Understanding Metropolitan Crowd Mobility via Mobile Cellular Accessing Data

HANCHENG CAO, Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China
JAGAN SANKARANARAYANAN, University of Maryland, College Park, MD
JIE FENG and YONG LI, Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China
HANAN SAMET, University of Maryland, College Park, MD

Understanding crowd mobility in a metropolitan area is extremely valuable for city planners and decision makers. However, crowd mobility is a relatively new area of research and has significant technical challenges: lack of large-scale fine-grained data, difficulties in large-scale trajectory processing, and issues with spatial resolution. In this article, we propose a novel approach for analyzing crowd mobility on a "city block" level. We first propose algorithms to detect homes, working places, and stay regions for individual user trajectories. Next, we propose a method for analyzing commute patterns and spatial correlation at a city block level. Using mobile cellular accessing trace data collected from users in Shanghai, we discover commute patterns, spatial correlation rules, as well as a hidden structure of the city based on crowd mobility analysis. Therefore, our proposed methods contribute to our understanding of human mobility in a large metropolitan area.

CCS Concepts: • **Social and professional topics → Geographic characteristics**; • **Information systems** → *Data mining*;

Additional Key Words and Phrases: Mobile data, human mobility, correlation detection, urban computing

ACM Reference format:
Hancheng Cao, Jagan Sankaranarayanan, Jie Feng, Yong Li, and Hanan Samet. 2019. Understanding Metropolitan Crowd Mobility via Mobile Cellular Accessing Data. *ACM Trans. Spatial Algorithms Syst.* 5, 2, Article 8 (July 2019), 18 pages.
https://doi.org/10.1145/3323345

Hancheng Cao is now with Stanford University.

This work was supported in part by the NSF under grants IIS-13-20791 and IIS-1816889, the National Key Research and Development Program of China under grant 2017YFE0112300, the National Nature Science Foundation of China under 61861136003, 61621091 and 61673237, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University—Tencent Joint Laboratory for Internet Innovation Technology.

Authors' addresses: H. Cao, Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China; email: hanchcao@stanford.edu; J. Sankaranarayanan, Department of Computer Science, University of Maryland (currently with Google Inc.); email: jagan@umiacs.umd.edu; J. Feng and Y. Li (corresponding author), Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China; emails: feng-j16@mails.tsinghua.edu.cn, liyong07@tsinghua.edu.cn; H. Samet, Department of Computer Science, University of Maryland; email: hjs@umiacs.umd.edu.

2374-0353/2019/07-ART8 $15.00
https://doi.org/10.1145/3323345

1 INTRODUCTION

Understanding human mobility benefits numerous applications in urban planning, traffic control, city management, and government decision-making [1–4]. Recent years have seen studies on human mobility using both traditional questionnaire-based sociology methods [5] and data-driven approaches [1, 6]. Problems such as single user mobility patterns and specific social group movement, e.g., tourists and students, have been widely studied [7–9]. These studies have focused on the regularity and daily motif of people. It has been found that individual human movement is a combination of periodic visits to a few important places such as home and a working place, plus occasional exploration of new places such as restaurants and parks [10, 11].

Different from single-user trajectory analysis and particular user group movement, on the other hand, general group movement, or crowd mobility problems provides deeper insights. Mobility viewer [12] is a visualization attempt on city level crowd flow. It provides a view of crowd flow based on the base station. Calabrese et al. [13] discovered a correlation between the location of a big event like a performance and the home locations of the event-goers. They found that local residents tend to participate in nearby events. Moreover, people living in different regions have different tastes of activities. For example, some neighborhoods prefer sports while other neighborhoods prefer music.

Despite these studies, there have been few studies on the general regularity of everyday crowd mobility in metropolitan areas. Mining regular crowd mobility patterns in metropolises is still an open research question. Cities are constantly on the move [3]. Everyday, at every moment, people go home, to work, shopping, or to entertainment by traveling from one block to another. Different blocks have completely different crowd commute patterns. A deeper understanding of city block level commute patterns contributes much to urban planning and smart transportation design [14]. Meanwhile, crowd mobility connects different regions of a city together. At the crowd level, people of similar living habits who roughly live or work in the same place generally share similar trajectories. Thus, certain regions are related by crowds, which form communities. Studying the correlations between blocks of the city, based on everyday crowd movement, can therefore help us understand human living habits and social structures and, in turn, contributes to better transportation system design and policy making.

Fine-grained crowd mobility analysis is challenging for three reasons. First of all, there is a lack of large-scale and fine-grained data. Traditional surveys [5] and GPS data [15] are limited in coverage while transportation data are biased and can only provide information on traffic instead of the metropolis's overall population. The characteristics of these data make crowd mobility analysis heavily depend on inference from limited and biased data to understand the global phenomenon, which yield unreliable results. Meanwhile, there are a few studies on crowd mobility using much more ubiquitous call detailed records (CDR) [1], yet CDR data are usually sparse in records as users are unlikely to make many phone calls in a single day, which is not suitable for the analysis of regular crowd mobility. Secondly, crowd mobility analysis requires careful selection of spatial resolution and division to aggregate crowds. Cellular towers partition the city solely through coverage and the coverage of one tower often cuts blocks into halves. Thus, crowd analysis on cellular tower regions is not meaningful for applications and suitable spatial resolution is needed. Furthermore, crowd mobility analysis requires trajectory modeling and processing techniques for sensible crowd analysis. It is tough to extract crowds from raw individual trajectories as everyone follows their own routes with random explorations, deviating from their usual trajectories, which makes it impossible to find groups of individuals sharing the same trajectories. How to extract semantic places (homes and working places, for example) and analyze sensible crowd mobility is therefore challenging.

Luckily, as the main function of a mobile phone shifts from the sparse use of phone calls/texts to much more frequent use of apps and Internet browsing, larger, finer-grained, and ubiquitous datasets become available. The resulting individual trajectories with higher sampling rates enable us to perform a direct metropolitan scale analysis on crowd mobility with fine granularity. Taking advantage of mobile cellular data accessing trace datasets collected from April 1 to April 7, 2016 on 0.85 million users in Shanghai, China, we make attempts to mine the everyday crowd mobility patterns on a city block level in a metropolitan area. To address the challenges, we propose a systematic pipeline of analyzing crowd mobility on a block level from cell phone trajectory data. Our contributions can be summarized as follows:

- We make use of a large-scale and fine-grained cellular tower accessing traces in a metropolitan area and analyzing everyday crowd mobility patterns on the citywide scale. We detect blocks from road networks and carry out crowd mobility analysis on a block level instead of at the cellular tower level, resulting in highly meaningful results for applications.
- We develop effective algorithms to detect home regions, working places, and stay regions from individual user trajectories, based on people's periodic visit patterns. We validate our algorithms via ground truth data labeled by volunteers.
- We visualize the distribution of home and working place, as well as commuting distance with block granularity on choropleth maps. Analyses discover the complex mixed functionality of Shanghai's city structure. By focusing on block correlation patterns caused by crowd mobility and with the aid of statistics and the community discovery method Infomap, we discover the hidden correlation rules and neighborhood structures of Shanghai. To the best of our knowledge, we are the first to do analyses on block commute patterns and spatial correlation from a crowd mobility perspective.

The rest of this article is organized as follows. Section 2 describes the utilized dataset and preprocessing procedure. Section 3 presents an overview of our crowd mobility analysis system, and related algorithms such as home/working place detection, stay region detection, and community discovery method. Section 4 carries out the block level everyday crowd mobility analysis. After reviewing related work in Section 5, we provide concluding remarks in Section 6.

2 DATASET, DEFINITION, AND PREPROCESSING

2.1 Mobile Cellular Data Accessing Trace

The trajectory data used in this study is mobile cellular data accessing traces collected by one of China's largest mobile operators, China Telecom, in Shanghai, one of the largest cities in China, from April 1 to April 7, 2016. As China's three major operators provide similar service, the users recorded in the dataset can be viewed as randomly sampled from the overall population. Whenever a user connects to a nearby base station via phone calls/traffic, the user's service information is recorded. Data are collected with the format of user ID, base station ID that the user gets access to, and the timestamp of the Internet connection. We define users' records as relation Raw Record R as follows.

Definition 1 (Raw Record R). R is a relation recording users' mobile cellular data accessing trace. A tuple in relation R is in the format of (u_i, b_i, t), where u_i represents the ID of the user, b_i represents the ID of the base station, and t represents the timestamp of the cellular tower access, meaning that user u_i accesses base station b_i at timestamp t.

The GPS information of all base stations are also available. We define relation Base Station Location L_b as follows.

Table 1. Cellular Data Information

Item	Value
Coverage	Shanghai
Number of Base Stations	8,573
Record Duration	April 1–7, 2016
Mean Number of Records per User	227.34
Total Record Numbers	193,115,587
Max Number of Records per User	10,012
Number of Users	849,439
Min Number of Records per User	1

Definition 2 (Base Station Location L_b). Base Station Location L_b is a relation recording the GPS information of each base station. A tuple in L_b is in the format of (b_i, x, y). b_i represents the ID of the base station while x and y represent the longitude and latitude of the base station.

Compared with GPS data that provides rather accurate longitude and latitude information of users, the spatial granularity of the cellular data accessing trace is lower as it only captures the ID of the base station that the user accesses. Yet base station level granularity is sufficient for crowd mobility analysis. In our dataset, cellular data accessing traces has been recorded on 8,573 base stations across Shanghai. As regulated by communication protocol, users mostly connect to the nearest base station when using the cellular network. Thus, we can determine the rough location of the users through the GPS information of the base station, whose coverage range in downtown Shanghai is about 200 meters to 500 meters while the suburban base station coverage is about 2,000 meters. In crowd mobility analysis, we care more about the approximate region where people are, rather than the exact location points of people. Furthermore, GPS data is seldom available for large group of people in a city as the user group is quite often limited and biased (e.g., taxi drivers). Therefore, GPS data are not suitable for studying general crowd movement and city zone features. On the other hand, cellular data accessing trace is ubiquitous and covers the entire population. Thus, it is the most ideal dataset to study region level crowd movement for now.

The dataset used in this work has 200 million records for 0.85 million users, with an average record number of 227 for 7 days, or 33 records per day, per user. Detailed statistics for our data is found in Table 1. We filter out users with record numbers less than 20 in the dataset to ensure that users' trajectories are well-sampled. The filtered data includes records for 0.75 million users. Compared with past work, our dataset is much larger and denser. It is ideal for crowd mobility analysis on urban blocks, enabling us to gain insights on a global scale with fine granularity.

2.2 City Block Division

Our mobile traffic accessing trace data is recorded on a cellular tower level. Making use of a Voronoi diagram, we obtained a division of the city where the center of each region is a cellular tower, based on which we can do analysis. However, urban subdivision based on a Voronoi diagram is not geographically meaningful enough, as a Voronoi cell usually goes through streets and divides the same block in halves, preventing further applications in urban planning.

Therefore, using the idea from Ref. [16], we subdivide a city into blocks based on the city's road network. First the city is divided into various small regions using the finest grained road networks through a raster-based model, where "0" stands for road segments and "1" stands for blank space. Then a dilation operation is performed so as to eliminate unnecessary details as lanes and overpassed roads. A connected component identification algorithm (e.g., see Ref. [17]) finds basic block units by clustering all consecutive "1" labeled grids. Finally, nearby block units located

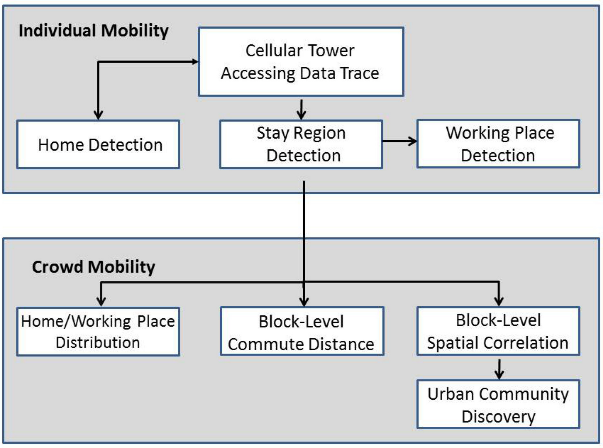

Individual Mobility
Cellular Tower Accessing Data Trace
Home Detection
Stay Region Detection
Working Place Detection
Crowd Mobility
Home/Working Place Distribution
Block-Level Commute Distance
Block-Level Spatial Correlation
Urban Community Discovery

Fig. 1. System architecture.

within major road networks (National Highway) are merged together to get blocks. In Shanghai, we are able to identify 17,056 basic block units and obtain 2,047 blocks for analysis.

The 2,047 blocks division partitions Shanghai into easily understood large blocks based on major road networks. Compared with the 17,056 basic block units partition, the 2,047 blocks division reduces errors when mapping Voronoi cells of cellular towers into blocks, which lays a solid foundation for our analysis and applications.

2.3 Data Preprocessing

The original geographical coordinates of the block division are in Baidu Map coordinates while cellular tower locations are in GPS coordinates. The two coordinates are not exactly the same and sometimes the same point when represented under the two coordinate systems have a deviation of a few hundred meters deviated from each other. Therefore, we first use the coordinates transformation API offered by the Baidu Map to ensure the two geographical data are under the same coordinate system.

Next, we join relations Raw Record R and Base Station Location L_b on b_{id}. Thus, we are able to get trajectory centers of all users for a week. We first define relation T_{BS}.

Definition 3 (Trajectory T_{BS}). T_{BS} is a relation recording the location centers of users at a certain timestamp. A tuple in T_{BS} is in the format of (u_i, t, x, y), where u_i represents the ID of users while x and y represent the longitude and latitude of the base station that the user accesses at timestamp t.

We further map relation T_{BS} into the obtained blocks, and define relation T_{Bl}.

Definition 4 (Trajectory T_{Bl}). T_{Bl} is a relation recording the blocks users access at a certain timestamp. A tuple in T_{Bl} is in the format of (u_i, t, b_i). u_i represents the ID of a user while b_i represents the block ID the user accesses at timestamp t.

3 SYSTEM ARCHITECTURE AND ALGORITHMS

The architecture of our crowd mobility analysis is shown in Figure 1. The input to our system is the individual cellular accessing trace and the output is the crowd mobility patterns. For better commute pattern analysis, we first label important places in the individual's trajectory as homes and working places. Based on these results, we aggregate them at the block level and carry out crowd mobility analysis.

3.1 Semantic Mining for Individual Trajectory

The preprocessed cellular data accessing trace is a series of individuals' trajectories. As revealed by many studies, individual trajectories show high periodicity both on weekdays and weekends [1, 18], as people tend to visit the same place at the same time slots of a day.

For crowd mobility analysis on the block level as identifying commute patterns and community structures, we need first to apply some processing procedures on the individual trajectories to filter out noises and extract semantic information. There are a few semantically important places for users like homes, working places, and their preferred restaurants, where they spend most of the time. Meanwhile, users pass by lots of less semantic places on their way. However, cellular data does not distinguish semantics. For example, a user may use the phone on the subway when going to work and be recorded at cellular tower A, despite the fact that the user is not really related to the activity happening in the region near A. We would like to extract semantic places in individual trajectories to make the analysis more meaningful.

Based on individual users' high periodicity in trajectories, it is feasible to detect important places of users in our week-long dataset. Labeling homes and working places is also possible. Now we first focus on methods to identify users' homes from trajectories. Next, we explain the stay point identification algorithm and a working place detection approach based on the stay point algorithm. Thanks to our fine granularity dataset, we are able to use simple modeling and achieve good results.

3.1.1 Home Detection Algorithm. Home is considered as the place where people rest at night. As a cell phone is normally inactive when the user is asleep, it is very likely that the first and last location appearing in a user's daily trajectory is recorded at the user's home. However, under some circumstances, a user may not use his/her phone directly after getting up or going to bed, but first uses the phone at other places such as on his way to work. We eliminate such a possibility by adding a temporal constraint that a "valid" candidate home location should be recorded either earlier than a threshold T_e or later than a threshold T_l. Furthermore, we drop all records from 12 a.m. to 4 a.m. in our dataset to avoid interference of unusual cellular network connection when the user is supposedly asleep.

Definition 5 (Candidate Home). We consider a location as *Candidate Home* of user i if all of the following criteria are met:

- The first point or the last point in the daily trajectory of user i.
- The record timestamp for the first point in the trajectory should be earlier than T_e, and the record timestamp for the last point in the trajectory should be later than T_l.
- The record timestamp for the first point in the dataset should not be a time when the user is supposedly asleep, which is automatically achieved as we first drop records with timestamp between 12 a.m. and 4 a.m.

For week-long data, we are able to get up to 14 candidate home locations for user i. Next, we label the most frequently appearing location in user i's *Candidate Home* lists as user i's home. As a block is a more meaningful spatial partition than a Voronoi cell, we use relation T_{Bl} as the algorithm's input. The output of the algorithm is the user's home block. The pseudo-code of the algorithm is shown in Algorithm 1.

3.1.2 Stay Region Detection Algorithm. As mentioned above, not all records in a user's trajectory are meaningful as a user may be recorded while passing a district. What's important in the trajectories are those stay regions where the user stays long enough.

Definition 6 (Stay Region S). A Stay Region S in a user's trajectory satisfies both temporal and spatial criteria:

ALGORITHM 1: Home Detection Algorithm.

```
Input:
    Block Level Trajectories $T_{Bl}$, number of users $n$, number of days $n_d$, early time threshold $T_e$,
    late time threshold threshold $T_l$
Output:
    User Home $H$
Initialize:
    $T_{Bl} \leftarrow T_{Bl}.drop(t \in [0\,am, 4\,am])$
    $H \leftarrow [\,]$
    for $i = 1$ to $n$ do
        Candidate_home: $H_C \leftarrow [\,]$
        for $j = 1$ to $n_d$ do
            $T_{Bl,i,j} \leftarrow T_{Bl}.select(u_i == i, t \in day\, j)$
            if $(T_{Bl,i,j}(1).t < T_e)$ then
                $H_C.append(T_{Bl,i,j}(1))$
            end if
            if $(T_{Bl,i,j}(length(T_{Bl,i,j})).t > T_l)$ then
                $H_C.append(T_{Bl,i,j}(length(T_{Bl,i,j})))$
            end if
        end for
        $H.append(i, mostcommon(H_C))$
    end for
```

—The user stays in S long enough, longer than temporal threshold T_0.
—The user does not move large distances in S. He/she should not leave the stay center farther than spatial threshold R_0.

Our proposed algorithm first compares two consecutive records in a user's base station level trajectory (extracted from T_{BS}). If the geographical distance between two base stations is lower than a spatial threshold R_0, then the two base stations are considered to be in candidate Stay Region S_C. We take the geographical mean of the two consecutive points as an estimated candidate S_C. Next, we compare S_C with the next record in the user's trajectory. If the geographical distance between the two points is larger than the spatial threshold R_0, which means the user moves a large distance to a distant region, then the algorithm checks if the candidate S_C meets the temporal criterion. If the user stays in the Candidate S_C longer than T_0, then the Candidate S_C is verified as a real stay region. The algorithm outputs the result and proceeds. If the temporal criterion is not met, then the algorithm moves on detecting the next stay center in the user's trajectory. On the other hand, if the geographical distance is smaller than spatial threshold R_0, then the weighted geographical mean of the new record point and the original candidate S_C center is calculated as the new candidate S_C center. The algorithm moves on detecting whether the next point in the trajectory satisfies the spatial constraint. If met, then the candidate S_C center is replaced by the weighted center of the original candidate S_C center result and the location of the new point. The merge for this candidate S_C center terminates when the new point fails to meet the spatial criterion. Finally, the algorithm checks if the candidate S_C center meets the temporal criteria. The pseudo-code of the algorithm is in Algorithm 2.

Note that our stay region detection algorithm is designed for trajectory data with a relatively high sampling rate. We assume the user's real origin-destination information is well-captured in the sampled trajectory as our cellular accessing data trace. On low-sampled trajectory data, this

ALGORITHM 2: Stay Region Detection Algorithm.

```
Input:
    $T_{BS}$, number of users $n$, spatial threshold $R_0$, temporal threshold $T_0$
Output:
    GPS coordinates of stay center: $S$
Initialize:
    $S \leftarrow [\,]$
    for $i = 1$ to $n$ do
        $T_i \leftarrow T_{BS}.select(u_i == i)$
        Candidate stay: $S_C \leftarrow T_i(1)$
        for $j = 2$ to $length(T_i)$ do
            if $(dist(S_C, T_i(j)) > R_0)$ then
                $t_s \leftarrow starttime(S_C)$, $t_e \leftarrow endtime(S_C)$
                if $(t_e - t_s > T_0)$ then
                    $S.append(i, S_C, t_s, t_e)$
                end if
            end if
            if $(dist(S_C, T_i(j)) < R_0)$ then
                $S_C \leftarrow weightedmean(S_C, T_i(j))$
            end if
        end for
    end for
```

algorithm could identify wrong stay regions if part of the real trajectory information is missing (for instance, if the data only captures that a user is at region A at 8 a.m. and 8 p.m., then our algorithm will identify A as a stay point, while in fact the user travels to various places far away between 8 a.m. and 8 p.m.).

Through the stay region detection algorithm, we are able to transform the input T_{BS} into stay center coordinates. We then map the GPS stay center into block level for further analysis.

In practice, we choose T_0 as 20 minutes, and R_0 as 400 meters in analyzing our dataset.

3.1.3 Working Place Detection Algorithm.

Definition 7 (Working Place W). W is considered the most frequently appearing location during a user's weekday trajectory; with constraint, the temporal duration of the stay includes the morning (9 a.m. to 11 a.m.) or the afternoon (2 p.m. to 4 p.m.).

We propose Algorithm 3. It inputs detected stay regions from Section 4.2 and outputs users' working places.

3.1.4 Detection Algorithm Evaluation. We evaluate our home and working place detection algorithms via ground truth data labeled by volunteers. We develop an interface visualizing a user's location as time changes. Volunteers can determine the home and working place of the user by viewing the trajectory of a user over time. We randomly selected 200 users' trajectories from our dataset, and asked 20 volunteers to label the home and working place location for each user. We compare the volunteer labeled ground truth data with the outputs of the proposed algorithms. The test results for home and working place detection are shown in Tables 2 and 3, respectively.

The precision for the home detection and working place detection algorithms are 100% and 89.4% (F1 score 0.94), respectively, while the recall rate for the home detection and working place detection algorithms are 89.5% and 86.1% (F1 score 0.88), respectively. The test shows that our

Table 2. Home Detection Diagnostic Test

	Prediction positive	Prediction negative
Condition positive	TP = 119	FN = 14
Condition negative	FP = 0	TN = 67

Table 3. Working Place Detection Diagnostic Test

	Prediction positive	Prediction negative
Condition positive	TP = 93	FN = 15
Condition negative	FP = 11	TN = 81

ALGORITHM 3: Working Place Detection Algorithm.

```
Input:
  Stay Block $S_B$, number of users $n$, WorkTime $T_W$
Output:
  Working Place $W$
Initialize:
  Candidate Working Place: $C_W \leftarrow S_B.select((t_e \textbf{ or } t_l) \in T_W)$
  for $i = 1$ to $n$ do
      $C_{W,i} = C_W.select(u_i == i)$
      $W.append(i, mostcommon(C_{W,i}))$
  end for
```

detection algorithms work well for mining semantics in trajectories. Note that labeling the working place for some users is difficult as the users' trajectories vastly vary from day to day, resulting in the relatively higher false negative in working place detection. Therefore, our detection algorithm proves to be satisfactory.

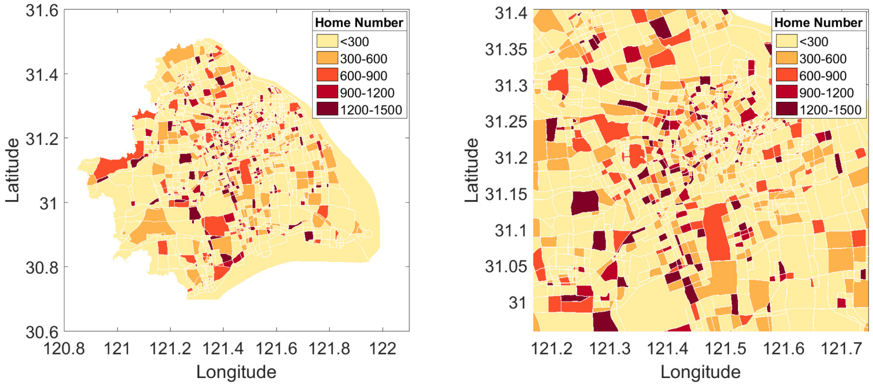
3.2 Community Discovery Algorithm

Through semantic mining of individual trajectories, we obtain stay regions, homes, and working places. By aggregating individuals by blocks, we can further analyze spatial correlation caused by crowd mobility and detect community structures in the city. The correlation between blocks can be modeled by a graph model, where nodes are blocks, and the weight associated with the edge between the two nodes represents the number of people moving between the blocks. Under the graph model, the problem of finding spatial clusters is equivalent to partitioning the correlation graph. We make use of the Infomap algorithm to carry out the clustering.

Infomap [19] is a classical algorithm for community detection. The idea of Infomap is to transform the graph partition problem into a minimum length coding problem. It uses the probability flow of random walks on a network as a proxy for information flows in the real system and decomposes the network into modules by compressing a description of the probability flow. Instead of manually assigning the cluster number, the Infomap algorithm can therefore automatically decide the cluster number through minimizing the probability flow.

4 BLOCK LEVEL CROWD MOBILITY ANALYSIS

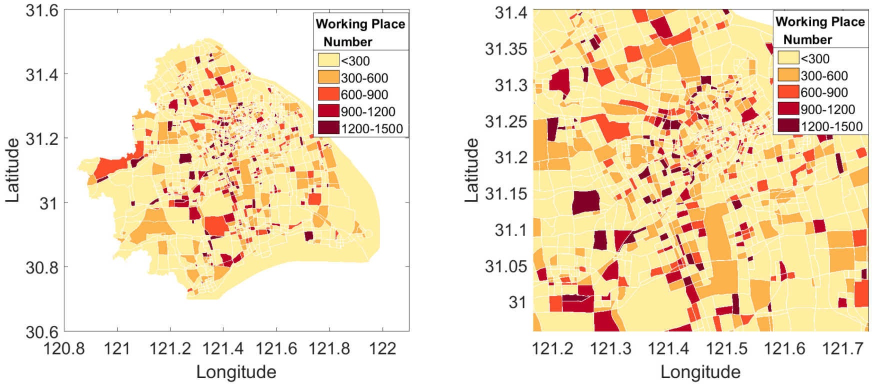
We have now extracted semantic information from raw cellular accessing trajectories corresponding to each individual's home block, working block, and stay regions. We can then aggregate the



(a) The Entire City

(b) City Center

Fig. 2. Home distribution in Shanghai. The color of each block represents the number of homes in the block. The darker the color, the higher the level of home distributions.



(a) The Entire City

(b) City Center

Fig. 3. Working place distribution in Shanghai. The color of each block represents the number of working places in the block. The darker the color, the higher level of working place distributions.

individual results on blocks and carry out crowd mobility analysis. We first look at homes and working place distributions in Shanghai, and then analyze commute distances of crowds. Finally, we focus on city blocks correlation both at the individual block and community levels.

4.1 Homes and Working Place Distributions

After obtaining individuals' homes and working places, we group the two relations by block ID and count the number of users in each block. Thus, we are able to get a distribution of home locations and working place locations in Shanghai, as visualized in Figures 2 and 3.

From the results, we can observe that the distribution of homes and working places in Shanghai is quite chaotic. Even neighboring blocks may have completely different distributions of homes and working places, suggesting that the city is mixed in functionality at the block level, especially in downtown. Downtown and suburban centers both show a high concentration of homes and working places, and the two distributions resemble each other. We use the distribution of home numbers in each block as Shanghai's home region feature, and the distribution of working place

Table 4. Metrics for Individual Commute Distance

Number of Users	427,041
Max. Distances	117.84 km
Mean Distance	5.394 km
Min. Distance	0

numbers in each block as the city's working place feature, and calculate the Pearson correlation value between the two 2,047-length feature vectors. Surprisingly, the correlation is as high as 0.9066, indicating a strong correlation between the distribution of homes and working places in the city. The functionality of city blocks in Shanghai, therefore, is quite vague. Under the 2,047 block-level partition, there is not a very clear difference in functionality such as residence or working zone (commercial, office, etc.). In fact, where there is a high concentration of population, there is also a high concentration of homes and working places. The function type of a district is dependent on time. For example, a district may represent residence function at night as people rest at home while it reflects a commercial function during working hours when people go to work in office buildings located in the block. Therefore, rather than identifying a district as of a static functional type (residence, working zone, entertainment, etc.) [20], our results highlight the necessity to dynamically identify the function of a district for better understanding of urban land use.

4.2 Commute Distances

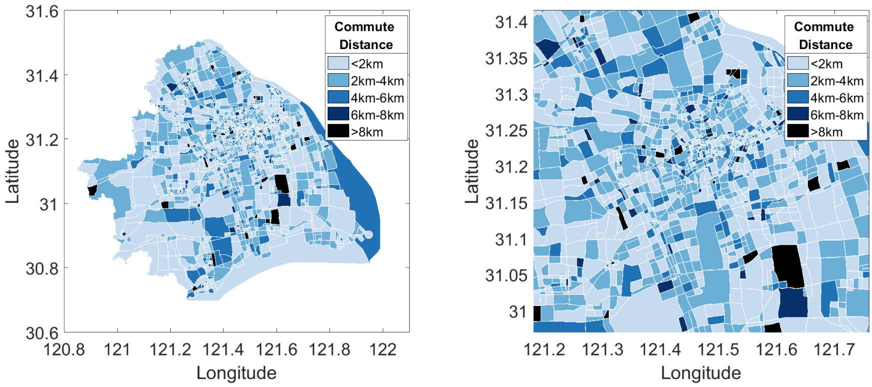
Commute distance, which measures the distance from home to work, is another important feature of human mobility. Based on home and working place detection results, we estimate the commute distance of individual users by calculating the geographic distance of home block center and working block center. Out of 849,434 users, 427,041 (50.3%) individuals' homes and working places are successfully detected.

The statistics of individual commute distance is shown in Table 4. On average, an individual in Shanghai travels 5.4km from home to work.

We can also track average crowd commute distance at the block level. We first group individual commute distances on their home block ID. Therefore, we can get a choropleth map, with the darkness of each color representing the average commute distance of the crowd who lives in the block, which enables us to learn the overall crowd commute patterns at the block level, as shown in Figure 4.

By aggregating individual commute distances on their working block ID, we are able to get another choropleth map, with the darkness of the color representing the average commute distance of the crowd working in the block, as shown in Figure 5.

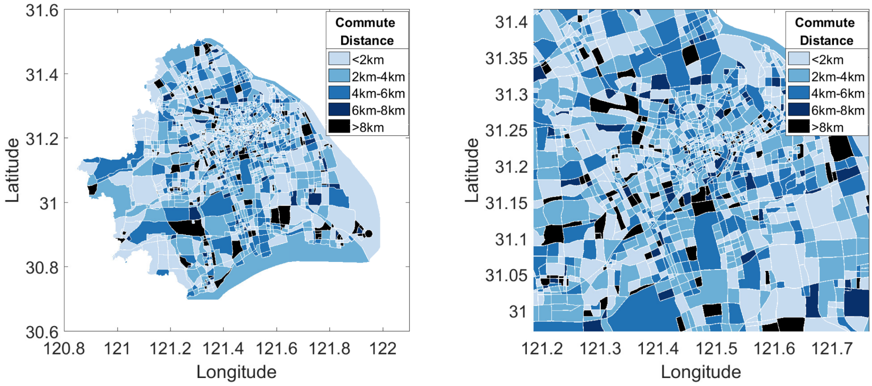
From the figures, we can observe distinct crowd mobility patterns for different blocks. In fact, even neighboring blocks can have very different commute distances. For home blocks, generally the color of downtown blocks is much lighter than suburban centers, suggesting that people living in downtown generally do not travel as large a distance for work while people living in the suburbs could travel long distances. The colors for the working block commute distance figures are much darker, particularly in the downtown regions. This suggests that people living outside downtown are going to the city center for work and cover a larger distance. As in Section 4.1, we use the distribution of mean home-to-work commute distance in each block as Shanghai's home commute distance feature, and the distribution of mean work-to-home commute distance in each block as Shanghai's working place commute distance feature, and calculate the Pearson



(a) The Entire City

(b) City Center

Fig. 4. Distribution of commute distance for people living in the block.



(a) The Entire City

(b) City Center

Fig. 5. Distribution of commute distance for people working in the block.

correlation value between the two 2,047-length feature vectors. The Pearson correlation value is 0.0493, suggesting little correlation between mean commute distance for people working in the block and mean commute distance for people living in the block. As commute distance is a good indicator of people's modes of transportation, which is further associated with people's living patterns and economic status, we can conclude that people living and working in the same block are very likely to be two different groups with different living habits and economic status.

4.3 Block Correlation Based on Crowd Mobility

People travel from one block to another throughout the day, making blocks correlate to each other through crowd flow. Our large-scale processed stay point trajectories of users make it possible for us to study such correlations between blocks.

4.3.1 Single Home/Working Place Correlation. Do people living in a block go to the same area to work? Do people who work in a block live in the same area? We can answer these questions by analyzing the single home/working place correlation pattern.

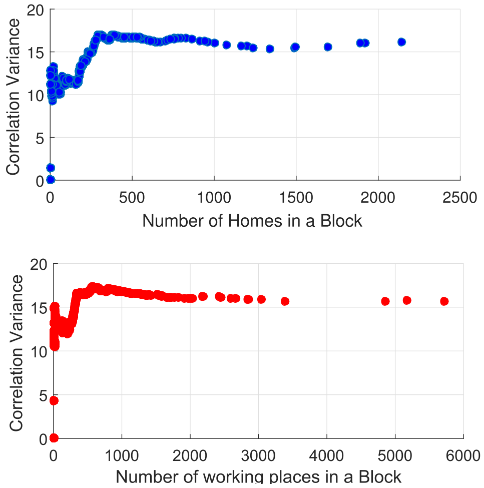


Fig. 6. Home/Working place correlation variance versus the number of home/working places.

To quantify the level of block correlation deviation, we first calculated the weighted spatial center of all correlated blocks of block i, denoted by C_i, as follows,

$$C_i = \frac{\sum_{j=1}^{N} (x_j, y_j) * n_j}{\sum_{j=1}^{N} (n_j)},$$

where x_j and y_j are the longitude and latitude centers of block j, and n_j is the number of people living in block i who work in block j. The total number of blocks in the city is denoted by N.

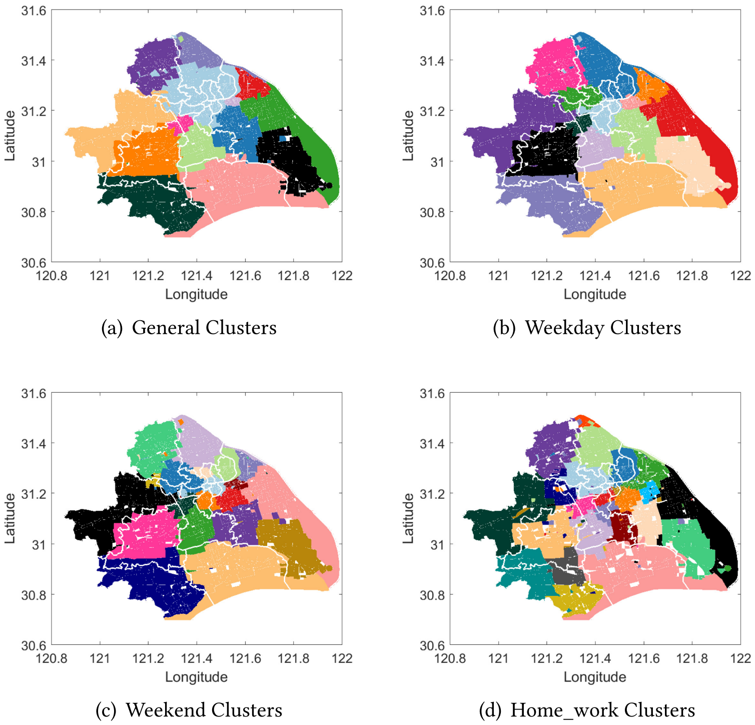
Next, we define the Correlation Variance of block i, denoted by Var_i, as follows,

$$Var_i = \sqrt{\frac{\sum_{j=1}^{N} (dist(C_i, (x_j, y_j)))^2 * n_j}{\sum_{j=1}^{N} (n_j)}},$$

which depicts the spatial deviation of correlated blocks. We adopt the Correlation Variance metric to both home and working place correlation.

We plot each blocks' spatial correlation deviation versus number of homes on the same figure, as shown in Figure 6. Surprisingly, we obtain the following result. The plot suggests that in Shanghai, there are no obvious spatial correlations as people living in the same place go to roughly the same place to work. The correlation between homes and working places are quite complex in that people living in the same block go to work in different places, and people working in the same place live in various residence districts. However, the complexity, or the variance of spatial correlation between home blocks and working places is in close relation with the number of people living or working in the block. In blocks with a small number of home/working places, the complexity of the spatial correlation remains almost the same. Above a certain threshold, the complexity grows quickly. As the home/working place numbers grow, however, the variance again remains stable.

4.3.2 Community Discovery. We now move onto a more global scale analysis. Is there a general correlation between groups of blocks in the city? Can we discover spatial clusters in the city where most users are active only inside the clusters? We adopt the Infomap algorithm described in Section 4.2.



(a) General Clusters

(b) Weekday Clusters

(c) Weekend Clusters

(d) Home_work Clusters

Fig. 7. Results of community discovery: Different clusters are shown with different colors. The white line represents political boundaries. (a) General cluster results based on all stay points, $Q = 0.37$, 14 clusters. (b) Weekday clusters based on all weekday stay points, $Q = 0.39$, 15 clusters. (c) Weekend clusters based on all weekend stay points, $Q = 0.49$, 20 clusters. (d) Home_work clusters based on detected user homes and working places, $Q = 0.50$, 24 clusters.

In community discovery, modularity Q is widely used to test the quality of detected clusters, where good partition has Q larger than 0.3. The larger Q, the better the partition result. We use Q to evaluate our results.

We first implement Infomap on an adjacency matrix of users' complete stay points, where an element e_{ij} represents the number of users who stay in both block i and block j in the week-long data. We end up with 14 clusters and $Q = 0.37$, indicating strong community characteristics. The good partition result suggests that crowd mobility in Shanghai has a community pattern, where users are active in certain zones. Although a single block can have correlations with lots of other blocks dotted in the city, the overall effect is that neighboring blocks form into the same clusters as nearby places have the most correlated crowd flows. We also observe that each crowd mobility cluster resembles a political boundary, as shown in Figure 7(a), yet clusters often go through boundaries. The similarity could be a result of different region's political positioning.

Next, we use the Infomap algorithm on an adjacency matrix of detected weekday and weekend stay centers. On weekdays, users normally cover larger distances, while during the weekend, people generally stay within their home regions to relax. The difference can be observed in the cluster results, as shown in Figure 7(b) and (c), where weekend correlation ends up with more clusters. For weekday and weekend partition, Q is 0.41 and 0.49, respectively, and shows a higher level of community pattern.

We finally detect the correlation between homes and working places. Compared with the correlations of all stay regions, home and working place correlation is even stonger and shows the greatest community pattern, resulting in the highest modularity ($Q = 0.50$). A visualization of cluster results is in Figure 7(d). We could also observe that in the cluster result of home and working place correlations, there exists some non-neighboring blocks in the same cluster, indicating that the crowd mobility patterns for homes and working places are somewhat less constrained by geographical distance constraints than other correlations as that of homes and shopping or homes and entertainment.

In conclusion, based on semantic mining on individual trajectory, we analyze home/working place distribution, commute patterns on block level, as well as spatial correlation patterns in this section, which uncovers hidden patterns in crowd mobility.

5 RELATED WORK

In the past decades, many researchers have explored the human mobility problem. Various datasets have been utilized in these works, e.g., GPS data such as taxi trajectories where the key is determining similarity [21] and minimizing its dis-similarity [22], geo-social records, and mobile phone data. Among these data, mobile phone data is an emerging data source with promising application prospects because of its long-term continuance and high coverage of the population.

The early studies of human mobility began with GPS trajectories. For example, Zheng has done much work in this field such as measuring the similarity of trajectories [23], inferring people's motion mode [15], mining interesting locations [24], detecting crowd flow [2], and so on. Other work includes extracting stay points from the trace [7], mining trajectory patterns [25], as well as even inferring the land usage [20]. These studies help us understand the human mobility based on the GPS trajectories. However, the fact that GPS devices consume too much energy limits their utility by precluding their use to sense the long-term behavior of a large population.

With the popularity of social networks, more and more people leave their location information when using services such as flight check-in and location sharing [26], which provide semantic texts for researchers to understand users' mobility behaviors. Parent et al. [27] summarized the semantic trajectories modeling and analyzing methods. Fan et al. [28] utilized the records of a search engine to detect the potential crowd. Cao et al. [18] studied how revisitation patterns of a place correlate with the place's function via semantic spatial temporal data. Based on the check-in information collected from a location-based online social network, Cranshaw et al. [3] tried to understand the dynamics of the city. Other works have dedicated to recognize user living pattern through check-in data [29, 30]. Because of its sparsity in temporal and spatial dimension, the social data can provide few details about the individual mobility. Thus, many mobility models based on the social network data are at the group-level mobility model [31, 32].

To our knowledge, mobile phone data with high coverage and easy-access for the operator is the best data source to analyze human mobility in both the individual and the aggregate level. Many researchers have done work in this area ranging from data preprocessing to pattern mining. The work of Gonzalez et al. [1] on understanding individual human mobility patterns is a classic. Further, Song et al. [6] analyzed the limitation of trajectory prediction and proposed a simple but effective prediction model. Isaacman et al. [10, 33] proposed a supervised system to identify important locations for the users and, further, built a simulation model to generate the anonymous trajectories. To better understand the human mobility mode, Gonzalez et al. [11, 34] defined motifs and proposed an interpretable model to simulate the human mobility. From the aggregate view, some researchers [16, 35] explored utilizing mobile phone data to estimate the population distribution and obtain competitive results compared with the traditional methods. Dong et al. [4] utilized mobile phone data to detect unusual crowd. Simini et al. [36] proposed a universal model to describe the migration among the US. Mobile phone data have been widely applied in the individual and aggregate level human mobility research, and achieved a great success.

In summary, our research differs from existing work as we focus on crowd mobility in a metropolitan area at the block granularity level. We are able to analyze block characteristics as commute patterns, correlation rules, and spatial network community based on crowd mobility. To the best of our knowledge, analyses on block commute patterns and spatial correlation from a crowd mobility perspective has not been put forward in the open literature. Our study offers a new direction for future crowd mobility analysis.

6 CONCLUSIONS

In this article, we mine everyday crowd mobility patterns on a city block level in a metropolitan area based on a dataset with 0.85 million users. We propose a systematic pipeline to analyze crowd mobility on the block level from cell phone trajectory data. By extensive analysis, we discover that the distribution of homes highly resembles that of working places, and that there is no correlation between people that live and work in the same block in terms of commute distance. Moreover, we discover the relationship between the complexity of home/working place block correlation and home/working place number in that block, and find hidden community structures caused by crowd mobility. In the future, we plan to focus more on temporal aspects of crowd mobility analysis to obtain insights into how people move and interact with one another. Other future work includes adding a spatial browsing capability [37–39] for the trajectories.

REFERENCES

[1] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nat.* 453, 7196 (2008), 779–782.

[2] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. 2016. DNN-based prediction model for spatio-temporal data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 92.

[3] Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman Sadeh. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city.

[4] Yuxiao Dong, Fabio Pinelli, Yiannis Gkoufas, Zubair Nabi, Francesco Calabrese, and Nitesh V. Chawla. 2015. Inferring unusual crowd events from mobile phone call detail records. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 474–492.

[5] Moshe Ben-Akiva and Michel Bierlaire. 1999. Discrete choice methods and their applications to short term travel decisions. In *Handbook of Transportation Science*. Springer, 5–33.

[6] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Sci.* 327, 5968 (2010), 1018–1021.

[7] Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Borriello. 2004. Extracting places from traces of locations. In *Proceedings of the 2nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*. ACM, 110–118.

[8] Sherif Akoush and Ahmed Sameh. 2007. Movement prediction using bayesian learning for neural networks. In *2007 Second International Conference on Systems and Networks Communications (ICSNC'07)*. IEEE, 6–6.

[9] Bob Mckercher and Gigi Lau. 2008. Movement patterns of tourists within a destination. *Tourism Geographies* 10, 3 (2008), 355–374.

[10] 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 *International Conference on Pervasive Computing*. Springer, 133–151.

[11] Christian M. Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C. González. 2013. Unravelling daily human mobility motifs. *J. R. Soc. Interface* 10, 84 (2013), 20130246.

[12] Yuxin Ma, Tao Lin, Zhendong Cao, Chen Li, Fei Wang, and Wei Chen. 2016. Mobility viewer: An Eulerian approach for studying urban crowd flow. *IEEE Trans. Intell. Transp. Syst.* 17, 9 (2016), 2627–2636.

[13] Francesco Calabrese, Francisco C. Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. 2010. The geography of taste: Analyzing cell-phone mobility and social events. In *International Conference on Pervasive Computing*. Springer, 22–37.

[14] Shan Jiang, Gaston A. Fiore, Yingxiang Yang, Joseph Ferreira Jr., Emilio Frazzoli, and Marta C. González. 2013. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM, 2.

[15] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *Proceedings of the 10th International Conference on Ubiquitous Computing*. ACM, 312–321.

[16] Fengli Xu, Pengyu Zhang, and Yong Li. 2016. Context-aware real-time population estimation for metropolis. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1064–1075.

[17] Hanan Samet and Markku Tamminen. 1986. An improved approach to connected component labeling of images. In *International Conference on Computer Vision And Pattern Recognition*, Vol. 318, 312.

[18] Hancheng Cao, Zhilong Chen, Fengli Xu, Yong Li, and Vassilis Kostakos. 2018. Revisitation in urban space vs. online: A comparison across POIs, websites, and smartphone apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 156.

[19] Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 4 (2008), 1118–1123.

[20] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 186–194.

[21] Gísli R. Hjaltason and Hanan Samet. 2000. Incremental Similarity Search in Multimedia Databases. Citeseer.

[22] Sarana Nutanong, Edwin H Jacox, and Hanan Samet. 2011. An incremental Hausdorff distance calculation algorithm. *Proceedings of the VLDB Endowment* 4, 8 (2011), 506–517.

[23] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 34.

[24] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 791–800.

[25] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 330–339.

[26] Hancheng Cao, Jie Feng, Yong Li, and Vassilis Kostakos. 2018. Uniqueness in the city: Urban morphology and location privacy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 62.

[27] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, et al. 2013. Semantic trajectories modeling and analysis. *ACM Comput. Surv. (CSUR)* 45, 4 (2013), 42.

[28] Zipei Fan, Xuan Song, and Ryosuke Shibasaki. 2014. CitySpectrum: A non-negative tensor factorization approach. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 213–223.

[29] Hancheng Cao, Fengli Xu, Jagan Sankaranarayanan, Yong Li, and Hanan Samet. 2019. Habit2vec: Trajectory semantic embedding for living pattern recognition in population. *IEEE Trans. Mob. Comput.* (2019).

[30] Fengli Xu, Tong Xia, Hancheng Cao, Yong Li, Funing Sun, and Fanchao Meng. 2018. Detecting popular temporal modes in population-scale unlabelled trajectory data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 46.

[31] Chao Zhang, Keyang Zhang, Quan Yuan, Luming Zhang, Tim Hanratty, and Jiawei Han. 2016. GMove: Group-level mobility modeling using geo-tagged social media. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1305–1314.

[32] Hongzhi Shi, Hancheng Cao, Xiangxin Zhou, Yong Li, Chao Zhang, Vassilis Kostakos, Funing Sun, and Fanchao Meng. 2019. Semantics-aware hidden Markov model for human mobility. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM.

[33] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. 2012. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*. ACM, 239–252.

[34] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. 2016. The TimeGeo modeling framework for urban motility without travel surveys. *Proceedings of the National Academy of Sciences* (2016), 201524261.

[35] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences* 111, 45 (2014), 15888–15893.

[36] Filippo Simini, Marta C. González, Amos Maritan, and Albert-László Barabási. 2012. A universal model for mobility and migration patterns. *Nat.* 484, 7392 (2012), 96–100.

[37] Frantisek Brabec and Hanan Samet. 2007. Client-based spatial browsing on the world wide web. *IEEE Internet Comput.* 11, 1 (2007), 52–59.

[38] Claudio Esperança and Hanan Samet. 2000. Experience with SAND-Tcl: A scripting tool for spatial databases. In *Proceedings of the 2000 Annual National Conference on Digital Government Research.* Digital Government Society of North America, 1–24.

[39] Hanan Samet, Houman Alborzi, František Brabec, Claudio Esperança, Gísli R. Hjaltason, Frank Morgan, and Egemen Tanin. 2003. Use of the SAND spatial browser for digital government applications. *Commun. ACM* 46, 1 (2003), 61–64.

Received December 2018; accepted March 2019