Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales *

Desheng Zhang† zhang@cs.umn.edu
Jun Huang‡ jun.huang@siat.ac.cn
Ye Li ‡ li.ye@siat.ac.cn
Fan Zhang‡ zhangfan@siat.ac.cn
Chengzhong Xu‡ cz.xu@siat.ac.cn
Tian He† tianhe@cs.umn.edu

†Department of Computer Science and Engineering, University of Minnesota, USA
‡Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China

ABSTRACT

Expanding our knowledge about human mobility is essential for building efficient wireless protocols and mobile applications. Previous human 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 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: (i) a data feed layer consisting of real-time data feeds from 24 thousand vehicles, 16 million smart cards and 10 million cellphones; (ii) a mobility abstraction layer exploring the correlation and divergence among the multi-source data to analyze and infer human mobility; and (iii) an application layer to improve urban efficiency based on the human mobility findings of the study. The evaluation shows that mPat achieves a 75% inference accuracy, and that its real-world application reduces passenger travel time by 36%.

1. INTRODUCTION

Human mobility patterns are of great importance for both the design and evaluation of wireless network protocols and mobile applications [4]. Recently, the study of human mobility has gained significant attention, thanks to the ubiquity of human location tracking devices, e.g., onboard GPS devices [8], cellphones [11], and Automatic Fare Collection (AFC) devices deployed by urban transit (e.g., subways [12], buses [3], and taxicabs [16]). A large collection of data from these devices can serve as a powerful macroscope to observe large-scale mobility patterns with high accuracy over long periods, which holds the potential to revolutionize the research on wireless networks and mobile services, such as optimization on Wi-Fi AP or cell tower deployment, mobile network pricing, and traffic-based advertising.

Based on various empirical datasets, several human mobility models have been proposed [7] [9] [10] [11] [15] [17] that capture human mobility to a certain degree yet all share the common drawback of biased sampling:

- The research based on cellphone data assumes that a high penetration of cellphones implies that cellphone users' mobility can serve as a general proxy for all residents' mobility. But based on our empirical results in Section 3, 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 residents who are not engaged in cellphone activities during their movements.

- The research based on urban transit data assumes that residents of large urban cities usually employ urban transit (i.e., taxicabs [7], buses [3] or subway [12]) for travel. But based on our empirical results in Section 3, an approach based on data from one type of transit has a bias against residents using other types of transit. To our knowledge, there has been no research based on data from all forms of urban transit (taxicabs, buses, subways and private vehicles).

In short, we argue that almost all state-of-the-art theory and practice on human mobility have focused on single-source empirical data in isolation from one another. Essentially, they all utilize just one portion of urban residents who are involved with a specific empirical dataset as a sample for the entire urban population, which inevitably introduces a sampling bias against the uninvolved residents. The key reason for these single-source approaches is that previous 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.

Now, however, the rapid expansion of urban infrastructure in recent years has offered new research opportunities to exploit real-time multi-source data to rectify sampling

*Prof. Tian He is the corresponding author of this paper.

MobiCom'14, September 7-11, 2014, Maui, HI, USA.

ACM 978-1-4503-2783-1/14/09 ...$15.00.
http://dx.doi.org/10.1145/2639108.2639116 .

bias. In particular, data from cellular networks can be consolidated with data from urban transit networks that integrate various sensors, communication devices and automatic fare collection devices. Therefore, such urban infrastructures are capable of capturing almost all residents traveling within an urban scale (e.g., taxicab users, bus users, subway users and cellphone users) under well-structured incentive and privacy-preserving mechanisms. Although these comprehensive multi-source data have the potential to revolutionize human mobility research, until now our knowledge of the correlation, divergence and integration among these multi-source data has been extremely limited.

To deepen our understanding on how these data can be used for the study of human mobility, we propose an architecture called mPat, which improves the urban efficiency by uniquely analyzing and inferring real-time Mobility PATterns based on the correlation and divergence among multi-source data provided by urban infrastructures, including cellular networks and transit networks. While cellphone and transit data have been studied in previous human mobility research, they have been exploited in isolation by separate researchers for different cities. In contrast, mPat incorporates data from transit networks and cellphone networks, allowing fellow researchers to examine human mobility within detailed transit contexts and to gain deeper insight. 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. In mPat, we uniquely establish multi-source data feeds through urban infrastructures to obtain mobility abstraction for application designs. The mobility abstraction essentially is the narrow waist of mPat, allowing a separation between the data feeds and the applications for fellow researchers to add more feeds or applications without redesigning the whole infrastructure. We provide a reference implementation of 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) 1 billion 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 used to pay subway and bus fares. To our knowledge, the feeds and datasets established for this study provide the highest quality of any previous human mobility research for the human mobility study in three aspects: the most detailed data including call detail records, fare records, and GPS records; the largest resident coverage, i.e., 10.4 million residents; the most complete urban data including cellular, taxicab, bus and subway networks for the same city. These datasets are available at [1].
- We make several key attempts to explore spatio-temporal correlations and divergence between multi-source data. We transparentize the heterogenous features of these data to abstract the high-level human mobility knowledge by which we are able to infer human mobility patterns. Based on our 91-day evaluation, mPat infers urban mobility with an average accuracy of 75%, and outperforms a statistical model by 57% and a single-source model by 39%. The ground truth of the urban mobility is based on a dataset about cellphone location updates.
- Based on the mobility patterns estimated by mPat, we select urban region pairs with high human mobility yet low transit mobility to optimize the effectiveness of Inter Region Transit (IRT) services. Based on our 31-day experiment, IRT reduces resident travel times by 36% on average, indicating mPat's real-world value.

The rest of the paper is organized as follows. Sections 2 and 3 give the related work and motivations. Section 4 presents an overview. Section 5 introduces the data feeds. Section 6 describes the mobility abstraction. Section 7 evaluates mPat. Section 8 presents our application, followed by a discussion and conclusion in Sections 9 and 10.

2. RELATED WORK

Recently, analyzing human mobility based on empirical data has received significant attention, due to the ubiquity of GPS devices and urban infrastructure upgrades. We summarize the related work by the utilized empirical 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 [11]; predicting where cellphone users will travel next [6]; and identifying cellphone users' important locations, e.g., work or home [10]. 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 GPS data are another important source for research in human mobility, e.g., identifying human mobility based on data from taxicabs [7], buses [3], subways [12], and private cars [8]. In contrast, 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 transit mode.

Other Data: Recently, human mobility has also been investigated by examining social networks or mobile ad hoc networks, e.g., with check-in data [5] and proximity data [2]. 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.

In short, despite the recent explosion in human mobility research, the bulk of work has focused on biased single-source datasets. This drawback points out a need for the integration and utilization of multi-source data. To meet this need, we aim to provide a novel architecture mPat, which is uniquely built upon real-time multi-source feeds to abstract human mobility for real-world applications. 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. As a result, we believe mPat will substantially assist the mobile computing community in designing better applications.

3. MOTIVATIONS

In this section, we introduce two design motivations based on the empirical data we collected in Shenzhen.

3.1 Drawback of Using Single-Source Data

The previous research on human mobility has relied on individual data generated from single-source feeds, e.g., cellphone or transit networks, which typically leads to biased sampling with inaccurate results. To provide such evidence, in Figure 1, we give the number of residents going to the Shenzhen airport during a 24 hour period based on four empirical single-source datasets, i.e., cellphone, taxicab, bus, and subway. We aggregate the residents using taxicabs, buses and subways as the urban transit residents for a clear comparison with the cellphone users.

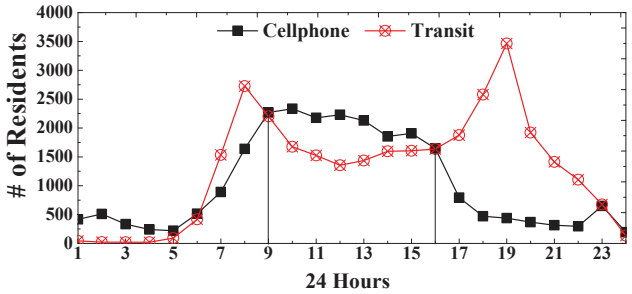

Fig 1: Tracked Residents to Airport in 24 Hours

As shown in Figure 1, we can use only cellphone activities to track most residents between 12AM and 5AM, during which the transit systems are not in service, except taxicabs. During non-rush hours, i.e., from 9AM to 4PM, the cellphone data track more residents than the transit data, because residents travel by private vehicles as well as by public transit during this period. Interestingly, we note that during the morning and evening rush hours, the transit data track more residents than cellphone data because it is difficult to make phone calls in the extremely crowded subway and buses in Shenzhen. Further, neither cellphone nor transit data can track a total resident count close to the daily departure passenger count of 51,000 provided by the Shenzhen airport. This is because neither of two datasets can track the residents using private vehicles and lacking cellphone activities. In short, Figure 1 indicates that both the cellphone and transit data are the single-source data that result in biased samples, a key drawback if used alone to infer residents' mobility, motivating us to explore multi-source data as follows.

3.2 Potential of Using Multi-Source Data

To investigate the potential of multi-source data, we render the daily mobility patterns obtained from cellphone and transit data in Shenzhen in Figures 2 and 3. The size of a vertex indicates the number of residents in an urban region, and the thickness of an edge indicates the volume of mobility. We found both similarities (e.g., travel to the downtown area) and differences (e.g., travel to the commercial area) in these two figures. Both the similarity and difference suggested to us to utilize the Pearson coefficients to explore their correlation in Figure 4. The correlation between the cellphone and transit data in the early morning is higher than at other time periods, which is because most residents use taxicabs to reach the airport in the early

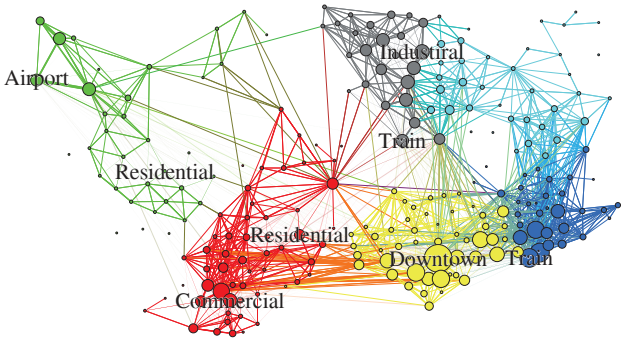

Fig 2: Human Mobility from Cellphone Data

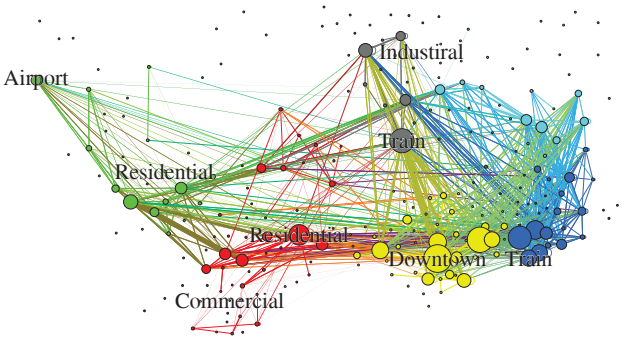

Fig 3: Human Mobility from Transit Data

morning, when public transit systems are not in service. Those residents are more likely to use their cellphones during such trips, leading to relatively high correlation. In contrast, during the rest of the day, the correlation is low because most residents reach the airport through public transit such as subway and buses. It is inconvenient to make phone calls in crowded environments, especially in the rush hour. The average Pearson coefficient value during the period is 0.22, indicating that cellphone and transit data are not highly correlated and have individual diversities.

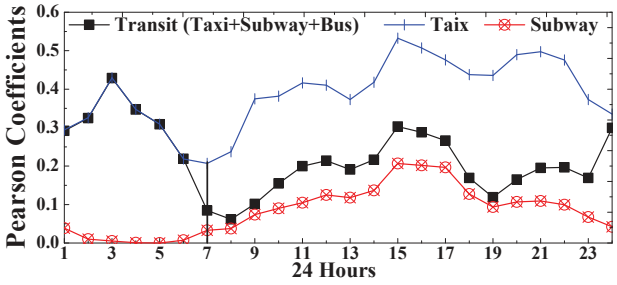

Fig 4: Correlation of Cellphone and Other Data

3.3 Summary

Our observation of our empirical data leads us to conclude that (i) the mobility obtained from cellular infrastructure or transit systems alone cannot unveil overall human mobility, and (ii) the correlation between these multi-source data suggests that they can be exploited together to compensate for the individual limitations thanks to their inherent diversities. These conclusions motivated us to investigate a novel approach using multi-source data. In the rest of this paper, we present our mPat architecture as well as a reference implementation in Shenzhen.

4. MPAT ARCHITECTURE

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

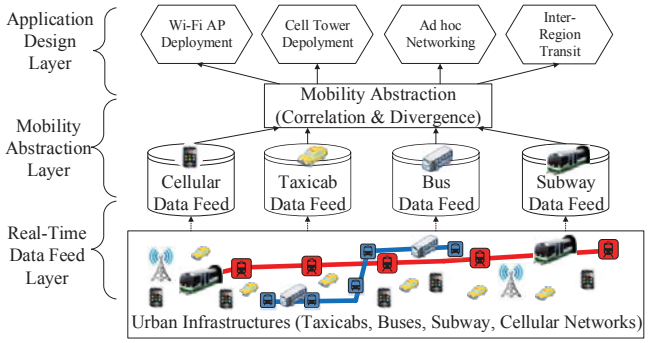

Fig 5: mPat Architecture

Real-Time Data Feed Layer ensures a secure and reliable feeding mechanism to establish multi-source data feeds through urban infrastructures. At the macro level, mPat establishes the data feed for anonymous cellphones in the cellular networks; at the micro level, mPat establishes the data feeds in the transit networks including taxicabs, buses and subways. The details are given in Section 5.

Mobility Abstraction Layer transparentizes heterogeneous features in the multi-source data to enable an effective mobility abstraction in an urban partition with extracted trips. As a result, mPat provides novel mobility analyses to unveil both the correlation and divergence between residents using cellphones and those using urban transit, which are exploited for a real-time mobility inference. The details are given in Sections 6 and 7.

Application Design Layer bridges our mobility abstraction to real-world applications to improve urban efficiency, such as 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 8.

Similar to the IP layer, i.e., the narrow waist of the Internet, the mobility abstraction layer essentially serves as the narrow waist of mPat, allowing a separation between the data feeds and applications. Based on the real-time input from the data feeds, the mobility abstraction provides appropriate service interfaces for accurate rendering of human mobility, which are then 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 urban infrastructures in Shenzhen as follows.

5. REAL-TIME DATA FEED LAYER

In this section, we first introduce the data feeds and then present our data storage and management.

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.

Cellphone Dataset		Taxicab GPS Dataset	
Collection Period	10/01/13-Now	Collection Period	01/01/12-Now
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-Now	Collection Period	07/01/11-Now
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

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.

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 6 as follows.

6. 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.

6.1 Building Blocks

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.

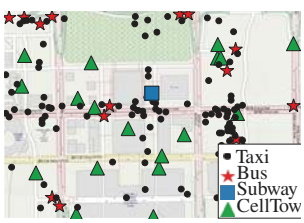

Fig 7: Granularity

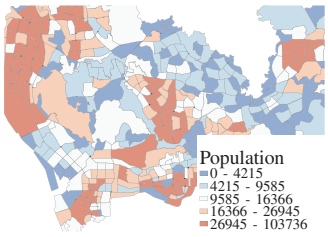

Fig 8: Spatial Partition

Spatial Partition: The spatial partition is dependent on the spatial granularity of the collected data, which is shown in Figure 7. 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 7. 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 [17]. 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, residential and traffic features, as shown in Figure 8 in which the color of regions indicates the population density.

6.2 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.

6.2.1 *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 9. 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.

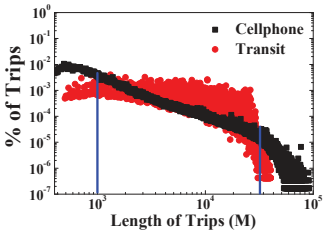

Fig 9: Length Distribution

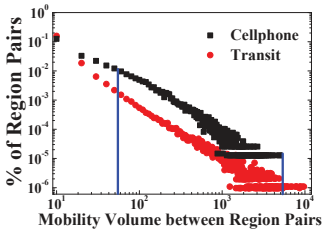

Fig 10: OD Distribution

To investigate the origins and destinations (OD) of trips, we study the volume distribution of the trips among 496 × 496 region pairs (i.e., OD combinations) based on their ODs, shown in Figure 10. 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 the cellphone trips are spatially distributed more evenly than the transit trips. This is because the cellphone data track residents between almost all region pairs, but the transit data only track residents between the 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.

6.2.2 *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 11. This probability is obtained by dividing the number of trips 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.

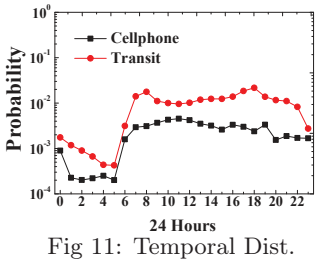

Fig 11: Temporal Dist.

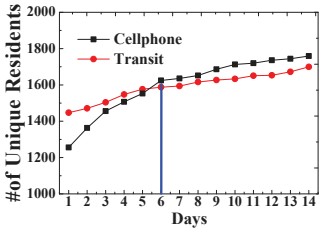

Fig 12: Cumulative Tracking

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 12, 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.

6.2.3 *Divergence*

In addition to the spatio-temporal correlation, 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 13 and 14.

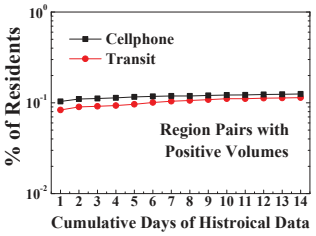

Fig 13: Divergence 1

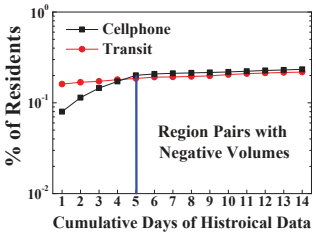

Fig 14: Divergence 2

For the region pairs with positive volumes (accounting for 31% of all region pairs), Figure 13 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 14 shows a similar plot to that in Figure 13, except that Figure 14 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.

6.2.4 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 6.2.1; (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 6.2.2; 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 6.2.3.

6.3 Online Human Mobility Inference

In this subsection, 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. To this end, 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 the mobility between two associated regions; (iii) a weight on an edge indicates the 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.

6.3.1 Inferring 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 works 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. In particular, we first obtain G^c for the residents with cellphone activities based on the real-time online cellphone data. 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 12. But due to the big size (several TBs) of our historical data and the real-time requirement for the online inference, we iteratively use partial historical data by batches as the retrospective data to estimate $G^{\bar{c}}$ in cumulation, instead of using all historical data at once. Such iterative estimation is challenging, because we could easily overestimate or underestimate $G^{\bar{c}}$ due to lack of knowledge on when to stop the iterative estimation.

To address this challenge, we use mobility graph G^t for the residents using urban transit as a threshold to stop the iteration if the cumulated $G^{\bar{c}}$ plus G^c completely covers G^t. The rationale behind this condition is based on our observation of the mobility divergence in Figures 13 and 14 in Section 6.2.3. For the region pairs with negative 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 only find few new residents compared to the residents they already found as in Figure 13. 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 G^t captured by the current transit data, as in Figure 14. Consequently, if $G^{\bar{c}} + G^c$ covers G^t, we stop continuing to iteratively train the historical cellphone data. Based on previous our analyses, if we stop the iteration earlier, we may underestimate the number of new residents; but if we stop the iteration later, we may obtain the limited new residents yet require much longer processing time. Such a balance is evaluated in Section 7.4.

Note that we use the transit data to assist in the utilization of the cellphone data instead of the other way around, which is because by our analyses in Sections 6.2.1 and 6.2.2, if combined with the historical data, the cellphone data are more effective than the transit data for tracking residents online.

6.3.2 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 6.3.3. Another input is the historical cellphone dataset $D[1, ..., K]$ of K days of the data related to the cellphone users without activities in the latest slot, which are used to infer the repeatable trips that were not captured by today's data due to lack of cellphone activities. Given them, our inferring algorithm is described as follows.

Algorithm 1 Inference of Mobility Graph G

Require: (i) G^c; (ii) G^t; (iii) $D[1, ..., K]$;
Ensure: Inferred G;
1: $k \leftarrow 0$;
2: repeat
3: Estimate $G^{\bar{c}}$ based on $D[K-k, ..., K]$;
4: $k \leftarrow k + 1$;
5: until $G^c + G^{\bar{c}}$ covers G^t or $k = K$
6: $G \leftarrow G^c + G^{\bar{c}}$;

The rationale behind our algorithm is as follows. We first infer $G^{\bar{c}}$ for the cellphone users without activities by the Kth related day of data in $D[K]$ as the retrospective data, e.g., the same day of the week in the previous week. On one hand, if $G^c + G^{\bar{c}}$ completely covers G^t, then it suggests that the mobility pattern of the cellphone users (shown by $G^c + G^{\bar{c}}$) has covered the mobility pattern of the residents using urban transit (shown by G^t). It suggests that the inferring process after this stage will track limited residents, so we use $G^c + G^{\bar{c}}$ to infer G and stop the algorithm. On the other hand, if $G^c + G^{\bar{c}}$ does not cover G^t, then it suggests that the training process after this stage will continue to track more residents, so the current $G^c + G^{\bar{c}}$ cannot be used to infer G. Thus, we continue to cumulatively estimate $G^{\bar{c}}$ based on one more previous related day of data as the increased retrospective data about these uncaptured cellphone users (i.e., the retrospective data increase to the data $D[K-1, K]$ including both $(K-1)$th and Kth related day), which is to compensate for the situation where a resident who still has no cellphone activities in the Kth related day. At the worst-case scenario, we go through the historical datasets of all related days to cumulatively estimate $G^{\bar{c}}$. Note that we increase the retrospective data from the latest to the earliest due to the freshness of the latest data.

These are three unresolved issues in the algorithm: (i) how to infer G^c and G^t, (ii) how to determine if $G^c + G^{\bar{c}}$ covers G^t, and (iii) how to iteratively estimate $G^{\bar{c}}$. These issues are discussed in the following three subsections.

6.3.3 G^c and G^t Inference

The inference about G^c and G^t is a classic topic for many data mining applications [7]. To focus on our key novelty, 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 and three real-time contexts, i.e., the current location, the time of day and the day of week. With these distributions, we obtain G^c and G^t with a straightforward probabilistic method. Note that this probabilistic method has a high accuracy because the majority of residents are regular commuters traveling between home and workplace. For example, Figure 16 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 15 gives the average subway travel time in a week between Shenzhen train station and 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.

6.3.4 G^t Coverage

In this work, $G^c + G^{\bar{c}}$ covers G^t, if the mobility in $G^c + G^{\bar{c}}$ spatially and temporally covers the mobility in G^t. This special spatio-temporal correlation can be validated by an edge-based weight comparison, i.e., $G^c + G^{\bar{c}}$ covers

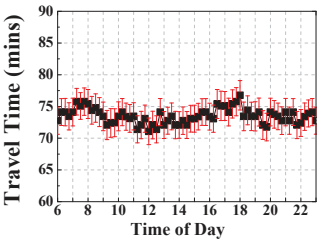

Fig 15: Travel Time

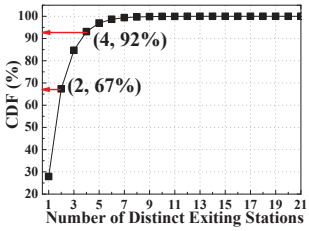

Fig 16: Distinct Stations

G^t, if the weight of every edge in $G^c + G^{\bar{c}}$ is larger than or equal to the weight of the corresponding edge in G^t. This edge-based weight comparison is possible because the two involved graphs (i.e., $G^c + G^{\bar{c}}$ and G^t) have the same vertices (due to self loops for every vertex associated to an urban region) and edges, and the differences are weights on edges (if no mobility exists between two vertices, the corresponding weight is 0). As a result, the fact that a graph $G^c + G^{\bar{c}}$ covers G^t can be validated by a condition: the weight (i.e., the volume) of every edge in $G^c + G^{\bar{c}}$ is higher than the weight of the corresponding edge in G^t. Such a validation is scalable and can be completed in the polynomial time due to the limited number of edges.

6.3.5 $G^{\bar{c}}$ Estimation

Iteratively estimating $G^{\bar{c}}$ by the cumulatively increasing retrospective data is equal to iteratively increasing the weights on edges of $G^{\bar{c}}$. As in Algorithm 1, we increase weights on edges of $G^{\bar{c}}$ by the retrospective data on a daily basis, which is based on the observation that most residents move among regions in regular daily patterns [17], e.g., daily commutes. So if they were not captured by the cellphone data today, they may be captured during a time with the similar contexts, e.g., the time of the day and the day of the week. As a result, for every uncaptured cellphone user due to lack of cellphone activities in the latest slot, we validate if this user was captured by the data of the previous related days in the retrospective data. If this user was indeed captured, we increase the weight on the corresponding edge of $G^{\bar{c}}$ to reflect this new captured mobility.

The key step for our estimation is to find the related historical data $D[1, ..., K]$. In this work, we categorize the cellphone data by two contexts, the time of the day and the day of the week. Then, for a particular time slot of a day, the related historical data are the historical data with the same 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, the related historical data are the historical data belonging to the 14 time slots from 7AM to 8AM, one for each of 14 previous Mondays. These related historical data $D[1, ..., K]$ would become the retrospective data $D[K-i, ..., K]$ during the ith iteration in Algorithm 1.

7. MPAT EVALUATION

7.1 Evaluation Methodology

We compare mPat with two state-of-the-art models. The Radiation model [14] infers the mobility based on the population density of an origin region and a destination region and that of the surrounding regions. To estimate the

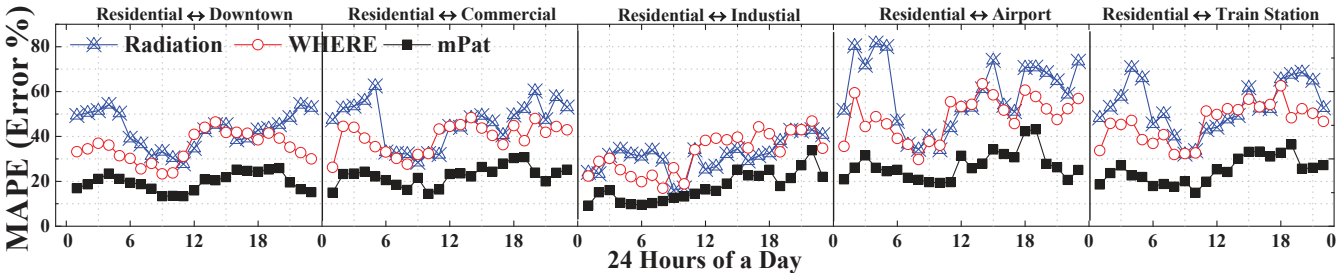

Fig 17: MAPE under One Hour Slot for 24 Hours of a day

region population, we allocate every cellphone user to a region where he/she stays the most at a particular slot based on three months of the cellphone data. Radiation serves as a statistical model suitable for the situation where the real-time data are not available. The WHERE model [11] takes the spatial and temporal probability distributions drawn from the cellphone data and produces synthetic cellphone data to indicate the inferred mobility. WHERE serves as a single-source approach for the situation where only the cellphone data are available. Further, we study the effect of the multi-source data correlation by comparing mPat to one of its variants called mPat-S, which uses all historical data as the retrospective data for the inference.

We utilize three months (91 days) of datasets from all feeds in Figure 6 on a rolling basis. We divide the data into two subsets: the testing set including the data for one particular day as the real-time streaming data; and the historical set including the data for the remaining of 90 days as the historical data. For a particular day, if we use one hour slots, at the end of the first slot, i.e., 1AM, we use mPat to infer the mobility for the slot from 12AM to 1AM, based on both the "real-time" data from 12AM to 1AM in the testing set, and all the data in the historical set. We test the models with Mean Average Percent Error (MAPE) in a time slot as 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 the inferred mobility between a region pair i; $\bar{\mathbf{T}}_i$ is the ground truth of the mobility between a region pair i. We move the 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 $\bar{\mathbf{T}}$ of urban mobility, unless we put a GPS tracker on every resident in a city for 24 hours a day. In order to infer the ground truth, we introduce another novel cellphone related dataset for the evaluation purpose. It contains regular location updates of 7.6 million active cellphone users in Shenzhen with 200 million records per day. 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, whether the users have cellphone activities or not. The only requirement is that the users have to turn on the cellphones. We use the mobility obtained from this dataset as the ground truth for the evaluation. We believe the mobility based on this new dataset is much closer to the actual ground truth than the previous methods.

Note that unlike call detail records (CDRs) which are stored for billing purposes during regular cellular operations, this "ground truth" dataset is obtained through extensive instrumentation of cellular infrastructures. It requires significant extra support in terms of software, hardware, and regulation policies. For example, this "ground truth" dataset grows more than 100GB per day and consolidating such a large dataset also requires a significant investment on computing infrastructure. Consequently, without immediate benefits, the cellphone service providers have no incentive to store this large-scale streaming dataset in regular operations, and hence it cannot be used as a regular input of mPat. In our particular case, this "ground truth" dataset was obtained during peak performance testing. Thus, unlike the generic datasets we introduced as the input of mPat, we obtained this new dataset from the providers offline and did not have regular real-time access to it.

With the "ground truth" dataset available, we first compare the inference models on two different spatial partitions, i.e., urban region level and road segment level, to show their accuracy under different granularity. Further, we investigate the impact of the historical data size on the accuracy and the running time of mPat to show its feasibility and robustness for the real-time inferences, justifying the mPat design. Finally, we present a summary.

7.2 Accuracy on Region Levels

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

Figure 17 plots the MAPE under one hour slots with the two-way mobility between a residential region and a downtown region, commercial region, industrial region, airport region, and train station region. In general, mPat outperforms WHERE, which outperforms Radiation. This is because Radiation considers only the population to infer the mobility instead of the historical trips, and WHERE uses only the cellphone data and does not correlate them to the transit data. In particular, the performance gain between mPat and others is lower during the rush hour. This is because the repeatable mobility patterns are higher in the rush hour, so all models have a better performance. Comparing the five region pairs, we found that for the region pairs on which the residents go for commutes (e.g., between the residential region and the industrial region), all models perform better than the region pairs between which residents go for traveling (i.e., between the residential region and the airport region). This is because the repeatable mobility pattern of such traveling is limited.

Figure 18 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 17. This is because urban mobility may change dramatically be-

tween different region pairs, and some remote region pairs with few transit services lead to higher MAPE. But the relative performance between the three models is similar to Figure 17. WHERE outperforms Radiation by 19% on average, but in the morning rush hour Radiation outperforms WHERE by 9%. mPat outperforms Radiation by 57% and WHERE by 39%, resulting from its utilization of the correlation between the cellphone and transit data.

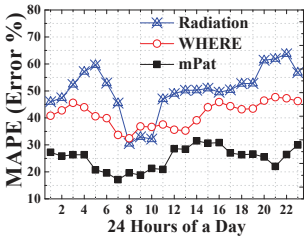

Fig 18: Hourly MAPE

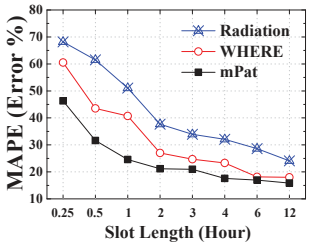

Fig 19: Effects of Lengths

Figure 19 plots the MAPE of all models with different slot lengths. The MAPE of all models reduces with an increase in the lengths of the time slots, because the urban mobility in a longer slot becomes more stable. mPat outperforms WHERE and Radiation significantly if the slot is shorter than 2 hours. But when the slot becomes longer than 6 hours, mPat and WHERE have a similar performance, because in such a long time slot, the cellphone data alone have a satisfactory performance to infer the mobility.

7.3 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 20 plots the MAPE of the inferred mobility from one of the busiest streets in the downtown Fu Hua Road to the busiest street in a residential area Le Yuan Road. 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.

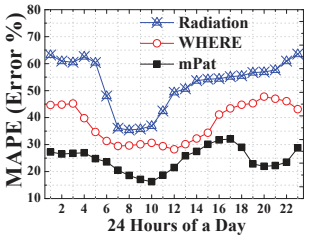

Fig 20: MAPE in One Pair

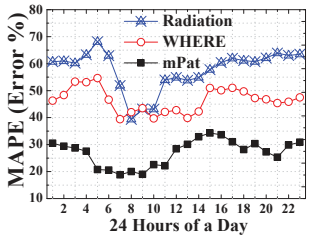

Fig 21: MAPE in 1000 Pairs

Figure 21 shows the average MAPE between 1,000 randomly selected segment pairs. The MAPE of all three models are higher than the MAPE in Figure 20 although their relative performance is similar to that in Figure 20. mPat is better than WHERE, but WHERE is normally better than Radiation. Compared to Figure 18, we found an 11% accuracy loss in mPat due to the finer granularity.

7.4 Justification of mPat's Iterative Design

We justify the design of mPat by comparing it to mPat-S, which does not use the transit data as a stop condition to iteratively utilize the historical data. Figures 22 and 23 plot the MAPE and running times on different data sizes. The more historical data, the lower the MAPE error and the longer the running time. mPat-S has a 9% accuracy gain over mPat, but mPat saves 71% of the running time. This is because with the same historical data, although mPat has the same historical data as mPat-S, mPat only uses a part of the historical data if the inferred mobility covers the transit mobility. This key feature of mPat leads to slightly higher MAPE but significantly shorter running times, which justifies the design of mPat.

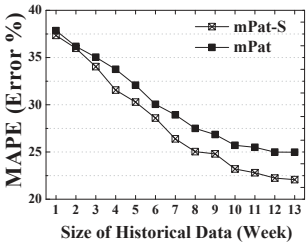

Fig 22: Data vs. MAPE

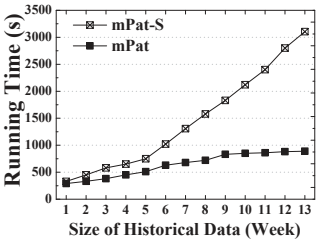

Fig 23: Data vs. Time

7.5 mPat Evaluation Summary

We make the following observations: (i) The inference accuracy is highly dependent on both locations and times as in Figures 17 and 18. On average, all models have better performance in the morning rush hour, due to the predictability of the morning commutes. (ii) The length of slots has a significant impact on the performance of all models as in Figure 19. It is intuitive that a longer slot has lower error rates, yet it also leads to low usability for real-time applications. (iii) The spatial granularity also has an impact on performance as seen by comparing Figures 20 and 21 to Figures 17 and 18, but the accuracy loss due to increasing granularity is not significant. (iv) Running time can be reduced by 71% with a 9% accuracy loss by using the correlation between cellphone and transit data as in Figures 22 and 23.

8. 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 [13]. 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.

8.1 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 times 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 24 gives an IRT example.

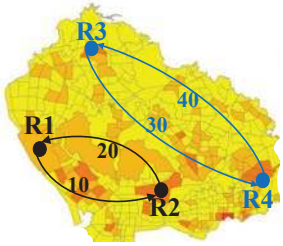

Fig 24: IRT Example

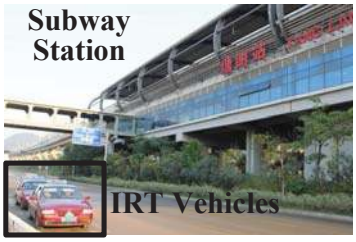

Fig 25: IRT Implementation

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.

8.2 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 25, 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 26, 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.

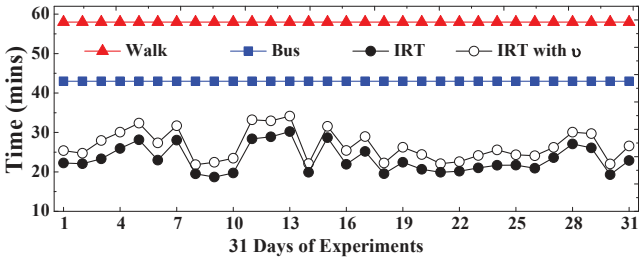

Fig 26: Average Travel Time in 31 days

8.3 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, 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 27 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 a similar travel time as IRT.

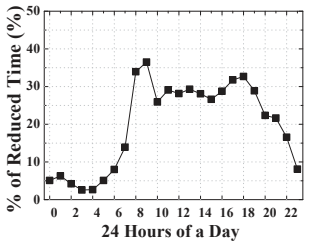

Fig 27: Reduced Time in 24H

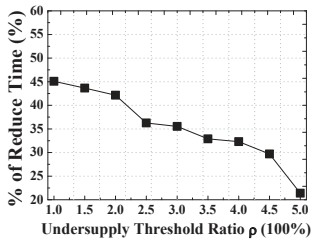

Fig 28: Reduced Time vs. ρ

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 28 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.

9. DISCUSSION

We discuss some issues about our architecture as follows.

Value of mPat Mobility Study: In this work, we use transportation as an example to show the benefit of human mobility models that serve as a key supporting layer for numerous applications. Basically, since most personal mobile devices are carried by users, the mobility patterns of the users are also the mobility patterns of these devices. Built upon the mobile devices' future locations and associated dwell time obtained from these mobility patterns, some typical mobile and wireless applications with great social or commercial benefits include Wi-Fi AP or cell tower deployments, ad hoc networking, infrastructureless message disseminations, and mobile network pricing.

Privacy Protections: Although data that can be used for the study of human mobility study are extremely valuable, privacy concerns regarding their utilization have inhibited their release and wider applications. We briefly discuss the active steps we took to protect privacy. (i) Anonymization: All data from feeds are anonymized by the employees of the service providers who are not involved in 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 information that is useful for our mobility analysis, and delete other information for 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 in aggregated results in a temporal-spatial partition and are not focused on individual cellphone or transit users.

Higher Dimension Multi-Source Traces: In the current Shenzhen implementation, we focus only on transit and cellphone activity datasets. As to future work, other urban trace data will be included in our mPat architecture to increase the accuracy of analyses and inferences. These datasets include (i) data of a growing bicycle network with 8,000 bicycles in Shenzhen for rental using smart cards; and (ii) data of the private vehicles that are tracked by onboard GPS installed by themselves or the manufacturers for location-based services. We envision combining these trace data about private transit with public transit data and then study their relationship with the cellphone data for a more accurate inference. To realize such a vision, we must provide incentives and privacy-preserving mechanisms to encourage residents to opt in.

10. CONCLUSION

In this work, we design, implement and evaluate an architecture for the analysis and inference of the human mobility with a 75% inference accuracy. Our endeavors offer a few valuable insights are that (i) the studies based on single-source data introduce biases into human mobility research, and the correlation as well as the divergence among multi-source data have the potential to address such biases; (ii) multi-source data can be used for cross-referencing in the real-time inferences to reduce the data needed to be processed for a shorter running time yet still maintain performance; and (iii) while it is challenging to integrate heterogenous large-scale feeds, it is more challenging to negotiate with service providers for real-time access and to protect the privacy of studied residents.

11. ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their valuable feedback. This work was supported in part by the US NSF Grants CNS-0845994, CNS-1239226 and China NSFC Grant 61100220, as well as the UMN Thesis Travel Grant and the visiting support by Cloud Computing Center in SIAT.

12. REFERENCES

[1] Sample Datasets. In *http://www-users.cs.umn.edu/~zhang/*.

[2] Backstrom, L., Sun, E., and Marlow, C. Find me if you can: Improving geographical prediction with social and spatial proximity. In WWW '10.

[3] Bhattacharya, S., Phithakkitnukoon, S., Nurmi, P., Klami, A., Veloso, M., and Bento, C. Gaussian process-based predictive modeling for bus ridership. UbiComp '13.

[4] Bogo, F., and Peserico, E. Optimal throughput and delay in delay-tolerant networks with ballistic mobility. MobiCom '13.

[5] Cho, E., Myers, S. A., and Leskovec, J. Friendship and mobility: User movement in location-based social networks. KDD '11.

[6] Dufková, K., Le Boudec, J.-Y., Kencl, L., and Bjelica, M. Predicting user-cell association in cellular networks. MELT'09.

[7] Ganti, R., Srivatsa, M., Ranganathan, A., and Han, J. Inferring human mobility patterns from taxicab traces. UbiComp '13.

[8] Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., and Trasarti, R. Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal.

[9] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. Understanding individual human mobility patterns. Nature.

[10] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Rowland, J., and Varshavsky, A. A tale of two cities. In HotMobile '10.

[11] Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., and Willinger, W. Human mobility modeling at metropolitan scales. MobiSys '12.

[12] Lathia, N., and Capra, L. How smart is your smartcard?: Measuring travel behaviours, perceptions, and incentives. UbiComp '11.

[13] Shenzhen Customized Public Transit . In http://house.people.com.cn/n/2014/0411/c164220-24882370.html.

[14] Simini, F., Gonzĺćlez, M. C., Maritan, A., and Barabĺćsi, A.-L. A universal model for mobility and migration patterns. Nature.

[15] Song, C., Qu, Z., Blumm, N., and Barabasi, A.-L. Limits of predictability in human mobility. Science.

[16] Zhang, D., Li, Y., Zhang, F., Lu, M., Liu, Y., and He, T. coRide: Carpool Service with a Win-win Fare Model for Large-scale Taxicab Networks. SenSys '13.

[17] Zheng, Y., Liu, Y., Yuan, J., and Xie, X. Urban computing with taxicabs. UbiComp '11.