Urban-Scale Human Mobility Modeling With Multi-Source Urban Network Data

Desheng Zhang, *Member, IEEE*, Tian He, *Fellow, IEEE*, Fan Zhang, *Member, IEEE*, and Chengzhong Xu, *Member, IEEE*

***Abstract*—Expanding our knowledge about human mobility is essential for building efficient wireless protocols and mobile applications. Previous mobility studies have typically been built upon empirical single-source data (e.g., cellphone or transit data), which inevitably introduces a bias against residents not contributing this type of data, e.g., call detail records cannot be obtained from the residents without cellphone activities, and transit data cannot cover the residents who walk or ride private vehicles. To address this issue, we propose and implement a novel architecture mPat to explore human mobility using multi-source urban network data. A reference implementation of mPat was developed at an unprecedented scale upon the urban infrastructures of Shenzhen, China. The novelty and uniqueness of mPat lie in its three layers: 1) a data feed layer consisting of real-time data feeds from various urban networks with 24 thousand vehicles, 16 million smart cards, and 10 million cellphones; 2) a mobility abstraction layer exploring correlation and divergence among multi-source data to infer human mobility with a context-aware optimization model based on block coordinate decent; and 3) an application layer to improve urban efficiency based on the human mobility findings of the study. The evaluation shows that mPat achieves a 79% inference accuracy, and that its real-world application reduces passenger travel time by 36%.**

***Index Terms*—Urban networks, human mobility, network modeling, smart cities.**

I. Introduction

HUMAN mobility is of great importance for both the design and evaluation of mobile network protocols [1]. Recently, the human mobility study gains great attention, thanks to the ubiquity of human location tracking devices, e.g., onboard GPS devices [2], handheld GPS devices [3], cellphones [4], and urban transit systems including subway [5], bus [6], taxicab [7] and smart cards [8]. The data from these devices offer empirically-driven momentum to the human mobility study, and serve as a new powerful mobility microscope, which holds the potential to revolutionize the research on mobile networks, e.g., Wi-Fi access point or cell tower deployment, *ad hoc* networking, mobile network pricing and marketing.

Based on various empirical datasets, several human mobility study has been proposed [4], [7], [9]–[12] and captures human mobility to a certain degree, yet with a common drawback, i.e., *biased sampling*, which is demonstrated by two kinds of the most popular methods.

- **The cellphone data based research** is under the observation that a high penetration of cellphones implies that the cellphone users' mobility can be seen as a general proxy to all resident mobility. But based on our empirical results in Section II, the mobility (at cell tower levels) captured by cellphone data depends crucially on how often residents have cellphone activities, and thus has a bias against the residents who do not have cellphone activities during movements.
- **The urban transit data based research** is under the observation that the residents in big urban cities usually employ the urban transit (i.e., taxicab [7], bus [6] or subway [5]) for travel. But based on our empirical results in Section II, an approach based on one kind of the transit data (e.g., taxicab) has a bias against the residents using other kinds of transit (e.g., bus or subway). To our knowledge, there is no approach based on the data from all urban transit (i.e., including taxicab, bus and subway), but even for such a comprehensive approach, it still has a bias against the residents using private vehicles.

In short, we argue that almost all state-of-the-art theory and practice on human mobility have typically focused on *single-source* empirical data in isolation from one another. Essentially, they utilize a part of the urban residents who are involved with a specific empirical dataset as a *sample* for the entire urban population, which inevitably introduces a *sampling bias* for the uninvolved residents. The key reason for the single-source approaches is that past researchers have been severely constrained by the capability of urban infrastructures to collect and consolidate large scale data in a timely and low-cost fashion.

We argue, however, that now quick expansion of urban infrastructures has offered new research opportunities by real-time *multi-source* data without the sampling bias. In particular, the cellular networks can be consolidated with the urban transit networks, which are integrated with various sensors, communication devices and automatic fare collection devices. Therefore, such urban infrastructures are capable of capturing almost all traveling residents at urban scale, e.g., taxicab users, bus users, subway users and cellphone users, if these residents opt to participate under incentive and privacy-preserving mechanisms. We argue that these unprecedented comprehensive real-time multi-source data have the potential

Manuscript received January 8, 2017; revised July 1, 2017; accepted August 16, 2017; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor J. Xie. Date of publication March 19, 2018; date of current version April 16, 2018. This work was supported in part by the U.S. NSF under Grant CNS-1544887 and Grant CNS-1446640 and in part by the China 973 Program under 2015CB352400. A preliminary result of this work was published at [13]. *(Corresponding author: Desheng Zhang.)*

D. Zhang is with the Department of Computer Science, Rutgers University, New Brunswick, NJ 08901 USA (e-mail: d.z@rutgers.edu).

T. He is with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: tianhe@cs.umn.edu).

F. Zhang and C. Xu are with the Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China (e-mail: zhangfan@siat.ac.cn; cz.xu@siat.ac.cn).

Digital Object Identifier 10.1109/TNET.2018.2801598

to revolutionize the human mobility research. However, our knowledge on correlation, divergence and utilization of these multi-source data is extremely limited, because no previous work has been proposed to explore these issues.

In this work, to deepen our understanding on how these data can be used for the human mobility study, we propose an architecture called mPat, which improves the urban efficiency by uniquely analyzing and inferring the real-time Mobility PATtern based on the correlation and divergence among multi-source data in urban infrastructures, including cellular networks and transit networks. While the cellphone and transit data have been studied before for human mobility, they were exploited by separate researchers for different cities in isolation. Differently, mPat incorporates the underlying transit networks in the cellphone networks, allowing the fellow researchers to put human mobility into detailed transit contexts and to gain deeper insights. In particular, mPat answers an essential question: given an urban spatial partition (*e.g.*, zip code zones), how many residents are moving from one region (i.e., origin) to another (i.e., destination) independently of transportation modes (i.e., subways, buses or taxicabs) in real-time. Such fine-grained mobility including transit modes has not been investigated. Specifically, our contributions are as follows:

- To our knowledge, we conduct the first work to design a generic architecture mPat to analyze and infer human mobility, instead of sampling a particular group of residents. Based on several empirical datasets, we provide direct evidence for two facts: (i) the mobility analysis based on single-source data has a bias against the residents who were not or only partially involved in the data; (ii) correlating multi-source data has the potential to compensate for introduced biases. In mPat, we uniquely establish the multi-source data feeds through the urban infrastructures (*e.g.*, cellphone, subway, bus and taxicab networks) to obtain *mobility abstraction* for application designs. We implement mPat in Shenzhen, the most crowded city in China (17,150 people per KM^2).
- We establish a feeding mechanism for multi-source data feeds and collect the data as follows: (i) 400 million calling records for a 10.4 million user cellular network, (ii) 22 billion GPS and fare records for a 14,000 taxicab network, (iii) 1 billion GPS records for a 10,000 bus network, and (iv) 6 billion transaction records for 16 million smart cards about subway and bus fares. To our knowledge, the established feeds and datasets have by far the highest standard for the human mobility study in three aspects: the most detailed data including call detail records, fare records, GPS records; the largest resident coverage (95%) of 11 million residents; the most complete urban data including cellular, taxicab, bus and subway networks for the same city. More importantly, we have been sharing anonymous sample datasets for the benefit of research community [13].
- We take initial attempts to explore spatio-temporal correlations and divergence between multi-source data. We transparentize the heterogenous features of these data (in terms of scale, timeliness, granularity and privacy) to abstract the high-level human mobility knowledge, by which we are able to infer the human mobility. In particular, we design an optimization model based on block coordinate decent to alternative optimize models based on cellphone data and transit data until the result converges. Based on our 91-day evaluation, mPat infers the urban mobility with an average accuracy of 79%, and outperforms a statistical model by 61% and a single-source model by 42%.

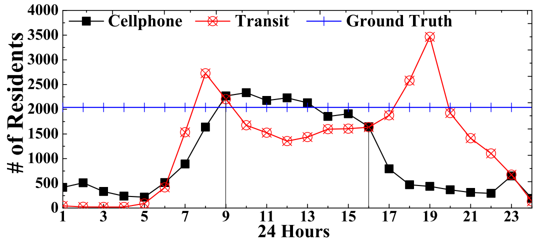


Fig. 1. Tracked residents to airport in 24 hours.

- To show mPat's real-world value, we propose, implement and evaluate a real-time transit service called IRT that delivers residents between urban regions with *high human mobility* yet *low transit mobility* inferred by mPat. Based on our 31-day experiment, IRT reduces resident travel times by 46% on average.

The paper is organized as follows. Section II gives the motivations. Section III presents the overview. Section IV introduces the feeds. Section V depicts the mobility model. Section VI evaluates mPat. Section VII presents our application, followed by related work and discussion in Sections VIII and IX. Section X concludes the paper.

II. Motivations

We introduce two design motivations based on the empirical data we collected in Shenzhen.

A. Drawback of Using Single-Source Data

The previous research on human mobility is relied on individual data generated from single-source feeds, *e.g.*, cellphone or transit networks, which typically leads to a biased sampling with inaccurate results. To provide such evidence, in Figure 1, we give the number of the residents going to the Shenzhen airport during 24 hours by four empirical single-source datasets, i.e., cellphone, taxicab, bus and subway. Details of datasets are given in Section IV. We aggregate residents using taxicabs, buses and subways as urban *transit residents* for a clear comparison. Further, we show the hourly average of the departure passenger number given by the airport, which serves as the lower bound of the ground truth.

During the regular daytime, i.e., from 9:00 to 16:00, we observed that the cellphone data track more residents than the transit data (i.e., the sum of the taxicab, bus and subway data). This is because some residents had phone activities but went to the airport with private vehicles, instead of the urban transit. They were tracked by the cellphone data but not the transit data. However, in the morning and evening rush hour, the transit data track more residents than cellphone data. This is because in the rush hour many residents who work and live near the airport will get on or off the urban

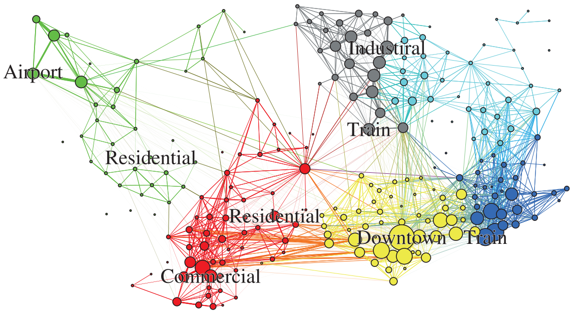


Fig. 2. Human mobility from cellphone data.

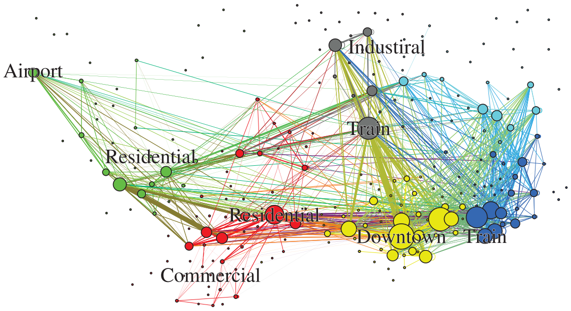


Fig. 3. Human mobility from transit data.

transit at the airport for daily commutes, and the most of these residents would not use their cellphones during the commutes. As a result, they were tracked by the transit data, but not the cellphone data. Noticeably, in some slots, e.g., 16:00, neither the cellphone data nor the transit data can track the residents close to the hourly average number of departure passengers given by the airport, proving that both of datasets cannot track the residents using private vehicles and lacking cellphone activities.

In short, we conclude that both the cellphone and transit data result in biased simples. These biased single-source data lead to drawbacks if used alone to infer the mobility, motivating us to explore multi-source data as follows.

B. Potential of Using Multi-Source Data

To investigate their potential, we render the daily mobility patterns obtained from cellphone and transit data in Shenzhen by Figures 2 and 3, where the size of a vertex indicates the resident number in an urban region; the thickness of an edge indicates the mobility volume. We found both similarity (e.g., the mobility to the airport, the train station, etc) and difference (e.g., the mobility to the commercial area) in these two figures, which motivates us to utilize the Pearson coefficients to explore their correlation in Figure 4. In the early morning, the correlation between cellphone and transit data is higher than other time. This is because the most residents use the urban transit (i.e., only taxicabs) for the mobility in the early morning, and they are more likely to use their cellphones during such unusual trips. But during the rush hour, the correlation is low due to the large number of commuting residents. The average value is 0.22, indicating that cellphone and transit data are not highly correlated, and have individual diversities.

C. Summary

Our results indicate that the mobility obtained from cellphone or transit data alone cannot unveil generic human mobility. But as shown in the previous section, they can be exploited together to compensate for the individual drawbacks due to their diversities. This motivates us to investigate a novel approach using multi-source data. But given the existence of a plethora of mobility models, instead of providing another model, we propose an architecture mPat from the lower-level data collection to the upper-level application design as follows.

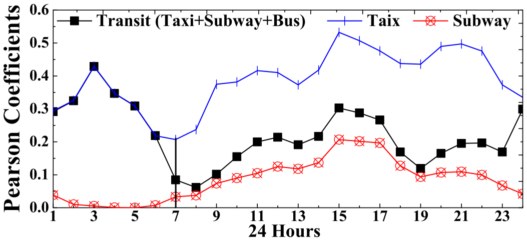


Fig. 4. Correlation of cellphone and other data.

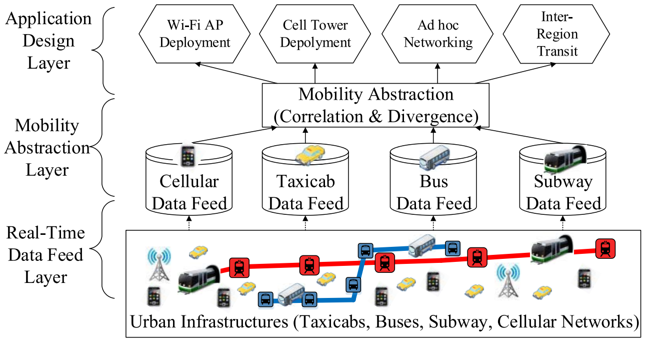


Fig. 5. mPat architecture.

III. mPat Architecture

In this section, we present mPat's architecture, which consists of three layers as in Figure 5.

Real-Time Data Feed Layer: ensures a secure and reliable feeding mechanism to establish multi-source data feeds through urban infrastructures in a privacy-preserving method. At a macro-level, mPat establishes the data feed for anonymous cellphones in the cellular networks; at a micro-level, mPat establishes the data feeds in the transit networks including taxicabs, buses and subways. The details are given in Section IV.

Mobility Abstraction Layer: transparentizes heterogeneous features in multi-source data to enable an effective mobility abstraction in an urban partition with extracted trips. mPat provides novel mobility analyses to unveil both correlation and divergence between residents using cellphones and residents using the urban transit, which are then exploited for a real-time mobility inference. The details are given in Sections V and VI.

Application Design Layer: bridges our mobility abstraction to real-world applications to improve urban efficiency, e.g., increasing urban transit ridership and reducing travel time for urban residents by uncovering urban region pairs with high human mobility yet low transit mobility. The details are given in Section VII.

Similar to the IP layer, i.e., the narrow waist of the Internet, the mobility abstraction layer essentially serves as a narrow waist of mPat, allowing a separation between data feeds and applications. Based on real-time input from data feeds, the mobility abstraction provides appropriate service interfaces for accurate rendering of human mobility, which are utilized by the applications to improve performance. mPat's three-level architecture suggests a horizontal view of building high-performance applications, but traditional stand-alone closed systems, e.g., cellular networks, do not have such capacities. The narrow waist allows fellow researchers to add more transit modes (e.g., bicycles) or applications without redesigning the whole architecture. Since cross-cutting design issues are better exposed when examined under a real-world implementation, we implement mPat based on Shenzhen infrastructures as follows.

IV. Real-Time Data Feed Layer

We have been collaborating with several Shenzhen government agencies and service providers, and establishing a reliable feeding mechanism that feeds mPat various data collected within Shenzhen infrastructures. This mechanism enables continuous capture and delivery of data from the service providers to mPat's data feed layer with end-to-end sub-second latency. We briefly introduce the established feeds in the layer as follows.

- **Cellphone Data Feed** is established for 10.4 million users in Shenzhen. The total records of data (including call detail records (CDR) among 17859 cell towers) are more than 5 million per day.
- **Taxicab Data Feed** is established through Shenzhen Transport Committee, to which all taxicab companies upload their taxicab status (GPS and occupancy) in real time by a cellular network used by all taxicabs in Shenzhen. The temporal granularity for this feed is extremely high, i.e., the uploading period is less than 30s. The daily size of all taxicab status data is 2 GB.
- **Subway Data Feed** is established by streaming entering and exiting records in smart card transactions. Such a feed accounts for more than 16 million smart cards, leading to 10 million daily transactions.
- **Bus Data Feed** consists of two parts: a GPS feed for all buses in real time (2 GB per day), and a transaction record feed from 16 million smart cards, generating 10,000 records per minute during the rush hour.

Our endeavor to consolidate the above feeds enables an extremely fine-grained mobility tracking that is unprecedented in terms of both quantity and quality. To facilitate mobility analyses based on real-time and historical data, we have stored the data from these feeds as in Figure 6.

Such big amounts of mobility data require significant efforts for the efficient storage and management. We utilize a 34 TB Hadoop Distributed File System (HDFS) on a cluster consisting of 11 nodes, each of which is equipped with 32 cores and 32 GB RAM. For daily management, we use the MapReduce based Pig and Hive. Pig is a high-level data-flow execution framework for parallel computation and Hive is a data warehouse infrastructure for data summarization and *ad hoc* querying.

Cellphone Dataset		Taxicab GPS Dataset	
Collection Period	10/01/13-3/1/14	Collection Period	01/01/12-3/1/14
Number of Users	10,432,246	Number of Taxis	14,453
Data Size	680 GB	Data Size	1.7 TB
Record Number	434,546,754	Record Number	22,439,795,235
Format		Format	
SIM ID	Date and Time	Plate Mumber	Date and Time
Cell Tower ID	Activities	Status	GPS Coordinates
Bus GPS Dataset		Smart Card for Subway & Bus	
Collection Period	01/01/13-3/1/14	Collection Period	07/01/11-3/1/14
Number of Vehicles	10,000	Number of Cards	16,000,000
Data Size	720 GB	Data Size	600 GB
Record Number	9,195,565,798	Record Number	6,212,660,742
Format		Format	
Plate Number	Date and Time	Card ID	Date and Time
Velocity	GPS Coordinates	Device ID	Station Name

Fig. 6. Datasets from real-time feeds.

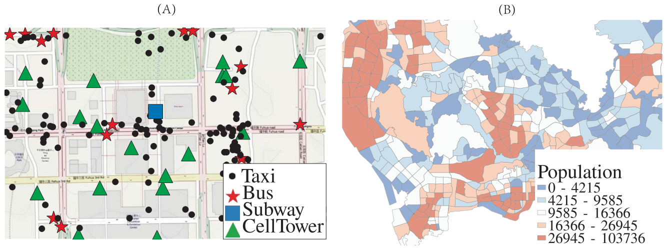


Fig. 7. (A) Granularity and (B) spatial partition.

Due to the extremely large size of our data, we found three main kinds of errant data. (i) Missing Data: *e.g.*, a taxicab's GPS data were not uploaded within a given time period. Such missing data are detected by monitoring the temporal consistence of incoming data for every data source, e.g., a taxicab. (ii) Duplicated Data: *e.g.*, the smart card datasets show two identical records for the same smart card. Such duplicated data are detected by comparing the timestamp of every record belonging to the same data source, e.g., the same smart card. (iii) Data with Logical Errors: *e.g.*, GPS coordinates show that a vehicle is off the road. Such data with logical errors are detected later when we analyze the data. The above errors may result from various reasons, *e.g.*, hardware malfunctions, software issues, and communications.

To address the above errors, for all incoming data, we first filter out the duplicated records and the records with missing or errant attributes. Then we correct the obvious numerical errors by various known contexts. We next store the data by dates and categories. Finally we compare the temporal consistence of the data to detect the missing records. Admittedly, the missing or filtered out data (which accounted for 11% of the total data) may impact the performance of our later analyses, but given the long time period, we believe we are still be able to provide insightful analyses in Section V as follows.

V. Mobility Abstraction Layer

We first describe two building blocks of our study, then study the correlation and divergence between the individual mobility, and finally provide the online mobility inference.

A. Building Blocks

1) Trip Extraction: We extract our basic mobility unit, i.e., *trip*, from the cellphone and transit data. For transit users, the trip extraction is straightforward because (i) for taxicab users, the Origins and Destinations (indicated as OD) are given by taxicab status records with GPS coordinates; (ii) for subway and bus users, the origins and destinations are given by the transaction records of smart cards and bus status records with GPS coordinates. But for cellphone users, it is more complicated because the cellphone data describe the mobile trace of a user by a sequence of cell towers with GPS coordinates, without the specific OD to define trips. Various methods have been proposed to divide a continuous trace into different trips based on the geometric feature of the trace. In this work, we focus on the mobility pattern inference, instead of dividing mobile traces. Therefore, we utilize one of the state-of-the-art methods [7] to obtain the trips based on the trace. In short, this method utilizes a graph theory concept called stretch factor to find several anchor points on a continuous trace as alternative origins and destinations, thus identifying the trips on the trace. Although our dataset includes records from several million cellphone users, the average number of cellphone records for every user is fewer than 10 per day. Our experiment indicates this method is scalable in cellphone record processing. With the obtained trips, we map all mobility into a spatial partition as follows.

2) Spatial Partition: The spatial partition is dependent on spatial granularity of collected data, which is shown in Figure 7(A). The ability to accommodate various levels of granularity means that mPat is not tied to a specific spatial partition. In other words, mPat works under a various range of spatial partitions with different granularity, as later indicated by our evaluation in Section VI. In our reference implementation, we aim to use a logical spatial partition on the Shenzhen urban area to study the spatial and temporal features of extracted trips in region levels. Some previous models use data driven methods to perform clustering algorithms (e.g., K-means) on digitally logged locations of interest for a particular group of residents, e.g., taxicab passengers [12]. But it is not suitable for our analysis of all urban residents. Further, some fixed grid based partitions are also popular, but such physical partitions lack real-world logical meaning. We argue that most urban areas have their own logical spatial partitions, *e.g.*, Zip+4 area partition, and they typically have a sophisticated logical meaning. Thus, we utilize an administrative region partition that divides Shenzhen urban area of 1,991 KM^2 into 496 administrative regions, based on geographical and residential features, as in Figure 7(B) in which the color of regions indicates the population density.

B. Offline Human Mobility Analysis

We explore spatio-temporal correlation and divergence between the mobility (in terms of trips) obtained by cellphone and transit data. Such correlation and divergence serve as the empirical guidelines for our later inference.

1) Spatial Correlation: We systemically study the mobility based on cellphone and transit data in terms of two key spatial components of a trip: the length and the OD.

To investigate the lengths of trips, we plot their distribution based on cellphone and transit data in Figure 8(A). We observed that the proportion of the cellphone trips shorter than 1KM or larger than 35KM is much higher than that of the transit trips. Further, the proportion of the cellphone trips with lengths from 1KM to 35KM is slightly lower than that of the transit trips. These observations make sense in the real world. This is because residents typically (i) walk for a trip shorter than 1KM, (ii) take personal vehicles or long-distance transit services for a trip longer than 35KM, and (iii) use either urban transit or other transportation for a trip between 1KM and 35KM. It suggests that cellphone data track more residents with short or long trips, while the transit data track more residents with medium-length trips.

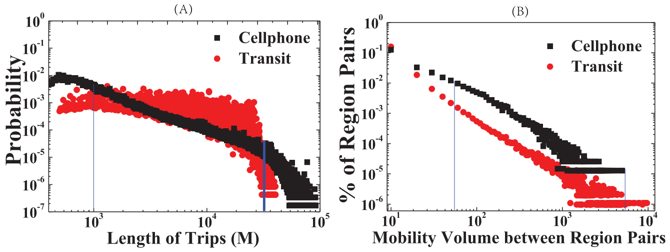


Fig. 8. (A) Length and (B) OD distribution.

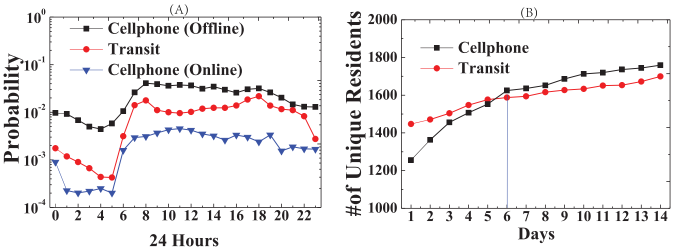


Fig. 9. (A) Temporal distribution and (B) accumulation.

To investigate trips' OD, we study volume distribution of trips among 496 × 496 region pairs (i.e., OD combinations) based on their ODs, as in Figure 8(B). For the most region pairs, the number of cellphone trips is fewer than 50 in a day, while among certain region pairs, the number of trips is much higher, e.g., 5,000. In terms of power-law distributions, the exponent of cellphone trips is smaller than the exponent of transit trips. It indicates that cellphone trips are spatially distributed more evenly than transit trips. This is because cellphone data track residents between almost all region pairs, but transit data only track residents between region pairs with public transit. It suggests that cellphone data are more effective when used to track the residents in terms of the OD diversity.

2) Temporal Correlation: We explore the temporal correlation between cellphone and transit trips to validate whether the cellphone data temporally outweigh the transit data for human mobility analyses. We plot the probability of mobility (i.e., a trip occurs for a resident) during 24 hours in Figure 9(A). This probability is obtained by dividing the number of trips (a resident can only be on one trip at most in a given moment) by the total Shenzhen population. We compare the probability based on the transit and cellphone data online (by one-hour slots). We found that in every hour, the probability of transit trips is higher than the probability of cellphone trips. This is because a cellphone resident is considered as being on a trip, only if this resident has cellphone activities involving different cellphone towers. As a result, this suggests that the transit data can track more trips than cellphone data online.

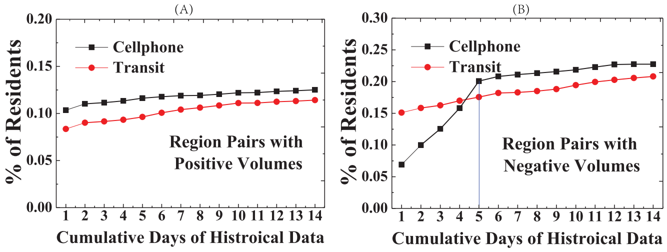


Fig. 10. (A) Divergence 1 and (B) Divergence 2.

The ineffectiveness of the cellphone data for online analyses is because many cellphone users did not have cellphone activities involving different cell towers during trips, i.e., the temporal sparseness. But we wonder if we cumulatively use both the historical and real-time data from cellphone users, can we track more trips or residents using the cellphone data, and if so, how many days of historical data we have to use. In fact, certain particular trips for the same resident have a highly-repeatable temporal pattern, e.g., daily commutes. Thus, if a resident did not have cellphone activities for a trip on one day, s/he may have cellphone activities for the same trip on other days. In Figure 9(B), we track the cumulative number of unique residents in trips by both cellphone and transit data between two representative regions from a residential region to an industrial park (e.g., daily commutes) during a one-hour slot 7:30AM-8:30AM. We found that increasing the use of the historical data leads to an increase in the number of residents on trips tracked by the cellphone or transit data. But when the historical data period used is longer than 6 days, the cellphone data track more residents than the transit data. This suggests that the cellphone data can be used for the online analyses, but they have to be used together with the historical data due to the temporal sparseness of cellphone activities.

3) Divergence: We also study another important metric: divergence which is defined as the difference between the numbers of residents tracked by the online cellphone and transit data. This difference offers insight about how much historical cellphone data we have to use together with the online cellphone data to track the actual number of residents. Theoretically, we can use all historical data, but that involves prolonged processing time, and may not be suitable for the applications requiring real-time inference.

To obtain such a divergence, we first individually obtain cellphone and transit trips based on the online data. Then, we deduct the transit trips from the cellphone trips to obtain the average divergence in region-pair levels. After the deduction, the trip volume for any region pairs could be positive or negative: the positive volume indicates the partial cellphone trips of the residents having cellphone activities but not using urban transit, whereas the negative volume indicates the partial transit trips of the residents lacking cellphone activities but using urban transit. We study the effectiveness of the cellphone or transit data to track residents on these two kinds of region pairs by two divergences in Figures 10(A) and 10(B).

For the region pairs with positive volumes (accounting for 31% of all region pairs), Figure 10(A) gives the percentage of residents in trips cumulatively tracked in 14 days among all the residents in Shenzhen. We found that for these region pairs, the effectiveness of cumulative tracking for both transit and cellphone data are fairly stable. After the first day among 14 days, the cumulative tracking for both the cellphone and transit data cannot find many new residents among these region pairs. This is because among these region pairs (usually far away from each other with the stable transit demand), the daily cellphone activities are intensive, so only few days of cellphone data may capture the majority of the cellphone users. This suggests that for region pairs with positive volumes, the mobility inference based on a short period of historical data may be sufficient, because a long period of the historical data can provide only few additional residents.

Figure 10(B) shows a similar plot to that in Figure 10(A), except that Figure 10(B) is for the region pairs with negative volumes (accounting for 69% of all region pairs). We found that the cumulative tracking based on the cellphone data seems to be more effective on a few earlier days (i.e., fewer than 5 days); whereas the cumulative tracking based on the transit data is not so effective, i.e., a 14-day historical dataset can track only a limited number of new residents. An important phenomenon we found is that the cellphone data-based cumulative tracking became less effective (as shown by the slope of the curve) after the total residents tracked by the cellphone data are close to or more than the total residents tracked by the transit data. One explanation for such an empirical result is that all transit residents may use the cellphone at least once during the five days of the transit, so the cumulative cellphone data can track almost all the transit residents within a few days along with additional private car users. This phenomenon serves as a key design guideline for our later mobility inference to balance the inference accuracy and the utilization of the historical data.

4) Remark: In this subsection, we empirically analyze both the correlation and divergence between the trips obtained by cellphone and transit data, leading to a few insights for our mobility inference design. In particular, (i) the historical cellphone data are more effective than the transit data in various spatial metrics as in Section V-B.1; (ii) the online cellphone data, due to their temporal spareness, have to be used together with the historical cellphone data to outweigh the transit data, as in Section V-B.2; and (iii) the effectiveness of the historical cellphone data is different for various region pairs, and such effectiveness can be indicated by the mobility divergence between the cellphone and transit data as in Section V-B.3.

C. Online Human Mobility Inference

We infer real-time online human mobility patterns based on the insights obtained in the last subsection. Such real-time mobility indicates how many residents are traveling between urban regions in a spatial partition. We propose a concept called a *mobility graph*. For a time slot τ, the mobility graph G is a directed graph where (i) a vertex indicates one region in a given spatial partition; (ii) an edge between two vertices indicates mobility between two associated regions; (iii) a weight on an edge indicates real-time mobility volume during the slot τ. As follows, we introduce how to infer such a mobility graph G for all the residents, based on the correlation and divergency between (i) the mobility graph G^c for the residents with cellphone activities in the previous slot, and (ii) the mobility graph G^t for the residents in the transit systems at the end of the previous slot.

1) Overview: There are two methods to infer G to cover all residents' mobility. First, G can be obtained by combining G^c and $G^{\bar{c}}$, i.e., combining the mobility patterns for the resident groups with and without cellphone activities at the end of the previous slot. Second, G can be obtained by combining G^t and $G^{\bar{t}}$, i.e., combining the mobility patterns for the resident groups using urban transit or private vehicles, respectively, in the previous slot. But in reality, neither of two methods will work since the current infrastructures can track neither the residents lacking cellphone activities nor the residents using private vehicles. Accordingly, in this work, we uniquely combine the above two methods together, based on our previous analyses. We aim to design an iterative optimization process where we alternatively update $G^{\bar{c}}$ and $G^{\bar{t}}$ until $G^c + G^{\bar{c}}$ is close to $G^t + G^{\bar{t}}$ (+ is the combination of these two mobility graphs).

- We first obtain G^c for the residents with cellphone activities based on the real-time online cellphone data (the details are given in the next subsection). We then estimate $G^{\bar{c}}$ for the residents without cellphone activities based on the historical cellphone data. This is built upon the observation that the residents who were not captured by the cellphone data today may be captured by the historical data before due to repeatable daily travel patterns as shown in Figure 9(B). Such an estimation based on historical data is challenging, because we could easily overestimate or underestimate $G^{\bar{c}}$ due to lack of knowledge on when to stop the iterative estimation.
- Alternatively, we can first obtain G^t for the residents with urban transit activities based on the real-time online transit data. Based on G^t, we can infer $G^{\bar{t}}$ by statistical methods with an assumption that the number of passengers using private vehicles traveling between two regions is proportional to the number of the passengers using public transit, i.e., $G^{\bar{t}} = \mathbf{W} \times G^t$ where $\mathbf{W}$ an unknown matrix indicating a relationship between $G^{\bar{t}}$ and G^t. $\mathbf{W}$ can be inferred by census data statistically but an accurate $\mathbf{W}$ is challenging to obtain.
- As a result, we have one way to infer G by $G^c + G^{\bar{c}}$ where $G^{\bar{c}}$ is inaccurate; whereas we have the other way to infer G by $G^t + G^{\bar{t}}$ where $G^{\bar{t}}$ is inaccurate. We can gradually update $G^{\bar{c}}$ or $G^{\bar{t}}$ but it is challenging to know when the updating is sufficient. In this paper, we argue that the updating of $G^{\bar{c}}$ and $G^{\bar{t}}$ is becoming sufficient if $G^c + G^{\bar{c}}$ is closer to $G^t + G^{\bar{t}}$ because both of them should be equal to G.

In this paper, we develop a technique based on Block Coordinate Decent [14] to alternatively update $G^{\bar{c}}$ and $G^{\bar{t}}$. Essentially, we use $G^t + G^{\bar{t}}$ for the residents using urban transit and private vehicles as a threshold to update $G^c + G^{\bar{c}}$, and vice versa. The rationale behind this condition is based on our observation of the mobility divergence in Figures 10(A) and 10(B) in Section V-B.3. For the region pairs with positive divergence volumes (i.e., the region pairs where the real-time cellphone mobility is more than the real-time transit mobility), the historical cellphone data based inference can find few new residents compared to the residents they already found as in Figure 10(A). However, for the region pairs with negative divergence volumes (i.e., region pairs where the real-time cellphone mobility is less than the real-time transit mobility), the historical cellphone data based inference can find many new residents in recent few days, but they become much less effective after the obtained cellphone mobility $G^{\bar{c}}$ plus G^c covers the mobility captured by the transit data, as in Figure 10(B).

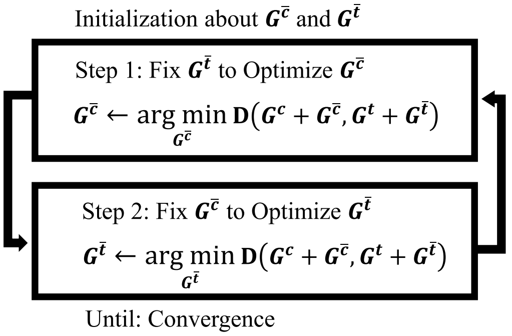

Initialization about $G^{\bar{c}}$ and $G^{\bar{t}}$
Step 1: Fix $G^{\bar{t}}$ to Optimize $G^{\bar{c}}$
$G^{\bar{c}} \leftarrow \arg\min_{G^{\bar{c}}} \mathbf{D}(G^c + G^{\bar{c}}, G^t + G^{\bar{t}})$
Step 2: Fix $G^{\bar{c}}$ to Optimize $G^{\bar{t}}$
$G^{\bar{t}} \leftarrow \arg\min_{G^{\bar{t}}} \mathbf{D}(G^c + G^{\bar{c}}, G^t + G^{\bar{t}})$
Until: Convergence

Fig. 11. Inferring algorithm.

The prerequisite for the inferring algorithm is to obtain G^c and G^t based on historical and real-time data, which is introduced in Section V-C.2. Another input is the historical cellphone and transit dataset. Given them, our inferring algorithm is described as in Figure 11.

The rationale behind our algorithm is as follows. We first initialize $G^{\bar{t}}$ and $G^{\bar{c}}$ based on historical data and census data, because the initialization does not affect the final results based on the property of the block coordinate descent [14]. In Step 1, we fix $G^{\bar{t}}$ and optimize $G^{\bar{c}}$ based on an objective function $\mathbf{D}$ to obtain the optimized $G^{\bar{c}}$ that minimizes $\mathbf{D}$; in Step 2, we fix the obtained $G^{\bar{c}}$, and optimize $G^{\bar{t}}$ with the same objective function $\mathbf{D}$ to obtain the optimized $G^{\bar{t}}$ that minimizes $\mathbf{D}$; and then we go back to Step 1, to further optimize $G^{\bar{c}}$, until the results converge. As follows, we discuss the convergence issue.

The above algorithm has an iterative process to alternatively optimize $G^{\bar{c}}$ and $G^{\bar{t}}$. Based on the property of the block coordinate descent, the convergence of the above iterative process is based on the objective function used. In this paper, we use Normalized Squared Loss function as our objective function $\mathbf{D}(G^c_{ab} + G^{\bar{c}}_{ab}, G^t_{ab} + G^{\bar{t}}_{ab})$ given as

$$\frac{((G^c_{ab}+G^{\bar{c}}_{ab})-(G^t_{ab}+G^{\bar{t}}_{ab}))^2}{\mathrm{STD}((G^c_1+G^{\bar{c}}_1-G^t_1+G^{\bar{t}}_1),\ldots,(G^c_M+G^{\bar{c}}_M-G^t_M+G^{\bar{t}}_M))}$$

where ab is an edge of a mobility graph, and M is the number of total edges of a mobility graph. As a result, we measure the distance between these two mobility graphs at edge levels. This normalized squared loss is an effective method to measure the distance between two variables and consider their distribution at the same time. Note that other distance functions can also be used in our iterative process but may not lead to the convexity of the optimization problem, and thus the convergence of the iterative process cannot be guaranteed. In this paper, we define the objective function for

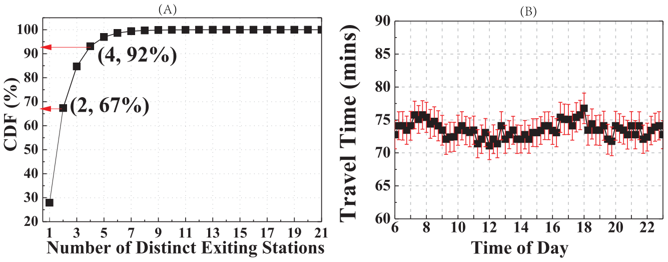


Fig. 12. (A) Distinct stations and (B) travel time.

one edge in order to parallelize the optimizations for all the edges in a mobility graph.

2) Context-Aware Inference: In the above optimization, we assume that G^c and G^t are given. As follows, we show how to obtain them. The inference about G^c and G^t is a classic topic for many data mining applications. In this work, we maintain and update the conditional probability distributions over the set of destinations that a particular cellphone or transit user can go to, given the previous trip history. We consider three real-time contexts, i.e., the current location, the time of day and the day of week. With these three contexts, we use a probabilistic method to obtain conditional probability distributions of destinations. This probabilistic method based on these three contexts has a high accuracy because the majority of residents are regular commuters traveling between home and workplaces. To validate this statement, Figure 12(A) gives the CDF of distinct exiting stations for residents taking subways in a week, and we found that 67% of all these residents only exit at two distinct stations or fewer, *e.g.*, home and workplace. As a result, it allows us to provide very accurate inference on which station a resident will exit, given the real-time contexts. In addition to the exiting station, we are also able to infer the exiting time. We notice that public transit systems have relatively stable travel time between stations in different time periods. Figure 12(B) gives the average subway travel time in a week between the Shenzhen train station and the Shenzhen airport, and we found that the travel time is stable about 74 minutes with a 2 minute variance. This nice feature allows us to use the timing information from smartcard transactions when residents *enter* the transit systems to infer the mobility. Essentially, we use the entering time and entering station as two real-time contexts to infer exiting time and exiting station. These two contexts increase inference accuracy.

In this work, we categorize the cellphone data by three contexts, the time of the day, the day of the week, and the transit volume. The first two contexts are classic temporal contexts. For a particular time slot of a day, the related historical data are the historical data with the same temporal contexts. For example, if we have the latest 14 weeks of the cellphone data as the total historical data, then for the slot from 7AM to 8AM on the next Monday, a possible set of the related historical data is the historical data belonging to the 14 time slots from 7AM to 8AM, one for each of 14 previous Mondays. In this work, we also consider another novel context called transit volume to select some highly-related historical data from this possible set. The transit volume is a novel logical context, which has not been investigated. Such a logical context cannot be replaced by others, because it accounts for irregular real-world events. For example, if we want to infer an uncaptured cellphone user's location, we find the previous data as the related training data in the same time of day on the same day of week (i.e., previous temporal contexts) with the similar transit volume, given by a deviation parameter ω, i.e., a volume V_1 is similar to another volume V_2 if $\frac{|V_1 - V_2|}{|V_1|} \leq \omega$. Similarly, we also categorize the transit data by three contexts, the time of the day, the day of the week, and the cellphone volume. The last context, i.e., the cellphone volume, is obtained by cellphone data and serves as a novel logical context to find related transit data from train purposes. The effectiveness of these parameters is evaluated in Section VI.

VI. mPat Evaluation

We evaluate the effectiveness of mPat's mobility inference based on the datasets introduced in Figure 6.

A. Evaluation Methodology

We compare mPat with two state-of-the-art models. **Radiation** model [15] predicts the mobility based on the population density of an origin region and a destination region and that of the surrounding regions. To estimate the region population, we allocate every cellphone user to a region where he/she stays the most at a particular slot based three months of the cellphone data. Radiation serves as a statistical model suitable for the situation where the real-time data are not available. **WHERE** model [4] takes the spatial and temporal probability distributions drawn from cellphone data and produces synthetic cellphone data, which indicates the inferred mobility. WHERE serves as a single-source approach suitable for the situation where only cellphone data are available. Further, we investigate the effect of utilizing the correlation between multi-source data by comparing mPat to its two versions: **mPat-S** which does not use any contexts to reduce the size of training data as in Section V-C.2; **mPat-C** which does not use transit or cellphone volumes as a context to select training data as in Section V-C.2.

We utilize three months (91 days) of datasets on a rolling basis. We divide the data into two subsets: *Testing Set* with data for one particular day as the real-time streaming data; *Historical Set* with data for 90 continuous days preceding the day in the testing set as the historical data. For a given day, if we use one hour slots, at the end of the first slot, *i.e.*, 1AM, we use mPat to infer mobility for the next slot from 1 to 2AM, based on both the "real-time" data from 12 to 1AM in the testing set, and all data in the historical set. We test models with **Mean Average Percent Error** (**MAPE**) in a time slot as

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|\bar{\mathbf{T}}_i - \mathbf{T}_i|}{\bar{\mathbf{T}}_i},$$

where $n = 496 \times 496 = 246016$ is the total number of region pairs; $\mathbf{T}_i$ is inferred mobility between a region pair i; $\bar{\mathbf{T}}_i$ is the ground truth of the mobility between a region pair i. An accurate model yields a small MAPE, and *vise versa*. We move data in the testing set forward for 90 days, leading to 90 experiments. The average results were reported.

Note that it is almost impossible to accurately obtain the ground truth of the urban mobility, unless we put a tracker

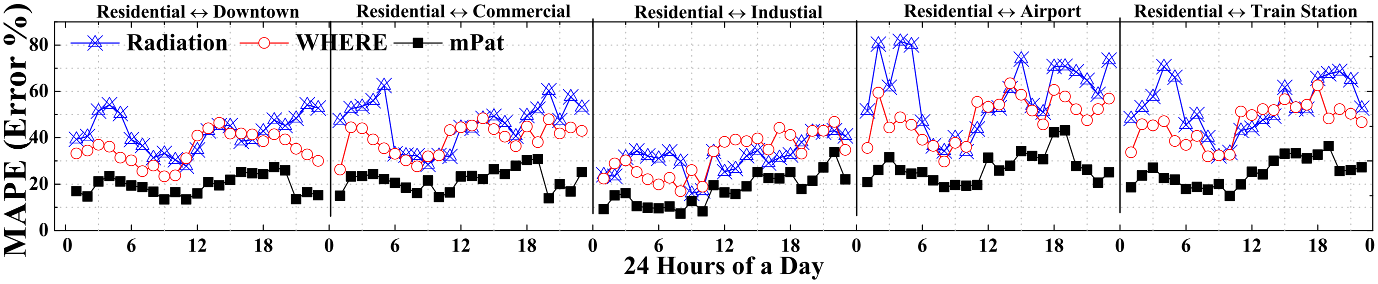


Fig. 13. MAPE under One Hour Slot for 24 Hours of a day.

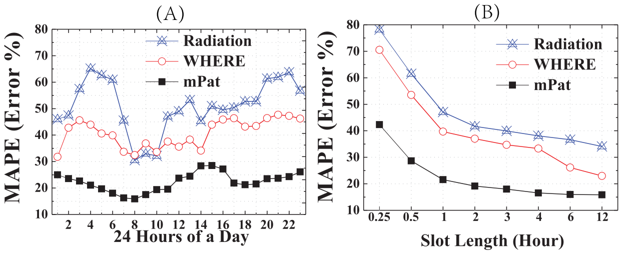


Fig. 14. (A) Hourly MAPE and (B) effects of lengths.

on every resident in a city 24 hours a day. In order to infer the ground truth, we introduce another new cellphone related dataset for the evaluation purpose, which contains regular location updates (in cell tower levels) of 7 million cellphone users in Shenzhen. The unique feature about this new dataset is that it was obtained by sampling locations of all the cellphone users every 15 minutes on average at the cell tower levels, no matter the users have the cellphone activities or not, and the only requirement is that the users have to turn on the cellphones. Note that the cellphone data feed is established for 10.4 million users but the ground truth is inferred based on 7 million users. This is because lots of cellphones do not have regular location updates due to low usage rates. We use the mobility obtained from this dataset as the ground truth for the evaluation. We believe our method based on new dataset is much closer to the actual ground truth than the previous methods. Note that we did not include this new dataset as an input of mPat in the inference design. This is because that this dataset requires extra supports from the cellphone network infrastructures and is not a generic dataset like CDR. Normally, the cellphone service providers normally will not store this large-scale streaming data since they grow more than 100GB per day and the most location updates are not useful to the providers for the billing purpose. In our case, this dataset is stored for a peak performance testing. Thus, unlike the datasets (CDR and transit) we introduced as the input of mPat, we obtained this new dataset from the providers offline and did not have the regular real-time access to it. Further, we investigate the impact of **Historical Data Size** on the accuracy and the running time of models to show the feasibility and robustness of different models for real-time inferences. Finally, we present the evaluation summary.

B. Accuracy on Region Levels

We compare three models' inferring accuracy in terms of MAPE values in different lengths of slots, by a low level comparison on five particular region pairs, and a high level comparison on all 246016 region pairs.

Figure 13 plots the MAPE under one hour slots with two-way mobility between a residential region and a downtown region, a commercial region, an industrial region, the airport region, and a train station region. In general, mPat outperforms WHERE, which outperforms Radiation. This is because Radiation only considers the population to infer the mobility instead of historical trips, and WHERE only uses the cellphone data and does not correlate it to the transit data. The performance gain between mPat and others is lower during the rush hour. This is because in the rush hour the repeatable mobility patterns are higher, so all models have better performance. Comparing the five region pairs, for the region pairs on which the residents go for commutes (i.e., between the residential region and the industrial, commercial or downtown regions), all models have better performance than the region pairs on which the residents go for travel (i.e., between the residential region to the airport or train station regions). This is because the repeatable mobility pattern of such travels is limited.

Figure 14(A) gives the MAPE on all region pairs under one hour slots during 24 hours. The MAPE of all three models are higher than the MAPE we observed in Figure 13. It is because the urban mobility may change dramatically between different region pairs, and some remote region pairs with few transit services or few cellphone activities lead to high MAPE. But the relative performance between three models is similar to Figure 13. WHERE outperforms Radiation by 19% on average, but at the morning rush hour Radiation outperforms WHERE by 9%. mPat outperforms WHERE by 42%, resulting from its utilization on the correlation between the cellphone and transit data.

Figure 14(B) plots the MAPE of all models with different slot lengths. The MAPE of all models reduces with the increase of the lengths, because the urban mobility in a longer slot becomes more stable. mPat outperforms WHERE and Radiation by 31% and 43% on average, respectively. When the slot becomes longer than 6 hours, mPat and WHERE have the similar performance, because in such a long time slot, the cellphone data alone is capable of inferring the mobility.

C. Accuracy on Street Levels

We study mPat at a different spatial granularity by showing the MAPE on the road segment level. Based on a digital map of Shenzhen, we assign the transit stations, taxicab locations, and cell towers to the closest road segments based on the Euclidean distance between them.

Figure 15(A) plots the MAPE of the inferred mobility from one of the busiest streets in downtown Fu Hua Road to the busiest street in a residential area Le Yuan Road.

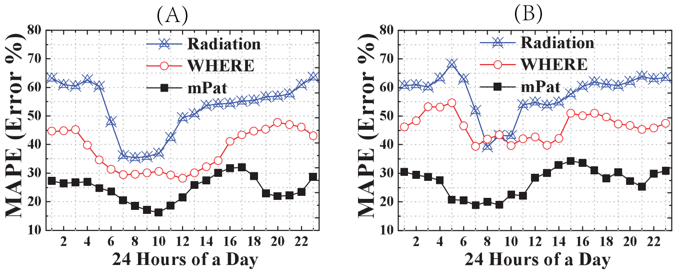


Fig. 15. MAPE with (A) one pair and (B) 1000 pair.

mPat performs better than WHERE and Radiation, especially at night when residents do not have many phone activities but still use urban transit. WHERE significantly outperforms Radiation during in the early morning and the regular day time, but Radiation has good accuracy during the morning rush hour. This is because during the rush hour, passenger demand is relative stable compared to other times.

Figure 15(B) shows the average MAPE between 1,000 randomly selected segment pairs. The MAPE of all three models are higher than the MAPE in Figure 15(A) although their relative performance is similar to that in Figure 15(A). mPat is better than WHERE, but WHERE is normally better than Radiation. Compared to Figure 14(A), we found an 11% accuracy loss in mPat due to the finer granularity.

D. Justification of mPat

We justify the design of mPat by (i) finding the optimal value for ω, which is used to decide the similarity (in terms of percentage) between transit volumes when using transit data to select training data as in Section V-C.2; (ii) comparing mPat to its two versions, i.e., mPat-C and mPat-S on different sizes of training data. The difference between mPat and mPat-C is whether we use transit or cellphone data as a Context to reduce the size of training data; whereas the difference between mPat and mPat-S is whether we use any contexts to reduce the size of training data.

Figure 16(A) gives effects of ω on mPat's MAPE. With the increase of ω, mPat's MAPE first decreases, and then increases. This is because when ω increases, mPat finds more slots with the similar transit volume to accurately expand the graph. But when ω becomes larger, mPat considers more slots with different transit volumes, leading to inaccurate expansion. The MAPE is minimized when $\omega = 0.5$ (i.e., the default value). If ω used leads to an empty training set, ω increases until it is not empty.

Figures 16(B) and 16(C) plot MAPE and running time on different data sizes in terms of weeks. As expected, the more the historical data, the lower the MAPE error, the longer the running time. mPat- C and mPat-S have a 7% accuracy gain to mPat, but the running time of mPat-C and mPat-S is much longer than mPat. This is because with the same historical data, (i) mPat has smaller training data than mPat-C, since mPat uses transit or cellphone volume as another context to select much compact training data; (ii) mPat has much smaller data than mPat-S, since mPat-S does not use any contexts to reduce the size of training data. The key feature of mPat using logical contexts leads to slightly higher MAPE but significantly shorter running time, which justifies the design of mPat.

E. mPat Evaluation Summary

We have the following observations. (i) The inference accuracy is highly depended on both locations and times as in Figures 13, 15(A) and 15(B). On average, all models have better performance in the morning rush hour, due to the predicability of morning commutes, and mPat outperforms other models as in Figure 15(A). (ii) The length of slots has significant impacts on performance of all models as in Figure 15(B). It is intuitive that a longer slot has lower error rates, yet leading to the low usability for many real-time applications. (iii) System parameters pose significant impacts of accuracy, and the optimal parameters have to be evaluated based on empirical data as in Figure 16(A). (iv) We maximally reduce 81% of the running time with a 19% accuracy reduction by using the correlation between cellphone and transit data to reduce the size of training data as in Figures 16(B) and 16(C).

VII. Inter Region Transit Service

Recently, the Shenzhen transport committee launched a pilot program to provide non-stop express transit services between several fixed location pairs with high passenger transit demand, but the passengers have concerns that the routes (i.e., the provided location pairs) are limited [16]. In this section, we facilitate this pilot program by proposing and evaluating a novel transit service called Inter Region Transit IRT based on our human mobility pattern analysis and inference. IRT is designed to identify the urban region pairs between which there is high human mobility yet low public transit mobility (i.e., the portion of human mobility supported by public transit) and then to provide non-stop transit services among these identified pairs. IRT is potentially capable of reducing the travel time between these undersupplied regional pairs.

A. IRT Design Overview

The design of IRT is described in terms of coverage, capacity, and schedule. Based on the inferred human and transit mobility, (i) Coverage: we find the undersupplied region pairs with high human mobility yet low transit mobility according to a given *undersupply ratio* ρ as a threshold, i.e., a region pair has IRT services only if its $\frac{\text{Human Mobility}}{\text{Transit Mobility}} > \rho$; (ii) Capacity: we individually decide the number of IRT vehicles for every region pair based on (a) the difference between human mobility and transit mobility among this region pair (i.e., potential passengers), (b) the vehicle capacity (e.g., 20-seat bus), (c) the schedule period (e.g., one hour) and (d) the travel time between this region pair; (iii) Schedule: we compute the departure time of vehicles leaving the service stop of a region (e.g., a logical centroid of a region) based on the number of vehicles and the schedule periods. Figure 17(A) gives an IRT example. Based on a given ratio ρ, we find two undersupplied region pairs (R_1, R_2) and (R_3, R_4) by the historical human and transit mobility patterns. Then, based on the historical mobility inference for the next service period, there are 10, 20, 30 and 40 residents unaccounted for, i.e., the difference between human and transit mobility. Finally, we set the number of vehicles and the departure time based on the schedule period, the travel time, and the vehicle capacity.

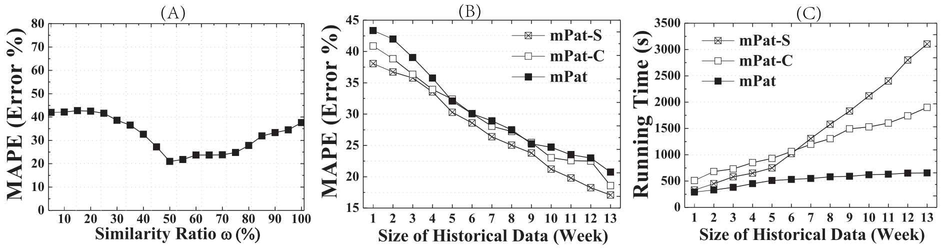


Fig. 16. Performance influence of ω and data. (A) ω vs. MAPE. (B) Data vs. MAPE. (C) Data vs. Time.

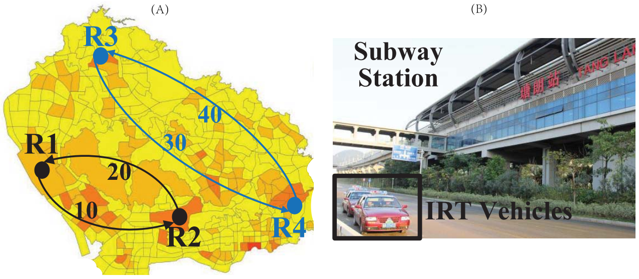


Fig. 17. (A) IRT Example and (B) implementation.

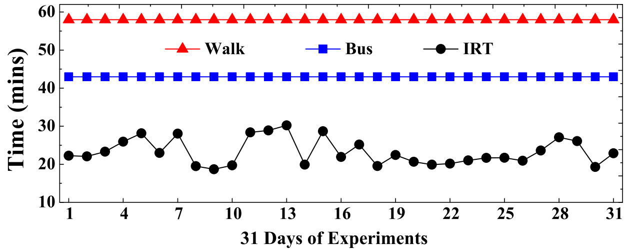


Fig. 18. Average Travel Time in 31 days.

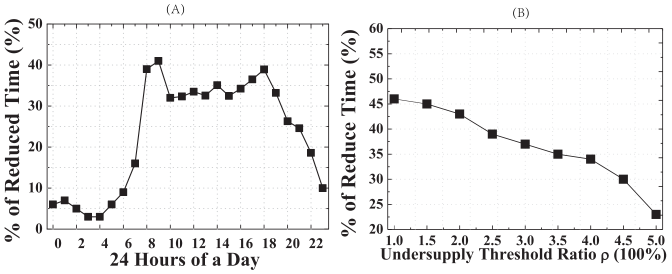


Fig. 19. (A) Reduced time and (B) Time vs. ρ.

B. Real World Experiment

We implemented IRT in one pair of undersupplied regions as a real-world effort. Since it requires a permit to deliver passengers in Shenzhen, we invited 12 volunteers who daily commute between these two regions for the evaluation. As shown in Figure 17(B), based on the number of volunteers, we rented 3 regular taxicabs to deliver them from a subway station as a service stop in one region to their workplace as another service stop in another region. At the subway station, 12 volunteers were picked up, and then were directly driven to their work. We calculated the departure times for every vehicle, but since these volunteers had to be driven to their work eventually, we only logged these departure times, instead having vehicles leave based on the calculated times. Then, we calculated what these passengers' travel times would be if the vehicles left based on departure times and went back to pick them up. But in reality, the vehicles waited and picked up them in one round at the subway station as one service stop and then dropped them off at their workplace as another service stop. We videotaped the service using three smart phones with which the travel time between two service stops was calculated. In Figure 18, we compare the travel time in IRT to the time of walking or taking a regular bus between the two service stops for a 31-day evaluation. We found that IRT reduces the travel time, compared to 43 mins of taking a bus or 57 mins of walking. But since the taxicabs are faster than the buses in terms of speed, we use a factor ν to account for the speed difference. In the experiment, ν is obtained based on our historical bus and taxicab GPS datasets. We found that IRT with ν still saves significant travel time for passengers. Note that we did not evaluate the speed factor for Wake and Bus solutions since we focused on evaluation of IRT itself.

C. Trace Driven Evaluation

We evaluate the performance of IRT by investigating the impact of the time of the day and the undersupply ratio ρ (default setting $\rho = 2$) on the percentage of reduced travel time. The original travel time between a pair of regions is given by the average travel time of three public transit modes, i.e., taxicab, bus, and subway; the new travel time in IRT is calculated based on the average travel time of taxicabs, since IRT is an express transit service without intermediate stops between regions, and thus IRT passengers' travel time is similar to that of taxicab passengers. But we still account for the speed difference with ν, which is obtained by the historical GPS data.

Impact of Time of Day: Figure 19(A) plots the percentages of the reduced travel time in 24 hours. During the 24 hours of a day, there is a high percentage of reduced travel time during the morning and evening rush hour, but in the non-rush hour period, e.g., during the early morning, the percentage of reduced travel time is low. This is because in the rush hour, many passengers use bus and subway services, leading to prolonged travel time due to intermediate stops between regions, whereas in the early morning, most travel passengers use taxicab services without intermediate stops, leading to similar time as IRT.

Impact of Undersupply Ratio: IRT services are deployed between a region pair only if between this pair, $\frac{\text{Human Mobility}}{\text{Transit Mobility}} > \rho$. Figure 19(B) plots the impact of ρ on the reduced travel time. With the increase of ρ, the average travel time decreases. This is because the larger the ρ, the fewer the region pairs that have IRT services. Further, the region pairs without IRT typically do not have an undersupply of urban transit services between them, so the residents traveling between these pairs typically use private or urban transit with a shorter travel time. Thus, the percentage of the reduced time decreases when ρ is larger.

VIII. Related Work

A. Human Mobility Modeling

Analyzing the human mobility in urban scales is crucial for mobile applications, urban planning [17], transportation [18] and social networks [19]. At first, the human mobility is studied by theoretical models, e.g., random way point, which are simple enough to be tractable, but no empirical evidence exists to prove the accuracy of such models [20]. Recently, analyzing human mobility based on empirical data has received significant attention, due to the ubiquity of GPS devices and urban infrastructure upgrades [4], [9]–[11], [21]–[26]. We summarize the related work by the utilized data.

- **Cellphone Data** Numerous methods have been proposed for the study of human mobility based on call detail records (CDR), e.g., modeling how cellphone users move [4], [27]; predicting where cellphone users will travel next [26]; estimating cellphone users' travel ranges [22]; identifying cellphone users' important locations, e.g., at workplaces or home [10], as well as in specific cities, e.g., Los Angles, New York City [24], and Rome [23]. However, the cellphone data is biased against a certain group of residents, leading to inaccurate analysis. To our knowledge, we are the first to correlate cellphone data with transit data to address the bias issue in order to increase analysis accuracy.
- **Transit Data** Transit data are another important source for research in human mobility, e.g., identifying human mobility based on data from taxicabs [7], [28]–[30], buses [6], subways [5], [31], bikes [32], [33], highway toll collection systems [34], [35], and private cars [2]. But our method is based on data from the entire set of urban transit networks correlated with data from cellular networks, instead of sampling residents using a specific mode.
- **Other Data** Other datasets have also been used for human mobility analysis, e.g., personal travel diaries [36], circulation of bank notes [37], studying statistical pattern from participants carrying GPS devices [3], analyzing urban resident mobility patterns from wireless network traces in WiFi access points [38], [39]. Recently, human mobility has also been investigated by examining social networks or mobile *ad hoc* networks, e.g., with check-in data [19] and proximity data [40]. While these data provide additional details, the number of involved residents is usually small compared to transit and cellphone users, which leads to a potential bias toward not only the residents who choose to reveal their locations but the places they choose to reveal. We argue that in the urban level, the social network data (e.g., check-ins) can be mostly covered by cellphone data, since an app user usually employs their cellular connectivity to access the social network.

B. Urban Data-Driven Applications

Many novel applications are proposed to improve urban mobility [41]–[46], *e.g.*, assisting mobile users to make transit decisions, such as taking a taxicab or not [47], inferring real-world maps based on GPS data [48], finding parking spots for drivers [49], predicting bus arrival times [50], informing drivers with smart routes based on those of experienced drivers [51], enabling passengers to query taxicab availability to make informed transit choices [52], predicting passenger demand for taxicab drivers [41], modeling the urban transit [44], recommending optimal pickup locations [53], suggesting profitable locations for taxicab drivers by constructing a profitability map where the nearby regions of drivers are scored serving as a metric for a taxicab driver decision making process [54], navigating new drivers based on GPS traces of experienced drivers [55], detecting the taxicab anomaly [56], and enabling us to better understand region functions of cities [57]. Yet the above research has not focused on general urban mobility modeling and and typically utilizes only one type of datasets.

C. Summary

Despite the recent explosion in human mobility research, the bulk of work has focused on biased single-source datasets. Any of them alone only involves a particular group of residents, who are a biased sample for the entire urban population. Further, the previous work has mostly utilized the static historical datasets already collected *offline*. These drawbacks point out a need for the online integration and utilization of multi-source data from urban infrastructures without marginal costs. To meet this need, in this paper, we provide a novel architecture mPat, which is uniquely built upon real-time multi-source feeds to abstract human mobility for real-world applications. Such an architecture has not been investigated before. Thus, for the first time, mPat is able to provide insights on not only *where* and *when* but also *how* (e.g., by bus) almost all urban residents move. Thus, we believe mPat will substantially assist the mobile computing community in designing better applications.

IX. Discussion

We discuss some issues about our architecture related to user privacy, data access, and data variety as follows.

Privacy Protections: While the data for the human mobility study have the potential for great social benefits, the privacy concerns regarding their utilization have inhibited their release and wider applications. We briefly discuss the active steps we took for privacy protections. (i) Anonymization: All data from feeds are anonymized by the employees of the service providers who were not involved in the our project, and each identifiable ID (e.g., SIM card IDs) is replaced by a serial identifier. (ii) Minimal Exposure: We only store and process the information which are useful for our mobility analysis, and drop other information for the minimal exposure, e.g., we store the cell tower IDs to infer locations in the cellphone data, but not durations of calls. (iii) Aggregation: Our mobility patterns are given at aggregated results in a temporal-spatial partition and are not focused on individual cellphones or transit users. (iv) Nature of Data Feeds: The nature of different feeds also provides a certain level of privacy protections, e.g., the taxicab and bus GPS feeds do not involve any identity about passengers; the smart card feeds only show passengers' locations at transit station levels; the cellphone feeds only generate temporally sparse and spatially coarse records.

Public Data Access: Accessing empirical datasets is vital to the mobility-related research, but such datasets are usually not available for the fellow researchers due to the privacy issues. As an initiative step, the partial aggregated data used in this work have been made for public access in the website of Transport Committee of Shenzhen Municipality [58]. Moreover, we will release more detailed data for the benefit of the mobile research community with privacy protection schemes. In particular, we are trying to assist taxicab companies to release the noisy distribution of their GPS records while retaining statistical features, which serves as a further step towards enabling the service providers to unlock their data's value for research and application designs with broad social benefits.

Additional Data Sources: In the currently Shenzhen implementation, we only focus on two data sources with large-scale datasets, i.e., cellphone and transit, due to the space limitation. We briefly introduce other urban data that will be included in mPat to increase the analysis and inference accuracy, if the residents opt to participate under privacy preserving mechanisms. (i) Image data from the traffic cameras or the cameras inside the transit systems. (ii) Data of a growing bicycle network with 8,000 bicycles in Shenzhen for rentals using the smart cards. (iii) Data of private vehicles that are tracked by the onboard GPS installed by themselves or the manufacturers for location-based services. Including these data into our architecture leads to more accurate analyses on the correlation between cellphone users and public transit users, since their difference is the private vehicle users. But these data typically have more sensitive information and are more complicated to be intergraded into our real-time data feed layer.

X. Conclusion

In this work, we design, implement and evaluate an architecture for the analysis and inference on human mobility with a 79% inference accuracy. Our technical endeavors offer a few valuable insights for the fellow researchers to utilize our architecture for not only the mobile network study but also real-world applications. Specifically,

1) the single-source data based study introduces biases for the human mobility research, and the correlation as well as divergence among multi-source data have the potential to address such biases;
2) the mobility patterns obtained from cellphone and transit users are slightly correlated but certainly have a divergence, which can be used to track for the residents who cannot be tracked by either of datasets alone;
3) the multi-source data can be used for cross-referencing in the real-time mobility inference in order to reduce the data need to be processed for shorter running time yet still maintaining the performance;
4) while it is challenging to integrate and utilize heterogenous large-scale data feeds, it is more challenging to negotiate with the service providers for the real-time data access and to protect the privacy of studied residents.

References

[1] F. Bogo and E. Peserico, "Optimal throughput and delay in delay-tolerant networks with ballistic mobility," in *Proc. MobiCom*, 2013, pp. 303–314.

[2] F. Giannotti *et al.*, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," *VLDB J. Int. J. Very Large Data Bases*, vol. 20, no. 5, pp. 695–719, 2011.

[3] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the levy-walk nature of human mobility," in *Proc. INFOCOM*, Apr. 2008, pp. 1597–1605.

[4] S. Isaacman *et al.*, "Human mobility modeling at metropolitan scales," in *Proc. MobiSys*, 2012, pp. 239–252.

[5] N. Lathia and L. Capra, "How smart is your smartcard?: Measuring travel behaviours, perceptions, and incentives," in *Proc. UbiComp*, 2011, pp. 291–300.

[6] S. Bhattacharya *et al.*, "Gaussian process-based predictive modeling for bus ridership," in *Proc. UbiComp*, 2013, pp. 1189–1198.

[7] R. Ganti, M. Srivatsa, A. Ranganathan, and J. Han, "Inferring human mobility patterns from taxicab location traces," in *Proc. UbiComp*, 2013, pp. 459–468.

[8] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system," in *Proc. UrbComp*, 2012, pp. 142–148.

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, Jun. 2008.

[10] S. Isaacman *et al.*, "A tale of two cities," in *Proc. HotMobile*, 2010, pp. 19–24.

[11] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, 2010.

[12] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proc. 13th Int. Conf. Ubiquitous Comput. (UbiComp)*, 2011, pp. 89–98.

[13] D. Zhang *et al.*, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proc. 20th ACM Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2014, pp. 201–212.

[14] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[15] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, p. 96, Apr. 2012.

[16] (2014). *Shenzhen Customized Public Transit*. [Online]. Available: http://house.people.com.cn/n/2014/0411/c164220-24882370.html

[17] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 38:1–38:5, Sep. 2014. [Online]. Available: http://doi.acm.org/10.1145/2629592

[18] J. Aslam, S. Lim, X. Pan, and D. Rus, "City-scale traffic estimation from a roving sensor network," in *Proc. 10th ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, 2012, pp. 141–154.

[19] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. KDD*, 2011, pp. 1082–1090.

[20] I. Rhee *et al.*, "On the levy-walk nature of human mobility," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 630–643, Jun. 2011.

[21] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 29:1–29:41, May 2015. [Online]. Available: http://doi.acm.org/10.1145/2743025

[22] S. Isaacman *et al.*, "Ranges of human mobility in Los Angeles and New York," in *Proc. PerCom Workshops*, Mar. 2011, pp. 88–93.

[23] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat, "Digital footprinting: Uncovering tourists with user-generated content," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 36–43, Oct. 2008.

[24] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti, "Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate," in *Proc. Int. Conf. Comput. Urban Planning Urban Manage.*, 2009, pp. 1–12.

[25] F. Girardin, F. D. Fiore, J. Blat, and C. Ratti, "Understanding of tourist dynamics from explicitly disclosed location information," in *Proc. Symp. LBS Telecartograph.*, 2007, pp. 1–11.

[26] K. Dufková, J.-Y. Le Boudec, L. Kencl, and M. Bjelica, "Predicting user-cell association in cellular networks from tracked data," in *Mobile Entity Localization and Tracking in GPS-less Environnments*. Springer, 2009, pp. 19–33.

[27] F. Xu, P. Zhang, and Y. Li, "Context-aware real-time population estimation for metropolis," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 1064–1075.

[28] F. Miao *et al.*, "Data-driven distributionally robust vehicle balancing using dynamic region partitions," in *Proc. 8th Int. Conf. Cyber-Phys. Syst.*, 2017, pp. 261–271.

[29] H. Rong, X. Zhou, C. Yang, Z. Shafiq, and A. Liu, "The rich and the poor: A Markov decision process approach to optimizing taxi driver revenue efficiency," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 2329–2334.

[30] X. Sun *et al.*, "Participatory sensing meets opportunistic sharing: Automatic phone-to-phone communication in vehicles," *IEEE Trans. Mobile Comput.*, vol. 15, no. 10, pp. 2550–2563, Oct. 2016.

[31] B. Du, C. Liu, W. Zhou, Z. Hou, and H. Xiong, "Catch me if you can: Detecting pickpocket suspects from large-scale transit records," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 87–96.

[32] J. Liu, L. Sun, W. Chen, and H. Xiong, "Rebalancing bike sharing systems: A multi-source data smart optimization," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1005–1014.

[33] Z. Yang *et al.*, "Mobility modeling and prediction in bike-sharing systems," in *Proc. 14th Annu. Int. Conf. Mobile Syst., Appl., Services*, 2016, pp. 165–178.

[34] S. Fang, K. Bian, K. Xie, D. Cui, and H. Hong, "The shortest path or not? Analyzing the ambiguity of path selection in China's toll highway networks," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 1964–1969.

[35] S. Wan *et al.*, "Predictability analysis on expressway vehicle mobility using electronic toll collection data," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 2589–2594.

[36] T. Hagerstrand, "What about people in regional science?" *Papers Reg. Sci. Assoc.*, vol. 24, no. 1, pp. 6–21, 1970.

[37] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, p. 462, Jan. 2006.

[38] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in *Proc. IEEE INFOCOM*, Apr. 2006, pp. 1–13.

[39] J. Yoon, B. D. Noble, M. Liu, and M. Kim, "Building realistic mobility models from coarse-grained traces," in *Proc. MobiSys*, 2006, pp. 177–190.

[40] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in *Proc. WWW*, 2010, pp. 61–70.

[41] Y. Ge *et al.*, "An energy-efficient mobile recommender system," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2010, pp. 899–908.

[42] Y. Huang and J. W. Powell, "Detecting regions of disequilibrium in taxi services under uncertainty," in *Proc. 20th Int. Conf. Adv. Geograph. Inf. Syst. (SIGSPATIAL)*, 2012, pp. 139–148.

[43] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger," in *Proc. 13th Int. Conf. Ubiquitous Comput. (UbiComp)*, 2011, pp. 109–118.

[44] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma, "Understanding transportation modes based on GPS data for Web applications," *ACM Trans. Web*, vol. 4, no. 1, p. 1, Jan. 2010.

[45] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the Web," in *Proc. 17th Int. Conf. World Wide Web (WWW)*, 2008, pp. 247–256.

[46] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on GPS data," in *Proc. 10th Int. Conf. Ubiquitous Comput. (UbiComp)*, 2008, pp. 312–321.

[47] W. Wu, W. S. Ng, S. Krishnaswamy, and A. Sinha, "To taxi or not to taxi?—Enabling personalised and real-time transportation decisions for mobile users," in *Proc. IEEE 13th Int. Conf. Mobile Data Manage. (MDM)*, Jul. 2012, pp. 320–323.

[48] J. Biagioni and J. Eriksson, "Map inference in the face of noise and disparity," in *Proc. SIGSPATIAL*, 2012, pp. 79–88.

[49] A. Nandugudi, T. Ki, C. Nuessle, and G. Challen, "PocketParker: Pocketsourcing parking lot availability," in *Proc. UBICOMP*, 2014, pp. 963–973.

[50] J. Biagioni, T. Gerlich, T. Merrifield, and J. Eriksson, "Easytracker: Automatic transit tracking, mapping, and arrival time prediction using smartphones," in *Proc. SenSys*, 2011, pp. 68–91.

[51] L.-Y. Wei, Y. Zheng, and W.-C. Peng, "Constructing popular routes from uncertain trajectories," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 195–203.

[52] R. K. Balan, K. X. Nguyen, and L. Jiang, "Real-time trip information service for a large taxi fleet," in *Proc. Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2011, pp. 99–112.

[53] Y. Ge, C. Liu, H. Xiong, and J. Chen, "A taxi business intelligence system," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 735–738.

[54] J. W. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps," in *Proc. 12th Int. Symp. Adv. Spatial Temporal Databases*, 2011, pp. 242–260.

[55] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 316–324.

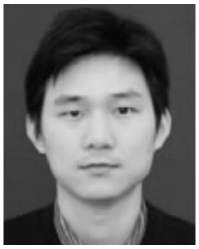
[56] R. Sen and R. K. Balan, "Challenges and opportunities in taxi fleet anomaly detection," in *Proc. SENSEMINE*, 2013, pp. 1–6.

[57] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 186–194.

[58] Transport Committee of Shenzhen. (2014). *Statistical Data for Transportation in Shenzhen*. [Online]. Available: http://www.sz.gov.cn/jtj/tjsj/zxtjxx/

Desheng Zhang (M'11) is currently an Assistant Professor with the Department of Computer Science, Rutgers University. He is broadly concentrated on bridging cyber-physical systems also known as Internet of Things under some contexts and big urban data by technical integration of communication, computation, and control in data-intensive urban systems. He is also focused on the life cycle of big-data-driven urban systems, from multi-source data collection to streaming-data processing, heterogeneous-data management, model abstraction, visualization, privacy, service design, and deployment in complex urban setting. His current research interests are real-time interactions among heterogeneous urban systems including cellphone, smartcard, taxi, bus, truck, subway, bike, personal vehicle, electric vehicle, and road networks.

Tian He (M'03–SM'12–F'18) received the Ph.D. degree from the University of Virginia, Charlottesville, VA, USA, under the supervision of Prof. J. A. Stankovic. He is currently a Full Professor with the Department of Computer Science and Engineering, University of Minnesota-Twin Cities. He has authored and co-authored over 200 papers in premier network journals and conferences with over 20 000 citations (h-index 59). His research includes wireless sensor networks, cyber-physical systems, intelligent transportation systems, real-time embedded systems, and distributed systems, supported by the National Science Foundation, IBM, Microsoft, and other agencies. He was a recipient of the NSF CAREER Award in 2009, the McKnight Land-Grant Chaired Professorship in 2011, the George W. Taylor Distinguished Research Award in 2015, the China NSF Outstanding Overseas Young Researcher I and II in 2012 and 2016, and five best paper awards in international conferences. He has held a few general/program chair positions in international conferences and on many program committees and also has been an Editorial Board Member of six international journals, including the *ACM Transactions on Sensor Networks* and the IEEE TRANSACTIONS ON COMPUTERS.



Fan Zhang received the B.S. degree in communication engineering and the M.S. and Ph.D. degrees in communication and information system from the Huazhong University of Science and Technology, Wuhan, China, in 2002, 2004, and 2008, respectively. He is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. From 2008 to 2009, he was a Senior Research Associate with the City University of Hong Kong. From 2009 to 2011, he was a Post-Doctoral Fellow with the University of New Mexico and the University of Nebraska-Lincoln, USA. His research topics include cloud computing, big data, cyber-physical systems, privacy, and security in cloud computing.

Chengzhong Xu is currently a Professor, the Chair Scientist, and the Director of the Institute of Advanced Computing and Digital Engineering and the Director of the Research Center for Cloud Computing, Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. His research Interests include parallel and distributed systems, Internet and cloud computing, and high-performance computing. He was a recipient of the 2011 National Thousand Talents Program, Guangdong Tech-Leaders, and the Outstanding Overseas Researchers Award from the National Science Foundation of China.