This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology

A Mobility Analytical Framework for Big Mobile Data in Densely Populated Area

Yuanyuan Qiao, *Member, IEEE,* Yihang Cheng, Jie Yang, Jiajia Liu, *Senior Member, IEEE,* and Nei Kato, *Fellow, IEEE*

***Abstract*—Due to the pervasiveness of mobile devices, a vast amount of geo-located data is generated, which allows us to gain deep insights into human behaviors. Among other data sources, the analysis of data traffic from mobile Internet enables the study of mobile subscribers' movements over long time periods at large scales, which is paramount to research over a wide range of disciplines, e.g., sociology, transportations, epidemiology, networking, etc. However, in order to efficiently analyze the massive data traffic from the view of user mobility, several technical challenges have to be tackled before releasing the full potential of such data sources, including data collection, trajectory construction, data noise removing, data storage, and methods for analyzing user mobility. This paper introduces a mobility analytical framework for big mobile data based on real data traffic collected from 2G/3G/4G networks covered nearly 7 million people. In order to construct user's history trajectories, we apply different rules to extract users' locations from different data sources, and reduce oscillations between the cell towers. The comparison of mobility characteristics between our mobile data and other existing data sources shows big potential of mobile Internet data traffic to study human mobility. In addition, our experiments discover the changing of city hotspots, the movement patterns during peak hours, and people with similar history trajectories, which uncover the common rules exist among huge population in city.**

***Index Terms*—Big Mobile Data, Human Mobility, Mobility Analytical Framework, Mobile Internet.**

I. Introduction

STUDY of human mobility would yield insight into a variety of social issues on geographical scales, such as urban planning [1], population distribution [2], and disease spreading [3]. As we know from our daily life, the movement of people in space is far from random. However, finding the formulation of quantitative laws explaining human mobility, which is essential to uncover the mechanisms governing human activities [4]–[9], still remains as an open question. There are essentially two ways for studying the nature of mobility: synthetic models and traces. Synthetic models attempt to represent the human behaviors by sets of mathematical equations, such as random mobility models (Lévy Walk [10], Random Walk [11]), models with temporal dependency (Gauss-Markov [12], Smooth Random [13]), models with spatial dependency (Probabilistic Random Walk [14]), models with geographic restriction (Pathway or City Section [15], Obstacle [16]) and so on [17]–[19]. Above models are tractable, scalable, easy to deploy, and especially useful in the field of ad hoc networks if traces have not yet been created [20]–[23]. Though synthetic models can easily reproduce human mobility patterns up to acceptable degree of accuracy, it is still quite difficult to assess to what extent they map reality, and the generated trajectories are different from those observed in real scenarios [24]. On the other side, real human traces provide accurate information, which traditionally are restricted by expensive data-collection methods, are of theoretical and practical significance in the area of mobility analytics.

Nowadays, smart devices bring us the ubiquitous mobile Internet access. People's movements could be sensed and easily collected by mobile phone, which generating large volumes of mobility data, such as Call Detail Records (CDRs), and Global Positioning System (GPS) tracks. CDRs provide the time a phone placed a voice call or received a text message, as well as the identity of the cell tower with which the phone was associated at the time [6], [8], [25]–[27]. Yet they are sparse in time and coarse in space, which limit the scope of their application to study human mobility. As for GPS tracks, the movements of individuals in latitude and longitude along with timestamp are recorded [28]. However, GPS signals may easily become unavailable in indoor or underground environments, GPS devices may get interferences in environments with high building density, and users are becoming more reluctant to share locations because continuously collecting GPS data may consume devices' energy quickly or make people uncomfortable considering privacy issues [29]. Due to above matters, up to now, there does not exist any GPS data source covering citywide population.

Recently, researchers found that data traffic from 2G/3G/4G data networks is extremely useful for studying human dynamics [30]–[32]. Passively collecting human movement trajectories while he/she is accessing to mobile Internet has lots of advantages: high cost efficiency, low energy consumption, covering a wide range and a large number of people, and with fine time granularity (people tend to surf mobile Internet frequently while moving, and many apps may send or receive network traffic packets periodically when running in the background). Collected trajectories are coarse in space because they record location only at the granularity of a cellular antenna

Yuanyuan Qiao, Yihang Cheng and Jie Yang are with Center for Data Science & Beijing Key Laboratory of Network System Architecture and Convergence & Beijing Laboratory of Advanced Information Network, the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China (e-mail: yyqiao@bupt.edu.cn; mrcheng@bupt.edu.cn; janeyang@bupt.edu.cn).

Jiajia Liu is with School of Cyber Engineering, Xidian University, Xi'an, China (e-mail: liujiajia@xidian.edu.cn).

Nei Kato is with Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan (e-mail: kato@it.ecei.tohoku.ac.jp).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology

(with average error of 175 meter [33], and the density of cell tower is much larger in urban area than in suburb or rural area due to the human population density). This error range is tolerable, and the analytical results are convincing enough to find fundamental laws in human dynamics [6], [34], to build individual mobility models [35] or aggregated mobility models [7], even enough to get a dynamic understanding of the population, activities, and environment [36]–[41].

According to the prediction in [42], monthly global mobile data traffic reached 2.5 exabytes at the end of 2014, and will surpass 24.3 exabytes by 2019. The explosion in data traffic amount brings many opportunities to obtain data source with rich information. However, existing methods are not prepared to deal with such huge volume of traffic data. New methods to solve great challenge for data collection, storage and analysis of big mobile data are needed urgently. As motivated by such observation, in this paper, our goal is to present a framework for efficiently analyzing massive data traffic from the view of user mobility in densely populated area. The contributions of our work are summarized as follow:

- To the best of our knowledge, we are the first to present a cloud computing based analytical framework to analyze big mobile data from the view of user mobility covering the mobile networks of 2G/3G/4G and the scale for nearly 7 million people. Big data technologies and analytical algorithms are used to store and process massive data traffic. In particularly, our framework is developed for analyzing user mobility pattern based on real mobile Internet data collected from 2G/3G/4G networks.
- Our framework is suitable for human trajectories consisting of a series of positions of cell towers, including CDRs. Since there is noise in raw data, different rules are required to construct human trajectory from different data sources. Towards this end, we define raw data processing rules for constructing human trajectory from different interfaces of 2G/3G/4G networks, and remove data noise by reducing oscillation between cell towers. In addition, in order to ensure the effectiveness of our data set, we calculate three widely accepted mobility indicators, i.e., the trip distance distribution, the radius of gyration, and the number of visited locations over time. Our results show that, for the same indicator, different data sources follow similar models but with different values of parameter.
- We further use our framework to explore human movement behavior in densely populated area. We employ a parameter free method to identify city hotspots from the view of population, apply a modified version of the Apriori algorithm to mine maximal sequential pattern, discover similar users based on their history trajectories, and predict users' future movements from both temporal and spatial perspectives. These functionalities are of significant meaning for improving user experience of Location Based Service (LBS), optimizing network resources, and advising city planning.

The remaining of the paper is organized as follows. In section II, related works in the field of mobility analytics are introduced. Section III provides the overall structure of our mobility analytical framework, including data collection methods for 2G/3G/4G networks, rules for constructing human trajectories, the design of database, and algorithms for mobility analysis. Section IV gives experimental results from our framework based on real data traffic. Conclusions are drawn in Section V.

II. Available Works Related to Mobility Analytical Frameworks

In order to study the inherent properties of human mobility in an efficient way, a mobility analytical framework, aiming to analyze big mobile data by providing data collection, data storage, data pre-processing function, and mobility functionalities, is essential. Mobility profiler [38] is a complete framework for discovering mobility profiles from raw cell tower connection data. It removes the cell tower oscillation, and constructs a cell tower topology to discover user's movement pattern. Framework "Jyotish" [43] constructs a predictive model by exploiting the regularity of people movement found in real joint Wifi/Bluetooth trace. With the rising of Social-Location-Mobile (SoLoMo), based on Big Data platforms of IBM, Heng Cao et al. presented a unified SoLoMo analysis approach from system-oriented view [44], which is designed to process the vast amount of data generated in the telecom area every day. Ying Zhang [30] proposed a systematic analysis methodology that considered inaccuracies from cellular data network. Although, above frameworks have solved some issues regarding mobility analytics, none of them covered the whole procedure for data traffic analysis (i.e., data collection, data storage, noise removing, trajectory construction, and data analysis from the view of user mobility), considered the methods for big data storage and analysis, or gave detailed experiments with real-world data traffic collected from citywide area.

With the rising of mobile Internet, research institutions and enterprises pay more and more attention to LBS, and try to apply their mobility analytical frameworks to practical applications. IBM developerWorks introduces an Advance Analysis Platform (AAP) for analyzing location to discover mobility pattern. They want to apply AAP to all kinds of location data, like GPS data from cars, planes or other equipments, the using of credit cards or public transportation cards, CDR, and Deep Packet Inspection (DPI) from operator. Microsoft Research developed "GeoLife" [45], to analyze GPS trajectory and provide location-based social-networking service. GeoLife is based on a framework, Hierarchical-Graph-based Similarity Measurement (HGSM), to uniformly model each individual's location history, and effectively measure the similarity among people. Different data sources usually have different features that require distinct analysis method. For example, GPS trajectory is usually generated with occasional outliers or some noisy points caused by the poor signal of location positioning systems [46]. Raw cell tower connection data usually have "cell tower oscillation", where even when the user is static he/she may be assigned to a number of neighboring cell towers because of load-balancing issue or changes in the ambient Radio Frequency (RF) environment [38]. Hence, for different kinds of spatial trajectories, specific rules must be considered.

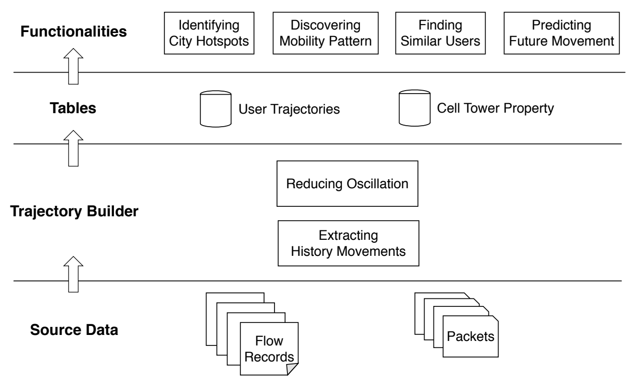


Fig. 1. The architecture of mobility analytical framework.

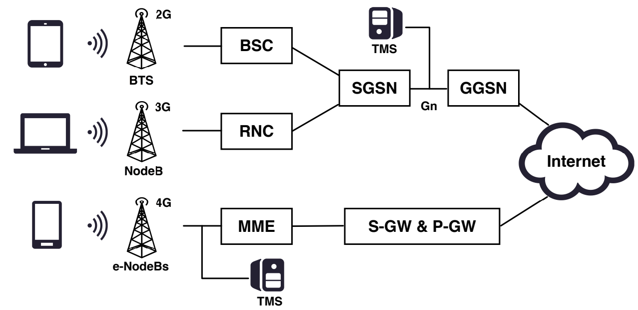


Fig. 2. Mobile Internet network architecture and the deployment of TMS.

Different from the previous work, this paper aims at developing a framework for user mobility analytics based on massive mobile Internet data traffic, which integrates (1) the techniques for big data collection, storage, and pre-processing, (2) the rules for extracting location data, and constructing people trajectories, (3) the methods for solving data noise (i.e., cell tower oscillation), and (4) the algorithms for discovering common mobility patterns in densely populated area. The real mobile Internet data traffic collected from 2G/3G/4G networks covering millions of people is used to verify the effectiveness of our framework. Our framework could be applied to analyzing human mobility with mobile Internet data traffic, and it is especially useful for efficiently processing big mobile data.

III. Methodology

In this section, we introduce a mobility analytical framework to analyze massive data traffic from mobile Internet. As shown in Fig. 1, we have two kinds of data sources, i.e., flow records and packets. Here, we define "flow" as bidirectional data transmission at the usual 5-tuple source Internet Protocol (IP), destination IP, source port, destination port, and transport protocol within a certain period of 64 seconds. Our framework is based on the cloud computing platform, which is the best tool to handle big data in present. It consists of trajectory builder (removing data noise, and extracting user history movements from cell tower ID sequence), database (storing users' trajectories, and cell tower property), and functionalities (mobility analytics based on users' trajectories).

A. Data Collection

By deploying our self-developed Traffic Monitoring System (TMS) at the core network edge connecting to the 2G/3G/4G network interfaces, data traffic generated by User Equipment (UE), such as smart phone, tablet, laptop computer equipped with mobile broadband adapter or any other devices that access to Internet through 2G/3G/4G networks, is collected. As shown in Fig. 2, in 2G or 3G networks, UE communicates with a Base Transceiver Station (BTS) or Node B which transmits its network traffic to Base Station Controller (BSC) or Radio Network Controller (RNC). The controllers (BSC/RNC) then deliver the network traffic to a Serving GPRS Support Node (SGSN) that establishes a tunnel on Gn interface (Interface between the GGSN and the SGSN) with Gateway GPRS Support Node (GGSN) through which the data enters the Internet (GPRS represents "General Packet Radio Service"). In 4G networks, evolved Node B (eNodeB) establishes the connection between UE and Mobility Management Entity (MME). Users' data traffic goes into Internet through Serving GateWay (S-GW) and Packet Data Network (PDN) GateWay (P-GW).

In 2G/3G networks, we collect mobile Internet traffic from Gn interface, and store the data traffic as flow records. In order to get people's location information from 4G networks, we collect LTE control-plane packets between eNodeBs and MME, which contain the whole signaling procedures, such as connection establishment, release procedure, or handover procedure. We can get a sequence of time-stamped records, each of which contains current service eNodeB ID, signaling procedure code, user ID, etc.

As network applications become increasingly complex and heterogeneous, there is an increasing need of application oriented traffic analysis. Most of the existing network traffic monitoring systems just equip physical probe to capture and store raw packets, because software-based traffic monitoring techniques are inadequate to achieve real-time monitoring. Yet TMS, which is based on a combined software/hardware architecture with flexibility to cope with the modification and addition of monitoring requirements as well as future rate increase, can conduct application oriented traffic analysis for a 10Gbps network line in real time using an 8-core machine [47].

B. Big Mobile Data Processing Platform

In order to store and process the massive data collected from big city cover large population, cloud computing based big mobile data platform with high storage capacity and computing power is essential. The mobility analytical framework is built on a Hadoop (an open-source software for reliable, scalable, distributed computing) [48] based cloud computing platform, which provides functions of data transmission, storage, processing, and management [47]. The system architecture of our platform is shown in Fig. 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology

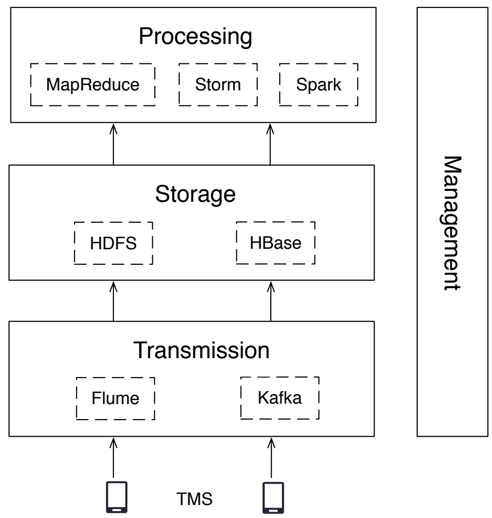


Fig. 3. System architecture of big mobile data processing platform.

1) Transmission Module: the data traffic collected by TMS is uploaded to cloud computing platform through a transmission module, which provides real-time and stable transmission by using Flume [49] and Kafka [50].

2) Storage Module: Hadoop Distributed File System (HDFS) [51] and HBase [52] are used to store massive data traffic in the form of flow records, or packets. All files in the platform are replicated for fault tolerance. Storage space of the platform can be easily extended by adding disks or new machines.

3) Processing Module: we use MapReduce [53], Spark [54] and Storm [55] to process the massive data traffic. MapReduce is a programming model and an associated implementation for processing and generating large data sets. Spark supports cyclic data flow and in-memory computing, which is very efficient for iterative and matrix computation. Storm is very useful to deal with realtime analysis.

4) Management Module: in order to monitor the whole platform, we developed a management module to monitor the status of all machines, equipments, softwares and modules. All monitoring data is collected by Flume, and stored in database. If the value of a monitoring item is over a set threshold, specific alarm information is sent to administrator via short message, email, and web interface immediately. In addition, we use Zookeeper [56] to modify configuration parameters of each machine and equipment (i.e., enabling/disabling the machines, equipments, softwares on machines, and modules of each software), and change the value of alarm threshold.

C. Trajectory Builder

In order to construct user's history trajectories, some rules should be applied while extracting user's location from data traffic. Meanwhile, data noise must be removed before analyzing data.

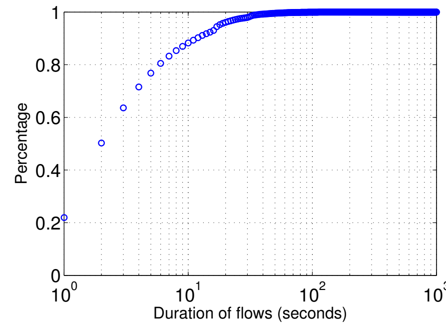


Fig. 4. CDF of flow durations for data traffic collected from 2G/3G networks.

1) Extracting History Movements: we extract user's trajectories by 4-tuple {user ID, cell tower ID, time stamp, duration}. A trajectory is constructed by a sequence of stays, and a stay is defined as,

$$S = (U, L, T, D),$$

where U is the user ID, L is the cell tower ID, T is the start time of this stay which is stored in Coordinated Universal Time (UTC), and D is the duration in seconds that a user accesses with a cell tower. Here we have user's trajectory as,

$$Traj = \langle S_1, S_2, \ldots, S_n \rangle,$$

where $S_k = (U, L_k, T_k, D_k), 1 \leqslant k \leqslant n$.

For different data sources (flow records or packets), we apply different rules to construct user's trajectories.

(a) Flow records as data source

For flow records, each flow will generate a stay, T is the start time of a flow, and D is the duration of that flow. Note that, one flow only records the cell tower that a user accessed to when the flow started. If the user switches to another cell tower before the current flow ends, this transition couldn't be captured. In order to examine this deviation, we further draw Cumulative Distribution Functions (CDFs) of flow duration of 2G/3G traffic data, i.e., we draw the value of flow duration on x-axis, and draw the cumulative percentage of each observed flow duration value on the y-axis. We can clearly see from Fig. 4 that over 80% of flows last less than 6 seconds, and the duration of 90% flows is less than 10 seconds. Since crossing over the coverage of a cell tower within 6 seconds is nearly impossible, we believe our dataset can capture user's movements perfectly. Note that, for those who move around the boarder area of a cell tower, we may not capture their next location correctly if he/she switches the cell tower in 6 seconds.

Since the changing of users' locations is not recorded in flows, in order to reduce the deviation, we use two rules to construct user's trajectories.

Rule 1 (merging overlapping flows with same location): if $L_k = L_{k+1}$, and $T_{k+1} < T_k + D_k < T_{k+1} + D_{k+1}$, remove S_{a_k+1}, we have $S_{a\_k} = (U_a, L_k, T_k, T_{k+1} - T_k + D_{k+1})$.

If user a generates two flows at the same cell tower, and the second flow starts before the first flow ends, two stays

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology

extracted from these two flows should be merged into one stay.

Rule 2 (identifying transitions from overlapping flows with different locations): if $L_k \neq L_{k+1}$ and $T_{k+1} < T_k + D_k$, then $S_{a\_k} = (U_a, L_k, T_k, T_{k+1} - T_k)$, $S_{a\_k+1} = (U_a, L_{k+1}, T_{k+1}, D_{k+1})$.

If user a generates two continuous flows at different cell towers, and the second flow starts before the first flow ends, the stay extracted from the first flow should end when second flow starts.

(b) Packets as data source

When a UE connects to LTE network, it will be either in active state, or in idle state. For different states, we apply two rules to construct the UE's trajectories.

In active state, LTE networks are aware of the ID of cell tower to which the UE is currently connected. Every time the UE switches to a cell tower, a stay is generated. For one user, there is no time interval between consecutive stays, and all movements (cell tower switching) can be captured.

Rule 3 (identifying transitions when a UE in active state in 4G networks): in active state, for two stays generated by user a, if T_k and T_{k+1} are the time that UE attaches to cell tower L_k and L_{k+1} respectively, we have $T_k + D_k = T_{k+1}$.

While in idle state, the UE shall initiate the tracking area updating procedure by sending a Tracking Area Update (TAU) request every 12 minutes (periodic tracking area updating is used to periodically notify the availability of the UE to the network).

Rule 4 (identifying transitions when a UE in idle state in 4G networks): in idle state, for each user, we have a stay $S_k = (U, L_k, T_k, 0)$ for every 12 minutes, and movements between each TAU is lost.

2) Reducing Oscillation: mobile phone may switch between different cells even user is not mobile. It usually happens when a user is in the overlapping area of two or more cells. This phenomenon is called "cell oscillation" or "ping-pong effect". For example, as shown in Fig. 5, if the user's real movement is $A \to B \to C$, when oscillations happened, the trajectory extracted from data traffic would be $A \to B \to D \to B \to D \to B \to C$ (this kind of oscillations is easy to identify) or $A \to B \to D \to C$ (this kind of oscillations is very hard to identify).

Some studies handle oscillation problem by clustering cell towers, which will reduce the position accuracy [38], [57]. Recent study proposes an algorithm framework called DECRE (Detect, Expand, Check, Remove) [58], which resolve oscillation by selecting a cell tower to approximate the mobile device's actual location.

In order to reduce the oscillations effectively under big mobile data environment, we only consider two features of cell tower oscillation: (1) it happens between adjacent cells, (2) the duration for oscillations is quite short (if the duration between switching is long, we can not tell it is an oscillation or real movement). Hence, we apply a simple method to handle the oscillation problem.

Rule 5 (reducing oscillations): calculate the average displacement (switching) time d_t for the given data set. If user changes location during d_t, oscillations may happen. Replace locations that user connects during d_t with the one that has the longest accessing time during d_t.

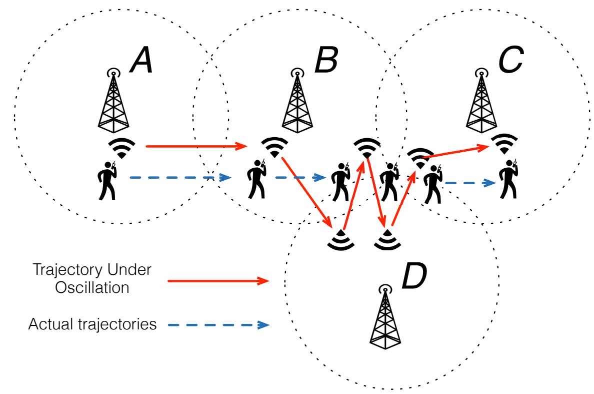


Fig. 5. An example of cell towers in mobile Internet.

Because we only capture user's movements when his/her smartphone generates data traffic, we only take consecutive stays, between which there is no time interval, into account when calculating average displacement time.

D. Data Storage

After removing data noise, we design two Tables to store the data of users' movement trajectories and information of cell towers.

1) Table 1 (User Trajectories): we store users' trajectories as 4-tuple {user ID, cell tower ID, time stamp, duration}, in order to draw user's history movements from spatial and temporal dimensions.

2) Table 2 (Cell Tower Property): for each cell tower, we store cell tower's property as 6-tuple {cell tower ID, network type, longitude, latitude, a list of adjacent cell tower, a list of semantic location tag}. Network type includes 2G/3G/4G networks, longitude and latitude are the geographical location of cell tower. Semantic location tag is the regional characteristic of cell tower (like "xx shopping mall", "xx station"), and usually includes the name of station, commercial/residential area, educational/industrial/government building, etc. A cell tower may have many semantic location tags.

Table 1 stores the most basic information of user's trajectory, which could be used to analyze mobility pattern, discover similar users or predict user's next location. After combining data in Table 1 and Table 2, advanced semantic information can be illustrated. We could predict user's future movement in spatio-temporal scene, and discover the most popular/hot/crowded place in a city.

E. Functionalities

In this section, we introduce the methods used by four mobility functionalities, including identifying city hotspots, discovering mobility pattern, finding similar users based on history movement path, and predicting user's future movements.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology

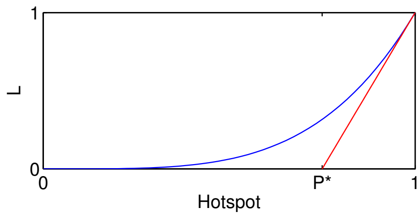


Fig. 6. Illustration of the criteria selection on the Lorenz curve. The threshold corresponds to the value of 1-P^*.

TABLE I
NOTATION DESCRIPTION OF MODIFIED APRIORI ALGORITHM

Notation	Description
δ	Minimum support threshold
$Traj\_S$	Set of mobility trajectories
H	Set of hotspots
C_k	Set of $length(k)$ candidate mobility patterns
P_k	Set of $length(k)$ mobility patterns
P	Set of all mobility patterns

1) Identifying City Hotspots: by collecting the mobile Internet data traffic generated by users, we could have a glance at structure and dynamics properties of a city. Especially, it is very important to identify the "heart" of the city, which is also called "city hotspot". If there are some abnormal changes of city hotspots, it may imply that unexpected events are happening or big event will happen soon.

City hotspots, the most significant locations along the human's trajectories, are made of the geographical area covered by one or many cell towers. Depending on the properties we want to focus on, many different kinds of city hotspots could be identified. Previous work identified city hotspots in the view of density of population [36], subscribers' mobile data [37], the residence time [59], and semantic location [46]. Above properties are characterized by a specified parameter. If the parameter value of a place is bigger than a threshold, the place could be identified as a city hotspot. Therefore, identifying hotspots is an issue of exploring an efficient threshold for a specified parameter. We employ a parameter free method proposed by Thomas Louail [36] to select the threshold. The employed method is based on the Lorenz curve.

For a specified time period, we can obtain the value of an indicator $r(j)$, such as population (number of users), from candidate hotspots. Here j is the sequence number of a candidate hotspot. We sort them in an increasing order, and then denote them as $r(1) < r(2) < \cdots < r(n)$ where n is the total number of locations. The Lorenz curve is constructed by the following way:

- On the x-axis, draw the proportion of candidate hotspots $P = j/n$, where $j = 1, 2, \ldots, n,$.
- On the y-axis, plot the corresponding proportion of the number of users with,

$$L(j) = \frac{\sum_{i=1}^{j} r(i)}{\sum_{i=1}^{n} r(i)}. \tag{1}$$

The method employs the natural way to identify the typical scale of the number of hotspots, which is to take the intersection point P^* between the tangent of $L(P)$ at point $P = 1$ and the horizontal axis $L = 0$ (see Fig. 6). Then the method gives $1 - P^*$ as the threshold for identifying the hotspots. This method is inspired from the classic scale determination for an exponential decay: if the decay from $F = 1$ is an exponential of the form $e^{(F-1)/a}$ where a is the typical scale we want to extract, this method would give $1 - F^* = a$.

2) Discovering Mobility Pattern: in the view of 2G/3G/4G networks, users' history movements are made of a series of stays, i.e., $Traj = \langle S_1, S_2, \ldots, S_n \rangle$, where $S_k = (U, L_k, T_k, D_k), 1 \leqslant k \leqslant n$. Discovering the mobility pattern of individuals or groups is the matter of finding maximal continuous trajectory. Therefore, we apply a modified version of the Apriori algorithm to discover maximal sequential pattern. A mobility pattern is identified only if the support value of discovered maximal continuous trajectory is larger than the support threshold [60]. In our case, for a set of trajectories $Traj\_S = \{Traj_1, Traj_2, \ldots, Traj_N\}$, the support value of pattern p is the ratio of the number of pattern p appeared in trajectories to the number of trajectories, which is defined as

$$supp(p) = \frac{|\{Traj_i | p \subset Traj_i \text{ and } 1 \leqslant i \leqslant N\}|}{N}. \tag{2}$$

For example, if a $Traj\_S$ contains 10 trajectories, and 6 trajectories contain the mobility pattern p, $supp(p)$ equals 0.6. Given a minimum support threshold, δ, the mobility pattern p is defined as a mobility pattern if and only if p has support value satisfying $supp(p) \geqslant \delta$.

In this paper, we discover the mobility pattern only between city hotspots. The main notations used in our method are listed in Table I and the pseudo-code of our algorithm is shown in Algorithm 1. A mobility pattern $p =< a_1, a_2, \ldots, a_n >$ is a candidate mobility pattern only if its sub-pattern $q =< a_1, a_2, \ldots, a_{n-1} >$ is discovered as a mobility pattern. For example, if $q =< a, b, c >$ is a mobility pattern, $p =< a, b, c, d >$ is a candidate mobility pattern. The main idea of this algorithm is discovering a continuous trajectory, the support value of which is larger than δ. We first calculate each hotspot's support value and the set of $length(1)$ mobility patterns are generated. And then the mobility patterns with $length(k)$ are generated through mobility patterns with $length(k-1)$. The iteration is ended when the set of $length(k)$ is $\emptyset$.

3) Finding Similar Users Based on Path: the history movements of users may reflect their relationship. If two different users have similar moving path every day, they may know each other or have the potential to be friends. The more unpopular locations (the locations that seldom people visit) they visit at the same time interval, the more likely they share similar interests. In [61], the authors mine user similarity based on GPS data collected from mobile phones to recommend friends or discover community. According to the features of our dataset, users' trajectories are extracted even when users are in big shopping mall or subway (GPS signal is not available in indoor places, underground, and the area of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology

Algorithm 1 Discovering Mobility Patterns

INPUT: Support threshold δ
  Set of mobility trajectories T
  Set of hotspots H
OUTPUT: Set of mobility patterns P

$k = 1$
$C_k = \{h | h \in H\}$
$P_k = \{h | h \in H \wedge supp(h) > \delta\}$
$P = \{\}$
repeat
  $k = k + 1$
  for all $mobility\ pattern\ p_{k-1} \in P_{k-1}$ **do**
    for all $frequent\ pattern\ p_1 \in P_1$ **do**
      $C_k = \{c_k | c_k = p_{k-1} \cup p_1\}$
    end for
  end for
  for all $trajectory\ Traj\_s \in Traj\_S$ **do**
    $C_t = subset(C_k, Traj\_s)$
    for all $candidate\ c \in C_t$ **do**
      $count(c) = count(c) + 1$
    end for
  end for
  $P_k = \{c | c \in C_k \wedge supp(c) > \delta\}$
  $P = \cup P_k$
until $P_k = \emptyset$
return P

intensive buildings), which ensure that users' daily movements in urban area are captured.

Firstly, we apply the improved Apriori algorithm to find the Maximum Similar Sequence (MSS) from two users' moving paths (path$_1$ and path$_2$).

Secondly, we calculate the "Inverse Document Frequency (IDF)" [62] for all the locations,

$$idf(s) = log\frac{N}{n}, \tag{3}$$

where N is the total number of users in the dataset, and n is the number of users visiting locations s. That is to say, if a lot of people visit location s, the value of $idf(s)$ will be very small.

Thirdly, IDF of i-th MSS for two paths are calculated as,

$$IDF(MSS_i) = 2^{|MSS_i|-1} \times \sum_{i=1}^{|MSS_i|} idf(s_i), \tag{4}$$

here, $|MSS_i|$ refers to the number of locations in i-th MSS.

As last, we have "Similar Score" for two paths as following:

$$SimScore(path_1, path_2) = \frac{\sum_{j=1}^{m} IDF(MSS_j)}{|path_1| \times |path_2|}, \tag{5}$$

where m is the number of MSS for $path_1$ and $path_2$, $|path_1|$ is the number of distinct locations in $path_1$ [61].

Based on different time intervals in a day we focus on, existing or potential relationship would be found. For example, paths of colleagues or family members tend to have high Similar Score during work time or night respectively. If two paths collected from two different phone numbers have very high Similar Score during weeks, we may assume that these two phone numbers belong to the same person.

4) Predicting User's Future Location: predicting users' future positions allows us to be ready for their movement and react in advance. In order to identify user groups according to their temporal and spatial characteristics, we discretize a day into 24 time segments, each segment lasts one hour long, as shown in Table II.

TABLE II
TIME SEGMENT OF THE CORRESPONDING TIME INTERVAL

Time segment	Time interval
0	0:00 $\sim$ 0:59
1	1:00$\sim$1:59
2	2:00$\sim$2:59
…	…
22	22:00$\sim$22:59
23	23:00$\sim$23:59

We use entropy to measure the activity of users and capture the degree of predictability, which is defined as follow:

$$H(X) = \sum_{i=1}^{n} (p(x_i)I(x_i)) = -\sum_{i=1}^{n} p(x_i)log_b^{p(x_i)}, \tag{6}$$

where n is the number of different locations a user visited in one time segment, i represents the location index the user visited, and b equals to 2. $p(x_i)$ is the probability of a user staying in a certain place in one time segment. The bigger the entropy value is, the more locations the user visits in current time segment. For each user, we build two entropy vectors: the entropy value for each segment in weekdays and in weekends respectively, i.e.,

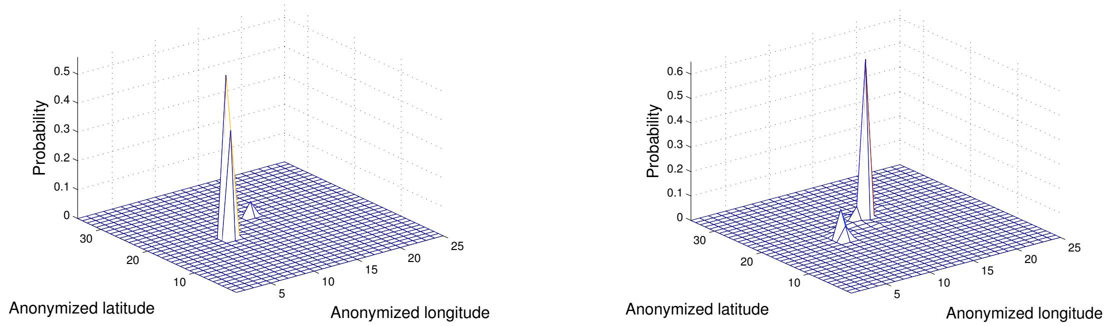
$$E_{weekday} = [e_{weekday}/segment_0, \ldots, e_{weekday}/segment_{23}],$$

$$E_{weekend} = [e_{weekend}/segment_0, \ldots, e_{weekend}/segment_{23}].$$

Considering the correlation between location and time, we group users into two groups (group 1 and group 2) with distinct mobility patterns by clustering them with k-means clustering.

(a) Group 1: users who have regular repeated patterns of movements.

For those who visit very limited locations everyday and follow regular pattern in different days, such as white collars who usually go to work around 8:00 am, and go home around 6:00 pm, as shown in Fig. 7, we apply "Intelligent Time Divisions (ITD)" [63] to predict not only their future movement, but also the time they may arrive. The method ITD takes spatial probability distribution as a significant characteristic to predict user's future mobility pattern. Spatial probability distribution of a user shows the probability that a user at a specific point in the space and is defined as,



(a) Working time (b) Non-working time

Fig. 7. Spatial probability distribution of a user with regular pattern of movements for workdays. The user basically stays in one place (probably his/her working place) in working time, and stays in another place (probably his/her home) in non-woking time.

$$P(X,Y) = prob(x = X, y = Y),$$

where (x, y) represents the location of a user. If we take time factor into consideration, we can define the spatial probability distribution as,

$$P_t(X,Y) = prob(x(t) = X \& y(t) = Y).$$

Actually, if we can get a user's history movements, the spatial probability distribution can be easily estimated.

(b) Group 2: users who move randomly.

For those who spend much time moving around the city everyday, such as postmen and taxi drivers, as shown in Fig. 8, we use a time-based Markov predictor to predict their next location. Although they travel relative randomly in the city, still, some patterns may be discovered due to personal habits, traffic conditions, and road planning in the city. For users in group 2, we predict not only the user's next location, but also the time interval he/she stays in this location.

For a trajectory $Traj = \langle S_1, S_2, \ldots, S_n \rangle$, in order to predict next location L_{n+1}, we find all $S_k = (U, L_k, T_k, D_k)$, $1 < k < n$, satisfying $L_k = L_n$, and $getHours(T_n) - T_{inter} < getHours(T_k) < getHours(T_n) + T_{inter}$. Here the value of T_{inter} depends on the length of the time interval we predict he/she may stay at the next location, and $getHours(T)$ equals to current time (24 hour format), since T is the start time of stay S which is stored as UTC time. As last, the location L_{k+1} that meet the above conditions and appear most frequently is the predicted next location. For example, a user just passes cell tower A at 9:00 am, and we want to know his/her next location, if we have $T_{inter} = 1$, we find all the cell tower A user passed by during 8:00 (a hour before 9:00) and 10:00 (a hour after 9:00) in his/her history trajectories. Then the next cell tower user connected to after cell tower A with highest probability of occurrence in his/her history trajectories is the prediction result.

IV. User Mobility Behavior

In order to test the effectiveness of the framework, we collected two datasets from real mobile Internet to do the experiments. In this part, firstly, we illustrate the basic characteristics of our data sets. Secondly, three human mobility indicators are calculated to show the mobility feature of our data sets. Then, the experimental results of mobility functionalities are introduced.

A. Data Set

We collected 2G/3G/4G data traffic from real mobile Internet, as shown in Table III. The 2G/3G data traffic is extracted as flow records from July 25^{th} 2015 to July 31^{st} 2015 that covers nearly 7 million people of a big northern city in China. The 4G data traffic is the control-plane packets from October 10^{th} 2013 to October 31^{st} 2013 with over 3 thousands people in a big city in southern China.

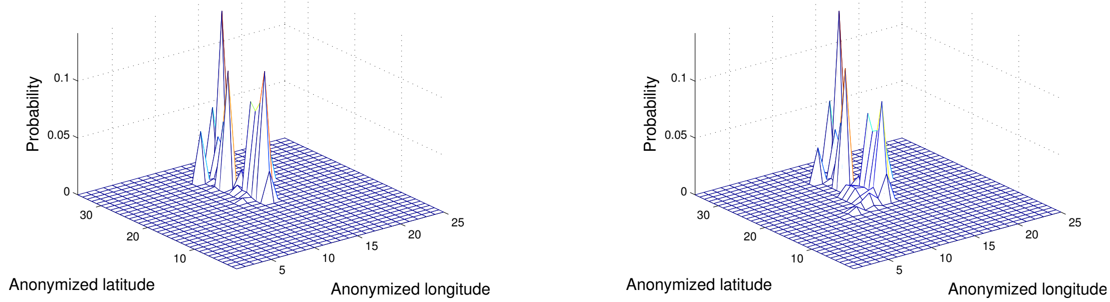
In our experiments, we use 2G/3G data traffic, which covers nearly 7 million people but lasts only 7 days, to study the mobility features of large-scale human, and detect the hotspots with large population in the city. The 4G data traffic, which captures the trajectories of thousands of people in 21 days, is more suitable to discover the mobility patterns, find similar users based on path, and predict user's future movements.

B. Mobility Features

Nowadays, large-scale human mobility is described by three widely accepted indicators: the trip distance distribution $p(r)$, the radius of gyration $r_g(t)$, and the number of visited locations over time $S(t)$ [64]. These three measures contain the basic ingredients to describe the individual trajectories, in which frequent travels occur among a limited number of places, with less frequent trips to new places outside each individual radius.

1) The Trip Distance Distribution $p(r)$: the trip distance distribution $p(r)$ quantifies the relative probability of finding a displacement of length r in a short time. By analyzing the circulation of bank notes in the United States, previous study [4] found the distribution of $p(r)$ decays as a power law, i.e., $p(r) \sim r^{-\beta}$ with $\beta \approx$1.59. In our case, as shown in Fig. 9, $p(r)$ follows power law with $\beta \approx$2.462, which implies that the proportion of large trip distance of bank note trajectories is bigger than our data sources. It is because the data source in [4] covers the nationwide area (United States), yet our data sources covers a tier-two city in China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology



(a) Working time

(b) Non-working time

Fig. 8. Spatial probability distribution of a user who moves randomly in workdays. The user moves among many places during working or non-woking time.

TABLE III
CHARACTERISTICS OF DATA SOURCES

Data sources	Networks	Duration	Number of users	Number of cell towers	Number of flows/packets
Flow records	2G/3G	7/25/2015-7/31/2015 (7 days)	6.90×10^6	85453	2.80×10^{10}
Packets	4G	10/10/2013-31/10/2013 (21 days)	3474	2252	3.60×10^7

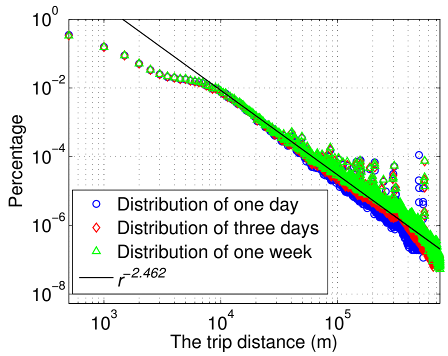


Fig. 9. The trip distance distribution $p(r)$.

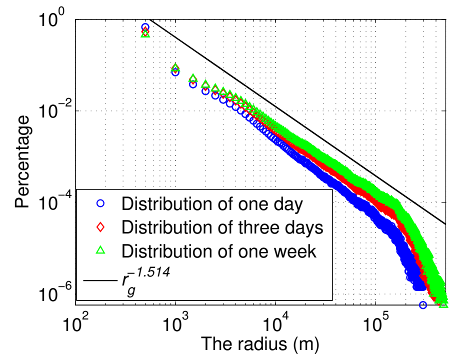


Fig. 10. The $p(r_g)$ distribution of the radius of gyration r_g for users in two days.

2) The Radius of Gyration r_g: the r_g reveals how extensively users move rather than capture the practical distance. Visiting the same sequence of locations in a circle continuously doesn't increase the radius of gyration' value while a straight line movement does [6]. r_g is defined as following,

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\vec{r_l} - \vec{r_{cm}})^2}, \tag{7}$$

where $\vec{r_l}$ is as i^{th} location in user's history trajectories, $i = (1, 2, \ldots, n)$, $\vec{r_{cm}} = \frac{1}{n}\sum_{i=1}^{n}\vec{r_l}$ is the center of a trajectory. As shown in Fig. 10, the distribution of $p(r_g)$ for users in seven days follows power law $p(r_g) \sim r_g^{-\beta}$ with $\beta \approx$1.514. In [35], power law is observed with $\beta \approx$1.55 for CDRs of 3 million users in one year. It implies that the movement range of mobile phone users in [35] is smaller than users in our data traffic collected from 2G/3G networks. A reasonable guess is that the geographical scope that covered by our dataset (about 53,840 square kilometers) is bigger than dataset in [35].

3) The Number of Visited Distinct Locations Over Time $S(t)$: the number of visited distinct locations over time describes how frequently user visits new places, which is expected to follow $S(t) \sim t^{\mu}$. $\mu < 1$ indicates a slow-down at large time cases, which implies a deceasing tendency of the user to visit previously unvisited locations. In our case, $S(t)$ grows as t^{μ} with $\mu = 0.807$.

In summary, above results show that, for the same indicator, different data sources follows similar models with different value of parameters. We can conclude that, human trajectories extracting from cell tower accessing records are able to capture the basic characteristics of human movements.

C. Hotspot Detection

In this part, we will identify the hotspots from the view of population. Intuitively, population at a place is directly proportional to the importance that is attributed to it by the users. Places with large population, such as big shopping mall, residential area, traffic hub, or the places for group

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology

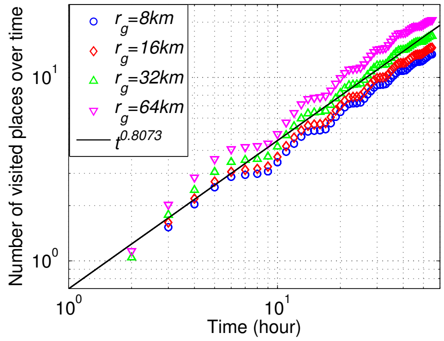


Fig. 11. The number of visited distinct locations $S(t)$ vs. time.

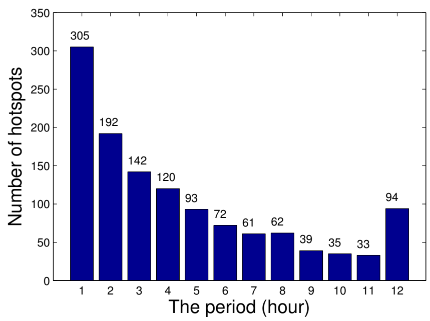


Fig. 12. Duration distribution of hotspots in one day.

TABLE IV
TOP 5 MOBILITY PATTERNS OF GROUPS IN THE CITY

Pattern number	Pattern	Support value
1	⟨transportation hub, residential area⟩	0.049
2	⟨residential area, transportation hub⟩	0.037
3	⟨government building, economic center, university⟩	0.034
4	⟨university, government building, economic center⟩	0.033
5	⟨megamall, mall, food street, residential area⟩	0.032

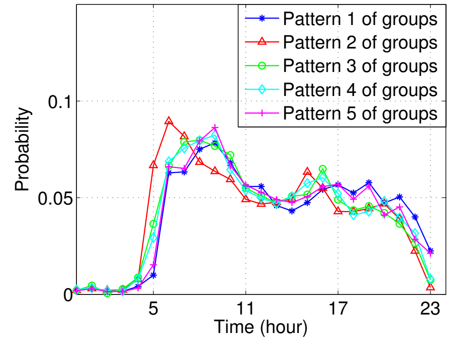


Fig. 13. The probability of the occurrences of patterns for groups varies with time.

TABLE V
TOP 4 MOBILITY PATTERNS OF USER x IN THE CITY

Pattern number	Pattern	Support value
1	⟨residential area, road 1, road 2, century mansion⟩	0.500
2	⟨century mansion, information mansion⟩	0.909
3	⟨information mansion, century mansion⟩	0.909
4	⟨century mansion, road 2, road 1, residential area⟩	0.519

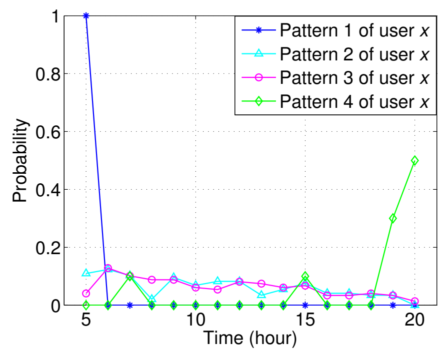


Fig. 14. The probability of the occurrence of patterns for user x varies with time.

activities, have significance for the city. However, hotspots always change with time, which shows the movements of population in different regions in the city, as shown in Fig. 12. We detect hotspots for each hour in one day. Over 24.4% of hotspots appear once in a day, only 7.5% of hotspots last more than 12 hours (like 24-hour eating streets, traffic hubs and universities).

D. The Mobility Pattern of Individuals and Groups

Understanding mobility pattern of groups in the city reveals the population stream among specific locations at a specified time, which has important practical applications to making better urban planning, such as, set new bus or subway routes, and increase cell towers in a more efficient way. Among all the mobility patterns for groups we evaluated, five patterns are most common in the city, as shown in Table IV.

In our experiments, the support threshold is set to 0.01 to make sure the average mobility pattern length achieves the longest. The most frequent patterns are a round-trip between a transportation hub and a residential area. It implies that there is a lot of people living in this residential area, and they usually go to a transportation hub when they need to travel in the city. As shown in Fig. 13, all patterns in Table IV start to occur at 4:00 am. Most of them happen between 6:00 am and 9:00 am in the morning, or between 3:00 pm and 7:00 pm in the afternoon. It indicates that some popular patterns appear during commute time in the city.

In addition, finding individual mobility pattern provides important information about personal habit and interest, which is of practical significance for Service Provider (SP), especially for location based SP. We apply the same algorithm to a sampled user x, and discover his/her daily patterns, as shown in Table V. The support threshold is set to 0.5, which means supported patterns emerge at least 10 days during 21 days.

As shown in Fig. 14, pattern 1 happened around 5:00 am for most of the time, and pattern 4 usually happened between

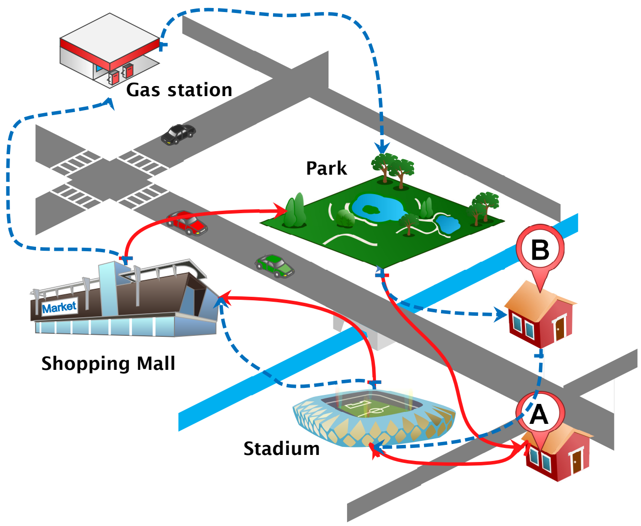


Fig. 15. The semantic path of two sampled users.

7:00 pm and 8:00 pm. Apparently, pattern 1 and pattern 4 are commute route for user x. In addition, century mansion and information mansion (the office building) may be his/her workplaces, and residential area is his/her home.

E. Relationship Among Users

User's history trajectories uncover his/her daily interests and living habits. For example, sport fans tend to visit gym and stadium more often, and fashion girls like going shopping during leisure time. The similarity of historical moving path among users draws potential or existed relationship. The more the users come to places that other users seldom visit, the closer they tend to be. It brings new strategy for SP to find target users, even helps authorities to locate suspicious who may have close relationship with target persons.

By applying method in section III (Finding Similar Users Based on Path), Similar Score value of every two users is calculated. Fig. 15 shows the semantic path of two sampled users that get 0.87 normalized Similar Score during a period of time in the weekend. User A and user B came to a stadium from two different places (probability their home). After a while, they went to the same shopping mall. Then, user B visited a filling station, and appeared in a park where A had already arrived. After about an hour, they returned to their "home" respectively. We can easily conclude that user A and user B may be friends, or have same interests.

F. Mobility Prediction

By applying ITD [63], time-based Markov, and Markov to different groups of users, we have experimental results as shown in Table VI. We use data traffic of 4G networks, which captures users' trajectories in 21 days. We select 2204 users from our data who generate more than 500 packets in 21 days (if user generates too little packets, the movements extracted from packets are not enough to do the experiments). By quantifying the activity of users, we identify 795 users with regular pattern of movements, and 1409 users as randomly moving people. For each trajectory, we use previous $n(n > 0)$ continuous movements to predict the $n+1$ movements. Then, prediction accuracy of each user is the proportion of correct predictions for his/her trajectories. The average value of prediction accuracy of all users is the prediction results.

We predict not only user's next location, but also the time he/she will arrive, and achieve better prediction accuracy than the benchmark (Markov algorithm), as shown in Table VI. ITD beats time-based Markov while predicting the next locations for users with regular pattern of movements. For predicting the future movements of randomly moving users, Time-based Markov achieves the highest prediction accuracy when compared with ITD and Markov. It implies that we should employ different prediction algorithms for distinct groups with different movement patterns.

V. Conclusion

In this paper, we propose a framework to analyze user mobility using big mobile data in densely populated area. The whole framework is based on cloud computing platform, which provides data collection, preprocessing, storage, and analysis function. We further introduce the rules for constructing users' trajectories from different data sources, methods for reducing data noise, algorithms for identifying hotspots, discovering mobility pattern of groups and individuals, finding similar users based on path and predicting user's future movements.

Some interesting findings come out after the experiments. Firstly, comparing with other studies based on trajectories extracted from bank notes, CDRs, and cellular networks, the three indicators (the distribution of trip distance, the radius of gyration, and the number of visited distinct locations over time) calculated by our datasets follow similar models with different value of parameters, which vary with the duration, the covered area and population of data. It indicates that users' trajectories extracted from data traffic of mobile Internet are very suitable for analyzing users' mobility in big city. Secondly, by applying our methods, collected data traffic reveals some interesting phenomenon in the city, such as, more than half hotspots last less than 3 hours, a large crowd moving between transportation hub and residential area during morning and evening peak, users with similar interests could be easily identified from their history trajectories. All the analysis results uncover the common rules exist among huge population in a city, which are of theoretical and practical significance for urban planning, traffic control, mobile network resource optimization, etc. Thirdly, people in the city usually have distinct mobility pattern. Considering mobility pattern while predicting user's future movements could improve the prediction accuracy.

In the future, we will apply the data stream algorithms and get the real time analysis result. In this way, we could make some applications more practical, such as predicting and monitoring large-scale events. In addition, for each mobility application, apply the most suitable algorithm to our dataset and improve existing methods are our future work too.

TABLE VI
THE ACCURACY OF PREDICTION ALGORITHMS FOR APPLYING DIFFERENT ALGORITHMS TO GROUPS WITH DISTINCT MOBILITY PATTERN

Groups	Number of users	Prediction accuracy of Intelligent Time Divisions	Prediction accuracy of time-based Markov	Markov
Users with regular pattern of movements	795	75.2%	22.5%	22.8%
Users move randomly	1409	26.3%	48.6%	46.9%
All users	2204	43.9%	39.2%	38.2%

VI. ACKNOWLEDGMENT

This work is supported in part by the Fundamental Research Funds for the Central Universities (2015RC11) , Director Foundation Project (2015BKL-NSAC-ZJ-01), 111 Project of China (B08004), and EU FP7 IRSES MobileCloud Project (Grant No. 612212).

REFERENCES

[1] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: Universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, May 2012.

[2] F. Giannotti, L. Pappalardo, D. Pedreschi, and D. Wang, "A complexity science perspective on human mobility," in *Mobility Data: Modeling, Management, and Understanding*. New York, NY: Cambridge University Press, Oct. 2012, pp. 297–314.

[3] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Natl. Acad. Sci.*, vol. 106, no. 51, pp. 21 484–21 489, Dec. 2009.

[4] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, Jan. 2006.

[5] A. L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, May 2005.

[6] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.

[7] F. Simini, M. C. González, A. Maritan, and A. L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, Apr. 2012.

[8] C. Song, Z. Qu, N. Blumm, and A. L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[9] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scient. Rep.*, vol. 3, no. 2923, pp. 1–9, Oct. 2013.

[10] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Trans. Networks*, vol. 19, no. 3, pp. 630–643, 2011.

[11] D. Aldous and J. Fill. (2002) Reversible markov chains and random walks on graphs. [Online]. Available: http://www.stat.berkeley.edu/∼aldous/RWG/book.html

[12] B. Liang and Z. J. Haas, "Predictive distance-based mobility management for pcs networks," in *Proc. INFOCOM'99. 18th Ann. Joint Conf. IEEE Comput. Commun. Soc.*, New York, NY, Mar. 1999, pp. 1377–1384.

[13] C. Bettstetter, "Smooth is better than sharp: A random mobility model for simulation of wireless networks," in *Proc. 4th ACM Int. Workshop Model.* New York, NY: ACM, Jul. 2001, pp. 19–27.

[14] C.-C. Chiang, "Wireless network multicasting," Master's thesis, Citeseer, 1998.

[15] V. A. Davies, "Evaluating mobility models within an ad hoc network," Master's thesis, Citeseer, 2000.

[16] A. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri, "Towards realistic mobility models for mobile ad hoc networks," in *Proc. 9th Ann. Int. Conf. Mobile Comput. Networking*. New York, NY: ACM, Sep. 2003, pp. 217–229.

[17] S. Batabyal and P. Bhaumik, "Mobility models, traces and impact of mobility on opportunistic routing algorithms: A survey," *Commun. Sur. & Tutor., IEEE*, vol. 17, no. 3, pp. 1679–1707, Apr. 2015.

[18] H. Nishiyama, M. Ito, and N. Kato, "Relay-by-smartphone: realizing multihop device-to-device communications," *Communications Magazine, IEEE*, vol. 52, no. 4, pp. 56–65, 2014.

[19] Y.-B. Lin, Y.-W. Lin, C.-Y. Chih, T.-Y. Li, C.-C. Tai, Y.-C. Wang, F. J. Lin, H.-C. Kuo, C.-C. Huang, and S.-C. Hsu, "Easyconnect: A management system for iot devices and its applications for interactive design and art," *Internet of Things Journal, IEEE*, vol. 2, no. 6, pp. 551–561, 2015.

[20] H. Cai and D. Y. Eun, "Crossing over the bounded domain: From exponential to power-law intermeeting time in mobile ad hoc networks," *IEEE/ACM Trans. Networks*, vol. 17, no. 5, pp. 1578–1591, Oct. 2009.

[21] J. Liu, S. Zhang, N. Kato, H. Ujikawa, and K. Suzuki, "Device-to-device communications for enhancing quality of experience in software defined multi-tier lte-a networks," *Network, IEEE*, vol. 29, no. 4, pp. 46–52, 2015.

[22] K. Zheng, L. Hou, H. Meng, Q. Zheng, N. Lu, and L. Lei, "Soft-defined heterogeneous vehicular network: Architecture and challenges," *arXiv preprint arXiv:1510.06579*, 2015.

[23] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5g," *Network, IEEE*, vol. 30, no. 1, pp. 44–51, 2016.

[24] J. Yeo, D. Kotz, and T. Henderson, "Crawdad: A community resource for archiving wireless data at dartmouth," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 2, pp. 21–22, Apr. 2006.

[25] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *ACM Commun.*, vol. 56, no. 1, pp. 74–82, Jan. 2013.

[26] Z. Smoreda, A. M. Olteanu Raimond, and T. Couronné, "Spatiotemporal data from mobile phones for personal mobility assessment," in *Transport survey methods: best practice for decision making*. Bingley, UK.: Emerald Group Publishing, Nov. 2013, pp. 745–766.

[27] B. C. Csáji, A. Browet, V. A. Traag, J. C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel, "Exploring the mobility of mobile phone users," *Phys. A: Stat. Mech. Appl.*, vol. 392, no. 6, pp. 1459–1473, Mar. 2013.

[28] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proc. 13th ACM SIGKDD Int. Conf.*, San Jose, California, Aug. 2007, pp. 330–339.

[29] M. Lin and W. Hsu, "Mining GPS data for mobility patterns: A survey," *Perv. Mobile Comput.*, vol. 12, pp. 1–16, Jun. 2014.

[30] Y. Zhang, "User mobility from the view of cellular data networks," in *Proc. INFOCOM, IEEE*, Toronto, Canada, Apr. 2014, pp. 1348–1356.

[31] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing User Behavior in Mobile Internet," *IEEE Trans. Emer. Top. Comput.*, vol. 3, no. 1, pp. 95–106, Mar. 2015.

[32] Y.-B. Lin and P.-K. Huang, "Prefetching for mobile web album," *Wireless Communications and Mobile Computing*, vol. 16, no. 1, pp. 18–28, 2016.

[33] A. Thiagarajan, L. Ravindranath, and H. Balakrishnan, "Accurate, low-energy trajectory mapping for mobile devices." in *Proc. NSDI*, Boston, MA, Mar. 2011, pp. 267–280.

[34] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Human mobility patterns in cellular networks," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1877–1880, Oct. 2013.

[35] C. Song, T. Koren, P. Wang, and A. L. Barabási, "Modelling the scaling properties of human mobility," *Nature Phys.*, vol. 6, no. 10, pp. 818–823, Sep. 2010.

[36] T. Louail, M. Lenormand, O. G. C. Ros, M. Picornell, R. Herranz, E. Frias Martinez, J. J. Ramasco, and M. Barthelemy, "From mobile phone data to the spatial structure of cities," *Scient. Rep.*, vol. 4, no. 5276, pp. 1–12, Jun. 2014.

[37] H. Klessig, V. Suryaprakash, O. Blume, A. Fehske, and G. Fettweis, "A

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2553182, IEEE Transactions on Vehicular Technology

framework enabling spatial analysis of mobile traffic hot spots," *IEEE*, vol. 3, no. 5, pp. 537–540, Aug. 2014.

[38] M. A. Bayir, M. Demirbas, and N. Eagle, "Mobility profiler: A framework for discovering mobility profiles of cell phone users," *Perv. Mobile Comput.*, vol. 6, no. 4, pp. 435–454, Aug. 2010.

[39] J. Yang, X. Zhang, Y. Qiao, Z. Md.Fadlullah, and N. Kato, "Global and individual mobility pattern discovery based on hotspots," in *Proc. ICC*, London, UK., Jun. 2015, pp. 1–6.

[40] X. Yan, X. Han, B. Wang, and T. Zhou, "Diversity of individual mobility patterns and emergence of aggregated scaling laws," *Scient. Rep.*, vol. 3, no. 2678, pp. 1–5, Sep. 2013.

[41] Telecomitalia. (2015) Big data challenge 2015. [Online]. Available: https://www.telecomitalia.com/tit/en/bigdatachallenge.html

[42] T. Cisco, "Global mobile data traffic forecast update, 2014–2019 white paper," *Cisco Public Information*, Feb. 2015.

[43] L. Vu, Q. Do, and K. Nahrstedt, "Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace," in *Proc. PerCom, IEEE Int. Conf.*, Seattle, Mar. 2011, pp. 54–62.

[44] H. Cao, W. Dong, L. Liu, C. Ma, W. Qian, J. Shi, C. Tian, Y. Wang, D. Konopnicki, M. Shmueli Scheuer, D. Cohen, N. Modani, H. Lamba, Dwivedi, N. A, K. AA, and M, "Solomo analytics for telco big data monetization," *IBM J. Res. Dev.*, vol. 58, no. 5/6, pp. 1–9, Nov. 2014.

[45] Y. Zheng, X. Xie, and W. Ma, "Geolife: A collaborative social networking service among user, location and trajectory." *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, Jun. 2010.

[46] Y. Zheng and X. Zhou, *Computing with spatial trajectories*. New York, NY: Springer Science & Business Media, 2011.

[47] J. Liu, F. Liu, and N. Ansari, "Monitoring and analyzing big traffic data of a large-scale cellular network with hadoop," *IEEE Networks*, vol. 28, no. 4, pp. 32–39, Jul. 2014.

[48] T. White, *Hadoop: The definitive guide*. Sebastopol, CA 95472: O'Reilly Media, Inc., 2012.

[49] Apache flume. [Online]. Available: https://flume.apache.org/

[50] Apache kafka. [Online]. Available: http://kafka.apache.org/

[51] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Proc. IEEE*, Washington D.C., May 2010, pp. 1–10.

[52] L. George, *HBase: the definitive guide*. Sebastopol, CA 95472: O'Reilly Media, Inc., 2011.

[53] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *ACM Commun.*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[54] Apache spark. [Online]. Available: https://spark.apache.org/

[55] Apache storm. [Online]. Available: https://storm.apache.org/

[56] P. Hunt, M. Konar, F. P. Junqueira, and B. Reed, "Zookeeper: Wait-free coordination for internet-scale systems." in *Proc. USENIX Annual Technol. Conf.*, Berkeley, CA, Jun. 2010, pp. 11–11.

[57] S. A. Shad, E. Chen, and T. Bao, "Cell oscillation resolution in mobility profile building," *ArXiv Prepr.*, vol. 9, no. 3, pp. 205–213, Jun. 2012.

[58] W. Wu, Y. Wang, J. B. Gomes, D. T. Anh, S. Antonatos, M. Xue, and P. Yang, "Oscillation resolution for mobile phone cellular tower data to enable mobility modelling," in *Proc. MDM*, Brisbane, Jul. 2014, pp. 321–328.

[59] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "Nextplace: A spatio-temporal prediction framework for pervasive systems," in *Perv. Comput.* San Francisco: Springer, Jun. 2011, pp. 152–169.

[60] T. Pangning, M. Steinbach, and V. Kumar, *Introduction to data mining*. Boston: Pearson Addison Wesley Boston, 2006.

[61] Y. Zheng, "Location-based social networks: Users," in *Computing with Spatial Trajectories*. New York: Springer, Oct. 2011, pp. 243–276.

[62] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972.

[63] S. Gatmir Motahari, H. Zang, and P. Reuther, "Time-clustering-based place prediction for wireless subscribers," *IEEE/ACM Trans. Networks*, vol. 21, no. 5, pp. 1436–1446, Dec. 2013.

[64] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *J. R. Soc. Interface*, vol. 10, no. 84, p. xxxviii, Aug. 2013.

Yuanyuan Qiao is currently a lecturer in the School of Information and Communication Engineering, BUPT. She received her B.E. degree from Xidian University in 2009, and received her Ph.D degree from BUPT in 2014. Her research focuses on traffic measurement and classification, mobile Internet traffic analysis and big data analytics.

Yihang Cheng is currently a graduate student in the School of Information and Communication Engineering, BUPT, and received his B.E. degree in communication engineering from BUPT in 2014. He is engaged in the research of users' relationships discovery and mobility prediction.

Jie Yang received her B.E., M.E., and Ph.D. degrees from Beijing University of Posts and Telecommunications, China in 1993, 1999, and 2007 respectively. She is now a professor and deputy dean of School of Information and Communication Engineering, BUPT. Her current research interests include broadband network traffic monitoring, user behavior analysis, big data analysis in Internet and Telecom, etc. She has published several papers on international magazines and conferences including IEEE JSAC, IEEE Trans. on Wireless Communications and IEEE Trans. on Parallel and Distributed Systems. Also, she is the Vice Program Committee Co-Chairs of IEEE IC-NIDC 2014 and 2012.

Jiajia Liu is currently a full professor at the School of Computer Science and Technology at Xidian University. He received the Yasujiro Niwa Outstanding Paper Award in 2012, and the Best Paper Awards of IEEE WCNC in 2012 and 2014. He was also a recipient of the Chinese Government Award for Outstanding Ph.D. Students Abroad in 2011, the Tohoku University RIEC Student Award, and the Tohoku University Professor Genkuro Fujino Award in 2012, as well as the prestigious Dean Award and President Award of Tohoku University in 2013. His research interests include wireless and mobile ad hoc networks, performance modeling and evaluation, stochastic network optimization, LTE-A and 5G networks.

Nei Kato is a professor with the Graduate School of Information Sciences, Tohoku University. He currently serves as the Chair of the IEEE Ad Hoc & Sensor Networks Technical Committee, Chair of the IEEE ComSoc Sendai Chapter, Associate Editor-in-Chief of IEEE Transactions on Parallel and Distributed Systems and the IEEE Internet of Things Journal, an Area Editor of IEEE Transactions on Vehicular Technology, and an Editor of IEEE Wireless Communications and IEEE Network. He is Strategic Adviser to the President of Tohoku University. He is a fellow of IEICE.