Sensing the Pulse of Urban Activity Centers Leveraging Bike Sharing Open Data

Longbiao Chen*†, Dingqi Yang‡*, Jeremie Jakubowicz*, Gang Pan†, Daqing Zhang*,Shijian Li†
*Institut Mines-Telecom; Telecom SudParis; UMR CNRS SAMOVAR, France
{longbiao.chen, dingqi.yang, jeremie.jakubowicz, daqing.zhang}@telecom-sudparis.eu
†Zhejiang University, China
{longbiaochen, gpan, shijianli}@zju.edu.cn
‡eXascale Infolab, University of Fribourg, Switzerland
dingqi.yang@unifr.ch

***Abstract*—Understanding social activities in Urban Activity Centers can benefit both urban authorities and citizens. Traditionally, monitoring large social activities usually incurs significant costs of human labor and time. Fortunately, with the recent booming of urban open data, a wide variety of human digital footprints have become openly accessible, providing us with new opportunities to understand the social dynamics in the cities. In this paper, we resort to urban open data from bike sharing systems, and propose a two-phase framework to identify social activities in Urban Activity Centers based on bike sharing open data. More specifically, we first detect bike usage anomalies from the bike trip data, and then identify the potential social activities from the detected anomalies using a proposed heuristic method by considering both spatial and temporal constraints. We evaluate our framework based on the large-scale real-world dataset collected from the bike sharing system of Washington, D.C. The results show that our framework can efficiently identify social activities in different types of Urban Activity Centers and outperforms the baseline approach. In particular, our framework can identify 89% of the social activities in the major Urban Activity Centers of Washington, D.C.**

***Keywords*-Urban Open Data; Bike Sharing System; Urban Activity Center**

I. Introduction

Urban Activity Centers (UACs) are urban areas designed for recreation, leisure, or tourism [1]. Large social activities, such as concerts and festivals, are often held in these areas. When such an activity takes place, there tends to be a big crowd of people gathering in the related UAC within a short period of time, which usually causes traffic congestion and potential security risks. For instance, during the opening game of the FIFA World Cup 2010, a traffic jam occurred for more than four hours around Johannesburg, South Africa where the game was held.

To implement better strategies for managing social activities in UACs, such as avoiding traffic congestion and ensuring public safety, urban authorities need to understand the characteristics of these activities. In particular, they need to investigate what kinds of activities are being held, when these activities start and end, and where people come from and go back to. Traditionally, such information is manually collected from heterogeneous sources, including activity hosts, transportation divisions, and sample participants. This often consumes substantial human labor and time.

Recently, researchers have proposed different approaches to automatically identify and characterize social activities in UACs using various kinds of *digital footprints* [2]–[4]. For instance, *Location-Based Social Networks (LBSNs)* contain large-scale user activity data, which can be used to understand user attendance in social activities [5]–[7]. However, since LBSNs largely rely on users' voluntary self-report, a considerable amount of critical information about human flow might be missing, making it difficult to characterize social activities [8]. For instance, after users check in at a concert, they might not check out when the concert ends. To infer such missing information, researchers have exploited *Taxi GPS traces* [9], since taxi is an important transportation means in urban areas [10], [11]. However, due to the competition among different taxi companies, the taxi GPS trace data are generally not openly accessible.

The emergence of bike sharing systems provides us with new opportunities to tackle this problem. Bike-share systems have been deployed in many cities to promote green transportation and healthy lifestyle [12]. The system data are often *openly accessible* to make it more convenient for riders to search for a bike station nearby, check the availability of bikes and docks, etc. From these data, researchers can extract rich information about human flow in urban areas, as a lens on the social activities occurring in UACs. For instance, if many people ride bikes to attend a concert, there will be a significant increase of bikes arrived at the stations near the concert venue before the opening time, and a significant decrease shortly after it ends.

In this paper, we attempt to sense the social activities in Urban Activity Centers by leveraging bike sharing systems as sensors. More specifically, by analyzing the bike sharing open data, we aim to identify social activities and characterize the bike users related to these activities. We propose a two-phase framework to achieve this goal. In the first phase, we extract bike usage patterns of each bike station, and detect the bike usage anomalies that are both *significant* compared to the bike usage in adjacent hours, and *unusual* compared to the station's daily bike usage

978-1-4673-7211-4 2015

DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.43

routine in the same hour. In the second phase, we pair and aggregate the anomalies detected in neighboring stations to identify social activities by considering both *spatial* and *temporal* constraints using two heuristic rules, respectively. By analyzing the bike trips related to these activities, we can further characterize the user types (e.g. residents or tourists) and their common origins and destinations.

The main contributions of this paper include:

1) A novel use case of bike sharing open data, namely social activity identification in Urban Activity Centers.
2) A two-phase framework to identify social activities from bike sharing open data. First, we detect significant and unusual bike usage patterns from historical bike trip data. Then, we derive two heuristic rules on spatial-temporal constraints to guide social activity identification from bike usage anomalies.
3) A performance evaluation of the proposed framework using the large-scale, real-world bike sharing open data from Washington, D.C. The results show that our framework effectively identifies 89% of the social activities in the major Urban Activity Centers, and outperforms the baseline approach.

The remainder of this paper is organized as follows. We review the related work in Section II, and describe the datasets in Section III. In Section IV we present the problem and our framework, and then detail the two phases in Section V and VI respectively. We report the evaluation results in Section VII. Finally, we discuss the limitations of our work in Section VIII and conclude the paper in Section IX.

II. Related Work

We survey the related work in two aspects: (1) social activity identification and analysis, and (2) bike sharing data applications.

A. Social Activity Identification and Analysis

Researchers have exploited different types of digital footprints for social activity identification and analysis, including *LBSN user activity data* and *taxi GPS trace data*. For instance, Liang et al. [6] used LBSN check-ins to model the size, duration, and temporal dynamics of short-lived crowds formed in social activities. However, the user-contributed data are often noisy and incomplete [5], [8], especially when users leave a social activity, they usually do not check out of the venue, making it challenging to identify the durations of social activities and capture the characteristics of users [6]. Zhang et al. [9] took a different approach, leveraging taxi GPS traces to extract human flow patterns in urban areas, and proposed a method to evaluate the social activeness of city-scale regions. However, since taxis are usually operated by different companies that do not want to make their operation data accessible to their competitors, it is usually difficult for researchers to obtain continuous and large-scale taxi trace data. In contrast, our work leverages publicly accessible bike sharing open data, which contain rich information about human flow in urban areas.

B. Bike Sharing Data Applications

The spread of of bike sharing systems in many cities enables a sustainable transportation means and a healthy lifestyle. Meanwhile, a large volume of bike usage data are generated, providing invaluable resource for researchers to enable new applications. For instance, some work [13] used bike usage data to optimize routes and strategies of bike sharing system re-balancing operation, which is usually done by transporting bikes among stations using other vehicles. Some other work [14] tried to improve bike station placement by modeling the bike demands in different areas from historical bike usage data. By predicting the bike usage in the future, Zhao et al. [15] proposed a system to help riders find available bikes and docks in the nearest stations. Nevertheless, none of the previous work inferred the connection between human flow and urban dynamics from the bike usage data. Our work leverages such data to capture the social activities in Urban Activity Centers.

III. Dataset Description

In this section, we present the two datasets used in this paper, viz., the *Bike Trip Dataset* and the *Urban Activity Center Dataset*. Both datasets are collected from the open data portals of the Washington, D.C. government.

A. Bike Trip Dataset

The data of many bike sharing systems have been openly accessible. In this paper, we use the historical bike trip dataset from the Capital Bikeshare System of Washington, D.C. The system consists of more than 200 bike stations, and each station is equipped with a specific number of bikes and docks, depending on local demand estimated by the planners. Since its opening in September 2010, Capital Bikeshare has become one of the largest bike sharing systems in the United States [16]. Figure 1 shows the coverage of the Capital Bikeshare stations in Washington, D.C. urban areas.

We retrieve the *Bike Trip Dataset* of Washington, D.C. from the Capital Bikeshare System Data Portal[1], including three years of bike trip records from 2012 to 2014. In summary, this dataset contains 3,296 bikes in 203 bike stations, among which 8,010,668 trips are recorded. Each trip record contains the following fields:

- **Trip ID**: the unique trip ID.
- **Departure Station ID**: the unique ID of the station where the corresponding bike departs from.
- **Departure Time**: the time when the corresponding bike is rent from a dock of the departure station.
- **Arrival Station ID**: the unique ID of the station where the corresponding bike arrives at.

[1] https://www.capitalbikeshare.com/system-data

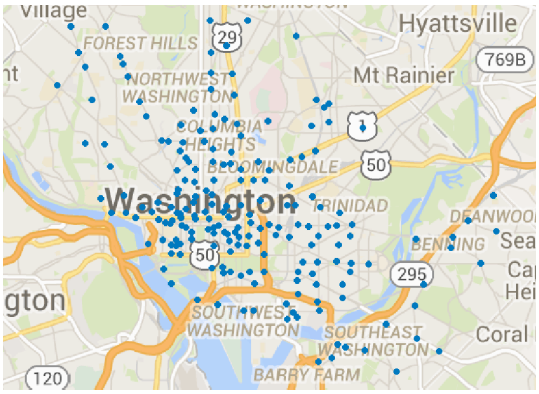

Figure 1: Station map of the Capital Bikeshare system in Washington, D.C. urban areas.

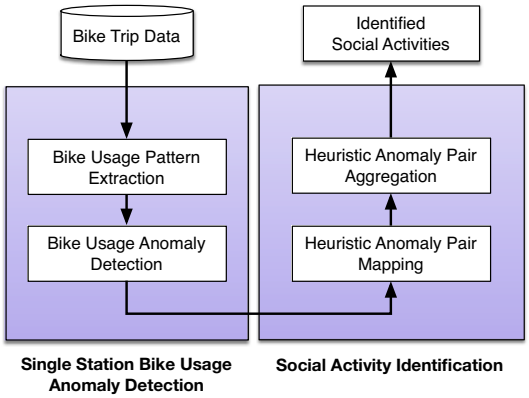

Figure 2: Framework overview.

- **Arrival Time**: the time when the corresponding bike is returned to a dock of the arrival station.
- **User Type**: being *casual* or *registered*, where a *casual* users refers to a rider who rents a bike with a temporal key, and a *registered* user is someone who has registered in the system and probably uses bikes on a regular basis.

B. Urban Activity Center Dataset

Urban Activity Centers often attract a large number of people for recreation, leisure, or tourism. For instance, the National Mall is an important UAC in Washington, D.C., featuring a large civic space and many iconic memorials. Large social activities, such as public concerts, civic art performances, and national festivals, are often held there.

The Metropolitan Washington Council of Governments (MWCOG) has defined and published a map of UACs covering major urban areas of the Washington, D.C.[2] We compile the *Urban Activity Center Dataset* based on the COG map, containing the names, the major social functions, and the locations of these UACs in Washington, D.C.

IV. Problem Description and Framework Overview

The objective of this work is to identify social activities in UACs using bike sharing open data. More specifically, given a set of bike trip data, our goal is to find out (1) the locations and durations of the social activities that are potentially relevant to bike usage anomalies observed in a certain stations, and (2) the characteristics of the bike riders related to the social activities, such as being tourists or local residents, and their common origins and destinations. Formally, we define the problem as follows.

Problem. *Given a set of bike trips R in a city, find the set of potentially relevant social activities $E = \{e\}$, where $e = (u, t_{start}, t_{end}, c)$ corresponds to a social activity occurring in the UAC u from time t_{start} to time t_{end}, and c denotes the characteristics of the corresponding bike riders.*

[2]http://www.mwcog.org/store/item.asp?PUBLICATION_ID=455

The connection between a single bike trip and a social activity is not significant nor straight-forward. However, with numerous bike trips observed during a long period of time, we can find out the bike usage patterns of stations, and capture the abnormal usage caused by social activities occurring near the stations. Based on this idea, we propose a framework to automatically identify social activities from bike trip data.

As shown in Figure 2, our framework consists of two phases, i.e., the *single station bike usage anomaly detection* phase, and the *social activity identification* phase. In the first phase, we extract bike usage patterns of each station from the historical bike trip data, as represented by the bike-arrival number and bike-departure number per hour. We then detect bike usage anomalies that are both *significant* compared to the station's usage in adjacent hours, and *unusual* compared to the stations daily usage routine. In the second phase, with the bike-arrival anomalies and bike-departure anomalies detected for all stations, we pair and aggregate anomalies from different stations to identify social activities, considering both spatial and temporal constraints using two heuristic rules.

V. Phase I: Single Station Bike Usage Anomaly Detection

In this phase, our objective is to detect bike usage anomalies of each station from the bike trip data. To this end, we first extract the normal bike usage patterns in each station, and then separate the *significant, unusual* bike usage from normal one. More specifically, we define the *bike usage* of a station as follows.

Definition. *The bike usage of a station is defined as the number of bikes arrived at a station in an hour, and the number of bikes departed from a station in an hour.*

Based on this definition, we extract the bike usage of each station from the bike trip data. We note that in general, urban human flow follows a *daily routine* [9], i.e., the human flow pattern in an area on one day is probably similar to some of the previous days. In this paper, we assume the bike usage in a station follows this daily routine. Therefore, instead of modeling the bike usage of a station as a continuous time series, we rearrange the bike usage data using a *day-hour* index to directly exploit the *daily routine* of bike usage.

For each bike station, we then detect bike usage anomalies that are both (1) *significant*, which means the bike arrival/departure number in a specific hour is much higher than that of the adjacent hours, and (2) *unusual*, which refers to an bike arrival/departure number that is generally unexpected to be observed in the same hour on other days. For instance, during the evening hours, normally only a few bike arrivals are expected at a station near a public park; however, when a concert is being held in that park from 7 p.m., the station might observe a *significant* number of bike arrivals at about 7 p.m. compared with the previous and next hours. Moreover, such an abnormal bike arrival number is also *unusual* compared with the bike arrival number at 7 p.m. on other days. We now describe the details of the two-step process in the following.

A. Bike Usage Pattern Extraction

In this step, we extract the bike usage patterns of each station from bike trip data. More specifically, we first construct two *bike usage data cubes*, $A = [a_{i,j,k}]$ and $D = [d_{i,j,k}]$, to denote the number of bikes arrived at station k at the jth hour of the ith day, and the number of bikes departed from station k in the jth hour of the ith day, respectively. We then fill the two data cubes with data from the bike trip dataset $R = \{r\}$, where $r = (s_{dep}, t_{dep}, s_{arr}, t_{arr})$ corresponds to a trip departed from station s_{dep} at time t_{dep} and arrived at s_{arr} at time t_{arr}. We calculate the day index i of t_{dep} and t_{arr} based on the time of the first trip respectively, as well as the hour index j.

Figure 3 shows the 30-day average of bike-arrival and bike-departure number in *Station #31219* near the National Mall from 06/01/2011 to 06/31/2011, where the shadows correspond to the standard deviation of different days. We can see clear daily routine of bike usage from this figure. First, for a specific hour, the number of bikes arrived at (or departed from) the station is generally stable. Second, the number of bike arrivals/departures in different hours vary significantly. For instance, the number is much higher in the day-time hour than in the mid-night or early-morning hour. Moreover, most bikes tend to arrive at this station in the morning and depart from there in the evening, which agrees to the human flow patterns of the station's area, i.e., the downtown center of Washington, D.C.

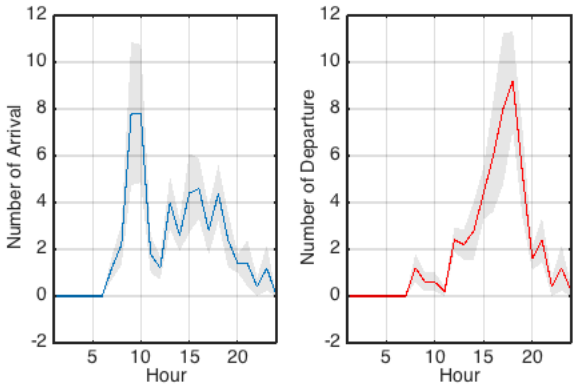

Figure 3: An illustrative example of the bike usage pattern in *Station #31219* near the National Mall.

B. Bike Usage Anomaly Detection

With the two bike usage data cubes extracted in the previous step, the goal of this step is to separate the significant, unusual bike usage from the normal one. The bike usage of station k can be denoted by $[A_k, D_k]$, while A_k is a matrix sliced from the bike-arrival data cube A along the station dimension, and D_k is a matrix sliced from the bike-departure data cube D along the station dimension. In the following, we describe the two-step method to detect bike-arrival anomalies from A_k, while the method for detecting bike-departure anomalies from D_k is the same.

1) Detecting significant bike usage: A significant bike usage means that the number of bike usage during a specific hour is much higher than the adjacent hours on the same day. We use a *sliding-window based method* to detect such elements from A_k. More specifically, let $A_k = [\boldsymbol{a}_{i*k}^T]$, where $\boldsymbol{a}_{i*k}$ denotes the vector of hourly bike-arrival number on the ith day at station k. We create a window of size 3 and slide it along $\boldsymbol{a}_{i*k}$ to find each central maxima that is larger than the values of its two neighboring hours, and has a value greater than a predefined threshold δ_n (e.g. the station's capacity). The detected significant bike-arrival elements of A_k is stored in a *set of anomaly candidate* $U'_A = \{u'_a\}$.

2) Selecting unusual bike usage from the candidates: An unusual bike usage means that the number of bike usage is generally unexpected to be observed in the same hour on other days. To select such unusual elements from the set of anomaly candidates, we first model the probability distribution of bike usage in the same hour on each day using a non-parametric model [17], and then filter unusual usage as the one with small probabilities. More specifically, let $A_k = [\boldsymbol{a}_{*jk}]$, where $\boldsymbol{a}_{*jk}$ is a vector of bike-arrival number at station k in the jth hour on all days. We first leverage kernel density estimation [18] to estimate the probability density function of each $\boldsymbol{a}_{*jk}$, and then calculate the probability for each candidate, $p(u'_a)$, using the corresponding density function. Finally, we select candidates with

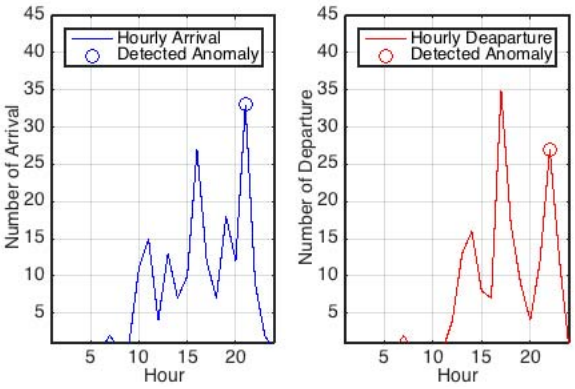

Figure 4: Bike usage anomalies detected on Independence Day of 2013 at *Station #31219* near the Nation Mall.

probability smaller than a given threshold δ_p as anomalies.

We perform the above-mentioned process for the bike-arrival and bike-departure data cubes to obtain a set of bike-arrival anomalies U_A and bike-departure anomalies U_D. In the next phase, we combine these two sets of anomalies together to identify the potential social activities.

VI. Phase II: Social Activity Identification

In this phase, we need to identify social activities from bike usage anomalies detected in the previous phase. Social activities are notable occurrences that involve a large number of participants and usually result in significant human flow within and around the activity venue [9]. In particular, before a social activity starts, a large crowd of people gather at the venue from other areas, while after the activity ends, these participants leave the activity venue. Such human flow patterns can be sensed by bike stations deployed within or around the Urban Activity Centers. For instance, Figure 4 shows the bike usage patterns on Independence Day (July 4th) of 2013 at *Station #31219* near the National Mall, where a fireworks show was held at 9 p.m. We can see a bike-arrival anomaly at 9 p.m. followed by a bike-departure anomaly at 10 p.m., corresponding to the large human flow gathering at and leaving from the National Mall for the specific moment. We note that the peaks in 3 p.m and 4 p.m. in the figures are not considered as anomalies, since they are not *unusual* compared to the daily routine of bike usage in the after-work rush-hours.

Based on the above observations, we propose a heuristic method to identify the corresponding social activities from the detected bike usage anomalies. First, we map bike-arrival and bike-departure anomalies into pairs to locate potentially relevant social activities in the temporal space. Then, we aggregate geographically close anomaly pairs into sets to identify the relevant social activities in the spatial space. In order to reduce the search space of possible spatial and temporal anomalies in pair mapping and aggregation, we build two heuristic rules for both steps respectively, as detailed in the following subsections. Finally, we extract the location, duration, and user characteristics from the identified social activities.

A. Heuristic Anomaly Pair Mapping

In this step, we iteratively select a bike-arrival anomaly and a bike-departure anomaly from the corresponding sets U_A and U_D to build a pair of anomalies to locate a potentially relevant social activity in the temporal space. Based on the intuition that a social activity may influence several bike stations around the activity venue, we derive a heuristic rule to guide the anomaly pairing procedure: the anomalies in a pair should occur in nearby stations within several hours range, and the bike-arrival anomaly should always occur before the bike-departure anomaly.

Formally, a bike-arrival anomaly u_a and a bike-departure anomaly u_d should obey the following criteria when considering them as an anomaly pair $p = (u_a, u_d)$.

- *Criterion 1: Nearby stations*

$$dist(u_a.k, u_d.k) < \delta_d$$

  where k is the corresponding station where the anomaly is detected, and δ_d is a distance threshold. We set the value of δ_d based on the average distance between bike stations.
- *Criterion 2: Short-range duration*

$$0 < (u_d.t - u_a.t) < \delta_t$$

  where t is the occurring time of the corresponding anomaly ($t = i * 24 + j$), and δ_t is a time threshold.

We note that if more than one candidate u_d for a u_a can meet the above-mentioned criteria, we select the one with the smallest geographic distance from u_a. We simply discard anomalies without adequate counterparts to form a pair. Finally, we obtain a set of anomaly pairs $P = \{p\}$.

B. Heuristic Anomaly Pair Aggregation

In this step, we aggregate geographically close anomaly pairs into subsets to identify the relevant social activities. Intuitively, when a social activity occurs in a UAC, the nearby stations may observe an anomaly pair during the activity hours. Therefore, we propose a heuristic rule to guide the aggregation of anomaly pairs: the anomaly pairs in a subset should correspond to stations that are geographically clustered, and their occurring time should have an overlap.

Formally, given the set of anomaly pairs P, the formed subset $P_i \subset P$ should meet the following criteria:

- *Criterion 1: Geographical cluster*
  $\forall p_1 = (u_{a1}, u_{d1}) \in P_i, p_2 = (u_{a2}, u_{d2}) \in P_i,$

$$dist(u_{a1}.k, u_{d2}.k) < \delta_d$$

  where u_{a1}, u_{a2} and u_{d1}, u_{d2} are the corresponding arrival and departure anomalies in a pair.

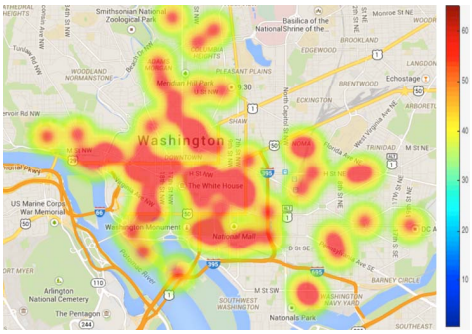

Figure 5: Heatmap of identified social activities in Washington, D.C. from 2012 to 2014.

- *Criterion 2: Temporal overlap*
  $\forall p_1 = (u_{a1}, u_{d1}) \in P_i, p_2 = (u_{a2}, u_{d2}) \in P_i,$

  $$\begin{cases} u_{d1}.t - u_{a2}.t > 0 \\ u_{d1}.t - u_{a2}.t \leq (u_{d1}.t - u_{a1}.t) + (u_{d2}.t - u_{a2}.t) \end{cases}$$

  This ensures that all the anomaly pairs in the subset share a couple of hours in common.

C. Social Activity Extraction

Finally, we obtain a set of social activities E by extracting key information from the aggregated anomaly subsets. More specifically, we extract a social activity $e = (u, t_{start}, t_{end}, c)$ from a subset of anomaly pair P_i, where:

- u is the UAC where the activity occurs. We obtain it by mapping the center location of the anomaly elements to a closest venue of UAC as defined in the dataset.
- t_{start} and t_{end} are the start and end hours, as determined by averaging the hours of arrival-anomalies and departure-anomalies.
- c is the user characteristics. By investigating the trip data details, we characterize (1) the bike user type, being casual or registered users, and (2) the common origins and destinations of users.

VII. Evaluation

In this section, based on a real-world bike sharing open data, we evaluate the performance of our framework by assessing its ability to identify social activities in UACs. In the following, we first describe the experiment setup, and then present the results of social activity identification. We also conduct two case studies to analyze the user characteristics.

A. Experiment Setup

We evaluate our framework by conducting experiments on social activity identification based on the *Bike Trip Dataset* in Washington, D.C. It contains bike trips between 203 bike stations in Washington, D.C from 2012 to 2014. The identified activities are bounded in areas defined in the *Urban Activity Center Dataset*, which contains 21 UACs in Washington, D.C.

We empirically set the parameters in our framework based on a series of experiments. Specifically, we set the usage anomaly significance threshold δ_n as the capacity of each bike station (which is 15 to 30 in general), and the distance threshold $\delta_d = 300m$ for searching geographically close stations, which is commonly adopted as an average distance between stations in many bike sharing systems [14]. We also set the *unusual* probability threshold $\delta_p = 3\%$ based on experiments. The maximum duration of an anomaly-pair, δ_t, is set to 3 hours, since most social activities, such as concerts and football games, usually last for only one or two hours.

We calculate the accuracy of the identification results to evaluate the proposed framework. First, we select a set of important UACs that have official Facebook or Twitter accounts, and compile a list of social activities occurred in these UACs in the three years based on the posts of these accounts. Then, we perform social activity identification in these UACs, and compare the results with the ground truth to compute the *precision* and *recall* of our framework. More specifically, for each identified activity, if an identified activity can be found in the ground truth list on the same day at the same UAC, we call it a hit. Based upon this, we define the *precision of identification* as:

$$precision = \frac{|\{\text{occurred activities}\}| \cap |\{\text{identified activities}\}|}{|\{\text{identified activities}\}|}$$

and the *recall of identification* as:

$$recall = \frac{|\{\text{occurred activities}\}| \cap |\{\text{identified activities}\}|}{|\{\text{occurred activities}\}|}$$

We also compare our framework with a baseline approach that considers only the spatial constraint in a simple way. For each UAC, the baseline detects bike usage anomalies from its nearby stations by applying a threshold on hourly bike arrival and departure number. However, in this way, the baseline approach is not able to capture the temporal variation and correlation of bike usage in stations.

B. Social Activity Identification Results

Figure 5 shows a heatmap of all the identified social activities in Washington, D.C. from 2012 to 2014. We can see that most social activities are found in the downtown areas, public parks, sports stadiums, and community centers. For instance, the National Mall, which receives approximately 24 million visitors each year, observes many social activities in the three years.

We select 10 important UACs to evaluate our framework. Table I presents the names, the major social functions, and some notable activities identified in these venues. The ground truth list contains 221 activities in these UACs. Meanwhile, we identify 237 activities using our framework, and 197 of them hit the ground truth. In summary,

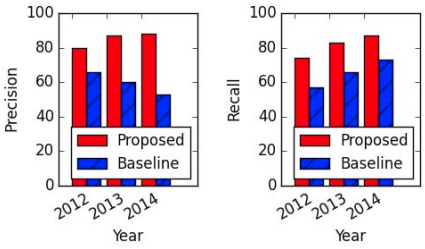

Figure 6: Yearly precision and recall of social activity identification using our framework and the baseline.

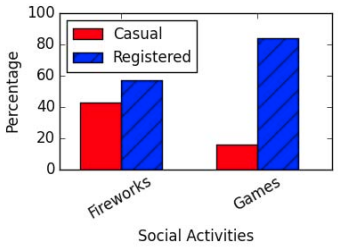

Figure 7: The bike user types of two social activities.

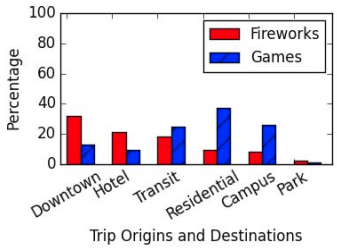

Figure 8: The bike trip origins and destinations of two social activities.

our method achieves a precision of 83% and a recall of 89%, respectively. In comparison, the baseline approach only achieves a precision of 61% and a recall of 63%, respectively.

We also present the yearly precision and recall for our framework and the baseline in Figure 6. We can see the recall of both our framework and the baseline increase over the years. This is probably due to the increasingly popularity of the Capital Bikeshare system among residents and tourists. As more and more people ride bikes to attend social activities, the human flow sensed by bike stations can be more significant and aligned with those activities. In addition, the precision of the baseline decreases over the years, which is attributable to the usage of a fixed threshold that can not adapt to the temporal variation when the overall bike usage increases yearly. In contrast, our framework considers both spatial and temporal constraints, and thus achieves better precision than the baseline.

C. Case Study on Bike User Characteristics

We analyze the characteristics of bike users related to two social activities, i.e. the *Independence Day Fireworks of 2014* in the National Mall, and the *Washington Nationals October Games of 2012* in the Nationals stadium. We focus on comparing the user types and bike trip origins/destinations by investigating the Bike Trip Dataset. First, we compare the percentage of user types in both activities (Figure 7). Then, we compare the area characteristics of the bike trip origins and destinations (Figure 8) by mapping the corresponding stations to the area functions using the open data from Washington, D.C. Zoning Map[3].

1) The Independence Day Fireworks of 2014: This famous fireworks show was held at around 9 p.m. on July 4th, 2014. It attracted hundreds of thousands of people to the National Mall and its surrounding areas. Figure 7 shows that 43% of the bike users found in the nearby stations are *casual* users, probably corresponding to a large population of tourists. Moreover, we find a wide range of bike trip origins and destinations (Figure 8), indicating the diverse user communities related to this social activity.

2) The Washington Nationals October Games of 2012: This game seasons attracted lots of baseball fans and observed a large volume of bike arrivals and departures around the Nationals stadium. Figure 7 shows that 84% of the bike users are registered users, most probably corresponding to local baseball fans. Moreover, most of the bike trip origins and destinations are local residential areas, university campus, and transit hubs to suburban areas (Figure 8), indicating a bias toward local residents.

VIII. Discussion

Our work may have the following limitations.

1) Data quality: Although bike trip data can be used to effectively identify social activities, the bike sharing user community is naturally biased. For instance, the 2013 Capital Bikeshare Survey [16] indicates that bike sharing users in Washington, D.C. tend to be considerably younger, highly educated, and relatively less affluent. Therefore, the social activities identified in this work have a bias towards the interests of this community. Nevertheless, we believe that by fusing more urban open data [19] from other means of transportation, our framework can be adapted to identify social activities attended by a broader audience.

2) Station capacity and bike availability: The number of bikes and docks in a station is limited. Therefore, when many people rent or return bikes in a station at the same time, the station will be exhausted or overflowed, and thus no bike rental or return can be handled. This makes it difficult to accurately estimate the scale of a social activity only via the bike usage data. To address this issue, we may consider integrating the LBSN data into our framework to help infer the scale of the crowds, as mentioned in [6].

IX. Conclusion

While enabling a green, healthy lifestyle, the emerging bike sharing systems worldwide generate a large volume of bike usage data, which provide invaluable resource for researchers to understand the social dynamics in the cities. In

[3]http://maps.dcoz.dc.gov/

Table I: List of the 10 important Urban Activity Centers and notable identified social activities

No.	UAC Name	Major Social Functions	Notable Identified Social Activities
1	National Mall	Memorial Park	Independence Day Fireworks
2	Nationals stadium	Baseball Stadium, Arena	Washington Nationals First Game
3	Verizon Center	Basketball Stadium, Arena	Taylor Swift Concert
4	RFK Stadium	Soccer Stadium	Nation's Football Classic
5	Capitol Hill	Landmark, Historic Site	A Capitol Fourth (Outdoor Concert)
6	Convention Center	Convention Center	USA Science & Engineering Festival
7	West Potomac Park	National Park	National Cherry Blossom Festival
8	Carnegie Library	Public Library, Museum	Mardi Gras Masquerade Gala
9	National Gallery of Art	Art Gallery, Museum	American Music Festival
10	The White House	Landmark, Historic Site	DC Emancipation Day Parades

this paper, we propose to identify social activities in Urban Activity Centers leveraging bike sharing open data. First, we detect bike usage anomalies that are both *significant* and *unusual* from the bike trip data. Then, we identify social activities from bike usage anomalies by considering both *spatial* and *temporal* constraints using two heuristic rules. Finally, we evaluate our framework using large-scale real-world bike trip data collected from Washington, D.C. The results show that our method can effectively identify social activities in major Urban Activity Centers.

In the future, we plan to incorporate more means of urban transportation to capture a wider variety of social dynamics in the cities. We also plan to evaluate our work in more cities such as New York City and Boston.

X. Acknowledgments

The authors would like to thank Vincent Gauthier, Xiaojuan Ma, Leye Wang and Wangsheng Zhang for their comments and inputs. This work was done when Longbiao Chen was visiting Institut Mine-Telecom, CNRS SAMOVAR, France.

References

[1] State Government of Victoria, "Melbourne 2030 - Planning for Sustainable Growth," 2002.

[2] D. Zhang, B. Guo, and Z. Yu, "The Emergence of Social and Community Intelligence," *Computer*, vol. 44, no. 7, pp. 21–28, 2011.

[3] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi GPS traces to social and community dynamics: A survey," *ACM CSUR*, vol. 46, no. 2, pp. 17–34, 2013.

[4] D. Zhang, B. Guo, and Z. Yu, "The emergence of social and community intelligence," *Computer*, vol. 44, no. 7, pp. 21–28, 2011.

[5] R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, and B. Guo, "Predicting activity attendance in event-based social networks: content, context and social influence," in *UbiComp 2014*, pp. 425–434.

[6] Y. Liang, J. Caverlee, Z. Cheng, and K. Y. Kamath, "How big is the crowd?: event and location based population modeling in social media," in *HyperText 2013*, pp. 99–108.

[7] Y. Zheng, L. Wang, R. Zhang, X. Xie, and W.-Y. Ma, "GeoLife: Managing and Understanding Your Past Life over Maps," in *MDM 2008*, pp. 211–212.

[8] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs," *IEEE TSMC*, vol. 45, no. 1, pp. 129–142, 2015.

[9] W. Zhang, G. Qi, G. Pan, H. Lu, S. Li, and Z. Wu, "City-scale Social Event Detection and Evaluation with Taxi Traces," *ACM TIST*, vol. 6, no. 3, 2015.

[10] G. Pan, G. Qi, W. Zhang, S. Li, Z. Wu, and L. T. Yang, "Trace analysis and mining for smart cities: issues, methods, and applications," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 120–126, 2013.

[11] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *UbiComp 2011*, pp. 89–98.

[12] P. DeMaio, "Bike-Sharing: History, Impacts, Models of Provision, and Future," *Journal of Public Transportation*, vol. 12, no. 4, pp. 41–56, 2009.

[13] L. Di Gaspero, A. Rendl, and T. Urli, "Constraint-Based Approaches for Balancing Bike Sharing Systems," *Lecture Notes in Computer Science*, vol. 8124, pp. 758–773, 2013.

[14] J. C. García-Palomares, J. Gutiérrez, and M. Latorre, "Optimizing the location of stations in bike-sharing programs: a GIS approach," *Applied Geography*, vol. 35, no. 1, pp. 235–246, 2012.

[15] Y. Zhao, L. Chen, C. Teng, S. Li, and G. Pan, "GreenBicycling: A Smartphone-Based Public Bicycle Sharing System for Healthy Life," in *CPSCom 2013*, pp. 1335–1340.

[16] Capital Bikeshare, "2013 Capital Bikeshare Member Survey Report," Washington, D.C., 2013.

[17] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric Model for Background Subtraction," in *ECCV 2000*, pp. 751–767.

[18] S. J. Sheather and M. C. Jones, "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, vol. 53, no. 3, pp. 683–690, 1991.

[19] L. Chen, D. Zhang, G. Pan, L. Wang, X. Ma, C. Chen, and S. Li, "Container throughput estimation leveraging ship GPS traces and open data," in *UbiComp 2014*, pp. 847–851.