Enhancing Traffic Congestion Estimation with Social Media by Coupled Hidden Markov Model

Senzhang Wang[1(✉)], Fengxiang Li[2], Leon Stenneth[3], and Philip S. Yu[4,5]

[1] Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
szwang@nuaa.edu.cn
[2] Peking University, Beijing 100871, China
1200012427@pku.edu.cn
[3] Nokia's HERE Connected Driving, Chicago, IL, USA
leon.stenneth@here.com
[4] University of Illinois at Chicago, Chicago, IL 60607, USA
psyu@uic.edu
[5] Institute for Data Science, Tsinghua University, Beijng 100084, China

Abstract. Estimating traffic conditions in arterial networks with GPS probe data is a practically important while substantially challenging problem. With the increasing availability of GPS equipments installed in various vehicles, GPS probe data is currently becoming a significant data source for traffic monitoring. 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. For the first time this paper studies how to explore social media as an auxiliary data source and incorporate it with GPS probe data to enhance traffic congestion estimation. Motivated by the increasing amount of traffic information available in Twitter, we first extensively collect tweets that report various traffic events such as congestion, accident, and road construction. Next we propose an extended Coupled Hidden Markov Model which can effectively integrate GPS probe readings and traffic related tweets to more accurately estimate traffic conditions of an arterial network. To address the computational challenge, a sequential importance sampling based EM algorithm is also introduced. We evaluate the proposed model on the arterial network of downtown Chicago. The experimental results demonstrate the superior performance of the model by comparison with previous methods.

Keywords: Social media · Traffic estimation · CHMM

1 Introduction

Conventional traffic monitoring methods rely on road sensor data collected from various sensors such as loop detectors [14], surveillance cameras [4], and radars.

P. Frasconi et al. (Eds.): ECML PKDD 2016, Part II, LNAI 9852, pp. 247–264, 2016.
DOI: 10.1007/978-3-319-46227-1_16

Due to the high cost of deploying and maintaining such devices, their spatialtemporal coverage is usually very limited. Recently, GPS based probe vehicle data have become a significant data source available for the arterials and highways not covered by dedicated sensing infrastructure. As such, there is considerable research interest in exploring GPS probes for conducting various traffic related applications [6,20]. However, the characteristics of probe data, including the lack of reliability, low sampling frequency, and the randomness of its spatiotemporal coverage, make it insufficient for fully estimating traffic states for large transportation networks [5].

Currently, it is a common practice for drivers and official transportation departments to release instant traffic information through social media [12,18]. By taking Twitter as an example, a large number of tweets that report traffic events like congestion and accident are posted instantly every day. Many such tweets, like "*Harrison St: accident at Kilbourn Ave, 2:04-4/2/2015*", explicitly give the type of traffic event, time, and location information. Motivated by the rich traffic information available in social media, many recent efforts have been devoted to exploring social media data to facilitate traffic related applications, such as traffic event location identification [16,19], traffic event detection [1,21], as well as traffic congestion estimation [2,3,10]. Chen et al. made the first attempt to estimate urban traffic congestions by relying only on the traffic information collected from Twitter [10]. To improve long-term traffic prediction, He et al. tried to use rich semantic information in online social media [7]. Wang et al. proposed a coupled matrix and tensor factorization model to integrate social media data, road features, and other information to better estimate traffic congestions of a city [2]. However, existing works mainly focus on studying how to utilize social media as the major data source for traffic monitoring. How to use social media data and fuse it with GPS probe data to improve traffic congestion estimation is still not fully explored.

For the first time, this paper incorporates traffic information extracted from Twitter with the sparse and noisy GPS probe data to enhance urban traffic congestion estimation. The challenges of the studied problem are two-fold. Firstly, the traffic information extracted from Twitter can be associated to multi-typed traffic events including congestion, accident, road construction, etc. It is nontrivial to model the potential impacts of the diverse traffic events on traffic congestion. For example, given a tweet that reports a traffic accident, how can we quantitatively measure its impact on traffic congestion? Secondly, it is also difficult to combine the two types of data with totally different data formats seamlessly. A piece of GPS probe reading normally contains the time, speed, heading, and the exact location (longitude, latitude) information of a vehicle; while a tweet that reports a particular traffic event typically will mention the traffic event type, the time, and the road or road segment information. The differences of the two types of datasets on both traffic information and location granularity make the effective combination of them very challenging.

To address the above challenges, we first extensively collect traffic related tweets from both traffic authority Twitter accounts (explain later) and regular

Twitter user accounts, and extract the traffic event, time, and location information by data processing. Through data analysis, we discover that (1) there is a high occurrence correlation between traffic events like accident and traffic congestion, and (2) the data of traffic event related tweets is an important complementary to GPS probe data. Both discoveries indicate that the estimation performance could potentially improved if Twitter data are properly incorporated. To effectively fuse the two types of data, we propose an extended Coupled Hidden Markov Model (E_CHMM). Different from traditional models with the GPS probe observations only [6,8], in this model we consider the GPS probe data and traffic related tweets as two types of observations generated from two different distributions independently. As the exact solution of the E_CHMM model is infeasible for a large network due to the exponential space and time consumption, we utilize a sequential importance sampling method to more efficiently solve the E-step of the EM algorithm. In the M-step, we formulate the original optimization problem decomposable into smaller problems that can be independently optimized. We evaluate our model on the arterial network of downtown Chicago with 1,257 road links whose total length is nearly 700 miles. The result shows that by incorporating Twitter data, about 15 % GPS probes can be reduced to achieve the comparable performance to previous method with all the GPS probes. This research provides us with a promising way to reduce the cost and improve the performance of urban traffic congestion estimation.

2 Preliminary

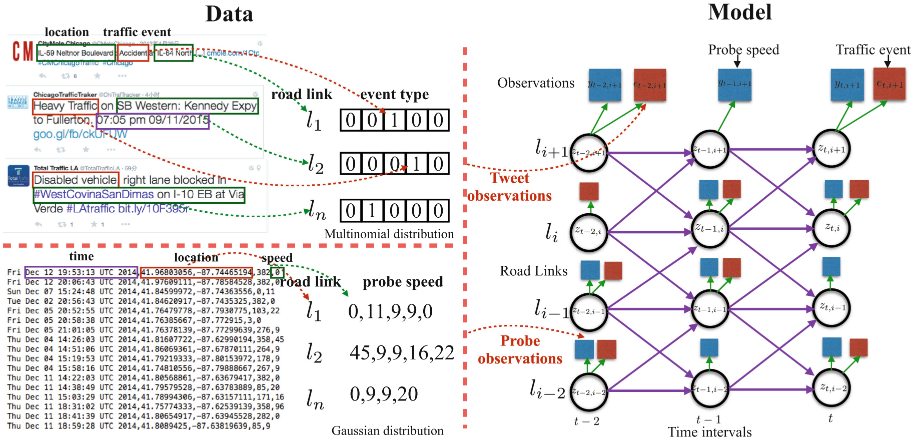
In this section, we will start with some definitions, and introduce the framework of our method. Next we will make some basic assumptions in traffic congestion estimation to facilitate us model the studied problem.

Definition 1. ***A tweet observation of traffic event*** $e_{t,l,i}$. *We represent a tweet observation of traffic event occurring on the road link* l *at time* t *as such a tuple* $e_{t,l,i} = (c, loc, t)$, *where* c *is the traffic event category,* loc *represents the location or road segment of the event, and* t *denotes the time.*

Definition 2. ***A GPS probe observation*** $y_{t,l,i}$. *We represent a GPS probe observation on the road link* l *at time* t *as such a vector* $y_{t,l,i} = (s, lat, lon, head, t)$, *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.*

Definition 3. ***A road link*** l. *We use the intersections to partition an arterial road* R *into several road links* $R = \{l_1, l_2, ...\}$. *Each road link* l *can be represented as such a tuple* $l = (Link\_ID, Start\_Inter, End\_Inter)$, *where* $link\_ID$ *is the ID of the road link,* $Start\_Inter$ *is the start intersection, and* End_Inter *is the end intersection.*

Definition 4. ***Neighbor road links.*** *Two road links* l_1 *and* l_2 *are called neighbor road links if they connect to each other, namely they share an intersection. Particularly, the road link* l *is also considered as a neighbor road link of itself. We denote all the neighbor links of road link* l *as* N_l.



CityMole Chicago @CMoleChicago · 2013年4月22日
IL-59 Neltnor Boulevard : Accident at IL-64 North [...] cmole.com/1Ctc
#CMChicagoTraffic #Chicago

ChicagoTrafficTraker @ChiTrafTracker · 4小时
Heavy Traffic on SB Western: Kennedy Expy to Fullerton. 07:05 pm 09/11/2015 goo.gl/fb/ckUFUW

Total Traffic LA @TotalTrafficLA · 59分
Disabled vehicle, right lane blocked in #WestCovinaSanDimas on I-10 EB at Via Verde #LAtraffic bit.ly/10F395r

```
Fri Dec 12 19:53:13 UTC 2014,41.96803056,-87.74465194,382,0
Fri Dec 12 20:06:43 UTC 2014,41.97609111,-87.78584528,382,0
Sun Dec 07 15:24:48 UTC 2014,41.84599972,-87.74363556,0,11
Tue Dec 02 20:56:43 UTC 2014,41.84620917,-87.7435325,382,0
Fri Dec 05 20:52:55 UTC 2014,41.76479778,-87.7930775,103,22
Fri Dec 05 20:58:38 UTC 2014,41.76385667,-87.772915,3,0
Fri Dec 05 21:01:05 UTC 2014,41.76378139,-87.77299639,276,9
Thu Dec 04 14:26:03 UTC 2014,41.81607722,-87.62990194,358,45
Thu Dec 04 14:51:06 UTC 2014,41.86069361,-87.67870111,264,9
Thu Dec 04 15:19:53 UTC 2014,41.79219333,-87.80153972,178,9
Thu Dec 04 15:58:16 UTC 2014,41.74810556,-87.79888667,267,9
Thu Dec 11 14:22:03 UTC 2014,41.80568861,-87.63679417,382,0
Thu Dec 11 14:38:49 UTC 2014,41.79579528,-87.63783889,85,20
Thu Dec 11 15:03:29 UTC 2014,41.78994306,-87.63157111,171,16
Thu Dec 11 18:31:02 UTC 2014,41.75774333,-87.62539139,358,96
Thu Dec 11 18:41:39 UTC 2014,41.80654917,-87.63945528,282,0
Thu Dec 11 18:59:28 UTC 2014,41.8089425,-87.63819639,85,9
```

Fig. 1. Framework of the proposed model (Color figure online)

Figure 1 shows the framework of our method. It contains two major parts: data collection and processing part, and the model part. There are two types of data sources in our model, traffic related tweets and GPS probe readings. From each traffic related tweet, we first extract the traffic event type, location, and time information, and then map it to the corresponding road link by geocoding. Similarly, we extract the exact location and travel speed information from each GPS probe reading, and then map it to the corresponding road link. For each road link, we assume the occurrence of traffic events on it follows multinomial distribution, and the traveling speed of vehicles in a particular time interval follows Gaussian distribution [6].

We model the spatiotemporal conditional dependencies of arterial traffic using a probabilistic graphical model *Coupled Hidden Markov Model*. A CHMM models a system of multiple interaction processes which are assumed to be a Markov process with unobserved states. In our model, the multiple processes evolving over time are the discrete traffic states of each link in the road network (the circles in the model part of Fig. 1). Since we do not observe the state of each link for all times, we consider them as *hidden*. We can observe the vehicle speed and traffic events from GPS probe and tweets (the blue and red squares in the model part of Fig. 1), and the traffic speed and event on each link are conditioned on its hidden state. In addition, a coupled structure to the HMM specifies the local dependencies between adjacent links of the arterial network. As shown in the model part of Fig. 1, ***the goal of this paper is to more accurately infer the hidden congestion states*** $z_{t,l}$ ***for each road link*** l ***in each time interval*** t ***by utilizing the traffic event observations*** $e_{t,l}$ ***and the probe observations*** $y_{t,l}$.

Following the classical traffic congestion estimation models [6,8], we make the following assumptions for computational tractability.

- *Discrete traffic states:* For each time interval t, the traffic condition on link l is represented by a discrete value s_t^l, which indicates the level of congestion.
- *Conditional independence of link travel speed:* Conditioned on the state s_t^l of a link l, the travel speed distribution on l is independent from all other traffic variables.
- *Conditional independence of traffic events:* Conditioned on the state s_t^l of a link l, the probability of traffic event $e_{t,l,i}$ occurring on link l is independent from all other traffic variables.
- *Conditional independence of state transitions:* Conditioned on the states of link l and its neighbor links in time interval t, the state of link l at time $t+1$ is independent from all other current link states, all past link states, and all past observations.

The second and third assumptions show that the two types of observations are independent to each other and only determined by the current traffic state of the road link. The last assumption implies that the traffic state of each road link is only related to its neighbor links in the last time interval, but independent of the states of the rest road links in earlier time intervals.

3 Twitter Data Collection

In this section we introduce how we collect traffic event information from Twitter. This paper focuses on studying the traffic conditions in Chicago, and we collect traffic event tweets in Chicago from two types of accounts as in [2]: traffic authority Twitter accounts and regular Twitter user accounts.

Traffic Authority Twitter Accounts. Traffic authority Twitter accounts refer to the Twitter accounts that specialize in posting traffic related information. Such accounts are mostly operated by official transportation departments. Tweets posted by these accounts are formal and easy to process, and the exact location and time information are explicitly given such as the tweet "*Heavy Traffic on NB Western: Fullerton to Kennedy Expy. 06:15 pm 02/13/2015*". We identify 10 such Twitter accounts that report real-time traffic information of Chicago: *ChicagoDrives*, *ChiTraTracker*, *roadnowChicago*, *traffic_Chicago*, *IDOT_Illinois*, *WGNtraffic*, *TotalTrafficCHI*, *GeoTrafficChi*, *roadnowil*, and *rosalindrossi*.

Regular Twitter user accounts. We also crawl the tweets posted by the regular users registered in Chicago. In all we target on more than 100,000 such users and crawl more than 32.3 million corresponding tweets. Next, we preprocess the data as follows. (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 can also identify the traffic event category. (2) *Tweet Geocoding.* We then geocode the tweets to the road links by matching their geo-tags and text content. By combing the geo-coordinates of tweets and the direction mentioned in the content, we can geocode the tweets to the road links. For most tweets without geo-tags, we first identify the streets, landmarks, and direction information from the content by using gazetteer, and then geocode them to the road segments.

Note that accurately identifying the locations of traffic events from tweets is itself a challenging task [1,16]. Traffic event location extraction from the short and noisy text is out of the scope of this work. In this paper we only keep the tweets that explicitly give the traffic event type and road segment information. For those with incomplete or obscure location information, we choose to omit them. In all we obtain 245,568 traffic event tweets from April 2014 to December 2014, around 80 % of which are collected from traffic authority accounts. Each tweet reports a traffic event. 163,742 of them are related to slow traffic, 77,454 are related to accident, and 4,372 report other traffic events such as road construction and road closure.

To investigate whether the traffic events reported by Twitter can reflect traffic conditions, we plot the probe speed observations on the road links with a traffic event reported by Twitter and on normal road links in Fig. 2. Each data point in the figure represents a probe speed observation on a road link. Blue data points represent the normal probe observations, while red data points represent probe observations on the road links where traffic congestions or accidents are reported by Twitter. One can see that the average probe speed on the road links with traffic events is lower than that on road links with normal traffic conditions. It implies that the traffic events reported by tweets usually indicate a slower traffic, and thus they can help us better estimate traffic conditions.

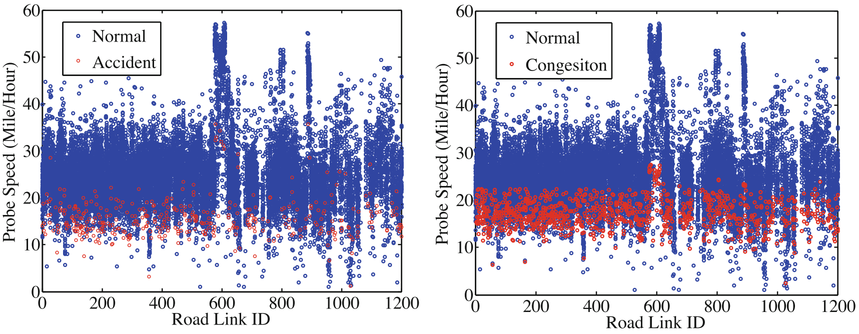


Fig. 2. Probe speed: Normal *vs* Accident (left figure), and Normal *vs* Congestion (right figure) (Color figure online)

4 Extended Coupled Hidden Markov Model: Incorporating Two Types of Observations

Before elaborating the method, we first give some notations and their meanings in Table 1. π_l^s denotes the initial probability of road link l in traffic state s.

A_l is the traffic state transition probability matrix for link l with respect to all its neighbor links. It is a matrix of size $S^{|N_l|} \times S$, where $S^{|N_l|}$ represents the number of all possible states of the neighbors N_l of link l. Based on our assumption, the state of link l in the time interval $t+1$ is only related to the states of its neighbor links N_l in the last time interval t. Hence each element $A_l(R_i, s)$ represents the probability of link l to be in state s in the time interval $t+1$ given that its neighbors N_l are in states $R_i = (r_{i1}, r_{i2}, ...r_{i|N_l|})$ in the time interval t. $g_l^s(\cdot)$ is the probability density function of vehicle speed for link l in state s. We assume it follows Gaussian distribution [6]. $f_l^s(\cdot)$ represents the distribution of traffic event number for link l in state s. We assume it follows Multinomial distribution. P_l^s contains all the parameters of the functions $f_l^s(\cdot)$ and $g_l^s(\cdot)$. $q_{t,l}^{R_i,s}$ is a variable to help estimate the transition probability matrix A_l. We use boldface capital letters to denote the observations or hidden state matrixes on all the road links in all the time intervals. For example, $\mathbf{Y}$ denotes all the GPS probe observations. We use capital letters with subscripts to denote the observations or hidden state vectors in a particular time interval. For example, Y_t denotes the GPS probe observations on all the road links from link l_1 to l_N in the time interval t.

Table 1. Notations and meanings

L	Number of road segment links
T	Number of time intervals
M	Number of traffic event types
S	Number of traffic states
N_l	The set of all the neighbor links of road link l
N_{li}	The *i-th* neighbor based on the lexicographical order of $link\_ID$, $link\_ID \in N_l$
$y_{t,l}$	The set of probe observations for link l in time slot t
$y_{t,l,i}$	One probe observation for link l in t, $y_{t,l,i} \in y_{t,l}$
$e_{t,l}$	The set of traffic event observations for link l in t
$e_{t,l,i}$	One traffic event observation for link l in t, $e_{t,l,i} \in e_{t,l}$
π_l^s	The initial probability that link l begins in state s
A_l	The state transition probability matrix for link l with respect to its neighbors N_l
$g_l^s(\cdot)$	The probability density function of travel speed for link l in traffic state s
$f_l^s(\cdot)$	The distribution function of traffic event for link l in traffic state s
P_l^s	The parameters of the probability density function $g_l^s(\cdot)$ and $f_l^s(\cdot)$
$z_{t,l}^s$	The probability of link l being in traffic state s in t
$q_{t,l}^{R_i,s}$	The probability of link l being in traffic state s for time period t given that its neighboring links N_l are in states $R_i = (r_{i1}, r_{i2}, ...r_{i\|N_l\|})$ in $t-1$

With above notations, we give the complete log likelihood of the observation data and hidden variables. Typically, the log likelihood of the hidden variables and observations of the CHMM can be written out as follows,

$$\begin{aligned} lnP(\mathbf{Y},\mathbf{E},\mathbf{Z}) &= lnP(Z_1)+\sum_{t=2}^{T} lnP(Z_t|Z_{t-1})+\sum_{t=1}^{T} lnP(Y_t,E_t|Z_t) \\ &= lnP(Z_1)+\sum_{t=2}^{T} lnP(Z_t|Z_{t-1})+\sum_{t=1}^{T} lnP(Y_t|Z_t)+\sum_{t=1}^{T} lnP(E_t|Z_t) \end{aligned} \tag{1}$$

The first term of the formula (1) represents the initial probability of traffic states Z_1 for all the road links, the second term is the probability that traffic states Z_{t-1} in time interval $t-1$ transit to the states Z_t in the next time interval t, and the third term is the probability of observations Y_t, E_t conditioned on the traffic states Z_t. Since the GPS probe observations are independent from the traffic event observations, we can further decompose $\sum_{t=1}^{T} lnP(Y_t, E_t|Z_t)$ as shown in the second line of formula (1).

The initial probability of the congestion states in the first time interval is

$$lnP(Z_1)=\sum_{l=1}^{L}\sum_{s=1}^{S} z_{1,l}^{s} ln\pi_l^s \tag{2}$$

The log probability of congestion state transiting from time interval $t-1$ to t can be further represented as follows,

$$lnP(Z_t|Z_{t-1})=\sum_{l=1}^{L}\sum_{s=1}^{S}\sum_{i=1}^{S^{|N_l|}}(\prod_{N_{lj}\in N_l} z_{t-1,N_{lj}}^{r_{ij}} z_{t,l}^{s} lnA_l(R_i,s)) \tag{3}$$

The third summation of formula (3) is over all the possible traffic states $S^{|N_l|}$ of the neighbors N_l, while the subsequent product is over terms on each of its individual neighbor state given the neighbor states $(r_{i1},...,r_{i|N_l|})$.

The probability of probe speed observations Y_t given the congestion states Z_t can be represented as

$$lnP(Y_t|Z_t)=\sum_{l=1}^{L}\sum_{s=1}^{S} z_{t,l}^{s}(\sum_{y_{t,l,i}\in y_{t,l}} ln(g_l^s(y_{t,l,i}))) \tag{4}$$

The probability of traffic event observations E_t given the congestion states Z_t can be represented as

$$lnP(E_t|Z_t)=\sum_{l=1}^{L}\sum_{s=1}^{S} z_{t,l}^{s}(\sum_{e_{t,l,i}\in e_{t,l}} ln(f_l^s(e_{t,l,i}))) \tag{5}$$

4.1 Solution of E_CHMM: EM Algorithm

Given the distribution function parameters P_l^s of observations and the state transition matrix A_l, it is possible to estimate the congestion states of the links based on the observations. Similarly, given the congestion states of the road links, we can estimate the parameters in the model. Motivated by this idea, EM algorithm can be applied to solve E_CHMM.

In the E-step, for road link l we compute the expected state probabilities $z^s_{t,l}$ and the transition probabilities $q^{R_i,s}_{t,l}$ given observations ($y_{t,l}$, $e_{t,l}$), distribution parameters P^s_l, and the state transition probability matrix A_l.

$$z^s_{t,l} \leftarrow E(z^s_{t,l}|y_{t,l}, e_{t,l}, P^s_l, A_l) \tag{6}$$

$$q^{R_i,s}_{t,l} \leftarrow E(q^{r,s}_{t,l}|y_{t,l}, e_{t,l}, P^s_l, A_l) \tag{7}$$

One can see that the traffic state $z^s_{t,l}$ is inferred based on both the GPS probe observation $y_{t,l}$ and the tweet observation $e_{t,l}$. To distinguish the importance of the two types of observations in estimating the traffic state $z^s_{t,l}$, we rewrite formula (7) as follows.

$$z^s_{t,l} \leftarrow \begin{cases} E(z^s_{t,l}|e_{t,l}, P^s_l, A_l) & \text{if } Cardinality(y_{t,l}) = 0 \\ w_{t,l}E(z^s_{t,l}|y_{t,l}, P^s_l, A_l) + (1 - w_{t,l})E(z^s_{t,l}|e_{t,l}, P^s_l, A_l) & \text{otherwise} \end{cases} \tag{8}$$

If only the tweet observation $e_{t,l}$ is available on road link l in time interval t, the congestion state $z^s_{t,l}$ is estimated only based on $e_{t,l}$. Otherwise, $z^s_{t,l}$ is estimated by using both types of observations. $w_{t,l}$ is the confidence of the probe observations. The idea is that if sufficient probe observations are available, we trust more on the traffic state $z^s_{t,l}$ estimated by probe observations. If the probe data are very spare, we trust more on the estimation results with the tweet observations. Here we use a sigmoid function to estimate the importance of the coefficient $w_{t,l} = \frac{1}{1+e^{\theta - Cardinality(y_{t,l})}}$, where θ is a predefined threshold of the probe observation size. More probe observations result in a large $w_{t,l}$, and thus the final estimation result $z^s_{t,l}$ relies more on the probe observations. In this paper we set $\theta = 3$.

In the M-step, we maximize the expected complete log-likelihood, given the probabilities $z^s_{t,l}$ and the transition probabilities $q^{R_i,s}_{t,l}$.

$$\begin{aligned} (P^s_l, A_l, \pi^s_l) &\leftarrow \underset{\mathbf{P},\mathbf{A},\mathbf{\Pi}}{\mathbf{argmax}} lnP(\mathbf{Y}, \mathbf{Z}, \mathbf{E}, \mathbf{P}, \mathbf{A}, \mathbf{\Pi}) \\ subject\ to &\sum_{s=1}^{S} A_l(R_i, s) = 1, A_l(R_i, s) \in [0,1], \forall l, R_i, s; \\ &\sum_{s=1}^{S} \pi^s_l = 1, \pi^s_l \in [0,1], \forall l, s. \end{aligned} \tag{9}$$

5 Parameter Inference

On small networks, it is possible to do exact inference in the CHMM by converting the model to an HMM. However, it is intractable to do exact inference for any reasonable traffic network with the naive solution due to the following reasons. (1) Computation of the forward variable involves S^L additions and N multiplications at each of T time steps; (2) each forward variable requires $8S^L$ bytes of memory to store, and all T of them must be stored; (3) the transition matrix itself is $S^L \times S^L$. Next we will introduce a sequential importance sampling based approach to more efficiently address the computational challenge.

5.1 E-Step: Particle Filtering

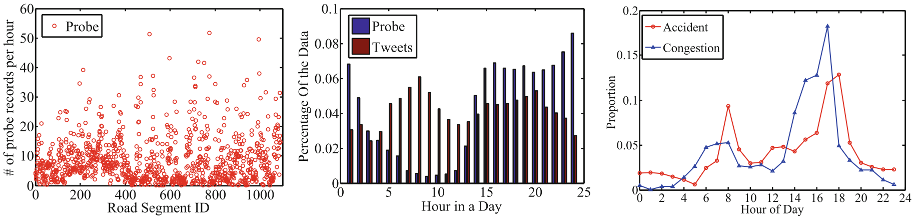
As a popular sequential importance sampling method, particle filtering is widely used to approximately estimate the internal states in dynamical systems such as signal processing and Bayesian statistical inference. Due to the extremely high computational cost of the CHMM, particle filtering is introduced in previous works [9]. In our setting, each particle or sample represents an instantiation of the traffic state evolution on the traffic network. Given the observed probe data and traffic events from tweets, each particle or sample is assigned a weight proportional to the probability of the observations. Using a large number of sampled particles, we can estimate the probabilities of the traffic states of each link in each time interval, and the probabilities of traffic state transition among the neighbor road links in successive time intervals. Details of the algorithm is given in Algorithm 1.

Algorithm 1. Particle Filtering to Estimate Congestion States

```
Input: Number of samples $K$ and time intervals $T$, the state transition matrix $A_l$, the
       parameters of the observation probability function $P_l^s$ for each road link $l$.
Output: The state probability distribution matrix $\mathbf{Z}$, and the transition probability $q_{t,l}^{R_i,s}$
Initialization: randomly sample $K$ samples $\{x_k^0\}_{k=1}^K$;
for $t = 1 : T$ do
    Generate $K$ samples of the state $x_k^t$ based on the sampled states $x_k^{t-1}$ and state
    transition matrix $A_l$: $x_k^t \sim q(x_k|x_k^{t-1})$;
    Compute the weights:
    $w_k^t = p(Y_t, E_t|x_k^t) = p(Y_t|x_k^t)p(E_t|x_k^t)$;
    Normalize the weights:
    $\hat{w}_k^t = \frac{w_k^t}{\sum_{j=1}^K w_j^t}$;
    Resample $K$ random samples $\{\hat{x}_k^t\}_{k=1}^K$ from $\{x_k^t\}_{k=1}^K$ with replacement in proportion to
    the weights $\{\hat{w}_k^t\}_{k=1}^K$;
    Replace the sample set with these new samples, i.e. $\{x_k^t\}_{k=1}^K \leftarrow \{\hat{x}_k^t\}_{k=1}^K$;
    Set the weights to be equal: $\hat{w}_k^t = \frac{1}{N}, k = 1, ..., N$
Estimate the state probability matrix $\mathbf{Z}$ and transition probability $q_{t,l}^{R,s}$ with the $K$ samples
return $\mathbf{Z}$, $q_{t,l}^{R_i,s}$;
```

5.2 M-step: Road Network Decomposition

In the M-step, we update three groups of parameters: the initial congestion state probability π_l^s, the observation distribution function parameters P_l^s, and the transition probability matrix A_l. To update these parameters, the expected



(a) # of probes (b) Hourly distributions (c) Traffic event correlation.

Fig. 3. Data Statistics: (a) Average # of probe readings for each road link in each hour. (b) Hourly distributions of probe readings and tweets on each road segment. (c) Hourly occurrence correlations between traffic accidents and congestions reported by tweets.

complete log-likelihood is maximized given the probability $z^s_{t,l}$ that each link l is in state s at time t and probability $q^{R_i,s}_{t,l}$ of link l to be in state s given that neighbors of link l are in states R_i at time $t-1$. Based on formulas (1)–(5), the expected complete log likelihood is as follows.

$$\begin{aligned}
&lnP(\mathbf{Y},\mathbf{E}|\mathbf{Z},\mathbf{Q},\mathbf{P},\mathbf{A},\boldsymbol{\Pi}) = \\
&\sum_{l=1}^{L}\sum_{s=1}^{S}\sum_{t=1}^{T} z^s_{t,l}\Big(\sum_{y_{t,l,i}\in y_{t,l}} ln(g^s_l(y_{t,l,i})) + \sum_{e_{t,l,i}\in e_{t,l}} ln(f^s_l(e_{t,l,i}))\Big) \\
&+\sum_{l=1}^{L}\sum_{t=2}^{T}\sum_{s=1}^{S}\sum_{i=1}^{S^{|N_l|}} q^{R_i,s}_{t,l} ln(A_l(R_i,s)) + \sum_{l=1}^{N}\sum_{s=1}^{S} z^s_{1,l} ln(\pi_{l,s})
\end{aligned} \tag{10}$$

We can simplify the computation of formula (10) in the following two ways. (1) One can see that formula (10) is comprised of three parts. Different parameters appear in different parts, and thus the three parts can be solved separately. (2) The optimization problem on the entire road network can be further decomposed into $S \times L$ smaller optimization problems with each one associated to a particular congestion state and road link of the network. For example, for the road link l in state s the first part in the right-hand side of formula (10) can be decomposed to such an optimization problem.

$$\max_{P_{l,s}} \sum_{t=1}^{T} z^s_{t,l}\Big(\sum_{y_{t,l,i}\in y_{t,l}} ln(g^s_l(y_{t,l,i})) + \sum_{e_{t,l,i}\in e_{t,l}} ln(f^s_l(e_{t,l,i}))\Big) \tag{11}$$

6 Evaluation

6.1 Experiment Setup

Datasets and analysis. The Twitter data are described in Sect. 3. From each tweet, we extract the road segment, time and traffic event information. We categorize these tweets into three types by keywords matching: congestion, accident and others. We also have more than 2 million GPS probe readings generated by various vehicles on 1,257 arterial road links of downtown Chicago in December 2014. The total length of these road links is nearly 700 miles.

Figure 3 gives the statistics of the two datasets. Figure 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, while for most road segments there are only very limited number of probe data. Figure 3(b) shows the percentages of probe data and traffic related tweets in each hour of day. One can see that most probe data are distributed in the time interval from 14:00pm to 0:00am. Most traffic related tweets are posted in two time intervals from 5:00am to 10:00am and from 15:00pm to 22:00pm. The hourly distributions of the two datasets are not perfectly consistent, which implies the combination of them could provide us with more comprehensive information. Figure 3(c) shows the proportion curves of the traffic accident and congestion reported by tweets in each hour of a day. One can see that the two curves show very similar increasing and decreasing trends, which indicates a strong occurrence correlation. Traffic congestions can cause more accidents, and accidents in turn can make traffic even worse. The high occurrence correlation between accident and congestion implies that other types of traffic events captured from tweets may potentially help us better estimate traffic congestions.

Ground Truth. Obtaining the ground truth itself is a challenge problem. The manually annotated ground truth is very expensive, and thus is not feasible for a large transportation network. Previous studies show that the bus probe data in urban areas can provide a good approximation of the real traffic conditions [5,22]. Thus we use the traffic conditions reported by Chicago Transit Authority (CTA) as the ground truth. The traffic conditions are estimated based on more than 5 million GPS traces generated by more than 2,000 CTA public passenger buses from 11/25/2014 to 12/30/2014[1]. CTA defines 5-state traffic conditions in Chicago: heavy congestion, medium-heavy congestion, medium, light, and flow conditions, with the corresponding traffic speeds as 0–10, 10–15, 15–20, 20–25, and over 25 miles per hour. We assign the 5 congestion states with values 1.0, 0.8, 0.6, 0.4, and 0.2 respectively. As the real time GPS traces for some links are sparse, we also consider the historical average traffic speed for each road link in the last 3 years. Given a time interval t and a road link l, the traffic speed can be estimated as $speed_{t,l} = w \sum_{i=1}^{n} \frac{speed_{t,l,i}}{n} + (1-w)speed^h_{t,l}$, where $speed_{t,l,i}$ is the ith real time probe speed record, $speed^h_{t,l}$ is the historical speed, and w is a weight. For simplicity, we consider a road segment is in congestion if the average speed is lower than 15 mph.

Competitive Methods. We compare E_CHMM with the following baselines.

- **CHMM with probe observations (P_CHMM) [6].** Herring et al. proposed a CHMM model to estimate arterial traffic conditions with probe data.

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

We use it as a baseline to evaluate whether incorporating the Twitter data can improve the performance.

- **CTCE model** [2]**.** CTCE is a recently proposed traffic congestion estimation model with social media as the primary data source. Instead of utilizing CHMM, CTCE models the traffic information on the road segments as matrices and tensors and apply matrix factorization technique to address the estimation task.
- **CHMM with tweet observations (T_CHMM).** In this model, only the tweet observations are available. We use this baseline to evaluate the performance of the CHMM model with the tweet observations only.
- **Linear combination of the two types of data (LC_CHMM).** We use two CHMMs with each one associated with one type of data to estimate the traffic conditions separately. Assuming the estimation results of the two models are $\mathbf{Z}_1$ and $\mathbf{Z}_2$, the final estimation is the linear combination of the two results, $\mathbf{Z} = \alpha \mathbf{Z_1} + (1-\alpha)\mathbf{Z_2}$.

Evaluation Metrics. We use the following metrics to evaluate the performance of the proposed model: *accuracy*, *precision@k*, and *Root Square Error (RMSE)*. We use *accuracy* to evaluate the estimation performance on all the road segments in all the time intervals. Normally, in a particular time interval only a small number of road segments are in congestion. Thus to better evaluate whether the proposed model can give good estimations on the road segments that are very likely to occur congestion, we also use *precision@k* as a metric. We first rank the congestion probabilities $z^s_{t,l}$ for all the road segments in all the time intervals. Then we only consider the road segments with the *top-k* congestion probabilities are in congestion. To further evaluate the performance of the model on the above mentioned 5-state traffic conditions, we use the *Root Mean Square Error (RMSE)* as the evaluation metric: $RMSE = \sqrt{\frac{\sum_{t,l}(z_{t,l} - \hat{z}_{t,l})^2}{L*T}}$, where $z_{t,l}$ is the estimated traffic state of link l in time interval t, and $\hat{z}_{t,l}$ is the ground truth.

6.2 Quantitive Evaluation Results

Evaluation with *precision@k.* Table 2 shows the average *precision@k* of different methods over various k. As the traffic conditions on weekdays and weekends can be quiet different, we present the results by weekday and weekend separately. We run the algorithm and calculate the *precision@k* on each day, and then average the results. The best results are highlighted in bold type. One can see that E_CHMM performs best among all the methods. LC_CHMM model is inferior to E_CHMM, but better than other methods. It is no surprise that T_CHMM presents the worst performance among all the methods. One can infer that the traffic event tweets are too sparse for the T_CHMM model to get an accurate estimation. P_CHMM can achieve comparable performance with CTCE, but both methods are inferior to LC_CHMM and E_CHMM. One can also see that in general the average *precision@k* on weekday is higher than

Table 2. Average Precision @k of different methods

Average Precision @k on weekday									
	top-10	top-20	top-30	top-50	top-100	top-150	top-200	top-250	top-300
P_CHMM	0.870	0.850	0.845	0.832	0.812	0.792	0.773	0.744	0.732
T_CHMM	0.690	0.665	0.624	0.613	0.585	0.532	0.473	0.464	0.452
LC_CHMM	0.890	0.850	0.852	0.842	0.832	0.817	0.792	0.784	0.775
CTCE	0.870	0.860	0.853	0.840	0.824	0.816	0.718	0.705	0.712
E_CHMM	**0.920**	**0.900**	**0.894**	**0.887**	**0.864**	**0.826**	**0.810**	**0.795**	**0.786**
Average Precision @k on weekend									
	top-10	top-20	top-30	top-50	top-100	top-150	top-200	top-250	top-300
P_CHMM	0.860	0.850	0.843	0.822	0.816	0.766	0.752	0.745	0.722
T_CHMM	0.660	0.650	0.612	0.625	0.570	0.464	0.453	0.415	0.425
LC_CHMM	0.870	0.850	0.845	0.825	0.820	0.812	0.805	0.785	0.768
CTCE	0.850	0.834	0.820	0.820	0.754	0.715	0.678	0.654	0.644
E_CHMM	**0.910**	**0.900**	**0.868**	**0.852**	**0.844**	**0.820**	**0.812**	**0.794**	**0.783**

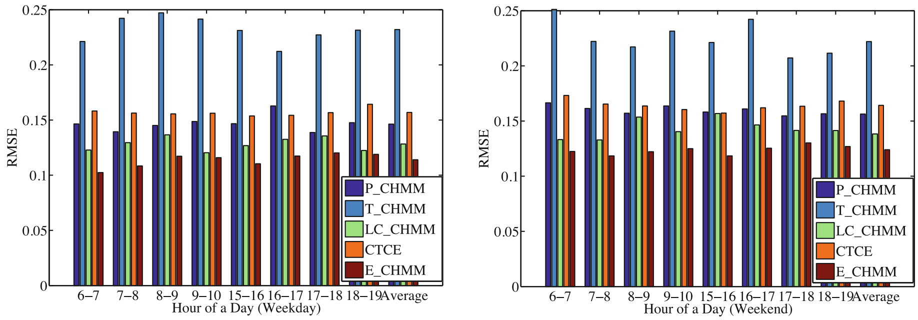


Fig. 4. RMSE of the four methods in rush hours

that on weekend. This is because most people travel on weekday more regularly than on weekend.

Performance evaluation in rush hours. People concern more on the traffic conditions in rush hours of a day. Thus we also evaluate the performance of different models in rush hours. Figure 4 shows the experiment results in the rush hours of 6:00–10:00 and 15:00–17:00 on weekday and on weekend, respectively. One can see that the RMSE of E_CHMM is mostly lower than all the baselines. The performance of T_CHMM is the worst among all the methods, which is consistent with the previous experiment results. LC_CHMM is consistently better than P_CHMM and CTCE, which means incorporating traffic event information from tweets does help us better estimate traffic conditions. However, LC_CHMM is inferior to the proposed E_CHMM. Thus we can conclude that E_CHMM is more efficient to fuse the two types of observations. By comparing the results on weekday and weekend, one can see that on average the RMSE of various methods on weekday is larger than that on weekend. This finding also verifies that traffic conditions on weekend is harder to estimate than on weekday.

Performance evaluation with various proportions of probe data. To examine how the probe data size affects the estimation performance, we display the estimation accuracy curves of the methods E_CHMM, LC_CHMM, and P_CHMM with different probe data sizes in Fig. 5. It shows that E_CHMM is consistently better than the two baselines. When the probe data are extremely sparse, say only 20 % probe data are available, the accuracy of P_CHMM is only 0.22 while E_CHMM is 0.42, which shows a significant improvement. However, with the increase of the probe data size, the difference between E_CHMM and the other two methods becomes smaller. This is probably because the information overlapping between the two datasets becomes larger when more probe date are available. When the probe data are sufficient, the traffic conditions inferred by traffic event tweets can also be captured by the probe readings. The LC_CHHM is better than P_CHHM but inferior to E_CHMM. One can see that E_CHMM only needs around 85 % probe data to achieve a comparable accuracy to P_CHMM with the whole probe data.

Scalability Analysis. As the optimization problem of the EM algorithm can be decomposed into many smaller optimization problems, we can easily solve it in parallel on multiple machines. Figure 6 shows the running time of solving the optimization problems by distributing them into multiple machines on the traffic data of a day on the studied road links. It shows a linearly decreasing trend of the running time with the increase of machine number. One can see that it needs more than 12 min for only one machine, but the time decreases to about 2 min if we distribute these independent smaller optimization problems on 5 machines. It demonstrates the proposed algorithm is very scalable to handle a large road network with thousands of road links.

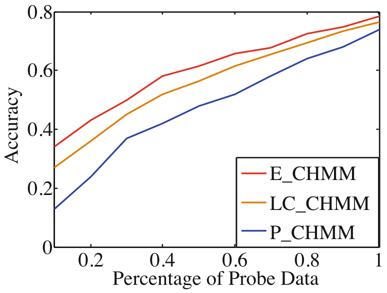


Fig. 5. Estimation accuracy *vs* probe data size

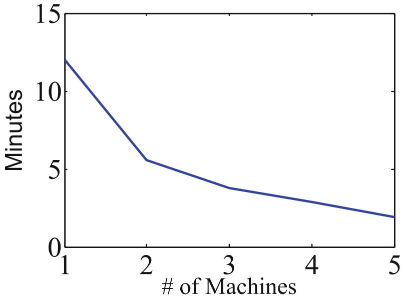


Fig. 6. Running time *vs* # of machines

7 Related Work

Traditionally, traffic monitoring and estimation mainly rely on various road sensors, and can be roughly categorized into traffic modeling on individual roads [11,13,14] and on a road network [20]. Helbing employed 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 [11]. 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 [14]. Porikli and Li proposed a Gaussian Mixture Hidden Markov Models to detect traffic condition with the MPEG video data [13]. 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 [6,15,20]. Such models mainly utilized the Floating Car Data (FCD) or probe data 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 [6]. Fabritiis et al. studied the problem of using FCD data based on traces of GPS positions to predict the traffic on Italian motorway network [15].

Recently, exploring traffic related information from social media like Twitter to detect traffic events or monitor traffic conditions has been a hot research topic [1,2,10,12]. Most previous works focused on investigating either how to extract and visualize the traffic event information from tweets [1,12] or how to locate the traffic events mentioned in the tweets [16,19]. As traffic event data are usually sparse and imbalanced, imbalanced learning techniques are usually explored [17]. The work in [10] 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 [2]. However, as the probe data are not explored, the performance are usually not desirable due to very sparse and noisy Twitter data [10].

8 Conclusion

In this paper, we studied the novel problem of incorporating social media semantics to enhance traffic congestion estimation. Motivated by the increasing availability of traffic information in social media, we first extensively collected traffic related tweets from Twitter. Then we extended the classical Coupled Hidden Markov Model to effectively combine the tweet observations and probe observations. To solve the proposed model, we also introduced an efficient EM algorithm to infer the parameters. Evaluation on the arterial network of Chicago showed the proposed model can both effectively combine the two types of observations and efficiently address the computational challenge.

Acknowledgments. This paper is supported by Beijing Advanced Innovation Center for Imaging Technology (No.BAICIT-2016001), the NSFC (Grant No.: 61272083, 61370126, 61202239), the NSF through grants III-1526499, the Natural Science Foundation of Hebei Province (No. F2014210068), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

1. Liu, M.L., Fu, K.Q., Lu, C.T., Chen, G.S., Wang, H.Q.: A search and summary application for traffic events detection based on twitter data. In: ACM SIGSPATIAL GIS (2014)
2. Wang, S.Z., He, L.F., Stenneth, L., Yu, P.S., Li, Z.J.: Citywide traffic congestion estimation with social media. In: ACM SIGSPATIAL GIS (2015)
3. Wang, S.Z., He, L.F., Stenneth, L., Yu, P.S., Li, Z.J., Huang, Z.Q.: Estimating urban traffic congestions with multi-sourced data. In: IEEE MDM (2016)
4. Ozkurt, C., Camci, F.: Automatic traffic density estimation and vehicle classification for traffic surveillance systems using neural networks. Math. Comput. Appl. **14**(3), 187–196 (2009)
5. Wang, Y., Zhu, Y.M., He, Z.C.: Challenges and Opportunities in Exploiting Large-Scale GPS Probe Data. Technical report, HPL-2011-109 (2011)
6. Herring, R., Hofleitner, A., Abbeel, P., Bayen, A.: Estimating arterial traffic conditions using sparse probe data. In: IEEE ITSC (2010)
7. He, J.R., Shen, W., Divakaruni, P., Wynter, L., Lawrence, R.: Improving traffic prediction with tweet semantics. In: IJCAI (2013)
8. Hofleitner, A., Herring, R., Abbeel, P., Bayen, A.: Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network. IEEE Trans. Intell. Transp. Syst. **13**(4), 1679–1693 (2012)
9. Cheng, P., Qiu, Z.J., Ran, B.: Particle filter based traffic state estimation using cell phone network data. In: IEEE ITSC (2006)
10. Chen, P.T., Chen, F., Qian, Z.: Road traffic congestion monitoring in social media with hinge-loss Markov random fields. In: ICDM (2014)
11. Helbing, D.: Traffic and related self-driven many-particle systems. Rev. Mod. Phys. **73**(4), 1067–1141 (2001)
12. Endarnoto, S.K., Pradipta, S., Nugroho, A.S., Purnama, J.: Traffic condition information extraction and visualization from social media twitter for android mobile application. In: ICEEI (2011)
13. Porikli, F., Li, X.K.: Traffic congestion estimation using HMM models without vehicle tracking. In: IEEE IV (2004)
14. Muoz L., Sun X.T., Horowitz R., Alvarez L.: Traffic density estimation with the cell transmission model. In: American Control Conference (2003)
15. Fabritiis, C.D., Ragona, R., Valenti, G.: Traffic estimation and prediction based on real time floating car data. In: ITCS (2008)
16. Ribeiro Jr., S.S., Davis Jr.,, C.A., Oliveira, D.R.R., Meira Jr., W., Gonalves, T.S., Pappa, G.L.: Traffic observatory: a system to detect and locate traffic events and conditions using twitter. In: ACM SIGSPATIAL LBSN (2012)
17. Wang, S.Z., Li, Z.J., Chao, W.H., Cao, Q.H.: Applying adaptive over-sampling technique based on data sparsity and cost-sensitive SVM to imbalanced learning. In: IJCNN (2012)
18. Bregman, S.: Uses of Social Media in Public Transportation. Transportation Research Board (2012)

19. Daly, E.M., Lecue, F., Bicer, V.: Westland row why so slow? Fusing social media and linked data sources for understanding real-time traffic conditions. In: IUI (2013)
20. Shang, J.B., Zheng, Y., Tong, W.Z., Chang, E., Yu, Y.: Inferring gas consumption and pollution emission of vehicles throughout a city. In: ACM SIGKDD (2014)
21. Sayyadi, H.: Event detection and tracking in social streams. In: ICWSM (2009)
22. Carli, R., Dotoli, M., Epicoco, N., Angelico, B., Vinciullo, A.: Automated evaluation of urban traffic congestion using bus as a probe. In: IEEE CASE (2015)