links adjacent to link i , including link i . We have $i' \in \pi_i \Leftrightarrow i' = i$, or i' and i have a common intersection. The equation of the dynamics is given by $\xi^{i,t} = f_d(\xi^{\pi_i, t-1}) + \epsilon_d^i \forall i \in I$, where ϵ_d^i represents the state noise of the dynamical model for link i . The dynamic equation can also be defined by a set of conditional independence assumptions:¹ $\xi^{i,t} \perp\!\!\!\perp \xi^{i',t'} | \xi^{\pi_i, t-1}$ for $(t', i') \in X(i, t)$, where $X(i, t) = \{t-1\} \times I \setminus \pi_i \cup \{0, \dots, t-2\} \times I$, and $A \setminus B$ denotes the set A without the elements of B . The mathematical formulation expresses that, given the state of the neighbors π_i at $t-1$, the state of link i at t is independent of the state of nonneighboring links at $t-1$ and is independent of the state of all links of the network at time intervals prior to $t-1$.
- 4) *Observation model*: The system is observed through noisy point to point travel-time measurements. A map-matching and path-inference algorithm [30] reconstructs the path of the vehicle between successive location reports and filters out the GPS noise. The map-matching algorithm provides the family of links $j(k)$ traversed between the k th pair of successive location reports and

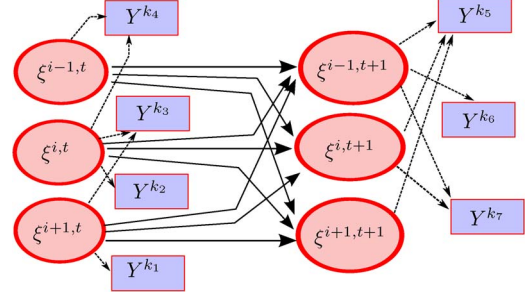


Fig. 2. Two-slice temporal Bayesian network representation of the model of arterial traffic dynamics. The circular nodes represent the (hidden) traffic states for each link at each time interval. The square nodes represent travel-time observations. There is an edge from the state of link i at time t to the state of link i' at time $t+1$ if i is a neighbor of i' ($i \in \pi_{i'}$). Observation Y_k , received at time t , represents the travel time of a probe vehicle on its path, defined by the set of traversed links $j(k)$ and the distances $x_{s,k}$ and $x_{e,k}$ to the downstream intersections on the first and last links of the path. There is an edge from the state of each link in $j(k)$ to Y_k .

the distances $x_{s,k}$ and $x_{e,k}$ to the downstream intersection of the first and last link traversed. Note that the path of the probe vehicle between consecutive location reports is fully specified by $x_{s,k}$, $x_{e,k}$, and $j(k)$. The travel time between $(x_{s,k}, x_{e,k})$ is an RV Y_k , with realization $y_k \in \mathbb{R}$. The observation equation is given by $Y_k = f_o(\xi^{j(k),t}, x_{s,k}, x_{e,k}) + \epsilon_o^Y(\xi^{j(k),t}, x_{s,k}, x_{e,k})$, where ϵ_o^Y represents the observation noise, that may depend on the state of the links of the path and the distance traveled on each of these links. We assume that the observation noise is a sum of independent RV representing the observation noise on each link of the path. The travel time on a path is then a sum of independent RVs representing the travel time on each link of the path. The measurements come from a small subset of vehicles traveling on the network and periodically sending their location in real time. Measurements from the past are stored and accessible in real time. The *Mobile Millennium* system, which has been developed by the University of California Berkeley (UC Berkeley) and Nokia [4], provides such data (see Fig. 1).

B. DBN Representation

The conditional independence introduced by the dynamic and observation equations are represented with a DBN [13]. DBNs are directed graphical models that represent the complex interdependence between the hidden state variables $\xi^{i,t}$ and the observations Y_k . The graphical structure specifies the conditional independence and provides a compact parameterization of the model. The model structure does not change over time, which means that the structure can be fully specified by a *two-slice temporal Bayesian network* (2TBN). It is common to assume that the parameters of the 2TBN do not change, i.e., the model is time invariant. The structure of the DBN induced by our assumptions on the dynamic and observation equations is shown in Fig. 2. The model is fully specified by the following conditional distributions.

- The transition probabilities: For each link i , we consider the conditional probability that $\xi^{i,t}$ has the realization

¹For sets of RVs A , B , and C , we denote by $A \perp\!\!\!\perp B|C$ that the assertion “ A is conditionally independent of B given C .”

$s^{i,t}$, given the state of its neighbors at the previous time interval $t - 1$. The state of a link at time t may depend on the state of its neighbors in any arbitrary way. Given that both the number of states and the number of neighbors are finite, the conditional probability is represented by matrix A^i where, for each row m , $A^i(m, 1)$ (resp. $A^i(m, 2)$) represents the probability of being congested (resp. undersaturated) given the state m of the neighbors, so that $A^i(m, 2) = 1 - A^i(m, 1)$. One possible choice for A^i is to consider all the possible state combinations of the neighbors, as done in [21], but the dimension of A^i grows exponentially with the number of neighbors, and the number of parameters to estimate does not reflect the amount of data available. We consider a more scalable model in which the state of link i at time interval t depends on the total number of undersaturated links among neighbors. With this model, there are $|\pi_i| + 1$ parameters to estimate for each link i , where $|\pi_i|$ is the cardinality of π_i . A wide variety of functions of the congestion indices of the neighbors can be used to predict the state of the link at the next time interval. Choosing the appropriate function of the congestion indices is called feature selection [18] and is not detailed in this paper. We experiment a few other choices for this function in the numerical analysis.

- The observation conditional probabilities: For each link i and each state s , we define the pdf of travel times on link i given the state s . We consider that, conditioned on the state s , the travel times on each link i are normally distributed and parameterized by a mean $\mu^{i,s}$ and a standard deviation $\sigma^{i,s}$. The normality assumption is not necessary for the derivations in the model but improves the computational efficiency, as discussed in Sections IV and VI. The k th travel-time measurement y_k is specified by the set of traversed links $j^{(k)}$ and the distance to the downstream intersection on the first and last links ($x_{s,k}$ and $x_{e,k}$, respectively). Given the state of the traversed links, the travel time on this path is normally distributed and denoted $f(y_k | s^{j^{(k)},t}, x_{s,k}, x_{e,k})$. The mean and the variance are the sum of the mean and of the variance of travel times on the (partial) links of the path, respectively. Note that probe vehicles may not report their location at the beginning or at the end of a link and that we need to properly scale the travel time on the fraction of link traversed (partial link), as detailed in Section III.
- The initial state probabilities: For each link i , we call $c^i(1)$ (resp. $c^i(2)$) the probability that link i is congested (resp. undersaturated) during the first time interval and have $c^i(2) = 1 - c^i(1)$.

The specification of the conditional distributions leads to the following decomposition of the joint probability of the model:

$$p(s, y | \theta) \prod_{\substack{t \in \mathcal{T} \setminus \{t_0\} \\ i \in \mathcal{I}}} A(\eta^{i,t-1}, s^{i,t}) \prod_{\substack{t \in \mathcal{T} \\ k \in \mathcal{K}(t)}} f(y_k | s^{j^{(k)},t}) \prod_{i \in \mathcal{I}} c^i(s^{i,0})$$

where $\eta^{i,t-1}$ represents the congestion state of the neighbors of link i at time interval $t - 1$.

C. Modeling Partial Link Travel Time Through Density Estimation

Since probe vehicles send their positions at any location on the network, the path can start and end at any location. The first and last links of the corresponding path are not fully traversed by the vehicle (*partial* links). In addition to the pdf of travel times on each link of the network, we need to define the pdf of travel times on partial links, i.e., the pdf of travel times on link i between any offsets x_1 and x_2 (where x_m , with $m = 1, 2$, represents the distance to the downstream intersection). We call Y_{x_1, x_2}^i as the RV representing the travel time on *partial* link i between offsets x_1 and x_2 ($x_1 \geq x_2$); then, $Y_{L^i, 0}^i$ represents the travel time on link i (between offsets L^i , length of link i , and 0). We assume that there exists $\alpha^i(x_1, x_2)$ such that $Y_{x_1, x_2}^i = \alpha^i(x_1, x_2) Y_{L^i, 0}^i$. The function α^i must satisfy the following conditions.

- The travel time on a partial link is a fraction of the link travel time: $\forall (x_1, x_2) \in [0, L]^2$, $\alpha^i(x_1, x_2) \in [0, 1]$. If the partial link spans the entire link, the partial travel time has the same distribution as the link travel time: $\alpha^i(L^i, 0) = 1$.
- If a partial link is included in another partial link, its travel time should be smaller: $\forall x_1, x_2 \mapsto \alpha^i(x_1, x_2)$ is a decreasing function of x_2 , and $\forall x_2, x_1 \mapsto \alpha^i(x_1, x_2)$ is an increasing function of x_1 .
- The probability for a vehicle to experience delay increases as the location gets closer to the downstream intersection. For the same distance traveled, travel times are longer close to the downstream intersection because of the presence of traffic signals. $\forall x_1, x_2 \mapsto \alpha^i(x_1, x_2)$ is a convex function of x_2 . Similarly, $\forall x_2, x_1 \mapsto \alpha^i(x_1, x_2)$ is a concave function of x_1 .

The function defined by $\alpha^i(x_1, x_2) = (x_1 - x_2)/L^i$ satisfies these conditions. However, it assumes that the travel time on a partial link is proportional to the distance traveled on the link but does not take into account the presence of traffic signals. In Section III, we derive a parametric model for α^i from a hydrodynamic model of traffic flow and learn the parameters from the sparse measurements of probe-vehicle locations: α^i is the *cumulative distribution function* (cdf) of a specific RV. For a probe vehicle sampled uniformly in time and reporting its position while traveling on link i , the RV represents the position of the vehicle on the link as it reports its location, and we denote by f_X its pdf. Because of the presence of traffic signals, f is a decreasing function of the distance to the downstream intersection (increasing function of the distance from the upstream intersection). We choose $\alpha^i(x_1, x_2) = \int_{x_2}^{x_1} f_X(x) dx$, which satisfies all the given assumptions.

III. MODELING THE SPATIAL DISTRIBUTION OF VEHICLES ON AN ARTERIAL LINK

Probe vehicles send periodic location measurements, which provide two sources of indirect information about the arterial traffic link parameters. First, as the location measurements are taken uniformly over time, more densely populated areas of the link will have more location measurements. Second, the time spent between two consecutive location measurements provides

information on the speed at which the vehicle drove through the corresponding arterial link(s).

We use the first source of information to define how the travel time on a partial link is related to the travel time on the entire link and derive the function $\alpha^i(\cdot, \cdot)$ introduced in Section II. We consider link i during time interval t . For notational simplicity, we omit the dependence on i and t .

A. Arterial Traffic Flow Model

We model vehicular flow as a continuum and represent it with macroscopic variables of *flow* $q(x, t)$ (veh/s), *density* $\rho(x, t)$ (veh/m), and *velocity* $v(x, t)$ (m/s). The definition of flow gives the following relation between these three variables: $q(x, t) = \rho(x, t) v(x, t)$. We make the assumption of a triangular fundamental diagram (FD) parameterized by v_f , which is the free flow speed (m/s); ρ_{\max} , which is the jam (maximum) density (veh/m); and q_{\max} , which is the capacity (veh/m). We define the critical density $\rho_c = q_{\max}/v_f$ and have the following expression for the FD:

$$q(x, t) = \begin{cases} v_f \rho(x, t), & \text{if } \rho(x, t) \in [0, \rho_c] \\ q_{\max} \frac{\rho_{\max} - \rho(x, t)}{\rho_{\max} - \rho_c}, & \text{if } \rho(x, t) \in [\rho_c, \rho_{\max}]. \end{cases}$$

We assume that the characteristics of the traffic light (red time R and cycle time C) and the arrival rate q_a remain constant during the time interval, leading to a periodic formation and dissolution of the queues [26]. We define two discrete traffic regimes, i.e., *undersaturated* and *congested*, depending on the presence (resp. the absence) of a remaining queue when the signal turns red.

Undersaturated regime: The queue fully dissipates within the green time. The queue is defined as the spatiotemporal region where vehicles are stopped on the link and called the *triangular queue*, with length l_{\max} .

Congested regime: There exists a part of the queue downstream of the triangular queue called *remaining queue* with length l_r corresponding to vehicles that have to stop multiple times before going through the intersection. All notations introduced up to here are shown for both regimes in Fig. 3.

B. Probability Distribution of Vehicle Locations

According to the assumptions, the density at location x is time periodic with period C . The density $d(x)$ at location x is the temporal average of the density $\rho(x, t)$ at location x and time t : $d(x) = 1/C \int_0^C \rho(x, t) dt$.

In practice, flow is never perfectly periodic, but we will assume that the given averaging over a duration C is a good proxy of a longer average. According to the assumptions, the density at location x and time t takes one of the three following values, numbered 1 to 3 for convenience: 1) $\rho_1 = \rho_{\max}$, when vehicles are stopped; 2) $\rho_2 = \rho_c$ when vehicles are dissipating from a queue; and 3) $\rho_3 = \rho_a$ when vehicles have not yet stopped in the queue. The average density at location x is $d(x) = \sum_{i=1}^3 \beta_i(x) \rho_i$, where $\beta_i(x)$ represents the fraction of the cycle time C during which density is equal to ρ_i at location x .

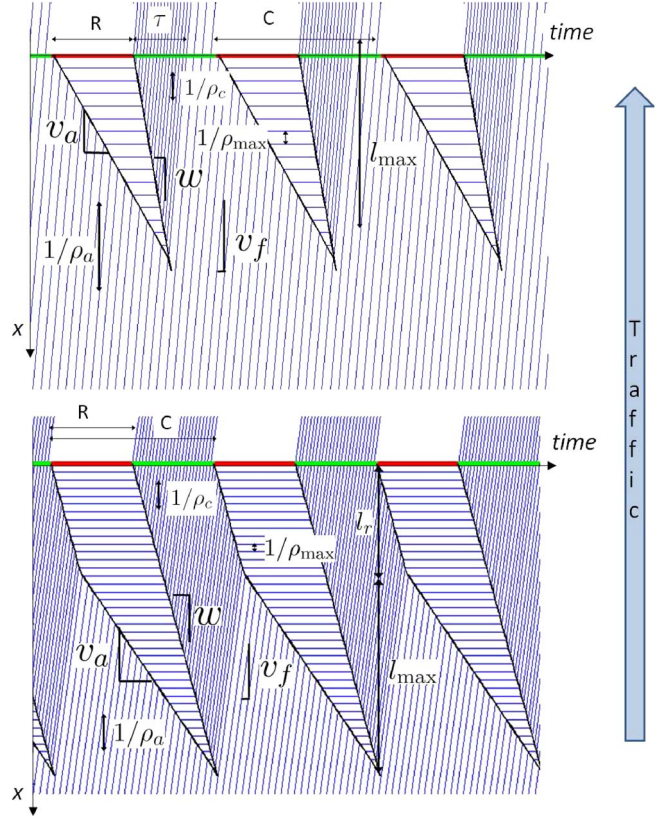


Fig. 3. Space-time diagram of vehicle trajectories with uniform arrivals under (top) an undersaturated traffic regime and (bottom) a congested traffic regime.

When vehicles are sampled uniformly in time, the pdf $f_X(x)$ of observing a vehicle at location x is proportional to the average density $d(x)$ at location x , with the proportionality constant given by $Z = \int_0^L d(x) dx$ so that $f_X(x) = d(x)/Z$.

1) *Undersaturated Regime:* Upstream of the maximum queue length, the density is equal to ρ_a throughout the entire cycle. Using the assumption that the FD is triangular and that the arrival density is constant, the average density linearly increases from ρ_a at $x = l_{\max}$ to the value it takes at the intersection, where $x = 0$. At the intersection, the density is ρ_{\max} for R seconds, when the light is red. The density is ρ_c when the queue dissipates, i.e., during the clearing time τ . The rest of the cycle has density ρ_a . The average density at the intersection is the sum of the arrival, maximum, and critical densities, which are weighted by the fraction of the cycle during which each of the densities is experienced. The average density at the intersection is $d(0) = 1/C(R\rho_{\max} + \tau\rho_c + (C - (R + \tau))\rho_a)$. From traffic theory, we have $\tau = R(\rho_a/\rho_c - \rho_a)$ [25]; thus, $d(0) = R/C\rho_{\max} + \rho_a$. The density at location x is given by

$$\begin{cases} d(x) = \rho_a, & \text{if } x \geq l_{\max} \\ d(x) = \rho_a + \frac{R}{C}\rho_{\max} \frac{l_{\max} - x}{l_{\max}}, & \text{if } x \leq l_{\max}, \end{cases}$$

i.e., $d(x) = \rho_a + R/C\rho_{\max} \max(l_{\max} - x, 0)/l_{\max}$. The constant $Z_u = \int_0^L d(x) dx$ is the temporal average of the number of vehicles on the link and is given by $Z_u = L\rho_a +$

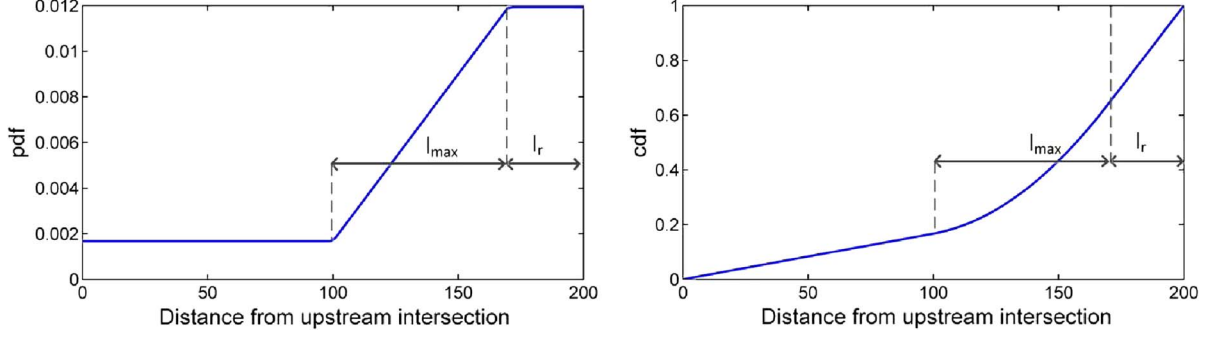


Fig. 4. (Left) Probability distribution of vehicle locations on the link as a function of the distance from the upstream intersection. (Right) Cumulative distribution.

$l_{\max}/2R/C\rho_{\max}$. The pdf of vehicle location $f_X^u(x) = d(x)/Z_u$ is given by

$$f_X^u(x) = \frac{\rho_a}{Z_u} + \frac{R}{CZ_u}\rho_{\max} \frac{\max(l_{\max} - x, 0)}{l_{\max}}.$$

2) *Congested Regime*: In the congested regime, the average density is constant upstream of the maximum queue length, i.e., equal to ρ_a , and increases linearly until the remaining queue. In the remaining queue, it is constant and equal to $R/C\rho_{\max} + (1 - R/C)\rho_c$. The probability distribution of vehicle location is given by

$$\begin{aligned} d(x) &= \rho_a, & \text{if } x \geq l_{\max} + l_r \\ d(x) &= \rho_a + d^{\#} \frac{-x + l_{\max} + l_r}{l_{\max}}, & \text{if } x \in [l_r, l_{\max} + l_r], \\ d(x) &= \frac{R}{C}\rho_{\max} + \left(1 - \frac{R}{C}\right)\rho_c, & \text{if } x \leq l_r \end{aligned} \quad (1)$$

where $d^{\#} = (R/C\rho_{\max} + (1 - R/C)\rho_c - \rho_a)$. We denote by Z_c the normalizing constant that ensures that the integral of the function $x \mapsto d(x)/Z_c$ on $[0, L]$ equals one as follows:

$$Z_c = L\rho_a + \left(\frac{l_{\max}}{2} + l_r\right) d^{\#}. \quad (2)$$

The undersaturated regime is a special case of the congested regime, in which the remaining queue length l_r is equal to zero. We consider the congested regime as the general case for the spatial distribution of vehicle location. The distribution is fully determined by three independent parameters, i.e., the remaining queue length l_r , the triangular queue length l_{\max} , and the normalized arrival density $\tilde{\rho}_a = \rho_a/Z_c$, as shown in Fig. 4 (left: pdf; right: cdf), and reads as follows:

$$\begin{aligned} f_X^c(x) &= \tilde{\rho}_a, & \text{if } x \geq l_{\max} + l_r \\ f_X^c(x) &= \tilde{\rho}_a + \frac{(l_r + l_{\max}) - x}{l_{\max}} \Delta_{\tilde{\rho}}, & \text{if } x \in [l_r, l_{\max} + l_r], \\ f_X^c(x) &= \tilde{\rho}_a + \Delta_{\tilde{\rho}}, & \text{if } x \leq l_r \end{aligned}$$

with $\Delta_{\tilde{\rho}} = \frac{1 - \tilde{\rho}_a L}{l_{\max}/2 + l_r}$.

The expression of $\Delta_{\tilde{\rho}}$ is obtained by noticing that $\int_0^L f_X^c(x) dx = 1$ or by direct computation from (1), by replacing Z_c by its expression (2) and ρ_a by $Z_c \tilde{\rho}_a$.

C. Density Estimation

We estimate the parameters of the distribution f_X by maximizing the likelihood of the set of location observations (denoted $(x_o)_{o \in O}$) provided by large amounts of historical data:

$$\underset{\tilde{\rho}_a, l_r, l_{\max}}{\text{maximize}} \sum_{o \in O} \ln(f_X(x_o)) \text{ s.t. } \begin{cases} 0 \leq \tilde{\rho}_a \leq \frac{1}{L} \\ l_r + l_{\max} \leq L \\ 0 \leq l_r, 0 < l_{\max}. \end{cases}$$

The constraints come from the physics of the problem. The first constraint is equivalent to $\rho_a \leq Z_c/L$, where Z_c/L is the average density on the entire 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 beyond the length of the link and that the triangular queue and the remaining queue must be nonnegative. The constraints on the queue lengths do not limit the generality of the model. Under spillover conditions (queue length extending beyond the upstream intersection), we consider that the queue length extends up to the upstream intersection; the rest of the queue is accounted for in the upstream links. Tighter bounds on the parameters can be added to ensure the physical interpretation of the results.

The objective function is not concave in the optimization variables. However, the search space is limited (three bounded parameters), and we perform a grid search followed by a local gradient ascent for the B best solutions of the grid search. We used 15 points in each dimension and chose $B = 10$ in the numerical experiments as a finer grid did not provide better results.

IV. TRAVEL-TIME ESTIMATION ALGORITHM

We observe a complex pattern of dependence among the travel times sent by the probe vehicles and want to learn the stochastic dependence between these observations to perform estimation and prediction on the arterial network. Rather than directly modeling the dependence between the observations, we introduce the variables $(\xi^{i,t})$, representing the discrete state of each link at each time interval, and model their dynamic evolution. They are introduced to exploit the underlying structure of the dynamical system and to simplify the estimation task. The variables $(\xi^{i,t})$ are called *latent* or *hidden* variables because they are not directly observed. Conditioned on $(\xi^{i,t})$, the observation variables are independent. The parameter

estimation problem would be simplified if we could directly observe the state variables $(\xi^{i,t})$. Without observing $(\xi^{i,t})$, the likelihood function is a marginal probability, obtained by summing (or integrating in the continuous case) over the latent variables. Marginalization couples the parameters and obscures the underlying structure of the likelihood function. The EM algorithm learns the dependence among the observations while exploiting the structure of the stochastic dynamic evolution [14]. Given the parameters of the dynamics and the observation model, we can compute the probability distribution of the latent variables. Similarly, given the probability distribution of the latent variables, we can compute the parameters that best explain the observed data. Section IV-A provides a mathematical justification for the use of this iterative algorithm. We detail the steps of the algorithm, i.e., the E step in Section IV-B and the M step in Section IV-C in the case of traffic estimation.

A. Introduction on the EM Algorithm

The EM algorithm allows us to exploit the underlying structure of the dynamical model, although the latent variables are not observed. It is an iterative algorithm consisting in two steps.

- *The E step* computes the joint probability distribution of the latent variables given the observed variables and the current values of the parameters. In the case of a DBN, this step corresponds to a smoothing step, in which, at each time t , we estimate the joint probability distribution of the state variables $(\xi^{i,t})_{i,t}$. In practice, the smoothing step is replaced by a filtering step for efficiency. The filtering step only uses observations received up to (and including) time t to compute the joint probability distribution of the state variables $(\xi^{i,t})_i$.
- *The M step* optimizes the parameters based on the estimation of the joint probability distribution of the latent variables. This step has the same complexity as if the latent variables were observed.

Let \mathcal{Y} denote the observable RV, with realization y (travel time from the probe vehicles) and ξ as the latent variables, with realization s (congestion state of the links of the network). Let θ be the set of unknown parameters, i.e., $\theta = \{(\mu^{i,s}, \sigma^{i,s}), i \in I, s \in \{0, \dots, S-1\}\} \cup \{A^i, i \in I\}$. The log likelihood of the data is the log of the marginal probability of the observations given the parameters, as given in

$$l(\theta; y) = \ln(p(y|\theta)) = \ln\left(\sum_s p(s, y|\theta)\right).$$

If ξ was observed, the maximum likelihood estimation would amount to maximizing $l_c(\theta; y, s) = \ln(p(y, s|\theta))$, which is referred to as the *complete log likelihood*, because it corresponds to the log probability of the complete set of RVs for a given value of the parameter θ . Given that ξ is in fact not observed, the complete log likelihood is a random quantity and cannot be directly maximized. Given a distribution, which is denoted as $q(s|y)$, we define a deterministic function of θ , which is denoted as $\langle l_c(\theta; y, s) \rangle_q$ and called the *expected complete log likelihood*. It corresponds to the average of the complete log

likelihood, over the realizations of ξ , when $q(s|y)$ is chosen as the averaging distribution, i.e.,

$$\langle l_c(\theta; y, s) \rangle_q = \sum_s q(s|y) \ln(p(y, s|\theta)).$$

We use Jensen's inequality and have

$$\begin{aligned} l(\theta; y) &\geq \sum_s q(s|y) \ln\left(\frac{p(s|y, \theta)}{q(s|y)}\right) \\ &= \sum_s q(s|y) \ln p(s, y|\theta) - \sum_s q(s|y) \ln q(s|y). \end{aligned}$$

To maximize $l(\theta; y)$, we iteratively maximize the right-hand side by 1) maximizing on the proposal distribution $q(s|y)$ given the current value of the parameter (E step) and 2) maximizing on the parameter θ given the proposal distribution (M step).

B. E Step

In the Bayesian approach to dynamic state estimation, one attempts to construct the probability distribution of the state (*belief state*) at time interval t based on all available measurements up to and including time interval t , which is known as the *posterior* distribution. The process of estimating the pdf of the state of the network at time interval t conditioned on the observations up to time t is called *filtering*. The E step of the EM algorithm technically requires *smoothing*, i.e., estimating the pdf of the state conditioned on all the observations available for the experiment. However, the smoothing is typically replaced by a filtering step for efficiency. Filtering consists of essentially two stages: *prediction* and *update*. The prediction uses the transition probabilities to compute the belief state from one measurement to the next. The update operation uses the latest available measurements to modify the state probability distribution using the Bayes rule.

The DBN used in this paper is a multiply connected belief network (at least one pair of variables has more than one undirected path connecting them), in which probabilistic inference is NP-hard [10]. In such networks, algorithms performing probabilistic inference have a time complexity that, in the worst case, is exponential in the number of hidden variables in the network. We need approximation algorithms to perform probabilistic inference. Algorithms such as Monte Carlo simulation [41], variational methods [31], and belief state simplification [7] are commonly used to approximate probabilistic inference. We investigate a Monte Carlo simulation approach (particle filter) described in the following.

Particle filtering is an approximation of a recursive Bayesian filter algorithm using Monte Carlo simulations, which has successfully been implemented for highway traffic estimation [9]. The belief state is represented by a set of random samples with associated weights (importance weights) such that, as the number of samples increases, the approximation tends to the true belief state. We simulate V particles ($V = 2000$ in the numerical experiments). Each particle v represents an instantiation of the *time evolution of the traffic state* of the network, i.e., a possible succession of traffic states for each

link and each time interval. A particle v at time t is represented by a vector of the states of each link and each time interval $(s_v^{i,t'})_{i \in I, t' \in \{0 \dots t\}}$. At t , each particle has a weight ω_v^t proportional to the probability of having this instantiation of the state evolution given the available data up to time t . The particles explore the possible state space and represent the belief state of the DBN. At time t , the spatiotemporal instantiations $s_v^t = (s_v^{i,t'})_{i \in I, t' \in \{0 \dots t\}}$ of the particles and their associated importance weight ω_v^t form an approximation $p_V(s^{1:t}|y^{1:t}, \theta)$ of the joint probability distribution $p(s^{1:t}|y^{1:t}, \theta)$ of the state of the links up to time t . According to this approximation, the probability of observing a state $s = (s^{i,t})_{i \in I, t \in \{0 \dots T\}}$ on the network throughout its time evolution is

$$\begin{aligned} p(s|y, \theta) &\approx p_V(s|y, \theta) \\ &= \sum_{v=1}^V \omega_v^T \mathbf{1}_s(s_v) \end{aligned}$$

where $\mathbf{1}_s(s_v)$ is equal to 1 if the particle has the state instantiation s and 0 if otherwise. In particular, we define the following *sufficient statistics* (SS):

- The path SS is the joint distribution of the states of the links $j(k)$, on path k , conditioned on the observations received up to time interval t . It is denoted as $p_V(s^{j(k),t}|y^{1:t}, \theta)$ and is computed by summing the weights of all the particles for which the links $j(k)$ are in state $s^{j(k),t} \in S^{|j(k)|}$ at t :

$$p_V(s^{j(k),t}|y^{1:t}, \theta) = \sum_{v=1}^V \omega_v \mathbf{1}_{s^{j(k),t}}(s_v^{j(k),t}).$$

- The link SS is the probability of the state of link i at time t , conditioned on the state of the neighbors π_i at time interval $t-1$ and the observations received up to t . It is denoted as $p_V(s^{i,t}|s^{\pi_i,t-1}, y^{1:t}, \theta)$ and is computed by summing the weights of all the particles for which link i is in state $s^{i,t} \in S$ at t and for which the neighbors of link i are in state $s^{\pi_i,t-1} \in S^{|\pi_i|}$ at $t-1$. To compute the conditional probability, this sum is normalized by the sum of the weights of the particles for which the neighbors of link i are in state $s^{\pi_i,t-1}$.

For each link and each time interval, the number of SS to compute is exponential in the number of neighbors of the link. We present a model that overcomes this computational cost by assuming that the state of a link at time t depends on the total number of undersaturated neighbors at $t-1$, defined by $\eta^{i,t-1} = \sum_{i' \in \pi_i} s^{i',t-1}$. The number of SS to compute for link i is $|\pi_i| + 1$ for each time interval, which significantly limits the complexity. Other functions could be used to compactly represent the state of the neighbors. We experiment with a few other choices in the numerical experiments. These functions do not need to be linear nor 1-D. The SS $p_V(s^{i,t}|\eta^{i,t-1}, y, \theta)$ are similarly computed as for $p_V(s^{i,t}|s^{\pi_i,t-1}, y, \theta)$: We sum the weights of the particles for which link i is in state $s^{i,t}$ at time interval t and for which the sum of the congestion of

the neighbors is $\eta^{i,t-1}$ at time interval $t-1$ and normalize as follows:

$$p_V(s^{i,t}|\eta^{i,t-1}, y^{1:t}, \theta) = \frac{\sum_{v=1}^V \omega_v^t \mathbf{1}_{s^{i,t}, \eta^{i,t-1}}(s_v^{i,t}, \eta_v^{i,t-1})}{Z(\eta^{i,t-1})}.$$

The constant $Z(\eta^{i,t-1}) = p_V(\eta^{i,t-1}|y^{1:t}, \theta)$ is computed from the particles or, with less computational cost, by summing the joint probabilities $p_V(s^{i,t}, \eta^{i,t-1})$ over the possible states of link i at time t .

Using these SS, the expected complete log likelihood $\langle l_c(\theta; y, s) \rangle_{p_V}$ is given by

$$\begin{aligned} &\sum_{\substack{t \in \mathcal{T} \setminus \{0\} \\ s^{i,t} \in S}} \sum_{\eta^{i,t-1}} p_V(s^{i,t}|\eta^{i,t-1}, y^{1:t}, \theta) \ln(A(\eta^{i,t-1}, s^{i,t})) \\ &+ \sum_{\substack{t \in \mathcal{T} \\ k \in K(t)}} \sum_{s^{j(k),t}} p_V(s^{j(k),t}|y^{1:t}, \theta) \ln(f(y_k|s^{j(k),t}, \theta)) \end{aligned}$$

where $s^{j(k),t} \in \{0, \dots, S-1\}^{|j(k)|}$, $\eta^{i,t-1} \in \{0, \dots, |\pi_i|\}$, $K(t)$ is the set of paths from probe vehicles received during time interval t , and $f(y_k|s^{j(k),t}, \theta)$ is the density of probability of the travel time y_k on the links of the path $j(k)$ that are in state $s^{j(k),t}$. The mean and variance of travel times are computed by summing the mean and variance travel times of the (partial) links of the path. We recall that the mean and variance of travel times on *partial* link i are scaled according to the function α^i . In the first sum, we remove 0 from the set \mathcal{T} since there is no transition prior to t_0 . To compute the SS, the filtering step is performed with the particles as follows:

- *Update at t* : Compute the *posterior* distribution using the measurements of time interval t . For each particle, ω_v^t is multiplied by the probability of each measurement given the states $\xi_v^{i,t}$ of the particle. The weights are normalized so that they sum to 1.
- *Prediction at $t+1$* : Predict the state distribution for time interval $t+1$ using the transition probabilities. For each link i and each particle v , we sample the state $\xi_v^{i,t+1}$ given the states $\xi_v^{\pi_i,t}$ (or any function of the states such as the sum of the congestion states) of its neighbors at time t according to the transition probabilities, i.e., the state $s^{i,t+1}$ is chosen with probability $A(s^{i,t+1}|\xi_v^{\pi_i,t})$.

This algorithm is known as a *sequential importance sampling* (SIS) particle filter [2]. A common problem with the SIS particle filter is the *degeneracy phenomenon* [15], [17]. After a few iterations, all but one particle have negligible weights. A large computational effort is devoted to updating particles whose contribution to the posterior distribution is almost zero. To reduce the effects of degeneracy, we *resample* the particles after the update step. The modified algorithm is known as *sequential importance resampling* or *sampling importance resampling*. To resample the particles, V particles are successively chosen randomly (with replacement) with a probability equal to its weight (the weights sum to 1). The new set of particles all have a weight equal to $1/V$ and is used to compute the *a priori* probability distribution of the states at time interval $t+1$.

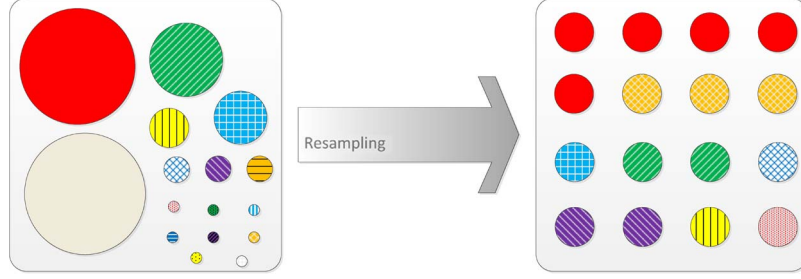


Fig. 5. Resampling algorithm: Each particle is represented by a circle with a diameter proportional to its weight. Each particle is chosen with a probability proportional to its weight, put in the *new* set of particles with weight $1/V$, and then replaced. This process is repeated V times. The intuition is that particles with a large weight are likely to be chosen several times, whereas particles with a small weight might not be present after the resampling step.

C. M Step: Update of the Model Parameters

The M step maximizes the expected complete log likelihood with respect to θ , representing the parameters of both the dynamics (transition probability matrices A^i , $i \in I$) and the observations (parameters of the travel-time distributions, which are conditioned on the state of the link). Given the structure of the complete log likelihood, this optimization can be independently performed for each transition probability matrix A^i and for the parameters of the joint Gaussian distribution. Note that because travel-time observations may span several links, the estimation of the travel-time distribution couples all the links of the network.

- The transition probability matrices are updated by maximizing with respect to the entries of A^i under the constraint that A^i is a stochastic matrix (all the lines have nonnegative entries and sum to 1). For the line j representing the transition probability when the neighbors are in state $m \in \{0, \dots, S-1\}^{|\pi_i|}$, we have

$$A^i(m, s) \propto \sum_{t \in \mathcal{T} \setminus \{0\}} p_V(s^{i,t} = s | s^{\pi_i, t-1} = m, y^{1:t}, \theta)$$

where the proportionality constant is computed for all m such that $\sum_s A^i(m, s) = 1$. A similar expression is obtained if the transitions depend on any functions of the states, such as the number of undersaturated neighbors.

- Given the discrete state of link i at time interval t , the travel time on link i , i.e., $Y^{i,t}$, is normally distributed. Remember that the pdf of a partial travel time is computed from the pdf of a link travel time using the scaling function $\alpha^i(\cdot, \cdot)$ shown in Section III, although the dependence does not explicitly appear for notational simplicity. The travel times are independent RV: Given the state s of the network at time t , $Y^{I,t}$ is a multivariate Gaussian variable with mean $\mu^s = (\mu^{i,s^i} : i \in I)$ and covariance $\Sigma^s = \text{diag}((\sigma^{i,s^i})^2 : i \in I)$, where s^i is the i th coordinate of s and represents the state of link i . The M step updates the mean $\mu = (\mu^{i,s} : i \in I, s \in \{0, \dots, S-1\})$. We also use the notation $\Sigma = \text{diag}((\sigma^{i,s})^2 : i \in I, s \in \{0, \dots, S-1\})$. It is the solution of the following optimization problem:

$$\begin{aligned} \underset{\mu \in \mathbb{R}^{|S| \times |I|}}{\text{minimize}} \quad & \sum_{k \in K(t)} \sum_{t \in \mathcal{T}} p_V(s^{j(k),t} | y, \theta) \\ & \times (y_k - \mu^{s^{j(k),t}})^T (\Sigma^{s^{j(k),t}})^{-1} (y_k - \mu^{s^{j(k),t}}). \end{aligned}$$

Given that $\Sigma^{s^{j(k),t}}$ is positive definite for all k , the objective function is convex in μ . However, the objective function is not jointly convex in μ and Σ , and we only optimize on μ ; the variances are estimated once at the beginning of the algorithm using a Gaussian mixture with two components. The number of variables grows linearly with the number of links. We add constraints to limit the feasible set to physically relevant values and implement an interior point algorithm [6].

Algorithm 1 Maximum likelihood estimation of the parameters of the dynamic and observation models.

Initialize the parameters: $(\mu^{i,s}, \sigma^{i,s})_{i,s}, (A^i)_i$.

EM algorithm for parameter estimation in DBN

while The algorithm has not converged **do**

 E Step (Section IV-B)

 Initialize the E Step: Simulate samples with weight $\omega_v = 1/V$ representing the state of the network at the initial time given the initial state probabilities.

for $t \in \mathcal{T}$ **do**

 Update: For each travel-time observation, multiply the weight of each particle with the probability of the observation given the state of the particle: $\omega_v \leftarrow \omega_v \prod_k f_{Y_k}(y_k | \xi_v^{j(k),t})$
 Normalize: divide the weight of each particle by the sum of the weights.

 Resample the particles to avoid degeneracy (see Fig. 5 and details in [2], [33]).

 Predict: For each link i and each particle v , sample the state at time interval $t+1$ using the transition probabilities A^i .

end for

 M Step (Section IV-C)

 Update the transition probabilities $A^i, i \in I$.

 Update the parameters of the observation model.

end while

V. EXPERIMENTS

The model formalizes an intuitive representation of the propagation of congestion throughout the network. This paper proposes a learning algorithm of the dynamics of traffic on a network and a real-time estimation framework. Our numerical results are organized as follows. First, we validate the use of the model shown in Section III, which is denoted as the *density*

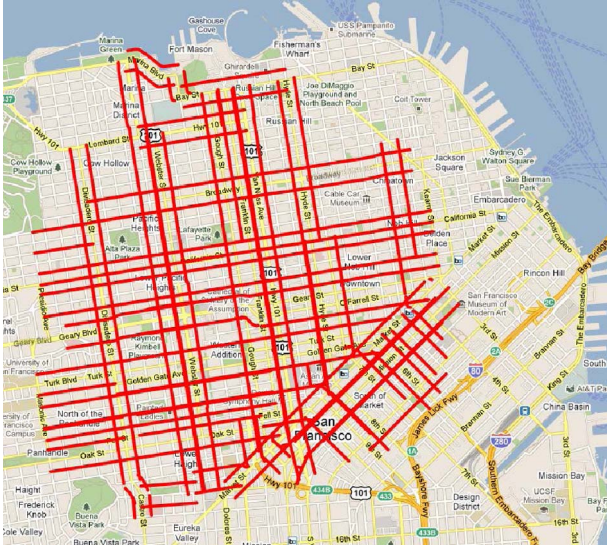


Fig. 6. Subnetwork of San Francisco used for validation.

model, to derive temporal averages of the probability distribution of vehicle locations and use it for scaling partial travel time. Second, we validate the DBN presented in this paper. Using cross-validation, we test the estimation and prediction accuracy of the model for different time horizons. We compare our results to a baseline model and investigate how the use of the density model improves the results. We also validate the pdf of travel times learned by the model. The estimation of the travel-time distribution (rather than mean values only) is crucial in arterial networks to accurately describe the variability of travel times.

A. Validation of the Density Model

Experimental setup: We use data collected by one of the feeds of the *Mobile Millennium* system: a fleet of 500 vehicles reporting their location every minute in San Francisco, CA. This paper focuses on a subnetwork of San Francisco (see Fig. 6) with 815 links and 527 intersections (more than 12.6 km of roadway). A *historical interval* is a tuple consisting of a day of the week, a start time, and an end time. For each historical interval and each link, we aggregate the locations reported by the vehicles and learn the parameters of the density model. We focus our numerical results on 15-min intervals representing Tuesdays from 4 P.M. to 8 P.M., i.e., (Tuesday, 4 P.M., 4:15 P.M.), ..., (Tuesday, 7:45 P.M., 8 P.M.).

Description of the statistical test of the model: For each link and each historical interval, we use the *Kolmogorov–Smirnov* (K–S) statistics to test if the locations of the probe vehicles are distributed according to the density model [39]. The K–S statistic is computed as the maximum difference between the empirical and the hypothetical cdf (density model). In contrast with other tests (e.g., T-test that tests uniquely the mean or the chi-squared test that assumes that the data is normally distributed), the K–S test is a standard nonparametric test to state whether samples are distributed according to a hypothetical distribution. We use this test to accept (or reject) the null hypothesis H_0 : “The measurements

TABLE I
OUTCOME OF STATISTICAL TESTS

Decision		True hypothesis	
		H_0 is true	H_0 is false
	Accepts H_0	Right decision	Wrong decision, Type II error, rate β
	Rejects H_0	Wrong decision, Type I error, rate α	Right decision

TABLE II
PERCENTAGE OF POSITIVE K–S TESTS FOR DIFFERENT VALUES OF α AND THE TWO HYPOTHESIS (DENSITY MODEL OR UNIFORM DISTRIBUTION)

Distribution of vehicles	α			Mean p-value
	0.1	0.05	0.01	
Density model	0.75	0.80	0.89	0.35
Uniform	0.46	0.55	0.67	0.15

of probe vehicles are distributed according to the density model.” When performing a statistical test, four situations described in Table I arise. The performance of a statistical test is defined by its *statistical significance* ($1 - \alpha$, where α is the probability to reject H_0 when it is actually true) and *statistical power* ($1 - \beta$, where β is the probability to accept H_0 when it is actually false). The p-value is used to decide if we accept or reject the null hypothesis H_0 . Low p-values indicate that the data do not follow the proposed distribution. We reject hypothesis H_0 at the α significance level if the p-value is smaller than α and accept it otherwise. The parameter α is commonly set to values ranging from 0.001 to 0.1 and often set to $\alpha = 0.05$. Smaller levels of α increase confidence in the determination of significance but increase the risk of Type II errors, and therefore have less *statistical power*. The K–S test has a probability of Type II error β that tends to zero as the number of samples tends to infinity. Since the number of samples is finite, we maintain the *power* of the statistical test by 1) not testing links that do not have enough measurements, 2) experimenting with different levels of significance, and 3) reporting the p-value for each decision.

We want to validate the capability of the density model to properly scale travel time on portions of arterial links. In particular, we want to show that vehicles are not uniformly distributed along the link since they are more likely to experience delay close to the downstream intersection. To illustrate this reasoning, we also perform the K–S test with a null hypothesis being that measurements are uniformly distributed along the link. We compare the results of the test on both hypotheses in Table II.

The results indicate that, for a majority of arterial links, the average location of vehicles is an RV that follows the density model. The spatial distribution of vehicle location is better represented by the density model than by a uniform distribution. A graphical representation of the data provides valuable qualitative information: For different links of the network, we represent the cumulative locations reported by the vehicles.² We

²The cumulative locations are computed as follows: 1) We order the locations reported by the probe vehicles; and 2) we plot the points $(x_i, i/N)$ for $i = 1 \dots N$, where N is the number of locations collected for the link and historic interval, and x_i is the i th location on the link (in meters from the upstream intersection).

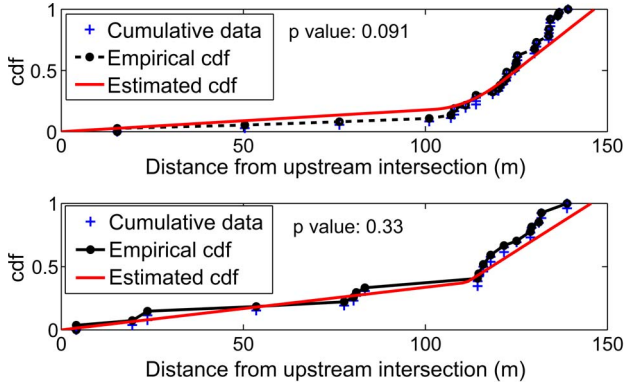


Fig. 7. Empirical and proposal cdf of the average vehicle locations. (Top) Link with a p-value equal to 0.09. The model predicts a sharp increase in the density of measurements toward the downstream extremity of the link, but no measurements are received on the last 15 m of the link. The digital map does not model the width of the road or the intersection, which might explain the absence of measurements on the last 15 m. (Bottom) Link with p-value equal to 0.33. The model learns the characteristics of the distribution of vehicle locations. We read an estimate of the historical queue length (around 30 m) that provides information on the average congestion of the link.

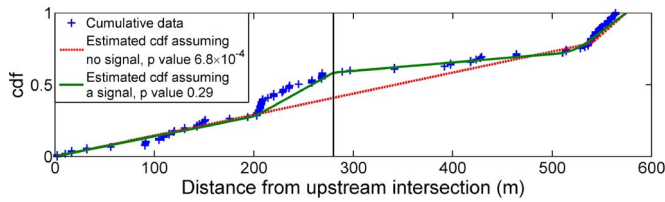


Fig. 8. Detecting signal locations using the average spatial distribution of vehicles. The figure shows an example of a very low p-value for a link of the network. Analyzing the results, we realized that a signal was missing in the database, explaining the poor fit of the model.

also represent the empirical (Kaplan–Meier) cdf [32] and the proposed cdf. In Fig. 7, we show the cumulative distributions obtained for two links of the network during the first historical interval. The first link shows a good qualitative fit. However, the p-value is only 0.091. The map discretization does not take into account the width of intersections and may be the reason why no measurements are received on the last 15 m of the link. The second link has an average p-value. In both cases, the data follows the sharp increase in the density of measurements close to the downstream intersection, as predicted by the model because of the presence of a traffic signal. The model also provides an estimate of the historical queue length on each link of the network that can be used for planning and network congestion analysis.

The analysis of the links with low p-values is also informative and valuable. Fig. 8 shows the result for a link with a p-value equal to 6.8×10^{-4} . We expect sharp increases in the density of measurements to occur upstream of traffic signals. The map database, provided to us by NAVTEQ, contains attributes of the transportation network, such as road characteristics, presence of traffic lights, and so on. On this link, the cumulative distribution of vehicle location exhibits two important increases, whereas only one signal was present in the map database.

Analyzing the location of the link in *Google Street View*, we confirmed that there was a signal that was not in the database. With the corrected information, we updated the proposed dis-

tribution and obtained a p-value equal to 0.29. A potential application of the algorithm is the automatic detection of traffic signals from probe data [24] but is not developed in this paper. Other sources of poor fitting are due to specific behaviors of the taxi, such as waiting in front of major hotels, which can be filtered, when considering successive locations of a taxi.

B. Validation of the Dynamic Bayesian Modeling

As probe vehicles report their location periodically in time, the duration between two successive location reports x_s and x_e represents an observation of the travel time of the vehicle on its path from x_s to x_e , i.e., the realizations y_k of the RVs Y_k . A map-matching and path-inference algorithm [30] that combines models of GPS emissions and of drivers' behavior into a conditional random field, filters out GPS noise, maps the GPS measurements to the road network, and reconstructs the most likely set of links traversed by the vehicle.

In our case study, we focus on learning the model parameters on Tuesdays from 4 P.M. to 8 P.M. in the subnetwork of San Francisco depicted in Fig. 6. We use 5 min as the time discretization Δ in the graphical model shown in Section II. We assume that, conditioned on the state of the links of the network, the travel times are independent Gaussian variables. The choice of a Gaussian distribution may restrict the flexibility of the model to capture unique traffic characteristics, but it is more computationally efficient in practice. In particular, the model relies on travel times from probe vehicles that typically traverse several links between successive observations. The travel time on the path is a sum of independent RVs, and its pdf is computed as the convolution of the pdf of the link travel times on the path. If the link travel times are normally distributed, the computation of the convolution is straightforward, whereas it requires numerical algorithms, if otherwise. Finding tractable approximation methods for using traffic-theory-inspired travel-time distributions is the subject of ongoing work [26]. We use the density model to compute the pdf of partial link travel times from the pdf of link travel times.

To validate the use of the DBN, we assess the estimation and prediction accuracy of the model. In traffic estimation (or prediction), access to ground-truth data is rare as it requires the monitoring of each vehicle on the entire network for the duration of the estimation. Instead, cross-validation [34] is commonly used in the machine learning community to assess how the results of a statistical model generalize to an independent data set, which is not used to develop the model but assumed to follow the same model. For each time interval, we randomly partition the available data (travel-time measurements of the probe vehicles) into complementary subsets. We learn the parameter of the model on one subset (training set) and validate the performance of the model on the other subset (validation or testing set). The training set constitutes 70% of the available data, and the remaining 30% is used for validation.

Estimation and prediction errors: We compare the travel times predicted by the model with the travel times reported by the probe vehicles and compute the average l_1 error. Given a set of observations $y_k, k \in K(t)$ received at time interval t and corresponding estimates (predictions) \hat{y}_k , the average l_p error

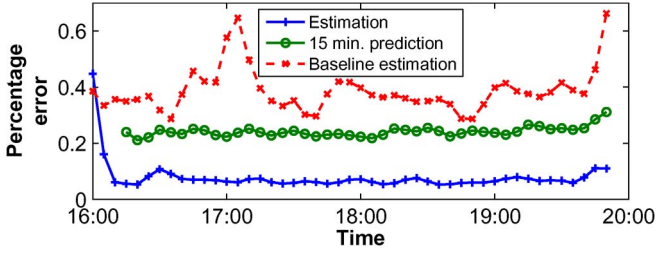


Fig. 9. Evolution of the estimation and prediction of the percentage l_1 error on the validation data set.

e_p is given by $e_p = (\sum_{k=1}^{K(t)} |y_k - \hat{y}_k|^p / |\mathcal{Y}|)^{1/p}$. The error is typically normalized by the average travel-time measurements \bar{y} (time between successive measurements), and we report the percentage of error $\tilde{e}_p = e_p / \bar{y}$. Without a reference, these values are hard to interpret: Travel times on arterial networks have a high variance due to, in particular, the presence of traffic signals [27]. Under similar traffic conditions, the travel times of vehicles on an arterial link significantly vary depending on the time at which the vehicle entered the link and the corresponding waiting time at the signal. To improve interpretability of the results of the model, we compare it to a *baseline model*: a time-series model adapted to probe vehicle data. If probe vehicles sent their travel times between defined positions, time series could be applied to estimate the travel time between these positions. However, no two distinct vehicles report their travel time between the same locations. We propose a baseline model that adapts the traditional time series approach to probe-vehicle data. Travel times are decomposed onto the links of the path, and partial link travel times are scaled onto link travel times. We then use a moving average estimation. We also validate the use of the density model to compute partial link travel times by comparing the errors of the DBN model with and without the density model (scaling of partial link travel times using the fraction of the link traversed). To study the importance of the DBN structure, we consider a model, which is denoted as *self only*, with no spatial dependence: In the 2TBN, the edges representing the dynamics only connect the same links. To show the generality of the spatial dependence allowed by the framework, we consider a model, which is denoted as *not self*, where we remove same link edges from time intervals t to $t + 1$ in the 2TBN in Fig. 2.

Fig. 9 compares the results of the proposed model (estimation and 15-min forecast capabilities) with the baseline model. We notice a significant improvement in the percentage of error compared with the baseline model. The prediction decreases with the horizon of prediction but remains better than the baseline. Note that the baseline model does not have prediction capabilities.

Table III compares the results of the DBN with or without the density model and validates the use of the density modeling to scale partial travel times and compute the pdf of travel times on partial links. The results also validate the short-term prediction capabilities of the DBN (both with and without the density modeling) and underline the importance of the rich DBN structure, as shown by the better results of the model compared with simpler DBN structures (not self and self only).

TABLE III
PERCENTAGE OF l_1 ERROR OF THE MODEL COMPUTED ON A VALIDATION DATA SET TO TEST THE ESTIMATION AND PREDICTION CAPABILITIES OF THE MODEL

	Percentage of l_1 error				
	with density	without density	not self	self only	baseline
Estimation	0.068	0.072	0.086	0.076	0.385
Prediction 15 min.	0.240	0.243	0.243	0.242	N/A

Validation of the estimated travel-time distributions: The algorithm produces more information than a single mean travel time: 1) It characterizes the pdf of travel times on the network; 2) it estimates the probability of congestion $p^{i,t}$ of each link i and time interval t ; and 3) it provides the parameters of the Gaussian distributions $(\mu^{s,i}, \sigma^{s,i})$. The distribution of travel times on any path $j(k)$ can be sampled and numerically approximated using algorithm 2. We use 1000 samples in the following. We define ζ_α as $\zeta_\alpha = \{y \in \mathbb{R} : \mathcal{P}(y_k \leq y) = 1 - \alpha/2, \mathcal{P}(y_k \geq y) = (1 + \alpha)/2\}$. The probability that y_k is in interval ζ_α is α . For a Gaussian distribution, $\zeta_{0.68}$ (resp. $\zeta_{0.95}$) is the interval centered around the median of length two (resp. four) standard deviations. If the estimation of the travel-time distribution is exact, the percentage of points in ζ_α is equal to α . The comparison of the percentage of points in ζ_α with α assesses the goodness of fit of the travel-time distributions with the testing data (see Fig. 10).

Algorithm 2 Travel-time sampling

```

 $\hat{y}^k = 0$  % Initialize the path travel-time sample
for  $l = 1 : j(k)$  do
     $r = \text{rand}()$ ; % Choose the congestion state
    if  $r < p^{c,l}$  then
         $g = \mu^{0,l} + \sigma^{0,l} \text{randn}()$ 
         $\hat{y}^k = \hat{y}^k + g$  % Add the sampled link travel time to the path travel time
    else
         $g = \mu^{1,l} + \sigma^{1,l} \text{randn}()$ 
         $\hat{y}^k = \hat{y}^k + g$  % Add the sampled link travel time to the path travel time
    end if
end for

```

We study the evolution of the percentage of points in ζ_α for different values of α over the validation period. The percentage of points in ζ_α varies over time but remains close to its theoretical value (α), as shown on the left side of Fig. 10. On the right of Fig. 10, we represent the percentage of points in ζ_α (averaged on the entire validation period) as a function of α . For all values of α , the percentage of points in ζ_α is slightly inferior to α . The difference between the theoretical and result curves is mostly due to small inaccuracies in the estimation of the mean and/or underestimation of the variance of the distribution. Note that if the curve produced by the model (dashed line with circles) was over the theoretical line, it would indicate an overestimation of the variance.

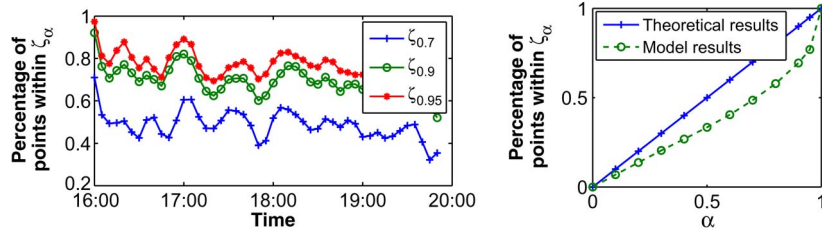


Fig. 10. Validation of the travel-time distributions computed by the model. (Left) Evolution of the percentage of points in ζ_α for $\alpha \in \{0.7, 0.9, 0.95\}$. (Right) Comparison of the percentage of points contained in ζ_α with the theoretical value.

VI. CONCLUSION AND DISCUSSION

Sparsely sampled probe vehicles come as a very promising source of data to develop ubiquitous traffic management systems on arterial networks. We have presented an algorithm that faces specific challenges of probe-vehicle data. In particular, we have addressed the following issues that emphasize the novelty of the estimation technique: 1) The location of measurements and quantity of measurements received in an area are unknown prior to receiving the measurements; 2) the travel-time measurements may span multiple links; and 3) the paths may include partial links for which pdf of travel times must be computed.

The algorithm leverages the massive amounts of data historically available to learn the dynamics of congestion on the network using an EM algorithm. Modeling assumptions on the observation model (independent travel times normally distributed) and the state dynamics (evolution depending on the state of the neighbors) maintain the tractability of the algorithm. In real time, the traffic conditions are estimated using streaming data. The historical training provides robustness to the model when little or no data are available in real time and provides short-term prediction capabilities. The algorithm significantly improves the estimation capabilities of a baseline time-series algorithm adapted to probe-vehicle data. The use of the density modeling to estimate partial link travel times from link travel times also provides an improvement compared with an approach consisting in scaling partial link travel times proportionally to the length of the partial links. Moreover, the algorithm estimates the pdf of travel times on the network rather than the mean travel times only, which is a valuable information given the variability of travel times on arterial networks.

The DBN provides the flexibility to adapt to the specifics of the data received and/or the requirements of the estimation by adapting some of the assumptions.

- The *time discretization* Δ is chosen as a tradeoff between the sparsity of the data and the information that can be reconstructed (fixed to 5 min in the numerical results). This time step can be adapted if more precise information is available (or increased if little information is available and if traffic conditions are known to have slow dynamics in the region and time period of interest).
- The state of traffic of each link is a discrete RV, on which depends the distribution of travel times. The number of traffic state is not theoretically limited and may not

be the same for all the links of the network. Increasing the number of states implies learning a significantly larger number of parameters to represent the dynamics of traffic on the network: parameters of the travel-time distribution for each state and each link of the network and parameters of the transition matrix representing the congestion dynamics of each link. As a tradeoff between the information provided by the probe-vehicle data and the complexity of the model, we have chosen a binary representation of traffic states and underline that the algorithm can be readily applied with a higher number of states.

- Conditioned on the discrete congestion state, the link travel times are RV, chosen to be normally distributed in this paper. The use of Gaussian RVs offers important model refinement possibilities (without increasing the computational complexity). First, the pdf of travel times on a path is analytically computed (conditioned on the state of the links on the path) as the sum of independent Gaussian variables. Second, the independence of link travel time, conditioned on the state of the corresponding links, can be interpreted as modeling link travel times on the network as a multivariate Gaussian RV with a diagonal covariance matrix. Allowing nonzero extradiagonal entries models correlation between the travel times on different links. However, Gaussian distributions do not characterize specific traffic features as distributions derived from a hydrodynamic model of traffic do [27].

A statistical model derived from hydrodynamic and horizontal queuing theory can be used in a DBN framework [26]. It provides interesting results for estimating the dynamics of queues and traffic parameters, such as signal timing, but requires more computation time and system implementation difficulties such as the implementation of global optimization solvers to estimate the parameters of the travel-time distributions.

ACKNOWLEDGMENT

The authors would like to thank T. Hunter from the University of California at Berkeley for providing filtered probe trajectories from the raw measurements of the probe vehicles and the members of the staff of the California Institute for Innovative Transportation for its contributions to develop, build, and deploy the system infrastructure of *Mobile Millennium* on which this paper relies.

REFERENCES

- [1] *Cabspotting*. [Online]. Available: <http://www.cabspotting.org>
- [2] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal. Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.
- [3] X. Ban, R. Herring, P. Hao, and A. Bayen, "Delay pattern estimation for signalized intersections using sampled travel times," in *Proc. 88th Transp. Res. Board Annu. Meeting*, Washington, DC, Jan. 2009, pp. 109–119.
- [4] A. Bayen, J. Butler, and A. Patire, "Mobile millennium final report," Univ. California, Berkeley, CA, CCIT Res. Rep. UCB-ITS-CWP-2011-6, 2011.
- [5] P. Bickel, C. Chen, J. Kwon, J. Rice, E. Van Zwet, and P. Varaiya, "Measuring traffic," *Stat. Sci.*, vol. 22, no. 4, pp. 581–597, 2007.
- [6] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [7] X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 33–42.
- [8] L. Chen, W. L. Jin, J. Hu, and Y. Zhang, "An urban intersection model based on multi-commodity kinematic wave theories," in *Proc. 11th IEEE Intell. Transp. Syst. Conf.*, 2008, pp. 269–274.
- [9] P. Cheng, Z. Qiu, and B. Ran, "Particle filter based traffic state estimation using cell phone network data," in *Proc. 9th Intell. Transp. Syst. Conf.*, Sep. 2006, pp. 1047–1052.
- [10] G. F. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artif. Intell.*, vol. 42, no. 2/3, pp. 393–405, Mar. 1990.
- [11] C. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transp. Res. B*, vol. 28, no. 4, pp. 269–287, Aug. 1994.
- [12] C. de Fabritiis, R. Ragona, and G. Valenti, "Traffic estimation and prediction based on real time floating car data," in *Proc. 11th IEEE Intell. Transp. Syst. Conf.*, 2008, pp. 197–203.
- [13] T. Dean and K. Kanazawa, "A model for reasoning about persistence and causation," *Comput. Intell.*, vol. 5, no. 3, pp. 142–150, Aug. 1990.
- [14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Stat. Soc. Ser. B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.
- [15] A. Doucet and N. de Freitas, *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.
- [16] C. Furtlehner, J. Lasgouttes, and A. de la Fortelle, "A belief propagation approach to traffic prediction using probe vehicles," in *Proc. 10th IEEE Intell. Transp. Syst. Conf.*, 2007, pp. 1022–1027.
- [17] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Proc. Inst. Elect. Eng. F—Radar Signal Process.*, vol. 140, no. 2, pp. 107–113, Apr. 1993.
- [18] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
- [19] B. Hellenga, P. Izadpanah, H. Takada, and L. Fu, "Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments," *Transp. Res. C*, vol. 16, no. 6, pp. 768–782, Dec. 2008.
- [20] R. Herring, "Real-time traffic modeling and estimation with streaming probe data using machine learning," Ph.D. dissertation, Univ. California, Berkeley, CA, 2010.
- [21] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen, "Estimating arterial traffic conditions using sparse probe data," in *Proc. 13th IEEE Intell. Transp. Syst. Conf.*, Madeira, Portugal, Sep. 2010, pp. 929–936.
- [22] R. Herring, A. Hofleitner, S. Amin, T. Abou Nasr, A. A. Khalek, P. Abbeel, and A. Bayen, "Using mobile phones to forecast arterial traffic through statistical learning," presented at the 89th Transportation Research Board Annual Meeting, Washington, DC, Jan. 2010, Paper 10-2493.
- [23] A. Hofleitner and A. Bayen, "Optimal decomposition of travel times measured by probe vehicles using a statistical traffic flow model," in *Proc. 14th IEEE Intell. Transp. Syst. Conf.*, Oct. 2011, pp. 815–821.
- [24] A. Hofleitner, E. Come, L. Oukhellou, J.-P. Lebacque, and A. Bayen, "Automatic signal detection leveraging sparsely sampled probe vehicles," in *Proc. 15th IEEE Intell. Transp. Syst. Conf.*, Anchorage, AK, Sep. 2012.
- [25] A. Hofleitner, R. Herring, and A. Bayen, "A hydrodynamic theory based statistical model of arterial traffic," Univ. California, Berkeley, CA, Tech. Rep. UCB-ITS-CWP-2011-2, Jan. 2011.
- [26] A. Hofleitner, R. Herring, and A. Bayen, "Arterial travel time forecast with streaming data: A hybrid flow model—Machine learning approach," *Transp. Res. B*, 2012, doi: 10.1016/j.trb.2012.03.006, to be published.
- [27] A. Hofleitner, R. Herring, and A. Bayen, "Probability distributions of travel times on arterial networks: A traffic flow and horizontal queueing theory approach," presented at the 91st Transp. Res. Board Annu. Meeting, Washington, DC, Jan. 2012, Paper 12-0798.
- [28] E. Horvitz, J. Apacible, R. Sarin, and L. Liao, "Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service," in *Proc. 21st Conf. Uncertainty Artif. Intell.*, Edinburgh, U.K., 2005.
- [29] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden, "CarTel: A distributed mobile sensor computing system," in *Proc. 4th Int. Conf. Embedded Netw. Sens. Syst.*, 2006, pp. 125–138.
- [30] T. Hunter, T. Moldovan, M. Zaharia, S. Merzgui, J. Ma, J. Franklin, P. Abbeel, and M. J. Bayen, "Scaling the mobile millennium system in the cloud," in *Proc. 2nd ACM SOCC*, 2011, vol. 28, pp. 1–8.
- [31] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, Nov. 1999.
- [32] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *J. Amer. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, Jun. 1958.
- [33] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *J. Comput. Graph. Stat.*, vol. 5, no. 1, pp. 1–25, Mar. 1996.
- [34] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Joint Conf. Artif. Intell.*, 1995, pp. 1137–1143.
- [35] A. Krause, E. Horvitz, A. Kansal, and F. Zhao, "Toward community sensing," in *Proc. 7th IEEE Conf. Inf. Process. Sens. Netw.*, 2008, pp. 481–492.
- [36] J. Kwon and K. Murphy, "Modeling freeway traffic with coupled HMMS," Univ. California, Berkeley, CA, Tech. Rep., 2000.
- [37] J. P. Lebacque, "First-order macroscopic traffic flow models: Intersection modeling, network modeling," in *Proc. 16th Int. Symp. Transp. Traffic Theory*, 2005, pp. 365–386.
- [38] M. Lighthill and G. Whitham, "On kinematic waves. II. A theory of traffic flow on long crowded roads," *Proc. Roy. Soc. Lond. Ser. A*, vol. 229, no. 1178, pp. 317–345, May 1955.
- [39] F. J. Massey, "The Kolmogorov–Smirnov test for goodness of fit," *J. Amer. Stat. Assoc.*, vol. 46, no. 253, pp. 68–78, Mar. 1951.
- [40] X. Min, J. Hu, Q. Chen, T. Zhang, and Y. Zhang, "Short-term traffic flow forecasting of urban network based on dynamic STARIMA model," in *Proc. 12th IEEE Intell. Transp. Syst. Conf.*, 2009, pp. 1–6.
- [41] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Univ. Toronto. Dept. Comput. Sci., Toronto, ON, Canada, Tech. Rep. CRG-TR-93-1, 1993.
- [42] T. Park and S. Lee, "A Bayesian approach for estimating link travel time on urban arterial road network," in *Proc. Comput. Sci. Appl.*, May 2004, pp. 1017–1025.
- [43] F. Porikli and L. Xiaokun, "Traffic congestion estimation using HMM models without vehicle tracking," in *Proc. IVS*, Jun. 2004, pp. 188–193.
- [44] A. Thiagarajan, L. Sivalingam, K. LaCurts, S. Toledo, J. Eriksson, S. Madden, and H. Balakrishnan, "VTrack: Accurate, energy-aware traffic delay estimation using mobile phones," in *Proc. 7th Conf. Embedded Netw. Sens. Syst.*, Berkeley, CA, Nov. 2009, pp. 85–98.
- [45] H. Van Zuylen, F. Zheng, and Y. Chen, "Using probe vehicle data for traffic state estimation in signalized urban networks," in *Proc. Traffic Data Collect. Stand.*, 2010, pp. 109–127.
- [46] D. B. Work, S. Blandin, O.-P. Tossavainen, B. Piccoli, and A. M. Bayen, "A traffic model for velocity data assimilation," *Appl. Math. Res. Exp.*, vol. 2010, no. 1, pp. 1–35, Apr. 2010.



Aude Hofleitner received the Engineering degree in applied mathematics from Ecole Polytechnique, Palaiseau, France; the M.S. degree in transportation engineering from the University of Marne La Vallée, Marne La Vallée, France; and the M.S. degree in applied mathematics from the Ecole Nationale des Ponts et Chaussées, Paris, France. She is currently working toward the Ph.D. degree in electrical engineering and computer science with the University of California, Berkeley.

Ms. Hofleitner was a recipient of the Dwight David Eisenhower Transportation Fellowship from the U.S. Department of Transportation and the Anita Borg Grace Hopper Conference Scholarship in 2011.



Ryan Herring received the Bachelor's degree in applied mathematics from Oberlin College, Oberlin, OH, in 2005 and the M.S. and Ph.D. degrees in industrial engineering and operations research from the University of California, Berkeley, in 2006 and 2010, respectively.

While he was a graduate student, he was with the California Center for Innovative Transportation, University of California, Berkeley. He is currently a Software Engineer with Apple, Inc. His research interests include big data, machine learning, and

estimation.



Pieter Abbeel received the B.S. and M.S. degrees in electrical engineering from Katholieke Universiteit Leuven, Leuven, Belgium, and the Ph.D. degree in computer science from Stanford University, Stanford, CA, in 2008.

Since 2008, he has been a Faculty Member with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. He has developed apprenticeship learning algorithms, which have enabled advanced helicopter aerobatics, including maneuvers such as tic-tocs, chaos,

and autorotation, which only exceptional human pilots can perform. His group has also enabled the first end-to-end completion of reliably picking up a crumpled laundry article and folding it. His work has been featured in many popular press outlets, including *BBC*, *MIT Technology Review*, *Discovery Channel*, *SmartPlanet*, and *Wired*.

Dr. Abbeel has received of various awards, including the Best Paper Awards at the International Conference on Machine Learning and the International Conference on Robotics and Automation, the Sloan Fellowship, the Okawa Foundation Award, and 2011's TR35.



Alexandre Bayen received the Engineering degree in applied mathematics from the Ecole Polytechnique, Palaiseau, France, in 1998 and the M.S. and Ph.D. degrees in aeronautics and astronautics from Stanford University, Stanford, CA, in 1999 and 2003, respectively.

From 2000 to 2003, he was a Visiting Researcher with NASA Ames Research Center. He worked as a Research Director with the Autonomous Navigation Laboratory, Laboratoire de Recherches Balistiques et Aerodynamiques, Ministère de la Defense, Vernon,

France, where he holds the rank of Major. He is currently an Associate Professor with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. He is the author of one book and over 100 articles in peer-reviewed journals and conferences.

Dr. Bayen received the Ballhaus Award from Stanford University in 2004, the CAREER award from the National Science Foundation in 2009, and the Presidential Early Career Award for Scientists and Engineers in 2010. He is also included in the NASA Top 10 Innovators on Water Sustainability in 2010. His projects Mobile Century and Mobile Millennium received the 2008 Best of ITS Award for "Best Innovative Practice" at the ITS World Congress and a TRANNY Award from the California Transportation Foundation in 2009. Mobile Millennium has been featured more than 100 times in the media, including television channels and radio stations (*CBS*, *NBC*, *ABC*, *CNET*, *NPR*, *KGO*, and the *BBC*), and in the popular press (*Wall Street Journal*, *Washington Post*, and *LA Times*).