City-Scale Social Event Detection and Evaluation with Taxi Traces

WANGSHENG ZHANG, GUANDE QI, and GANG PAN, Zhejiang University
HUA LU, Aalborg University
SHIJIAN LI and ZHAOHUI WU, Zhejiang University

A social event is an occurrence that involves lots of people and is accompanied by an obvious rise in human flow. Analysis of social events has real-world importance because events bring about impacts on many aspects of city life. Traditionally, detection and impact measurement of social events rely on social investigation, which involves considerable human effort. Recently, by analyzing messages in social networks, researchers can also detect and evaluate country-scale events. Nevertheless, the analysis of city-scale events has not been explored. In this article, we use human flow dynamics, which reflect the social activeness of a region, to detect social events and measure their impacts. We first extract human flow dynamics from taxi traces. Second, we propose a method that can not only discover the happening time and venue of events from abnormal social activeness, but also measure the scale of events through changes in such activeness. Third, we extract traffic congestion information from traces and use its change during social events to measure their impact. The results of experiments validate the effectiveness of both the event detection and impact measurement methods.

Categories and Subject Descriptors: J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Social event, social impact, traffic condition, taxi traces

ACM Reference Format:
Wangsheng Zhang, Guande Qi, Gang Pan, Hua Lu, Shijian Li, and Zhaohui Wu. 2015. City-scale social event detection and impact measure with taxi traces. ACM Trans. Intell. Syst. Technol. 6, 3, Article 40 (May 2015), 20 pages.
DOI: http://dx.doi.org/10.1145/2700478

1. INTRODUCTION

Informally, social events are notable occurrences that usually involve lots of people, such as traffic jams, pop concerts, school openings, and exhibitions. Different from general occurrences, a social event not only has a start time and a venue, but also has an unusually large number of participants. In this article, we define "social event" as an occurrence that involves lots of people and is accompanied by an obvious rise in human flow. We think the two major elements that characterize a social event are its time and venue. Before (after) the event, these participants gather to (depart from) the event

This work was partly supported by National Key Basic Research Program of China (2013CB329504), Program for New Century Excellent Talents in University (NCET-13-0521), and the Fundamental Research Funds for the Central Universities. Hua Lu's work was partly supported by the International Network Programme of The Danish Agency for Science, Technology and Innovation under grant no. 12-132707.
Authors' addresses: W. Zhang, G. Qi, G. Pan (corresponding author), S. Li, and Z. Wu, College of Computer Science, Zhejiang University; emails: {zws10, qiguande, gpan, shijianli, wzh}@zju.edu.cn; H. Lu, Office 3.2.03, Department of Computer Science, Aalborg University; email: luhua@cs.aau.dk.

DOI: http://dx.doi.org/10.1145/2700478

venue and generate abnormal traffic flow in the proximity of the venue. Motivated by this, this research uses large-scale taxi trace data to detect social events and measure the potential impact of the detected events.

Social events have great impact on people's lives. First, social events are significant occurrences involving a large number of people, and therefore they cause obvious changes of city dynamics, such as increasing traffic flow and crowds. Second, such increases put severe pressure on city administrations. For example, when a pop concert ends, considerably more traffic than usual is needed to evacuate the large audience quickly and safely. Third, the whooping crowds and uproars during an event bring about threats to public security.

Although many big events are preplanned, there are also many social events that happen suddenly. Even those planned events may be changed,and the change may not be announced to the public in a timely manner. As a result, for many social events, people only become aware of them when they happen. Meanwhile, the end time is usually flexible and hard to precisely plan beforehand. In addition, the exact scale and end time of a social event is sometimes unknown to administrations that need to manage the event and to people who the event may impact.

We believe, social event analysis—the detection of social events and the evaluation of their impact—is very meaningful in practice as a complement to the event plan information. Modeling of social events can help to reveal dynamics of city life and to understand crowd behavior. Detection of social events can highlight urban hot spots for recommendation. Appropriate measuring of social event impacts can help city administration be well prepared for crowd evacuation and emergency handling.

Traditional methods for social event analysis rely on investigation of public news and government statistics. Such simple methods are limited in that (i) event detection has a large time lapse, since it costs much time to collect the data and to disseminate detection result. For example, in China, an event can be reported to the government and made public only through a step-by-step procedure. (ii) There is no general method for impact measure. For example, traffic accidents or natural disasters are investigated in detail, whereas concerts or exhibitions are overlooked.

Recently, there have been more data available for social event analysis, such as mobile phone data, surveillance camera data, and traffic data. However, using these data has limitations. Mobile data may involve serious privacy issues. Surveillance camera data analysis involves huge computing resource consumption, and the surveillance camera infrastructure may not cover the whole city. Traffic data are more suitable for monitoring traffic conditions than for analyzing human behaviors.

Twitter and LBSN data are also used for social event analysis. Using smartphones, people can record their social life, upload the records to social network, and forward news to their friends on the Internet. Such behaviors generate massive message datasets that can be analyzed to detect social events.

Nevertheless, using Twitter or LBSN data for social event analysis also has its shortcomings. First, it is generated by users and has bias due to user preference. For example, a social event will have little related content on Twitter if few users are interested in sharing it on Twitter. Second, it may have a delay due to user response. For example, a user may record his or her place and share it on LBSN after the event starts.

Today, ubiquitous localization devices make people record their traces easily and generate large-scale trace data [Guo et al. 2014]. Such data reflect human mobility in the real world and characterize city dynamics, and therefore they can be used for analyzing social events. Social events involve a lot of participants; the gathering and departure of people create abnormal human flows around the event venue. Taxi GPS trace data can capture human flow information, as well as traffic flow fluctuations caused by social events [Castro et al. 2013]. It is automatically collected by taxis and not affected by user preferences or delays. We think it is a reliable record of human flow. Therefore, by analyzing large-scale trace data, we can detect an obvious rise in human flows, most likely caused by social events, and then directly measure their impacts on traffic and transportation. Moreover, taxi positioning systems are mature and have been operating for several years in many cities in China. Because taxis are a public transportation system, taxi trace involves less privacy issues than mobile phone data.

In this article, we use taxi GPS traces to detect social events and evaluate their impacts. First, we present a method to model regional social activeness by taxi passenger number (pick-up and drop-down number). Pick-up/drop-down number reflects the increase/decrease of population in a region. Therefore, we detect venues and happening times of social events based on mining abnormal regional social activeness. Furthermore, we quantify the scale of each event and analyze its impact on transportation systems with the increase of social activeness in the event's duration.

Instead of focusing on a certain kind of event or even a specific event, the aim of this article to detect various city-scale social events and measure their impacts by using taxi traces. These events can happen at any place or any time and have different scales. The start and end time of events may be very diverse. In addition, the happening venues of events may be irregular and of different sizes. Also, city-scale social events involve different number of participants. Therefore, they are diverse in scale and impact on city life. All these features make the problem studied in this article a challenging one.

We propose a method to adaptively detect the time and place an event happens based on taxi GPS traces. The taxi GPS traces reveal temporal and spatial passenger flow distribution. Using such information, we design a probabilistic model for the routine social behavior dynamics of a region. This model depicts the probability of passenger flow when no social event happens in a region. The abnormal passenger flow corresponds to abnormal social behavior and social events. We present a bottom-up method for event detection, which merges small temporal and spatial segments into the actual happening place and time of an event.

Moreover, we quantitatively measure the scale and impact of social events. In particular, we calculate the extra amount of passenger flow, which depicts additive human flow during events, to measure the scale of a social event. The occurrences of social events directly impact traffic systems. Thus, to analyze the impact of social events, we compute traffic conditions to reveal the relation between increase in traffic congestion and scale of social events.

This article's contributions are threefold: (i) We formulate a social event detection and evaluation problem that takes taxi GPS trace data as input, (ii) we propose a probabilistic model for social event detection from the taxi GPS trace data, and (iii) we design a quantitative method to evaluate the scale and impact of social events.

2. RELATED WORKS

The ubiquity of localization techniques makes trace data increasingly available. Traces of individuals can be used to mine meaningful places [Isaacman et al. 2011] and frequent trajectory patterns [Liu et al. 2007], predict future traces [Monreale et al. 2009; Li et al. 2012], recognize location semantics [Lian and Xie 2011] and recommend trips [Zheng et al. 2013] for individuals, study urban lifestyles [Yuan et al. 2013], measure regional economic development [Chen et al. 2014], and build social ties between individuals [Eagle et al. 2009]. Traces from public transportation characterize urban traffic dynamics, social activeness, and regional dynamics. For a brief survey of research issues, methods, and applications in trace analysis and mining, refer to Pan et al. [2013b].

Traces are related to individuals; thus, large-scale trace data can be used to characterize the social behavior dynamics of a region. Research in this aspect is mainly based on mobile phone call data and traffic data. By using mobile phone call records, the call volume of each cell tower can be used to measure the social activeness of the cell that the tower covers [Sagl et al. 2012]. Such records can also be used to analyze whether the social behavior pattern of the region is repeated daily [Sevtsuk and Ratti 2010] to extract the eigen-behaviors of each region [Reades et al. 2009] and to cluster regions according to their eigen-behaviors [Becker et al. 2011]. With trace data from traffic systems, researchers have proposed extracting passenger flow volume to characterize regional social activeness and similar results [Kaltenbrunner et al. 2010; Pan et al. 2013a; Yue et al. 2009]. Such research shows that statistics of human flow in a region can depict regional social activeness. Nevertheless, cell towers are sparsely located, and cell tower data are too coarse to reveal events in small regions.

There has been research on social event detection from online data. People discuss real-world events in cyberspace and leave corresponding hot topics. Social events can be regarded as a special kind of social behavior, which involve lots of participants and a set of individual activities. Sakaki et al. [2013] extract earthquake-related messages from Tweet, detect the location of earthquake by the location data within the messages, and find the center and the trajectory of it. Zhou and Chen [2014] propose a novel framework to detect composite social events over Twitter social media streams, which fully exploits the information of social data over multiple dimensions. Lee and Sumiya [2010] develop a geo-social event detection system by monitoring crowd behaviors indirectly via Twitter. Zhao et al. [2007] extract keywords from blogs and emails, build a graph to model information flows between users and multigraphs to form information streams on a network, and use graph cuts to detect social events. Petkos et al. [2012] propose a novel multimodal clustering algorithm to fuse multimedia items with the purpose of detecting social events. Bao et al. [2013] propose a robust high-order co-clustering algorithm to detect social, real-world events from the sharing of images/videos on social media sites like Flickr and YouTube. People will ask search engines for help when an epidemic starts; Ginsberg et al. [2008] show that the number of related queries in Google can be used to detect the outbreak of influenza. However, online data, such as tweets, are generated by social network users and is biased due to user preference. For example, a social event will have no related content on Twitter if no user wants to share it on Twitter. LBSN data may have a delay due to user response. For example, a user may record his or her place and share it on LBSN only after the event starts.

Social event detection, however, can benefit from the information of real-world social dynamics hidden in traces. Social events involve lots of participants and are accompanied by abnormal peaks of social activities. The number of phone calls can characterize social activeness and thus help to detect social events. Candia et al. [2008] use the number of mobile phone calls to depict regional social activeness and find that it follows a daily pattern. Thus, an anomaly detection method can directly highlight outbreaks of social events. Traag et al. [2011] analyze Call Detail Records (CDR) to detect social events like festivals, measure their impact on human life, and discover the areas where assistance is needed. For mobile data, fine-grained position data (e.g., GPS) involves serious privacy issue whereas coarse-grained position data (e.g., localization by cell towers) is unable to support subsequent fine analysis.

Trace data not only allow detecting outbreaks of events, but also make it possible to measure their impacts on the physical world. Vaccari et al. [2010] use CDRs to measure urban dynamics and to visualize the impacts of special events, such as a presidential inauguration, on the number of calls. Wen et al. [2008] apply vehicle traces to measure urban traffic congestion and visualize the impact of the Beijing Olympic Games on local transportation systems. Di Lorenzo et al. [2013] propose an intelligent tool for exploration of social events dynamics from augmented trajectories along the spatial, temporal, and organizational dimensions. Researches in these aspects, however, are preliminary visualization work and have not characterized the relativeness between social events and traffic congestion.

To sum up, the data that current research on social event analysis use has limitations. Twitter and LBSN data are generated by users and exhibit (i) bias due to user preference and (ii) delay due to user response. Mobile data may involve serious privacy issues. Traffic data are more suitable for monitoring traffic conditions than for analyzing human behaviors.

3. PROBLEM DESCRIPTION

Happening time, venue, and impact are the three main properties of social events. Knowing the happening time and venue of events can help to quickly locate an event and find its participants. Impact measurement allows us to understand how an event would affect the urban system, especially urban traffic. Thereby, our work aims to solve two problems: namely, event detection and impact measurement. Social event detection must extract the happening time and venue of events from continuous time and space. Impact measurement must quantify the scale of events and their relation to traffic congestion. Formally, these two problems can be defined as:

Problem 1: Given taxi traces as input, the detection of social events must find the place and time events that happen. From the mathematical view, given taxi history traces Tr of a city-scale area, we are to find the set of events $\{e\}$ where $e = (R, d_{min}, d_{max}, t_{min}, t_{max})$ means that an event happens in region R from day d_{min} to d_{max}, time t_{min} to t_{max}.

Problem 2: The impact measurement of events must measure the scale of events. Mathematically, given taxi traces Tr and the set of events $\{e\}$, we are to calculate an event's scale I_e.

It is not straightforward to find the relation between a single trace and social events, but large-scale trace data can characterize the social activeness of a region and even highlight the abnormal dynamics caused by events. The pick-up/set-down numbers, which can be extracted from taxi traces, indicate the increase/decrease in numbers of people in a region. Such information is a direct indication of fluctuation in the social activeness of a region. The social activeness of a region is generally regular at the same time on different days and weeks when there is no social event. The occurrence of social events, however, obviously breaks such regularity (cf. Figure 1). Therefore, by detecting the abnormal dynamics of a region, we can detect social events.

Therefore, the event detection problem for a given region can be transformed into two steps: (i) how to characterize social activeness of the region and (ii) how to detect abnormal dynamics to find events. The first step is a feature extraction problem. Given a region, we characterize the social activeness by its human flow dynamics D extracted from traces Tr. We denote the human flow dynamics on day d_i at time t_j as $D_{i,j}$; Moreover, since $D_{i,j}$ often has a daily routine, we extract the regular social activeness $\overline{D_j}$ in time t_j by average $D_{i,j}$ with same j over days. Therefore, social events can be detected by finding abnormal dynamics. We assume social activeness $D_{i,j}$ obeys a probability distribution:

$$Pr\left(D_{i,j} - \overline{D_j} \leq \epsilon\right) = \int_{-\infty}^{\epsilon} f_j(x)\,dx. \tag{1}$$

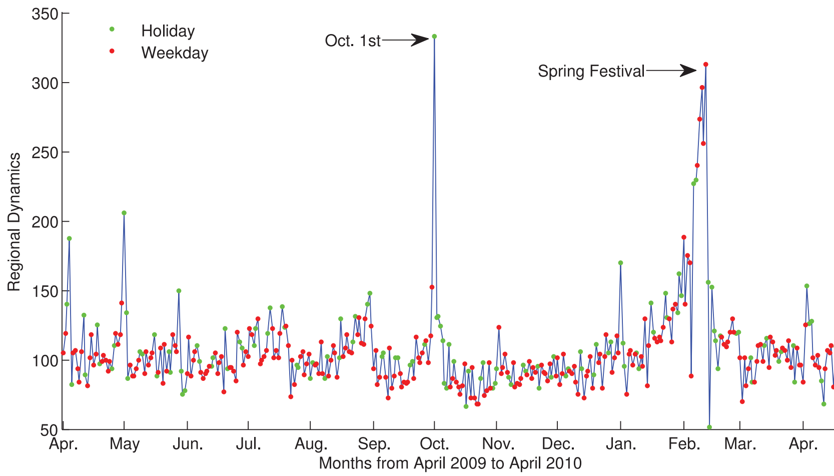


Fig. 1. An example of social event: Lots of merchants come to a wholesale market in Hangzhou, which causes irregular dynamics during important holidays.

The probability that an event happens at time t_j on day d_i is:

$$Pr\left(e_{i,j}|Tr\right) = 1 - \int_{D_{i,j}-\overline{D_j}}^{+\infty} f_j\left(x\right)dx. \tag{2}$$

Nevertheless, the venue where an event happens is not given in advance. It is a challenge for event detection to determine the venue within a city area. To solve this problem, we propose a bottom-up method that reconstructs the venue from tiny parts. Namely, the city road network is divided into small segments r_k, for each of which the probability that an event happens is calculated. Let $a_{i,j,k} = Pr(e_{i,j}^k|Tr)$ be the probability that an event happens by road segment r_k, at time t_j on day d_i. The probability of small segments can be arranged into a matrix A according to spatial and temporal order. This matrix can be viewed as an image, where each pixel (i, j, k) (in row i, column j, layer k) has the probability $a_{i,j,k}$ as a gray value (1: surely an event, 0: no event). This image has many peak regions that indicate social events. Therefore, by using an image segmentation algorithm, we can detect the social events as objects in the image and obtain their venues and happening time from the rows, columns, and layers.

Given the happening time and venue of an event e, we calculate its scale I_e as the difference between the actual human flow during the event and the regular human flow. More specifically, human flow in each time segment is captured as pick-up/set-down numbers in the region. During the event, lots of participants come to the region, which leads to an increase of human flow over regular times. Using the difference between human flow during the event and that during normal days, we can measure the scale of the social event quantitatively.

4. SOCIAL EVENT ANALYSIS

4.1. Social Event Detection

The aim of event detection is to extract the happening time and venue of potential social events. To solve the problem, in this section, we first present a probabilistic model for the routine social activity of each region. This model depicts the probability of passenger flow when no social event is happening in a region. Second, we present a

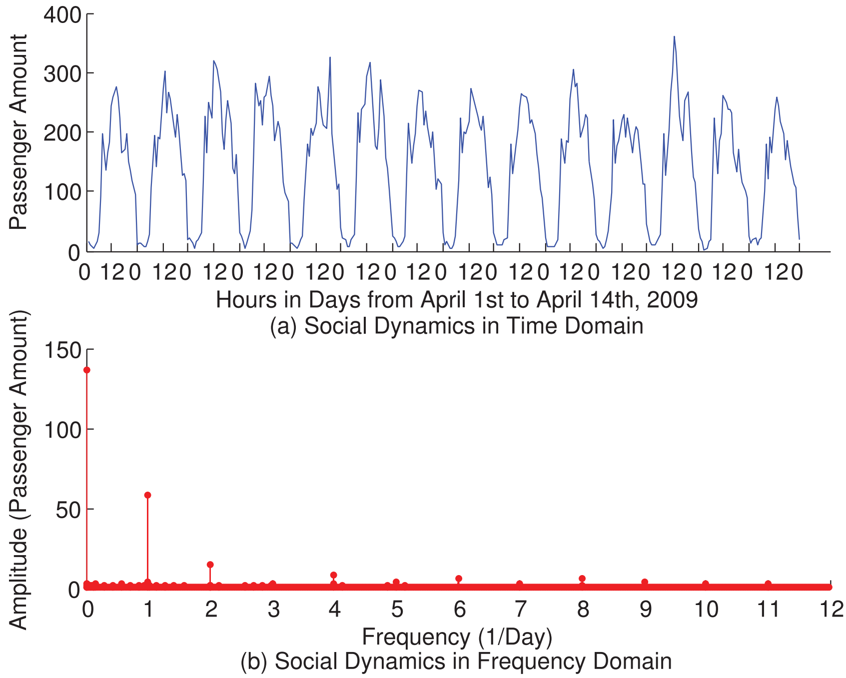


Fig. 2. DFT result of a sample region, (a) origin social activeness; (b) amplitude and period in frequency domain.

bottom-up method for event venue determination. Namely, for each small road segment, we calculate the probability that an event happens. We adopt an image segmentation algorithm to mine the happening time and venue of events from these probabilities.

4.1.1. Modeling Routine Social Activeness. To calculate the probability that a social event happens in a region, we need to model the region's social activeness. Previous research has shown that urban regions' routine social activeness changes periodically. Thus, we model regional social activeness as fluctuations around the regular pattern.

Specifically, social activeness in a region relates to the number of people in that region. Therefore, the social activeness of a region is characterized by taxi set-down numbers that depict the number of people coming to a region. In other words, the set-down number contributes to the increase of people in a region. For each time segment t_n with length of a half-hour, the set-down number is extracted from taxi traces to depict the corresponding social activeness:

$$D(t_n) = SN(t_n) - SN(t_{n-1}).$$

Here, $\{SN(t_n),\ n \geq 0\}$ is the counting process that counts the accumulative number of set-downs in each time segment.

Given the regional social activeness, we can analyze its periodic variation using Discrete Fourier Transform (DFT). As is shown by the sample in Figure 2, the DFT method constructs a representation of social activeness in the frequency domain. Let the social activeness be a discrete time series $\{D(t_n)\},\ n = 0, \ldots, N-1$, the amplitude of the mth sinusoidal component in the frequency domain is:

$$|y_m| = \left| \sum_{n=0}^{N-1} D(t_n) e^{-\frac{2\pi i}{N} mn} \right|, m = 0, \ldots, N-1.$$

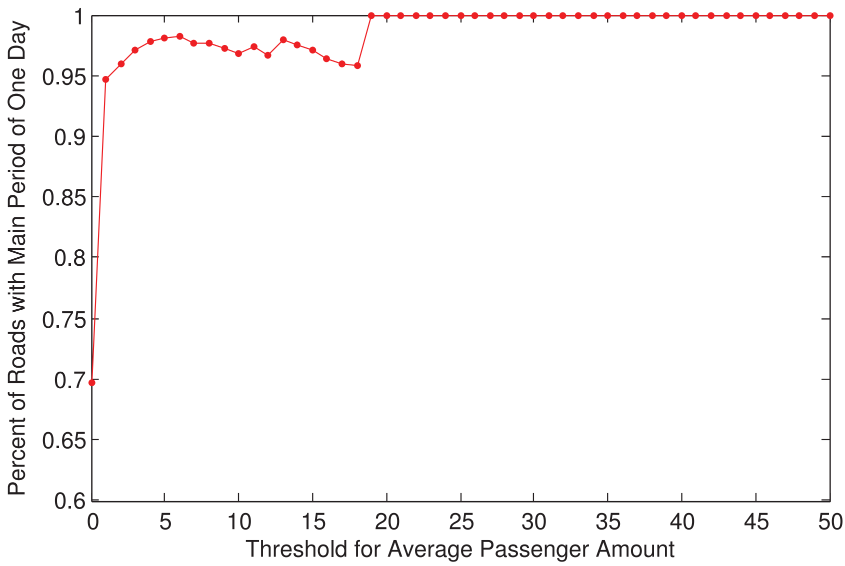


Fig. 3. Percentage of road segments with main period of 1 day divided by segments that have more passengers on average than the threshold.

As a result, the main frequency and period of social activeness can be determined by the sinusoidal component that has the largest amplitude.

In our trace data, we divide the road network into segments that intersect each other only at their end points. To determine the main period of social activeness in each segment, we extract each segment's social activeness during the period of time from April 1, 2009 to April 20, 2010. We then apply DFT to these social activeness periods and obtain the main component that has the largest amplitude (except the DC component). We then investigate the relation between segment's passenger and its main period. For a certain threshold, we filter out segments with average passenger per hour less than that threshold and calculate the percentage of segments with main period 1 day in the rest segments. Figure 3 plots the percentage versus the threshold. According to Figure 3, for segments with average passengers per hour of more than one. More than 95% of them have a pattern of social activeness that follows a daily routine.

Therefore, we model the social activeness using daily patterns and random noises. We change sequence $\{D(t_n), n = 0, \ldots, N-1\}$ to a day/time fashion sequence $\{D_{i,j}\}$ as we mentioned earlier. The regular daily pattern of social activeness is expressed as a vector, whose jth element is estimated unbiasedly as:

$$\overline{D_j} = \frac{1}{\#(i)} \sum_i D_{i,j}.$$

Here, $\#(i)$ is the number of days in trace data Tr. Since it follows a daily routine, social activeness at the same time on different days can be modeled by adding white Gaussian noises to such regular pattern. That is, social dynamic at time j on day i is modeled as:

$$D_{i,j} = \overline{D_j} + \epsilon;\ \epsilon \sim N\left(0, {\sigma_j}^2\right). \tag{3}$$

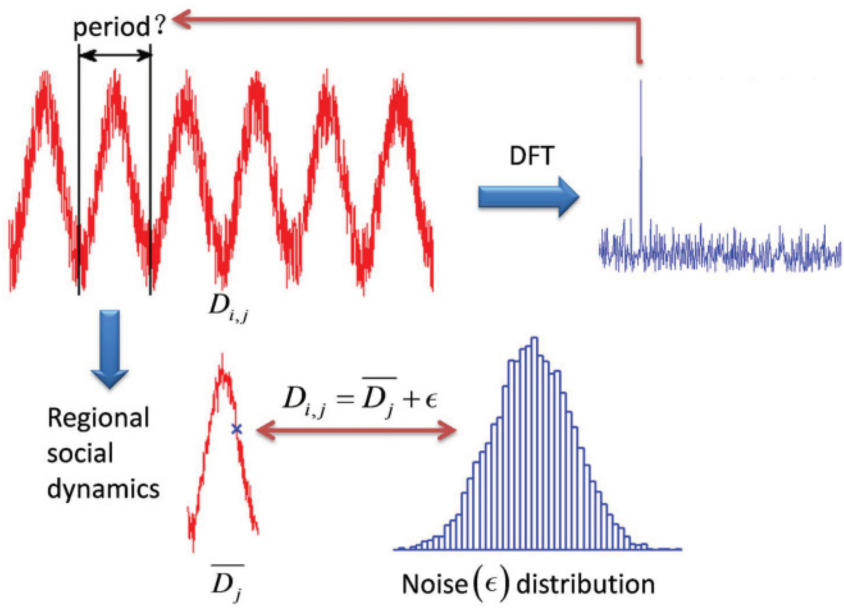


Fig. 4. Illustration of modeling regional social activeness.

The unbiased estimation of ${\sigma_j}^2$ is:

$$ {\sigma_j}^2 = \frac{1}{\#(i) - 1} \sum_i \left(D_{i,j} - \overline{D_j}\right)^2. $$

4.1.2. Extracting Social Events. This section introduces a bottom-up method of social event detection. Based on modeling routine social activeness, we first calculate the probability that no events happen in a small unit (in a road segment at a certain time). Subsequently, we extract the happening time and venue by merging the units where events are much likely to occur.

By modeling regional social activeness, we can characterize the regular activeness of a region. When social events happen, participants gather around the venue, which leads to abnormal activeness in the region. Therefore, given a region, we can use one-tailed statistical hypothesis testing to calculate the probability that an event happens at a certain time. The null hypothesis of this problem is that a region is regular during the jth time interval of the ith day. The alternative hypothesis is that an event happens. Let its social activeness at that time be $D_{i,j}$; the null hypothesis is written mathematically as:

$$ D_{i,j} \sim N\left(\overline{D_j}, {\sigma_j}^2\right). $$

Therefore, according to one-tailed hypothesis testing, the p-value for accepting the null hypothesis is:

$$ p_0 = 1 - \Phi\left(\frac{D_{i,j} - \overline{D_j}}{\sigma_j}\right). $$

It is exactly the probability that no event happens during the jth time interval of the ith day. Therefore, the probability that an event actually happens is:

$$ Pr\left(e_{i,j} | Tr\right) = 1 - p_0. $$

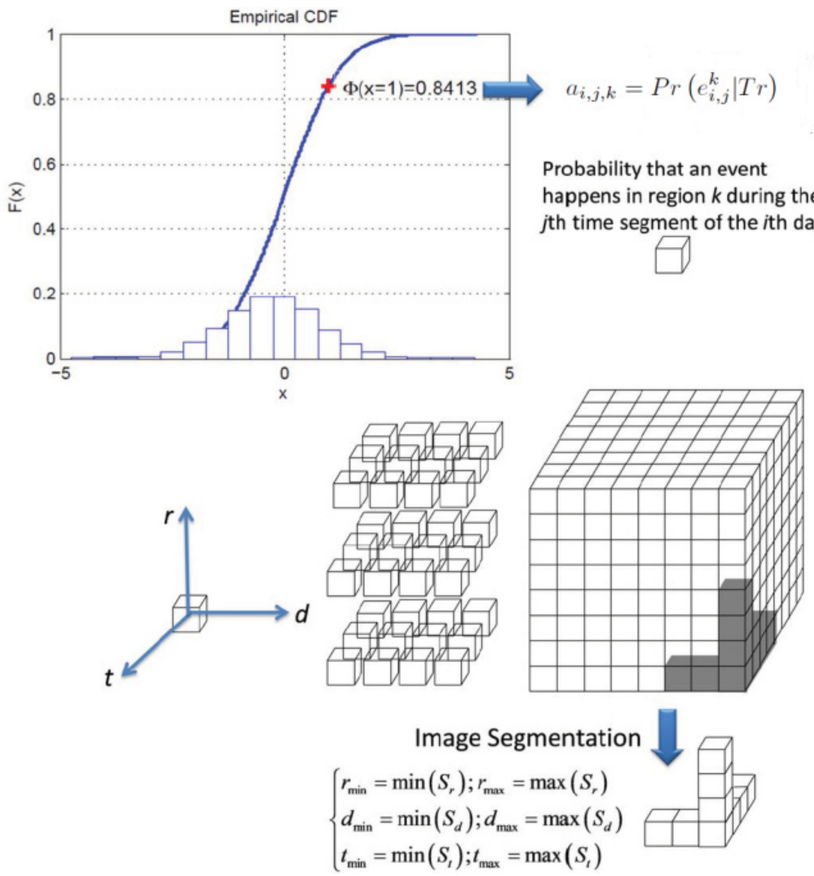


Fig. 5. Illustration of social event detection.

Next, we extract the events by merging the regions and times associated with event happenings. In particular, we use a 3D matrix to represent the aggregation of event probabilities. We divide the city road network into road segments. For each road segment, we calculate the probability that an event happens in the jth time interval of the ith day. By arranging the probabilities according to time and space order, we could retrieve a matrix:

$$A = \left[a_{i,j,k}\right],\ a_{i,j,k} = Pr\left(e_{i,j}^{k}|Tr\right),$$

where the entry in the ith row, jth column, kth layer is the probability that an event happens in the kth region, during the jth time segment of the ith day.

The 3D matrix of event probabilities can be viewed as analogous to a 3D image whose pixels have gray levels representing the probability of event (0 for white and 1 for black). We defined the neighborhood of pixels according to six adjacent connections. With this definition, the temporal and spatial proximity of pixels is retained. That is, neighboring pixels in rows, columns, and layers represent exactly the probability for neighbor time intervals, days, and regions, respectively. Such a neighborhood is reasonable since we know that events may span several regions and last for several hours and days.

The extraction of social events from the 3D matrix is similar to retrieving objects from a 3D image. Thus, we can use image segmentation algorithms to detect social events.

The gray value of each small pixel $a_{i,j,k}$ is the probability of an event; a darker pixel means an event is more like to exist. In this image, each of the dark blocks $b = \{a_{i,j,k}\}$ can be mapped to a spatial ($S_r = \{r_k\}$) and temporal span $S_t = \{t_j\}$ and $S_d = \{d_i\}$ of an event. An image segmentation algorithm can highlight the dark blocks; therefore, it can be used to detect the span of an event. The temporal and spatial border of the events (the temporal border of an event is when it starts and ends) can be calculated as:

$$S_e = (r_{min}, r_{max}, d_{min}, d_{max}, t_{min}, t_{max});$$

$$\begin{cases} r_{min} = min(S_r), \; r_{max} = max(S_r) \\ d_{min} = min(S_d), \; d_{max} = max(S_d) \\ t_{min} = min(S_t), \; t_{max} = max(S_t) \end{cases}$$

Specifically, we use the watershed algorithm for image segmentation in this article. This algorithm treats a grayscale image as a topological surface, where each point has as its altitude as its gray level. Thereby, when putting the surface into water, water rises along the basin of this surface and immerses the surface. Before the water from different basins merges together, we can find the borders of the basins. These borders exactly outline the contours of the objects in the image.

4.2. Event Impact Measurement

Given a social event with the time and venue of its occurrence, this section measures the event's scale and impact. Understanding the scale of social events is very useful in public security in order to evaluate the potential effect of an event on city life. To measure the scale of a social event, the extra number of passengers (which depicts additional human flow during events) is calculated and used. Moreover, the occurrence of social events directly impacts the traffic system because masses of participants gather or leave the venue before or after the event. To analyze the impact of social events, we compute traffic conditions to reveal the relation between increases in traffic congestion and scale of social events.

4.2.1. Scale of Social Events. The increased number of passenger flow is used to measure the scale of social events. Passenger flow reflects the increase of people in a region. When a social event happens, participants gather at the venue, which causes an increase of passenger flow compared to the usual case during the same time on other days. Therefore, this increase in passenger number during events can be used to measure the number of participants and thus also characterize the scale of the event.

Concretely, for an event e, suppose its temporal and spatial span is $S_e = (r_{min}, r_{max}, d_{min}, d_{max}, t_{min}, t_{max})$. Let the social activeness of region r_k during time t_j in day d_i be $D^k_{i,j}$ and its regular social dynamics pattern of time t_j be $\overline{D^k_j}$. The scale of this event is calculated as:

$$I_e = \sum_{t_j=t_{min}}^{t_{max}} \sum_{d_i=d_{min}}^{d_{max}} \sum_{r_k=r_{min}}^{r_{max}} \left(D^k_{i,j} - \overline{D^k_j}\right),$$

where $D^k_{i,j} - \overline{D^k_j}$ is the increase of passenger flow in region r_k during time t_j in day d_i. Thus, this summation is a direct indication for the number of participants of an event.

4.2.2. Impact on Traffic Systems. Social events can affect many aspects of city life. For example, the gathering of crowds can cause many security issues, and the masses of participants in an event call for extraordinary traffic supply. Therefore, it is important to evaluate the impact of events on the real world. In this article, taxi traces are mined to analyze event impact on urban traffic systems. We first estimate the traffic condition of each road segment from taxi traces. Subsequently, the relationship between traffic congestion and social events is revealed.

Traffic congestion specific to a road is calculated as the indication of traffic conditions. Taxis are mobile sensors in a city that report the travel speed of the roads on which they move. Therefore, it is natural to use the average speed of taxis on a road at the same time to calculate traffic congestion. However, an important problem is to consider the different qualities of roads in the congestion value. That is to say, a narrow street may be easily crowded, and the actual number of cars on the street may be small. In such a case, we should calculate a congestion level separately for each road. To that end, we use the method proposed by Liu et al. [2010] to calculate traffic congestion. In particular, let the traffic congestion of the kth road segment at time t_j in day d_i be $c^k_{i,j}$, and the average speed of this road at that time be $v^k_{i,j}$, $c^k_{i,j}$ is calculated as:

$$c^k_{i,j} = \alpha c^k_{i,j-1} + (1-\alpha)\left(1 - P\left(v \leq v^k_{i,j}\right)\right) \tag{4}$$

Here, $P(v \leq v^k_{i,j})$ is the probability that the speed v on this road on historical days is lower than $v^k_{i,j}$, and α is the smoothing parameter, which is recommended to be 0.5.

Therefore, analyzing the relationship between traffic congestion and scale of events needs to address two problems. The first problem is whether the occurrence of an event is accompanied by traffic jams. If so, we further prove the importance of social event analysis and verify the common sense that the gathering and evacuation of participants will bring about traffic congestion around the event venue. The second problem is to evaluate the event's impact on traffic systems. Namely, traffic conditions under usual cases and during a social event are different. Such difference is mainly caused by the social event. A social event with greater scale is supposed to have a larger impact and cause a larger difference in traffic conditions. Thus, the second problem is to reveal the relationship between the scale of social events and traffic condition difference.

5. EXPERIMENTS

5.1. Dataset

In the experiments, we use taxis trace data from two cities, namely Shanghai and Hangzhou. Shanghai, located in the Yangtze River Delta of East China, is the financial center of China with an urban population of 12 million in 2010. Hangzhou, a popular tourism city located in East China, has 6 million urban residents in 2009. The Shanghai dataset contains 2 years' traces from September 19, 2009 to December 31, 2011. This dataset has nearly 10 billion records from more than 10,000 taxis. The Hangzhou dataset contains nearly 1 year's data from April 1, 2009 to April 20, 2010. This dataset has nearly 3 billion records from more than 5,000 taxis in Hangzhou. All these taxi traces are in the same format. In particular, for each taxi, trace data are sampled nearly once per minute; each sample mainly contains these fields:

—VEHICLE_ID: unique taxi ID in the dataset;
—LONGITUDE: current longitude of the taxi;
—LATITUDE: current latitude of the taxi;
—STATE: current status (occupied/vacant) of the taxi;
—Velocity: current driving speed of the taxi;
—TIME: sampling timestamps in the format of "YYYY-MM-DD HH:MM:SS".

5.2. Setup

There are two parts of experiments: namely, social event detection and impact evaluation.

First, this article employs a bottom-up method for event detection. The bottom component for event detection is the probability that an event happens in a 10-meter road segment during an hour. At the bottom level, the urban space is divided into small parts. That is, given the road network in Shanghai and Hangzhou, we divide it into roads that only intersect with others at end points. These roads are further separated into 10-meter road segments. Next, the data are broken into parts according to time. In particular, we divide each day into 24 time segments by hours. Accordingly, for each road segment, we calculate its social activeness and the probability of an event during each hour. Note, however, that we use the first 30 days of data for training the regular social activeness and calculate the remaining days' probability of an event.

The first part of our experiments is social event detection. We detect social events with these bottom components by using the watershed algorithm. For each road, we depict the probability that an event happens in one of the segments of the road during the time span of data by constructing a 3D probability matrix with the bottom components. For example, given a road with length of 1 kilometer, we get 100 road segments. If we have 385-days' data, we get a $385 \times 24 \times 100$ matrix. This matrix is the input for the watershed algorithm for social event detection, and the output consists of the occurrence time and venue of events.

We validate the event detection method by investigating real-world social events. We put the detection result, namely occurrence time and venue, into a search engine. We call it a hit if a real-world event appears in the first 10 search results. Each hit validates that the detected result is an actual event. The percentage of hits on events is used to measure event detection accuracy. We evaluate our method by calculating the recall of these events returned by our detection method.

The second part of our experiments is event impact evaluation. We measure the scale of each event and its impact on traffic conditions. The congestion of each road in each hour is calculated using an existing method [Liu et al. 2010]. Given an event, with its occurrence time and venue, we measure its scale with the increase of passenger number during the studied time compared to a regular time in the venue. Therefore, we measure the relation between event scale and the increase of traffic congestion from regular times and to the time when an event happens.

Both of the event scale and impact are evaluated. To verify that the measurement of the event scale is reasonable, we regress the event scale with actual participants of an event. For the analysis of event impact, we further verify (i) whether the occurrence of events is accompanied by traffic jams and (ii) evaluate an event's impact on the traffic system.

5.3. Event Detection: Result and Analysis

We use Hangzhou taxi traces for the validation of the proposed event detection method. There were 2,596 social events detected in the whole city during the period from May 1, 2009 to April 20, 2010. An overview of the geographical distribution of detected events is shown in Figure 6.

First, we verify the event detection method with the query result in the search engine Google. The query result from Google is the online information that is published by a social event organizer or news agency and crawled by a search engine. It is more credible than unverified user-generated contents and therefore we use it as ground truth. For each detection result, we put its occurrence time and venue into the search engine. If a real-world event appears in the first 10 query results, we judge that this

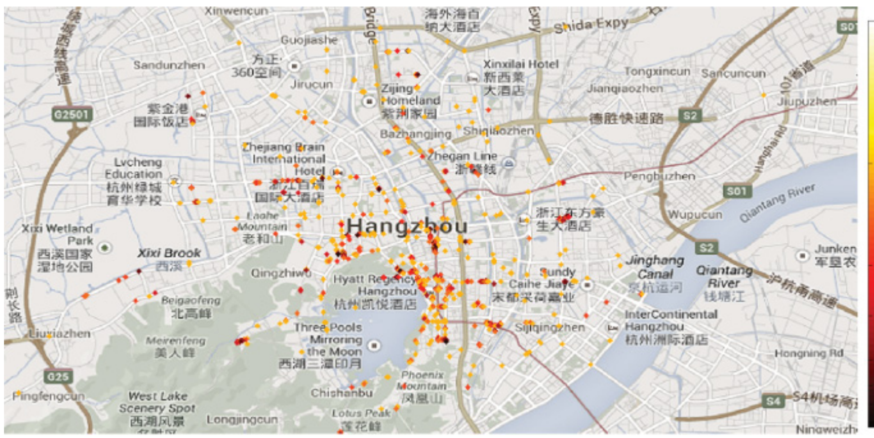


Fig. 6. Heat map of number of detected events over Hangzhou city from May 1, 2009 to April 20, 2010.

Table I. List of the Top 10 Events Detected in This Article

Rank	Venue	Occurrence time	Actual event in search engine
1	Huanglong sports center	Nov. 13–14, 2009	Concerts by Fish Leong and other singers
2	Hangzhou Tower	Dec. 31, 2009 to Jan. 1, 2010	Shopping festival
3	Zijingang Campus of Zhejiang University	Sep. 28–30, 2009	Start of the autumn term
4	Qianjiang College of Hangzhou Normal University	Sep. 29–30, 2009	Start of the autumn term
5	Hangzhou Tower	Nov. 13–14, 2009	Sales promotion
6	Hangzhou City Sport Center	June 6, 2009	National badminton match
7	Qianjiang College of Hangzhou Normal University	Nov. 7–8, 2009	NA
8	Zijingang Campus of Zhejiang University	Jul. 4–5, 2009	Japanese-Language Proficiency Test (JLPT)
9	The northeast corner of Westlake	Oct. 24, 2009	Annual fireworks show of Hangzhou
10	Xixi Wetland	May 28 to 29, 2009	Dragon Boat race of Hangzhou

detection result hits. Thus, in this article, we define accuracy by Google searching hits as follows:

$$accuracy = \frac{number\ of\ events\ in\ top\ ten\ query\ results}{total\ number\ of\ events}.$$

The top 100 detected events are retrieved by order of event scale and labeled with this procedure. According to the definition, the accuracy of our method is 80%. For comparison, we also carry out a baseline method that segments image by threshold. The baseline method uses the same matrix as our proposed method but treats a matrix element as a part of an event if it is higher than a fixed threshold. The accuracy of the baseline approach with an optimal threshold for accuracy (0.83) is 65%. We list the top 10 detected events in Table I.

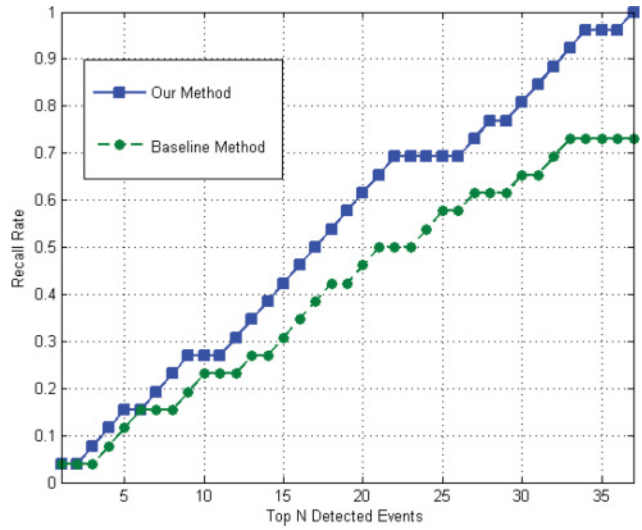


Fig. 7. The recall rate of top N detected events in Huanglong Sports Center by different methods. We find that our method can cover all 26 events in Huanglong Sports Center in its top 37 detected events better than baseline approach.

Second, we define *the recall rate of top N detected events* by the number of ground truth events in the top N detected events divided by the number of all ground truth events. We collect the events that happened in Huanglong Sports Center from its official website (http://www.hlsports.net). There are 14 concerts and 12 sport matches during the period from May 1 to April 20. We use these 26 events as ground truth to calculate recall rate. The result of our method, as well as the baseline approach with optimal threshold for recall (0.91), are shown in Figure 7. We find that our method can cover all 26 events in Huanglong Sports Center in its top 37 detected events, better than baseline approach with optimal threshold for recall.

To better understand the result of our method, we illustrate two location examples of event detection: Zijingang Campus of Zhejiang University and Xixi Wetland, a popular sightseeing place.

Zijingang Campus: The major event of Zijingang Campus is its autumn semester beginning day, when freshmen enroll. We plot the probability of an event in the regions near Zijingang Campus's main entrance from August 27, 2009 to August 31, 2009, in Figure 8. We find that the probability of an event has an impulse in August 28–30, which is consistent with the real semester beginning day.

Xixi Wetland: Xixi Wetland is a popular sightseeing place in Hangzhou. Here, we plot the occurrence time of detected events in Figure 9. We find that all detected events start after 6:00 AM and finish before 18:00 PM, which is consistent with Xixi Wetland's open time.

The detection method's efficiency is an important concern. In our experiments on a Windows-based Genuine Intel(R) CPU i3 @ 3.20GHz machine with 4.00GB memory, the preprocessing of all trace data in Hangzhou costs 78 minutes, and the preprocessing of all trace data in Shanghai costs 102 minutes. The postprocessing of all segment results in Hangzhou costs 33 minutes, and the postprocessing of all segment results in Shanghai costs 67 minutes. The average time cost for the watershed algorithm to process a road over all time periods is 12.6 seconds. Considering that the time cost is

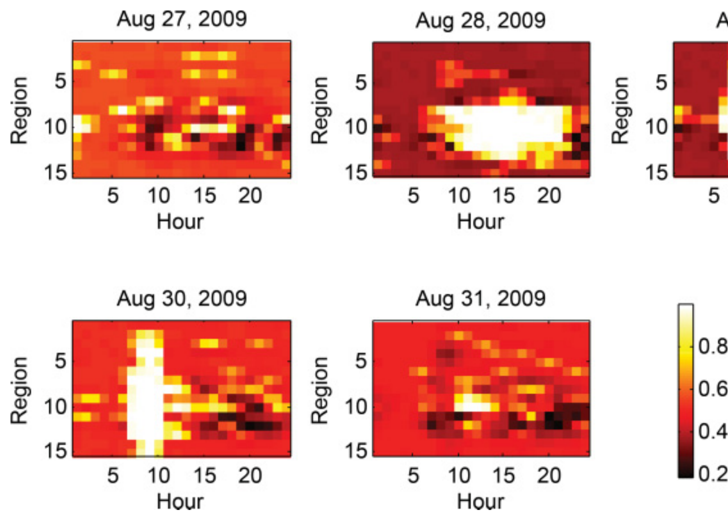


Fig. 8. The probability of event in the regions near Zijingang Campus's main entrance from August 27, 2009 to August 31, 2009. We find that the probability of an event has an impulse in August 28–30, which is consistent with the real semester beginning day.

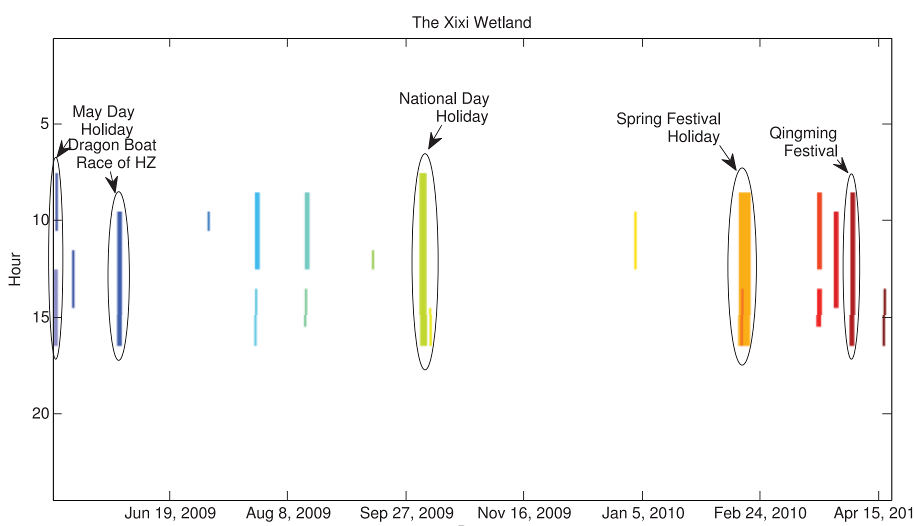
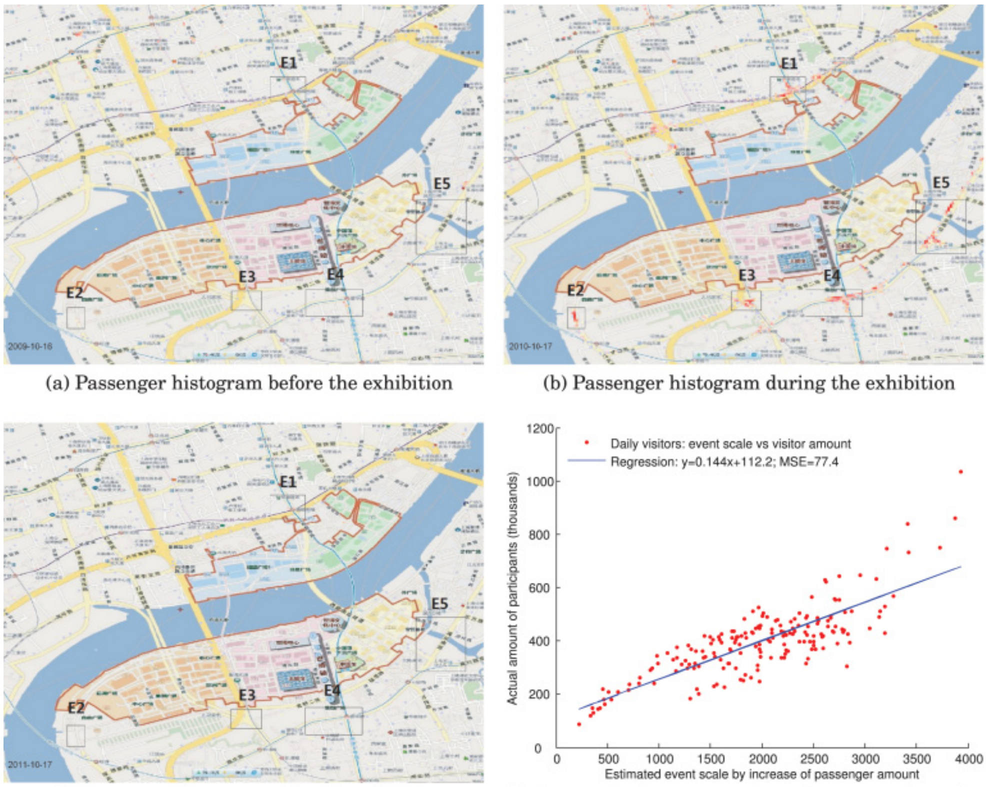


Fig. 9. The occurrence time of detected events in Xixi Wetland. Different events are distinguished by different colors. The events in Top 100 detected events are highlighted with ellipses. We find that all detected events start after 6:00 AM and finish before 18:00 PM, which is consistent with Xixi Wetland's open time.

for all the data of more than 1 year, it is possible to build a real-time application for the real-time taxi GPS traces.

5.4. Impact Evaluation: Result and Analysis

First, we use Shanghai taxi traces to demonstrate our measurement of event scale. In particular, we focus on the occurrence of the Shanghai World Expo from May 1, 2010 to October 1, 2010. Our event detection method finds several places around the pavilion (Figure 10(a), (b), (c)). These places are entrances where people can gather



(a) Passenger histogram before the exhibition

(b) Passenger histogram during the exhibition

(c) Passenger histogram after the exhibition

(d) Regression between estimated event scale and actual participants amount

Fig. 10. (a)–(c) The passenger histogram of 1 day before, during, and after the exhibition. Black bordered rectangles highlight the five entrances (E1–E5) of the World Expo. Red dots in five rectangles represent the scales of increase of passengers in these places. We can see that the exhibition brings about an obvious increase of passengers in these five entrances during the exhibition. (d) Linear regression between estimated event scale and actual participant number. We find that our measurement of event scale has a strong correlation with the actual number of participants (the correlation coefficient is 0.8103).

to enter the exhibitions. Thus, for each day during the World Expo, we calculate the scale of events by the increase of passengers in these places. The actual number of participants for the World Expo can be find online (http://www.expo2010.cn/). We find that our measurement of event scale has a strong correlation with the actual number of participants (the correlation coefficient is 0.8103). The linear regression result is shown in Figure 10.

Second, we use Hangzhou taxi traces to analyze the impact of social events on traffic systems. The traffic congestion of this city is calculated with Equation (4), with 1 (0) to represent that a road has more (less) vehicles than at regular times. To verify that social events are often accompanied by traffic jams, we calculate the average traffic congestion for a given event scale. In particular, we divide the events into m groups according to the event scale. Events with scale between $10(n-1)$ and $10n$ are put into the nth group; the last group is composed of events whose scale is larger than $10(m-1)$. For each group, we calculate the average traffic congestion around the venue where events in that group happen. Figure 11 illustrates the result when $m = 17$. The result shows that as the event scale becomes larger, the traffic congestion around the venue is

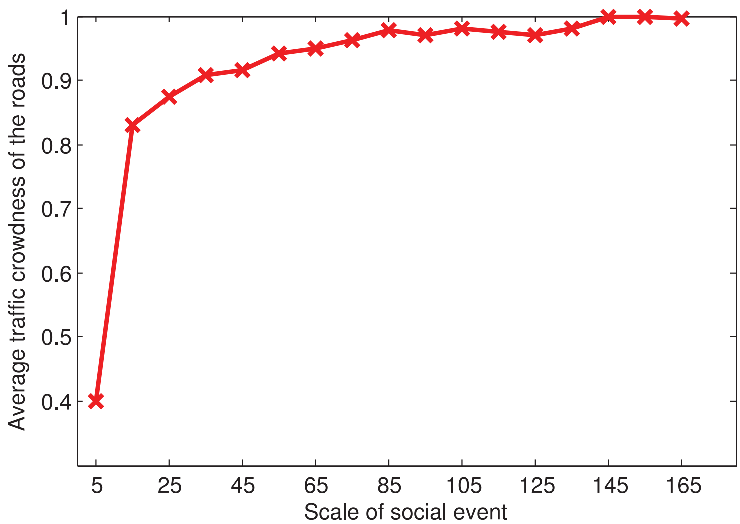


Fig. 11. The average traffic congestion of the roads that event, whose scale is between $[x-5,\ x+5)$ happens.

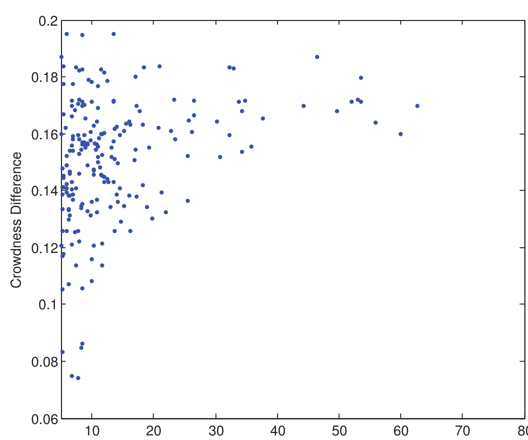


Fig. 12. Visualization of the relationship between scale and increase of traffic congestion for each event (by point)

also higher. Moreover, traffic congestion often accompanies social events whose scales are larger than 15.

Third, we analyze the impact of social events on increase of traffic congestion. We calculate the increase of traffic congestion during an event by subtracting it from regular times. Such increase is supposed to be caused by social events. We visualize the event scale and the increase in Figure 12. This figure shows that a larger social event (with larger scale) also brings about greater impact to traffic systems (much increased traffic congestion).

6. CONCLUSION

Social events are notable occurrences that involve lots of participants and thus bring about great impacts on different aspects of city life. Taxi traces are very useful in providing the human flow dynamics of a region, which depict the regional social activeness. This article mines taxi traces to detect social events and evaluate their impacts. In particular, we first propose a method that can not only discover the happening time and venue of events from the taxi trace data, but can also measure the scale of events with the change of human flow within a region. Second, we extract traffic congestion information from taxi traces and use its change during social events to evaluate the impact of social events. Through the experiments, we validate the effectiveness of the event detection and impact evaluation methods.

REFERENCES

Bing-Kun Bao, Weiqing Min, Ke Lu, and Changsheng Xu. 2013. Social event detection with robust high-order co-clustering. In *Proceedings of the 3rd ACM International Conference on Multimedia Retrieval*. ACM, New York, NY, 135–142. DOI:http://dx.doi.org/10.1145/2461466.2461491

Richard A. Becker, Ramón Cáceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2011. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing* 10, 4 (Apr. 2011), 18–26. DOI:http://dx.doi.org/10.1109/MPRV.2011.44

Julián Candia, Marta C. González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41, 22 (Nov. 2008), 224015. DOI:http://dx.doi.org/10.1088/1751-8113/41/22/224015

Pablo Samuel Castro, Daqing Zhang, Chao Chen, Shijian Li, and Gang Pan. 2013. From taxi GPS traces to social and community dynamics: A survey. *ACM Computing Survey* 46, 2, Article 17 (Dec. 2013), 34 pages. DOI:http://dx.doi.org/10.1145/2543581.2543584

Longbiao Chen, Daqing Zhang, Gang Pan, Leye Wang, Xiaojuan Ma, Chao Chen, and Shijian Li. 2014. Container throughput estimation leveraging ship GPS traces and open data. In *Proceedings of the 2014 ACM Conference on Ubiquitous Computing (UbiComp'14)*. 847–851. DOI:http://dx.doi.org/10.1145/2632048.2632050

G. Di Lorenzo, M.L. Sbodio, V. Lopez, and R. Lloyd. 2013. EXSED: An intelligent tool for exploration of social events dynamics from augmented trajectories. In *Proceedings of the 14th IEEE International Conference on Mobile Data Management (MDM)*, Vol. 1. 323–330. DOI:http://dx.doi.org/10.1109/MDM.2013.47

Nathan Eagle, Alex Sandy Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 36, 106 (Aug. 2009), 15274–15278. DOI:http://dx.doi.org/10.1073/pnas.0900282106

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2008. Detecting influenza epidemics using search engine query data. *Nature* 457, 7532 (Nov. 2008), 1012–1014. DOI:http://dx.doi.org/10.1038/nature07634

Bin Guo, Zhiwen Yu, Daqing Zhang, Huilei He, Jilei Tian, and Xingshe Zhou. 2014. Toward a group-aware smartphone sensing system. *IEEE Pervasive Computing* 13, 4 (2014), 80–88. DOI:http://dx.doi.org/10.1109/MPRV.2014.80

Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. Identifying important places in people's lives from cellular network data. In *Proceedings of the 9th International Conference on Pervasive Computing*. Springer, Berlin, 133–151. DOI:http://dx.doi.org/10.1007/978-3-642-21726-5_9

Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing* 6, 4 (Aug. 2010), 455–466. DOI:http://dx.doi.org/10.1016/j.pmcj.2010.07.002

Ryong Lee and Kazutoshi Sumiya. 2010. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks(LBSN'10)*. ACM, New York, NY, 1–10. DOI:http://dx.doi.org/10.1145/1867699.1867701

Xiaolong Li, Gang Pan, Zhaohui Wu, Guande Qi, Shijian Li, Daqing Zhang, Wangsheng Zhang, and Zonghui Wang. 2012. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science in China* 6, 1 (Feb. 2012), 111–121. DOI:http://dx.doi.org/10.1007/s11704-011-1192-6

Defu Lian and Xing Xie. 2011. Learning location naming from user check-in histories. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, New York, NY, 112–121. DOI:http://dx.doi.org/10.1145/2093973.2093990

Siyuan Liu, Yunhuai Liu, Lionel M Ni, Jianping Fan, and Minglu Li. 2010. Towards mobility-based clustering. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 919–928. DOI:http://dx.doi.org/10.1145/1835804.1835920

Yunhao Liu, Lei Chen, Jian Pei, Qiuxia Chen, and Yiyang Zhao. 2007. Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays. In *Proceedings of the 5th IEEE International Conference on Pervasive Computing and Communications*. IEEE, Piscataway, NJ, 37–46. DOI:http://dx.doi.org/10.1109/PERCOM.2007.23

Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. 2009. WhereNext: A location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 637–646. DOI:http://dx.doi.org/10.1145/1557019.1557091

Gang Pan, Guande Qi, Zhaohui Wu, Daqing Zhang, and Shijian Li. 2013a. Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems* 14, 1 (Mar. 2013), 113–123. DOI:http://dx.doi.org/10.1109/TITS.2012.2209201

Gang Pan, Guande Qi, Wangsheng Zhang, Shijian Li, Zhaohui Wu, and Laurence Tianruo Yang. 2013b. Trace analysis and mining for smart cities: Issues, methods, and applications. *IEEE Communications Magazine* 51, 6 (2013). DOI:http://dx.doi.org/10.1109/MCOM.2013.6525604

Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2012. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR'12)*. ACM, New York, NY, Article 23, 8 pages. DOI:http://dx.doi.org/10.1145/2324796.2324825

Jonathan Reades, Francesco Calabrese, and Carlo Ratti. 2009. Eigenplaces: Analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design* 36, 5 (2009), 824–836. DOI:http://dx.doi.org/10.1068/b34133t

Günther Sagl, Bernd Resch, Bartosz Hawelka, and Euro Beinat. 2012. From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments. In *Proceedings of the 6th Geoinformatics Forum*. 54–63.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2013. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering* 25, 4 (2013), 919–931. DOI:http://dx.doi.org/10.1109/TKDE.2012.29

Andres Sevtsuk and Carlo Ratti. 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology* 17, 1 (Mar 2010), 41–60. DOI:http://dx.doi.org/10.1080/10630731003597322

V. A. Traag, A. Browet, F. Calabrese, and F. Morlot. 2011. Social event detection in massive mobile phone data using probabilistic location inference. In *Proceedings of the 3rd IEEE International Conference on Social Computing (SocialCom)*. 625–628. DOI:http://dx.doi.org/10.1109/PASSAT/SocialCom.2011.133

Andrea Vaccari, Mauro Martino, Francisca Rojas, and Carlo Ratti. 2010. Pulse of the city: Visualizing urban dynamics of special events. In *Proceedings of the 20th International Conference on Computer Graphics and Vision*. ACM, New York, NY.

Huimin Wen, Zhongwei Hu, Jifu Guo, Liyun Zhu, and Jianping Sun. 2008. Operational analysis on Beijing road network during the Olympic Games. *Journal of Transportation Systems Engineering and Information Technology* 8, 6 (Dec. 2008), 32–37. DOI:http://dx.doi.org/10.1016/S1570-6672(09)60003-9

Nicholas Jing Yuan, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie. 2013. We know how you live: Exploring the spectrum of urban lifestyles. In *Proceedings of the 1st ACM Conference on Online Social Networks*. 3–14.

Yang Yue, Yan Zhuang, Qingquan Li, and Qingzhou Mao. 2009. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In *Proceedings of the 17th International Conference on Geoinformatics*. IEEE, Piscataway, NJ, 1–6. DOI:http://dx.doi.org/10.1109/GEOINFORMATICS.2009.5293469

Qiankun Zhao, Prasenjit Mitra, and Bi Chen. 2007. Temporal and information flow based event detection from social text streams. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. AAAI Press, Palo Alto, CA, 1501–1506.

Kai Zheng, Shuo Shang, Nicholas Jing Yuan, and Yi Yang. 2013. Towards efficient search for activity trajectories. In *Proceedings of the 29th IEEE International Conference on Data Engineering*. 230–241. DOI:http://dx.doi.org/10.1109/ICDE.2013.6544828

Xiangmin Zhou and Lei Chen. 2014. Event detection over twitter social media streams. *The VLDB Journal* 23, 3 (Jun. 2014), 381–400. DOI:http://dx.doi.org/10.1007/s00778-013-0320-3

Received January 2014; revised September 2014; accepted October 2014