Estimating Urban Traffic Congestions with Multi-sourced Data

Conference Paper · June 2016

DOI: 10.1109/MDM.2016.25

CITATIONS
21

READS
122

6 authors, including:

Senzhang Wang
Nanjing University of Aeronautics & Astronautics
107 PUBLICATIONS **773** CITATIONS

SEE PROFILE

Lifang He
Lehigh University
78 PUBLICATIONS **960** CITATIONS

SEE PROFILE

Philip S. Yu
University of Illinois at Chicago
1,529 PUBLICATIONS **69,444** CITATIONS

SEE PROFILE

Estimating Urban Traffic Congestions with Multi-Sourced Data

Senzhang Wang[1], Lifang He[2*], Leon Stenneth[3], Philip S. Yu[4,5], Zhoujun Li[6], Zhiqiu Huang[1]
[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China, zqhuang@nuaa.edu.cn
[2] Shenzhen University, Shenzhen, China, lifanghescut@gmail.com
[3] Nokia's HERE Connected Driving, Chicago, USA, leon.stenneth@here.com
[4] University of Illinois at Chicago, Chicago, USA, psyu@uic.edu
[5] Institute for Data Science, Tsinghua University, Beijing, China
[6] Beihang University, Beijing, China, lizj@buaa.edu.cn

***Abstract*—This paper studies the novel problem of more accurately estimating urban traffic congestions by integrating sparse probe data and traffic related information collected from social media. Limited by the lack of reliability and low sampling frequency of GPS probes, probe data are usually not sufficient for fully estimating traffic conditions of a large arterial network. To address the data sparsity challenge, we extensively collect and model traffic related data from multiple data sources. Besides the GPS probe data, we also extensively collect traffic related tweets that report various traffic events such as congestion, accident, and road construction from both traffic authority accounts and general user accounts from Twitter. To further explore other factors that might affect traffic conditions, we also extract auxiliary information including road congestion correlations, social events, road features, as well as point of interest (POI) for help. To integrate the different types of data coming from different sources, we finally propose a coupled matrix and tensor factorization model to more accurately complete the very sparse traffic congestion matrix by collaboratively factorizing it with other matrices and tensors formed by other data. We evaluate the proposed model on the arterial network of downtown Chicago with 1257 road segments. The results demonstrate the effectiveness and efficiency of the proposed model by comparison with previous approaches.**

I. Introduction

Estimating traffic conditions in urban area is a practically important while substantially challenging problem. Traditional methods rely on various road sensors such as loop detectors [1], surveillance cameras [2], radars, etc. A major issue of these methods is that the spatiotemporal coverage of the road sensors is usually limited due to the high cost of deploying and maintaining them [3]. Currently GPS based probe vehicle data have been widely used to illuminate traffic conditions for applications including travel time estimation, map building and congestion detection [4]. However, limited by the lack of reliability and low sampling frequency of GPS probes, probe data are usually not sufficient for fully estimating traffic conditions of a large arterial network [5].

Currently, sharing real-time traffic information through social media platforms such as Twitter is becoming a common practice for both individual users and the official transportation departments [6]. As shown in Fig. 1, *Roadnow Chicago* is a Twitter account that focuses on posting traffic related tweets in Chicago. Such tweets can be about road congestions, accidents, road constructions, and other traffic events. Each tweet reports a traffic event with the event type (green rectangles) and the road segments where the event happens (red rectangles). By taking Twitter users as traffic sensors, the monitoring coverage of traffic conditions can be largely expanded as Twitter users including pedestrians, drivers, and passengers can spread over the entire city. In addition, traffic events like road construction and accident can be directly mentioned in the tweets, but are difficult to infer from GPS data. Recently, some efforts have been devoted to utilizing social media data to help understand traffic conditions [7]–[10]. These works mainly focused on studying how to extract the traffic event information from tweets [8], [9], [11], how to locate the traffic events mentioned in the tweets [10], or how to monitor traffic congestions with real-time traffic event tweets [12], [13]. These methods mostly ignored some auxiliary information including historical data, congestion correlation, and road features, while these information is very helpful especially when the traffic related social media data are sparse [13]. How to incorporate the sparse GPS probe data, social media data, and other information to build a better traffic congestion estimation model is a practically important and challenging research issue.

Fig. 1: Examples of tweets that report traffic events

In this paper, we propose a framework which can effectively

*Corresponding author.

2375-0324/16 $31.00 © 2016 IEEE
DOI 10.1109/MDM.2016.25

combine different types of traffic related information including GPS probe data, traffic related tweets, social events, road features, as well as POI features to more accurately estimate traffic congestions in urban area. Specifically, we first widely collect and process traffic related tweets from both traffic authority accounts (explain later) and general user accounts. We then regard the real-time tweets that report various traffic events and GPS probe readings as two types of primary information in the studied task. Based on the road segments mentioned in the tweets and the exact locations of the probe readings, we map both the traffic events and probe readings to the corresponding road segments by geocoding. As the real-time data are very sparse, we apply the method proposed in [13] to mine which groups of road segments are likely to co-occur congestion from a large volume of historical data. To further explore other factors that might affect traffic conditions, we also extract auxiliary information including social events, road physical features, and POI features. We finally propose a coupled matrix and tensor factorization scheme to integrate the above mentioned multi-sourced data. With the collaboratively factorized low rank matrices, we can effectively estimate the citywide traffic conditions by completing the sparse traffic congestion matrix.

The main contributions of this work are as follows.

- We propose a novel urban traffic congestion estimation framework to fuse traffic related data from multiple data sources. Compared with previous methods, the proposed method can cover a much larger area of the city and give a more accurate estimation.
- To alleviate data sparsity issue, we calculate the congestion correlations among the road segments with a spatiotemporal frequent pattern mining method. Other useful information including social events, road features, and POIs are also explored as auxiliary information.
- We model the traffic congestion estimation problem as a congestion matrix completing task, and extend the coupled matrix and tensor factorization scheme proposed in [13] to more effectively complete the congestion matrix by integrating multi-sourced data.
- We utilize the Chicago Transit Authority (CTA) public bus GPS data as the ground truth to evaluate our model. The promising experimental results demonstrate the effectiveness of the proposed model in fusing multi-sourced data for better estimating urban traffic congestions.

The remainder of the paper is organized as follows. In Section II, we give a definition of the studied problem. Section III introduces the multi-sourced data and how to collect them. Section IV describes how to mine the road segment correlations in congestion. The coupled matrix and tensor factorization schema is presented in Section V. Evaluations are given in Section VI followed by related work in Section VII. Finally, we conclude this paper in Section VIII.

II. Preliminary and Problem Definition

In this section, we first given some definitions to help state the problem. Then we briefly describe the insight of our method, and finally give a definition of the studied problem.

Definition 1: **An arterial road** R_i [13]. An arterial road R_i can be represented as a tuple $R_i = (name_i, dir_i, \mathbf{L}_i)$, where $name_i$ is the name of the arterial road, dir_i denotes the road direction, and $\mathbf{L}_i = (l_{i,1}, ..., l_{i,n})$ is the set of intersections. The intersection $l_{i,j}$ contains the exact location information and can be represented as $l_{i,j} = (lat_{i,j}, lon_{i,j})$.

Definition 2: **A road segment** r_i [13]. A segment r_i of the arterial road R_i is a continuous part of R_i. Formally, we define $r_i = (ID_i, name_i, \mathbf{l}_i)$, where ID_i is the road segment ID, $name_i$ is the name of the arterial road r_i belongs to, and $\mathbf{l}_i$ is a subset of $\mathbf{L}_i$.

Definition 3: **Road network $\mathcal{G}$.** A road network $\mathcal{G} = (E, V)$ is comprised of a set of road segments connected to each other in a graph format. $E = \{r_i\}$ is the set of the edges with each edge associated with a road segment, and $V = \{v_i\}$ is the set of the nodes with each node associated with an intersection.

If we consider each Twitter user as a sensor that monitors nearby traffic conditions, the posted traffic event related tweets can be regarded as traffic event signals which can be related to congestion, car accident, blocked road segments, etc. A traffic event tweet is defined as follows.

Definition 4: **A traffic event tweet** e_i [13]. We represent a traffic event tweet e_i as such a tuple $e_i = (c_i, \mathbf{w}_i, t_i)$ where $c_i \in C$ is the traffic event category, $\mathbf{w}_i = (w_{i,1}, ..., w_{i,N_{e_i}})$ represents the words mentioning the locations of the event, and t_i denotes the event time.

Definition 5: **A GPS probe reading** p_i. We represent a GPS probe reading p_i as such a vector $p_i = \{s, lat, lon, head, t\}$, where s is the vehicle speed, lat is the latitude, lon is the longitude, $head$ is the heading of the probe, and t denotes the time of the probe reading.

The framework of our model is comprised of three parts: 1) real-time Twitter data and probe data acquisition and processing, 2) constructing the following matrices and tensor: congestion probability matrix $\mathbf{Y}^h$, congestion correlation matrix $\mathbf{Z}$, event tensor $\mathcal{A}$, congestion matrix $\mathbf{Y}$, and confidence matrix $\mathbf{P}$, and 3) the coupled matrix and tensor factorization algorithm to complete the congestion matrix $\mathbf{Y}$.

We instantly collect tweets that report traffic events and social events, and extract the event category, location, and time information. For each GPS probe reading, we extract the time, location, and probe speed information, and then map it to the corresponding road segment. Based on these information, we construct the congestion matrix $\mathbf{Y}$, the event tensor $\mathcal{A}$, the probe congestion matrix $\mathbf{H}$, and the confidence matrix $\mathbf{Q}$. The two dimensions of $\mathbf{Y}$ are road segment ID and time slot. $y_{ij} = 1$ means that the road segment r_i is in congestion in the jth hour of a day. The three dimensions of $\mathcal{A}$ are road segment ID, event category, and time slot. $A_{ijk} = 1$ means an event with category k happens on the road segment r_i in the jth hour of a day. The matrix $\mathbf{H}$ is the probe congestion matrix with each entry h_{ij} denoting the traffic state of the road segment i in the time slot j estimated with the probe data. The matrix $\mathbf{Q}$ is the confidence matrix with each entry

q_{ij} denoting how reliable the estimated traffic state h_{ij} is. We will explain how to construct the two matrices later.

We also construct the historical congestion probability matrix $\mathbf{Y}^h$ and congestion correlation matrix $\mathbf{Z}$ based on a large volume of historical data. The former provides us with the prior knowledge of which road segments are more likely to be in congestion in some time intervals. The later shows us which groups of road segments are more likely to be congested simultaneously. The entry y^h_{ij} of $\mathbf{Y}^h$ denotes the probability of road segment r_i congested in the jth hour of a day, and the entry z_{ij} of $\mathbf{Z}$ denotes the probability of road segments r_i and r_j co-occurring congestion.

As the congestion matrix $\mathbf{Y}$ is very sparse and most entries are unknown, our goal is to perform matrix completion by utilizing coupled matrix and tensor factorization which makes full use of above mentioned rich information. Formally, the studied problem can be defined as follows.

Urban Traffic Congestion Estimation with Multi-Sourced Data: *Given the GPS probe data $\mathcal{P} = \{p_1, ..., p_l\}$, traffic event tweets $\mathcal{E} = \{e_1, ..., e_m\}$, the road segments $\mathcal{L} = \{r_1, ..., r_n\}$, the road related features $\mathcal{F} = \{f_1, ..., f_n\}$, and the time slots $\mathcal{T} = \{t_1, ..., t_k\}$, the problem is how to estimate the traffic congestions on the road segments $\mathcal{L}$ in $\mathcal{T}$, namely how to complete the sparse congestion matrix $\boldsymbol{Y}$.*

III. Data Collection

In this section, we take downtown Chicago as an example to show how we collect and process multi-sourced data. We firstly describe how we collect and process traffic related tweets and GPS probe data. To fully utilize other related information for better estimating traffic conditions, we next briefly introduce how we extract road features and social events as auxiliary information. In this paper we focus on studying the traffic congestions in Chicago. However, our method can be easily generalized to study traffic congestions of other big cities.

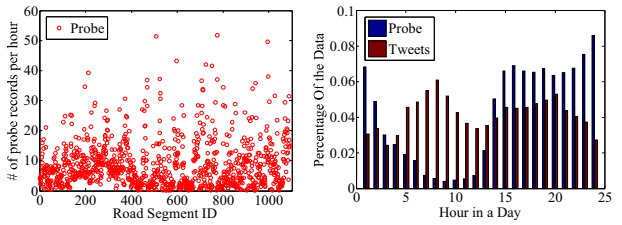
A. Twitter Data Collection

We collect traffic event tweets from the following two types of accounts as in [13]: the Twitter accounts operated by official traffic departments and general user accounts.

TABLE I: Statistics of the data in Chicago

Traffic related tweets			
#Traffic related tweets	Congestion	Accident	Others
245,568	163,742	77,454	4,372
GPS probe data			
#Probe readings		Readings per link per hour	
2,351,647		2.6	
Social event related tweets			
#Social even tweets	Parties	Shows	Sports
5,196	2,412	1,723	1,061

Traffic Authority Account. Traffic authority accounts refer to the Twitter accounts that specialize in posting traffic related information. For example, *traffic_Chicago* is such an account that focuses on releasing real-time traffic information of Chicago. The tweets posted by these accounts are formal and easier to process, and the exact location and time information are usually explicitly given. Taking the tweet "*Heavy Traffic on NB Western: Fullerton to Kennedy Expy. 06:15 pm 02/13/2015*" as an example, we can easily extract the road segment, traffic event category, and time information. We identify 11 such Twitter accounts related to Chicago, and for each account we crawl all the posted tweets from April 2014 to December 2014.



(a) Hourly # of probes for each road segment on average (b) Percentage of probes and tweets in each hour of a day

Fig. 2: Data distributions.

General Sensor User. We also collect traffic related tweets from regular users. We selected 100,000 Twitter users registered in Chicago, and crawled more than 32.3 million tweets posted by these users. Next, two major steps are conducted for data preprocessing. 1) *Traffic Event Tweets Identification.* We select traffic event tweets from all the crawled tweets which match at least one term of the predefined vocabularies: "stuck", "congestion", "jam", "crowded", "pedestrian", "driver", "accident", "crash", "road blocked", "road construction", "slow traffic", "heavy traffic", and "disabled vehicle". Based on the keywords contained in the tweets, we also can identify the traffic event category. 2) *Tweet Geocoding.* We then geocode tweets to the road segments by matching their geo-tags and text content. Some tweets are geo-tagged. By combing the geo-coordinates of tweets and the direction mentioned in the content, we can geocode the tweets to the road segments. For most tweets without geo-tags, we need to first identify the streets, landmarks, and direction information from the content by using gazetteer, and then geocode them to the road segments. We omit the tweets without any location information.

As shown in Table I, in all we collected 245,568 traffic event tweets. 163,742 of them are related to traffic congestion, 77,454 are related to accident, and 4,372 are related to other traffic events such as road construction and road closure. We categorize these tweets into three types: congestion, accident and others.

B. Probe Data

We also have more than 2.4 million GPS probe readings in the period of December 2014. These probes cover most arterial roads of downtown Chicago. Each probe reading contains the

TABLE II: Road features and social events

Road physical features	*road segment length*, *number of lanes*, *one-way road*, *heading*, *number of intersections*
POI features	*Schools*, *Hospitals*, *Museums*, *Libraries*, *Parks*, *Police Stations*, *Parking zones*, *Market & Malls*
Types of social events	*Show*, *Party*, *Concert*, *Exhibition*, *Performance*, *Live music*, *Nightlife*, *Meeting*, *Festival*, *Game*

following key information: time, latitude, longitude, heading, and speed. As the road segment information cannot directly obtained from the probe readings, we first need map the probe readings to the corresponding road segments. The road segment mapping process contains two steps. In the first step, we map the probe readings to the road segments by Google geocoding API. Google geocoding API can return the road segment name when the latitude and longitude of the probe reading is used as the input. In the second step, we confirm the traveling direction of the vehicle based on the heading information of the probe.

Fig. 2 shows the distribution of our probe data on each road segment, and the distribution of traffic related tweets in each hour of a day. Fig. 2(a) shows the average numbers of probe readings in each hour of a day for each road segment. One can see that the probe data are unevenly distributed on the arterial network. Probes frequently appear on only a small number of road segments, and for most road segments there are only very limited number of probe data. One can also see that the probe data are actually very sparse. Fig. 2(b) shows the percentages of both probe data and traffic related tweets in each hour of a day. One can see that most probe data are collected in the time interval from 14:00 to 0:00. Most traffic related tweets are posted in two time intervals from 5:00 to 10:00 and from 15:00 to 22:00. The hourly distributions of the two types of data are not perfectly consistent, which implies the combination of them could potentially improve the performance of traffic congestion estimation.

C. Road Features and Social Events

Road features are widely used for traffic estimation and prediction [13]–[15]. Wang et al. discovered that social events such as football matches and concerts can significantly affect nearby traffic conditions [13]. Thus following the work in [13], we incorporate road features and social events crawled from Twitter into our estimation model. More specially, we use the following two types of road features: road physical features, and point of interest (POI) features.

Table II shows the extracted road physical features, POI features, and various types of social events. The physical features of the road segments can be easily obtained as they are publicly available. We formulate the physical features of a road segment r as a vector f_r with each element representing a physical feature shown in Table II. For the POI features, we first extract the locations of the POIs in downtown Chicago. For each road segment we then identify the POIs nearby and formulate them as a POI feature vector f_p^r. Each element of f_p^r denotes the number of corresponding POIs nearby the road segment r. *Chicago Events* is a Twitter account that specializes in sharing various social events in Chicago. Taking the tweet "*Merger Party @ The Velvet Lounge 67 E Cermak Rd - 6pm-9pm*" as an example, the event type, time, and location information are usually given in a posted tweet. For each social event se extracted from tweets, we formulate it as such a tuple $se = \{c_{se}, p_{se}, t_{se}\}$, where c_{se} is the event category, p_{se} is the event location, and t_{se} is the event time. As shown in Tables I and II, in all we crawled 5,196 different types of social events in Chicago. Due to the data sparsity issue, we group the social events into three types: *parties*, *shows*, and *sports*. Then we use the two-dimensional Gaussian model proposed in [13] to measure the impact intensity of the social events on the nearby traffic conditions based on the Euclidean distance from the road segment to the social event location.

IV. Mining Prior Knowledge From Historical Data

As we mentioned above, the real-time probe data and traffic related tweets are both sparse. Relying on real-time data only are far from enough to fully estimate traffic conditions of an arterial network. To address this issue, previous works [12], [13] utilized a large volume of historical data as important reference information to facilitate the real-time estimation task. In this paper, we mine three types of prior knowledge from historical tweets and incorporate them into our model: the historical congestion probability and the historical event occurrence probability for each road segment in each hour of a day, and the congestion correlations among the road segments. To model the prior knowledge, we construct the historical congestion probability matrix $\mathbf{Y}^h$, the historical event occurrence probability tensor $\mathcal{A}^h$, and the congestion correlation matrix $\mathbf{Z}$. Each entry y_{ij} denotes the empirical probability of the road segment r_i being in congestion state in t_j based on statistical analysis on historical tweets. Each entry a^h_{ijk} of $\mathcal{A}^h$ denotes the possibility that an traffic or social event e_k occurs on or near road segment r_i in hour t_j of a day. Each entry z_{ij} of $\mathbf{Z}$ is the probability of road segments r_i and r_j co-occurring congestion. Following the work [13], we use the proposed spatiotemporal frequent pattern mining method to quantitively measure the congestion correlations among the road segments. Due to space limitation, we omit the technique details here.

Fig. 3 plots three typical co-congestion patterns on the road segments of Chicago. Red lines in the figure denote the road segments that are in congestion. The thick yellow lines represent express ways and the thin lines represent arterial streets. Fig. 3(a) shows a pattern with two cross road segments: *Michigan* and *Roosevelt*. Two cross road segments are more likely to co-occur congestions as they are connected to each other. Fig. 3(b) shows a pattern with three parallel road segments: *Michigan*, *State*, and *Wacker*. The three road segments all cross the Chicago Loop downtown and head to

(a) Pattern 1: roads cross to each other (b) Pattern 2: two parallel roads (c) Pattern 3: roads connecting to the same express way

Fig. 3: Three typical co-congestion patterns in Chicago.

the same direction. The Chicago Loop is the central business district of Chicago, and it is a high-traffic area. Therefore, it shows that road segments close to each other and heading to the same direction are also likely to co-occur congestion. Although the three road segments in Fig. 3(c) neither cross to each other nor parallel to each other, they are all close to the *John F. Kennedy Expressway I-90* and have entries or exits to *I-90*. It implies that the traffic conditions of the express ways and the connected arterial roads are highly correlated. When the express way is congested, the arterial roads entering or exiting the express way are also likely to be in congestion.

V. CEMD: Congestion Estimation with Multi-sourced Data

In this section we introduce the proposed urban traffic **C**ongestion **E**stimation model with **M**ulti-sourced **D**ata by coupled matrix and tensor factorization. For short, in the following parts of the paper we call our model as CEMD.

As our model uses tensor factorization techniques to facilitate matrix factorization, we first give a quick review of some notations and tensor operations. The order of a tensor is the number of dimensions, also known as ways or modes. In this paper, we use the third-order tensor. Scalars are denoted by lowercase letters, *e.g.*, a. Vectors are denoted by boldface lowercase letters, *e.g.*, $\mathbf{a}$. Matrices are denoted by boldface capital letters, *e.g.*, $\mathbf{X}$. Tensors are denoted by calligraphic letters, *e.g.*, $\mathcal{A}$. The ith entry of a vector $\mathbf{a}$ is denoted by a_i, element (i, j) of a matrix $\mathbf{X}$ is denoted by x_{ij}, and element (i, j, k) of a third-order tensor $\mathcal{A}$ is denoted by a_{ijk}. The ith row and the jth column of a matrix $\mathbf{X}$ are denoted by $\mathbf{x}_{i:}$ and $\mathbf{x}_{:j}$, respectively.

The *norm* of a tensor $\mathcal{A} \in \mathbb{R}^{N\times M\times L}$ is defined as:

$$\|\mathcal{A}\| = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} a_{ijk}^2}$$

This is analogous to the matrix *Frobenius norm*, which is denoted $\|\mathbf{X}\|$ for a matrix $\mathbf{X}$.

The *outer product* of two vectors $\mathbf{a} \in \mathbb{R}^N$ and $\mathbf{b} \in \mathbb{R}^M$, denoted by $\mathbf{a}\circ\mathbf{b}$, is a matrix of size $N \times M$ with the elements $(\mathbf{a}\circ\mathbf{b})_{ij} = a_i b_j$. The *n-mode product* of a tensor $\mathcal{C} \in \mathbb{R}^{I_1\times I_2\cdots\times I_N}$ with a matrix $\mathbf{U} \in \mathbf{R}^{I_n\times J}$, denoted by $\mathcal{C}\times_n \mathbf{U}$, is a tensor of size $I_1\times\cdots\times I_{n-1}\times J\times I_{n+1}\times\cdots\times I_N$ with the elements $(\mathcal{C}\times_n \mathbf{U})_{i_1\cdots i_{n-1}ji_{n+1}\cdots i_N} = \sum_{i_n=1}^{I_n} a_{i_1 i_2\cdots i_N} u_{i_n j}$.

The *Tucker factorization* of a tensor $\mathcal{A} \in \mathbb{R}^{N\times M\times L}$ is defined as:

$$\mathcal{A} = \mathcal{C}\times_1 \mathbf{U}\times_2 \mathbf{V}\times_3 \mathbf{W}$$

where $\mathbf{U} \in \mathbb{R}^{N\times R}, \mathbf{V} \in \mathbb{R}^{M\times S}$ and $\mathbf{W} \in \mathbb{R}^{L\times T}$ are the factor matrices. The tensor $\mathcal{C} \in \mathbb{R}^{R\times S\times T}$ is the core tensor and its entries show the level of interaction between the different components.

A. Modeling Sparse GPS Probe Data

Given a road segment r_i and the time interval t, assume the set of GPS probe readings is $\{s_1, s_2..., s_n\}$. With the probe readings, we estimate the traffic conditions of r_i in t as follows. We first average the probe speed $s_{average} = \frac{1}{n}\sum_{i=1}^{n} s_i$, and then estimate the traffic conditions based on the estimated average probe speed and 5-state traffic conditions defined by Chicago Transit Authority (CTA). CTA defines the traffic conditions of downtown Chicago as the following 5 states: heavy congestion, medium-heavy congestion, medium, light, and flow conditions. The corresponding traffic speeds are 0-10, 10-15, 15-20, 20-25, and over 25 mph, respectively. Note that except for a very few road segments, speed on arterials of downtown Chicago is limited to 30 mph by ordinance. We assign the 5 traffic states with values 1.0, 0.8, 0.6, 0.4 and 0.2, respectively. A higher value means heavier congestion. We then fill the probe congestion matrix $\mathbf{H}$ with the traffic state values. Each entry $h_{i,j}$ of $\mathbf{H}$ denotes the congestion state of the road segment r_i in time interval j. For example, $h_{i,j} = 1.0$ means that the probe speed on the road segment r_i in j is less than 10 miles per hour, and thus it is in heavy congestion state.

Note that the reliability of the estimated traffic states largely relies on how many probe readings are available for each road segment and in each time interval. More readings imply a more reliable estimation; otherwise the estimation is unreliable. As shown in Fig. 2(a), the probe data are very sparse. Give a time interval there may be only be a few readings or even

no readings on a particular road segment. To quantitively measure the reliability of the traffic states estimated by probes, we construct a confidence matrix $\mathbf{Q}$. Each entry q_{ij} of $\mathbf{Q}$ is calculated by such a sigmoid function

$$q_{ij}(n) = \frac{1}{1 + e^{n - Cardinality(p_{ij})}} \tag{1}$$

where n is a predefined threshold of the probe reading size. In this paper we set n to 3. One can see that more probe readings can result in a larger q_{ij}, which means the estimated traffic state is more reliable.

B. Coupled Matrix and Tensor Factorization to Integrate Multi-Sourced Data

The insight of using matrix and tensor factorization to estimate urban traffic congestion is: given a very sparse road congestion matrix $\mathbf{Y}$, try to complete $\mathbf{Y}$ by factorizing it into two low rank latent matrices [13]. Before we introduce our method, we first describe some symbols as follows. $\mathcal{A} \in \mathbb{R}^{N \times M \times L}$ represents the event tensor, $\mathbf{X} \in \mathbb{R}^{N \times K}$ represents the road feature matrix, $\mathbf{Y} \in \mathbb{R}^{N \times M}$ is the congestion matrix, $\mathbf{Q} \in \mathbb{R}^{N \times M}$ is the confidence matrix, $\mathbf{H} \in \mathbb{R}^{N \times M}$ is the probe congestion matrix, and $\mathbf{Z} \in \mathbb{R}^{N \times N}$ is the congestion correlation matrix. Here N is the number of road segments, M is the number of time slots (hour) per day, K is the number of road features, and L is the number of event categories.

As the congestion matrix $\mathbf{Y}$ is very sparse, factorizing it directly usually cannot achieve promising results. As shown in Fig. 4, to fully utilize other information and more accurately factorize $\mathbf{Y}$, we factorize it collaboratively with the matrices $\mathbf{H}$, $\mathbf{X}$, $\mathbf{Z}$ and the tensor $\mathcal{A}$. The road feature matrix $\mathbf{X}$ can be factorized into the multiplication of two matrices, $\mathbf{X} = \mathbf{U} \times \mathbf{F}$, where $\mathbf{U} \in \mathbb{R}^{N \times R}$ and $\mathbf{F} \in \mathbb{R}^{R \times K}$ are low rank latent factors for road segments and geographical features, respectively. We factorize the road feature matrix $\mathbf{X}$ based on the idea that road segments with similar road features are more likely to present similar traffic conditions. Likewise, the congestion matrix $\mathbf{Y}$ and $\mathbf{H}$ can be both factorized into the multiplication of two matrices, $\mathbf{Y} = \mathbf{U} \times \mathbf{V}^{\mathrm{T}}$ and $\mathbf{H} = \mathbf{U} \times \mathbf{V}^{\mathrm{T}}$, where $\mathbf{V} \in \mathbb{R}^{M \times R}$ is a low rank latent factor matrix for time slots. We assume the congestion matrix $\mathbf{Y}$ and the probe congestion matrix $\mathbf{H}$ share the same low rank latent matrices $\mathbf{U}$ and $\mathbf{V}$, because the final congestion matrix $\mathbf{Y}$ should be similar to the probe congestion matrix $\mathbf{H}$ as much as possible. The event tensor can be factorized as $\mathcal{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{L \times T}$ is a low rank latent factor matrix for event categories. The idea is that road segments with similar traffic events occurring in the same time interval are more likely to present similar traffic conditions.

The objective function is defined as follows,

$$\begin{aligned}
&\mathcal{L}(\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathcal{C}, \mathbf{F}) = \frac{1}{2}||\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}||^2 + \\
&\frac{\lambda_1}{2}||\mathbf{Q} \odot (\mathbf{H} - \mathbf{U}\mathbf{V}^{\mathrm{T}})||^2 + \frac{\lambda_2}{2}||\mathcal{A} - \mathcal{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}||^2 \\
&+ \frac{\lambda_3}{2}||\mathbf{X} - \mathbf{U}\mathbf{F}||^2 + \frac{\lambda_4}{2}\mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{L}_z\mathbf{U}) \\
&+ \frac{\lambda_5}{2}(||\mathbf{U}||^2 + ||\mathbf{V}||^2 + ||\mathbf{W}||^2 + ||\mathcal{C}||^2 + ||\mathbf{F}||^2)
\end{aligned} \tag{2}$$

where $\odot$ represents the Hadamard product of two matrices, and $\mathrm{tr}(\cdot)$ denotes the matrix trace. $||\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}||^2$ is to control the error of factorization of $\mathbf{Y}$. $||\mathbf{Q} \odot (\mathbf{H} - \mathbf{U}\mathbf{V})||^2$ is to make the final estimated congestion matrix $\mathbf{Y}$ similar to the probe congestion matrix $\mathbf{H}$. $||\mathcal{A} - \mathcal{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}||^2$ is to control the error of factorizing tensor $\mathcal{A}$. $||\mathbf{X} - \mathbf{U}\mathbf{F}||^2$ is to control the error of factorization of $\mathbf{X}$. $||\mathbf{U}||^2 + ||\mathbf{V}||^2 + ||\mathbf{W}||^2 + ||\mathcal{C}||^2 + ||\mathbf{F}||^2$ is a regularization penalty to avoid overfitting, λ_1, λ_2, λ_3, λ_4, and λ_5 are parameters to control the contribution of each part. $\mathbf{L}_z = \mathbf{D} - \mathbf{Z}$ is the Laplacian matrix of the road segment congestion correlation graph in which $\mathbf{D}$ is a diagonal matrix with diagonal entries $d_{ii} = \sum_i z_{ij}$. $\mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{L}_z\mathbf{U})$ is used to guarantee two road segments r_i and r_j with a higher congestion correlation (*i.e.*, z_{ij} is big) should also have a closer distance between the vector $\mathbf{u}_i$ and $\mathbf{u}_j$ in the matrix $\mathbf{U}$.

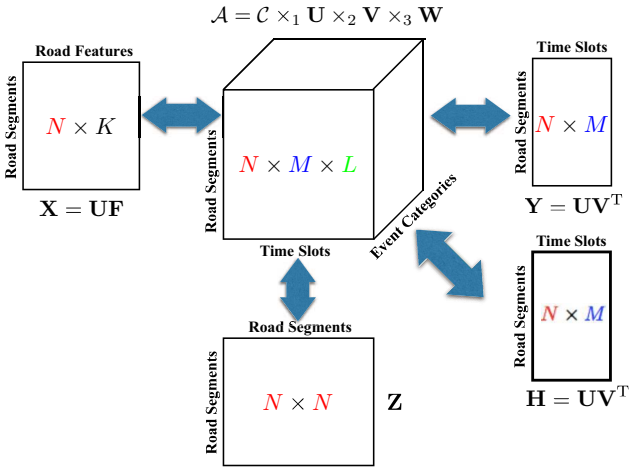


Fig. 4: Context-aware matrix and tensor factorization

C. Integrating Historical Data

To integrate the prior knowledge mined from historical data to further alleviate the sparsity issue of real-time data, we incorporate the historical congestion probability matrix $\mathbf{Y}^h$ and the historical road event probability tensor $\mathcal{A}^h$ into our model. By combining the historical prior knowledge, we have the following objective function

$$\begin{aligned}
&\mathcal{L}(\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathcal{C}, \mathbf{F}) = \frac{1}{2}||\mathbf{Y} - \mathbf{U}\mathbf{V}^{\mathrm{T}}||^2 + \\
&\frac{\lambda_1}{2}||\mathbf{Q} \odot (\mathbf{H} - \mathbf{U}\mathbf{V}^{\mathrm{T}})||^2 + \frac{\lambda_2}{2}||\mathbf{Y}^h - \mathbf{U}\mathbf{V}^{\mathrm{T}}||^2 + \\
&\frac{\lambda_3}{2}||\mathcal{A} - \mathcal{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}||^2 + \frac{\lambda_4}{2}||\mathbf{X} - \mathbf{U}\mathbf{F}||^2 + \\
&\frac{\lambda_5}{2}||\mathcal{A}^h - \mathcal{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}||^2 + \frac{\lambda_6}{2}\mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{L}_z\mathbf{U}) + \\
&\frac{\lambda_7}{2}(||\mathbf{U}||^2 + ||\mathbf{V}||^2 + ||\mathbf{W}||^2 + ||\mathcal{C}||^2 + ||\mathbf{F}||^2)
\end{aligned} \tag{3}$$

where $||\mathbf{Y}^h - \mathbf{U}\mathbf{V}^{\mathrm{T}}||^2$ is to control the error of factorizing the historical congestion probability matrix $\mathbf{Y}^h$, and $||\mathcal{A}^h - \mathcal{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}||^2$ is to control the error of factorizing the historical event tensor $\mathcal{A}^h$. The insight is that the congestion states of the roads should be similar to their

historical congestion states. Therefore, we assume that the congestion matrix $\mathbf{Y}$ should be similar to $\mathbf{Y}^h$, and the two matrices should share the same low rank factor matrices $\mathbf{U}$ and $\mathbf{V}$. Likewise, we also assume that the event tensor $\mathcal{A}$ and the historical event tensor $\mathcal{A}^h$ share the same factor matrices $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and core tensor $\mathcal{C}$.

The objective function is not jointly convex to all the variables $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathcal{C}$, and $\mathbf{F}$. Thus it is very hard to get closed-form solutions to minimize the objective function. We use an element-wise optimization algorithm to iteratively update each entry in the matrices and tensor independently by gradient descent [13]–[16]. We omit the algorithm detail here, and one can refer to the works [13], [14] for more details for solving this problem. Here we only list the gradient for each variable as follows.

$$\begin{aligned}&\nabla_{\mathbf{u}_{i:}}\mathcal{L} = (\mathbf{u}_{i:}\mathbf{V}^{\mathrm{T}} - \mathbf{y}_{i:})\mathbf{V} + \\ &\lambda_1[\mathbf{q}_{i:} \odot (\mathbf{u}_{i:}\mathbf{V}^{\mathrm{T}} - \mathbf{h}_{i:})]diag(\mathbf{q}_{i:})\mathbf{V} + \lambda_2(\mathbf{u}_{i:}\mathbf{V}^{\mathrm{T}} - \mathbf{y}^h_{i:})\mathbf{V} \\ &+ \lambda_3(\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} - a_{ijk})\mathcal{C} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} + \\ &\lambda_5(\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \times_4 \mathbf{w}_{k:} - a^h_{ijk})\mathcal{C} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} + \\ &\lambda_4(\mathbf{u}_{i:}\mathbf{F} - \mathbf{x}_{i:})\mathbf{F}^{\mathrm{T}} + \lambda_6(\mathbf{L}_z\mathbf{U})_{i:} + \lambda_7\mathbf{u}_{i:}\end{aligned} \tag{4}$$

$$\begin{aligned}&\nabla_{\mathbf{v}_{j:}}\mathcal{L} = (\mathbf{v}_{j:}\mathbf{U}^{\mathrm{T}} - \mathbf{y}_{:j}{}^{\mathrm{T}})\mathbf{U} + \lambda_2(\mathbf{v}_{j:}\mathbf{U}^{\mathrm{T}} - \mathbf{y}^h{}_{:j}{}^{\mathrm{T}})\mathbf{U} + \\ &\lambda_1[\mathbf{q}^{\mathrm{T}}_{:j} \odot (\mathbf{v}_{j:}\mathbf{U}^{\mathrm{T}} - \mathbf{h}^{\mathrm{T}}_{:j}]diag(\mathbf{q}_{:j})\mathbf{U} + \lambda_7\mathbf{v}_{j:} + \\ &\lambda_3(\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} - a_{ijk})\mathcal{C} \times_1 \mathbf{u}_{i:} \times_3 \mathbf{w}_{k:} + \\ &\lambda_5(\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} - a^h_{ijk})\mathcal{C} \times_1 \mathbf{u}_{i:} \times_3 \mathbf{w}_{k:}\end{aligned} \tag{5}$$

$$\begin{aligned}&\nabla_{\mathbf{w}_{k:}}\mathcal{L} = \lambda_6\mathrm{w}_{k:} + \\ &\lambda_2(\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} - a_{ijk})\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \\ &+ \lambda_3(\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} - a^h_{ijk})\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:}\end{aligned} \tag{6}$$

$$\begin{aligned}&\nabla_{\mathcal{C}}\mathcal{L} = \lambda_2(\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} - a_{ijk})\mathbf{u}_{i:} \circ \mathbf{v}_{j:} \circ \mathbf{w}_{k:} \\ &+ \lambda_3(\mathcal{C} \times_1 \mathbf{u}_{i:} \times_2 \mathbf{v}_{j:} \times_3 \mathbf{w}_{k:} - a^h_{ijk})\mathbf{u}_{i:} \circ \mathbf{v}_{j:} \circ \mathbf{w}_{k:} \\ &+ \lambda_6\mathcal{C}\end{aligned} \tag{7}$$

$$\nabla_{\mathrm{F}}\mathcal{L} = \lambda_4\mathbf{u}_{i:}{}^{\mathrm{T}}(\mathbf{u}_{i:}F - \mathbf{x}_{i:}) + \lambda_6\mathrm{F} \tag{8}$$

VI. Experimental Results

In this section, we evaluate the performance of the proposed CEMD model on estimating traffic congestions of downtown Chicago. We first describe the experiment setup, including the ground truth data, the baselines, and evaluation metrics. Next we perform quantitively evaluations of different methods. Especially, we test the performance of different methods in rush hours of a day. Finally, we evaluate the scalability of the proposed CEMD model.

A. Ground Truth

We use the GPS traces data of Chicago Transit Authority (CTA) public passenger buses as our ground truth. This dataset includes the traffic speed data of Chicago's arterial streets in real-time by continuously monitoring and analyzing GPS traces received from more than 2,000 CTA public passenger buses. The GPS probes of the CTA buses report their current locations and speed information in every ten minutes. The entire bus routes contain 1,257 road segments covering nearly 700 miles of arterial roads. We use the publicly available historical traffic data collected from 11/25/2014 to 12/30/2014 which contains more than 500 million records. Each record contains time, bus ID, road segment ID, the number of buses on the road segment, and the average speed. For each road segment in each hour, we use the real-time average speed as the ground truth speed if there are more than 5 probe readings; otherwise we use a weighted average between the historical and real time speeds. We use the method described in Section 5.1 to assign values denoting the traffic conditions based on the bus speed. Roughly we consider a road segment is in congestion if the speed is lower than 15 mph. Note that the GPS data of CTA buses is different from the probe data used in our model. The probes used in our model are installed on general vehicles rather than buses.

B. Baselines and Evaluation Metrics

We use the following methods as baselines.

- CF. Collaborative filtering (CF) is widely used in recommendation [17]. We can consider the congestion estimation task as a CF problem by factorizing the road congestion matrix only. In CF, only the congestion matrix $\mathbf{Y}$ is used and factorized.
- TSE. TSE is a Traffic Speed Estimation model based on a context aware matrix factorization approach [14]. As the basic idea of TSE is similar to our method, we apply this method to the congestion estimation task as a comparison. We apply TSE model to co-factorize the following matrices: the road feature matrix $\mathbf{X}$, the traffic congestion matrix $\mathbf{Y}$, the historical congestion matrix $\mathbf{Y}^h$, and the congestion correlation matrix $\mathbf{Z}$.
- CTCE. CTCE is our previously proposed model [13]. In this model, Twitter is used as the major data source to estimate urban traffic congestions. We conduct a comparison with this model to investigate whether incorporating probe data can further improve the performance.

We use the evaluation metric *precision@k* to conduct a coarse-grained estimation with only 2 traffic states: congestion and flow. Specifically, we first complete the road congestion matrix $\mathbf{Y}$ by multiplying the low rank matrices $\mathbf{U}$ and $\mathbf{V}^{\mathrm{T}}$, and then rank the values of all the entries in $\mathbf{Y}$. As in a particular time slot, usually only a small portion of road segments are in congestion and most others are not, we consider the road segments with top-k entry values are in congestion.

To conduct a fine-grained evaluation, we further evaluate the proposed CEMD model in estimating traffic conditions with 5 traffic states in rush hours. Hence we also use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)

https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/77hq-huss

TABLE III: Average precision @k of different methods

Average precision @k on Weekday					
	top-10	top-20	top-30	top-50	top-100
CF	0.500	0.437	0.414	0.400	0.412
TSE	0.821	0.786	0.774	0.720	0.676
CTCE	0.873	0.866	0.853	0.800	0.824
CEMD	**0.892**	**0.874**	**0.866**	**0.840**	**0.838**
Average precision @k on Weekend					
	top-10	top-20	top-30	top-50	top-100
CF	0.485	0.436	0.472	0.440	0.415
TSE	0.812	0.821	0.785	0.800	0.735
CTCE	0.854	0.834	0.822	0.820	0.754
CEMD	**0.872**	**0.847**	**0.838**	**0.822**	**0.767**

as evaluation metrics by normalizing the entry values in the range from 0.2 to 1.

$$MAE = \frac{\Sigma_i |y_i - \hat{y}_i|}{N} \text{ and } RMSE = \sqrt{\frac{\Sigma_i (y_i - \hat{y}_i)^2}{N}}$$

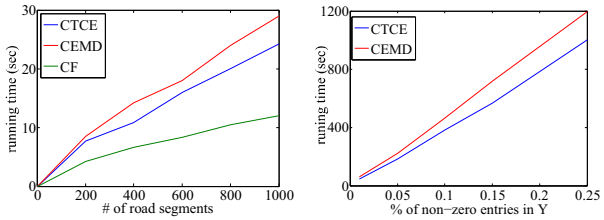
where y_i is the ground truth and $\hat{y}_i$ is the estimated value associated to the traffic state.

C. Estimation Accuracy with 2 Traffic States

We first evaluate whether the proposed CEMD model can more accurately estimate whether a road segment is in congestion (speed smaller than 15 mph) or not. Table III shows the average *precision*@k of the four methods over various k. As the traffic conditions on weekdays and weekends are different, we present the results by weekday and weekend separately. The best results are highlighted by bold. One can see that CEMD model achieves the best performance in all the cases. The performance of CF model is significantly lower than all the other methods. This is because the available information is very limited if other information are ignored. One can see that CEMD model outperforms CTCE model, which demonstrates the combination of probe data and social media data can further improve the estimation performance. TSE is much better than CF, but inferior to CTCE and CEMD. One can also see that the estimation performance on weekdays is better than on weekends, which means the traffic conditions on weekends are harder to estimate due to the irregular travel routes of people.

D. Quantitive Evaluation in Rush Hours

To conduct a fined grained evaluation, we report the average MAE and RMSE of different methods in rush hours of a day. Table IV gives the results on the rush hours 6:00-10:00 and 15:00-19:00, respectively. The figures in bold denotes the best results. One can see that in most cases the proposed CEMD model achieves the best performance. Both CTCE model and CEMD model are significantly better than the other two methods, which implies combing multiple traffic related data do help us better estimate traffic congestions. By comparing the results of CTCE and CEMD, one can see that in general CEMD works better. However, in some hours CEMD is inferior to CTCE model. The possible reasons could be as follows. First, the probe data are very noisy and unevenly distributed among the road segments. Second, there might be a large overlap between the information provided by Twitter data and probe data, and thus the new information provided by probe data can be limited.



(a) time *vs* # of segments (b) time *vs* sparsity of Y

Fig. 5: Scalability study

Similar to our previous experimental result, it also shows that the traffic conditions in weekday are easier to estimate than that in weekend. For example, the MAE in the hour 6:00-7:00 estimated by CTCE and CEMD are 0.1259 and 0.1224 for weekday, and the number is 0.1375 and 0.1253 for weekend. Similar result can also be found for RMSE. CF is significantly inferior to other methods. TSE is better than CF, but significantly inferior to other methods.

E. Scalability Analysis

To study the scalability of the proposed CEMD, we compare the running time of CEMD with CF and CTCE, and the results are shown in Fig. 5. Fig. 5(a) shows the increase trend of the running time for the three methods CTCE, CEMD and CF with the rising of road segment number. One can see that the running time of the three models linearly increases with the increase of road segment size. The running time of CEMD is consistently longer that CTCE and CF. This is because CEMD combines the probe data and Twitter data, and thus CEMD needs to factorize more matrices. However, the figure shows that CEMD is not time consuming. For the congestion matrix $\mathbf{Y}$ with 1000 road segments, the running time of CEMD is only around 30 seconds due to the very sparse matrices. To further study the effect of congestion matrix $\mathbf{Y}$ sparsity on the running time, we plot the running times of CTCE and CEMD under various proportions of non-zero entires in $\mathbf{Y}$. One can see that a denser $\mathbf{Y}$ leads to a much longer running time of CTCE, and the running time of CEMD is always slightly longer than CTCE. In real case the congestion matrix $\mathbf{Y}$ is usually very sparse with less than 5% of non-zero entries. Thus, both CTCE and CEMD models can easily scale to large data with thousands of road segments in urban area.

VII. Related Work

Compared to traffic monitoring on highways, traffic monitoring on an arterial network of the urban area presents additional challenges due to the lack of ubiquity and a uniform penetration rate of the probe data and the randomness of their spatiotemporal coverage [18], [19]. Previous researches on traffic estimation on an arterial network can be roughly categorized into traffic modeling on individual roads [1], [20]–[22] and on a road network [14], [23], [24]. Helbing employed

TABLE IV: MAE and RMSE of different methods in rush hours

	6:00-7:00	7:00-8:00	8:00-9:00	9:00-10:00	15:00-16:00	16:00-17:00	17:00-18:00	18:00-19:00
MAE on Weekday								
CF	0.2233	0.2343	0.2682	0.2737	0.2250	0.2284	0.2109	0.2566
TSE	0.2178	0.2047	0.2016	0.2260	0.1653	0.2016	0.1627	0.1699
CTCE	0.1259	0.1215	0.1208	**0.1232**	0.1201	0.1180	0.1233	0.1267
CEMD	**0.1224**	**0.1204**	**0.1138**	0.1243	**0.1167**	**0.1098**	**0.1232**	**0.1203**
MAE on Weekend								
CF	0.2542	0.2913	0.2252	0.2313	0.2136	0.2223	0.2008	0.2187
TSE	0.2335	0.2123	0.2059	0.1980	0.1674	0.1378	0.1866	0.1968
CTCE	0.1375	0.1281	**0.1345**	0.1271	**0.1287**	0.1260	0.1250	0.1268
CEMD	**0.1253**	**0.1266**	0.1348	**0.1224**	0.1354	**0.1142**	**0.1136**	**0.1246**
RMSE on Weekday								
CF	0.2568	0.2715	0.2953	0.3108	0.2659	0.2650	0.3162	0.3167
TSE	0.2348	0.2163	0.2495	0.2695	0.1977	0.2284	0.1865	0.1932
CTCE	0.1582	0.1563	0.1556	0.1562	0.1537	**0.1543**	0.1567	0.1643
CEMD	**0.1426**	**0.1522**	**0.1489**	**0.1425**	**0.1527**	0.1556	**0.1534**	**0.1582**
RMSE on Weekend								
CF	0.2967	0.3387	0.3452	0.3114	0.2664	0.2649	0.2565	0.3127
TSE	0.2690	0.2701	0.2484	0.2331	0.2105	0.1842	0.2361	0.2360
CTCE	0.1732	0.1654	**0.1637**	0.1604	0.1572	0.1620	0.1634	0.1681
CEMD	**0.1684**	**0.1527**	0.1662	**0.1557**	**0.1526**	**0.1583**	**0.1624**	**0.1667**

a Fundamental Diagram to learn the relations among vehicle speed, traffic density, and volume for a particular road to estimate traffic condition on an individual road [21]. Muoz et al. proposed a macroscopic traffic flow model SMM by utilizing the loop detector data to estimate the traffic density at unmonitored locations along a highway [1]. Porikli and Li proposed a Gaussian Mixture Hidden Markov Models (GM-HMM) to detect traffic condition with the MPEG video data [20]. Researches on traffic monitoring on a road network usually need to capture and model the correlations of the traffic conditions among the road segments connected to each other [14], [22]–[25]. Such models mainly utilized the Floating Car Data (FCD) [24] or probe data [25] generated by the GPS sensors equipped in vehicles. Herring et al. proposed a coupled Hidden Markov Model which can effectively capture the traffic congestion correlations among the road segments [25]. Fabritiis et al. studied the problem of using real-time FCD data based on traces of GPS positions to predict the traffic on Italian motorway network [22]. Yuan investigated how to use the trajectories of taxis collected by GPS to efficiently find driving directions for drivers [24]. With the availability of other rich information like POIs and road features, some recent researches tried to explore these features to help estimate traffic condition. Shang et al. proposed a context-aware matrix factorization algorithm to estimate the traffic speed of the road network in Beijing by integrating POIs and road geographic features [14].

Recently, exploring traffic related information from social media like Twitter to detect traffic events or monitor traffic conditions has been a hot research topic [6], [8], [9], [11]–[13]. Zheng surveyed the methodologies for cross-domain data fusion, including how to fuse different types of traffic related data to help various traffic related applications [26]. Most previous works focused on investigating either how to extract and visualize the traffic event information from tweets [6], [8], [9], [11] or how to locate the traffic events mentioned in the tweets [7], [10]. As the traffic events are usually imbalanced data which means that only a small number of road segments are congested or occur traffic accidents in a particular time interval, imbalanced classification [27] and event detection techniques are usually used to detect and categorize traffic related tweets with a large volume of tweets [8], [11]. The work in [12] is the first to estimate traffic congestion of an arterial network by collecting traffic related tweets from Twitter. Wang et al. further incorporated other information such as social events and road features with social media data to more effectively estimate citywide traffic congestions [13]. However, as the probe data are not explored, the performance are usually not desirable due to very sparse and noisy Twitter data [12]. How to effectively fuse social media data with traditional road sensor data to further improve the performance of various traffic related applications remains an open problem.

VIII. Conclusion

This paper proposes a novel framework to effectively integrate multi-sourced data for better estimating urban traffic congestions. As the traditionally used GPS probe data are usually sparse and noisy, their spatiotemporal coverage is very limited. Thus relying on the probe data only is not sufficient to precisely estimate traffic conditions of a large arterial network. To address this issue, we extensively collect other traffic related information from different data sources. The multi-sourced data include GPS probe data, traffic information from social media, road features, social events, as well as traffic congestion prior knowledge mined from historical data. To effectively integrate above information, we extend the previously proposed coupled matrix and tensor factorization method to estimate traffic conditions through completing the congestion matrix with the help of other matrices and tensors formed by the multi-sourced data. Evaluations on the real arterial network of downtown Chicago verify the effectiveness and efficiency of the proposed method.

IX. Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant Nos. 61370126, 61202239, 61303017, 61503253), National High Technology Research and Development Program of China under grant (No. 2015AA016004), and US NSF through grants (III-1526499, CNS-1115234, OISE-1129076).

References

[1] L. Muñoz, X. Sun, R. Horowitz, and L. Alvarez, "Traffic density estimation with the cell transmission model," in *Proceedings of the 2003 American Control Conference*, 2003.

[2] C. Ozkurt and F. Camci, "Automatic traffic density estimation and vehicle classification for traffic surveillance systems using neural networks," *Mathematical and Computational Application*, vol. 14, no. 3, pp. 187–196, 2009.

[3] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. On Intelligent Systems and Technology*, vol. 5, no. 3, 2014.

[4] S. Tao, V. Manolopoulos, S. Rodriguez, and A. Rusu, "Real-time urban traffic state estimation with a-gps mobile phones as probes," *Journal of Transportation Technologies*, vol. 2, no. 1, pp. 22–31, 2012.

[5] Y. Wang, Y. Zhu, and Z. He, "Challenges and opportunities in exploiting large-scale gps probe data," in *Technical Report, HPL-2011-109*, 2011.

[6] S. K. Endarnoto, S. Pradipta, A. S. Nugroho, and J. Purnama, "Traffic condition information extraction and visualization from social media twitter for android mobile application," in *Proceedings of International Conference on Electronics Engineering and Informatics*, 2011.

[7] E. M. Daly, F. Lecue, and V. Bicer, "Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions," in *Proceedings of International Conference on Intelligent User Interfaces*, 2013.

[8] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from twitter stream analysis," *Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1–15, 2015.

[9] M. Liu, K. Fu, C.-T. Lu, G. Chen, and H. Wang, "A search and summary application for traffic events detection based on twitter data," in *Proceedings of ACM SIGSPATIAL International Conferences on Advances in Geographic Information Systems*, 2014.

[10] J. Sílvio S. Ribeiro, J. Clodoveu A. Davis, D. R. R. Oliveira, J. Wagner Meira, T. S. Gonçalves, and G. L. Pappa, "Traffic observatory: a system to detect and locate traffic events and conditions using twitter," in *Proceedings of ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN)*, 2012.

[11] A. Schulz, P. Ristoski, and H. Paulheim, "I see a car crash: Real-time detection of small scale incidents in microblogs," in *ESWC*, 2013.

[12] P.-T. Chen, F. Chen, and Z. Qian, "Road traffic congestion monitoring in social media with hinge-loss markov random fields," in *Proceedings of IEEE International Conference on Data Mining*, 2014.

[13] S. Wang, L. He, L. Stenneth, P. S. Yu, and Z. Li, "Citywide traffic congestion estimation with social media," in *Proceedings of ACM SIGSPATIAL International Conferences on Advances in Geographic Information Systems*, 2015.

[14] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[15] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, and E. Chang, "Diagnosing new york city's noises with ubiquitous data," in *Proceedings of ACM Joint International Conference on Pervasive and Ubiquitous Computing*, 2014.

[16] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse recommendation: n-dimensional tensor factorization for context aware collaborative filtering," in *Proceedings of ACM Recommender Systems*, 2010.

[17] B. Sarwar, G. karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of International World Wide Web Conference*, 2001.

[18] T. Park and S. Lee, "A bayesian approach for estimating link travel time on urban arterial road network," in *Proceedings of the International Conference on Computational Science and Its Applications*, 2004.

[19] H. J. van Zuylen, F. Zheng, and Y. Chen, "Using probe vehicle data for traffic state estimation in signalized urban networks," in *Traffic Data Collection and its Standardization*. Springer New York, 2010, vol. 144.

[20] F. Porikli and X. Li, "Traffic congestion estimation using hmm models without vehicle tracking," in *Intelligent Vehicles Symposium*, 2004.

[21] D. Helbing, "Traffic and related self-driven many-particle systems," *Reviews of modern physics*, 2001.

[22] C. de Fabritiis, R. Ragona, and G. Valenti, "Traffic estimation and prediction based on real time floating car data," in *Proceedings of International Conference on Information Technology and Computer Science*, 2008.

[23] W. Pattara-Atikom, P. Pongpaibool, and S. Thajchayapong, "Estimating road traffic congestion using vehicle velocity," in *ITS Telecommunications*, 2006.

[24] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving directions based on taxi trajectories," in *Proceedings of ACM SIGSPATIAL International Conferences on Advances in Geographic Information Systems*, 2010.

[25] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen, "Estimating arterial traffic conditions using sparse probe data," in *Proceedings of International IEEE Conference on Intelligent Transportation Systems*, 2010.

[26] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," in *IEEE Transactions on Big Data*, vol. 1, no. 1, 2015.

[27] S. Wang, Z. Li, W.-H. Chao, and Q. Cao, "Applying adaptive oversampling technique based on data density and cost-sensitive svm to imbalanced learning," in *Proceedings of International Joint Conference on Neural Networks*, 2012.