Mining Road Network Correlation for Traffic Estimation via Compressive Sensing

Zhidan Liu, *Member, IEEE*, Zhenjiang Li, *Member, IEEE*, Mo Li, *Member, IEEE*, Wei Xing, and Dongming Lu

***Abstract*—This paper presents a transport traffic estimation method which leverages road network correlation and sparse traffic sampling via the compressive sensing technique. Through the investigation on a traffic data set of more than 4400 taxis from Shanghai city, China, we observe nontrivial traffic correlations among the traffic conditions of different road segments and derive a mathematical model to capture such relations. After mathematical manipulation, the models can be used to construct representation bases to sparsely represent the traffic conditions of all road segments in a road network. With the trait of sparse representation, we propose a traffic estimation approach that applies the compressive sensing technique to achieve a city-scale traffic estimation with only a small number of probe vehicles, largely reducing the system operating cost. To validate the traffic correlation model and estimation method, we do extensive trace-driven experiments with real-world traffic data. The results show that the model effectively reveals the hidden structure of traffic correlations. The proposed estimation method derives accurate traffic conditions with the average accuracy as 0.80, calculated as the ratio between the number of correct traffic state category estimations and the number of all estimation times, based on only 50 probe vehicles' intervention, which significantly outperforms the state-of-the-art methods in both cost and traffic estimation accuracy.**

***Index Terms*—Correlation modeling, traffic estimation, compressive sensing.**

I. Introduction

UNDERSTANDING traffic conditions is crucial in urban cities, which used to incur heavy road infrastructure constructions, e.g., inductive loop detectors [8], and close-circuit cameras [18]. Due to the high deploying costs, however, it is prohibitive to densely adopt them in the city scale, which largely limits the coverage. Recent studies leverage roving vehicles on roads as probes to estimate the traffic conditions [11], [24], [30]. Vehicles, equipped with GPS, can periodically report their current locations, driving speeds or directions via certain data delivery scheme [14]. With such on-site traffic information, we can instantly estimate traffic speeds of the roads covered by probe vehicles. For example, Google employs probe cars to measure road speeds and summarizes the traffic speeds of different roads on Google Map. Using roving vehicles largely extends the coverage of traffic estimation. Existing approaches employ extensive probe vehicles to cover the interested roads, and they are often limited by the number of participating vehicles in acquiring the complete traffic map due to privacy concerns or energy expenditure. A particular road may not have probe vehicles at all times.

In this paper, we explore the correlations among traffic conditions of different roads and propose a method that recovers the entire traffic map from sparse traffic samplings. We perform a thorough investigation on a traffic data set of more than 4400 taxis from Shanghai city, China, where we observe interesting traffic correlations. The traffic conditions among road segments may relate to each other *directly* or *indirectly*. For example, the traffic conditions of two cascaded road segments are usually directly correlated. At a crossroad, however, the traffic of one road segment may not only relate to the traffic coming from its cascaded road segment but also to the traffic redirected from the intersecting road segment in an indirect way. For two parallel roads to a same direction, their traffics may alternatively shift and get balanced between each other and jointly correlate with the traffics of the intermediate road segments connecting them. In most cases, the traffic condition of one road is not solely related to any particular individual road but several ones. The detailed traffic correlations among roads are non-trivial. Different road segments impose different impacts on the local traffic, and some critical road segments may have impacts on the traffic conditions of many other road segments. Generally, it is not easy to reveal the complicated traffic affinities among road segments with straightforward lookup over the raw traffic data. After comprehensively investigating the traffic correlations, we turn to mathematical modeling and capture them with an Multiple Linear Regression (MLR) model. The correlation model is effective in two aspects: (1) It implicitly indicates the set of key road segments in a road network which impact the overall traffic most. (2) Based on the model we could form a representation space, which sparsely represents the traffic conditions of all roads in the road network. Unveiling the sparse property, hidden inside the raw traffic condition data, makes it

Manuscript received May 18, 2015; revised September 22, 2015; accepted December 28, 2015. Date of publication January 25, 2016; date of current version June 24, 2016. This work was supported in part by Singapore MOE AcRF under Grant MOE2012-T2-1-070 and Grant MOE2013-T1-002-005, by the NTU NAP under Grant M4080738.020, by the NSFC under Grant 61303233, by the City University of Hong Kong under Project 7200480/CS, and by the National Key Research and Development Program under Grant 2014BAG01B02 and Grant 2014BAG01B03. The Associate Editor for this paper was H. Van Lint.

Z. Liu was with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. He is now with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: liuzhidan@ntu.edu.sg).

Z. Li is with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: zhenjiang.li@cityu.edu.hk).

M. Li is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: limo@ntu.edu.sg).

W. Xing and D. Lu are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: wxing@zju.edu.cn; ldm@zju.edu.cn).

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.2016.2514519

possible that the traffic conditions of the entire road network can be timely estimated by only sparse traffic sensing, leveraging the recent compressive sensing technique.

The contributions of this paper can thus be summarized as follows. Through a comprehensive study on a real-world traffic data set, we develop an MLR based mathematical model to unveil the hidden structure of urban traffics and capture the traffic correlations, which is useful for urban traffic analysis, e.g., traffic estimation, and city planning, e.g., identifying the critical road segments for the road network optimization. The development of the MLR model is computationally efficient and the traffic estimation method is quite lightweight. We further introduce a region partition mechanism for large scale road networks to preserve high estimation accuracy yet with a significant computation overhead reduction. Experimental results in Section V for a large road network with 1826 road segments covering 50.70 km^2, the majority of central areas of Shanghai city, show that the MLR models can be trained with about 8 mins, and the traffic estimation for the whole road network can be finished within 0.3 seconds, which demonstrate the applicability and scalability of our methods. The MLR models are used to construct a representation basis. Benefiting from the sparse representation, we design a traffic estimation method via the compressive sensing technique with only sparse traffic samplings. We evaluate the correlation model and the traffic estimation method with extensive trace-driven experiments. The results show that the developed model effectively captures the underlying traffic correlation. More than half of the road segments in our tested area correlate with other road segments indirectly. The experimental results also validate the effectiveness of our traffic estimation method. Based on the speed reports collected from only 50 probe vehicles, we can accurately estimate the traffic conditions of all 386 road segments in a road network spanning 7.68 km^2 with the average estimation accuracy as 0.80, which is calculated as the ratio between the number of accurate estimated traffic states and the total number of all estimations. Experimental result shows that the proposed method significantly outperforms the baseline and state-of-the-art methods.

The rest of this paper is organized as follows. Related works are reviewed in Section II. The traffic estimation problem and basic method are presented in Section III. The traffic correlation model and estimation method are described in Section IV. Trace-driven experiments are conducted to evaluate our design in Section V. Section VI finally concludes this paper.

II. Related Work

Traffic Sensing and Monitoring: There are tremendous efforts made to traffic sensing and monitoring of urban cities. With the deployment of sensing devices, e.g., close-circuit cameras or inductive loop detectors, on some well selected road segments, information about traffic volumes [2], [15], traffic congestions [1], [10] and traffic queues [19] of the monitored roads can be finally obtained. Due to the high deployment and maintenance costs of these sensing infrastructures, the idea of using roving vehicles to probe traffic conditions has been proposed as an efficient alternative. Crowdsourcing vehicles/participants help to realize the prediction of vehicle travel time [5], [28], and classification of road traffic states [25] through analyzing their collected data. None of above works, however, focus on acquiring the city-scale traffic with limited probes. They are primarily designed based on the assumption that sufficient amount of traffic data could be collected. In this paper, our traffic model exploits a sparse representation of global traffic conditions and assists to achieve a large scale traffic estimation with only sparse roving probes.

Traffic Estimation and Correlation: Idé *et al.* [11] estimate the traveling time of all road segments based on the traffic histories. They formulate it as a trajectory regression problem and propose a weight propagation mechanism to overcome the data sparsity issue. Both [24] and [27] improve the method in [11] by considering the traffic time-varying property. By exploiting the spatial-temporal correlation in the traffic matrix, SEER [29] and SVD-TE [13], [20], [30] recover the missing traffic elements using the multiple singular spectrum analysis (MSSA) technique and the singular value decomposition (SVD) technique, respectively. To our best knowledge, none of those existing works explicitly mine the traffic correlation among road segments under a general setting and enable the online traffic estimation for the whole road network.

There have been attempts made to explicitly study the traffic correlation in road networks. [9], [23] model the traffic correlation using Hidden Markov Model (HMM). However, to fit HMM, the derived models describe correlations only in quite coarse traffic levels, e.g., "high" and "low" two different states, which largely limits the model utility. In addition, most of them can only obtain the traveling time for those road segments with sufficient probe vehicles, which makes the solution not scalable. Their computational overhead is also high, e.g., several hours to derive a model for a road network with 1916 road segments, which challenges the model updating and system maintenances. Benefiting from the linear property of MLR, our model is computational lightweight, e.g., several minutes to complete the model construction for a road network with the same scale. Moreover, by incorporating with compressive sensing, our model enables a timely traffic estimation for the whole road network with sparse samplings.

III. Problem and Basic Method

Traffic Estimation Problem: We divide roads to road segments by intersecting points for better estimation granularity. The average traffic speed v_i within a time frame is adopted to assess the traffic condition of a road segment r_i [8]. The objective of traffic estimation is to recover the average traffic speeds $V = [v_1, v_2, \ldots, v_n]$ of n road segments in a road network based on the traffic samplings from probe vehicles. A typical traffic sampling contains a time stamp, current GPS position, instant speed and etc. [21]. Fig. 1 depicts an example with 3 probe vehicles in a road network of 60 road segments. Within a time frame, we collect the traffic samplings from the three probe vehicles, i.e., h_1 to h_3. According to the time stamps in the first and last reports from each vehicle within the time frame, we can get their traveling time $T = [t_1, t_2, t_3]^T$, where t_j indicates the traveling time of vehicle h_j. Besides, based on

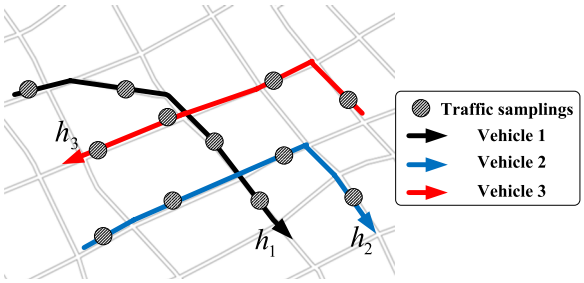


Fig. 1. An example of the traffic estimation problem with 3 probe vehicles in a road network of 60 road segments.

the GPS positions reported by each vehicle, we can recover their traveling trajectories on the road map. For each vehicle h_j, we can construct a vector $L_j = [l_j^1, l_j^2, \ldots, l_j^{60}]$. Each l_j^i in the vector indicates the distance h_j travels on road segment r_i. When h_j completely travels through road segment r_i, l_j^i equals to the length of this road segment. If h_j only travels a portion of road segment r_i, l_j^i is prorated based on the map matching technique. Based on each L_j, we can construct a 3×60 matrix $L = [L_1; L_2; L_3]$. In general, a statistical average traffic speed vector $V = [v_1, v_2, \ldots, v_{60}]^T$ of all road segments shall satisfy the following constraint:

$$T = \frac{L}{V} + e \tag{1}$$

where L/V means the element-wise division between L and V. T and L/V may not be exactly equal since V records the average traffic speeds of each road segment. An error vector e is hence introduced in Eq. (1). We can estimate V with the least-square method. Such an estimation approach degrades to mapping all probe vehicles' instant speeds to their running road segments if we set the time frame to the exact period interval of traffic samplings. The usual case, however, is that we could not directly calculate V as Eq. (1) is usually underdetermined, where the unknowns in V are much more than the measurements in T, i.e., $60 \gg 3$. Precisely solving such a problem requires at least as many linear independent measurements as the number of road segments. We can further partition each vehicle trajectory into multiple shorter pieces and consequently obtain more measurements (which will be further discussed in Section IV-B, and evaluated with experiments in Section V-C). Even that, the actual number of probe vehicles is yet far from adequate as the overall coverage is not sufficient and many of their observations are linear dependent. The goal of this paper is thus to reliably recover more unknowns from far fewer measurements when we have only sparse samplings. To tackle this issue, we mine and take advantages of the hidden correlations of traffic conditions in road network, and seek help from compressive sensing technique.

Traffic Correlation: Intuitively, the traffic conditions of nearby road segments are correlated. We refer to such relationship as *traffic correlation*. We first look at the traffic correlations from the one month traffic data collected from Shanghai city, China, and reveal the traffic correlations (which will be comprehensively explored in next section). For the ease of presentation, we examine the traffic correlations in the Jing'an

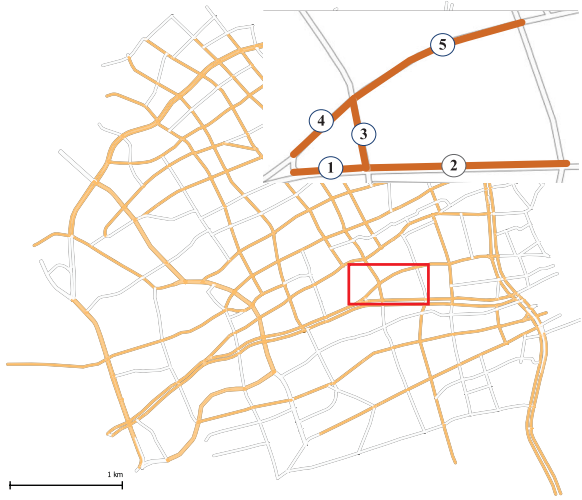


Fig. 2. Road map of the Jing'an area in Shanghai city, China. The traffic data set covers all primary and secondary roads highlighted in yellow. The zoom-in subfigure for the rectangle region contains 5 road segments.

area as depicted in Fig. 2, which occupies 7.68 km^2. We look at the well covered road segments (yellow roads in Fig. 2)[1] and filter out the minor road segments (gray roads in Fig. 2). To further ensure the accuracy, we divide the roads into 386 road segments according to the intersecting points of roads (more details in Section V-A). For investigation in this section, the average traffic speed v_i of each road segment r_i is treated as the average speed of all taxis running on that road segment for each time frame. We further introduce a traffic condition metric *congestion rate*. Congestion rate c_i of road segment r_i is defined as the reciprocal of its average traffic speed v_i in each time frame, and c_i is formally calculated as follows:

$$c_i = \begin{cases} \frac{1}{v_i}, & v_i \neq 0 \\ \infty, & v_i = 0. \end{cases}$$

Obviously, smaller congestion rates imply higher traffic speeds and thus better traffic conditions. In this paper, we use average traffic speed v_i and congestion rate c_i to describe the traffic condition of each road segment r_i alternatively. They can be easily calculated from each other.

We look at a concrete example of five road segments, r_1 to r_5, in the zoom-in area from Fig. 2. We plot the congestion rates (in 15 min time frames) of the five road segments for one week in Fig. 3. According to the map, r_1 and r_2 are cascaded with each other. From Fig. 3(a) and (b), we observe similar distributions of their congestion rates, which indicates clearly a direct traffic correlation between the cascaded road segments. Fig. 3(d) and (e) suggest a similar traffic correlation between the cascaded road segments r_4 and r_5. While road segments r_1 and r_4 are not physically connected, they are parallel with each other and lead to a same destination. From the congestion rate in Fig. 3(a) and (d), we observe that although the absolute values of their congestion rates are quite different, they experience a similar varying

[1]We refer to the road attributes in a digital map and find that yellow roads completely cover all primary and secondary roads in the Jing'an area.

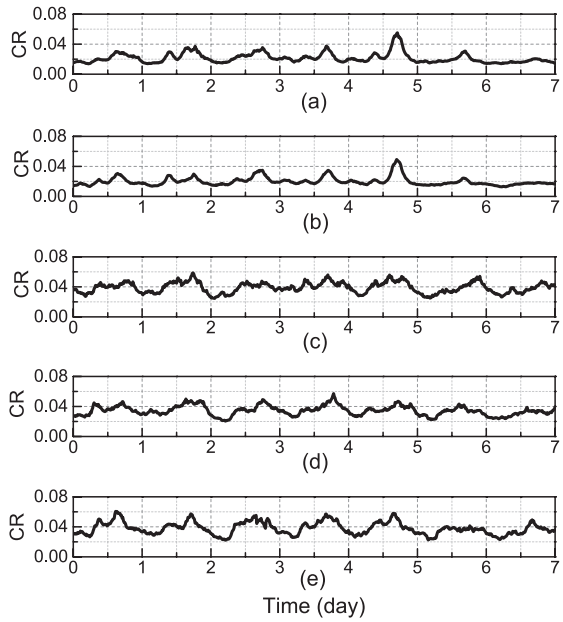


Fig. 3. Congestion rates of the 5 road segments from the subfigure of Fig. 2 in one week. Subfigures (a) to (e) correspond to road segment r_1 to r_5, respectively. Each congestion rate point is computed from the average value of reported speeds of all taxis passing by within a 15 min time frame.

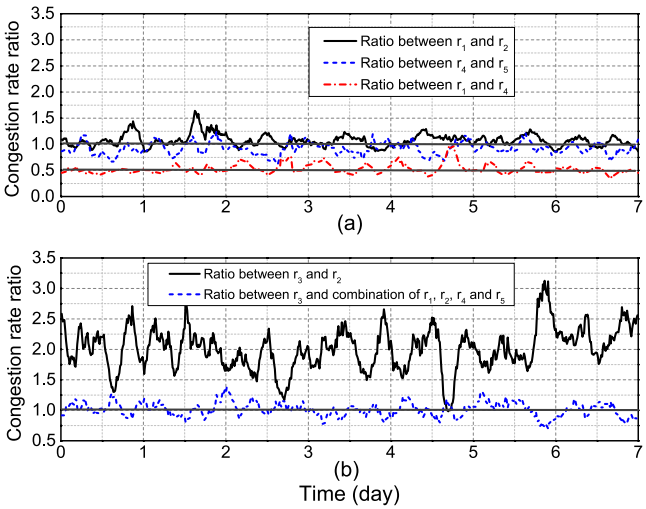


Fig. 4. Congestion rate ratios between different pairs of road segments. (a) ratio between r_1 and r_2, r_4 and r_5, r_1 and r_4 with variance 0.0135, 0.0136, and 0.0089, respectively; (b) ratio between r_3 and r_2, r_3 and a linear combination of r_1 to r_5 (except r_3) with variance 0.1193 and 0.0134, respectively.

pattern. To validate above statements, we plot the ratio of their congestion rates in each time frame in Fig. 4(a). We see that the congestion rate ratios between each pair of cascaded road segments are well represented by a straight line, implying that their traffic condition varyings tend to be synchronous. Since r_1 and r_4 lead to a same destination, their traffic burdens are alternatively shifted and balanced with each other. They thus demonstrate a direct traffic correlation, as shown in Fig. 4(a).

Road segment r_3 intersects with the above four road segments. From its congestion rate data, it is not easy to observe any correlation with other road segments. In Fig. 4(b), we plot

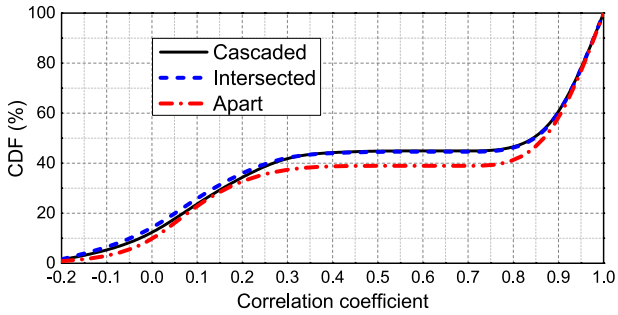


Fig. 5. CDF of pair-wise road segment correlations in the Jing'an area.

the congestion rate ratios between r_3 and r_2. The figure shows that the ratio indeed diverges as the time elapses, with the variance up to 0.1193. The congestion rate ratios between r_3 and other three road segments are also tested and no clear correlations are found. Our further investigation, however, discovers that though r_3 is not directly related to any particular intersected road segment, it does correlate to them in a more sophisticated manner. In Fig. 4(b), we plot the congestion rate ratios between r_3 and a linear combination of the rest four road segments, i.e., $c_3 = 0.0125 + 0.3421c_1 - 0.1773c_2 + 0.3053c_4 + 0.3206c_5$, which shows a much stronger traffic correlation. The ratio is stabilized around the straight line 1.0 and the variance reduces to 0.0134. This can be explained by the fact that r_3 bridges the rest four road segments and the traffic amount on r_3 is related to all but not any particular one of them.

Through statistical analysis, we plot the pair-wise traffic correlation distribution of all 386 road segments in Fig. 5. The correlation is measured by the *Pearson correlation coefficient*. According to the geographic relationship between each pair of road segments, we classify the results into three categories: cascaded, intersected, and apart. We define road segments connected with the same road name as cascaded, while connected with different road names as intersected. From the figure, we see that around 55% cascaded and intersected road segments appear strong correlations (e.g., ≥ 0.8). Those correlations mainly reflect the direct traffic correlations between two road segments. The remaining 45% may contain rich indirect correlation opportunities, like the case of r_3 in Fig. 4(b). The correlations of apart road segments (including the "parallel" cases) have a similar distribution with even more direct correlations (more than 60%). As those road segments are not directly connected, such correlations are possibly due to the existence of critical road segments, whose traffics have broader impacts on other road segments. In addition, the rest 40% may still contain plenty of indirect correlation opportunities. Therefore, to mine the traffic correlations for the city-scale traffic estimation, both direct and indirect correlations should be explored and a generic traffic correlation model is desired to capture such relations which we will present in next section.

Compressive Sensing Based Solution: The observed traffic correlation makes it possible to find a solution to the underdetermined system in Eq. (1) due to recent advances in the compressive sensing technique. The compressive sensing theory states that a sparse signal X of size n, i.e., $\|X\|_{l_0} \ll n$, can be reconstructed from m projection measurements, where $m \ll n$

[6]. The measurements are performed using a linear transform Φ on signal X, i.e., $Y = \Phi X$. In our traffic estimation problem, T is the measurement vector Y, and L is the measurement matrix Φ. The signal $1/V = \mathbf{c} = [c_1, c_2, \ldots, c_{60}]$, however, is usually not sparse, and we should find an alternative domain Ψ in which $\mathbf{c}$ can be sparsely represented. It is usually desired to construct such a domain by considering the characteristics in signal X [16]. For our case, traffic correlation in $\mathbf{c}$ is a valuable trait. In domain Ψ, one expects to achieve a k-sparse representation $\mathbf{s}$ for $\mathbf{c}$, i.e., $\mathbf{c} = \Psi\mathbf{s}$, where vector $\mathbf{s}$ contains k nonzero coefficients. As a result, the measurement vector Y can be rewritten as $Y = \Phi\Psi\mathbf{s}$. The compressive sensing theory shows that when $k \ll n$ and measurement matrix Φ and representation basis Ψ satisfy the *Restricted Isometry Property* (RIP) [3], the k-sparse $\mathbf{s}$ can be precisely recovered with high probability by solving the following l_1-minimization problem:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}\in R^N} \|\mathbf{s}\|_{l_1} \quad \text{s.t.} \quad Y = \Phi\Psi\mathbf{s}$$

when $m \geq a\mu^2(\Phi, \Psi)k \log n$, where a is a positive constant, and $\mu^2(\Phi, \Psi)$ is the coherence between Φ and Ψ. To further consider the measurements in Y which are usually polluted with error e, i.e., $Y = \Phi\Psi\mathbf{s} + e$ like Eq. (1), the equation above can be rewritten as:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}\in R^N} \|\mathbf{s}\|_{l_1} \quad \text{s.t.} \quad \|Y - \Phi\Psi\mathbf{s}\|_{l_2}^2 \leq \epsilon \tag{2}$$

where ϵ is the error tolerance [7]. Eq. (2) can be efficiently solved by using a standard compressive sensing solver, and thus the global traffic conditions can be recovered as $\mathbf{c} = \Psi\hat{\mathbf{s}}$.

IV. Traffic Estimation Method

In this section, we will introduce the two components of our traffic estimation method, i.e., the offline traffic correlation mining and the online traffic estimation.

A. Traffic Correlation Mining

Traffic correlations among road segments are mined based on the history traffic data of the road network. Mining can be performed offline and infrequently to update traffic correlations with relatively long time intervals (e.g., several weeks or months). In the following, we use congestion rates to build the traffic correlation model.

MLR Model: From the practical example in previous section, it appears that the traffic condition of one road segment can be linearly approximated by the traffic conditions of its nearby road segments. We extend such an observation and use the Multiple Linear Regression (MLR) model to capture the hidden correlations among different road segments (even not physically connected) at a global scale. For each road segment r_i, we build the traffic correlation model for its congestion rate c_{r_i} with respect to the traffic conditions of all other road segments in the road network as:

$$c_{r_i} = \beta_{r_i,0} + \sum_{j=1, j\neq i}^{n} \beta_{r_i,r_j} \times c_{r_j} = \mathbf{c}_{\mathbf{r_i}}^T \cdot \boldsymbol{\beta}_{\mathbf{r_i}} \tag{3}$$

where $\mathbf{c}_{\mathbf{r_i}}$ and $\boldsymbol{\beta}_{\mathbf{r_i}}$ are two $n \times 1$ vectors, and n is the total number of road segments in the road network. The vector $\mathbf{c}_{\mathbf{r_i}} = [1, c_{r_1}, \ldots, c_{r_{i-1}}, c_{r_{i+1}}, \ldots, c_{r_n}]^T$ represents the congestion rates of the rest $n-1$ road segments except r_i. The element 1 is added to represent the constant item $\beta_{r_i,0}$ in the MLR model. On the other hand, the vector $\boldsymbol{\beta}_{\mathbf{r_i}} = [\beta_{r_i,0}, \beta_{r_i,r_1}, \ldots, \beta_{r_i,r_{i-1}}, \beta_{r_i,r_{i+1}}, \ldots, \beta_{r_i,n}]^T$ records the corresponding coefficients of each road segment. With sufficient training data, we can estimate the regression coefficient vector $\boldsymbol{\beta}_{\mathbf{r_i}}$ using the least-square method, which minimizes:

$$\sum_{q=1}^{W} \left(c_{r_i}^q - \left(\beta_{r_i,0} + \sum_{j=1, j\neq i}^{n} \beta_{r_i,r_j} \cdot c_{r_j}^q \right) \right)^2$$

where W is the total number of time frames in training data set and $c_{r_j}^q$ is the congestion rate of road segment r_j in the q-th time frame.

A representation matrix can thus be constructed and used to describe the vector $\mathbf{c} = [1, c_{r_1}, c_{r_2}, \ldots, c_{r_n}]^T$, with $(n+1)$ congestion rates, as follows in Eq. (4). The element 1 added in $\mathbf{c}$ takes over the constant item in the MLR model.

$$P = \begin{bmatrix} \gamma & 0 & 0 & \cdots & 0 \\ \beta_{r_1,0} & -1 & \beta_{r_1,r_2} & \cdots & \beta_{r_1,r_n} \\ \beta_{r_2,0} & \beta_{r_2,r_1} & -1 & \cdots & \beta_{r_2,r_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{r_n,0} & \beta_{r_n,r_1} & \beta_{r_n,r_2} & \cdots & -1 \end{bmatrix} \tag{4}$$

where the i-th row corresponds to the MLR model of road segment r_i if we move c_{r_i} to the right-hand side in Eq. (3). The coefficients added in the first row is for the constant item in the MLR model and the first element γ is set to be $0 < \gamma < 1$, ensuring that P is invertible. In principle, if our MLR model indeed captures the traffic correlations of all road segments, one would see the projection of $\mathbf{c}$ on P, $\mathbf{s} = P\mathbf{c}$, to be a vector containing many zero/near-zero entries. In other words, $\mathbf{c}$ is transformed to a sparse representation, and the non-zero/significant entries in $\mathbf{s}$ capture the major traits of current traffic conditions of the whole road network. If this is the case, P^{-1} serves as the representation basis, i.e., $\Psi = P^{-1}$, which maps $\mathbf{s}$ back to $\mathbf{c}$ and can be used in compressive sensing technique. We will evaluate the representation capability of P^{-1} and the computation complexity of the MLR model later.

Different Traffic Scenarios: According to recent measurement studies, urban traffics are not consistent across time and demonstrate different patterns [2], [24], [27]. The underlying traffic correlation may be similar but not identical due to the varying traffics, e.g., in Fig. 3, we thus consider such patterns in our MLR modeling to better capture the traffic correlation, and derive a set of different models for different scenarios. We classify traffic scenarios based on two aspects: different times during a day and different days in a week. We first distinguish traffics of workdays from non-workdays considering the commuter traffics. We further classify each day (including both workdays and non-workdays) into two periods: period 1

(21:00 pm to 07:00 am and 13:00 pm to 16:00 pm: non-peak hours) and period 2 (07:00 am to 13:00 pm and 16:00 pm to 21:00 pm: peak hours). In summary, we consider varied traffic scenarios of four types. Note that the traffic scenario classification is not necessarily fixed to be four to train our models, which can be adjusted in different cities. The four periods used here fit the citizen's daily routines and the traffic patterns in Shanghai city according to our traffic data set.

To train traffic correlation models using actual traffic data set, we first prepare four training data groups according to the time and days. Specifically, we divide the entire time duration covered by the traffic data set into a series of time frames, and map each time frame into one of the four groups, i.e., period 1 (and 2) of workdays (and non-workdays). For each time frame, we compute an average congestion rate based on the associated speeds for each individual road segment. Thus, we have four training data groups with respect to four traffic scenarios at each road segment. Finally, for each data group, we iteratively train the traffic model using MLR in Eq. (3) and derive all regression coefficients β_{r_i,r_j} for the matrix in Eq. (4). We thus obtain four representation matrices P.

Coefficient Pruning: For each road segment r_i, we do not exactly know which set of road segments it actually correlates with before model training. Its congestion rate c_{r_i} in Eq. (3), expressed using traffic conditions of all other $n-1$ road segments, can include all possible combinations. In reality, one road segment tends to be more correlated with only a certain number but not all of the road segments, e.g., nearby ones. Correlating all road segments to each particular road segment, the overfitting issue may occur and huge computation overhead is unnecessarily triggered. Therefore, for any road segment r_i, the traffic correlation would be more reasonably represented by its top κ correlated segments rather than all other $n-1$ road segments. We can simplify the model in Eq. (3) by learning $\kappa+1$ coefficients merely. To effectively select the top κ correlated road segments for r_i, we consider both the traffic condition similarity and the geographic distance between r_i and any other one road segment r_j. We introduce a selection factor $f_{r_i,r_j} = (d_{r_i,r_j}/\rho_{r_i,r_j})$ to differentiate other $n-1$ road segments. d_{r_i,r_j} is the geographic distance between the center points of two road segments r_i and r_j, and ρ_{r_i,r_j} is the *Pearson correlation coefficient* of their traffic conditions. A road segment r_j with a smaller f_{r_i,r_j} is possibly more correlated with r_i. After selecting the top κ correlated segments, we prune the number of unknown coefficients from $n+1$ to $\kappa+1$. For the rest, we simply set their coefficients to 0. In our current implementation, we use the same κ for all road segments in the same traffic scenario, i.e., $\kappa+1$ non-zero coefficients in each row of the representation matrix P in Eq. (4). A customized selection of κ for each individual road segment (each row in P) will further improve the accuracy, which we do not include in this paper due to page limit.

To obtain the appropriate κ values, we use the accuracy of the derived traffic model to quantify the quality of each κ selection. We demonstrate the κ selections for the Jing'an area based on four weeks traffic data, and the selection process can be easily extended to other areas. The data set of the first three weeks are used for the model training, and the data of the last week

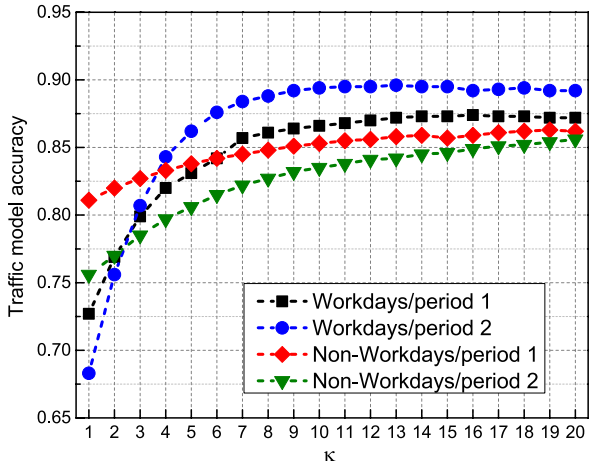


Fig. 6. The accuracy of traffic models for the four traffic scenarios when κ varies from 1 to 20.

is used to evaluate the accuracy of the models. The accuracy of the traffic models can be evaluated by:

$$1 - \frac{\sum_{i=1}^{n}\sum_{q=1}^{W}|v_{r_i,q} - \tilde{v}_{r_i,q}|}{\sum_{i=1}^{n}\sum_{q=1}^{W} v_{r_i,q}}$$

where $v_{r_i,q}$ is directly calculated from the traffic samplings, $\tilde{v}_{r_i,q}$ is obtained by the estimated congestion rate $c_{r_i,q}$ from the MLR model, and W is the total number of time frames. A higher accuracy value indicates a better κ selection.

Fig. 6 depicts the traffic model accuracy for each traffic scenario when κ varies from 1 to 20. From the figure, we see that when κ is small, the accuracy is low. This is because for each road segment, not all their correlated road segments have been included in the model when κ is small. As κ increases, the traffic model accuracy increases rapidly, and finally tends to be stable. It implies that the contribution from including additional road segments in the MLR model is negligible after κ becomes sufficiently large. Thus, we select a relatively small κ value to reduce the computation overhead, while still preserve acceptable accuracy. We set κ to 10 for all the four traffic scenarios, as the traffic model accuracy is insensitive to the κ value beyond that.

Validation of the Traffic Sparsity: We evaluate representation capability of obtained representation matrices P, i.e., to evaluate the sparsity of $\mathbf{s}$ that meets $\mathbf{s} = P\mathbf{c} = \Psi^{-1}\mathbf{c}$. We normalize vector $\mathbf{s}$ as vector $\mathbf{s}' = [(|\boldsymbol{o_i}|/\|\mathbf{s}\|_{l_1}), o_i \in \mathbf{s}]^T$. The sparsity of $\mathbf{s}$ is the number of elements in $\mathbf{s}'$ with values greater than a small threshold (elements less than the threshold are viewed as negligible entries [22]).

We test the traffic sparsity with the 4th-week traffic data. For each time frame, we can obtain a congestion rate vector $\mathbf{c}$. According to which traffic scenario the time frame belongs to, we calculate $\mathbf{s} = P\mathbf{c}$ with the corresponding representation matrix, and examine the sparsity of $\mathbf{s}$. Different MLR models of road segments in different traffic scenarios lead to different P in Eq. (4), and thus different Ψ. We find that $\mathbf{s}$ is of good sparsity in all four traffic scenarios as shown in Fig. 7. With a threshold 0.01, the sparsity in any traffic scenario is less than 25 that is much smaller than the total number of road segments

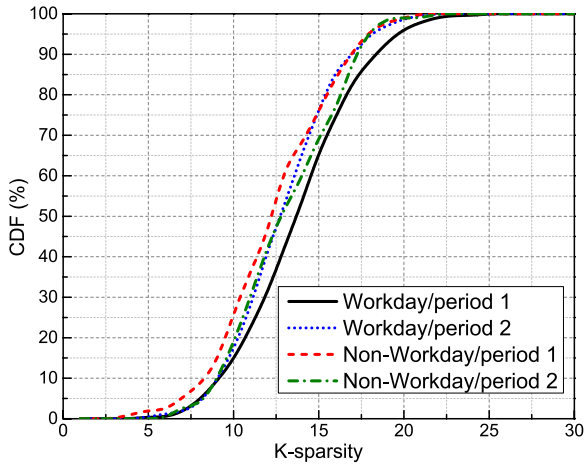


Fig. 7. The sparsity of representation $\mathbf{s}$ in domain Ψ for the four traffic scenarios.

386. As a result, the representation matrix P, constructed by our traffic models, can sparsely describe the traffic conditions. We thus obtain a set of suitable representation bases $\Psi = P^{-1}$ for all the four traffic scenarios, which can be used in compressive sensing technique to recover $\mathbf{c}$.

Computation Complexity: If a road network consists of n road segments and we correlate the traffic condition of each road segment with all other $n-1$ road segments, the computation complexity to mine traffic correlations using MLR is $\mathcal{O}(n^3W)$, where W is the total number of training examples. With the coefficient pruning, however, the computation complexity can be significantly reduced to $\mathcal{O}(\kappa^2 nW)$, where κ is a small constant defined in the "Coefficient pruning" subsection and $\kappa \ll n$. We will further investigate the MLR computation complexity experimentally in Section V-E, which shows that it takes only several minutes to establish our MLR model for a large road network with thousands of road segments.

B. Traffic Estimation via Compressive Sensing

We have constructed the representation bases Ψ for the traffic estimation problem $Y = \Phi\Psi\mathbf{s} + \boldsymbol{e}$. In this subsection, we will introduce how to obtain the measurement vector Y and the measurement matrix Φ using a small number of probe vehicles, and then obtain the timely traffic condition estimation of all road segments via a compressive sensing solver.

Measurement Vector Y: A server continuously collects traffic samplings from a fleet of probe vehicles in each time frame and estimates the global traffic conditions at the end of the time frame. In other words, the size of time frame is the time granularity of traffic estimation. A smaller time frame leads to more timely updating of traffic states. In any time frame, suppose the server collects samplings from m probe vehicles, h_1 to h_m. According to the time stamps contained in the traffic samplings, we get their traveling times, and for each probe vehicle h_j, we denote t_{h_j} as its traveling time. $Y = [t_{h_1}, t_{h_2}, \ldots, t_{h_m}]^T$ is thus the measurement vector in the compressive sensing formulation.

Measurement Matrix Φ: Based on the GPS positions reported by each probe vehicle, we can calculate their traveling trajectories on the road map. To overcome the noise of GPS data, we adopt an Hidden Markov Model (HMM) based map matching algorithm [17] to match a sequence of GPS positions sparsely sampled by one probe vehicle to the most likely sequence of road segments. By viewing actually travelled road segments as hidden states and traffic samplings as observations, [17] has shown that the HMM based map matching method achieves high mapping accuracy and efficiency. Thus, we use the output of HMM based map matching method as the final vehicle trajectory. Specifically, for probe vehicle h_j, we can construct a vector $L_{h_j} = \{l_{h_j}^{r_i}, i = 1, 3, \ldots, n\}$ according to its trajectory. Each $l_{h_j}^{r_i}$ in L_{h_j} indicates the distance h_j travels on the road segment r_i, which can be calculated as:

$$l_{h_j}^{r_i} = \begin{cases} d_{h_j}^{r_i}, & r_i \text{ is passed by } h_j \\ 0, & \text{otherwise} \end{cases}$$

where $d_{h_j}^{r_i}$ is the actual traveled distance by probe vehicle h_j on road segment r_i. If r_i is fully covered by the trajectory of h_j, then $d_{h_j}^{r_i}$ is the length of r_i; otherwise, the traveled distance is computed as the *great circle distance* via an map matching technique. Based on each L_{h_j}, we can construct an $m \times n$ matrix as $[L_{h_1}; L_{h_2}; \ldots; L_{h_m}]^T$. Considering the constant item $\beta_{r_i,0}$ in MLR model, we add a *zero* value at the first position of each distance vector, and obtain the final $m \times (n+1)$ matrix as Eq. (5) that is the measurement matrix Φ for the compressive sensing formulation. Since each probe vehicle travels freely in the city, the trajectories of any two vehicles are independent [2]. As a result, the measurement matrix Φ constructed in Eq. (5) is a random matrix, which satisfies the requirement by compressive sensing on Φ, whose elements should be randomly chosen.

$$\Phi = \begin{bmatrix} 0 & l_{h_1}^{r_1} & l_{h_1}^{r_2} & \cdots & l_{h_1}^{r_n} \\ 0 & l_{h_2}^{r_1} & l_{h_2}^{r_2} & \cdots & l_{h_2}^{r_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & l_{h_m}^{r_1} & l_{h_m}^{r_2} & \cdots & l_{h_m}^{r_n} \end{bmatrix}. \quad (5)$$

The compressive sensing theory requires the measurement matrix Φ to be incoherent with the representation basis Ψ for accurate recovery results. In our formulation, the two matrices Φ and Ψ are constructed independently and vary in each time frame. As a result, it is highly possible that they are incoherent. We adopt the metric of *dual-incoherence* [22] to examine whether Φ and Ψ are incoherent for most of the time. Dual-incoherence measures the correlation between Φ and Ψ indirectly. The larger dual-incoherence, the more incoherent the two matrices are. Specifically, we compute the *dual-incoherence* between Φ and Ψ for all the time frames in the 4th-week traffic data, and plot the statistical results for all the four traffic scenarios in Fig. 8(a). The statistical results suggest that the varied pairs of Φ and Ψ are quite incoherent. In more than 90% cases, the dual-coherence of Φ and Ψ is greater than 380, which approaches the maximum possible dual-incoherence between any pair of Φ and Ψ, i.e., $(n+1) = 387$. As such an incoherent property is a sufficient but not a necessary condition in compressive sensing theory, the l_1-minimization method can still recover the sparse representation $\mathbf{s}$ in most of the time, even if Φ and Ψ are not fully incoherent [16]. In our experiments, we

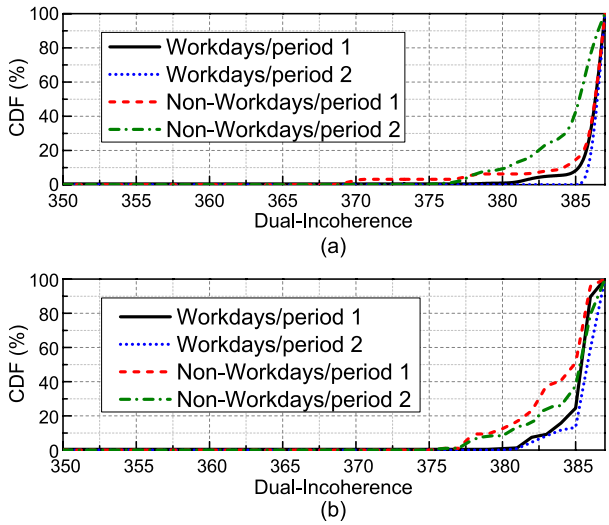


Fig. 8. Dual-incoherence between Φ and Ψ with (a) original trajectories; (b) partitioned trajectories.

observe those two matrices always preserve a good incoherent property.

As earlier mentioned in Section III, during the estimation stage, we can partition each vehicle trajectory into multiple shorter pieces to increase the number of independent measurements. After trajectory partitioning, we obtain more traveling time measurements in Y, and accordingly increase the rows in the measurement matrix Φ by splitting each original row in Eq. (5) into multiple new rows according to the partitioned trajectories. With the default setting in Section V, we validate the dual-incoherence between the expanded Φ and Ψ for all the time frames in the 4th-week traffic data, and plot the results in Fig. 8(b), where we find the trajectory partitioning operation does not affect the incoherence between Φ and Ψ, and they are still always incoherent. The natural benefit of trajectory partitioning is that more rows of measurements and constraints are obtained in Y and Φ, thus more traffic details for estimation accuracy. However, the length of partitioned trajectories is generally short. Map matching errors and instant velocity jitters will become obvious and severe. In Section V-C, we do detailed experiments and find that the benefit outweighs the harm in achieving higher accuracy when we partition the original trajectories into more individual pieces.

Compressive Sensing Based Recovery: We have obtained all the required elements, i.e., Y, Φ, and Ψ, in the compressive sensing formulation. We have also verified the sparse representation capability of $\Psi = P^{-1}$, and the incoherence between Φ and Ψ. Thus, by solving the corresponding l_1-minimization problem of Eq. (2) using some standard compressive sensing solvers, e.g., linear programming [4], we can recover $\mathbf{s}$. Then we calculate the congestion rate vector $\mathbf{c} = \Psi\mathbf{s}$, and further obtain the average traffic speeds of all road segments from the reciprocal of congestion rates.

C. Discussion

Although the experiments and analysis focus on the Jing'an area, the principles can be directly extended to other areas or cities (e.g., Beijing or New York), with the same parameter tuning and system execution procedures. To control the computation overhead and traffic estimation accuracy, a large city could be partitioned into regions and our method is then applied to each region. The region can be aligned with the jurisdictional area of each local traffic management bureau. As the connection between two regions is usually based on cascaded roads, the inaccuracy of the traffic correlation models for roads along the region boundary is minimal.

One concern about the application of our approach to other cities is the size of training data, i.e., traveling speed. To learn an MLR model for each road segment in one traffic scenario, the training data size should be $\geq (\kappa+1)^2$ if we correlate one road segment with κ road segments [12]. Therefore, the minimum training data size would be $W = pn(\kappa+1)$ for a road network of n road segments with p traffic scenarios. We prefer more training data to weaken the influences of noises and variances of traffic data. According to our analysis and experiments in the Shanghai city case, we expect the training data size would be 10 times of the basis W. It is worthy to note that even if we have not that sufficient training data in other cities, we can still train the MLR models for the initial use. Once we have accumulated enough traffic data during live usages, we can update MLR models with more training data.

In summary, our approach captures the hidden traffic correlation using a simple yet effective MLR model, which is computationally efficient and enables traffic estimation of the whole road network even with a small number of probe vehicles. By exploiting and analyzing the big traffic data, our approach demonstrates the feasibility of data-driven intelligent transportation system [26]. To further improve the traffic estimation accuracy, we can enrich the proposed approach by encoding some domain knowledge, e.g., network behavior.

V. Experimental Evaluation

We perform extensive trace-driven experiments to analyze traffic correlations among road segments in a road network and examine the performance of our traffic estimation method.

A. Traffic Data Set

The traffic data set is collected from a fleet of more than 4400 taxis running in Shanghai city, China. One month traffic data set of all taxis is available for us. Each taxi reports traffic sampling every minute through the cellular network. Equipped with GPS devices in taxis, each reported sampling includes a time stamp, latitude, longitude, instant speed, and other information. The pair of GPS latitude and longitude represents the position where the taxi reports.

Road Network: We first use the Jing'an area, a downtown region of Shanghai city, China, as an example to evaluate the performance of our traffic correlation model and traffic estimation method. We further extend our solution to a much larger area about 50.70 km^2, covering the major central area of Shanghai city, to investigate its applicability and scalability. We exploit the OpenStreetMap (OSM)[2] map to export all the roads

[2]http://www.openstreetmap.org/

in those areas, and segment the roads according to their intersecting points. In OSM map, two driving directions of the roads are represented separately. The trajectory of each vehicle could be mapped to the corresponding directional road segments. We classify the geographic structure of all road segments into three groups: "cascaded," "intersected," and "apart." "Cascaded" includes road segments connected with the same road name, while "intersected" contains connected road segments with different road names. OSM contains 926 road segments totally in the Jing'an area and we select a subset of 386 road segments with sufficient coverage by the taxi data, which form a road network for our experiments. We refer to road attributes and find that the selected 386 roads completely cover all primary and secondary roads in this area. Similarly, we select 1826 out of totally 3451 road segments from the central area with sufficient taxi data coverage. Note that road segment selection aims to obtain the ground truth for the performance evaluation only. In practice, after the traffic models are constructed, our method can estimate the traffic conditions for the whole road network. Throughout the one month traffic data, the number of probe taxis within the Jing'an area (central area) ranges from 65 to 498 (175 to 906) in each time frame. As probe vehicles travel along road segments freely in a road network, we thus use the traffic samplings from randomly selected taxis for the traffic estimation to best match the real scenario. By default, we use the traffic samplings from 50 randomly taxis for the traffic estimation in each time frame. Probe taxis generate adequate traffic data for both model training and performance testing, e.g., we have about 5 millions traffic samplings for the Jing'an area and about 14 millions traffic samplings for Shanghai central area.

Traffic Data Preprocessing: We divide the one month traffic data for each 15 min time frame and classify into four traffic scenarios as described in Section IV-A. Time frames with size of 15 min are long enough to accumulate sufficient traffic samplings for accurate map matching for each probe taxi, and also provide a fine granularity for timely traffic estimation [13], [23], [30]. By default, we use the data of all taxis in the first three weeks to train the MLR models and the data of a small subset of those taxis in the last week to evaluate the performance of our approach. In each time frame, we calculate the average traffic speed of each road segment as the average value of all reported speeds by all taxis passing by. In case there is no reported speeds for a road segment in one time frame, we use the average speed of this road segment in the previous 4 time frames. After the traffic data preprocessing, every road segment has a speed in each time frame, and thus the congestion rate. Those average speeds are treated as the *pseudo ground truth*, and used to evaluate the performance of our traffic estimation method.

B. Traffic Correlation Analysis

To train MLR models for each road segment r_i, r_i selects its top κ correlated segments from the rest $n-1$ road segments. κ is set to 10 according to the prior empirical investigation in Fig. 6. We first study the geography distribution of the top correlated road segments derived from our traffic models. The physical distance distribution of pair-wise correlated road segments is plotted in Fig. 9(a). We find that 90% correlated road segments locate within the range of 0.5 km, where the distance between two road segments is measured from their midpoints. For each road segment, correlated segments are mainly from its vicinity giving good locality in traffic correlation. On the other hand, we further examine the *logic* relationship between each road segment and its κ correlated road segments. We observe 774 cascaded pairs (20%), 1711 intersected pairs (44%), and 1375 apart pairs (36%). As "cascaded" and "intersected" account for 64% in total, it implies that around 26% of "apart" pairs also preserve the correlation locality (e.g., the "parallel" case). In addition, remote segments (>0.5 km) contribute non-negligible portions as well, e.g., 10%. Above correlations are hidden and largely omitted in previous works, e.g., [13], [23], [29], [30]. They cannot be unveiled unless with an explicitly derived traffic model.

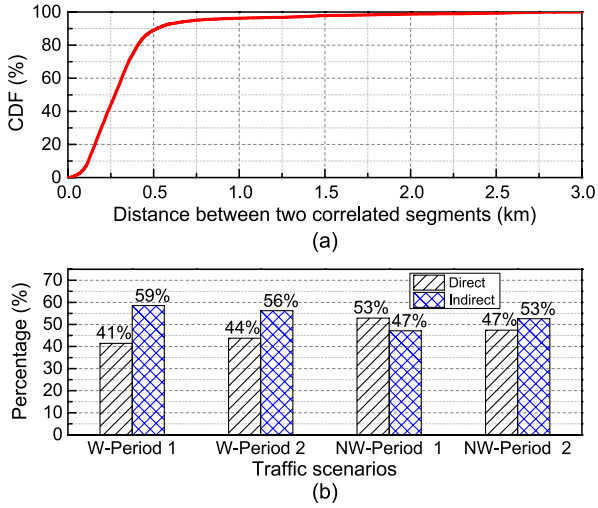


Fig. 9. (a) CDF of distance between top correlated road segments. (b) Statistical results of correlation types in the four traffic scenarios.

We then investigate how each road segment r_i is related to other road segments by examining the normalized regression coefficients in the MLR model. If there is a dominant coefficient, r_i is viewed as directly correlated with that particular road segment. Otherwise, r_i is indirectly correlated with a set of road segments. We empirically select 0.35 as the threshold to find the dominant coefficient for road segment r_i. We believe 0.35 is great enough to filter the most predominant road segment from the $\kappa = 10$ correlated road segments of r_i. Fig. 9(b) depicts the distributions of correlation types in all the four traffic scenarios. Although the percentage of direct correlation is high, in around half of the cases, road segments are indirectly correlated with other segments. The results imply that plenty of implicit correlations are hidden among different road segments. Such correlations are not directly visible unless explored by a method such as the proposed traffic model.

C. Basic Evaluation

Trajectory Separation: We conduct experiments to explore the impacts of trajectory partitioning, and show its effectiveness.

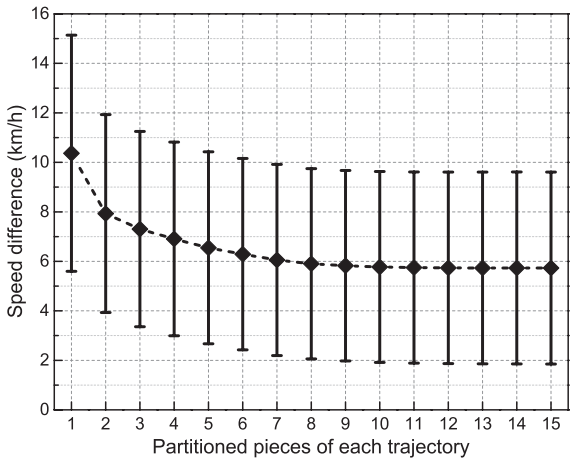


Fig. 10. Average absolute speed differences between pseudo ground truths and estimated speeds with various partitioned pieces of each trajectory.

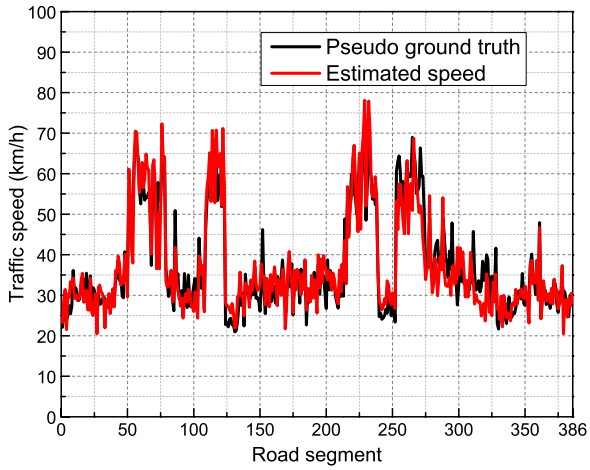


Fig. 11. The traffic estimation snapshot of a randomly selected time frame, with comparison of the pseudo ground truths.

A trajectory generally can be partitioned into multiple pieces according to the amount of traffic samplings. For the original trajectories of randomly selected probe taxis, we uniformly partition each of them into x pieces and perform traffic estimation based on the partitioned trajectories. Fig. 10 plots the average absolute speed differences between the pseudo ground truths and estimated speeds when we partition each original trajectory into x pieces of short trajectories (if possible), where $x = 1$ means we perform the traffic estimation with the original trajectories. From Fig. 10, we find that when each original trajectory is partitioned into more pieces, the average absolute speed difference tends to be smaller. These results imply that benefits outweighs the harm in achieving higher traffic estimation accuracy. According to the experiment, we obtain better estimation results when we partition each original trajectory into maximum possible number (15 pieces in each 15 min time frame) of short trajectories. We see negligible performance gain beyond $x = 10$ pieces of partition, and we use this as the default setting for all following experiments.

Traffic Estimation Accuracy: We first show a snapshot of estimated traffic speeds of all road segments in one randomly selected time frame in Fig. 11. This figure shows that the estimated speeds cross different road segments match the trend of the pseudo ground truths very well. To quantify the difference between the pseudo ground truths and the estimated speeds, we plot the CDF of absolute speed differences between them over the four traffic scenarios across the entire 4th-week traffic data in Fig. 12(a). Due to the inherently higher traffic dynamics in workdays, e.g., the high taxi velocity fluctuation and traffic variance due to traffic jams in workdays, the performance in non-workdays is generally better. The 90-percentile and 60-percentile speed differences are 10.5 km/h and 5.2 km/h for non-workdays and 11.0 km/h and 5.6 km/h for workdays respectively. In general, the estimated speed for each road segment is in good accuracy but always slightly smaller than its pseudo ground truth with overall difference less than 5.0 km/h. It is because in the measurement vector Y, the time difference between the last and the first samplings within a time frame may contain certain time that should not be included, e.g., the waiting

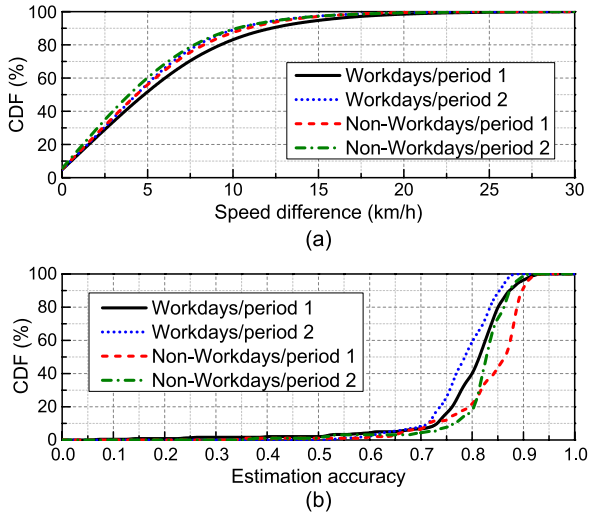


Fig. 12. CDF of (a) the absolute speed differences; (b) the estimation accuracy with respect to four different traffic condition indicators.

time for traffic lights. The measured traveling time is thus greater than the actual one and thus a smaller estimated speed.

Instead of directly giving the speed estimations, we translate the estimated speed to a more meaningful traffic indicator for each road. It is worthy to note that different from previous methods only providing coarse traffic levels [9], [23], our method can provide the detailed traveling speed information for each road segment. Similar with Google Map, we classify the traffic conditions of each road segment r_i to four categories according to its traffic speed v_i (in km/h), i.e., *Congested* ($v_i < 20$), *Slow* ($20 \leq v_i < 40$), *Normal* ($40 \leq v_i < 60$) and *Fast* ($v_i \geq 60$). We then compute the estimation accuracy as (# of estimation hits/# of total time frames), where an estimation hit means both the pseudo ground truth and the estimated speed are classified to a same category in one time frame. From the results in Fig. 12(b), We find that the four traffic scenarios have similar estimation accuracy distribution, which concentrates within a high accuracy range between 0.75 and 0.90. The accuracy achieves up to 0.94 and the overall average

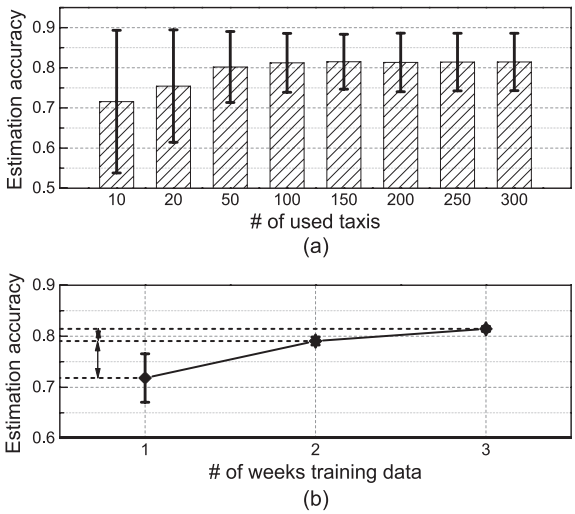


Fig. 13. (a) Estimation accuracy with different numbers of probe taxis used; (b) Estimation accuracy with different number of weeks training data.

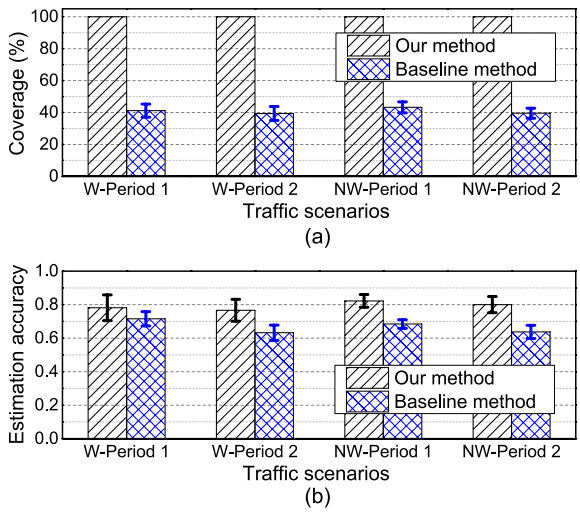


Fig. 14. (a) Road network coverage by different methods; (b) Estimation accuracy of road segments directly covered by taxis with different methods.

is more than 0.80, which provides quite accurate estimation results.

Number of Probe Vehicles: In this experiment, we investigate the impact of the probe taxi number on the estimation performance of our method. We vary the taxi number used in the traffic estimation (denoted as m) from 10 to 300, and plot the estimation accuracy in Fig. 13(a). Since probe vehicles in practice travel independently in a road network, m taxis are selected randomly in each time frame to best match the reality. If the total amount of taxis in the road network is smaller than m (e.g., in the early morning), we use all available taxis for the traffic estimation. From the figure, we find when m is very small (e.g., 10 or 20), the estimation already achieves high accuracy on average (> 0.70), but with a large variance. After the number of taxis used is greater than 50, the average estimation accuracy stabilizes around 0.80 while the standard deviation continuously decreases as more taxis are used in the estimation. Based on the statistics, we find that the accuracy improvement and standard deviation reduction are only 1.54% and 19.16% respectively, when the taxis number increases from 50 to 300. With our method, a small group of probe taxis can offer comparable accuracy as that from a large number of probe taxis. Hence, our default setting $m = 50$ leads to both good estimation performance and low system operation cost.

Training Data Size: We vary the training data size from one week to three weeks and use a small subset of data of the subsequent week from 50 probes as the testing data to investigate the impacts of training data size. Fig. 13(b) plots the estimation accuracy, where we find more training data can lead to higher accuracy, which is attributed to the better trained MLR models. On the other hand, we also find that even more data may bring slight improvement. For example, when we increase the training data size from two weeks to three weeks, the accuracy improvement is only 0.02. It is worthy to note that even with one week training data our method can achieve high accuracy as 0.72. This implies that our method can work well even with a small set of initial training data.

D. Performance Comparison

We compare our traffic estimation method with a baseline method and the state-of-the-art method SVD-TE [13], [30].

With Baseline Method: Within a time frame, traffic conditions of the road segments, on which one or more taxis report traffic samplings, can also be measured by the average of instant speeds in traffic samplings, we name such a method as the *baseline method*. Since sparse taxis partially cover a road network in each time frame, baseline method provides an alternative way to obtain traffic conditions of those road segments directly covered by probe taxis only. With sparse probe taxis (i.e., 50 taxis), Fig. 14(a) shows that such naive method covers less than 45% road segments of the road network in all the four traffic scenarios, while our method can always estimate the traffic conditions of the whole road network. In Fig. 14(b), we compare their estimation accuracies. For a fair comparison, in each time frame, we compare the estimation accuracy only on the road segments with reported traffic samplings [with the same metric as Fig. 12(b)]. Due to high variance of each instant speed, the accuracy of the baseline method is only around 0.66. Our method achieves an accuracy around 0.80 in all the four scenarios and outperforms the baseline method by 17.13%. This is because our method uses traveling time and distance of taxi trajectories to estimate the average traveling speed of each road segment, which avoids the biased speed observations in traffic samplings due to dynamic taxi behaviors.

With State-of-the-Art Method: SVD-TE utilizes the spatial temporal correlation in traffic matrix constructed by recent traffic samplings to recover missing elements. Specifically, SVD-TE accumulates traffic samplings in an $n \times t$ traffic matrix and then leverages the singular value decomposition (SVD) technique to recover the missing traffic conditions in the traffic matrix. n is fixed as the total number of road segments in a road network, and t can be varying which determines a window size to accommodate traffic conditions in the SVD recovery. Fig. 15(a) shows that the estimation accuracy of SVD-TE is low when the matrix width (i.e., t) is small and stabilizes around

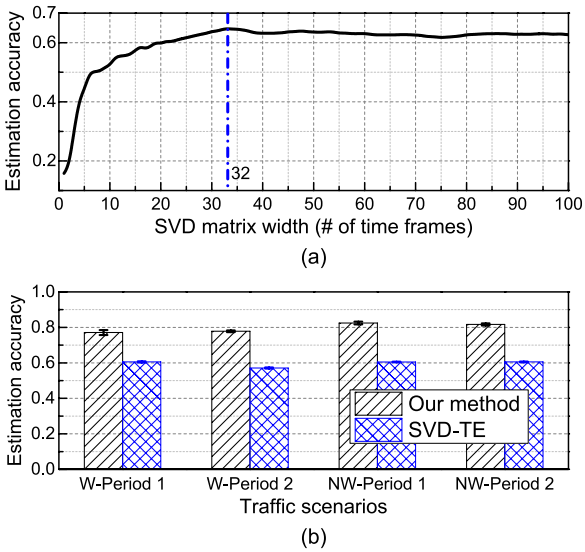


Fig. 15. (a) Estimation accuracy of SVD-TE with various matrix width t; (b) Comparison on estimation accuracy of our method and SVD-TE.

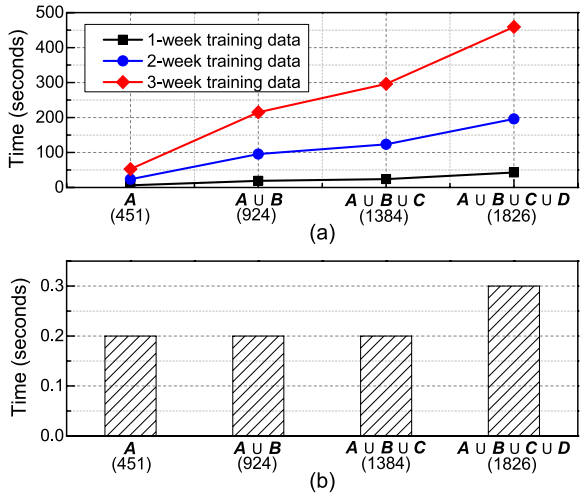


Fig. 16. Execution time of (a) the construction of MLR models; (b) compressive sensing recovery with reports from 50 taxis within one time frame.

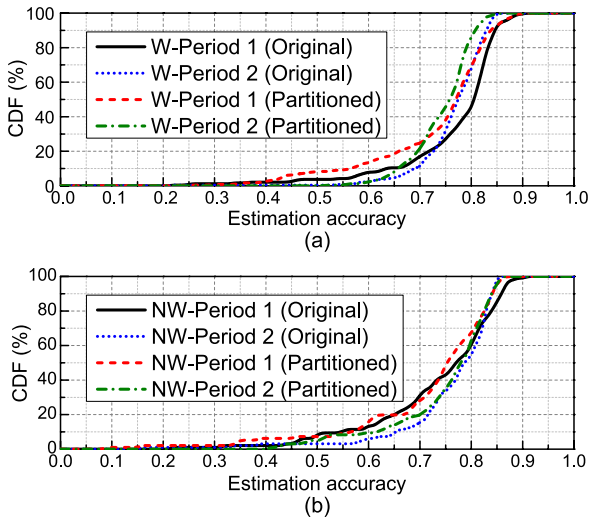


Fig. 17. Estimation accuracy in the large road network with 50.70 km^2 coverage when our method is applied using region partition and directly.

0.63 when it is greater than 32 time frames (i.e., 480 minutes). In Fig. 15(b), we compare our method with SVD-TE when t adopts 32. Our method outperforms SVD-TE on the estimation accuracy about 20%. In addition, our method can estimate traffic conditions of the entire road network every 15 minutes (i.e., one time frame). Such a short delay is necessary in practice due to the timeliness of traffic estimation. The major difference between our work and SVD-TE is that we have explicitly built the traffic correlation model and thus strengthened efficiency of compressive sensing technique, to which the great advantages of our method are attributed.

E. Scalability Evaluation

We evaluate our method in a larger road network to understand its applicability and scalability in practice. The network details have been introduced in Section V-A. To control the computation overhead, a large area could be partitioned into several regions and our method is then applied to each region, just as described in Section IV-C. The large road network in this experiment trial is divided into 4 regions, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$. Each region is slightly larger than the Jing'an area.

Computation Overhead: We will first investigate the computation overhead. We apply our method to four road networks $\mathcal{A}$, $\mathcal{A} \cup \mathcal{B}$, $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$, and $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C} \cup \mathcal{D}$, with different scales, where the union of multiple regions means regions together form a road network. All experiments in this section are conducted on a desktop with quad-core 2.66Ghz CPU and 4GB RAM, and we use the program execution time as the performance metric to evaluate the computation overhead. Two phases of our method may incur high computation overhead: MLR modeling and compressive sensing recovery, and results are detailed in Fig. 16. In a single region, e.g., $\mathcal{A}$ with 451 road segments, the construction of the MLR models can be finished within one minute. As the network scales, the MLR construction overhead linearly increases. However, the MLR model construction is infrequent and offline in our method, e.g., the MLR model could be updated every one or two months, the computation overhead is still practically acceptable, e.g., less than 8 minutes for a road network covering 50.7 km^2. If the region partition mechanism is used, the MLR model construction in each region can be parallel, which leads to a significant improvement, e.g., overall less than 1 minute overhead. On the other hand, Fig. 16(b) further indicates that the overhead due to compressive sensing recovery is negligible which is less than 0.3 seconds with traffic samplings from 50 randomly selected taxis within each 15 min time frame.

Estimation Accuracy: We finally examine the estimation accuracy of our method using the region partition mechanism on the large road network with 50.70 km^2 coverage, denoted as "Partitioned" in Fig. 17. As a benchmark, we also plot the performance when directly applying our method to the whole network, denoted as "Original." Fig. 17 shows that "Partitioned" achieves comparable accuracy with "Original" in both workdays and non-workdays. Both Figs. 16 and 17 validate the applicability and scalability of our method.

VI. Conclusion and Future Work

This paper presents a transport traffic estimation method which applies compressive sensing technique to achieve city-scale traffic estimation with only sparse traffic probes. The strong correlations among the road network is captured by an explicit model and further exploited to form a space basis that can sparsely represent the road traffic conditions. Through extensive trace-driven study and experiments, we validate the effectiveness of our traffic correlation model and show that our approach achieves accurate and scalable traffic estimation with only sparse probes.

As future works, we plan to remove the traffic estimation bias due to the inaccurate traveling time calculated from traffic data. It is also of interest to encode some transportation domain knowledge, e.g., routing behavior and network equilibrium, into our approach to further improve the estimation accuracy.

References

[1] M. Abdel-Aty and A. Pande, "ATMS implementation system for identifying traffic conditions leading to potential crashes," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 78–91, Mar. 2006.

[2] J. Aslam, S. Lim, X. Pan, and D. Rus, "City-scale traffic estimation from a roving sensor network," in *Proc. ACM SenSys*, 2012, pp. 141–154.

[3] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[4] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[5] C. Chen *et al.*, "Tripplanner: Personalized trip planning leveraging heterogeneous crowdsourced digital footprints," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1259–1273, Jun. 2015.

[6] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[7] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[8] J. Han, J. W. Polak, J. Barria, and R. Krishnan, "On the estimation of space-mean-speed from inductive loop detector data," *Transp. Plann. Technol.*, vol. 33, no. 1, pp. 91–104, Feb. 2010.

[9] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen, "Estimating arterial traffic conditions using sparse probe data," in *Proc. IEEE ITSC*, 2010, pp. 929–936.

[10] E. Horvitz, J. Apacible, R. Sarin, and L. Liao, "Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service," in *Proc. UAI*, 2005, pp. 275–280.

[11] T. Idé and M. Sugiyama, "Trajectory regression on road networks," in *Proc. AAAI*, 2011, pp. 203–208.

[12] M. H. Kutner, C. Nachtsheim, and J. Neter, *Applied Linear Regression Models*. New York, NY, USA: McGraw-Hill, 2004.

[13] Z. Li, Y. Zhu, H. Zhu, and M. Li, "Compressive sensing approach to urban traffic sensing," in *Proc. IEEE ICDCS*, 2011, pp. 889–898.

[14] Y. Liu, Z. Yang, T. Ning, and H. Wu, "Efficient quality-of-service (QoS) support in mobile opportunistic networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4574–4584, Nov. 2014.

[15] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[16] P. Misra, W. Hu, M. Yang, and S. Jha, "Efficient cross-correlation via sparse representation in sensor networks," in *Proc. ACM/IEEE IPSN*, 2012, pp. 13–24.

[17] P. Newson and J. Krumm, "Hidden Markov map matching through noise and sparseness," in *Proc. ACM GIS*, 2009, pp. 336–343.

[18] T. N. Schoepflin and D. J. Dailey, "Dynamic camera calibration of roadside traffic management cameras for vehicle speed estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 2, pp. 90–98, Jun. 2003.

[19] R. Sen *et al.*, "Kyun queue: A sensor network system to monitor road traffic queues," in *Proc. ACM SenSys*, 2012, pp. 127–140.

[20] Y. Wang, Y. Zheng, and Y. Xue, "Travel time estimation of a path using sparse trajectories," in *Proc. ACM SIGKDD*, 2014, pp. 25–34.

[21] Y. Wang, Y. Zhu, Z. He, Y. Yue, and Q. Li, "Challenges and opportunities in exploiting large-scale GPS probe data," HP Labs, Palo Alto, CA, USA, HP Tech. Rep. HPL-2011-109, 2011.

[22] X. Wu and M. Liu, "In-situ soil moisture sensing: Measurement scheduling and estimation using compressive sensing," in *Proc. ACM/IEEE IPSN*, 2012, pp. 1–11.

[23] B. Yang, C. Guo, and C. S. Jensen, "Travel cost inference from sparse, spatio temporally correlated time series using Markov models," in *Proc. VLDB*, 2013, pp. 769–780.

[24] B. Yang, M. Kaul, and C. S. Jensen, "Using incomplete information for complete weight annotation of road networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1267–1279, May 2014.

[25] J. Yoon, B. Noble, and M. Liu, "Surface street traffic estimation," in *Proc. ACM MobiSys*, 2007, pp. 220–232.

[26] J. Zhang *et al.*, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[27] J. Zheng and L. M. Ni, "Time-dependent trajectory regression on road networks via multi-task learning," in *Proc. AAAI*, 2013, pp. 1048–1055.

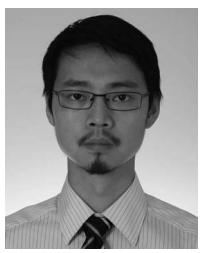
[28] P. Zhou, Y. Zheng, and M. Li, "How long to wait?: Predicting bus arrival time with mobile phone based participatory sensing," in *Proc. ACM MobiSys*, 2012, pp. 379–392.

[29] H. Zhu, Y. Zhu, M. Li, and L. M. Ni, "SEER: Metropolitan-scale traffic perception based on lossy sensory data," in *Proc. IEEE INFOCOM*, 2009, pp. 217–225.

[30] Y. Zhu, Z. Li, H. Zhu, M. Li, and Q. Zhang, "A compressive sensing approach to urban traffic estimation with probe vehicles," *IEEE Trans. Mobile Comput.*, vol. 12, no. 11, pp. 2289–2302, Nov. 2013.

Zhidan Liu (M'15) received the B.E. degree in computer science and technology from Northeastern University, Shenyang, China, in 2009 and the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2014. He is currently a Research Fellow at Nanyang Technological University, Singapore. His research interests include wireless sensor networks, mobile computing, and big data analytics.

Zhenjiang Li (M'12) received the B.E. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2007 and the M.Phil. degree in electronic and computer engineering and the Ph.D. degree in computer science and engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2009 and 2012, respectively. He is currently an Assistant Professor of computer science at City University of Hong Kong, Hong Kong. His research interests include mobile sensing and computing, wireless networks, and distributed networking systems.



Mo Li (M'06) received the B.S. degree in computer science and technology from Tsinghua University, Beijing, China, in 2004 and the Ph.D. degree in computer science and engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2009. He is currently an Assistant Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include wireless sensor networks, pervasive computing, and mobile and wireless computing.

Wei Xing received the B.E., M.E., and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1989, 1992, and 2009, respectively. In 1992, he joined the Department of Control, College of Information Technology, Zhejiang University. Since 2002, he has been with the College of Computer Science and Technology, Zhejiang University, where he is currently an Associate Professor. His research interests include multimedia technology and Internet of Things.

Dongming Lu received the B.E., M.E., and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1989, 1991, and 1994, respectively. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University. His research interests include Internet of Things, multimedia technology, and digital preservation of cultural heritage.