Cooperative Bayesian Estimation of Vehicular Traffic in Large-Scale Networks

Alessandra Pascale, *Member, IEEE*, Monica Nicoli, *Member, IEEE*, and Umberto Spagnolini, *Senior Member, IEEE*

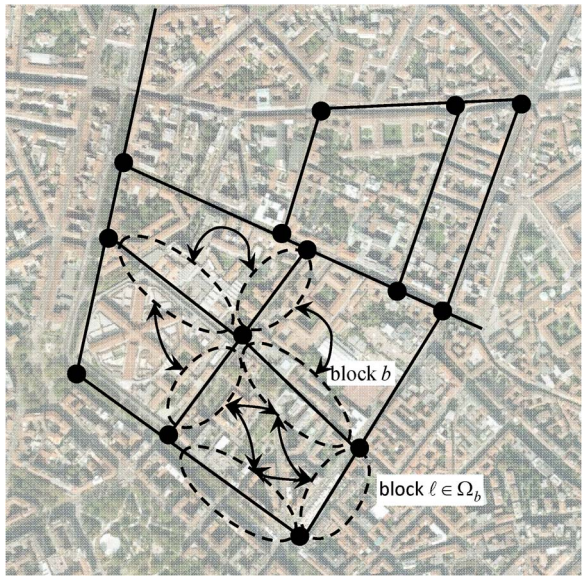

Fig. 1. Example of decomposition of a large-scale traffic estimation problem into smaller but interdependent problems (map courtesy of Google Maps).

I. Introduction

INTELLIGENT transportation systems (ITS) are expected to improve the quality, safety, and sustainability of mobility, by integrating information and communication technologies with transport engineering. ITS rely on a capillary network of devices, either road-embedded or mobile probes, that are distributed over the roads providing local measurements of traffic variables such as speed, flow, and density. Measurements collected by the sensors, which are often heterogeneous, are fused and elaborated to draw information on the global traffic state and take corresponding actions for mobility optimization [1], [2].

A traffic network can be described as a set of interconnected road links, as illustrated in Fig. 1 for the metropolitan area of Milan, Italy, with roads represented as thick lines and intersections as dots. Typically, on some links, the traffic variable is directly observed by one or more sensing devices (i.e., loops or cameras), while other links could be completely unmonitored. Focusing on a large-scale network, the objective of this paper is the estimation of the overall traffic field starting from the set of sparse, irregularly deployed, and noisy measurements provided by the available sensors. Notice that *a priori* information on the traffic dynamics is crucial for a reliable reconstruction, in order to compensate for incomplete and noisy observations. The use of an ensemble of *a priori* information, ranging from traffic modeling (e.g., fluid dynamic) to historical data, motivates the employment of Bayesian methods.

The Bayesian approach proposed in this paper jointly exploits real-time sensor data and analytical traffic modeling, to enable the recovery of the traffic field over the whole network as for a virtual sensor deployment covering all links. The inference problem is very challenging due to the nonlinear nature of the traffic process and the high complexity required by the large-scale estimation that involves a huge number of variables [3], [4]. It is even more challenging if we consider that measurements are often heterogeneous for nature and accuracy, and they are typically acquired by different operators that are

Manuscript received July 19, 2013; revised November 18, 2013 and February 18, 2014; accepted February 23, 2014. Date of publication April 17, 2014; date of current version September 26, 2014. This work was supported by a Ph.D. Grant of the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) and the Scuola Interpolitecnica di Dottorato (SIPD), and by the project Mobilità Intelligente Ecosostenibile (MIE), funded by MIUR within the framework Cluster Tecnologico Nazionale "Tecnologie per le Smart Communities". The Associate Editor for this paper was S. Sun.

A. Pascale was with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milano, Italy. She is now with the Smarter Cities Technology Centre, IBM Research–Ireland, Dublin Technology Campus, Dublin 15, Ireland (e-mail: apascale@ie.ibm.com).

M. Nicoli and U. Spagnolini are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milano, Italy (e-mail: monica.nicoli@polimi.it; umberto.spagnolini@polimi.it).

Color versions of one or more of the figures in this paper are available online.

not prone to share traffic data for obvious economic reasons, which justifies the decomposition of the estimation into a set of small-scale interdependent problems, as illustrated in Fig. 1. The network is divided into subnetworks, here referred to as blocks, whose dimensions vary depending on the specific application (e.g., from urban crosses to huge portions of cities/highways). Cooperative processing is carried out by sharing information on the reliability of the local traffic estimates among neighboring subnetworks (as shown by the arrows in Fig. 1).

Cooperative Bayesian estimation finds a variety of applications in ITS. It is suited for networks with heterogeneous data sources (loops, probe vehicles, cameras, and wireless sensor networks) or with subnetworks monitored by different road operators. Fusion of data from hybrid sources/operators is known to increase the coverage and accuracy of traffic monitoring [5]. Furthermore, the cooperative Bayesian approach avoids the exchange of raw data between subnetworks, passing nonsensitive beliefs that embed the traffic information. Possible applications range from data reconstruction in the case of missing or failured sensors, to accurate monitoring for local control operation (e.g., for traffic light management with few reliable sensors), traffic forecast for dynamic route guidance (where observation is incomplete also over time), and distributed traffic monitoring from geolocalized interconnected vehicles (e.g., in vehicular ad-hoc networks) [1], [6].

As an example of application, in this paper we consider a simulated single-lane road scenario, with fragmented traffic monitoring by loops, characterized by several sensor failures. This paper shows how the proposed approach can overcome the problem related to the sensor malfunctions exploiting information shared with neighboring subnetworks, as for a virtual global (or centralized) processing. The proposed method relies on the cell transmission model (CTM) [7] for traffic description, a discrete version of the fluid-dynamic Lighthill–Whitham–Richards (LWR) theory, here extended as proposed by the authors in [8] with a stochastic component that accounts for the randomness of traffic. For estimation of the traffic variables, Bayesian filtering is employed in each subnetwork, which is integrated with belief propagation (BP) [9] for information sharing between subnetworks. Performances are evaluated simulating different traffic scenarios.

This paper is organized as follows. A brief overview of the literature on traffic estimation and the relation to prior work is in Section II. The original contribution is outlined in Section III, explaining the key idea behind the proposed approach. Section IV contains a description of the traffic model used for the development of the estimation method, with the adaptation of CTM to the multiblock structure needed for cooperative processing. Section V details the proposed estimation technique. Sections VI and VII present the performance assessment in terms of computational complexity and estimate accuracy. Finally, Section VIII presents the conclusion and future developments.

II. Related Works

Statistical inference methods proposed in the literature mainly deal with centralized approaches for traffic reconstruction. Conventional methods exploit analytical models or historical data set [10]–[14]. Motivated by the stochastic nature of traffic, the scientific community has recently started to investigate nonlinear Bayesian models [15]–[17], particle filter (PF) [17], [18], Bayesian networks (BNs) [19]–[22], mixtures of Gaussian processes [23], Gaussian process regression [24], and graph theory [25]. Recently, some works have investigated PF with centralized [18] or parallel computation [26], using a second-order nonlinear model, i.e., the compositional stochastic model.

The focus of this paper is to move from a centralized traffic estimation approach to a new cooperative approach. With this perspective, graph theory is a promising tool for describing the statistical relations existing among traffic observations at different, and typically nonuniformly distributed, locations or time instants. Unlike other contexts such as image processing, wireless sensor networks, wireless positioning, and decision making, in which Markov random fields (MRFs) and BNs [27]–[30] are now mature tools, the approach is relatively new in traffic estimation. Traffic can be view as an MRF defined over an undirected graph, whose nodes are associated with the traffic variable observed at different locations/instants, while edges define the relations existing among the variables according to the fluid–dynamic theory [31].

To cope with nonlinearity, a centralized estimation solution based on the fluid–dynamic model CTM and PF has been proposed in [8]; the method, however, becomes too complex (due to the increasing number of particles) in the case of large road dimensions. A different approach, relying on a linearized CTM, has been developed in [15]–[17], in which a jump Markov linear system (JMLS, e.g., [32]) has been introduced to model traffic using mixture Kalman filtering for traffic density estimation [15]. Kalman filtering is very convenient because it provides a practical and easily implementable closed-form solution for traffic estimation. Its employment, however, is suited only for small-scale networks, due to the fact that it requires linearization by JMLS and the number of variables to be jointly handled in the linearized system increases with the road size, becoming unfeasible from a computational viewpoint in large-scale networks.

In this paper, we propose to extend the JMLS approach to a cooperative framework, which is based on the division of the estimation problem into subnetworks and on the cooperation between the estimation processes in different subnetworks, in such a way to enable the application to large-scale networks with feasible complexity. Cooperative processing, which has been widely investigated in recent years for statistical inference in sensor networks [33], [34], has a great deal of potential for monitoring of large-scale traffic networks. In this context, designing the cooperative algorithm is a challenging task due to the nonlinearity of traffic and the complicated spatial–temporal network structure.

III. Cooperative Approach

Here, we outline the cooperative paradigm for traffic estimation highlighting the main contribution of this paper. Without loss of generality, traffic density is assumed as the variable to be

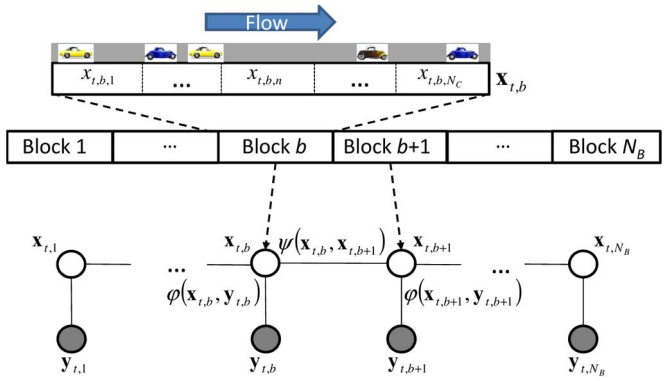

Fig. 2. Road traffic modeling by Bayesian network.

estimated (velocity or flow could be considered as well, either as separated or jointly handled variables). In order to sketch the estimation problem, we consider the single-road scenario in Fig. 2, which is general enough to allow the extension to more complex networks with straightforward adaptations (as done hereinafter in the numerical performance analysis). The road is divided into a set $\mathcal{B}$ of $N_B = |\mathcal{B}|$ subnetworks or blocks. Each block is composed of N_C cells. At time t, the traffic densities observed in the N_C cells of block b are denoted by the state vector $\mathbf{x}_{t,b} = [x_{t,b,1} \cdots x_{t,b,N_C}]^T$ and the corresponding noisy measurements provided by sensors as $\mathbf{y}_{t,b} = [y_{t,b,1} \cdots y_{t,b,N_C}]^T$.

Conventional approaches for traffic estimation perform a separate analysis in each road block, taking into account only the flows on boundary sections. However, density variables observed over closely located blocks are likely to be statistically related due to the diffusion process [31]. We propose to account for this interdependence by modeling the density over the network as an MRF. Its estimation is decomposed into a set of local estimations, each performed over a subnetwork in cooperation with neighboring subnetworks by an iterated exchange of traffic beliefs. Compared to the global centralized estimation, the message-passing procedure enables a full exploitation of all measurements in the network and, thus, the attainment of nearly optimal performances[1] with lower complexity.

We introduce a graph model in which nodes are associated with blocks $b \in \mathcal{B}$, while edges $(b, \ell) \in \mathcal{E}$ represent the conditional dependencies between densities in neighboring blocks b and ℓ, with $\mathcal{E}$ being the set that collects all the neighbor pairs. For the estimation of the traffic densities $\mathbf{x}_{t,b}$, we propose to exploit the ensemble of all the measurements collected up to time t by all sensors: $\mathbf{Y}_{1:t} = [\mathbf{y}_{1:t,1} \cdots \mathbf{y}_{1:t,N_B}]$ with $\mathbf{y}_{1:t,b} = [\mathbf{y}_{1,b} \cdots \mathbf{y}_{t,b}]$. Assuming a cycle-free graph, the information about the global state $\mathbf{X}_t = [\mathbf{x}_{t,1} \cdots \mathbf{x}_{t,N_B}]$ can be represented by the joint distribution [29]

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \propto \prod_{b\in\mathcal{B}} \underbrace{p(\mathbf{x}_{t,b}|\mathbf{Y}_{1:t-1})p(\mathbf{y}_{t,b}|\mathbf{x}_{t,b})}_{\varphi(\mathbf{x}_{t,b},\mathbf{y}_{t,b})} \times \prod_{(b,\ell)\in\mathcal{E}} \underbrace{\frac{p(\mathbf{x}_{t,b},\mathbf{x}_{t,\ell}|\mathbf{Y}_{1:t-1})}{p(\mathbf{x}_{t,b}|\mathbf{Y}_{1:t-1})p(\mathbf{x}_{t,\ell}|\mathbf{Y}_{1:t-1})}}_{\psi(\mathbf{x}_{t,b},\mathbf{x}_{t,\ell})} \quad (1)$$

[1]Convergence rate depends on the network connectivity [35].

where the function $\varphi(\mathbf{x}_{t,b}, \mathbf{y}_{t,b})$ embodies the information on the traffic state $\mathbf{x}_{t,b}$ in block b drawn from local sensor measurements, while $\psi(\mathbf{x}_{t,b}, \mathbf{x}_{t,\ell})$ denotes the *a priori* information that comes from neighboring blocks.

The proposed method draws the global information $p(\mathbf{X}_t|\mathbf{Y}_{1:t})$ by means of a distributed BP procedure [9] that runs on the graph in Fig. 2. At each time instant t, the estimate of the traffic density $\mathbf{x}_{t,b}$ in each block b is carried out by the computation of the local belief $\varphi(\mathbf{x}_{t,b}, \mathbf{y}_{t,b})$. A phase-conditioned linear model is employed for the description of the traffic evolution: traffic is seen locally as a JMLS switching over a set of linear propagation modes associated to the possible configurations of traffic phases (free flow or congestion); a PF is used for the estimation of phases and a set of Kalman filters (KF) for densities. The estimate is then refined based on the information $\psi(\mathbf{x}_{t,b}, \mathbf{x}_{t,\ell})$ provided by neighboring blocks via the BP procedure. The local traffic phase is selected as the key information to be shared for cooperation because it rules the backward/forward propagation linking the measurements of different subnetworks. This parameterization has the advantage of avoiding an intensive exchange of data between subnetworks.

The **original contribution** can be summarized as follows:

1) The introduction of the cooperative paradigm in traffic estimation (as outlined earlier), based on cooperation between estimation processes in different subnetworks and sharing of statistical information, in such a way to enable the application of Bayesian traffic estimation to large-scale road networks with feasible complexity;
2) The definition of a new traffic model (see Section IV) relying on a stochastic version of the CTM and a multi-block extension of the JMLS framework [16], [17], with Markovian modeling of the phase evolution over closely located blocks and inclusion of new phase configurations in each block for Bayesian description of traffic behavior in large-scale networks;
3) The development of a BP-PF procedure (see Section V) for the implementation of the cooperative approach. The performances of the method are compared to those of conventional Bayesian estimation in terms of accuracy (see Section VI) and computational complexity (see Section VII).

IV. Traffic Model

Here, a detailed description of how traffic is modeled on a local and global basis is presented. A brief recall of the CTM is given in the following, which is followed by the definition of the traffic model in the single road block (see Section IV-A) and the extension to multiple blocks for large-scale environments (see Section IV-B).

Traffic is modeled according to the CTM, with road space divided into cells and time sampled with interval Δt. As illustrated in Fig. 3, at time step t, the traffic state in cell n of length l_n is described by the density $\rho_n(t)$, i.e., number of vehicles per space unit, which evolves following the conservation law:

$$\rho_n(t) = \rho_n(t-1) + \frac{\Delta t}{l_n}\left(f_n(t-1) - f_{n+1}(t-1)\right) \quad (2)$$

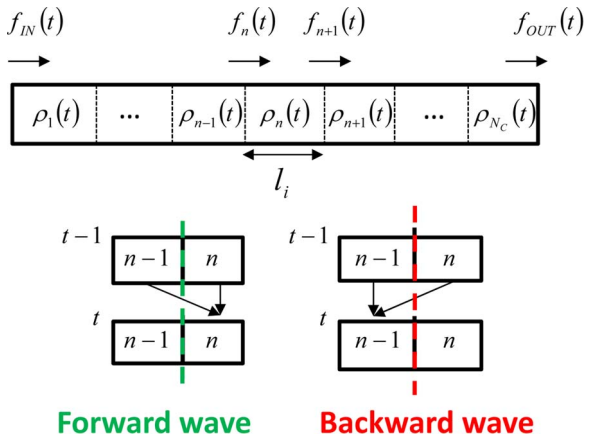

Fig. 3. CTM modeling: (top) section of road divided into cells; (bottom) linear models for forward/backward-propagating waves across two cells.

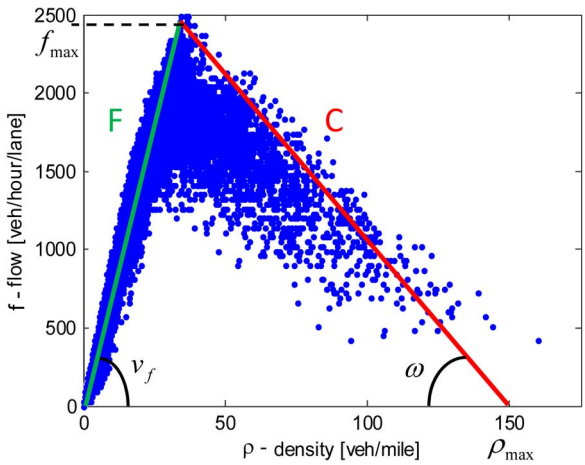

Fig. 4. Fundamental diagram.

where $f_n(t)$ and $f_{n+1}(t)$ denote the flow entering and exiting the cell n at time step t, respectively. The flow entering cell n depends on the traffic demand generated by cell $n-1$, the supply of space provided by cell n, and the road capacity $f_{\max}$ (i.e., the maximum number of vehicles that can move from one cell to the next), according to the relation

$$f_n(t) = \min(v_f \rho_{n-1}(t), \omega(\rho_{\max,n} - \rho_n(t)), f_{\max}) + w_n(t). \quad (3)$$

Here, v_f denotes the free-flow velocity, ω denotes the jam upstream velocity, and $\rho_{\max,n}$ denotes the maximum density in cell n. The zero-mean random variable $w_n(t)$ has been introduced, with respect to the deterministic model [7], to account for the randomness of traffic related to the granular characteristic of flows and the unpredictable driver behaviors.

It can be seen from (3) that the relation existing in stationary conditions between the mean density and flow variables, which is defined by the function $\min(\cdot)$, is nonlinear. This relation is illustrated in the fundamental diagram in Fig. 4, in which the two linear branches (first two arguments in the $\min(\cdot)$ function) are associated with the two main traffic phases: the straight line with slope v_f represents the free-flow condition (F), with flow increasing as density increases, and the straight line with slope ω represents the backward traffic propagation when congestion (C) occurs. Randomness of traffic can be appreciated by looking at the dots in Fig. 4 representing flow–density measurements collected by a traffic sensor. It can be observed that data samples are randomly distributed around the fundamental diagram, with a higher spread over the congestion branch rather than the free-flow branch, due to the increased interaction observed between vehicles in congested conditions. The noise term $w_n(t)$ in (3) accounts for such a dispersion, with variance selected to model the different degree of variability observed in the two phases.

The space–time process $\rho_n(t)$ in (2) and (3) is complex to handle because it is nonlinear. Still, as highlighted earlier, it can be seen as a composition of two main propagation phases: a forward propagation with velocity v_f in the F phase (see the left scheme in Fig. 3) and a backward propagation with velocity ω in the C phase (right scheme in Fig. 3). The phase F/C describes the propagation direction of the traffic wave at the intersection between two cells. This property can be used to simplify the traffic model and the related estimation over a block of cells, as described in the following section.

A. Modeling a Block of Cells

To model a road section of N_C cells, monitored by a set of sensors, we define as overall traffic state the vector of densities $\mathbf{x}_t = [\rho_1(t) \cdots \rho_{N_C}(t)]^T$ that evolve over time according to (2) and (3) for $n = 1, \ldots, N_C$. Without loss of generality, we assume that densities are measured by N_C sensors, i.e., one sensor for each cell, with accuracy that may vary from cell to cell.[2] The measurement of the density in the nth cell is defined as $\tilde{\rho}_n(t) = \rho_n(t) + r_n(t)$, with the random term $r_n(t)$ modeling the measurement error. Let $\mathbf{y}_t = [\tilde{\rho}_1(t) \cdots \tilde{\rho}_{N_C}(t)]^T$ denote the whole set of measurements, $\boldsymbol{\rho}_{\max} = [\rho_{\max,1} \cdots \rho_{\max,N_C}]^T$ denote the maximum densities at the two end cells, and $\mathbf{u}_t = [f_{\text{in}}(t) \;\; f_{\text{out}}(t)]^T$ denote the inflow/outflow conditions, the Bayesian model describing the entire road section (or block) is given by

$$\begin{aligned} \mathbf{x}_t &= \mathbf{g}(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\rho}_{\max}, \mathbf{w}_t) \\ \mathbf{y}_t &= \mathbf{x}_t + \mathbf{r}_t \end{aligned} \quad (4)$$

where $\mathbf{g}(\cdot)$ is the nonlinear vector function obtained from models (2) and (3), for $n = 1, \ldots, N_C$; $\mathbf{r}_t = [r_1(t) \cdots r_{N_C}(t)]^T$ is the overall measurement noise; and $\mathbf{w}_t = [w_0(t) \cdots w_{N_C}(t)]^T$ is the random process driving the model evolution. The widely adopted zero-mean uncorrelated Gaussian distribution is assumed for the two mutually independent random vectors $\mathbf{r}_t$ and $\mathbf{w}_t$, as an approximation of the real probability density functions (pdfs) [15], [17], with limited values to ensure that the density remains within the range $[0, \rho_{\max}]$ and the flow within $[0, f_{\max}]$.[3] Particular care should be taken to ensure that the samples of the driving noise fulfill the feasibility constraint of the CTM: $v_f \leq v^{\max}$ and $\omega \leq v^{\max}$ with $v^{\max} = (l_n / \triangle t)$ for

[2]The model can be extended to cases in which some cells are not monitored by sensors. Ideal condition for traffic estimation would be to have at least one sensor per block, but the proposed estimation method also remains valid in the case of the presence of nonmonitored blocks, as it is able to exploit measurements coming from neighboring blocks. This aspect will be explained in detail by simulations in Section VII, where a fragmented monitoring scenario is considered.

[3]The noise variances used in the simulations are sufficiently low to consider the effect of the limitation on the Gaussian pdf as negligible.

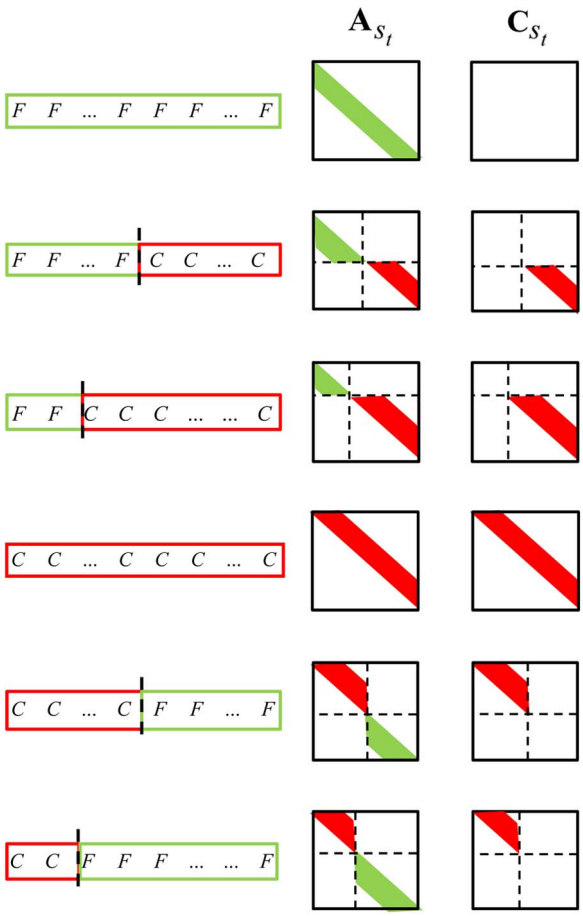

Fig. 5. (Left) Phase configurations. (Right) Structure of the matrices $\mathbf{A}_{s_t}$ and $\mathbf{C}_{s_t}$ of the JMLS model.

cell n (i.e., a stream of vehicles cannot pass two cells in one time-step). See [8] for details.

Since the CTM model (2) conditioned to the (either F or C) traffic phase is linear in $\rho_n(t)$, we can see the density as the output of a switched linear model ruled by the phase variable, as originally proposed in [16]. In our specific scenario, the phase for the block of N_C cells, i.e., the *block* phase, is defined by the set of F/C binary conditions over the N_C+1 interfaces between cells. The block phase at time t is denoted by the discrete-valued variable $s_t \in \mathcal{S}$ that takes values over a set of $|\mathcal{S}| = 2^{N_C+1}$ possible combinations of the N_C+1 phases, ranging from $s_t = \{F \dots F\}$ (when the whole block is in free flow) to $s_t = \{C \dots C\}$ (when the block is fully congested). For a small road section (i.e., for small N_C), we can reduce the cardinality by excluding some unlikely traffic conditions from the set $\mathcal{S}$. Specifically, we assume a reduced set $\mathcal{S}'$, where the block phase is all "F", all "C", or mixed with one single breakpoint (or wavefront) "FC" or "CF", to account for homogeneous traffic conditions with at most one phase change. This assumption leads to $N_{\text{ph}} = 2N_C + 2 = |\mathcal{S}'|$ resulting phase configurations, as illustrated in the left panel in Fig. 5. As an example, for $N_C = 2$, it is $N_{\text{ph}} = 6$ with $\mathcal{S}' = \{\text{FFF}, \text{FFC}, \text{FCC}, \text{CCC}, \text{CCF}, \text{CFF}\}$. Different from previous works where the front "CF" has not been considered as unlikely to occur in small road sections [17], here we need to account also for this case to enable the extension to large-scale networks and to properly model the backward propagation of platoons of vehicles in congestion.

TABLE I
JMLS MATRICES FOR DIFFERENT VALUES OF THE TRAFFIC PHASE FOR A BLOCK OF $N_C = 2$ CELLS

s_t	$\mathbf{A}_{s_t}$	$\mathbf{B}_{s_t}$	$\mathbf{C}_{s_t}$
FFF	$\begin{bmatrix} 1-a^F_{1,1} & 0 \\ a^F_{1,2} & 1-a^F_{2,2} \end{bmatrix}$	$\begin{bmatrix} b_1 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
FFC	$\begin{bmatrix} 1-a^F_{1,1} & 0 \\ a^F_{1,2} & 1 \end{bmatrix}$	$\begin{bmatrix} b_1 & 0 \\ 0 & -b_2 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
FCC	$\begin{bmatrix} 1 & a^C_{2,1} \\ 0 & 1-a^C_{2,2} \end{bmatrix}$	$\begin{bmatrix} b_1 & 0 \\ 0 & -b_2 \end{bmatrix}$	$\begin{bmatrix} 0 & -a^C_{2,1} \\ 0 & a^C_{2,2} \end{bmatrix}$
CCC	$\begin{bmatrix} 1-a^C_{1,1} & a^C_{2,1} \\ 0 & 1-a^C_{2,2} \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & -b_2 \end{bmatrix}$	$\begin{bmatrix} a^C_{1,1} & -a^C_{2,1} \\ 0 & a^C_{2,2} \end{bmatrix}$
CCF	$\begin{bmatrix} 1-a^C_{1,1} & a^C_{2,1} \\ 0 & 1-a^C_{2,2} \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} a^C_{1,1} & -a^C_{2,1} \\ 0 & a^C_{2,2} \end{bmatrix}$
CFF	$\begin{bmatrix} 1-a^C_{1,1} & 0 \\ 0 & 1-a^F_{2,2} \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} a^C_{1,1} & 0 \\ 0 & 0 \end{bmatrix}$

It is worth noticing that the assumption of a single breakpoint that allows to reduce the number of phases from 2^{N_C+1} to $2N_C+2$ is realistic only for small road sections; a similar condition has been considered in [15] and [17], in which the road sections have four cells as a maximum. For larger networks, this assumption does not hold anymore, and the relation of the number of phases to the number of cells cannot be considered as linear anymore, causing a fast increment in the system dimensions and inevitably in the computational complexity of the traffic model. A way to adapt the model to large-scale systems is the extension to multiple blocks through cooperative processing, as proposed in the next section.

As regards the phase evolution in time, we assume that the block phase s_t evolves according to the Markov chain in Fig. 6(a), with transition probabilities $p(s_{t+1}|s_t)$ and starting state $s_0 = \text{FFF}$. Given the block phase s_t, the densities follow the linear model

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}_{s_t}\mathbf{x}_{t-1} + \mathbf{B}_{s_t}\mathbf{u}_{t-1} + \mathbf{C}_{s_t}\boldsymbol{\rho}_{\max} + \mathbf{w}_{t,s_t} \\ \mathbf{y}_t &= \mathbf{x}_t + \mathbf{r}_t \end{aligned} \tag{5}$$

where system matrices $\{\mathbf{A}_{s_t}, \mathbf{B}_{s_t}, \mathbf{C}_{s_t}\}$ and the covariance matrix of the driving process $\mathbf{w}_{t,s_t} = [w_{s_t,1}(t) \cdots w_{s_t,N_C}(t)]^T$ depend on the specific value of s_t. The generating noise for the density model is $w_{s_t,n}(t) = (\Delta t/l_n)(w_n(t-1) - w_{n+1}(t-1))$. The traffic process is thus modeled as a JMLS [32] that switches among N_{ph} linear systems according to the evolution of the traffic phase s_t.

The structure of the matrices $\{\mathbf{A}_{s_t}, \mathbf{B}_{s_t}, \mathbf{C}_{s_t}\}$ is obtained from (2) and (3) based on the traffic phase value s_t. Matrix $\mathbf{B}_{s_t}$ is used to impose the input/output flows $\mathbf{u}_t$ on the road section boundaries, while $\mathbf{A}_{s_t}$ and $\mathbf{C}_{s_t}$ describe the traffic propagation over the road, as illustrated in the right plane in Fig. 5. For $N_C = 2$, they model a forward traffic propagation for $s_t = \text{FFF}$ and a backward propagation for $s_t = \text{CCC}$, while they have a mixed structure depending on the breakpoint position in all the other cases. Matrix elements are written in Table I, for the case $N_C = 2$, as a function of the following parameters: $a^F_{n,m} = v_{f_n}/v^{\max}_m$, $a^C_{n,m} = \omega_n/v^{\max}_m$, and $b_n = 1/v^{\max}_n$ (velocities can change from cell to cell and are thus indicated by the subscript n).

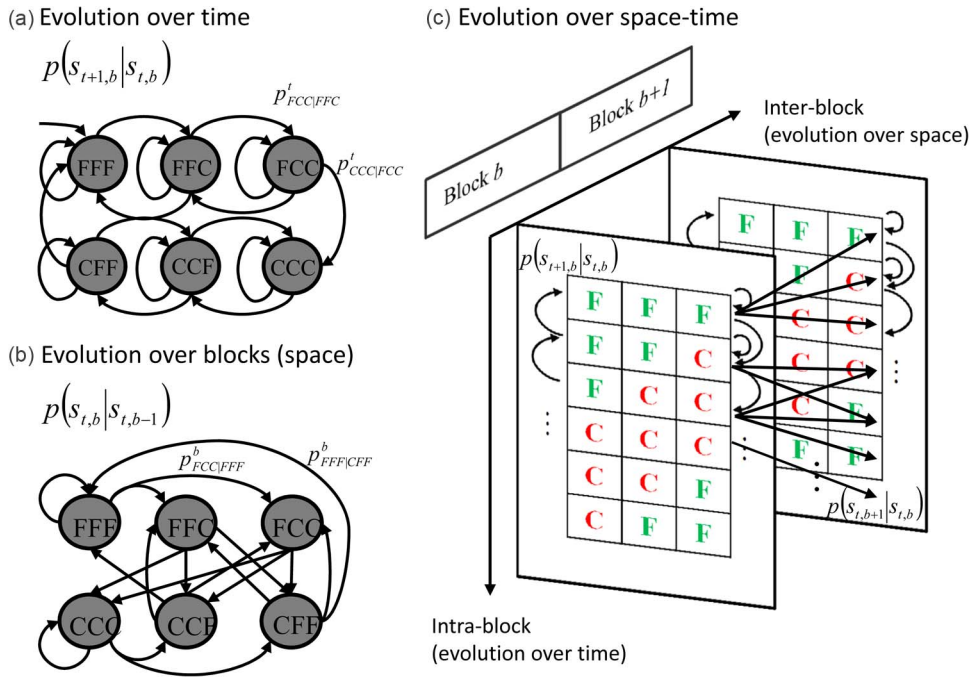

Fig. 6. Space–time Markovian model for the evolution of the traffic phase within a block and over multiple blocks. (a) Markov chain for the temporal evolution within the block. (b) Markov chain for the spatial evolution over the blocks. (c) Space–time Markov chain obtained as composition of the two chains: the arrows within the block represent the evolution of the phase in time and the arrows between the blocks represent the evolution in space. Probabilities associated to the arrows denote the transition probabilities of the Markov chain.

As concerns the noise terms, we assume that the variance of the traffic generating process $w_n(t)$ is larger in congestion ($\sigma^2_{w_C}$) than in free flow ($\sigma^2_{w_F}$), i.e., $\sigma^2_{w_C} > \sigma^2_{w_F}$. We thus set $\mathbf{w}_{t,s_t} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{s_t})$, where $\mathbf{Q}_{s_t}$ is a $N_C \times N_C$ matrix whose elements are linear combinations of the two variances $(\sigma^2_{w_F}, \sigma^2_{w_C})$, with weights of the combination depending on the phase s_t. Matrix elements can be easily found for all cases in Table I, by considering the definition of $w_{s_t,n}(t)$ as a function of $w_n(t)$ and taking into account that $w_n(t)$ is uncorrelated. The distribution of the measurement error is $\mathbf{r}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2_r \mathbf{I}_{N_C})$.

B. Modeling Multiple Blocks

For modeling large-scale systems, here we propose the extension of the previous model to a road of N_B blocks (with N_C cells each), using $\mathbf{x}_{t,b}$ and $s_{t,b}$ to denote the densities and the phases in the bth block, for $b = 1, \ldots, N_B$. We assume that the traffic densities $\mathbf{x}_{t,b}$ in each block b, when conditioned to the local traffic phase $s_{t,b}$ and for given boundary conditions, evolve independently from the other blocks according to the linear model (5). The traffic phase, on the other hand, is statistically related to the traffic phases in the two adjacent blocks due to the colocation and the backward/forward propagation mechanism. We propose to model the phase evolution over the blocks according to a first-order Markov model with transition probabilities $p(s_{t,b}|s_{t,b\pm1})$, as depicted in Fig. 6(b).

To illustrate the concept, we consider a road composed of blocks having $N_C = 2$ cells, $N_C + 1 = 3$ interfaces, and $N_{\text{ph}} = 6$ block-phase values, as in the previous example. Since phases defined on consecutive blocks partially overlap, as shown in Fig. 7, only a subset of phase configurations are admissible over the blocks. Fig. 7 depicts three examples of feasible configurations $[s_{t,b-1}, s_{t,b}, s_{t,b+1}]$, for block $b-1$, b, and $b+1$ at time t: in the first example, the propagation of two wavefronts is represented, and the second and third examples show congestion at the beginning and end of the section, respectively. The feasible phase configurations among two adjacent blocks, i.e., $[s_{t,b-1}, s_{t,b}]$, can be drawn from the Markov chain in Fig. 6(b) that shows all the admissible phase transitions from $b-1$ to b by arrows. The chain ruling the phase evolution from $b+1$ to b can be drawn symmetrically for the pair of phases $[s_{t,b}, s_{t,b+1}]$. Transition probabilities can be estimated from data by computing the frequencies of state transitions. In the simulated scenarios considered in this paper, the analysis of the transition frequencies showed that all transitions departing from the same starting phase $s_{t,b-1}$ have approximately the same probability $p(s_{t,b}|s_{t,b-1})$.

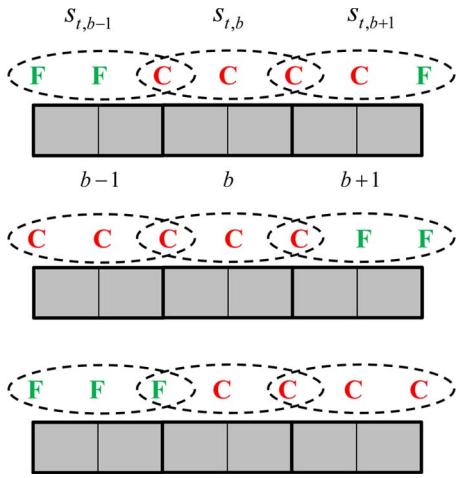

Fig. 7. Spatial evolution of traffic phases over a road section composed of three blocks of $N_C = 2$ cells. Example of configurations of block phases for adjacent blocks.

The overall space–time Markovian chain for the phase evolution is represented in Fig. 6, as the composition of the two separate evolutions: over time within a single block [see Fig. 6(a)], over space spanning different blocks [see Fig. 6(b)], and over space and time [Fig. 6(c)].

V. Estimation of Traffic Densities

This section details the proposed procedure for cooperative traffic estimation. The objective is the estimation of the traffic densities $\mathbf{x}_{t,b}$ over each block b, exploiting *all* the measurements $\mathbf{Y}_{1:t}$ collected in the N_B blocks (up to time t) and relying on the Bayesian model defined in Section IV-A and B. The optimal estimate is based on the *a posteriori* pdf

$$p(\mathbf{x}_{t,b}|\mathbf{Y}_{1:t}) = \sum_{\mathbf{s}_{1:t,b}} p(\mathbf{x}_{t,b}|\mathbf{s}_{1:t,b}, \mathbf{Y}_{1:t}) p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t}) \quad (6)$$

that accounts for all possible phase evolutions $\mathbf{s}_{1:t,b} = [s_{1,b} \cdots s_{t,b}]^T$ over space and time. For the previous computation, we propose a recursive procedure that, at each step t, updates the pdf starting from the previous one, i.e., $p(\mathbf{x}_{t-1,b}|\mathbf{Y}_{1:t-1})$, based on the new measurements made available at time t by the local sensors and also by the sensors in neighboring blocks $\ell \in \Omega_b$. The symbol Ω_b denotes the set of neighboring blocks for block b, here limited to the two adjacent blocks: $\Omega_b = \{b-1, b+1\}$. The recursive procedure is based on a local filtering of the traffic densities in the block b for the computation of $p(\mathbf{x}_{t,b}|\mathbf{s}_{1:t,b}, \mathbf{Y}_{1:t})$ and a cooperative filtering of the traffic phases based on information sharing between blocks for the computation of $p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t})$.

We first observe that, according to the JMLS in Section IV-A, the *a posteriori* pdf of the density conditioned to the phase sequence, i.e., $p(\mathbf{x}_{t,b}|\mathbf{s}_{1:t,b}, \mathbf{Y}_{1:t}) = p(\mathbf{x}_{t,b}|\mathbf{s}_{1:t,b}, \mathbf{y}_{1:t,b})$, can be derived from the local sensor measurements $\mathbf{y}_{1:t,b}$. Being the model of the traffic density linear Gaussian, the pdf and the related density estimate can be updated at each time t, using a local KF conditioned to the phase sequence $\mathbf{s}_{1:t,b}$. On the other hand, taking into account the space–time Markovian chain ruling the phase evolution, the pdf $p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t})$ requires for the update a combination of the local observations in block b, with information provided by the other blocks. This is performed by an interblock BP procedure that allows block b to receive phase beliefs from neighboring blocks $\ell \in \Omega_b$ and optimally fuse these beliefs with the local information. Since the phase sequence distribution is non-Gaussian, PF is used for belief representation.

Once $p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t})$ has been evaluated, to reduce the complexity of the computation of the traffic density pdf in (6), we select the maximum *a posteriori* (MAP) estimate of the phase sequence as

$$\hat{\mathbf{s}}_{1:t,b} = \arg\max_{\mathbf{s}_{1:t,b}} p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t}) \quad (7)$$

and we approximate the pdf (6) as $p(\mathbf{x}_{t,b}|\mathbf{Y}_{1:t}) \approx p(\mathbf{x}_{t,b}|\hat{\mathbf{s}}_{1:t,b}, \mathbf{y}_{1:t,b}) p(\hat{\mathbf{s}}_{1:t,b}|\mathbf{Y}_{1:t})$, with the distribution associated to the most probable phase sequence.[4] One step ahead of a set of parallel KFs is performed for evaluating the pdf's $p(\mathbf{x}_{t,b}|\mathbf{s}_{1:t,b}, \mathbf{y}_{1:t,b})$ for a number of candidate phase trajectories $\mathbf{s}_{1:t,b}$ (the particles, as detailed later) and the corresponding *a posteriori* probabilities $p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t})$; the KF associated to the phase sequence $\hat{\mathbf{s}}_{1:t,b}$ with maximum probability is then selected, providing $p(\mathbf{x}_{t,b}|\hat{\mathbf{s}}_{1:t,b}, \mathbf{y}_{1:t,b})$ (see [32] for details).

The BP procedure follows the guidelines in Section III, with the only difference that the estimation refers to the phase sequence $\mathbf{s}_{1:t,b}$ instead of the density, as the density estimate is performed separately by the KF. The BP algorithm is illustrated in Fig. 8 and detailed in the subsequent section.

A. BP for Cooperative Estimation

The BP procedure requires the definition of node potential $\varphi(\mathbf{s}_{1:t,b}, \mathbf{y}_{t,b})$ for each block b and the pairwise potentials $\psi(\mathbf{s}_{1:t,b}, \mathbf{s}_{1:t,\ell})$ for any pair of adjacent blocks b and $\ell \in \Omega_b$, as introduced in Section III. The node potential is obtained as

$$\varphi(\mathbf{s}_{1:t,b}, \mathbf{y}_{t,b}) = p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t-1}) p(\mathbf{y}_{t,b}|\mathbf{s}_{1:t,b}) \quad (8)$$

where $p(\mathbf{y}_{t,b}|\mathbf{s}_{1:t,b})$ is the current measurement likelihood, and $p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t-1})$ is the *a priori* pdf. The latter is calculated from the *a posteriori* pdf of the previous step, using the transition probabilities $p(s_{t,b}|s_{t-1,b})$ of the Markovian chain describing the temporal evolution within the block, i.e.,

$$p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t-1}) = \sum_{\mathbf{s}_{1:t-1,b}} p(\mathbf{s}_{1:t,b}|\mathbf{s}_{1:t-1,b}) p(\mathbf{s}_{1:t-1,b}|\mathbf{Y}_{1:t-1}). \quad (9)$$

The procedure used for the implementation of the previous equations is in Appendix A.

On the other hand, for the pairwise potential between the adjacent blocks b and $\ell \in \Omega_b$, we exploit the Markovian chain describing the spatial evolution over the blocks [see Fig. 6(b)], as

$$\begin{aligned} \psi(\mathbf{s}_{1:t,b}, \mathbf{s}_{1:t,\ell}) &= \frac{p(\mathbf{s}_{1:t,b}, \mathbf{s}_{1:t,\ell}|\mathbf{Y}_{1:t-1})}{p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t-1}) p(\mathbf{s}_{1:t,\ell}|\mathbf{Y}_{1:t-1})} \quad (10) \\ &\approx \frac{p(s_{t,b}, s_{t,\ell}|\mathbf{Y}_{1:t-1})}{p(s_{t,b}|\mathbf{Y}_{1:t-1}) p(s_{t,\ell}|\mathbf{Y}_{1:t-1})} \\ &= \psi(s_{t,b}, s_{t,\ell}) \quad (11) \end{aligned}$$

where the *a priori* probabilities are obtained from (9). We approximated the computation from (10) and (11) by considering only the current value of the phase on blocks b and ℓ according to the Markovian chain model in Section IV-B, where the interdependence between the phases in adjacent blocks is assumed to be limited to the simultaneous phase configuration. Notice that considering the pairwise potential as dependent on the whole sequences $(\mathbf{s}_{1:t,b}, \mathbf{s}_{1:t,\ell})$ would require handling a number of states increasing exponentially with time, leading to intolerably high computational complexity.

Node and pairwise potentials are used by the BP procedure to evaluate the *a posteriori* pdf $p(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t})$ through repeated approximations. The pdf at iteration k, i.e., $p^{(k)}(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t})$, is obtained by the following two steps.

1) **Message fusion for computation of the block belief.** Block b receives a message $m^{(k)}_{\ell \to b}(s_{t,b})$ from each neighbor $\ell \in \Omega_b$ representing the belief about the block-b phase

[4] The scaling term $p(\hat{\mathbf{s}}_{1:t,b}|\mathbf{Y}_{1:t})$ is irrelevant for density estimation and it is not considered in the KF computations.

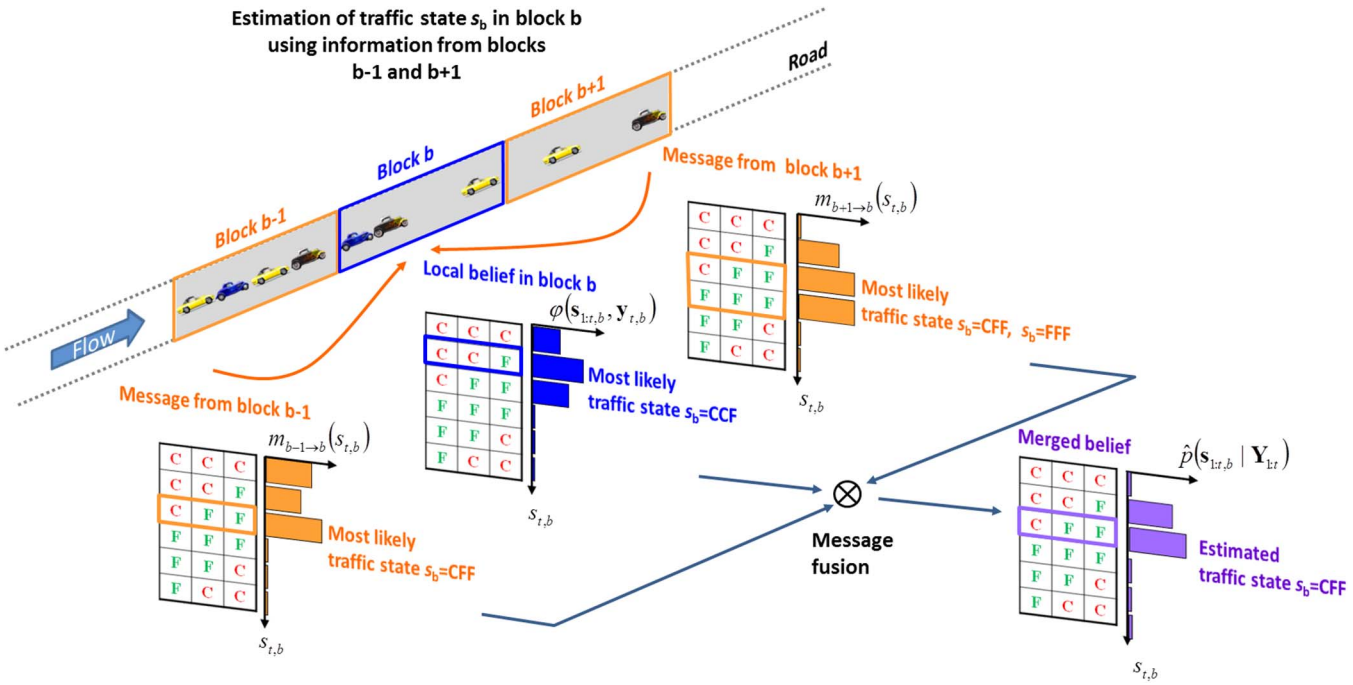

Fig. 8. Illustration of the BP algorithm for the estimation of the traffic state in block b. The estimate is obtained by merging belief information (or messages) provided by neighboring road blocks ($b-1$ and $b+1$) with the information coming from local observations (block b). Beliefs are probability distributions (here represented as bar diagrams) evaluated over all possible traffic phase values, ranging from "CCC" to "FCC". In this example, information from sensors in neighboring blocks allows the system to modify the initial assessment of traffic condition "CCF" into the correct one "CFF", shifting upstream the "CF" breakpoint, due to information propagation over the road blocks.

based on the block-ℓ observations. Block b approximates its own *a posteriori* pdf, fusing the messages from all neighbors, i.e.,

$$p^{(k)}(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t}) \propto \varphi(\mathbf{s}_{1:t,b}, \mathbf{y}_{t,b}) \prod_{\ell\in\Omega_b} m^{(k)}_{\ell\to b}(s_{t,b}) \qquad (12)$$

as shown in Fig. 8.

2) **Update of messages to be sent to neighbors**. Given all messages received from neighbors, block b computes the message to send to each neighbor $\ell \in \Omega_b$ as

$$m^{(k+1)}_{b\to\ell}(s_{t,\ell}) \propto \sum_{\mathbf{s}_{1:t,b}} \varphi(\mathbf{s}_{1:t,b}, \mathbf{y}_{t,b})\psi(s_{t,b}, s_{t,\ell}) \times \prod_{h\in\Omega_b\setminus\ell} m^{(k)}_{h\to b}(s_{t,b}). \qquad (13)$$

The approximated pdf at each block b is used to update the MAP estimate of the phase sequence as

$$\hat{\mathbf{s}}^{(k)}_{1:t,b} = \arg\max_{\mathbf{s}_{1:t,b}} p^{(k)}(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t}). \qquad (14)$$

The BP procedure ends when $|\hat{\mathbf{s}}^{(k+1)}_{1:t,b} - \hat{\mathbf{s}}^{(k)}_{1:t,b}| < \varepsilon$, where ε is a judiciously chosen convergence parameter.

For the implementation of the BP procedure, we represent the beliefs as weighted sets of random particles according to the importance sampling principle

$$p^{(k)}(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t}) = \sum_{i=1}^{N_p} \bar{q}^{(i)}_{t,b}\delta(\mathbf{s}_{1:t,b} - \mathbf{s}^{(i)}_{1:t,b}|\mathbf{Y}_{1:t}) \qquad (15)$$

where $\{\mathbf{s}^{(i)}_{1:t,b}\}_{i=1,\dots,N_p}$ are the particles and $\{\bar{q}^{(i)}_{1:t,b}\}_{i=1,\dots,N_p}$ are the corresponding weights. A finite memory T_{est} is used for the phase sequence tracking, as detailed in Section VII, keeping track of $\mathbf{s}_{t-T_{\text{est}}:t,b}$ instead of $\mathbf{s}_{1:t,b}$. The BP algorithm for the computation of particles and weights is presented in Appendix B.

Remark: For the estimation of the densities $\mathbf{x}_{t,b}$ in the KF, given the phase sequence estimate $\hat{\mathbf{s}}_{1:t,b}$, we need to know also the boundary conditions $\mathbf{u}_{t,b} = [f_{b,\text{in}}(t)\ f_{b,\text{out}}(t)]^T$ for the block b. To draw such information, the data provided to block b by the neighboring blocks $\ell \in \Omega_b$ must include, in addition to the previously defined messages, the traffic densities estimated in the two adjacent blocks. Each neighboring block thereby sends the belief about the block-b phases $\mathbf{s}_{1:t,b}$ and the estimate of its own traffic densities $\mathbf{x}_{t,\ell}$ on the boundary cells. When more complex networks with intersections are considered, the estimated densities need to be exchanged among all the neighbors in the graph. In this case, the computation of boundary conditions has to take into account the splitting matrix defining how flows split at each intersection.

VI. Computational Complexity

The computational complexity is a key factor of our approach to large-scale traffic estimation. Here, we consider the complexity of traffic estimation in a large-scale road with $N_B N_C$ cells. We compare the cooperative approach, which relies on the division of the road into N_B blocks of N_C cells each, with an equivalent centralized approach that applies the PF procedure (without BP) on the whole road as a single block of $N_B N_C$ cells.

Considering the BP procedure (8)–(15), we can divide the estimation process in two steps: the computation of the local

TABLE II
JMLS MATRICES FOR DIFFERENT VALUES OF THE TRAFFIC PHASE FOR A BLOCK OF $N_C = 2$ CELLS

Method	N. phases N_{ph}	N. particles N_p	Complexity per block	Total complexity
Centr.	$2^{N_B N_C+1}$	$O\left(N_{ph}^t\right)$	$O\left(N_p\right)$	$O\left(2^{N_B N_C t}\right)$
Coop.	$2\left(N_C+1\right)$	$O\left(N_{ph}^t\right)$	$O\left(N_p\right)$	$O\left(N_B 2^t N_C^t\right)$

belief or node potential in (8) and (9) and the update of the belief using the information propagated by neighbors through the BP procedure (12)–(14). In the first step, the node potential represented by N_p particles is evaluated following the PF procedure in Appendix A, with computational complexity $O(N_p)$ (see also [32]). In the BP procedure, messages represented by N_p particles are exchanged and multiplied in each block with the node potential. The complexity of this operation is again linear in the number of particles N_p. It follows that the overall complexity per block of both the cooperative approach and the centralized estimation approach is $O(N_p)$. The amount of information exchange can be evaluated, observing that each message is composed by N_p particles and that the convergence is reached usually after 2 to 3 iterations. The load of information exchange is then mainly related to the number of particles as the system reaches convergence in few steps. This is another reason to maintain small the dimension of blocks.

However, the critical point is that the number N_p needed to represent each phase belief is strictly related to the dimension of the phase space, which in turn is linked to the road block size (here expressed in number of cells). For cooperative estimation, we showed in Section IV-A that, in case of small block size, the number of phases that can be defined in a single block grows linearly with the number of cells in the block, i.e., $N_{\mathrm{ph}} = 2(N_C + 1)$. For the centralized approach, on the other hand, the number of cells is $N_B N_C$; since this corresponds to a large block size, the single breakpoint assumption is not realistic anymore and the number of phases grows exponentially as $N_{\mathrm{ph}} = 2^{N_B N_C+1}$. Moreover, the algorithm attempts to estimate the sequence of phases $\mathbf{s}_{1:t,b}$, and it is well known in the literature that the number of possible trajectories grows with time as N_{ph}^t [32], [36] (notice that a finite memory t can be typically assumed for practical tracking). Since the number of particles required for belief representation is, to the first-order approximation, linear in the support of the phase-sequence variable [37], it follows that, in the worst case (i.e., assuming the whole set of sequences as support of the distribution), the number of particles N_p used to track the trajectories grows exponentially as $O(N_{\mathrm{ph}}^t)$. In practical situations, the support is smaller and the complexity is reduced. The aforementioned is therefore an overestimation of the computational cost, as the particles sample only the most likely trajectories in the space, not all of them. In any case, the relevant point is the relative gain between the computational complexity of the centralized and cooperative approaches that can be inferred from the results summarized in Table II.

We can conclude that, in the worst case, the complexity is $O(N_C^t)$ per block for the distributed approach and $O(2^{N_B N_C t})$ for the centralized approach. The latter is clearly unfeasible,

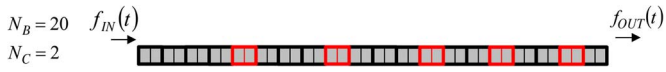

Fig. 9. Simulation settings for $N_B = 20$ with $N_C = 2$ cells/each. Central blocks (red) may have larger measurement errors.

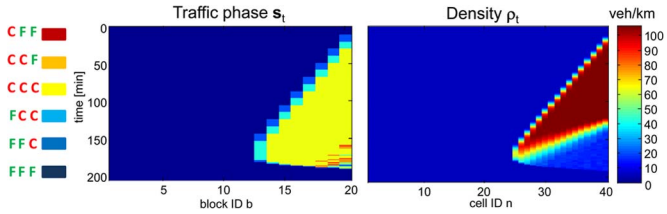

Fig. 10. Real phase sequence and densities for a scenario with generation and dissolution of traffic congestion.

and this is the reason that motivated the adoption of the JMLS-PF procedure only for small blocks in previous works [15], [17]. The cooperative approach enables the application to large-scale networks, as the overall complexity is only linear in the dimension of the network in number of blocks, rather than exponential.

Furthermore, we observe that the size of the road block is typically a tradeoff between the desired estimation accuracy and the computational complexity: the larger is the block, the higher is the accuracy (as the estimate exploits measurements provided by a larger set of sensors) and the higher is the computational cost (as the number of variables that need to be jointly handled increases). The cooperative approach allows extending the range of observation to multiple blocks, still maintaining the complexity to the level of a single-block processing and linear in the number of blocks (see Table II).

VII. PERFORMANCE ASSESSMENT

Here, we present numerical results on traffic density estimation. Since joint estimation of traffic densities over N_B blocks would be unfeasible by conventional Bayesian estimation, as explained in the previous section, here we compare the proposed cooperative method with the estimation obtained by N_B separate tracking filters, i.e., one for each block, each based on the Bayesian model in Section IV-A. To gain an insight into the estimation procedure, we consider a simple synthetic scenario that will be incrementally expanded, adding more elements and traffic situations. The scenario is shown in Fig. 9: it is a single-lane road divided into $N_B = 20$ blocks, with $N_C = 2$ cells each of length $l = 750$ m. The parameters used for the traffic simulation are as follows: free-flow speed $v_f = 25$ m/s $= 90$ km/h, backward propagation speed $\omega = 4$ m/s $= 14.4$ km/h, sampling interval $\Delta t = 20$ s (equal to the measurement interval), and maximum density $\rho_{\max} = 80$ veh/cell $= 106.67$ veh/km. The traffic densities and the related measurements are simulated according to the model (4). We assume that the boundary conditions $f_{\mathrm{IN}}(t)$ and $f_{\mathrm{OUT}}(t)$ are known and error-free. Performance is evaluated in terms of the root-mean-square error (RMSE) of the density estimate.

We consider a first simulation in which the measurement error $\mathbf{r}_{b,t} \sim \mathcal{N}(0, \sigma_{r,b}^2 \mathbf{I}_{N_C})$ in block b is stationary over the blocks with $\sigma_{r,b}^2 = \sigma_r^2$. We simulate a scenario with generation, backpropagation, and dissolution of a congestion. The

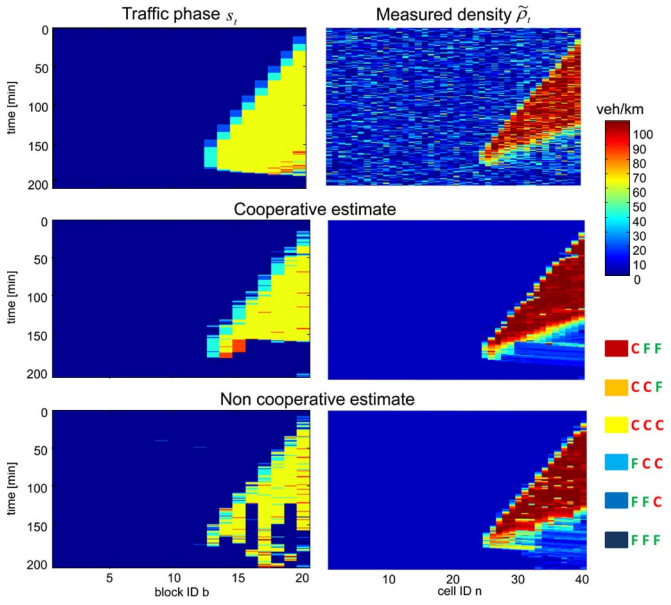

Fig. 11. Traffic density estimation and phase sequences for $\sigma_r = 20$ veh/km. Top: true traffic phase and measured densities. Center: cooperative estimation. Bottom: noncooperative estimation.

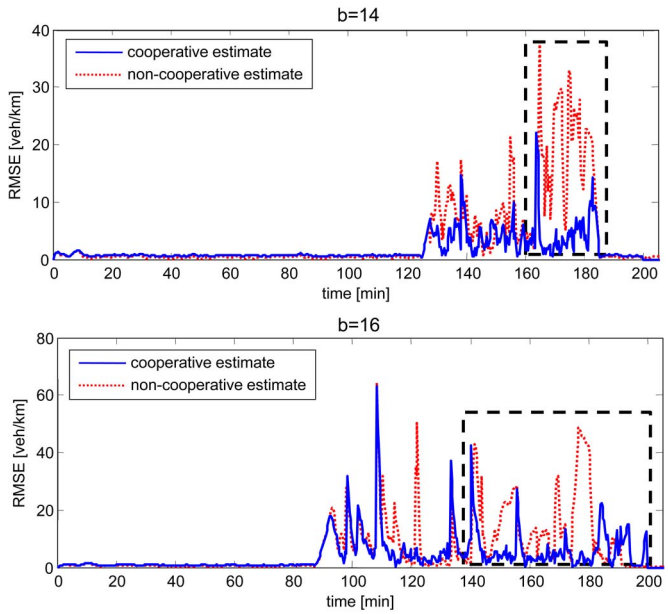

Fig. 12. RMSE on blocks $b = \{14, 16\}$, for the scenario in Fig. 11. Squares mark sections where the cooperative method provides major gain.

simulated traffic densities and the corresponding phase sequences are shown in Fig. 10. The number of particles is optimized studying the RMSE. Numerical results show that an increasing number of particles allows for a more accurate approximation of the traffic pdf and consequently for a lower estimation error, up to the limit $N_p = 100$, where a floor is observed on the RMSE versus N_p performance. No significant improvements are observed above this value.

In Fig. 11, a comparison of the density estimate accuracy for the two considered approaches is shown for the case $\sigma_r = 20$ veh/km, with $N_p = 100$. The figure shows the true phase sequences (top left figure), the density data (top right figure), and the related estimates obtained by the cooperative (central figures) and noncooperative approaches (bottom figures). In both the estimation approaches, blocks exchange estimates of the boundary flows to let neighbors compute inflows and outflows. The cooperative approach is shown to provide accurate tracking results on both the phase sequence $\mathbf{s}_{1:t}$ and the densities $\mathbf{X}_{1:t}$. Higher accuracy is observed during the congestion dissolution: blocks use the message-passing procedure to share information about the backward/forward propagation of traffic waves through the subnetworks; therefore, if a block is estimating traffic congestion, it can inform neighboring blocks about the possible propagation or dissolution of the jam wave through the boundaries (i.e., about the predicted traffic evolution over nearby blocks). In particular, the RMSE for blocks $b = 14$ and $b = 16$ is shown in Fig. 12. Black squares mark areas where the cooperative estimate provides a significant gain over the noncooperative approach; they mostly correspond to the congestion dissolution process.

The second simulation describes a realistic situation, in which a subset of sensors experiences lower accuracy due to failure or to the use of heterogeneous sensing systems with different reliability. The cooperative method reveals here its major advantages because it allows the noisy blocks to exploit measurements collected by other (possibly more reliable) sensors. In our simulation scenario, the measurements collected on blocks $b \in \mathcal{B}_0 = \{5, 9, 13, 16, 19\}$ are affected by a higher error, with $\sigma_{r,b} = 20$ veh/km, while the other blocks $b \notin \mathcal{B}_0$ have uncertainty $\sigma_{r,b} = 2.67$ veh/km. To appreciate the advantages of the cooperative method, we analyze in detail one of these noisy blocks, with index $b = 16$, for $N_p = 100$. The two methods are compared in Fig. 13. The figure shows the MAP estimate of the phase sequence $\hat{\mathbf{s}}^{\mathrm{MAP}}_{1:t,16}$ (left panels) and the density estimate obtained by a KF conditioned on $\hat{\mathbf{s}}^{\mathrm{MAP}}_{1:t,16}$ (right panels) for the cooperative method (bottom) and the noncooperative method (top). When the cooperative approach is used, the error is lower because the 16th block can exploit more reliable information coming from other blocks in the network. The cooperative method provides a higher accuracy of density estimation compared to the noncooperative approach with comparable computational complexity. This is very useful for real-time traffic estimation in road networks in which some critical areas are covered by less reliable sensors. The proposed method allows to overcome problems related to wrong values of the traffic variables sensed on those areas. In practice, we do not exclude the noisy sensors; we rather attempt to integrate all the data, taking into account the different reliability degrees and relying more on measurements provided by less noisy nearby sensors.

The estimation results for all the blocks are shown in Fig. 14, in which the two approaches are compared with the true data. The interesting point is that, in the cooperative approach, not only the noisy blocks $b \in \mathcal{B}$ take advantages of the cooperation but also the neighboring blocks. This behavior can be explained by considering that, in both the approaches, blocks exchange estimates of the boundary flows to let neighbors compute inflows and outflows. In the cooperative algorithm, block $b = 16$ sends to its neighbors reliable estimates, while in the noncooperative approach, the block estimate is affected by a large error that propagates to the neighbors through the boundary flows.

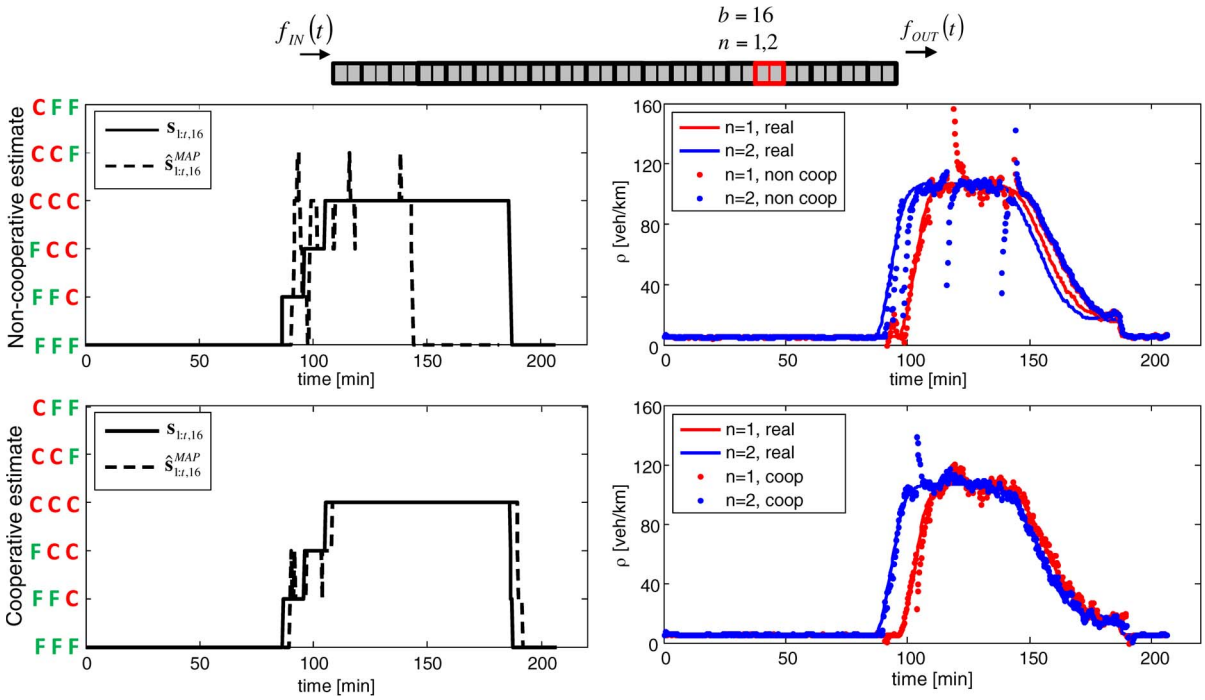

Fig. 13. Estimation on the block $b = 16$ equipped with corrupted sensors with $\sigma_{r,b} = 20$ veh/km. Results from noncooperative estimation (top) and cooperative estimation (bottom) methods.

These results underline the advantage of cooperative processing in the case of a lack or malfunction of some sensors, as the method allows to obtain good estimation results over the whole road also in areas not covered by good sensors. This is important in both urban and freeway environments, in which the problem of density and flow estimation in presence of bad data is of high relevance.

As final example, we consider a more complex scenario, with a longer road section of $N_B = 45$ blocks and a couple of congestions backpropagating along the road. We use the same traffic parameters as in the previous cases. We assume first that all sensors have the same accuracy $\sigma_r = 13.33$ veh/km. In Fig. 15, we compare the density estimates for the cooperative and noncooperative estimation. The cooperative estimation is able to reconstruct the traffic field, while the noncooperative approach is subject to systematic errors in the phase estimate (highlighted by circles).

The cooperative estimate is general enough to be applied to a large variety of traffic networks, by simply adapting the model presented in Section IV-A to the specific case. As an example, we analyze here the scenario on the top of Fig. 16 obtained from the previous scenario, by adding an onramp and an offramp modeled as described in [38]. The figure shows the true density (top subfigure), the measured density (central subfigure), and the estimation results (bottom subfigure). All the blocks are affected by the same accuracy $\sigma_r = 6.67$ veh/km. From numerical results, we can conclude that the proposed method can reach a high accuracy of estimation even in complex networks similar to the one presented. Moreover, the computational cost in each subnetwork grows only linearly with the number of neighboring blocks (in this case, 2).

We finally consider the case of the presence of malfunctions in some sensors modeled with a higher measurement error, i.e., $\sigma_r = 13.33$ veh/km, as shown in Fig. 17. In this case, we compare the estimate obtained by the cooperative and noncooperative approaches. As in previous examples, when subnetworks cooperate, problems related to malfunctions are overcome using information from more reliable sensors. This prevents the arising of situations (highlighted with white circles in the figure) in which the reconstruction becomes unreliable.

VIII. Conclusion

In this paper, we have presented a cooperative method for large-scale traffic estimation based on the splitting of the road network into small subnetworks. As a first step, each subnetwork performs its own traffic estimation, and then it updates the estimate by exploiting traffic information coming from neighboring subnetworks. This procedure allows mutual sharing of traffic data as for a global estimation, but with a much lower computational cost, which is proportional to the dimension of the subnetwork and to the number of neighboring subnetworks.

We used a macroscopic approach to traffic modeling, which is based on the CTM extended with a stochastic component to account for the randomness of traffic. For the estimation, we used in each subnetwork a JMLS based on a linearized version of the CTM and a PF approach to track the traffic phase sequence. For making the subnetworks share information on their estimates, we proposed an iteratively message-passing procedure based on BP.

We evaluated the performance simulating different traffic scenarios, ranging from simple cases to more complex situations with onramp/offramp and sensor failures. Our approach was shown to provide a significant gain, in terms of accuracy,

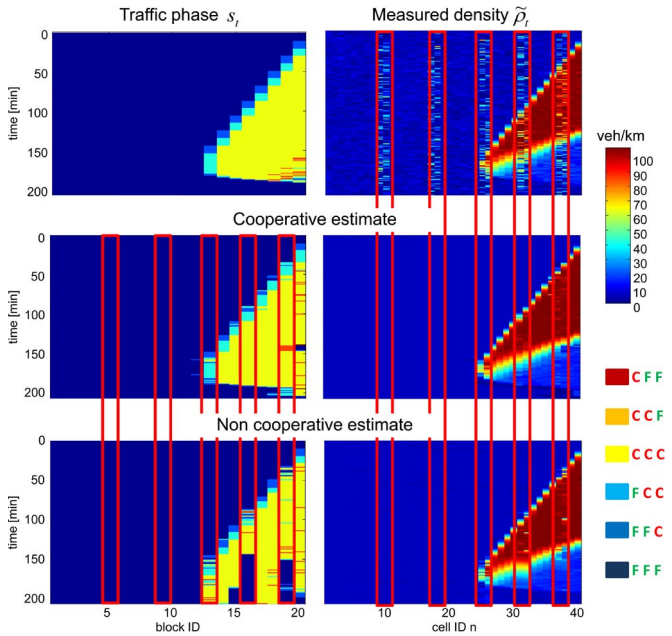

Fig. 14. Traffic density estimation and phases sequence in case of $\sigma_{r,b} = 20$ veh/km, with $b \in \mathcal{B}_0 = \{5, 9, 13, 16, 19\}$. Top: true traffic phase and measured densities. Center: cooperative estimate. Bottom: noncooperative estimate. The simulation scenario is the same as in Fig. 11 but with varying measurement accuracy over the blocks. Red boxes indicate sensors with larger error.

when compared to noncooperative methods without information sharing among subnetworks (i.e., where only flows on boundaries are communicated). The method was shown to be particularly useful in the case of fragmented or low-density observation of the traffic field due to sensor failures (e.g., loops). Furthermore, in large-scale networks, the cooperative method has the potential to exploit measurements collected in different blocks of the road network, as for an equivalent centralized estimation that jointly handles all the data, but with much lower complexity. As a future development, the model will be calibrated and validated on real data and large-scale scenarios to show the benefits provided by the distribution of processing over the subnetworks.

APPENDIX A

We follow the PF approach in [32] for the computation of the node potential for node b, i.e., $\varphi(\mathbf{s}_{1:t,b}, \mathbf{y}_{t,b})$. The method consists of three steps.

1) **Update.** The node potential is represented as

$$\varphi(\mathbf{s}_{1:t,b}, \mathbf{y}_{t,b}) = \sum_{i=1}^{N_p} q_{t,b}^{(i)} \delta \quad \mathbf{s}_{1:t,b} - \mathbf{s}_{1:t,b}^{(i)} \tag{16}$$

with particles $\{\mathbf{s}_{1:t,b}^{(i)}\}_{i=1}^{N_p}$ and weights $\{q_{t,b}^{(i)}\}_{i=1}^{N_p}$, for N_p particles. The evaluation of the weights requires a number of one-step-ahead KFs on the state variable $\mathbf{x}_t$; see [32] for details.

2) **Resampling**. Particles are resampled according to the weights $\{q_{t,b}^{(i)}\}_{i=1}^{N_p}$ obtaining the new set of particles $\{\tilde{\mathbf{s}}_{1:t,b}^{(i)}\}_{i=1}^{N_p}$ with weights $\tilde{q}_{t,b}^{(i)} = (1/N_p)$.

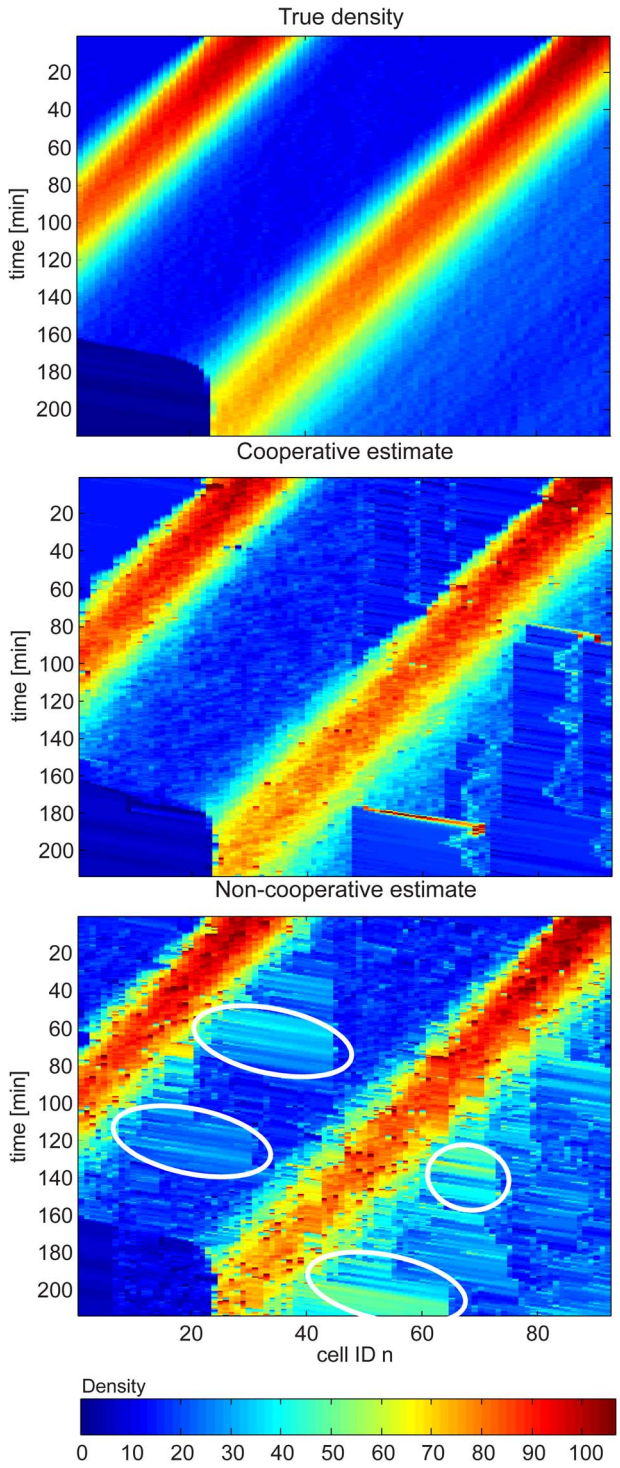

Fig. 15. Traffic density estimation on a longer road in case of $\sigma_r = 13.33$ veh/km. Top: true densities. Center: cooperative estimate. Bottom: non-cooperative estimate. Circles reveal areas with systematic error in the phase estimation. Densities are expressed in vehicles per kilometer.

3) **Propagation**. Particles are propagated from t to $t+1$, using the transition probabilities of the Markov chain $p(s_{t+1,b}|s_{t,b})$ [see Fig. 6(a)] and computing the next phase $s_{t+1,b}^{(i)}$ for each particle $s_{t,b}^{(i)}$. We thus obtain the new sequences of phases $\{\mathbf{s}_{1:t+1,b}^{(i)}\} = \{\mathbf{s}_{1:t,b}^{(i)}, s_{t+1,b}^{(i)}\}$.

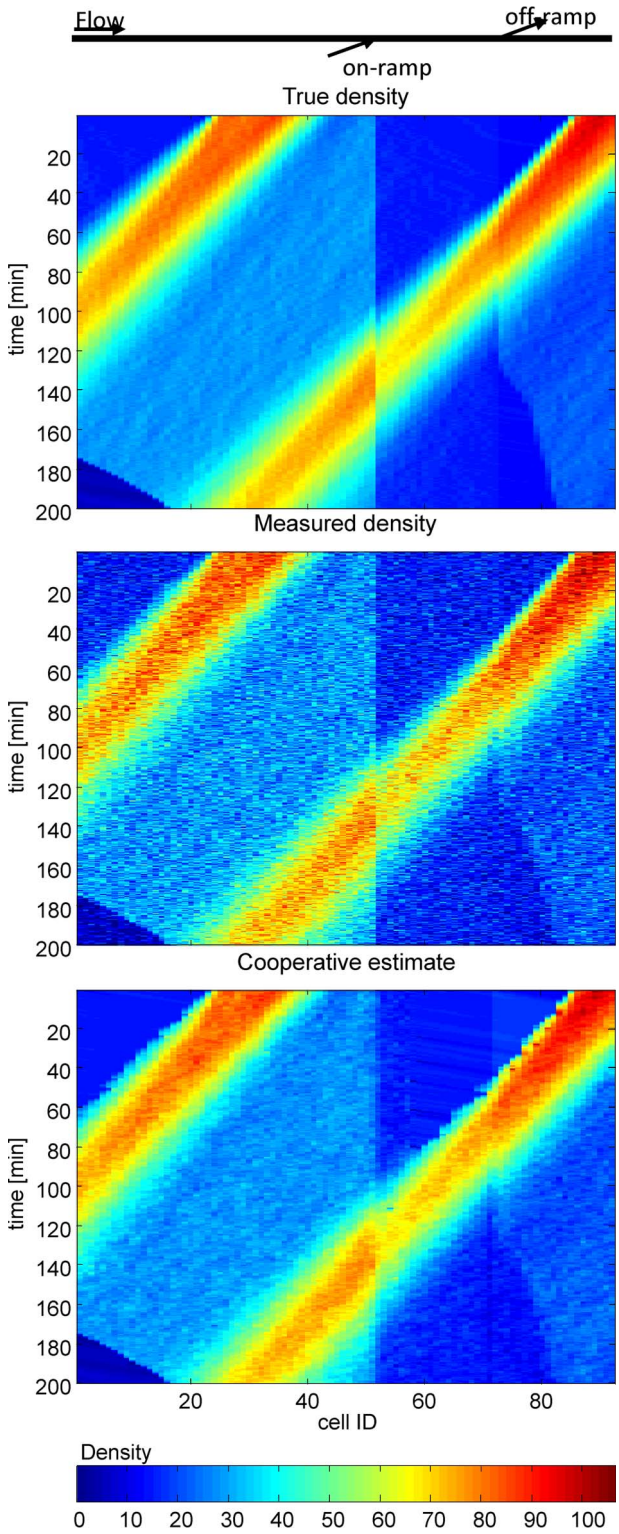

Fig. 16. Traffic density estimation on a road with ramps in case of $\sigma_r = 6.67$ veh/km. Top: true densities. Center: measured densities. Bottom: cooperative estimate. Densities are expressed in vehicles per kilometer.

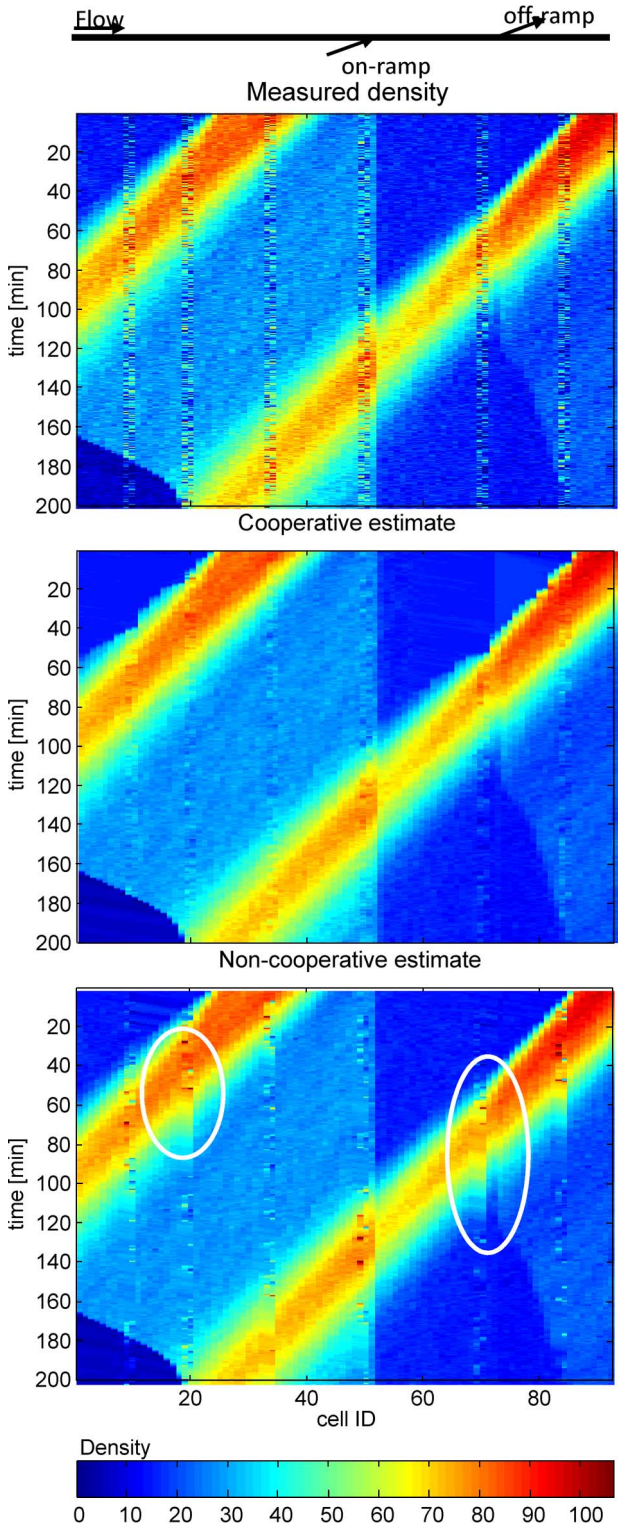

Fig. 17. Traffic density estimation on a longer road in case of sensor malfunctions, with $\sigma_r = 13.33$ veh/km. Top: measured densities. Center: cooperative estimate. Bottom: noncooperative estimate. Circles reveal areas with systematic error in the phase estimation. Densities are expressed in vehicles per kilometer.

APPENDIX B

We present here the BP algorithm used to implement the two steps presented in Section V-A. We focus our attention on the BP on block b at time t.

1) At iteration k, the node b receives messages from nodes $\ell \in \Omega_b$ as

$$m^{(k)}_{\ell \to b}(s_{t,b}) = \sum_{i=1}^{N_p} \gamma^{(i,k)}_{\ell \to b} \delta \quad s_{t,b} - s^{(i,k)}_{t,\ell \to b} \tag{17}$$

where $\{s_{t,\ell\to b}^{(i,k)}\}_{i=1}^{N_p}$ are the particles, which are defined in the set $\mathcal{S}'$, and $\{\gamma_{\ell\to b}^{(i,k)}\}_{i=1}^{N_p}$ are the related weights. Using these messages and its own node potential, node b computes its belief as in (12). Note that the node potential is written as in (16) with particles $\mathbf{s}_{1:t,b}^{(i)}$ taking value in the Cartesian product $\mathcal{S}'^t$, while messages (17) coming from the neighbors $\ell \in \Omega_b$ have weights $\gamma_{\ell\to b}^{(i,k)}$ associated only to the time-t phase $s_{t,\ell\to b}^{(i,k)} \in \mathcal{S}'$, not the entire sequence of t phases. The application of (12) requires the two distributions to be defined over the same set of samples. To solve this problem, we consider two points. The first point is that each message $m_{\ell\to b}^{(k)}(s_{t,b})$ is equivalent to a discrete distribution defined on $\mathcal{S}'$ and can be written as

$$m_{\ell\to b}^{(k)}(s_{t,b}) = \sum_{s=1}^{N_{\text{ph}}} \eta_{\ell\to b}^{(s,k)} \delta(s_{t,b} - s) \tag{18}$$

where $\eta_{\ell\to b}^{(s,k)} = \sum_{\{i:s_{t,\ell\to b}^{(i,k)}=s\}} \gamma_{\ell\to b}^{(i,k)}$. As second point, we consider the weight $q_{t,b}^{(i)}$ in (16) as associated to the last phase of the sequence $\mathbf{s}_{1:t,b}^{(i)}$ that we call $s_{t,b}^{(i)}$. Applying these approximations to (12), we can associate to each particle $\mathbf{s}_{1:t,b}^{(i)}$ the weight

$$\bar{q}_{t,b}^{(i,k)} \propto q_{t,b}^{(i)} \prod_{\ell\in\Omega_b} \eta_{\ell\to b}^{s_{t,b}^{(i)},k} \tag{19}$$

that needs to be normalized. The belief is thus written as $p^{(k)}(\mathbf{s}_{1:t,b}|\mathbf{Y}_{1:t}) = \sum_{i=1}^{N_p} \bar{q}_{t,b}^{(i,k)} \delta(\mathbf{s}_{1:t,b} - \mathbf{s}_{1:t,b}^{(i)}|\mathbf{Y}_{1:t})$ at each block b. The MAP estimate of the phase sequence at iteration k is $\hat{\mathbf{s}}_{k,1:t,b} = \mathbf{s}_{1:t,b}^{(\bar{i})}$, with $\bar{i} = \arg\max_i \bar{q}_{t,b}^{(i,k)}$. The BP procedure ends when $|\hat{\mathbf{s}}_{k+1,1:t,b} - \hat{\mathbf{s}}_{k,1:t,b}| < \varepsilon$. When the convergence is reached, weights are renamed simply $\{\bar{q}_{t,b}^{(i)}\}_{i=1}^{N_p}$.

2) The message that node b sends to its neighbor ℓ at iteration $k+1$ is computed following the approach in [27]. Particles are propagated from node b to node ℓ using the distribution $\psi(s_{t,b}, s_{t,\ell})$ to obtain the new set $\{s_{t,b\to\ell}^{(i,k+1)}\}_{i=1}^{N_p}$ and compute their weights as

$$\gamma_{b\to\ell}^{(j,k+1)} = \frac{\sum_{i=1}^{N_p} \psi\quad s_{t,b}^{(i)}, s_{t,b\to\ell}^{(j,k+1)}\quad \bar{q}_{t,b}^{(i)} \prod_{h\in\Omega_b\setminus\ell} \eta_{h\to b}^{s_{t,b}^{(i)},k}}{\sum_{u=1}^{N_p} \psi\quad s_{t,b}^{(u)}, s_{t,b\to\ell}^{(j,k+1)}}.$$

The new message can be written as

$$m_{b\to\ell}^{(k+1)}(s_{t,\ell}) = \sum_{i=1}^{N_p} \gamma_{b\to\ell}^{(i,k+1)} \delta\quad s_{t,\ell} - s_{t,b\to\ell}^{(i,k+1)}\quad . \tag{20}$$

As the last step, the message is resampled in order to obtain N_p new samples with weights equal to $1/N_p$.

REFERENCES

[1] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang, "Review of road traffic control strategies," *Proc. IEEE*, vol. 91, no. 12, pp. 2043–2067, Dec. 2003.

[2] A. Pascale, M. Nicoli, F. Deflorio, B. Dalla Chiara, and U. Spagnolini, "Wireless sensor networks for traffic management and road safety," *IET Intell. Transp. Syst.*, vol. 6, no. 1, pp. 67–77, Mar. 2012.

[3] D. Ngoduy, S. P. Hoogendoorn, and R. Liu, "Continuum modelling of cooperative traffic flow dynamics," *J. Phys. A, Stat. Mech. Appl.*, vol. 388, no. 13, pp. 2705–2716, Jul. 2009.

[4] F. Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 630–638, Sep. 2010.

[5] N.-E. El Faouzi, H. Leung, and A. Kurian, "Data fusion in intelligent transportation systems: Progress and challenges—A survey," *Inf. Fusion*, vol. 12, no. 1, pp. 4–10, Jan. 2011.

[6] S. P. Hoogendoorn, V. L. Knoop, and H. J. van Zuylen, "Robust control of traffic networks under uncertain conditions," *J. Adv. Transp.*, vol. 42, no. 3, pp. 357–377, Jul. 2008.

[7] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transp. Res. B, Methodol.*, vol. 28, no. 4, pp. 269–287, Aug. 1994.

[8] A. Pascale, G. Gomes, and M. Nicoli, "Estimation of highway traffic from sparse sensors: Stochastic modeling and particle filtering," in *Proc. 38th IEEE ICASSP*, Vancouver, BC, Canada, 2013, pp. 6158–6162.

[9] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*. San Mateo, CA, USA: Morgan Kaufmann, 1988.

[10] S. P. Hoogendoorn and P. H. L. Bovy, "State-of-art of vehicular traffic flow modelling," *Proc. Inst. Mech. Eng. I, J. Syst. Control Eng.*, vol. 215, no. 4, pp. 283–303, Jun. 2001.

[11] B. L. Smith, B. M. Williams, and R. Keith Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transp. Res. C, Emerging Technol.*, vol. 10, no. 4, pp. 303–321, Aug. 2002.

[12] R. Chrobok, O. Kaumann, J. Wahle, and M. Schreckenberg, "Different methods of traffic forecast based on real data," *Eur. J. Oper. Res.*, vol. 155, no. 3, pp. 558–568, Jun. 2004.

[13] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transp. Res. B, Methodol.*, vol. 18, no. 1, pp. 1–11, Feb. 1984.

[14] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short-term traffic forecasting: Overview of objectives and methods," *Transp. Rev.*, vol. 24, no. 5, pp. 533–557, Sep. 2004.

[15] X. Sun, L. Munoz, and R. Horowitz, "Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control," in *Proc. 42nd IEEE Conf. Decision Control*, 2003, vol. 6, pp. 6333–6338.

[16] L. Munoz, X. Sun, R. Horowitz, and L. Alvarez, "Traffic density estimation with the cell transmission model," in *Proc. IEEE Amer. Control Conf.*, 2003, vol. 5, pp. 3750–3755.

[17] K. Stanková and B. De Schutter, "On freeway traffic density estimation for a jump Markov linear model based on Daganzo's cell transmission model," in *Proc. 13th IEEE ITSC*, 2010, pp. 13–18.

[18] L. Mihaylova, R. Boel, and A. Hegyi, "Freeway traffic estimation within particle filtering framework," *Automatica*, vol. 43, no. 2, pp. 290–300, Feb. 2007.

[19] A. Pascale and M. Nicoli, "Adaptive Bayesian network for traffic flow prediction," in *Proc. IEEE SSP Workshop*, Nice, France, 2011, pp. 177–180.

[20] S. Sun, C. Zhang, and G. Yu, "A Bayesian network approach to traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 124–132, Mar. 2006.

[21] C. M. Queen and C. J. Albers, "Intervention and causality: Forecasting traffic flows using a dynamic Bayesian network," *J. Amer. Stat. Assoc.*, vol. 104, no. 486, pp. 669–681, Jun. 2009.

[22] E. Castillo, J. M. Menéndez, and S. Sànchez-Cambronero, "Predicting traffic flow using Bayesian networks," *Transp. Res. B, Methodol.*, vol. 42, no. 5, pp. 482–509, Jun. 2008.

[23] S. Sun and X. Xu, "Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 466–475, Jun. 2011.

[24] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphical lasso and neural networks," *J. Transp. Eng.*, vol. 138, no. 11, pp. 1358–1367, Nov. 2012.

[25] C. Furtlehner, J. M. Lasgouttes, and A. de La Fortelle, "A belief propagation approach to traffic prediction using probe vehicles," in *Proc. IEEE ITSC*, Seattle, WA, USA, 2007, pp. 1022–1027.

[26] L. Mihaylova, A. Hegyi, A. Gning, and R. K. Boel, "Parallelized particle and Gaussian sum particle filters for large-scale freeway traffic systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 36–48, Mar. 2012.

[27] D. Fontanella, M. Nicoli, and L. Vandendorpe, "Bayesian localization in sensor networks: Distributed algorithm and fundamental limits," in *Proc. IEEE ICC*, Cape Town, South Africa, 2010, pp. 1–5.

[28] A. T. Ihler, J. W. Fisher, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for self-localization of sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 809–819, Apr. 2005.

[29] L. E. Doyle, A. C. Kokaram, S. J. Doyle, and T. K. Forde, "Ad hoc networking, Markov random fields, and decision making," *IEEE Signal Process. Mag.*, vol. 23, no. 5, pp. 63–73, Sep. 2006.

[30] M. Cetin, L. Chen, J. W. Fisher, A. T. Ihler, R. L. Moses, and M. J. Wainwright, "Distributed fusion in sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 42–55, Jul. 2006.

[31] J. Lighthill and G. B. Whitham, "On kinematic waves. II. A theory of traffic flows on long crowded roads," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 229, no. 1178, pp. 317–345, May 1955.

[32] A. Doucet, N. J. Gordon, and V. Krishnamurthy, "Particle filters for state estimation of jump Markov linear systems," *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 613–624, Mar. 2001.

[33] R. Couillet and M. Debbah, "Signal processing in large systems: A new paradigm," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 24–39, Jan. 2013.

[34] M. Nicoli and D. Fontanella, "Fundamental performance limits of TOA-based cooperative localization," in *Proc. IEEE ICC*, Dresden, Germany, 2009, pp. 1–5.

[35] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky, "Nonparametric belief propagation," in *Proc. IEEE Conf. CVPR*, 2003, vol. 1, pp. 605–612.

[36] A. Logothetis and V. Krishnamurthy, "Expectation maximization algorithms for MAP estimation of jump Markov linear systems," *IEEE Trans. Signal Process.*, vol. 47, no. 8, pp. 2139–2156, Aug. 1999.

[37] D. Fox, "Adapting the sample size in particle filters through KLD-sampling," *Int. J. Robot. Res.*, vol. 22, no. 12, pp. 985–1003, Dec. 2003.

[38] G. Gomes, R. Horowitz, A. A. Kurzhanskiy, P. Varaiya, and J. Kwon, "Behavior of the cell transmission model and effectiveness of ramp metering," *Transp. Res. C, Emerging Technol.*, vol. 16, no. 4, pp. 485–513, Aug. 2008.