# UrbanFACET: A Multifaceted City Panorama By Large-Scale Human Mobility Analysis

Lei Shi, Tao Jiang, Ye Zhao, Zhichun Guo, Xiatian Zhang, and Yao Lu

**Abstract**—Cities are living systems where urban infrastructures and their functions are defined and evolved due to population behaviors. Visualizing, profiling and comparing the human mobility in such cities has been an important topic in urban design and planning. This paper studies a unique big mobile data set which includes the longitudinal movement pattern of tens of millions of city residents. The data set, acquired from mobile users' agnostic check-ins at hundreds of thousands of mobile apps, is well utilized in a visualization study on long-term human mobility, integrative with urban structure (e.g., administrative division) and POI (Point of Interest) distributions. In particular, we develop a set of novel information-theoretic metrics. These metrics include Fluidity, vibrAncy, Commutation, divErsity, and densiTy (FACET), which characterize the mobility of city residents and functional regions. Finally, we develop a visual analytics system, namely UrbanFACET, that assembles a multifaceted city panorama to visually analyze these mobility metrics, profile and compare urban functions over different cities and time periods. The system is evaluated through case studies in real-world big cities, which brings myriad value to potential domain users such as urban planners, business managers and security officers in their key responsibilities.

**Index Terms**—Urban data visualization, human mobility analysis, information theory.

## 1 Introduction

This paper studies the visualization of metropolitan-scale human mobility with a new kind of mobile sensing data. The data set, called the big mobile data, is acquired by synthesizing urban mobile user's agnostic check-ins from hundreds of thousands of mobile applications (see details in Section 3.1). In comparison to previous large-scale urban data sets collected in special-purpose applications (e.g., taxi [1] [2]) or from cell phone call details (CDR) [3] [4], the big mobile data possesses several positive properties for the analysis of longitudinal human mobility. Spatially, it covers the whole city area, in contrast to the taxi data confined to transportation lines. Temporally, it is a more uniform sampling of human mobility over time, compared with the sparse event of voice calls [5]. Population-wise, it has a much larger penetration rate ($> 50\%$) and is more representative than the geo-tagged social media by crowdsourcing ($\sim 5\%$). With this big mobile data locating millions of city residents in the long term, we consider the comprehensive visualization of their individual-level mobility that can unveil the high-level characteristics of residents (e.g., their life abundance, commutation pattern), and city regions (e.g., population diversity, degree of fluidity).

With this study, we are able to accomplish many critical urban analysis tasks that are impossible, or at least inaccurate, previously. First, how does the high-level resident/region characteristics from the human mobility distribute in a major city (*the overview task*)? For example, whether the life abundance follows a similar distribution to the population density, i.e., degrading from the city center to the remote areas radially. How does the terrain type impact this distribution? Second, whether the land use of a city (i.e., the functionality) can be inferred by the mobility of people living in that area and how does this profiling correlate with the administrative city divisions (*the profiling task*)? Third, how does the human mobility vary across city and time (*the comparison task*)? For example, is the human mobility pattern in a small city different from that of a big city? How does the hotspot of human mobility evolve over time, e.g., from morning to evening in a single day? Answering these questions add myriad value to: 1) urban planners and public security officers, to improve their situation and risk awareness on the target city; and 2) a wide spectrum of business people, to determine their enterprise or real estate site selection, and to find the optimal place for commercial billboards.

In this work, we propose a visual analytics system called UrbanFACET to accomplish these promising tasks. Designing and implementing UrbanFACET raises many research questions as well as technical challenges. First, though the short-term population-level human mobility visualization has been widely available, e.g., Origin-Destination (OD) flows [6] [7], trips [8] [2], and trajectories [9] [10], the theoretical foundation for visualizing individual's long-term mobility has not yet been built up. There were several researches on the behavior of human mobility, in particular the two-state model of scale-free jump and wait [11] [12] [13] [14], but how to synthesize these effective models into mobility metrics that can be visualized in a single map (i.e., the city panorama) remains a daunting task. The challenge lies in the balance between the abstraction of mobility behavior and the interpretability of visualization results. Second, given appropriate mobility metrics for visualization, efficiently computing them in the big data scenario composed of billions of mobility records introduces additional challenges. Compromise should be made among abstraction granularity, visualization fidelity and the user interactivity. Third, designing the visualization interface to accommodate the spatial, temporal and multifaceted human


- *Lei Shi and Tao Jiang are with SKLCS, Institute of Software, Chinese Academy of Sciences. Email: {shil,jiangt}@ios.ac.cn.*
- *Ye Zhao is with the Department of Computer Science, Kent State University. Email: zhao@cs.kent.edu.*
- *Zhichun Guo is with the School Of Computer Science, Fudan University. Email: shirley.gzc@gmail.com.*
- *Xiatian Zhang and Yao Lu are with the Beijing Tendcloud Tianxia Technology Co., Ltd. Email: {xiatian.zhang,yao.lu}@tendcloud.com.*

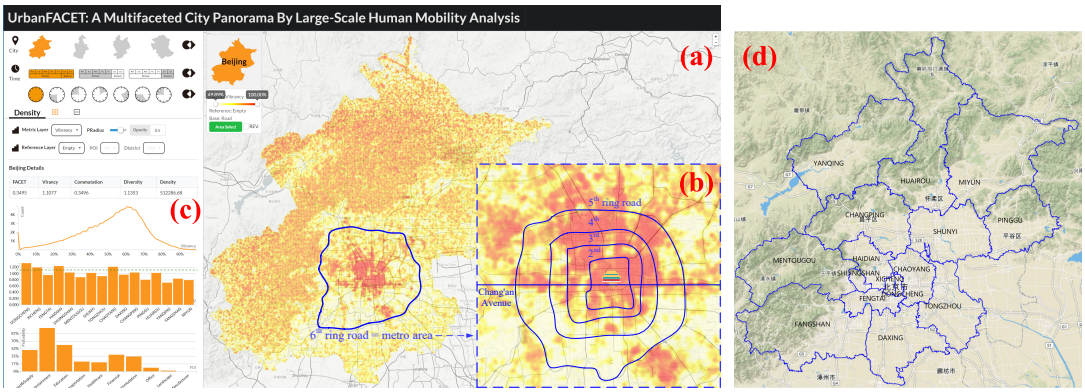



Fig. 1: UrbanFACET visualization of the long-term user mobility in Beijing: (a) the full scale distribution of the vibrancy metric; (b) the metro area is enlarged, i.e., inside the 6th ring road annotated in (a); (c) the skewed value distribution on vibrancy; (d) the terrain map of Beijing, overlaid with district boundaries. Two classes of high vibrancy regions are discovered: an asymmetric region within the metro of Beijing, and the western, northern and northeastern mountain regions where travelers will visit.

mobility metrics into a single view is nontrivial and involves appropriate interaction models to explore these metrics, in order to fulfill the targeted urban analysis tasks.

The contribution of this work can be summarized as below.

- We invent a class of information-theoretic metrics to capture the individual's long-term mobility within the metropolitan scale. These metrics are based on the fundamental concept of Shannon entropy [15] over the distribution of individual's location records in spatial dimensions enriched by urban references. Moreover, a decomposition of the original entropy metric leads to a new definition of record entropy that can represent the mobility of city regions (Section 4.2).
- We propose a grid-based data binning technique to efficiently compute the entropy-based mobility metrics over the big mobile data up to billions of records (Section 4.3). These mobility metrics are further analyzed by a three-step hierarchical clustering algorithm to visually profile the land use of city regions through a flower-shape metaphor design (Section 4.4).
- We design and implement the UrbanFACET system which applies a kernel density estimation (KDE) based method to display a multifaceted city panorama on human mobilities. Versatile interactions are integrated which allow users to select, filter, and compare the mobility patterns in any regions across multiple big cities and in different time periods (Section 5). The proposed visual analytics system and mobility metrics are evaluated through case studies on several real-world big cities (Section 6).

## 2 Related Work

We summarize related works according to the pipeline of UrbanFACET, from urban data processing to human mobility metrics, and finally the visualization.

### 2.1 The Analysis of Urban Data

In the field of urban computing, big and heterogeneous urban data have been processed and analyzed to tackle the major issues that modern cities face [16]. For example, the real-time vehicle movement data collected in the city can be used to predict traffic flows [17] [18], estimate travel times [19] [20], and finally recommend the fastest route for drivers [1] [10]. Meanwhile, the mobile device localization data can be used for urban planning [21], tourism analysis [22], road usage understanding [23], and human mobility analysis [24] [4] [25] [26] [13] [11] (see details in the next part). In addition, passenger's smart card data can be analyzed to understand the efficiency of public transportation system [27] and recommend the best ticket purchasing for travellers [28].

In the survey by Zheng et al. [16], the urban data acquisition methods can be categorized into three classes. The first class refers to the traditional sensing and measurement, e.g., the commutation census [29] and road-side traffic sensors, such as inductive loops [17], laser scanners [30], and Bluetooth device detectors [31]. The second class refers to passive crowd sensing which includes short-range indoor localization, GPS (on taxis [8] [9] [32] [10] [33] [1] [2], vehicles [18] [34], travelers [22] [20]), smart cards (for public transportation [28] [35] [27] and docked bicycles [6]), and mobile cell phone sensing (e.g., Call Detail Records (CDR) [3] [21] [23] [24] [36] [37] [4]). The third class refers to participatory sensing, in which people actively contribute their own information. For example, the geo-tagged social media, such as messages on Twitter [38] and pictures on Flickr [39] [40], can be collected to reveal the spatiotemporal pattern of a particular group of people in the city.

Compared with the existing classes of urban data, the big mobile data set studied in this work is more suitable for the long-term human mobility study due to its unique data characteristics (see Section 3 for details).

### 2.2 The Study of Human Mobility

Most urban data studied above captures the movement of people in the city, so it is straightforward to analyze human mobility patterns using these data sets.

In the city level, all the location records can be aggregated to reveal the spatiotemporal distribution of urban activity [21] [38]. The Mobile Landscapes project mapped the cell phone usage of a city to display its spatiotemporal dynamics [21]. When the context of urban activities is available, thematic city patterns in space and time can also be examined [38]. After decomposing the urban data into trips, the city-level mobility can be modeled by the OD flow matrix among regions where the region is defined by municipal districts [29] [34], cell towers [23], or major roads [33].

To capture the mobility of individuals, the OD pair of each single trip is processed. For example, Liu et al. analyzed the OD pairs of taxi cabs to study the intra-urban trip patterns [8]. Song et al. constructed the mobility network of each mobile user to

understand the predictability of human mobility [4]. When the temporal information of OD trips is considered [2], the concept of temporal OD flow can be aggregated from individual OD trips, which becomes a popular design for urban data visualization [6] [7] [39] [37].

In case each single trip is measured continuously (e.g., by the floating car technique [32]), individual routes or trajectories are available for the detailed human mobility study [9] [10] [1]. For instance, the trajectories of taxis can be used to classify drivers by their job performance [9]. These trajectories can be aggregated to construct a time-dependent landmark graph [1] or a trajectory visualization [10], in order to compute the fastest route for drivers.

The statistics of single trips can be analyzed holistically to reveal the fundamental pattern of human travel. Based on over one million bank note circulation reports in US, Brockmann et al. explained human travel as the combination of a scale-free jump and a heavy-tailed wait [11]. They proposed a random-walk model to characterize these findings. On a more accurate mobility data set, Rhee et al. validated the scale-free nature of human mobility and proposed a similar truncated Lévy-walk model [12]. In the recent studies by the group of Barabási, they observed a high degree of temporal and spatial regularity in human mobility because of the tendency to avoid visiting new places and return to previously visited locations (e.g., home and workplace) [13] [14]. In this sense, the human mobility is highly predictable. Using a large-scale mobile phone trace, Song et al. derived a 93% predictability on mobile user's mobility [4]. In a similar study, de Montjoye demonstrated that four spatio-temporal mobile records were enough to uniquely identify 95% of mobile users [24].

In summary, a large portion of existing works on human mobility focus on statistical and predictive models, which are not suitable for visualization because of the lack of semantic urban information. On the other hand, there are several mobility representations designed for visualization, e.g., the temporal OD flows. However, most designs stay at short-term aggregated levels, few considers the mobility metric at the individual level. In comparison, UrbanFACET aims at visually analyzing the individual-level human mobility in the long term and we propose a suite of entropy-based metrics to serve this need. In fact, the Shannon entropy has been adapted in [41] [4] [25] [26] to capture the predictability of human mobility. For the visualization purpose, we have extended these entropy metrics with the spatial and temporal semantics associated with each mobility record. This semantic information in return helps to visually profile the societal and functional aspects of modern cities and their regions.

### 2.3 The Visualization of Movement

A lot of studies have been conducted to visualize the OD relationships and OD flows from massive vehicle data [42] [43]. Flowstrates allowed the users to visually query OD data by regions of interest and to analyze the temporal changes through heatmap based flow ordering, filtering and aggregation [44]. Ferreira et al. modelled a wide range of spatiotemporal queries and applied different data aggregations and visual representations for users to explore OD taxi trajectories and compare query results [2]. Zeng et al. visualized isochrone/isotime map, and OD-pair journeys for the study of mobility factors from the public transit data [27]. Ma et al. visualized the human flow across a city with flow volumes, links, and communities over a large cell phone tracking data set. [37].

Meanwhile, visualizations have been developed to analyze the human mobility. Bicycle hire patterns are discovered by flow maps and OD maps [6]. They were also studied by aggregating individual OD data according to trip directions and distance [7]. Bluetooth based OD data is used to find the dynamics of vehicles and characterize urban networks [31]. Smart card data recording human behaviors is visualized for analyzing subway routes and reachable regions [27]. Social media data with geo-tags can be used to find places and events related to urban movement [38] [39]. Urban Pulse captured the spatiotemporal activity in a city from multiple urban data sets [40].

Many other approaches focus on the entire trajectories and visualize the hidden traffic patterns using various metaphors and interactions. These approaches can be categorized into several types. First, individual trajectories and microscopic traffic patterns are interactively visualized. TripVista was designed for users to explore vehicle trajectories in three different views from the spatial, temporal and multidimensional perspectives [30]. Second, the large and complex trajectory data is clustered, analyzed, and then visualized for the study of mobility patterns. Kohonen map applied the Self-Organizing Map (SOM) based analysis on trajectory data, which combines the automatic data clustering with human interaction [45]. This approach can visually manipulate the trajectory clustering process at different levels. Third, time-varying traffic patterns in the city are visually explored by studying the relationship of massive trajectories and urban regions. The route diversity patterns were visualized by displaying the high-dimensional attributes and statistics of different routes [10].

In summary, the visualization of long-term human mobility in multiple facets and cities is rarely studied in the literature. There is no existing approach to visually discover the long-lasting schemes of intra- and inter-city mobility patterns over different regions and cities in the metropolitan scale. UrbanFACET is designed to characterize such long-term mobility patterns by visualizing a set of newly designed metrics on an extremely large mobile data set that has not been utilized in previous visual analytics systems.

## 3 Urban Data and Mobility Analysis Task

### 3.1 Data Source

Our urban data is provided by TalkingData, a mobile analytics company who keeps the real-time tracking of billions of smart devices in China, including mobile phones, tablets, wearable devices, etc. This data is collected by registering third-party APIs inside the mobile apps of each device. When a registered mobile app is activated (not necessarily being used), the API will report location records to the company server. The metadata of location records is shown in Table 1. There are four fields in each record: the time of recording, location information (longitude and latitude), unique ID of the smart device, and the localization method (i.e. source, including GPS, Wi-Fi, base station and Internet IP). Here is a sample record:"20:10/07/03/2016, 116.3336266,39.890955,1470076020481,GPS".

In this work, we conduct analysis on four data sets extracted from the company's data repository. As listed in Table 2, each data set corresponds to a 90-day collection of one Chinese city, including Beijing (capital of China), Tianjin (one of five national central cities in China), Tangshan and Zhangjiakou (two major cities in Hebei province surrounding Beijing and Tianjin). These four cities form the so-called national capital region of China. To obtain more accurate results, we only keep the location records collected by GPS and Wi-Fi, from which sources the spatial localization error is kept below 100 meters.

TABLE 1: The metadata of location records in our data source.

| Field | Description | Sample |
|---|---|---|
| Time | Timestamp of the record | 20:10/07/03/2016 |
| Lon. | Longitude of location | 116.3336266 |
| Lat. | Latitude of location | 39.890955 |
| Mid | Unique ID of the device | 1470076020481 |
| Src | Source of the location record | GPS |

TABLE 2: Statistics on four data sets used in this research.

| City | #Device | #Record | Size | Time |
|---|---|---|---|---|
| Beijing | 31849742 | 8407648917 | 738.1G | 90 days |
| Tianjin | 8011128 | 2858575880 | 206.8G | 90 days |
| Tangshan | 2786668 | 920364499 | 64.8G | 90 days |
| Zhangjiakou | 1392236 | 317252149 | 23.1G | 90 days |

Compared with previously published urban data sources [46] [40], our data set is unique in three aspects. First, spatially it has a larger coverage and closer distribution with respect to the actual population distribution in a city (Figure 14). In the largest city studied (i.e., Beijing), the data set measures the everyday mobility of 31.8 million devices (∼50% penetration rate), with 93.4 million records per day. In comparison, the GPS localization of taxis, the road-side traffic sensors, and the smart cards on public transportation only report the human mobility along roads and transportation lines. Second, temporally our data set represents a more uniform sampling of the human mobility compared with the CDR data in which the position information is recorded only when a call is made or a short message is sent. This result has been analyzed and reported in [41]. Third, categorically our data set assembles a more comprehensive sampling of different types of people in a city. The location records are collected by the third-party APIs installed with more than 120,000 kinds of mobile apps, covering a wide spectrum of application domains, such as entertainment, education, information, etc. On the other hand, the geo-reference social media only covers the group of people who would like to share their locations. These data characteristics are studied further in the next part.

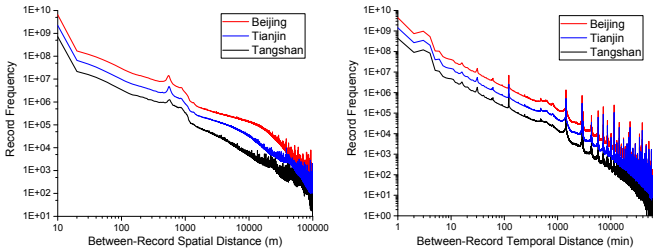
## 3.2 Feasibility for Long-Term Human Mobility Analysis

### 3.2.1 Conformity to human mobility model

The seminal works on the scaling law of human mobility [11] [12] [13] discovered that the human travel and mobility pattern can be captured by the combination of a Lévy flight like jump and a waiting period. Both the jump distance and the waiting time length follow power-law distributions, i.e., $P(r) \sim r^{-\beta}$ for the distance and $P(t) \sim t^{-\beta}$ for the waiting time.

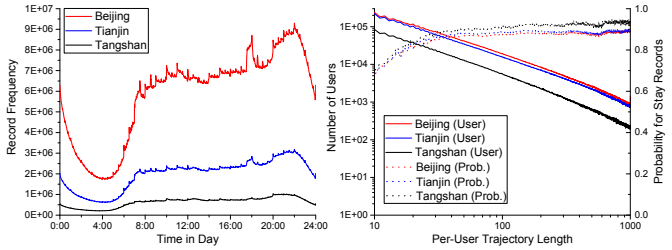
We validate whether this scaling law on human mobility still holds in our data sets. After splitting the data set into the trajectory of each user, we plot the distribution of between-record spatial distance and temporal distance in Figure 2(a)(b). In both space and time, the overall distribution in three largest cities (Beijing, Tianjin and Tangshan) can be fit into power-law distributions. The scaling exponent in spatial distance is $1.8 \pm 0.05$, close to the exponent of $1.75 \pm 0.15$ on mobile CDR data [13] and $1.59 \pm 0.02$ on bank note displacement data [11]. The scaling exponent in temporal distance is $1.63 \pm 0.01$, very close to the waiting time exponent of $1.6 \pm 0.03$ on bank note displacement data.

The conformity to the scaling law indicates that the mobile application based sampling does not distort the distributional pattern on individual's spatiotemporal mobility, which makes it feasible to conduct our study.




(a) Spatial distance (b) Temporal distance

Fig. 2: The scaling law of human mobility still holds in our data sets: (a) between-record spatial distance distribution; (b) between-record temporal distance distribution.




(a) Temporal uniformity (b) Stay probability

Fig. 3: Sampling uniformity: (a) the distribution of location records within a day, relatively stable except the wee hours; (b) the probability to sample stay records (right Y) and the number of users (left Y), with respect to the user's trajectory length (X).

### 3.2.2 Uniformity in space, time, app and mobility state

On the sampling of individual's mobility, another question is whether the sampling is uniform in long-term over space, time and other dimensions. The negative examples include the taxicab's trajectory, which only covers records on transportation lines [8] [10] [2]; the CDR data, which can be biased in the overall spatiotemporal pattern of individual's mobility due to the sparsity of phone call behaviors [5], compared with more fine-grained sampling by data network services [41]; and the driving assistance apps, which mostly report location records during travel.

In our data set, the spatial uniformity is straightforward to check. As shown in Figure 14, the record density distribution in all cities reveal a radial diffusion pattern from the city center, conforming well to the prior knowledge of population distributions. Temporally, the record time distribution within a day is given in Figure 3(a), which shows a relative stable sampling rate except the wee hours. In Figure 3(b), we compute the probability to be in the mobility state of stay, in contrast to the travel state. The mobility classification method follows the work of Ref. [3] [25]. First, the trajectory of each user is split into multiple trips by a maximal spatial distance threshold $\Delta S$. Second, the records of each trip are classified as stay points if the duration time of the trip is larger than a minimal time threshold $\Delta T$. Here we use $\Delta S = 800\ meter$, $\Delta T = 15\ min$. Figure 3(b) indicates that as the user's trajectory length increases beyond 10 (X axis), the probability to sample stay points also increases from around 0.7 to 0.9 (right Y axis). The fact of having a large share of stay points allows us to compute the distribution of user's stays in Section 4. Finally, the mobile app. share in our data set shows a fat-tailed distribution (Figure 4), which ensures a uniform sampling among population groups.

In summary, the big mobile data applied here represents a uniform sample in space, time, mobility state and application domains, thus is feasible for the long-term human mobility study.

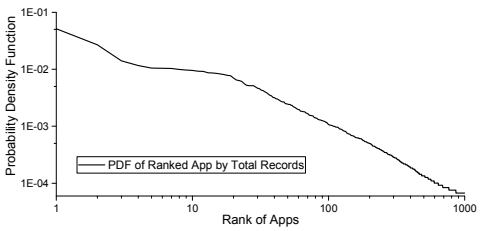



Fig. 4: The probability density function of top 1000 mobile apps measured in TalkingData, ordered by the number of total records.

### 3.3 Task Description

Over the big mobile data that carries individual's long-term mobility in the city, we aim to construct a city panorama for users to visually analyze these mobility information and complete three classes of urban analysis tasks. These tasks can be beneficial in several key urban applications such as city planning, public security, and business site configuration.

- **Overview** analysis of the distribution of all the individual's long-term mobility in the metropolitan scale, and the correlation analysis between the mobility distribution and city landmarks, POIs and terrain types. This can help to understand the distribution of high-level characteristics of corresponding city residents and regions in a single city panorama (e.g., life abundance, region fluidity, etc.). City planners and security officers will gain situation and risk awareness by both the static, dynamic, and categorical mobility distributions in completing these tasks.
- **Profiling** city lands by clustering adjacent regions through their mobility metrics. This mobility-based land profiles can be correlated with administrative divisions, city landmarks, and POIs to infer the fine-grained land use of cities. Moreover, the up-to-date land profile can be compared with the city planning blueprint to characterize the trend of land use and potential misuses. Both government and business people can identify risks and opportunities for their key decisions.
- **Comparison** analysis among cities and across time (e.g., morning and night) on the mobility panorama to reveal the city-level distinction and the potential temporal pattern on the overall mobility. The mobility panorama can also be compared with regional city demographics (e.g., GDP, house price, etc.) to infer the hidden linkage and contributing factors of the mobility metrics. Governments can optimize their city plannings, including facility and transportation management, based on the comparison result with advanced cities.

## 4 Large-Scale Mobility Analysis

Figure 5 summarizes the mobility analysis pipeline in UrbanFACET. The pipeline takes the raw data described in Section 3.1 as input, which sums up to more than 7,300 billion records (Figure 5(a)). In the first step, the big data is partitioned according to the record location into cities by their administrative boundary (Figure 5(b)). Each city's data set is then split into user-level record trajectories by their device ID (Figure 5(c)). For effective visualization, we further extract spatiotemporal features from each location record to enrich the semantic representation of user's trajectories (Figure 5(d)). A suite of information-theoretic mobility metrics are proposed and computed over the long-term spatiotemporal distribution of user's trajectories (Figure 5(e)). Finally, city regions are profiled based on a cluster analysis of these multifaceted mobility metrics (Figure 5(f)). Below we describe more details on the last three stages, i.e., feature extraction, mobility metric design, and the city profiling. The scalability issue in the mobility analysis is also discussed.

### 4.1 Feature Extraction

As shown in Table 2, we obtain 90-day sample data sets on four Chinese cities. To make sure that each device is linked to a unique mobile user, a data cleansing step is employed over the device's record trajectories. The trajectories with smaller than one record per month or larger than 2,500 records per month are removed. The over-low record number can indicate a secondary device for its owner or the owner is not a local resident. The over high record number mostly reveals an abnormal usage type of the mobile device, such as robots, fixed sensors. After the cleansing, "user" will be used interchangeably with "device".

On the trajectory of each user, the major issue for mobility analysis is the data sparsity that there are hardly two records with exactly the same location or timestamp. The data sparsity makes it hard to synthesize all the trajectories into a single city panorama and to conduct clustering and comparison analysis. In this work, we propose to employ spatiotemporal feature extraction to compose a dense representation of each user's mobility out of their long-term trajectory. The benefits are two-fold: first, the huge amount of data can be aggregated in appropriate spatial and temporal granularities to reduce the data size; second, semantic information in space and time can be embedded into the mobility representation for interpretable visualization and analysis.

In details, we consider two types of spatial features. The first is the standard Point of Interests (POI) information on the map. Figure 6(a) lists ten classes of POIs used in this work, adapted from OpenStreetMap POI types [47]. Consider a user with his trajectory denoted as $N$ location records $\{L_i\}_{i=1,\cdots,N}$, we compute the probability of $L_i$ belonging to the $j$th POI class as

$$q_{ij} = \frac{\sum_k \varphi_{0,\sigma_{j,k}^2}(dist(L_i, POI_{j,k}))}{\sum_j \sum_k \varphi_{0,\sigma_{j,k}^2}(dist(L_i, POI_{j,k}))} \quad (1)$$

where $\varphi_{0,\sigma_{j,k}^2}$ indicates the standard Gaussian probability density function with zero mean and a variance of $\sigma_{j,k}^2$. $dist(L_i, POI_{j,k})$ indicates the spatial distance between the record's location $L_i$ and the $k$th POI of class $j$ ($POI_{j,k}$). The probability is in fact the sum of influences from all $j$th-class POIs, before normalized across all POI classes. An illustration is given in Figure 7(a). By default, $\sigma_{j,k}$, the influence variance of $POI_{j,k}$, is set to 1.5 times of its enclosure radius if the POI is area-based; or 100 meters if the POI is point-based.

The second type of spatial feature is the city's administrative division (DIV) of each location record. Because DIVs are mostly non-overlapping, the feature extraction is straightforward, i.e., $q_{ij} = 1$ or 0 for $L_i$ depending on whether or not $L_i$ is within the $j$th DIV.

The temporal features take a similar design to DIV by dividing time into non-overlapping intervals. We employ two types of time divisions. As shown in Figure 6(b), ID 0∼6 separate one day into 7 intervals; ID 7∼8 separate one week into weekday and weekend.

After the spatiotemporal feature extraction, each location record is represented by several vectors, as shown in Figure 5(d). Denote the POI-based vector of $L_i$ as $\vec{x}_i$, we have $\vec{x}_i =< q_{i1}, \cdots, q_{iM} >$, where $M$ is the number of POI classes ($M = 10$). Finally, the POI-based vector representation of the target user's trajectory, denoted

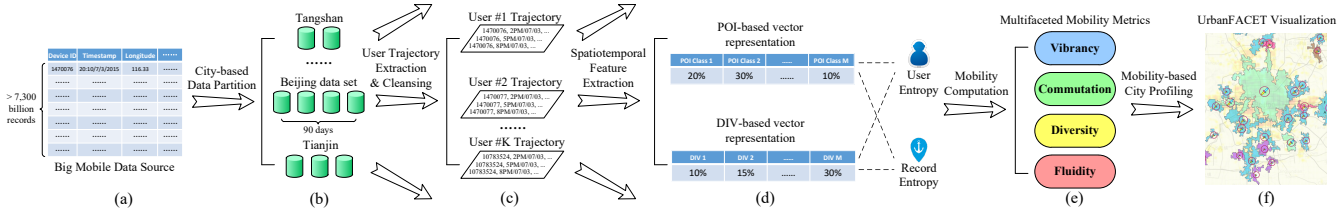



Fig. 5: The mobility analysis pipeline in UrbanFACET. The big mobile data is first partitioned into city-based data sets. The user-level trajectory and spatiotemporal features are then extracted to compute the multifaceted mobility metrics. Finally, the city regions are profiled by these mobility metrics through multivariate and spatial clustering.

| POI Class | POI Class Type |
|---|---|
| 1 | Food & Supply |
| 2 | Entertainment & Leisure |
| 3 | Education |
| 4 | Transportation |
| 5 | Healthcare & Emergency |
| 6 | Financial & Bank |
| 7 | Accommodation |
| 8 | Office & Commercial |
| 9 | Natural Landscape |
| 10 | Factory & Manufacturer |

(a)

| ID | Name | Time Interval |
|---|---|---|
| 0 | Morning | 6:00 ∼ 9:00 |
| 1 | Forenoon | 9:00 ∼ 12:00 |
| 2 | Noon | 12:00 ∼ 14:00 |
| 3 | Afternoon | 14:00 ∼ 17:00 |
| 4 | Evening | 17:00 ∼ 21:00 |
| 5 | Night | 21:00 ∼ 24:00 |
| 6 | Midnight | 0:00 ∼ 6:00 |
| 7 | Weekday | Monday ∼ Friday |
| 8 | Weekend | Saturday, Sunday |

(b)

Fig. 6: The list of (a) POI classes and (b) time interval divisions used in the feature extraction.

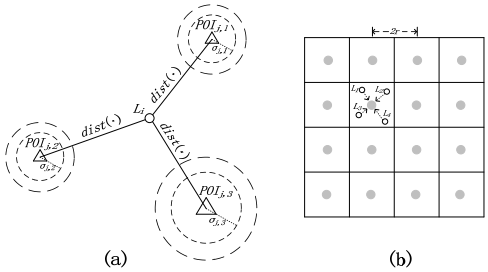



Fig. 7: Computing the POI class probability for each record: (a) sum of multiple Gaussian probability density functions; (b) the scalable computation method using grid-based data binning.

as $\vec{x} =< p_1, \cdots, p_M >$, is computed by summing all the record's vector together and normalizing on each class.

$$p_j = \frac{\sum_i q_{ij}}{\sum_j \sum_i q_{ij}} \tag{2}$$

$p_j$ denotes the probability for the user to access $j$th-class POIs. As $\sum_j p_j = 1$, the vector $\vec{x}$ is also the distribution of user's trajectory over POI classes. Similarly, each user will also have a DIV-based vector representation and two time-based representations.

### 4.2 Information-Theoretic Mobility Metrics

Over the dense vector representation of urban users, we follow-up to construct user-level mobility metrics, which can be visualized for human mobility analysis within the city. In fact, there have been plenty of literature on the study of human mobility. Nevertheless, they either focus on statistical and predictive models that are not suitable for urban mobility visualization [11] [13], or are designed to capture the short-term human mobility (e.g., temporal OD flows [6] [7] [37]). In comparison, the spatial or temporal distributions of user's trajectory, i.e., their vector representations, are most suitable to illustrate user's long-term mobility patterns.

A natural approach to compute such mobility metrics is to classify the feature vectors into multiple classes or user groups. However, after extensive studies, we discover two major difficulties in applying the classification method. First, in real-life big mobile data, all user's feature representation form a complex distribution on the vector space such that it is hard to determine both the number of user groups and their exact class boundaries. The second difficulty lies in extracting the common mobility metric from each user group after classification. This leads to the incapability to interpret the classification result for visualization.

In this work, we propose to use continuous user-level mobility metrics instead of the discrete urban user group. We introduce the notion of Shannon entropy [15] (or entropy in short) in defining user-level mobility metrics. This design brings three key advantages for the completion of targeted mobility analysis tasks (Section 3.3). First, by information theory, the entropy metric captures the uncertainty of distribution, which is ideal for measuring user's mobility. Meanwhile, entropy quantifies the average information acquired in seeing an instance of the random variable, i.e., the spatial/temporal category of a location record. Mapping user's entropy onto spatial regions can reveal the information content of this region, thus is feasible for city profiling tasks. Second, the entropy-based mobility metric is scalar when attached to each user's location record. It is easy to aggregate these mobility metrics over space and time for the overview analysis. Third, as the entropy is defined over distributions, the resulting metric is insensitive to the data sampling rate. Consequently, the city with a higher penetration rate can be effectively compared with another city with lower penetration rate.

Mathematically, take the POI-based vector representation $\vec{x} =< p_1, \cdots, p_M >$ as an example, we define **user entropy** as

$$H_p = -\sum_{j=1}^{M} p_j \cdot \log p_j \tag{3}$$

To visualize the user entropy on the geospatial map, a key choice is to determine the mobility of each record from their user entropy value. We propose two strategies here. The baseline strategy applies the user entropy directly to each record belonging to this user. By the baseline strategy, the mobility variation within a single user's trajectory is discarded, though in fact a user can bring more uncertainty (disorder) to their less-visited spatial classes. To resolve this deficiency, in the second strategy we decompose the quantity of user entropy across all of his location records, while maintaining the sum of entropy value unchanged.

$$\begin{aligned} N \cdot H_p = & \quad -\textstyle\sum_{j=1}^{M} N \cdot p_j \cdot \log p_j \\ = & \quad -\frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{M} q_{ij}} \cdot \textstyle\sum_{j=1}^{M}\sum_{i=1}^{N} q_{ij} \cdot \log p_j \\ \approx & \quad -\textstyle\sum_{j=1}^{M}\sum_{i=1}^{N} q_{ij} \cdot \log p_j \\ = & \quad \textstyle\sum_{i=1}^{N}(-\sum_{j=1}^{M} q_{ij} \cdot \log p_j) \end{aligned}$$

Based on this decomposition, we define **record entropy** by

$$H_r = -\sum_{j=1}^{M} q_{ij} \cdot \log p_j \tag{4}$$

Compared with the user entropy focusing on people's mobility property, the metric of record entropy emphasizes the property of spatial regions. Take the POI-class based record entropy as an example, a region will have a high average entropy if most people in this region are casual visitors.

Finally, the user/record entropy definition is combined with the candidate feature vector representations to assemble multifaceted user-level mobility metrics. In this work, we focus on the spatial mobility visualization for city profiling and comparison. Therefore, we will detail the case with POI-based and DIV-based vector representations. This leads to four mobility metrics (Figure 5(e)):

- *Vibrancy (user entropy over POI representation)*: this metric computes the degree of a user switching among different classes of POIs. As shown in Figure 6(a), the POI class indicates the type of resource connected to a particular region (e.g., food, leisure, education, etc). A high vibrancy people means s/he is connected to a larger number of resource types, therefore can be inferred as richer people. Consequently, the region with high vibrancy can be interpreted as the land where rich people live or visit.
- *Commutation (user entropy over DIV representation)*: this metric computes the degree of a user switching among different administrative divisions. As most records of a normal user's trajectory can be classified as home and work places. A high commutation people means s/he needs to commute a lot for everyday jobs. The region with high commutation indicates the work or home places of high commutation people, or the transportation lines.
- *Diversity (record entropy over POI representation)*: this metric computes the scarcity of a record's POI class in its user's POI distribution. A high diversity record means the user rarely visits this type of region/resource, thus raises security concerns. A high diversity region indicates that most users on that land are rare visitors to this type of region/resource. They can have very diversified POI distributions, and therefore interpreted as a group of very different people.
- *Fluidity (record entropy over DIV representation)*: this metric computes the scarcity of a record's administrative division in its user's DIV distribution. A high fluidity record means the user does not live or work in this region. A high fluidity region indicates that most users on that land can come from quite different work/home places. Typical high fluidity areas include national tourism POIs and transportation lines.

In the follow-up spatial visualization, the mobility metric of a region is computed as the average of all records in that region. The sum of metrics is not used because it is sensitive to the population density over regions.

### 4.3 Scalability Issues in Mobility Computation

In computing the user mobility metrics, the main challenge is to process the huge amount of urban mobility records. Denote the number of total records as $R$, in the city of Beijing, there are over 8 billion records within 90 days ($R = 8 \times 10^9$). On each mobility metric, the computation complexity is at least $O(R)$, hence the key is to minimize the time to process each record. According to Section 4.2, the most costly step is to calculate the probability of each record belonging to each POI class, e.g., $q_{ij}$ for the $i$th record belonging to the $j$th POI class. In theory, we need to fit the distance between each record location and all POIs using their respective Gaussian distribution, which involves a complexity of $O(R \cdot P)$, where $P$ is the number of POIs. This is infeasible in processing billions of mobility records and thousands of POIs.

In UrbanFACET, we apply a grid-based binning technique to resolve this scalability issue. As shown in Figure 7(b), we put a point set of square lattice on the map, with an intra-point interval of $2r$. The Voronoi diagram over this lattice gives a regular tessellation of squares, i.e., the grid-based space partition with a grid radius of $r$. According the property of Voronoi diagram, the center of each grid is the closest central point to all the other points in the same grid. Based on this property, we approximate the probability of each location affiliated with any POI class by the probability of the central point in the same grid. In this way, we only need to calculate the POI probability for all the grid centers. For the city of Beijing, there are about 0.2 million grids ($r = 100$ meters), four magnitudes smaller than the number of records. Also, the probability information of grid centers can be pre-processed for multiple grid settings. Finally on each record, the online computation is reduced to two multiplications on longitude/latitude to determine which grid the record belongs to.

Note that the binning technique usually introduces a reduced data precision and an aliasing effect on visualization. In our setting, these side effects are alleviated because the location record has an inherent error of 100 meters, which is used as the grid radius. Meanwhile, we apply a kernel density estimation based metric rendering in Section 5.1 to smooth the aliasing effect.

### 4.4 Mobility-based City Profiling

An important task of UrbanFACET is to visually profile city regions (e.g., their functionality) through the mobility metrics. The classical visualization design of heat map can display the spatial distribution of a single mobility metric (Figure 8(f)), but can not scale to support the panorama of multifaceted metrics. In Section 5.2, we introduce a flower-shape glyph design to visually encode multifaceted mobility metrics. To further reduce the overlap among these glyphs and alleviate information overload in the panorama, we propose to employ mobility-based region clustering. The cluster of regions with similar mobility metrics can be overlaid with a single flower-shape glyph for the mobility profiling and visualization.

In the region clustering, the regular grids used in the metric computation are set as the basic spatial unit. We propose a three-step hierarchical clustering method. First, all the grids in a city are clustered based on the pairwise distance between their multifaceted mobility metric vectors. In the implementation, Euclidean distance and the k-means clustering algorithm [48] is applied. Each component of the mobility vector (i.e., a single mobility metric) is normalized into $[0,1]$ across all grids before used in the clustering. This step ensures that the grids in the same cluster will have similar mobility metric values. Second, for each region cluster generated by k-means, all of its underlying grids are clustered again by the spatial clustering algorithm, in our implementation, DBSCAN [49]. This ensures the spatial affinity of the second-level clusters. Third, all these second-level clusters are filtered by the total grid number (i.e., cluster area size) so that insignificant clusters are removed in the visualization.

For the sake of interactive city profiling, we allow users to set the number of k-means clusters, the internal density threshold of spatial clusters (by fixing *eps* and adjusting *minPts* in DBSCAN), and the minimal cluster area size in different analysis scenarios.

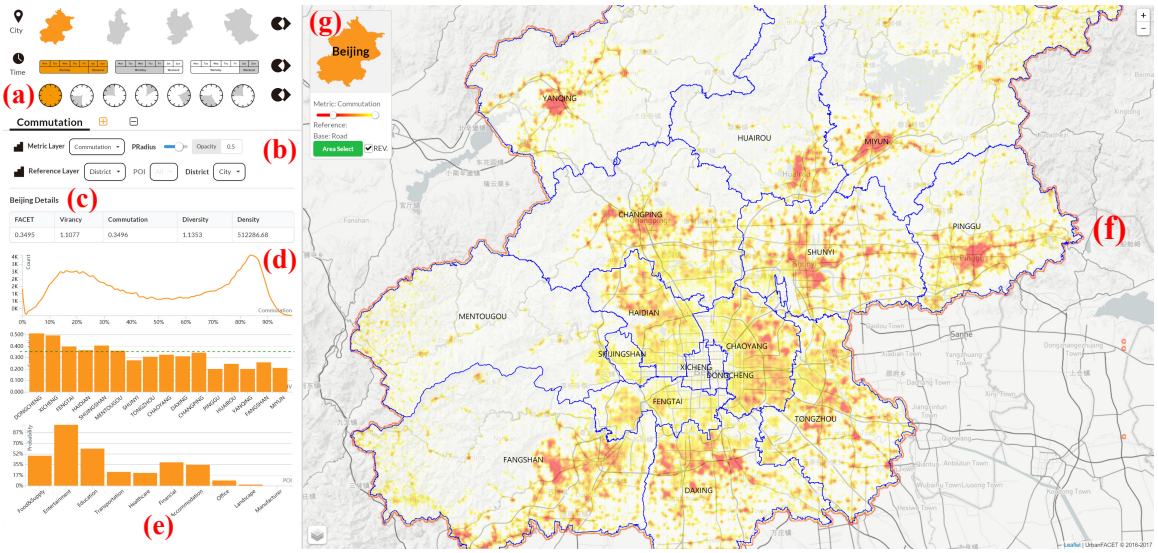



Fig. 8: The visualization interface of UrbanFACET: (a) the selection on city and time; (b) the selection of mobility metrics, and the visualization options; (c) detailed information on the selected region or the city if nothing is selected; (d) probability density function of the selected mobility metric; (e) average probability distribution on POI/DIV classes to compute the mobility metric; (f) the spatial visualization of mobility metrics (commutation in this case, the color mapping is inversed); (g) legend and controller of the map.

## 5 Visualization Design

The UrbanFACET visualization interface (Figure 8) is designed to fulfill three human mobility analysis tasks in Section 3.3. First, the overview task, users start by picking the target city, time period, and mobility metric to analyze in the top-left control panel (Figure 8(a)(b)). The overall distribution of this metric is then displayed in the main panel (Figure 8(f)). The exact value and probability distribution of the selected metric is displayed in the bottom-left detail panel for fine-grained analysis (Figure 8(c)(d)(e)). Second, the profiling task, multiple mobility metrics can be displayed simultaneously by a flower-shape glyph design, which is overlaid on the main panel to visually profile the spatial regions of a city. Third, the main panel can be split into juxtaposed sub-views to compare between cities, time periods, or different metrics.

In the following, we describe the detailed design of metric visualization in the context of geospatial maps (Section 5.1), the mobility-based city profiling (Section 5.2), and the interactions for spatiotemporal comparison analysis (Section 5.3).

### 5.1 Mapping User's Mobility Metrics

The map-based visualization in the main panel has two default layers: the base layer and the metric layer. The base layer in the background gives the geospatial information about the city, serving as location references. Several base layer types are available for separate analysis scenarios, e.g., road network, terrain and satellite imaging maps. On the other hand, the metric layer is overlaid on top of the base layer and displays the key information of metric distribution within the city. A number of metrics can be selected to display (Figure 8(b)), including the four mobility metrics proposed in Section 4.2, the classical record density metric, and three regional city demographics (GDP, population, house price).

On the metric layer, the default scatterplot method does not scale to render billions of records on a single map. In UrbanFACET, we propose a kernel density estimation (KDE) based visualization method over the data binning technique in Section 4.3. Take the vibrancy metric as an example, we first aggregate all the records according to the $Q$ regular grids in data binning. Then on each grid $G_i$, one data point is sampled at the grid center $x_i$, with the average vibrancy $c_i$ on this grid as the metric value. Finally, the density function by KDE is computed as

$$\hat{f}_h(x) = \frac{1}{h}\sum_{i=1}^{Q} c_i \cdot \varphi_{0,1}\left(\frac{x - x_i}{h}\right) \tag{5}$$

where $\varphi_{0,1}$ is the standard normal kernel function, $h$ is the bandwidth parameter controlling the smoothing level.

The proposed visualization method is considered better in our scenario than other aggregation methods. First, binning without KDE, i.e., filling each grid with the average mobility metric, is shown to produce aliased visualization, as shown in Figure 9(a). This is detrimental to the display of spatial trends and patterns. Second, the standard KDE over individual records [50] [51] can create a more accurate visualization with respect to the raw data, but will induce a large online computation overhead. The time cost can reach to several minutes to estimate the distribution from 8 billion records in Beijing with the GPU-based acceleration [50] and much longer without GPU [51]. In comparison, KDE over 0.2 million grids only needs about two seconds in our implementation without GPU. Last, as the grid radius has been set to the potential localization error, this grid-based approach will not introduce artifacts if the KDE bandwidth is appropriately chosen.

The selection of the bandwidth parameter is critical to the visualization result. Our finding coincides with Ref. [50] that there is no single optimal parameter because the actual density distribution is unknown. An appropriate bandwidth should be larger than the grid radius, and be adaptive to the zoom factor of the current view. Therefore, we set the initial bandwidth as two times of the grid radius according to Ref. [50], and the bandwidth decreases/increases by the same ratio with the increase/decrease of zoom factor. Figure 9(b) shows the resulting KDE visualization with the initial bandwidth setting. When the view is zoomed in, if the bandwidth is fixed, there can be a granularity problem so that the display resolution is reduced, as shown in Figure 9(c). By the adaptive bandwidth setting, the visualization in Figure 9(d) illustrates finer-grained details upon the zoom-in operation.


(a) (b) (c) (d) (e)

Fig. 9: Metric layer visualization methods: (a) grid-based binning with the single-color palette; (b) KDE over grids with the intial bandwidth and range filters, the two-color palette is used; (c) zoom-in view with the fixed bandwidth setting; (d) zoom-in view with the adaptive bandwidth setting; (e) the flower-shape glyphs to represent the multifaceted mobility metrics.

In rendering the KDE visualization, there are two additional design choices: the color palette and the mapping function from the metric value to colors. As shown in Figure 9(b)(c)(d), we adopt a two-color palette to display the metric value. The highest and lowest values are mapped to red and yellow respectively, the intermediate value is interpolated in the color hue space. We do not use the single-color palette as recommended in Ref. [52] [53] for the comparison task of sequential data. As shown in Figure 9(a), the single-color palette can introduce ambiguity between the low metric value and the empty value (e.g., no value or filtered out). The standard rainbow palette with many base colors is also considered harmful due to the lack of perceptual color ordering [54]. It is only used in comparing very similar distributions, as shown in Figure 15. The two-color palette features a compromise that low values can be noticed while the color ordering issue is minimized. By default, users interpret red as hot spot in heat map.

On the color mapping function, the main challenge is the skewed metric value distribution in both space and statistics. As shown in Figure 1(c), the vibrancy distribution in Beijing concentrates in a small value range, and different metrics/cities can have rather diversified value hot spots. Mapping metric values linearly to the color palette will make it hard to detect distributional patterns. We propose a ranking based color mapping approach. All the metric values (in grids) are first ranked and their percentiles are used in a linear mapping to the color palette. This approach naturally supports the comparison among juxtaposed mobility maps by ranking the grids of all the maps together.

Meanwhile, the metric distribution is also skewed in space, e.g., the density in the city metro is much higher than other regions. All grids in the city metro will have similar high percentiles, which are hard to distinguish even with linear color filters. In this work, we resolve the issue by applying a log-scale color filters. As shown in the legend panel of Figure 1(a), the high-percentile value range corresponds to a larger interval on the filter. The color filter is designed as double-ended to select the visible range by the percentile of the metric. The grids in visible metric percentiles are mapped to the entire color palette and then displayed in the metric layer visualization. This color mapping can also be reversed for the analysis of low percentile regions, as controlled by the "REV." checkbox in Figure 8(g).

To enable fine-grained analysis, we implement three statistical charts in the detail panel of UrbanFACET. The top view (Figure 8(c)) lists the average value of all metrics in the currently selected region. If no region is selected, the entire city is considered. The middle view (Figure 8(d)) depicts the probability density function of grid-based metric values. This distribution helps users to set the color filter in the analysis. The bottom view (Figure 8(e)) depicts user's spatial distribution among POI/DIV classes.

### 5.2 Profiling Cities with Multifaceted Mobilities

Over base and metric layers, UrbanFACET designs another reference layer to display the side information (Figure 8(b)), such as POIs, DIV-based regions, demographics and multifaceted mobility metrics. This reference layer helps to visually profile city regions through the correlation between spatial and human mobility information.

For DIV references, by default the district partition is overlaid on top of the metric layer to show the correlation of DIVs with the metric value (Figure 8(f)). The region granularity can be adjusted by the region controller into sub-districts and streets. For multifaceted mobility metrics, we propose a novel flower-shape glyph design, as shown in Figure 9(e). The glyph is composed of four colored petals and one ring surrounding the petals, which visualizes the 4-tuple of mobility metrics plus density: $(Fluidity, vibrAncy, Commutation, divErsity, densiTy)$. The area size of the ring is designed to be perceptually proportional to the normalized density metric across the current view. Each petal within the ring represents one mobility metric, with the area size of the petal proportional to the normalized metric value. The largest mobility metric in the view will have its petal stretched to the inner rim of the ring, as shown by the four petals of Dongcheng district in Figure 9(e). Upon mouse hovering of a glyph, the detailed value of all five metrics are shown in a pop-up label. The hovered glyph also grows to the same size with the largest glyph in the current view, in order to observe the multifaceted mobility patterns for low-density areas. In some cases, all glyphs are set to the same size to only compare the mobility metrics (Figure 13(a)). There are two options for the multifaceted metric visualization, one by overlaying the flower-shape glyphs on each DIV to compare the mobility of administrative regions (Figure 13(a)). Another is to visualize the result of mobility-based region clustering (Section 4.4), in which adjacent regions similar in mobility metrics are put together and profiled with their mobility metrics (Figure 13(b)(c)).

### 5.3 Comparing among Metrics, City and Time

On the comparison task, UrbanFACET relies on a side-by-side interface design that juxtaposes multiple spatial maps as sub-views (e.g., Figure 12(a)). This is achieved by the split operation using the "+" button on top of the metric layer configuration in Figure 8(b). Each time, the current base/metric/reference layer setting is replicated in the newly created sub-view. The new map can be configured in the same manner to a different metric visualization

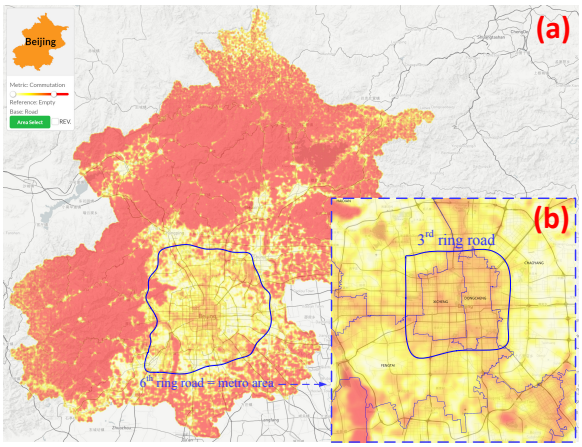



Fig. 10: The commutation metric distribution in Beijing: (a) the full scale visualization of the city; (b) the metro area is enlarged.

for comparison. By this design, multiple mobility metrics can be compared, the mobility metric and the regional demographics (GDP, population, house price) can also be correlated.

On comparing among cities or different time periods, we provide shortcut buttons as shown in the rightmost column of Figure 8(a). The system supports the comparison of up to four cities (i.e., Beijing, Tianjin, Tangshan, Zhangjiakou, as shown in Figure 14), and six time periods based on the division in Figure 6(b) excluding midnight (Figure 15). Switching among cities and filtering by time periods are also supported by the iconized selector in the middle column of Figure 8(a).

A suite of interaction methods are designed for the side-by-side comparison. First, on time-based comparison, zoom&pan on each sub-view is synchronized with all the other sub-views, in order to preserve the spatial context. Because cities will have different spatial maps, the zoom&pan operation is not synchronized in the city-based comparison. Second, the range filter on the color mapping works in a centralized manner. Specifying the filter on one sub-view will affect all the other sub-views. This is achieved by the ranking-based percentile filtering mechanism.

## 6 Case Study

We deployed UrbanFACET to analyze the human mobility in four representative cities across China (Table 2), which forms the so-called national capital region. Recently, the unification of national capital region has become a hot topic in China. Profiling and comparing these cities in terms of the human mobility provides key values to the coordinated development of this pivotal region. In this case study, we will focus on Beijing for the mobility analysis and profiling of a single city, and then extend to all four cities in the comparison analysis.

### 6.1 Overview of City Mobility

First, we conduct a detailed analysis on the distribution of four long-term mobility metrics over Beijing, i.e., vibrancy, commutation, diversity and fluidity. Traditionally, the metropolitan area of Beijing, aka metro in short, is defined by the region within the 6th ring road (the blue enclosure in Figure 1(a), enlarged in Figure 1(b)). The metro is centered at Tiananmen, and can be divided into the northern and southern city by the Chang'an avenue, and also divided into five ring-shaped regions by the 2nd to 6th ring roads (Figure 1(b)). The population density of Beijing spreads radially from the center to the remote area of the city along these ring-shaped regions.

**Vibrancy:** As discussed in Section 4.2, the vibrancy metric indicates the degree of a resident switching among different classes of POIs. A high vibrancy region means the residents there in average access many different types of resources, thus potentially lives a richer life. The overall vibrancy distribution in Beijing is shown in Figure 1(a). Two classes of high vibrancy regions can be found: a) the central area of the metro of Beijing; b) the western, northern and northeastern regions outside the metro area.

We drill down to these two classes to obtain more findings on the vibrancy distribution. First, in the metro area, we observe an asymmetric pattern that is quite different from the population density distribution, as shown by the enlarged vibrancy distribution in Figure 1(b). The core of the high vibrancy region is rectangle-shaped, from the 5.5th ring in the north to the 3rd ring in the south, and from the 4th ring in the east to the 3.5th ring in the west. In this sense, it can be inferred that the northern people lives a more abundant life than the southern people. This finding agrees well with the unbalanced wealth distribution in Beijing ever since the Ming dynasty: the forbidden palace and the inner city both locate in the northern city. Second, by comparing with the terrain map in Figure 1(d), we discover that the three high vibrancy regions outside the metro correspond to three major mountains surrounding Beijing in the west, north and northeast. To reason this result, we sample 200 users that ever visit these mountain areas and compute their overall record distribution. It can be found that the sampled distribution has almost 70% records in the metro area of Beijing. In other words, the location records in the mountains are mostly made by the travelers who live in the Beijing metro area, thus spreads the high vibrancy pattern.

It is also interesting to observe the vibrancy valleys of Beijing. As shown in Figure 1(b), between the 2nd and 3rd ring roads in the north, a vibrancy valley locates in a famous military region known as the Huangsi Compound, having no POI on the map. In comparison, the Central and Southern Sea, serving as headquarters to the central government, has a low record density (Figure 14(c)) and medium vibrancy (Figure 1(a)).

**Commutation:** The distribution of the commutation metric in Beijing is shown in Figure 10(a). The mountain areas outside the metro have a large commutation need, which is understandable. Meanwhile, unlike the vibrancy distribution, the commutation in the metro is relatively low. If we take a closer look, as shown in Figure 10(b), the metro area can be divided into two sub-regions. The sub-region within the 2.5th ring, mainly the old city districts of Dongcheng and Xicheng, has a higher commutation. We hypothesize that people mainly work in this sub-region, which is the center of Beijing, and do not live there because of the overhigh house price and inconvenient life in the old city. The second sub-region is between the 2.5th ring and 5.5th ring, where the commutation is among the lowest in Beijing. This is because people here can afford the house price near their workplace and also find these modern city districts attractive for life. Some outliers in this sub-region include the Xiangshan park, a famous city recreation site, and Yizhuang, the Beijing Economic Development Area, the workplace for a lot of people who do not live there.

With an inversed color mapping of the commutation distribution (Figure 8(f)), more valleys outside the metro of Beijing are found to be the center of suburb districts (e.g., Yanqing, Miyun, Pinggu). The residents there can meet their everyday demands in the district center and do not need to commute to the metro.

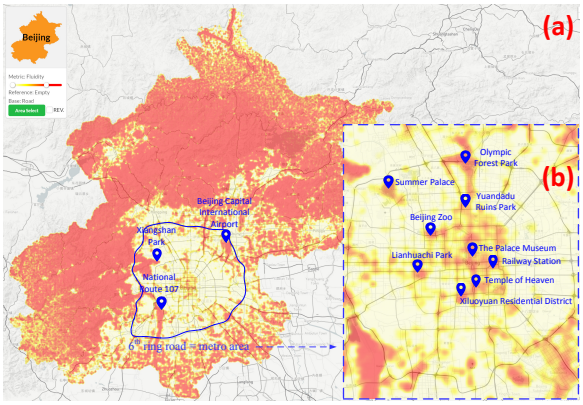



Fig. 11: The fluidity metric distribution in Beijing: (a) the full scale visualization of the city; (b) the metro area is enlarged. Several high fluidity areas are annotated with their corresponding POIs.

**Diversity:** Within the metro area, the diversity distribution in Beijing is similar to that of the vibrancy, as shown in Figure 15. Meanwhile, in the mountain regions, the diversity is mostly zero. This is because the diversity serves as the property of land, in comparison to the vibrancy metric measuring people. Despite of many vibrant visitors, the mountain regions do not provide POIs to visit, thus has a low diversity.

**Fluidity:** In the fluidity distribution of Beijing (Figure 11(a)), there is a three-layer pattern similar to the commutation distribution. Outside the metro, the fluidity is high in the mountain regions, except the valley on suburb centers. This validates our hypothesis that the records on the mountain are made by occasional travellers whose base are in different districts of the metro area. Inside the metro, the region outside the 2.5th ring road mostly has a low fluidity. Some exceptions happen in the airport, Xiangshan park and the National Route 107 (the busiest road connecting northern and southern China), where people from different districts come for recreation or transportation.

Meanwhile, the inner region of the metro (within the 2.5th ring) has a relatively high fluidity. In an enlarged distribution (Figure 11(b)), the fluidity hotspots are found to be co-located with several world-class tourism attractions, such as the Forbidden City, the Temple of Heaven, the Summer Palace, the Olympic Park, etc.

The overview of mobility metrics in Beijing provides many recommendations: for city governors and planners, more investments can be made in the southern area to foster a balanced development of the whole city; for enterprise managers, new sites and offices can be opened between the 2.5th and 5th ring roads so that employees can benefit from fewer commutations; for security officers, more attentions should be paid on famous tourism attractions where the risk is higher with visitors from different places.

### 6.2 Profiling City Regions

Beyond the overview visualization, the mobility metrics can be correlated with other city demographics to profile the land style of the city in multiple granularities. For example, Figure 12(a) juxtaposes the average vibrancy of Beijing by administrative divisions with the city's house price distribution. While the vibrancy and house price are positively correlated in most districts, four remote districts, i.e., Shunyi/Huairou/Changping/Daxing, are shown to have relatively higher vibrancy ranks than their house prices among remote districts. These districts can be inferred as more livable by having a better balance in life abundance and living cost. This finding is validated by the per capita GDP distribution of Beijing in Figure 12(b). Shunyi/Huairou/Daxing are among the top in districts outside the metro, and Shunyi even has a higher GDP than some districts in the metro (e.g., Haidian and Chaoyang).

In a more detailed study, the four mobility metrics can be used altogether to classify the functionality of city districts. As shown in Figure 13(a), all the districts of Beijing are clustered into four classes. The mobility metrics of each district are depicted by the flower-shape glyph on the map. Dongcheng/Xicheng, the two districts in the inner/old city, have the largest mobility in all four metrics. They are diversified and fluid, their residents are vibrant and commuting. Moving outside, Haidian/Chaoyang are also core districts in the Beijing metro. Their vibrancy and diversity are still high, but the commutation and fluidity are low. People there do not need to commute much and tend to be local residents. Moving further outside, Changping/Shunyi/Huairou/Shijingshan/Fengtai/Daxing are remote districts similar to Haidian/Chaoyang, but with a lower vibrancy and diversity. People there tend to have downgraded life quality. The other six districts are rural areas having the lowest vibrancy and close to zero diversity. There are little POIs to ensure the life quality except around the district center.

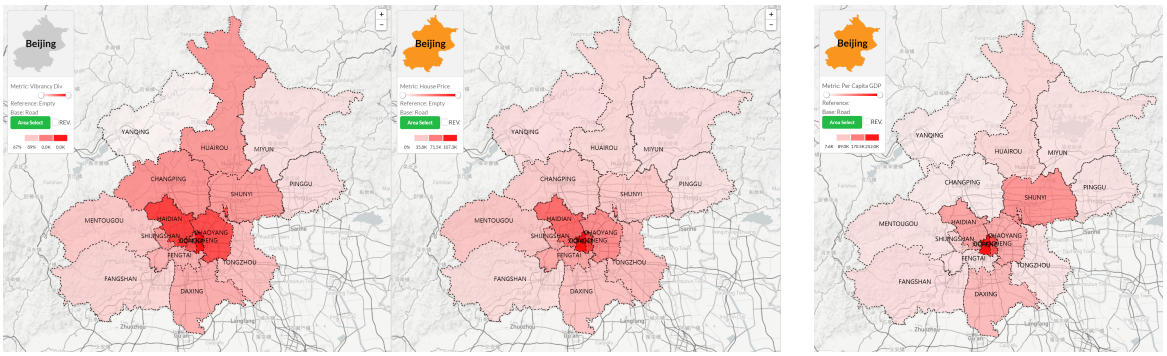
In the finest granularity, the city can be profiled on the basis of regular grids used in the metric computation. By the region clustering algorithm in Section 4.4, spatial clusters of grids can be detected and visualized. As shown in Figure 13(b), there are six major cluster types by the default setting, which are indicated by their filled colors. The green cluster encompasses the core area of Beijing metro plus some well-known wealthy regions in the surrounding, e.g., Shangdi (China's Silicon Valley), Yizhuang. The signature of this cluster is the very high vibrancy and diversity, corresponding to a high quality of life and prosperous land. The blue cluster mainly locates outside the green cluster, featuring a medium vibrancy and low diversity. This shows that people there still lives a good life, but the land is not as prosperous as the core of metro, indicating major area for accommodations. The pink cluster is similar to the blue cluster, but is much smaller in size and higher in vibrancy, which corresponds to upper-class accommodation for rich people. The house price in the pink cluster is generally higher than the surroundings.

Meanwhile, the purple cluster co-locates with the center of suburb districts, having low values in all four metrics. The orange cluster appears mostly in the mountain regions, corresponding to several tourism sites such as the Great Wall, the Tanzhe Temple, the Phoenix Mountain, etc. They have a high vibrancy, fluidity and commutation value, indicating a composite of tourists from different metro districts. The last brown cluster features a high fluidity and vibrancy, but medium commutation, interpreted as a mixture of business people that commutes less than the tourists. The only brown cluster is found at the Beijing airport. These clusters can be extended by setting a smaller cluster size threshold. As shown in Figure 13(c), more clusters are detected, notably the increased orange tourism sites in the mountain area.

The above profiling result provides the city governor a better classification of urban regions than the default administrative division. The urban regions in the same cluster can be managed with the same policy and development plan because of their uniformity in residents and land style.

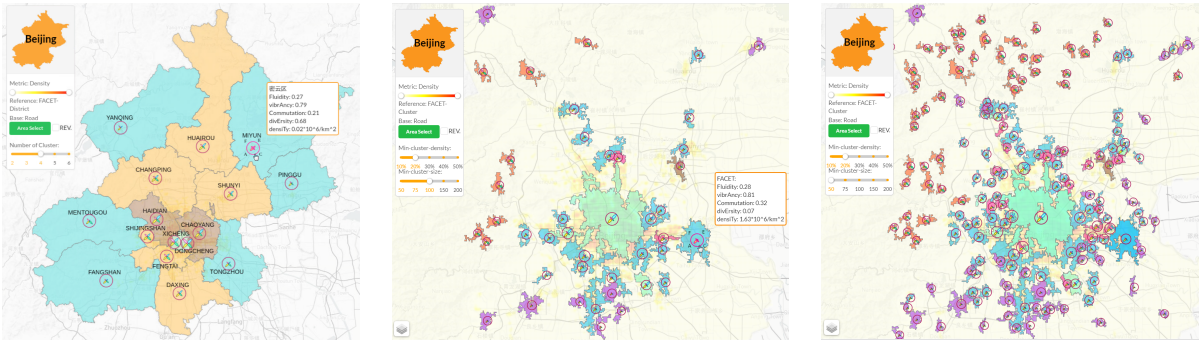
### 6.3 Comparison among Time and City

UrbanFACET supports the comparison of mobility distributions in multiple cities. Figure 14(a) displays the vibrancy distribution




(a) (b)

Fig. 12: The correlation between mobility metrics and city demographics: (a) vibrancy and house price distribution in all districts of Beijing; (b) per capita GDP distribution.




(a) District clusters (b) Grid clusters (medium cluster area) (c) Grid clusters (small cluster area)

Fig. 13: Profiling the city of Beijing by all four mobility metrics: (a) the clusters on districts; (b) the clusters on grids; (c) more clusters on grids by setting a smaller cluster area threshold.

of the four cities in the national capital region. The spatial grids with the top 25% vibrancy are selected to show. Beijing, as the capital of China, enjoys a large high-vibrancy core in the metro. In comparison, Tianjin/Tangshan only have a couple of small areas with vibrant people; on the other hand, the city of Zhangjiakou is relatively weak in life abundance even in the metro area.

On commutation, as shown in Figure 14(b), the distribution in Beijing and Tianjin are similar in that the core of metro area has a medium commutation, which is surrounded by the low commutation metro area. The remote area has a high commutation because of the resident's need to travel to the metro for life essentials. On the other hand, Tangshan and Zhangjiakou only have a two-layer commutation pattern. The metro area is low and the outer area is high. Note that both cities have multiple metro areas. The central difference is, the residents in the core metro area of Beijing and Tianjin still need to commute, which can be seen as an outstanding feature of big cities. In small to medium cities like Tangshan and Zhangjiakou, people have much shorter commutation distance even they work/live in the city center.

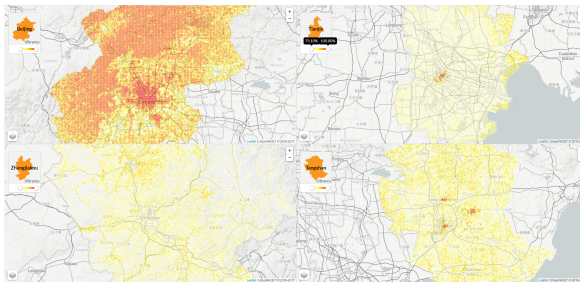
The mobility patterns discovered in the above comparison reveal a brand new insight on the life style of these four cities, which are impossible to acquire from the traditional population density based map. As shown in the record density distribution of Figure 14(c), all cities populate from the center to the surroundings, and the city size decreases from Beijing to Tianjin, then to Tangshan and Zhangjiakou. The density comparison between Beijing and Tianjin are not so overwhelming as their comparison in vibrancy (Figure 14(a)).

UrbanFACET also supports the time-based comparison. In Figure 15, it can be noticed that Beijing is less diversified in the early morning and the late midnight, when most people are probably at home. The highest diversity happens in the evening, when a large number of residents will have some entertainments after their daily work.

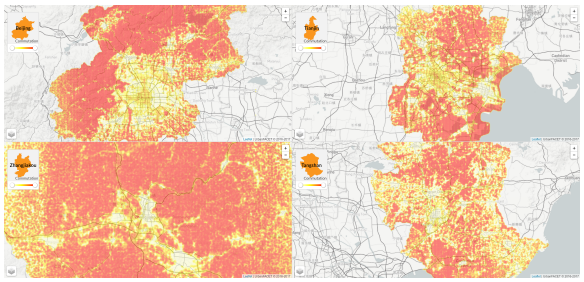
The above comparison result provides several valuable suggestions. For city planners, the traffic problem of Beijing and Tianjin can be studied together due to their similar commutation pattern; for luxury sellers, it is better to target Beijing as the potential market because of the larger number of rich people there; for security officers, the period in the evening is the most noteworthy time as many city regions get crowded with different kinds of people.

## 7 Conclusion

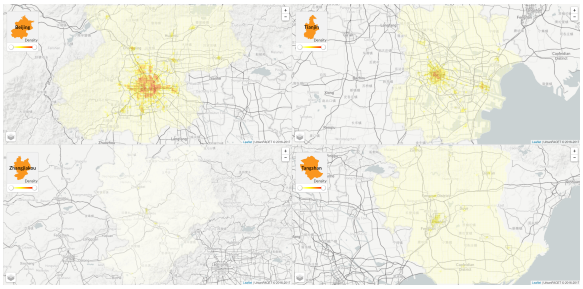
We present a metropolitan-scale visual analytics study over a new urban big data in China. Compared with previous works, our data set is a better sample in space, time and population-wise for long-term human mobility study in the city. To efficiently visualize, profile, and compare the mobility pattern across multiple regions and cities, we have proposed: 1) a suite of information-theoretic metrics to effectively characterize long-term individual-level mobility patterns; 2) a scalable, grid-based data analytics pipeline to compute these mobility metrics and profile functional city regions; 3) an integrated visual analytics system, namely UrbanFACET, to support the interactive, profiling and comparative analysis on the multifaceted mobility panorama and its correlation with urban structure and POI distributions. Real-world case studies are conducted over big cities in China, which provides recommendations to many important urban users such as city planners, governors, and security officers.



(a) Vibrancy



(b) Commutation



(c) Density

Fig. 14: The comparison of mobility metric distribution among Beijing, Tianjin, Tangshan and Zhangjiakou.

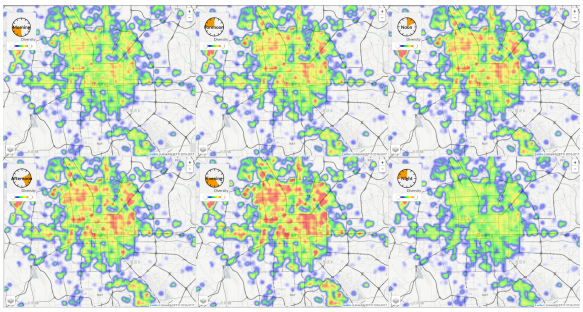


Fig. 15: The diversity metric distribution in Beijing. The standard rainbow palette is used to compare similar distributions over time.

In future, we plan to extend UrbanFACET in three directions. First, beyond the macroscopic city profiling, it is also important to model the fine-grained city regions and their local residents. Second, besides the spatial entropy metrics, we are interested to define new mobility metrics over the temporal distribution of user's movement records. Third, visual storytelling with multifaceted urban panorama poses a demanding challenge to extend the user base of UrbanFACET beyond domain experts.

## REFERENCES


[1] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 220–232, 2013.

[2] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, 2013.

[3] S. Phithakkitnukoon, T. Horanont, G. Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: Identifying human daily activity pattern using mobile phone data," *Human Behavior Understanding*, pp. 14–25, 2010.

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

[5] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?" *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, no. 3, pp. 33–44, 2012.

[6] J. Wood, A. Slingsby, and J. Dykes, "Visualizing the dynamics of london's bicycle hire scheme," *Cartographica*, vol. 46, no. 4, pp. 239–251, 2011.

[7] G. Andrienko, N. Andrienko, G. Fuchs, and J. Wood, "Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data," *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[8] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, "Understanding intra-urban trip patterns from taxi trajectory data," *Journal of Geographical Systems*, vol. 14, no. 4, pp. 463–483, 2012.

[9] L. Liu, C. Andris, and C. Ratti, "Uncovering cabdrivers' behaviour patterns from their digital traces," *Computers, Environment and Urban Systems*, vol. 34, no. 6, pp. 541–548, 2010.

[10] H. Liu, Y. Gao, L. Lu, S. Liu, L. Ni, and H. Qu, "Visual analysis of route diversity," *VAST'11*, pp. 171–180, 2011.

[11] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462–465, 2006.

[12] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM transactions on networking (TON)*, vol. 19, no. 3, pp. 630–643, 2011.

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

[14] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, pp. 818–823, 2010.

[15] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[16] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, p. 38, 2014.

[17] B. Jiang and C. Liu, "Street-based topological representations and analyses for predicting traffic flow in gis," *International Journal of Geographical Information Science*, vol. 23, no. 9, pp. 1119–1137, 2009.

[18] D. Tong, C. J. Merry, and B. Coifman, "New perspectives on the use of gps and gis to support a highway performance study." *Transactions in GIS*, vol. 13, no. 1, pp. 69–85, 2009.

[19] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "Vtrack: Accurate, energy-aware road traffic delay estimation using mobile phones," in *SenSys'09*, 2009, pp. 85–98.

[20] J. Zimmerman, A. Tomasic, C. Garrod, D. Yoo, C. Hiruncharoenvate, R. Aziz, N. R. Thiruvengadam, Y. Huang, and A. Steinfeld, "Field trial of tiramisu: Crowd-sourcing bus arrival times to spur co-design," in *CHI'11*, 2011, pp. 1677–1686.

[21] C. Ratti, R. Pulselli, S. Williams, and D. Frenchman, "Mobile landscapes: using location data from cellphones for urban analysis," *Environment and Planning B: Planning and Design*, vol. 33, no. 5, pp. 727–748, 2006.

[22] N. Shoval, "Tracking technologies and urban analysis," *Cities*, vol. 25, no. 1, pp. 21–28, 2008.

[23] P. Wang, T. Hunter, A. Bayen, K. Schechtner, and M. González, "Understanding road usage patterns in urban areas," *Scientific reports*, vol. 2, p. 1001, 2012.

[24] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, p. 1376, 2013.

[25] K. Chaogui, L. Yu, and W. Lun, "An analysis of entropy of human mobility from mobile phone data," *Geomatics and Information Science of Wuhan University China*, vol. 42, no. 1, pp. 63–69, 2017.

[26] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *MobiCom'07*, 2007, pp. 123–134.

[27] W. Zeng, C.-W. Fu, S. Müller Arisona, A. Erath, and H. Qu, "Visualizing Mobility of Public Transportation System," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1833–1842, 2014.

[28] N. Lathia and L. Capra, "Mining mobility data to minimise travellers' spending on public transport," in *KDD'11*, 2011, pp. 1181–1189.

[29] S. Sang, M. OḰelly, and M.-P. Kwan, "Examining commuting patterns: results from a journey-to-work model disaggregated by gender and occupation," *Urban Studies*, vol. 48, no. 5, pp. 891–909, 2011.

[30] H. Guo, Z. Wang, B. Yu, H. Zhao, and X. Yuan, "Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection," in *PacificVis'11*, 2011, pp. 163–170.

[31] P. A. Laharotte, R. Billot, E. Come, L. Oukhellou, A. Nantes, and N. E. E. Faouzi, "Spatiotemporal analysis of bluetooth data: Application to a large urban network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1439–1448, 2015.

[32] Q. Li, T. Zhang, H. Wang, and Z. Zeng, "Dynamic accessibility mapping using floating car data: a network-constrained density estimation approach," *Journal of Transport Geography*, vol. 19, no. 3, pp. 379–393, 2011.

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

[34] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti, "Discovering the geographical borders of human mobility," *KI-Künstliche Intelligenz*, vol. 26, pp. 253–260, 2012.

[35] E. van der Hurk, L. Kroon, G. Marti, and P. Vervesti, "Deduction of passengers' route choices from smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 430–440, 2015.

[36] S. Gao, Y. Liu, Y. Wang, and X. Ma, "Discovering spatial interaction communities from mobile phone data," *Transactions in GIS*, vol. 42, no. 1, pp. 463–481, 2013.

[37] Y. Ma, T. Lin, Z. Cao, C. Li, F. Wang, and W. Chen, "Mobility viewer: An eulerian approach for studying urban crowd flow," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 96, pp. 2627–2636, 2016.

[38] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom, "Thematic patterns in georeferenced tweets through space-time visual analytics," *Computing in Science and Engineering*, vol. 15, no. 3, pp. 72–82, 2013.

[39] P. Jankowski, N. Andrienko, G. Andrienko, and S. Kisilevich, "Discovering landmark preferences and movement patterns from photo postings," *Transaction in GIS*, vol. 4, no. 6, pp. 833–852, 2010.

[40] F. Miranda, H. Doraiswamy, M. Lage, K. Zhao, B. Gonçalves, L. Wilson, M. Hsieh, and C. T. Silva, "Urban pulse: Capturing the rhythm of cities," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 791–800, 2017.

[41] Y. Zhang, "User mobility from the view of cellular data networks," *IEEE INFOCOM*, pp. 1348–1356, 2014.

[42] W. Chen, F. Guo, and F. Y. Wang, "A survey of traffic data visualization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 2970–2984, Dec 2015.

[43] G. Andrienko, N. Andrienko, W. Chen, R. Maciejewski, and Y. Zhao, "Visual analytics for transportation: State of the art and further research directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2229–2231, 2017.

[44] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne, "Flowstrates: An approach for visual exploration of temporal origin-destination data," in *Computer Graphics Forum*, vol. 30, no. 3, 2011, pp. 971–980.

[45] T. Schreck, J. Bernard, T. Von Landesberger, and J. Kohlhammer, "Visual cluster analysis of trajectory data with interactive kohonen maps," *Information Visualization*, vol. 8, no. 1, pp. 14–29, 2009.

[46] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining user similarity based on location history," in *SIGSPATIAL'08*, 2008, p. 34.

[47] O. Wiki, "Map features - openstreetmap wiki," 2016, http://wiki.openstreetmap.org/wiki/Map_Features.

[48] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[49] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *KDD'96*, 1996, pp. 226–231.

[50] O. D. Lampe and H. Hauser, "Interactive visualization of streaming data with kernel density estimation," in *PacificVis'11*, 2011, pp. 171–178.

[51] N. Willems, H. Van De Wetering, and J. J. Van Wijk, "Visualization of vessel movements," in *Computer Graphics Forum*, vol. 28, no. 3. Wiley Online Library, 2009, pp. 959–966.

[52] C. Tominski, G. Fuchs, and H. Schumann, "Task-driven color coding," in *IV'08*, 2008, pp. 373–380.

[53] S. Mittelstädt, D. Jäckle, F. Stoffel, and D. A. Keim, "Colorcat: Guided design of colormaps for combined analysis tasks," in *EuroVis'15 short papers*, vol. 2, 2015.

[54] D. Borland and R. M. T. Ii, "Rainbow color map (still) considered harmful," *IEEE Computer Graphics and Applications*, vol. 27, no. 2, 2007.


**Lei Shi** is a professor in the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences. He holds B.S., M.S. and Ph.D. degrees from the Department of Computer Science and Technology, Tsinghua University. His research interests span Information Visualization, Visual Analytics and Data Mining. He has published 70 papers in refereed conferences and journals. He serves program committee member on several international conferences, a note paper chair for PacificVis'18, and was the recipient of VAST Challenge Award in 2010 and 2012.

**Tao Jiang** is currently a graduate student in University of Chinese Academy of Sciences, and the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences. He received his B.S. degree from the School of Software, Beijing Institute of Technology. His research interests include Computer Graphics, Data Visualization and Visual Analytics.

**Ye Zhao** is a professor in the Department of Computer Science at the Kent State University, Ohio, USA. He has co-authored more than 50 refereed technical papers and served in many program committees including IEEE SciVis and VAST conferences. His current research projects include visual analytics of urban transportation data, multidimensional, text, and animated information visualization, medical image processing, and computational hemodynamics modeling.

**Zhichun Guo** is an undergraduate student in Fudan University, majoring in computer science. Her research interests include data mining, social network analysis, and visual analytics. She is a recipient of Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment.

**Xiatian Zhang** is the chief data scientist of TalkingData. He has long engaged in machine learning research and has dozens of research papers in publication. He also has much of experience in the applications of machine learning, such as recommender systems and computing advertising. He used to work for IBM Research - China, Tencent data platform and Huawei Noah's Ark Lab.

**Yao Lu** is a data scientist in TalkingData Ltd co., which is the largest third party mobile data service platform in China. Her research interest lies in the coverage of artificial intelligence, social computing and urban computing. Before joining in TalkingData, Yao worked as an algorithm expert in Alibaba on recommender system and social network. Yao graduated from Department of Automation in Tsinghua University and also worked as a research visit in EPFL.