# Forecasting the Subway Passenger Flow under Event Occurrences with Social Media

Ming Ni, Qing He, *Member, IEEE*, and Jing Gao, *Member, IEEE*[1]

***Abstract*— Subway passenger flow prediction is strategically important in metro transit system management. The prediction under event occurrences turns into a very challenging task. In this paper, we adopt a new kind of data source -- social media to tackle this challenge. We develop a systematic approach to examine social media activities and sense event occurrences. Our initial analysis demonstrates that there exists a moderate positive correlation between passenger flow and the rates of social media posts. This finding motivates us to develop a novel approach for improved flow forecast. We first develop a hashtag based event detection algorithm. Further, we propose a parametric and convex optimization based approach, called Optimization and Prediction with hybrid Loss function (OPL), to fuse the linear regression and the results of seasonal autoregressive integrated moving average (SARIMA) model jointly. The OPL hybrid model takes advantage of the unique strengths of linear correlation in social media features and SARIMA model in time series prediction. Experiments on events nearby a subway station show that OPL reports the best forecasting performance compared with other state-of-the-art techniques. In addition, an ensemble model is developed to leverage the weighted results from OPL and support vector machine regression together. As a result, the prediction accuracy and robustness further increases.**

***Index Terms*—Social media; Event identification; Subway passenger flow prediction; Social sensing; Transit ridership**

## I. Introduction

Passenger flow prediction is critical for planning, management and operations of public transit systems [1]. The output from the prediction can benefit transit network design, route scheduling, and station crowd regulation operations [2]. The majority of the previous studies lie in forecasting day-to-day recurrent passenger flow [3][4][5][6]. However, when it comes to non-recurrent events (e.g. sporting game, concert, running race, etc.), because of its irregularity and inconsistency, passenger flow prediction turns into a very challenging task. Very limited methods have been proposed in the literature.

For solving this problem, instead of revising existing methods, we intend to leverage a new kind of data -- social media. User-generated contents on social media strengthen linkage and interactions between users, meanwhile provide a large amount of information. The vast information is able to capture the public attention, which is one of the common traits of events.

However, social media data is much difficult to process compared with traditional relational data. There still exist several major challenges in handling social media data, which is unstructured, noisy, gigantic, and contains a variety of information. Take Twitter data for example. Only in 2014, we have collected over 29.7 million geo-tagged posts bounded in the New York City Area. At individual post level, a fundamental question of data mining arises: what it is talking about, and what event information it contains. Thus the first challenge (C1), within a transportation context, is how to identify transportation-related events that each post refers to. An individual geo-tagged post is able to provide social activity analysis at spatial-temporal aggregated level. Transportation authorities can leverage such information to identify hot spots and further indicate passenger flows in near future for public gathering. Therefore, the second challenge (C2) is how to develop a method to coordinate social media for forecasting passenger flow, especially under event occurrences.

This paper aims to address challenges (C1) and (C2). More specifically, under event occurrences, we intend to extract event information from geo-tagged social media data, and leverage both historical transit data and real-time social media data to forecast future passenger flow at subway stations. The following questions will be investigated: (i) Can social media be used to identify public events in real life? (ii) How to build the prediction model by the features extracted from social media? To the best of our knowledge, there has not been considerable published research on the effects of passenger flow prediction with social media.

The paper has the following structure. Section II summarizes related works about recent popular transportation prediction techniques and the uses of social media in transport applications. An overview of the data, including subway passenger flow and social media, is given in Section 0. Section IV describes the setup of event detection approach. Section V

[1] This work was supported by the U.S. Department of Transportation through Transportation Informatics (TransInfo) University Transportation Center at University at Buffalo, The State University of New York (Corresponding author: Q. He).

Ming Ni is with Department of Industrial and Systems Engineering, University at Buffalo, the State University of New York, Buffalo, NY 14260 (e-mail: mingni@buffalo.edu)

Qing He is with Department of Civil, Structural, and Environmental Engineering and Industrial and Systems Engineering at University at Buffalo, the State University of New York, Buffalo, NY 14260 (e-mail: qinghe@buffalo.edu)

Jing Gao is with Department of Computer Science and Engineering at the University at Buffalo, the State University of New York, Buffalo, NY 14260 (e-mail: jing@buffalo.edu)

presents a detailed analysis of the relationship between event passenger flow and social media. Section VI presents the technical details of prediction modeling and experiments on real-world datasets. Finally, Section VII provides concluding remarks.

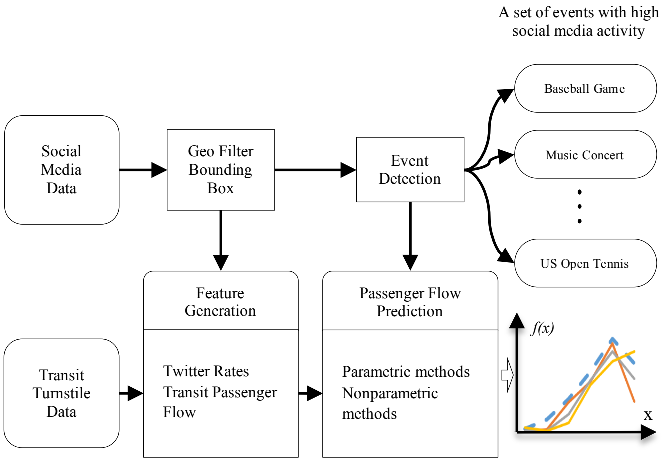


Fig. 1. System architecture for passenger flow prediction from social media

## II. Related Works

There is a vast literature in short-term transportation forecasting [7]. Generally, there are two groups of approaches receiving wide attention, namely, parametric and non-parametric techniques.

The common parametric techniques include autoregressive integrated moving average model (ARIMA), exponential smoothing [8], and historical average [9]. Especially, ARIMA has been fully developed for various transportation prediction purposes, including traffic occupancy [10], travel time [11] and traffic flow [12]. Previous research [13][14] shows ARIMA performs well for stationary and non-event time series. With the rise of data mining and science, non-parametric techniques also have been widely adopted recently. Neural network [15][16], support vector machine for regression (SVR) [17] and k-nearest neighbor [18] were used to build the traffic volume prediction model for the time-series data.

The passenger flow prediction belongs to the subcategory of short-term transportation prediction. Some researchers adopted both kinds of prediction techniques to forecast the passenger flow for railway [15][19][20], bus stop [21][22], and subway stations. Specifically for passenger flow prediction at subway stations, there are different prediction levels, respectively, at whole transit lines [3][4], at one station with passenger transfer flow [5], and at one station with entrance and transfer flow [6]. All of them obtained a desirable predict result of typical commuting volumes. However, none of them adds consideration of atypical conditions.

Recently, more and more attempts have been made to implement The Internet and social media analysis in the domain of transportation [23][24][25]. A huge group of people in the online community generates a tremendous amount of content. Chaniotaks and Antoniou [26] proposed a generic methodological framework for collecting and analyzing the data from social media. And other researchers took advantages of using crowdsourcing these resources to capture the incoming non-recurrent events [27], to explain the causes of transport overcrowding [28], to investigate intelligent transportation systems services [29], and to utilize deep learning approach [30][31]. Studies are trying to exploit this area mainly fall into two applications, traffic detection, and traffic prediction, with supervised learning techniques.

In the application of traffic detection, Wanichayapong et al. [32] used synthetic analysis to classify the traffic incident information into spatial categories from the social media data. Schulz et al. [33] extracted features from part-of-speech tagging and words in Twitter posts and developed classifiers to detect car accident occurrences. They applied spatial and temporal filtering to locate the accidents. Daly [34] built a system called Dublin's Semantic Traffic Annotator and Reasoner to use natural language processing techniques to analyze social media contents in order to capture real-time traffic conditions. Mai and Hranac [35] explored the time and location of the related Twitter posts after traffic incidents occurred. They found that the majority of tweets are posted within 5 hours and 25 miles for freeway incidents. Gal-Tzur et al. [36] used the Twitter messages sent from transportation authorities to develop classifiers to identify the posts related to transportation information. Moreover, they presented a keyword-based hierarchical schema to categorize these posts. Chen et al. [37] tried to detect traffic congestion and location solely based on social media data by using topic modeling and hinge-loss Markov random fields. D'Andrea et al. [38] utilized Twitter data and developed a support vector machine model to recognize useful keywords from tweets and detect traffic events in the area of highway road network. Kumar et al. [39] incorporated social media to detect road hazards by sentiment and language analysis. Most recently, Zhang et al. [40] studied and revealed the characteristics of traffic flow surge near the tweet concentration, which is defined as a cluster of keywords for traffic related events. Further, Zhang and He proposed analytical models to detect on-site traffic accidents [41] and decode people's travel behavior with geo-mobility clustering [42].

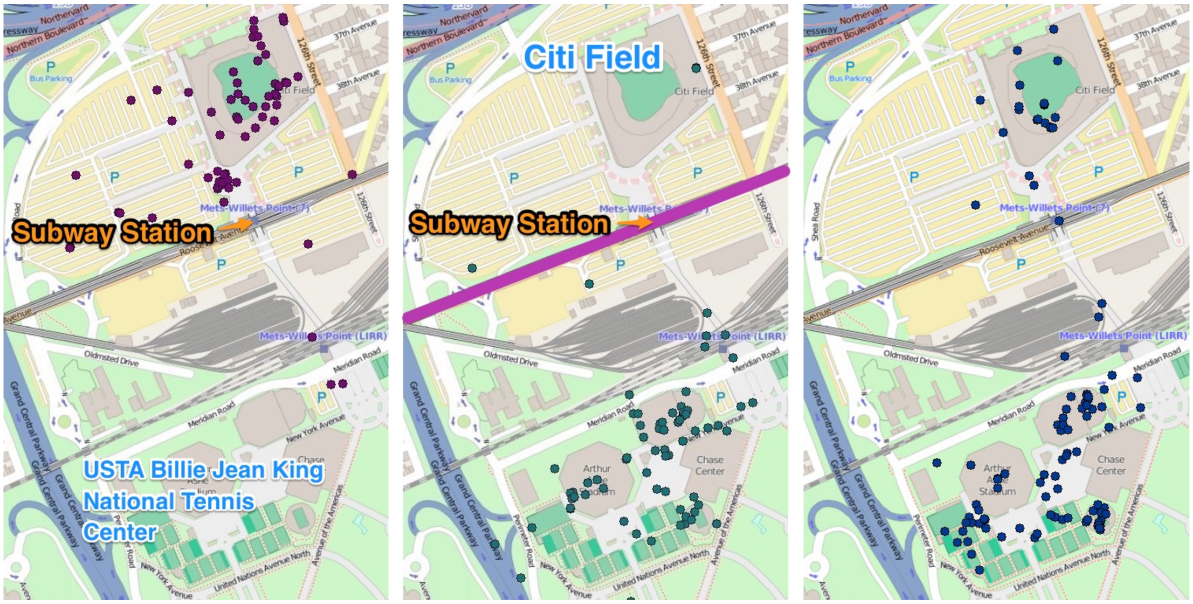
For traffic prediction, He et al. [43] proposed a long-term traffic prediction models with social media features for a freeway network in San Francisco Bay area. They found that there exists a negative correlation between social activity on the web and traffic activity on the roads. Ni et al. [44] tried to forecast freeway traffic flows under special event conditions by taking into account information derived from social media. Lin et al. [45] applied linear regression models for predicting the impact of inclement weather on freeway speed with the help of social media.

For subway and transit, Collins et al. [46] used sentiment analysis of transit riders' short messages on social media to measure their satisfaction about transit. They found that the social media posts with the sharp increased negative sentiment indicated some transit incidents, like fire and delays.

Above studies show that there is great potential to use social media to locate right information for transportation applications. However, none of the previous studies explores

TABLE I
SAMPLE TWEETS BEFORE EVENTS

| **Event** | | | **Sample Twitter Message** | |
|---|---|---|---|---|
| *Type* | *Start Time* | *Details* | *Create at* | *Text content* |
| Baseball game | 2014-05-14<br>19:10 | Mets vs. Yankee | 2014-05-14<br>18:22:22 | Checked in CITI field for the yankees vs mets game w yankees mets |
| Tennis games | 2014-08-25<br>19:00 | US Open 1st round | 2014-08-25<br>17:49:46 | I'm at 2014 usopen tennis championships in flushing ny |
| Baseball game + Tennis games | 2014-08-28<br>19:00 (T) 19:10 (B) | US Open 2nd round & Mets vs. Braves | 2014-08-28<br>18:29:10 | love this place billy jean king national tennis centre us open |



(a) Baseball game (b) Tennis games (c) Baseball game + Tennis games

Fig. 2. Geographic distribution of tweets two hours before the events

the effectiveness of using social media for passenger flow prediction in public metro transit systems.

## III. DATASET

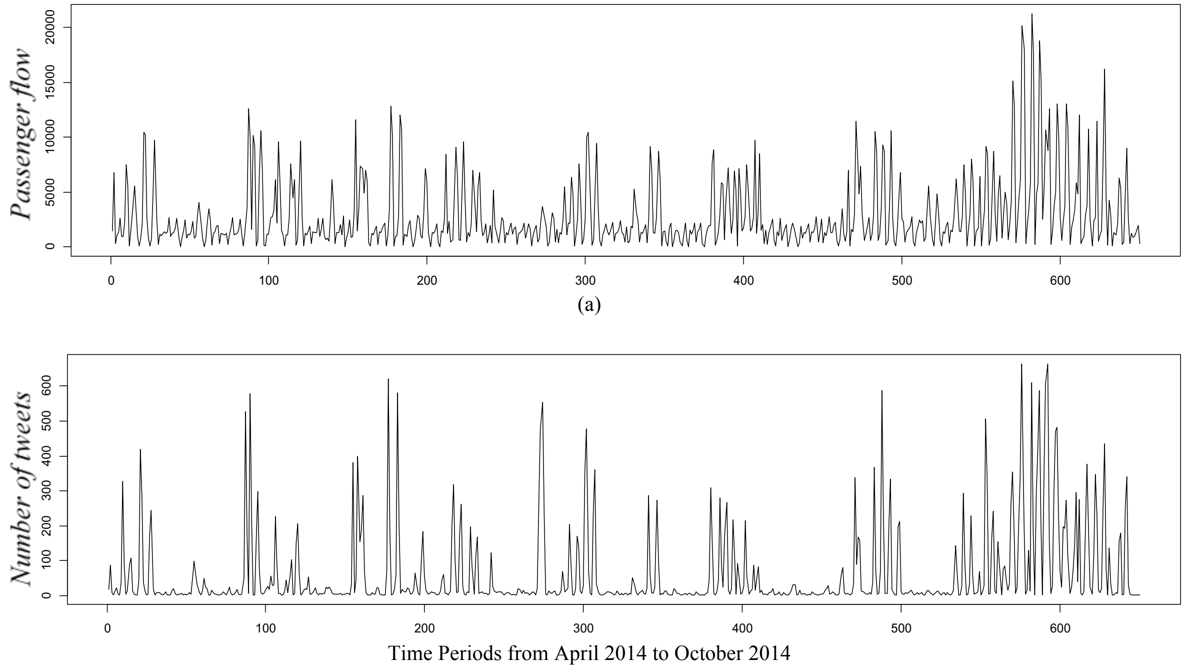
This study expands the successful applications of social media data to predict passenger volume at a subway station. We focus our study on subway station "Mets – Willets Point" on Line 7 in New York City. The station is selected based on two main reasons. First, "Mets – Willets Point" is adjacent to not one but two stadiums, Citi Field and USTA Billie Jean King National Tennis Center (NTC). Citi Field is the home stadium of New York Mets Baseball team, and NTC hosts the annual US Open grand-slam tennis tournament. Second, the sports events always obtain public attention. From our observation, there is a substantial volume of social media posts referring to the events.

We collected the turnstile usage at "Mets – Willets Point" subway station from Metropolitan Transportation Authority (MTA) [47]. In order to cover various types of events, the time range is set from April 2014 to October 2014, in which various events occur nearby.

Turnstile devices record passengers passing each turnstile for either entry or exit, and it reports the aggregated number every four hours. In this paper, we aggregate both entry and exit flows as total passenger flow, which is of transit agency's interest.

We collected Twitter data, known as tweets, as social media data. Twitter message is an online text post limited to 140 characters by Twitter users. Tweets were collected in the same temporal window through Twitter Streaming API with geo-location filter [48]. The spatial bounding box was set to cover only the subway station and two stadiums. Because of the location filter, besides text content, username and timestamp, each tweet contains its geographic coordinate. Inside the post, users are able to prefix by a # symbol with words, which is called the Twitter hashtag. A hashtag provides unique tagging convention to facilitate tweets with certain topics, contexts or events. The aforementioned information from Twitter messages defines a tweet in this paper.

Fig. 2 shows the locations of tweets sent two hours before different types of events start. As it can be seen, tweets were mostly sent from the stadium in which the event was held. Moreover, different events correspond to different social media activities, and to various levels of public attention. From social



(a)

(b)

Fig. 3. Comparisons of passenger flow (a) and number of tweets (b)

media data perspective, the characteristics of tweets, like time stamps, geolocations, text content, quantity ratios, etc., lead to such differences. Our objective is to find ways to measure these differences in social media data and leverage them into prediction models to forecast subway passenger flow.

## IV. Hashtag-based Event Identification

The events held in stadiums were well attended. The attendance not only brings a high volume of passenger flows but also activities on Twitter, shown in Fig. 3. As one can see, event scenarios generate large spikes of social media activity and passenger flow at the same time.

We assume that the complete schedule of all events is unknown for transit operators. The subway station Mets-Willets Point could coordinate transit passengers for two major sports events, US Open Tennis Championships and Major League Baseball for New York Mets. The former was held late August and early September over a two-week period, and the latter was held from April to September 2014. However, after initial examinations, we found that there were other events like concerts and speeches being held nearby as well. Therefore, we need to identify the events by social media data.

Instead of detecting the exact topic of the events [49][50][51], we would like to examine tweets within the area and probe whether there will exist events involving high social activities. To correctly identify the events, rather than using the complex machinery of latent variable topic models (e.g. Latent Dirichlet Allocation [52]), we employ the Twitter hashtags to measure social media activities and provide the context for them [53].

Hashtag extraction is the first step of the proposed event detection algorithm. We denote $t$ as one of the time intervals, with $t = 1, \dots, T$, where $T$ is the total number of four-hour intervals. $HL_t$ is the list of hashtags during $t$. $HL_t = \{H_{t1}, \dots, H_{tj}, \dots, H_{tJ_t}\}$, where $H_{tj}$ is the $j^{th}$ hashtag and $J_t$ is the total number of hashtags labeled by Twitter users during $t$.

Furthermore, let $\boldsymbol{M}^H \in \mathbb{R}^{T \times S}$ denote the hashtag matrix, where $S$ is the number of hashtags. Its element $M^H_{t,s}$ corresponds to the occurrence of the $s^{th}$ hashtag in the $t^{th}$ time interval. Since hashtag matrix contains all the $T$ time intervals, $S \geq \max_{t \in T} J_t$. In the hashtag matrix, all the hashtags over time intervals merge into the columns. Various words and phrases depict different aspects of social activities. In sum, the column names of hashtag matrix are the hashtags, the rows stand for the time intervals, and each entry in the matrix corresponds to the frequency of the hashtag.

The summary of the notations present as follows:

- $t$ is the index of time intervals, $t = 1, \dots, T$.
- $p$ is the index of the tweet.
- $s$ is the index of the hashtag.
- $J_t$ is the total number of hashtags labeled by Twitter users during $t$.
- $\boldsymbol{M}^H \in \mathbb{R}^{T \times S}$ denote the hashtag matrix, where $S$ is the total number of hashtags.
- $HL_t \in \boldsymbol{M}^H$ is the list of hashtags during $t$, which is essentially one row of $\boldsymbol{M}^H$.
- $H_{tj} \in HL_t$ is the $j^{th}$ hashtag in the list $HL_t$.
- $OC_t$ is the occurrence of each element $H_{tj}$ in $HL_t$ .for time interval $t$
- $TW_{p,s}$ is the occurrence of $s^{th}$ hashtag in $HL$ of $p^{th}$ tweet

Below are the steps of event detection by hashtags.

**Algorithm 1**: Hashtag-based Event Identification

```
Input: Tweets within the area
Output: Hashtag matrix $\boldsymbol{M}^H \in \mathbb{R}^{T\times S}$
  1. Hashtags extraction
     $HL_t = \{H_{t1}, \dots, H_{tj}, \dots, H_{tJ_t}\} \;\; \forall t \in [1, T]$
  2. Lexical analysis
     $HL \equiv \bigcup_{t=1}^{T} HL_t$
     Remove stop words, punctuation and duplicated strings from HL
  3. Label all collected tweets by hashtag
     $TW_{p,s}$ ≡ calculate the occurrence of $s^{th}$ word in $HL$ of $p^{th}$ tweet
     for $p^{th}$ tweet $p = 1$ to $P$ do
        for $s^{th}$ word in $HL$
           Append the $TW_{p,s}$ as a new column for $p^{th}$ tweet
  4. Build hashtag matrix ($\boldsymbol{M}^H \in \mathbb{R}^{T\times S}$)
     Each row of $\boldsymbol{M}^H$ represents the vector of $HL$
     $OC_t \in \mathbb{R}^S$ ≡ the occurrence of each element in $HL_t$ for time interval $t$
     for $t = 1$ to $T$ do
        $OC_t = \sum_{p\in T} \sum_{s\in S} TW_{p,s}$
        $\boldsymbol{M}_t^H = OC_t$
  5. Peak detection
     for $t = 1$ to $T$ do
        Rank $OC_t$ based on $\sum_s OC_{t,s}$ from the largest to the smallest.
        for $s = 1$ to $S$ do
           Sort $OC_{t,s}$ from largest to smallest.
```

Since there could be different hashtags for different time intervals, it is trivial to see that $\boldsymbol{M}^H$ is originally a sparse column-wise matrix, and each column corresponds to the frequency of hashtag in each time interval. By concatenating hashtag list $HL_t$ over *t*, it converts $\boldsymbol{M}^H$ to a full storage matrix in order to sort the hashtag matrix row by row for peak detection afterward.

Moreover, instead of directly utilizing the occurrence of hashtags labeled by Twitter users, we extract the string vector of hashtags and use it to label the text content of each tweet. It will facilitate the approach to capture those tweets about a similar topic without hashtags.

TABLE II
SAMPLE EVENTS AND THEIR TOP HASHTAGS

| Date | Hour | No. of EH | Top Hashtags | | |
|---|---|---|---|---|---|
| 3/31 | 17 to 21 | 65 | mets | openingday | ny |
| 4/5 | 13 to 17 | 306 | mets | reds | baseball |
| 4/9 | 17 to 21 | 34 | amaluna | cirquedusoleil | citifield |
| 5/14 | 17 to 21 | 710 | mets | yankees | subwayseries |
| 5/31 | 9 to 13 | 85 | happiest5k | queens | ny |
| 6/7 | 17 to 21 | 75 | digifestnyc | nyc | selfie |
| 8/25 | 17 to 21 | 437 | usopen | tennis | usopen2014 |
| 8/31 | 13 to 17 | 609 | usopen | mets | tennis |

Finally, we implement peak detection to extract most frequently occurring hashtags as event hashtags, representing social media activities with context. In TABLE II, the top 3 frequently occurring hashtags are presented. Moreover, we use the sum of all occurring hashtags for each time interval to measure the social media activity. High-rank number of hashtags indicates that the corresponding time interval is under event occurrence.

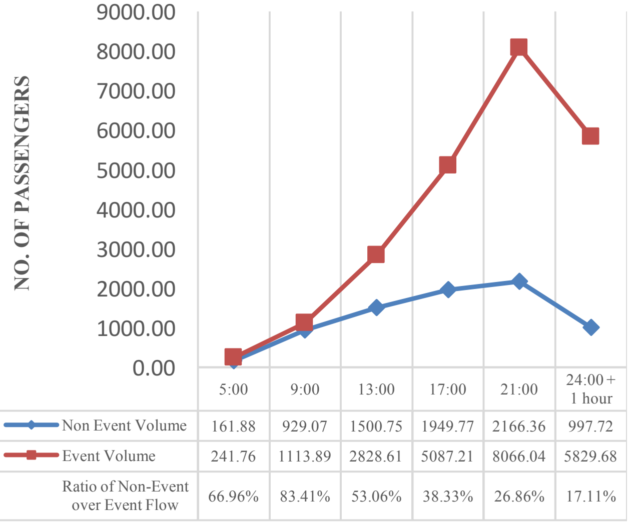
TABLE II shows the various detected events, including US Open, baseball games, music shows, running races, etc. In order to justify the method, we compare the detection results with the true home game schedule of New York Mets, which had long time range and a decent number of games. There were 81 game days during April 2014 to October 2014 for New York Mets. After eliminating the days with missing Twitter data, 65 game days remain. Since the objective of the event detection is to sense the positive events instead of non-events, we evaluate the identification results with *precision*, *recall* and F$_1$ score.

The proposed method achieves good performance in identifying those baseball events, i.e., the *precision* is 98.27%, *recall* 87.69% and F$_1$ score 0.9268.

Note that there are two reasons to use event hashtags instead of the quantity of tweets directly. First, there is a chance that high volume of tweets does not necessarily indicate event and attendance. In our observation, a conversation between users, commercial promotions or information dissemination could also generate a high quantity of tweets. The proposed hashtag-based method is able to diminish the effects of these unrelated tweets. Second, the top event hashtags can describe what the event is about, though the hashtags might not be formal English words. It can be seen in ***Error! Reference source not found.***TABLE II, different kind of events and baseball teams can be easily recognized by the top event hashtags.

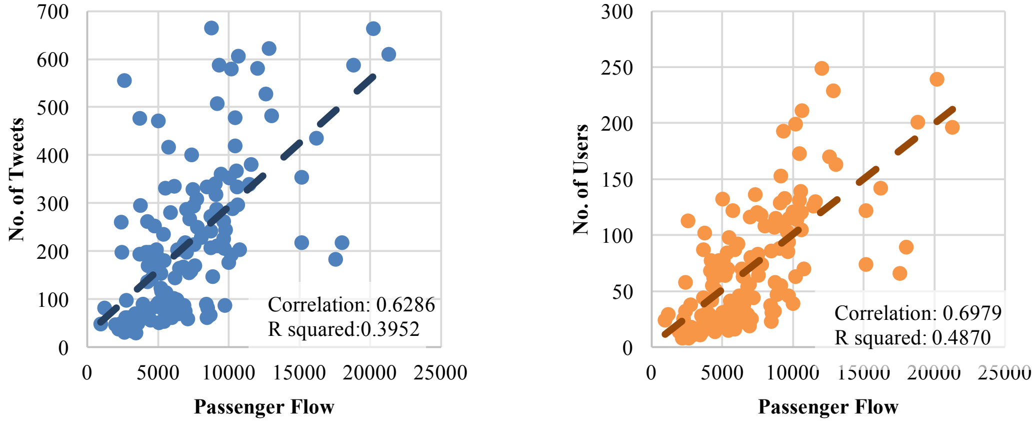
## V. EVENTS CHARACTERISTICS

Different events in stadiums bring different size of audience to the sites, in which the passenger flow at the subway station varies accordingly.



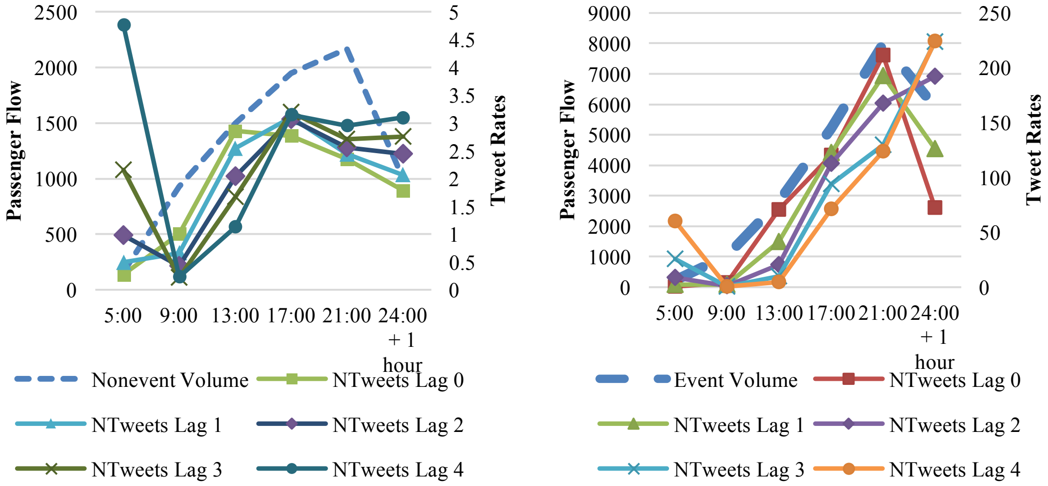
| | 5:00 | 9:00 | 13:00 | 17:00 | 21:00 | 24:00+ 1 hour |
|---|---|---|---|---|---|---|
| Non Event Volume | 161.88 | 929.07 | 1500.75 | 1949.77 | 2166.36 | 997.72 |
| Event Volume | 241.76 | 1113.89 | 2828.61 | 5087.21 | 8066.04 | 5829.68 |
| Ratio of Non-Event over Event Flow | 66.96% | 83.41% | 53.06% | 38.33% | 26.86% | 17.11% |

Fig. 4. Average event/nonevent daily passenger flow at Mets-Willies Point station



(a) Number of Tweets V.S. Passenger flow (b) Number of Users V.S. Passenger flow

Fig. 5. The correlation between tweet rates and passenger flow under events



(a) Nonevent (b) Event

Fig. 6. Average Passenger flow V.S. Average Tweet Rates at Citi Field Station

As shown in Fig. 4, there are huge differences between event and ordinary transit traffic in quantity, more importantly in variation. This difference inevitably leads to the difficulty of transit prediction by traditional time series models (e.g. ARIMA).

On the other hand, in Fig. 5 we plotted the number of tweets against passenger flow under event occurrences in (a), and the number of Twitter users against passenger flow in (b). As one can see, a linear trend is observed between tweet counts on passenger flow. The correlation coefficient is above 0.62 and adjusted $R^2$ value is above 0.39. The $R^2$ values indicate that the number of users is a more robust predictor. We reasonably believe that there exists a moderate positive correlation between tweet counts and event passenger flows. This result gives us the confidence to explore further the prediction modeling of social media on the event passenger flow.

Note that our study is restrained to the extent that the geo-tagged tweet is available. For some of the time periods, the amount of tweets is very small despite the time of day. In this case, event identification measures social media activities and automatically excludes these time periods from the correlation study and the following analysis.

## VI. Prediction Modeling

In this section, we intend to investigate whether or not the content of social media will assist in forecasting event passenger flow. The first step is to identify the best time lags for the prediction models.

To measure the tweets quantifiable, we define two types of feature as tweets rates from social media data:

- *NTweets(t)*: Number of event-related tweets at time step *t*.
- *NUsers(t)*: Number of unique tweet users at time step *t*.

Because the record time interval of transit passenger flow is four hours, we also aggregate the tweets data in four-hour intervals. If the predicted passenger flow is at time *t*, we shift tweet rates to earlier hours: *t-1,t-2,…t-L,* since prediction requires features ahead of passenger flow time. Based on the positive correlation of tweet rates and passenger flow in Fig. 6, we construct a linear regression (LR) model, where passenger flow is the dependent variable, and tweet rates over different hours are independent variables.

The highest predictive correlation is achieved when the tweet rates are calculated based on one hour prior to event time range. We obtain an adjusted $R^2$ value of 0.616 in lag one-hour case. For comparison, the $R^2$ values in lag zero and two-hour cases are, respectively, 0.488 and 0.512. Also, shown in Fig. 6, one can see that the curve of tweets rates with one hour lag fits best to the curve of event passenger flow, whereas for non-event passenger flow there are no obvious patterns between tweets rate and passenger flow. Based on such analysis, we will include tweet rates with one-hour lag into the base prediction model in the following analysis.

Next, we implement cross validation to compare the results of LR model and two popular prediction models: average prediction (AVG) and seasonal autoregressive integrated moving average (SARIMA). We generate an experiment with 100 runs of datasets from the event detection result, and each run takes inputs by randomly splitting the entire dataset into training (70%) and test (30%) sets.

The prediction performance is evaluated by two metrics, namely Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE).

In our experiment with 100 runs, the LR model with tweet rates improves the MAPE by 33.08% comparing with SARIMA (See Fig. 7 for details). Notice that such good performance is achieved by the LR with two variables only. However, the LR model does not capture the relation between time steps, since the passenger flow data are time series in nature.

We conduct a comparison of $R^2$ values between two models: 1) the Tweets-based LR model and 2) the historical-flow-based SARMIA model. The experiment obtains adjusted $R^2$ value of 0.616 for the LR, 0.400 for the SARMIA, and 0.696 for combined features of both. As one can see, around 60% of the event passenger flow variance can be explained by the number of tweets variation. And around 40% of the variance comes from historical time-series flow data, which includes a large portion of day-to-day recurrent passenger flow and a small portion of the non-recurrent event flow. The combination of these two methods shows better $R^2$ value since the LR provides event-related features while the SARIMA presents the features related to time series and routine flow.

Inspired by the above experiment with two modeling methods, we propose a convex optimization based approach, called Optimization and Prediction with hybrid Loss function (OPL), to fuse the LR model and the SARIMA model in the objective function jointly. The OPL model aims to take advantage of unique strengths of line regression in social media features and SARIMA model in time series prediction.

The hypothesis of the proposed model is a parametric linear model, defined as:

$$h_w(x) = 1 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \qquad x_0 = 1$$

Where $x_i$ is $i^{\text{th}}$ feature and its corresponding coefficient is $w_i$. In total, the experiment runs for $m$=100 times. Each entry of the experiment is one of the four-hour intervals from the event detection result. Following our experiment design, we randomly split the $m$ runs into training $m_{train}$ (70%) and test $m_{test}$ (30%). The two tweet rates, *NTweets* and *NUsers*, with one-hour lag act as features in the model.

We construct the total loss function as:

$$J(w,\hat{y}) = \textstyle\sum_j^{m_{train}} (y^{(j)} - h_w(x^{(j)}))^2 + \alpha \cdot \sum_j^{m_{test}} \left(\hat{y}^{(j)} - h_w(x^{(j)})\right)^2 + \beta \sum_j^{m_{test}} \left(\hat{y}^{(j)} - y^{*(j)}\right)^2 \quad (1)$$

The idea behind the loss function is to combine the modeling of the predictions on both training and test data as well as the predictions from time series model. Equation (1) contains three main parts. The first component is the sum of least square for the training set, which is the same as linear regression. The second component incorporates the prediction part directly into the loss function in order to minimize the square error from test data. In addition, to fuse the results of SARIMA, we manage to add the sum of least square between OPL predicted $\hat{y}^{(j)}$ and SARIMA predicted $y^{*(j)}$ into Equation (1) as the third component. $y^{*(j)}$ plays the role of regularization to leverage the whole loss function. Since OPL only includes two independent variables, in the trail experiments, it shows that it is not necessary to equip L1 regularization to prevent overfitting. In sum, OPL adopts the moderately large correlated social media features, and incorporates the prediction results from conventional time series model.

To minimize Equation (1), we first vectorize all variables and coefficients:

$$W \in \mathbb{R}^n \qquad Y \in \mathbb{R}^{m_{train}}$$
$$X^{train} \in \mathbb{R}^{m_{train} \times n} \qquad \hat{Y} \in \mathbb{R}^{m_{test}}$$
$$X^{test} \in \mathbb{R}^{m_{test} \times n} \qquad Y^* \in \mathbb{R}^{m_{test}}$$

Then, the loss function is transformed into:

$$J(W,\hat{Y}) = tr(Y - X^{train} \times W^T) \times \left(Y - X^{train} \times W^T\right)^T) + \alpha \cdot tr(\hat{Y} - X^{test} \times W^T) \times \left(\hat{Y} - X^{test} \times W^T\right)^T) + \beta \cdot tr((\hat{Y} - Y^*) \times \left(\hat{Y} - Y^*\right)^T)$$

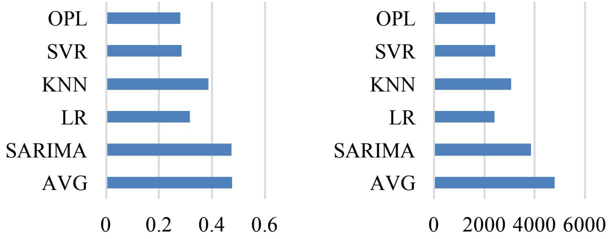
Take partial derivative of the above equation with respect to $W$ and $\hat{Y}$, respectively and we get:

$$\nabla_W J(W,\hat{Y}) = [\left(X^{train}\right)^T \times X^{train} + \alpha \cdot (X^{test})^T \times X^{test}] \times W^T - \alpha \cdot (X^{test})^T \times \hat{Y}^T - (X^{train})^T \times Y^T = 0 \quad (2)$$

$$\nabla_{\hat{Y}} J(W,\hat{Y}) = \alpha \cdot X^{test} \times W^T - (\alpha + \beta) \cdot \hat{Y}^T + \beta \cdot Y^{*T} = 0 \quad (3)$$

Then, we use the gradient descent method to solve Equations (2) and (3) to find a local minimum of $\hat{Y}$. Given Equation (1), gradient descent starts with an initial set of $(W,\hat{Y})$ and iteratively moves toward a set of values that minimize the function. Each iteration takes a step in the negative direction of the function gradient. Because the Equation (1) is convex, the result of OPL shall be the global optimal values.

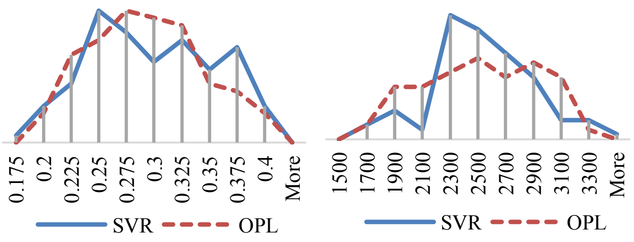
In order to benchmark our proposed method against existing popular prediction approaches, we introduce two nonparametric methods, including SVR and k-nearest neighbors (KNN). The prediction process utilizes cross-validation as well.



(a) MAPE (b) RMSE

Fig. 7 Performance metrics of the prediction models

Fig. 7 illustrates that the OPL yields better prediction accuracy than other methods. Compared with the LR, the OPL improves MAPE by 11.4%. Also, one can see that the SVR presents desirable prediction performance as well. The SVR and the OPL have different characteristics. The SVR is a nonparametric technique that considers tweet rates only. The OPL is a parametric method and incorporates the prediction results from conventional time series model. Further, a detailed comparison is conducted by another 100 randomly generated runs.



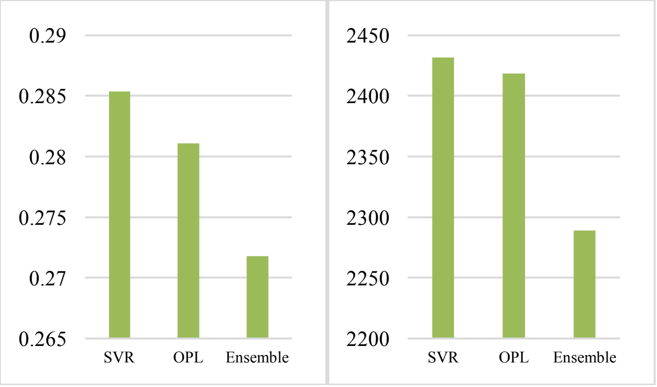
(a) MAPE (b) RMSE

Fig. 8. The distributions of test errors to compare the SVR and OPL

Fig. 8 depicts the distributions of test errors for both SVR and OPL. While either method performs relatively well on its own, it shows the distributions are heterogeneous for both metrics, MAPE and RMSE. The heterogeneity of error distributions encourages us to combine the merits from both techniques. Inspired by the aggregation approach proposed by [54], we implement stacking -- an ensemble learning approach to merge the prediction results of the SVR and OPL.

$$\hat{Y} = P(X^{train}|OPL) \cdot \underset{\hat{Y}}{\operatorname{argmin}} J(W, \hat{Y}|OPL) + P(X^{train}|SVR) \cdot \underset{\hat{Y}}{\operatorname{argmin}} J(W, \hat{Y}|SVR) \quad (4)$$

We estimate $\hat{Y}$ by Equation (4). The weighted probabilities come from normalized root mean square error of training data. The output averages the argument of the minimum for both SVR and OPL.



(a) MAPE (b) RMSE

Fig. 9. Improvement from ensemble learning from the OPL and SVR

As one can see from Fig. 9, the ensemble approach yields better prediction accuracy than either OPL or SVR. It is worth mentioning that the improvement over the conventional SARIMA is more than 40%. Notice that tweet features are obtained from no-cost and real-time social media data. The results indicate the promising value of using social media for passenger flow prediction under event conditions.

## VII. Conclusions

In this paper, we have addressed two important questions, in brief, whether social media data is able to signify public gathering events, and what techniques can be used to model the passenger flow prediction by the features extracted from social media.

First, we exploit social media to detect various events with hashtags. In order to capture events precisely, the hashtags from the Twitter users have been analyzed, tuned, adapted and applied with lexical processing techniques and peak detection. Our approach achieves good performance with precision 98.27% and recall 87.69% for the baseball games. It is a simple but efficient method to capture the events related to public gathering with high social media activity.

Second, we propose a convex optimization model called Optimization and Prediction with hybrid Loss function (OPL) to fuse the least squares of linear regression and the prediction results of SARIMA in the same objective function. The OPL hybrid model aims to take advantage of the unique strengths of line regression in social media features and SARIMA model in time series prediction. Among several popular prediction methods, OPL shows the best results in terms of MAPE and RMSE. In addition, by comparing the distribution of prediction errors of OPL with SVR, which is a popular nonparametric and nonlinear method, it is found that their performance shows heterogeneous error patterns. Therefore, an ensemble model is developed to leverage the weighted results from OPL and SVR jointly. As a result, the prediction accuracy and robustness further increases.

Overall, social media data show the capability in passenger flow prediction under event conditions. Social media offers a cost-effective way to obtain real-time traveler related data, and fills the gap between day-to-day passenger flow volume and abruptly changing non-recurrent event volume. The positive correlation between passenger flow and social media activity plays a significant role as transit demand indicator in the public transit system.

In future, one could further explore the minimum percentage of social media use in an event that leads to a respectable accuracy, and how such minimum can be estimated in order to compute a trust index for the regression result.

## References

[1] M.-C. Chen and Y. Wei, "Exploring time variants for short-term passenger flow," *J. Transp. Geogr.*, vol. 19, no. 4, pp. 488–498, Jul. 2011.

[2] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González, "Spatiotemporal Patterns of Urban Human Mobility," *J. Stat. Phys.*, vol. 151, no. 1–2, pp. 304–318, Dec. 2012.

[3] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transp. Res. Part C Emerg. Technol.*, vol. 21, no. 1, pp. 148–162, Apr. 2012.

[4] B. Leng, J. Zeng, Z. Xiong, W. Lv, and Y. Wan, "Probability Tree Based Passenger Flow Prediction and Its Application to the Beijing Subway System," *Front Comput Sci*, vol. 7, no. 2, pp. 195–203, Apr. 2013.

[5] Y. Sun, G. Zhang, and H. Yin, "Passenger Flow Prediction of Subway Transfer Stations Based on Nonparametric Regression Model," *Discrete Dyn. Nat. Soc.*, vol. 2014, p. e397154, Apr. 2014.

[6] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, Oct. 2015.

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

[8] B. Williams, P. Durvasula, and D. Brown, "Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1644, pp. 132–141, Jan. 1998.

[9] A. G. Hobeika and C. K. Kim, "Traffic-flow-prediction Systems Based on Upstream Traffic," in *Vehicle Navigation and Information Systems Conference, 1994. Proceedings., 1994*, 1994, pp. 345–350.

[10] M. S. Ahmed and A. R. Cook, "Analysis of Freeway Traffic Time-series Data by Using Box-Jenkins Techniques," *Transp. Res. Rec.*, no. 722, pp. 1–9, 1979.

[11] X. Zhang and J. A. Rice, "Short-term Travel Time Prediction," *Transp. Res. Part C Emerg. Technol.*, vol. 11, no. 3–4, pp. 187–210, Jun. 2003.

[12] B. Williams, "Multivariate Vehicular Traffic Flow Prediction: Evaluation of ARIMAX Modeling," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1776, pp. 194–200, Jan. 2001.

[13] "Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results," *J. Transp. Eng.*, vol. 129, no. 6, pp. 664–672, 2003.

[14] S. Lee and D. Fambro, "Application of Subset Autoregressive Integrated Moving Average Model for Short-Term Freeway Traffic Volume Forecasting," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1678, pp. 179–188, Jan. 1999.

[15] T.-H. Tsai, C.-K. Lee, and C.-H. Wei, "Neural Network Based Temporal Feature Models for Short-term Railway Passenger Demand Forecasting," *Expert Syst. Appl.*, vol. 36, no. 2, Part 2, pp. 3728–3736, Mar. 2009.

[16] R. Yasdi, "Prediction of Road Traffic using a Neural Network Approach," *Neural Comput. Appl.*, vol. 8, no. 2, pp. 135–142, May 1999.

[17] C.-H. Wu, J.-M. Ho, and D. T. Lee, "Travel-time prediction with support vector regression," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 276–281, 2004.

[18] F. Guo, R. Krishnan, and J. Polak, "A computationally efficient two-stage method for short-term traffic prediction on urban roads," *Transp. Plan. Technol.*, vol. 36, no. 1, pp. 62–75, Feb. 2013.

[19] W. Gong, "ARMA-GRNN for passenger demand forecasting," in *2010 Sixth International Conference on Natural Computation (ICNC)*, 2010, vol. 3, pp. 1577–1581.

[20] X. Jiang, L. Zhang, and X. (Michael) Chen, "Short-term forecasting of high-speed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in China," *Transp. Res. Part C Emerg. Technol.*, vol. 44, pp. 110–127, Jul. 2014.

[21] C.-H. Zhang, R. Song, and Y. Sun, "Kalman Filter-Based Short-Term Passenger Flow Forecasting on Bus Stop," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 11, no. 4, p. 154, 2011.

[22] M. Gong, X. Fei, Z. Wang, and Y. Qiu, "Sequential Framework for Short-Term Passenger Flow Prediction at Bus Stop," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2417, pp. 58–66, Dec. 2014.

[23] F. Y. Wang, "Scanning the Issue and Beyond: Real-Time Social Transportation with Online Social Signals," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 3, pp. 909–914, Jun. 2014.

[24] F. Y. Wang, "Scanning the Issue and Beyond: Transportation Games for Social Transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1061–1069, Jun. 2015.

[25] X. Zheng, W. Chen, P. Wang, D. Shen, S. Chen, X. Wang, Q. Zhang, and L. Yang, "Big Data for Social Transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 620–630, Mar. 2016.

[26] E. Chaniotakis and C. Antoniou, "Use of Geotagged Social Media in Urban Settings: Empirical Evidence on Its Potential from Twitter," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC)*, 2015, pp. 214–219.

[27] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Using Data From the Web to Predict Public Transport Arrivals Under Special Events Scenarios," *J. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 273–288, Jul. 2015.

[28] F. C. Pereira, F. Rodrigues, E. Polisciuc, and M. Ben-Akiva, "Why so many people? Explaining Nonhabitual Transport Overcrowding With Internet Data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1370–1379, Jun. 2015.

[29] F. Y. Wang, J. J. Zhang, X. Zheng, X. Wang, Y. Yuan, X. Dai, J. Zhang, and L. Yang, "Where does AlphaGo go: from church-turing thesis to AlphaGo thesis and beyond," *IEEECAA J. Autom. Sin.*, vol. 3, no. 2, pp. 113–120, Apr. 2016.

[30] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic Flow Prediction With Big Data: A Deep Learning Approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[31] X. Wang, X. Zheng, Q. Zhang, T. Wang, and D. Shen, "Crowdsourcing in ITS: The State of the Work and the Networking," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1596–1605, Jun. 2016.

[32] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in *2011 11th International Conference on ITS Telecommunications (ITST)*, 2011, pp. 107–112.

[33] A. Schulz, P. Ristoski, and H. Paulheim, "I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs," in *The Semantic Web: ESWC 2013 Satellite Events*, P. Cimiano, M. Fernández, V. Lopez, S. Schlobach, and J. Völker, Eds. Springer Berlin Heidelberg, 2013, pp. 22–33.

[34] F. L. Elizabeth Daly, "Westland Row Why So Slow? Fusing Social Media and Linked Data Sources for Understanding Real-Time Traffic Conditions," 2013.

[35] E. Mai and R. Hranac, "Twitter Interactions as a Data Source for Transportation Incidents," presented at the Transportation Research Board 92nd Annual Meeting, 2013.

[36] A. Gal-Tzur, S. M. Grant-Muller, T. Kuflik, E. Minkov, S. Nocera, and I. Shoor, "The potential of social media in delivering transport policy goals," *Transp. Policy*, vol. 32, pp. 115–123, Mar. 2014.

[37] P.-T. Chen, F. Chen, and Z. Qian, "Road Traffic Congestion Monitoring in Social Media with Hinge-Loss Markov Random Fields," in *2014 IEEE International Conference on Data Mining (ICDM)*, 2014, pp. 80–89.

[38] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-Time Detection of Traffic From Twitter Stream Analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. PP, no. 99, pp. 1–15, 2015.

[39] A. Kumar, M. Jiang, and Y. Fang, "Where Not to Go?: Detecting Road Hazards Using Twitter," in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, New York, NY, USA, 2014, pp. 1223–1226.

[40] Z. Zhang, M. Ni, Q. He, J. Gao, J. Gou, and X. Li, "An Exploratory Study on the Correlation between Twitter Concentration and Traffic Surge," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2553, 2016.

[41] Z. Zhang and Q. He, "On-site Traffic Accident Detection with Both Social Media and Traffic Data," *9th Trienn. Symp. Transp. Anal. TRISTAN IX*, 2016.

[42] Z. Zhang and Q. He, "Exploring Travel Behavior with Social Media: An Empirical Study of Abnormal Movements Using High Resolution Tweet Trajectory Data," *Submitted to. Transp. Res. Part C Emerg. Technol.*, 2016.

[43] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, "Improving Traffic Prediction with Tweet Semantics," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, China, 2013, pp. 1387–1393.

[44] M. Ni, Q. He, and J. Gao, "Using social media to predict traffic flow under special event conditions," in *The 93rd Annual Meeting of Transportation Research Board*, 2014.

[45] L. Lin, M. Ni, Q. He, J. Gao, and A. W. Sadek, "Modeling the Impacts of Inclement Weather on Freeway Traffic Speed," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2482, pp. 82–89, 2015.

[46] C. Collins, S. Hasan, and S. Ukkusuri, "A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data," *J. Public Transp.*, vol. 16, no. 2, Jun. 2013.

[47] "Metropolitan Transportation Authority Date Feed," *mta.info*. [Online]. Available: http://web.mta.info/developers/developer-data-terms.html. [Accessed: 13-Jul-2015].

[48] "Twitter Streaming APIs," *Twitter Developers*. [Online]. Available: https://dev.twitter.com/streaming/overview. [Accessed: 13-Jul-2015].

[49] D. Rampage, S. Dumais, and D. Liebling, "Characterizing Microblogs with Topic Models," *Proc. Fourth Int. AAAI Conf. Weblogs Soc. Media*, 2010.

[50] W. Weerkamp and M. Rijke, "Credibility-inspired Ranking for Blog Post Retrieval," *Inf Retr*, vol. 15, no. 3–4, pp. 243–277, Jun. 2012.

[51] M. Cordeiro, "Twitter Event Detection: Combining Wavelet Analysis and Topic Inference Summarization," in *Doctoral Symposium on Informatics Engineering, DSIE*, 2012, vol. 8, pp. 11–16.

[52] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J Mach Learn Res*, vol. 3, pp. 993–1022, Mar. 2003.

[53] P. Giridhar, M. T. Amin, T. Abdelzaher, L. M. Kaplan, J. George, and R. Ganti, "ClariSense: Clarifying sensor anomalies using social network feeds," in *2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2014, pp. 395–400.

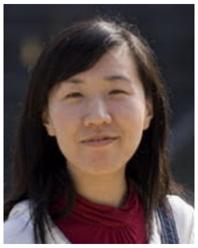
[54] M.-C. Tan, S. C. Wong, J.-M. Xu, Z.-R. Guan, and P. Zhang, "An Aggregation Approach to Short-Term Traffic Flow Prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 60–69, Mar. 2009.

**Ming Ni** received B.S. degree in quality and reliability engineering from College of Reliability and Systems Engineering, Beijing University of Aeronautics and Astronautics, Beijing, China, in July 2011, and M.S. degree in industrial and systems engineering from University at Buffalo, in May 2013. Since 2013, he has been working towards the Ph.D. degree in industrial and systems engineering at University at Buffalo.

His research interests include supply chain management and logistics, and data mining for social media data.

Dr. **Qing He** (S'10-M'13**)** received his B.S. and M.S. in Electrical Engineering at Southwest Jiaotong University, and received his Ph.D. degree in Systems and Industrial Engineering from University of Arizona in 2010. From 2010 to 2012, he worked as a postdoctoral researcher in IBM T. J. Watson Research Center.

He is currently the Stephen Still Assistant Professor in Transportation Engineering and Logistics, affiliated with both Civil Engineering and Industrial Engineering at University at Buffalo, The State University of New York since 2012. His research interests include traffic signal control and freeway operations, social media and transportation, railway predictive maintenance, transportation data analytics, and supply chain management and logistics. Dr. He is a member of Transportation Research Board Standing Committee in Freeway Operations. He won IBM Faculty Partnership Award in 2012 and 2014, respectively.



Dr. **Jing Gao** (S'07-M'12) is currently an Assistant Professor in the Department of Computer Science at the University at Buffalo (UB), State University of New York. She received her PhD from Computer Science Department, University of Illinois at Urbana Champaign in 2011, and subsequently joined UB in 2012. She is broadly interested in data and information analysis with a focus on information integration, crowdsourcing, ensemble methods, mining data streams, transfer learning and anomaly detection. More information about her research can be found at: http://www.cse.buffalo.edu/~jing.

`