Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment

Fengli Xu, Yong Li, *Senior Member, IEEE*, Huandong Wang, Pengyu Zhang, and Depeng Jin, *Member, IEEE*

***Abstract*— Understanding mobile traffic patterns of large scale cellular towers in urban environment is extremely valuable for Internet service providers, mobile users, and government managers of modern metropolis. This paper aims at extracting and modeling the traffic patterns of large scale towers deployed in a metropolitan city. To achieve this goal, we need to address several challenges, including lack of appropriate tools for processing large scale traffic measurement data, unknown traffic patterns, as well as handling complicated factors of urban ecology and human behaviors that affect traffic patterns. Our core contribution is a powerful model which combines three dimensional information (time, locations of towers, and traffic frequency spectrum) to extract and model the traffic patterns of thousands of cellular towers. Our empirical analysis reveals the following important observations. First, only five basic time-domain traffic patterns exist among the 9600 cellular towers. Second, each of the extracted traffic pattern maps to one type of geographical locations related to urban ecology, including residential area, business district, transport, entertainment, and comprehensive area. Third, our frequency-domain traffic spectrum analysis suggests that the traffic of any tower among 9600 can be constructed using a linear combination of four primary components corresponding to human activity behaviors. We believe that the proposed traffic patterns extraction and modeling methodology, combined with the empirical analysis on the mobile traffic, pave the way toward a deep understanding of the traffic patterns of large scale cellular towers in modern metropolis.**

***Index Terms*— Mobile data traffic, measurement study, traffic patterns, clustering, geographical location.**

I. Introduction

THE past few years have seen a dramatic growth in cellular network traffic, contributed by billions of mobile devices as the first-class citizens of the Internet. The global cellular network traffic from mobile devices is expected to surpass 24 exabytes (10^{18}) per month by 2019 [1], 9× larger than the traffic served by existing cellular network. While we are embracing a world with ambient cellular connectivity, however, we are facing a critical and challenging problem — we have limited understanding about the patterns of traffic experienced by cellular towers deployed in urban areas, especially when 3G and LTE networks are widely available in current modern metropolis [1]–[3]. We do not completely understand how urban functional regions and ecologies, such as business district, affect the mobile traffic of cellular towers [2]. In addition, the dominant factors that affect their traffic variations are still unknown. Such limited knowledge significantly increases the cost of operating thousands of cellular towers in big cities.

Despite of the aforementioned lack of knowledge, understanding the traffic patterns of cellular towers in the large scale urban environment is extremely valuable for Internet service providers (ISP), mobile users, and government managers of modern cites [4]–[6]. If we can *identify and model* the patterns of cellular towers, instead of using the same strategy to provide services, such as using the same load balancing and data pricing algorithms on each tower, an ISP can exploit the modeled traffic patterns and customize the strategies for individual cellular towers. For example, an ISP can potentially have different pricing on individual cellular tower based on the traffic it experiences. In addition, mobile users will benefit from the traffic modeling as well because they can choose towers with predicted lower traffic and enjoy better services. Surprisingly, management departments of government will benefit from the traffic modeling as well because they may infer the land usage and human economy activities by looking at the patterns of cellular traffic [7].

On the other hand, understanding the traffic patterns of cellular towers is challenging for three reasons. First, the traffic experienced by thousands of cellular towers deployed in large scale modern cities is complicated and hard to analyze. For example, our dataset includes 9,600 cellular towers and 150,000 subscribers, where lots of redundant and conflict logs are observed. To identify traffic patterns embedded in the thousands of towers, we need to design a system that is able to clean and handle the data of large scale cellular traffic. Second, we do not have the priori about the existence of patterns that can be used for representing the behavior of thousands of cellular towers. To make matters worse, even if such patterns exist, we do not know their profiles. Without these profiles, it is challenging to group thousands of cellular towers into a small number of patterns. Third, the traffic of a cellular tower is affected by many factors, such as time and locations, etc. These factors, sometimes, compound with

Manuscript received November 25, 2015; accepted September 28, 2016; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor S. Date of publication November 29, 2016; date of current version April 14, 2017. Sen. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2013CB329105; in part by the National Nature Science Foundation of China under Grant 61301080, Grant 91338203, Grant 91338102, and Grant 61321061; and in part by the Research Fund of Tsinghua University under Grant 20161080099. A conference version of this paper was published in ACM IMC 2015. *(Corresponding author: Yong Li.)*

F. Xu, Y. Li, H. Wang, and D. Jin are with the State Key Laboratory on Microwave and Digital Communications and the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: liyong07@tsinghua.edu.cn).

P. Zhang is with the Department of Electrical Engineering and Computer Science, Stanford University, Stanford, CA 94305, USA.

Digital Object Identifier 10.1109/TNET.2016.2623950

each other and further complicate our analysis. For example, significant traffic variation is observed at both fine-grained (hours) and coarse-grained (days) time scale, and across towers deployed in different locations [5], [8]. By addressing these challenges, in this paper, we investigate how to extract and model the mobile traffic patterns of thousands of cellular towers in a large scale urban environment via credible dataset collected by one of the largest commercial mobile operators.

Our core contribution is a powerful model which combines three dimensional information, including time, locations of towers, and traffic frequency spectrum, for extracting and modeling the traffic patterns of thousands of cellular towers. A breakdown of the core contribution comprises three parts. First, we design a system which leverages machine learning to identify and extract five patterns from the traffic of thousands of cellular towers. Our system is built with processing large scale data in mind and is able to process the traffic of thousands of towers with granularity of 10 minutes. Second, we identify the geographical context of traffic experienced by cellular towers by investigating the correlation between time-domain traffic characteristics and geographical locations of towers. Therefore, by looking at the traffic pattern of a tower, we can infer the type of location where it is deployed and the type of users it serves. Third, our frequency-domain traffic spectrum analysis reveals that any traffic of the 9,600 cellular towers can be constructed using a linear combination of four primary components corresponding to human activity behaviors. This observation provides an unique angle (frequency) for analyzing cellular traffic and significantly simplifies the process of analysis by a linear model.

Through investigating the traffic of 9,600 cellular towers, we find following interesting observations. First, the 9,600 cellular towers can be classified into five groups using features extracted from time-domain traffic. This experimental result confirms our motivation that a small number of patterns do exist among thousands of cellular towers. Second, each of the traffic pattern maps to one type of geographical locations, including resident, office, transport, entertainment, and comprehensive area. Therefore, the traffic pattern of a cellular tower does suggest the urban ecology and geographical location context where it is deployed as well as the type of users it serves. Third, our frequency-domain analysis reveals that the transition between the five traffic patterns encodes the mobility of human. For example, when the phase of residential pattern moves toward the phase of transport pattern, people start their commute from home to work. In summary, we believe that the proposed traffic patterns extraction and modeling, combined with the empirical study on large scale cellular towers, pave the way toward a deep understanding of the traffic patterns of large scale cellular towers.

This paper is structured as follows. In Section 2, we provide details about the utilized dataset, and present some basic observations of traffic spatio-temporal distributions. In Section 3, we design our traffic processing system and identify the key traffic patterns of the large scale cellular towers. Based on the discovered five traffic patterns, in Section 4 and 5, we conduct a deep analysis and reveal the correlation among data traffic, urban ecology and human behaviors in the time and frequency domain respectively. After discussing related work in Section 6, we summarize our discoveries and discuss potential investigations in Section 7.

II. Dataset and Visualization

In this section, we provide details about the dataset we investigate as well as the needed preprocessing. In addition, we visualize the spatial-temporal distribution of cellular traffic.

A. Dataset Description

The dataset is an anonymized cellular trace collected by an ISP from Shanghai, a big city in China, between Aug 1st and Aug 31st 2014. Each entry of the trace contains detailed mobile data usage of 150,000 users, including the ID of devices (anonymized), start-end time of data connection, base station ID, address of base station, and the amount of 3G or LTE data used in each connection. The trace logs 1.96 billion tuples of the described information, contributed by approximately 9,600 base stations all over Shanghai. The trace contains 2.4 petabytes (10^{15}) logs, 77 terabytes (10^{12}) per day and 8 gigabytes (10^{9}) per base station on average. This large scale and fine-grained dataset guarantees the credibility of our traffic pattern analysis and modeling.

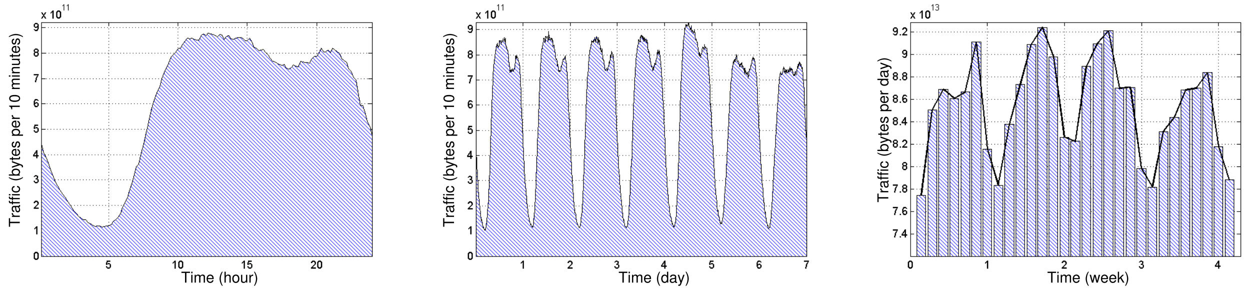
B. Preprocessing

The trace collected by the ISP needs to be preprocessed because of the existence of redundant and conflict traffic logs as well as the incomplete information of base stations' locations. The preprocessing includes three steps. First, we eliminate the redundant and conflict logs, such as the identical traffic logs, introduced by technical issues. Second, to solve the problem of incomplete information, we convert the addresses of base stations to their geographical longitudes and latitudes through APIs provided by Baidu Map, the most popular online map service provider in China. This conversion gives us the precise geographical location of a base station, which is important for analyzing the ground truth of urban functional regions. The last step of preprocessing is computing the traffic density (byte/km^2) across the city. The obtained traffic density allows us to understand the spatial distribution of cellular traffic.

C. Data Visualization

Before diving into a deep analysis of mobile data traffic, we first visualize the spatial-temporal traffic distribution of the 9,600 base stations, where we find two interesting observations.

First, the data embeds fundamental temporal patterns of mobile data traffic. Figure 1 shows the aggregated traffic of the 9,600 towers at different time scales. Figure 1(a) shows the traffic distribution of a day (Aug 7th 2014, Thursday) where we observe that the aggregated network traffic is tightly coupled with the sleep pattern of humans. High cellular traffic is observed during the day and low traffic is experienced during midnight. There are two traffic peaks in each day: one

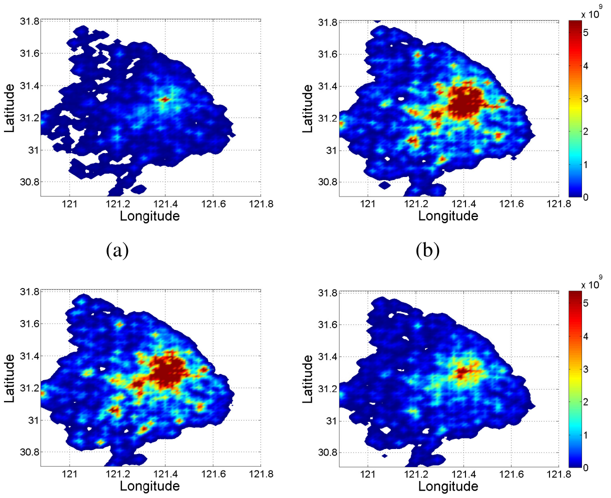


(a)

(b)

(c)

Fig. 1. The temporal distribution of cellular traffic at different time scales. (a) Hourly. (b) Daily. (c) Weekly.



(a)

(b)

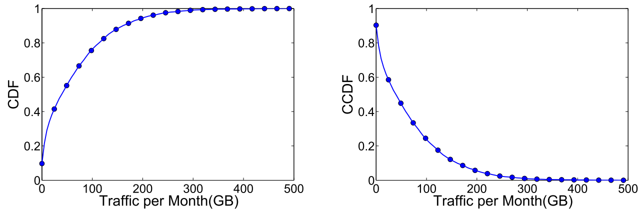
(c)

(d)

Fig. 2. The spatial distribution of cellular traffic at different time. (a) 4AM. (b) 10AM. (c) 4PM. (d) 10PM.

around 12PM and the other around 10PM. Similar patterns are observed in Figure 1(b). The timing of the two peaks suggests that most people tend to consume data traffic heavily after lunch and before sleep. Figure 1(b) shows the traffic distribution of a week (from Aug 4 to Aug 10 2014) and Figure 1(c) shows the traffic distribution of a month (from Aug 3 to Aug 31 2014). Both figures show that the traffic exhibits a periodical pattern on the scale of a week, where weekend's traffic is less than weekday's traffic. Such traffic variation comes from people's weekly work schedule.

On the other hand, our trace also records the spatial distribution of mobile data traffic. Surprisingly, we find that the spatial and temporal characteristics of traffic are correlated. Figure 2 shows the geographical traffic density (bytes transmitted per hour per km^2) at 4AM, 10AM, 4PM and 10PM. As shown in the color bar, the red one indicates higher traffic and the blue one stands for lower traffic. We find the following observations. First, towers deployed at the center of the city experience high traffic despite of the time of a day. Second, at 4AM, most areas of the city are covered by dark color, which suggests that traffic demand is small because of human sleep. In contrast, at 10AM, most areas of the city are covered by light color, suggesting that traffic demand becomes high because people start working. Therefore, the areas of peak traffic map to areas occupied by human, such as residential housing or central business district (CBD). Third, the traffic demand of different area varies significantly. In Figure 3, we present the cumulative distribution function(CDF) and complementary cumulative distribution function(CCDF) of the mobile traffic on cellular towers. We can observe that more than 70% of cellular towers have a traffic demand lower than 100GB per month. However the distribution also exhibit a "long tail", which suggests significant differences on traffic demand of different cellular towers.



(a)

(b)

Fig. 3. The (a) CDF and (b) CCDF of cellular traffic.

III. Identifying Traffic Patterns of Cellular Towers

Now, we investigate the data traffic of the thousands of 3G/LTE cellular towers and design a system that is able to identify key traffic patterns of large scale cellular towers. We start from understanding the traffic patterns of a few cellular towers to motivate our study.

A. Motivation and Problem Statement

Our cellular network traffic measurement and analysis are motivated by a key observation — the traffic pattern of one cellular tower is vastly different from another. Through online map service, we randomly select four towers from the positions of residential areas and four towers from business districts, and plot their normalized traffic profile in the left and right column of Figure 4, respectively. We can clearly observe the difference of traffic between these two types of cellular towers, where the traffic profiles of residential towers have two peaks within a day and remain high across night, while

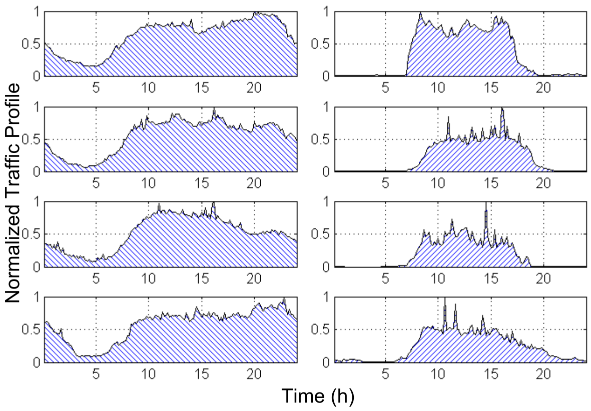
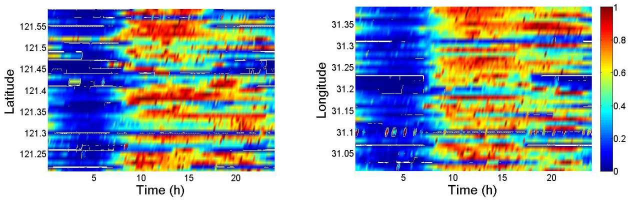


Fig. 4. Cellular traffic experienced by base stations deployed in the residential area and business district.

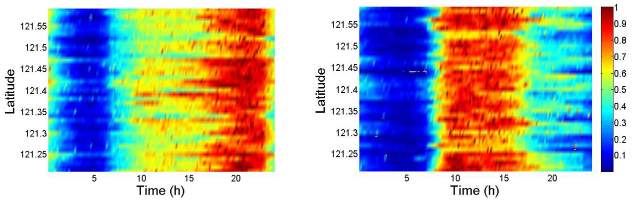


(a) (b)

Fig. 5. Cellular traffic experienced by base station randomly selected from different latitudes and longitudes. Large traffic variations are observed. (a) Latitudes. (b) Longitudes.

the traffic profiles of towers in business district experience only one peak within a day and get close to zero across night. This comparison clearly reveals the difference of traffic patterns between the two specific types of cellular towers. However, from the perspective of an ISP, which manages thousands of cellular towers, is the traffic pattern of one cellular tower vastly different from another? To understand this problem, we conduct a large scale measurement and investigate the recorded 9600 cellular towers in our dataset of Shanghai.

Figure 5 shows the normalized traffic variations within one day with 40 randomly selected cellular towers for each 0.01 degree latitudes or longitudes respectively. The x-axis shows the time in hours and y-axis shows the logical positions of the selected cellular towers in terms of latitude (a) or longitude (b). For example, the first row of pixels in Figure 5(a) represent the traffic variations of one cellular towers, of which the latitude is around 121.60 and the longitude is randomly selected. Traffic measured on each cellular tower is normalized by its maximum, and the color presents the normalized value where red color indicates higher traffic and blue color stands for lower traffic as shown in the color bar. In these measurements, we find two observations. First, the peak hour of one cellular tower, which is marked as red, is vastly different from another during the day time when serving mobile users. In fact, the variance of the peak among the selected towers is about 10 hours. Second, while most of towers experience low traffic in early morning, the first several towers in Figure 5(a), of which the latitude is around 121.60, also have low traffic during evening. Therefore, significant differences of data traffic are observed across cellular towers. Such differences cause troubles for an ISP to manage its cellular network. For example, because of the unique pattern of individual traffic, an ISP cannot obtain the optimal performance by using the same load balancing strategy, which is built on top of traffic patterns, on different towers. Therefore, a natural question to ask is that is it possible to model the traffic pattern of thousands of cellular towers? More specifically, can we utilize a few simple patterns to present the traffic of thousands of cellular towers? Identifying these patterns of cellular towers would give an ISP significant benefits on network management, including load balancing, pricing, etc.



(a) (b)

Fig. 6. Cellular traffic experienced by base stations selected from residential and business district. (a) Residential Area base stations. (b) Business District base stations.

Our investigation suggests that at least two, maybe more, traffic patterns exist among thousands of cellular towers. Figure 6 shows the normalized traffic profile of 40 selected cellular towers deployed in residential area and in business district for each 0.01 degree latitudes. Compared with the disorder in both temporal and spatial dimension exhibited in Figure 5, traffic variations for cellular towers in a single kind of regions are more regular and similar to each other. In addition, we find other two observations in this investigation. First, in terms of the traffic of residential area, all residential towers experience similar traffic patterns where the peak traffic is present around 9PM. In addition, only a small amount of traffic is observed between 8AM and 4PM because most users leave home for work. Similar conclusion can be drawn for towers deployed in business district. Second, the traffic pattern of residential towers is different from towers deployed in business district where peak hour appears around 1PM. Inspired by these two observations, we conclude that traffic patterns do exist among thousands of cellular towers. One key question addressed by this paper is finding out how many traffic patterns exist among thousands of cellular towers and how to identify them.

B. Identifying Traffic Patterns of Cell-Towers

Investigating traffic patterns among thousands of cellular towers is extremely challenging for three reasons. First, we have little prior knowledge about the data traffic, and do not know which cell towers may share the same traffic pattern and how the pattern may look like. Second, the measured cellular traffic data is huge in terms of tracing 9,600 cellular towers for a month. To make matters worse, the measured data is not clean in terms of unstructured logs. Last but not least, the

Algorithm 1: Traffic Patterns Identifier

```
Input: Cell towers number $M$, Threshold value $T$,
Traffic vector $X_i$, for $i = 1, 2, 3...M$
Output: Cluster labels of tower $i$, $L_i$, for $i = 1, 2, 3...M$
Initialize:
    Clusters: $c_k \leftarrow [X_k]$, for $k = 1, 2, 3...M$,
    Cluster set: $C \leftarrow [c_1, c_2...c_M]$
    Cluster number: $N \leftarrow M$
    Distance matrix: $D \leftarrow Inf$
    Stop index: $stop \leftarrow false$
while $stop == false$ do
    $D \leftarrow Inf$
    for $\forall c_i, c_j \in C, i \neq j$ do
        $D_{i,j} \leftarrow compute\_distance(c_i, c_j)$
    $[Mindistance, index1, index2] \leftarrow find\_min(D)$
    if $Mindistance > T$ then
        $stop \leftarrow true$
        $break$;
    $merge(c_{index1}, c_{index2})$
    $N \leftarrow N - 1$
for $i = 1$ to $N$ do
    for $\forall X_k \in c_i$ do
        $L_k \leftarrow i$
Return $L$
```

measured cellular traffic data is noisy where large variation of traffic is observed because the absolute traffic depends on the number of mobile users served. All these factors make the analysis of cellular traffic patterns extremely challenging. To tackle these challenges, we design, implement, and evaluate a system which is able to identify the key traffic patterns of such large scale cellular towers. Our system is composed by three key elements: traffic vectorizer, pattern identifier and metric tuner.

Traffic Vectorizer: We implement a traffic vectorizer on Hadoop platform to convert the large scale unstructured traffic logs into traffic usage vectors. The key of designing the traffic vectorizer is a parallel transformer, which takes the time-domain traffic logs of thousands of cellular towers as its input and converts each cell tower's logs into a time-domain traffic vector. The vector is constructed in two phases — aggregation and normalization. In the first phase, each cellular tower's traffic logs are segmented into thousands of chunks, with each chunk contains 10-minutes traffic logs. Then we aggregate the traffic logs in each chunk and generate a traffic usage vector. In the second phase, since we aim to identify the similar traffic patterns without the interference of different amplitude, we perform zero-score normalization on each vector to eliminate their differences in amplitude, while the difference of traffic amplitude is analyzed after the key patterns are identified. We define the traffic vector of cellular tower j as $X_j = (x_j[1], ..., x_j[N])^T$, with $x_j[i]$ stands for the normalized traffic amount in the i_{th} 10-minute time slot. We remove 3 days from the month to make the duration consist of four entire weeks. Thus, N is number of 28 days' 10-minutes segmentation, i.e., 4032 in our analysis.

Pattern Identifier: Pattern identifier takes the vectorized data from the vectorizer and runs an unsupervised machine learning algorithm for identifying the key patterns of cellular tower traffic. The pattern identifier addresses one key challenge of the mining process — unknown patterns, by exploiting hierarchical clustering [9]. The algorithm of our system is shown in Algorithm 1. We first considers each input point as a cluster and then bottom-up iteratively merges the nearest two clusters until the stop condition is met. In the clustering, we use the euclidean distance as the distance metric and define the distance between clusters as average-linkage distance. In each iteration, we compute the distances between each pair of clusters, and find out the pair of clusters corresponding to the minimum distance and merge them into a new cluster. In addition, we set a threshold value as stop condition, which stops the clustering when the minimum distance between two clusters, $Mindistance$ is above the threshold value.

Metric Tuner: As the number of traffic patterns is unknown, a key question is when the identifier should stop its clustering. In our system, we use Davies-Bouldin index [10] to explicitly inform the identifier that the optimum number of patterns have been identified. Davies-Bouldin index is utilized because it measures both the separation of clusters and cohesion within clusters, which mathematically guarantees good clustering result. The mathematic formulation of Davies-Bouldin index is as follows,

$$\text{minimize} \quad \frac{1}{R}\sum_{i=1}^{R} \max_{j=1, j\neq i}^{R} \frac{S_i + S_j}{M_{i,j}},$$

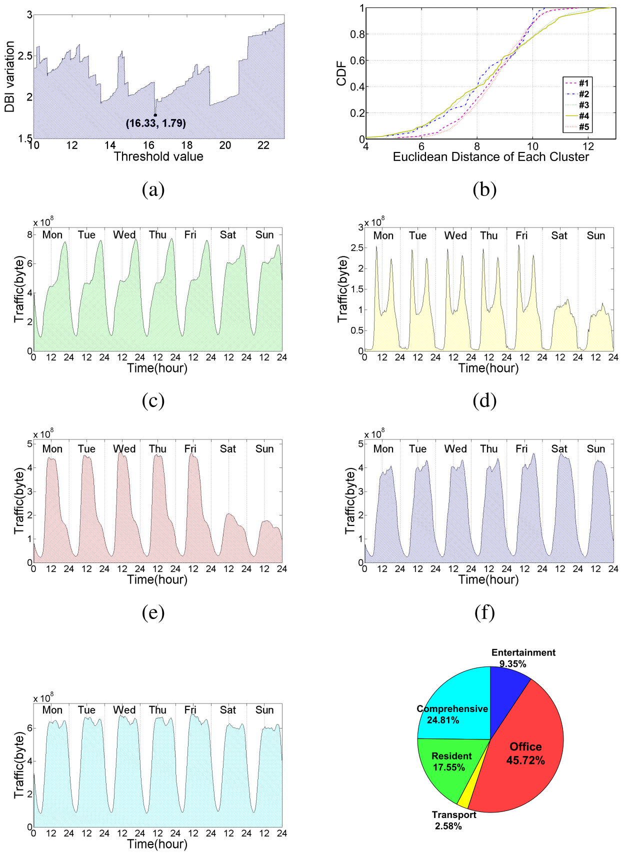
$$\text{subject to}$$

$$M_{i,j} = ||A_i - A_j||_2,$$

$$S_i = \frac{1}{T_i}\sum_{k=1}^{T_i} ||X_k - A_i||_2,$$

where the objective function is the Davies-Bouldin index, X_i is the vectorized data of cellular tower i, A_i is the centroid of each cluster, R is the number of clusters and T_i is the numbers of towers within the i_{th} cluster. We minimize the Davies-Bouldin index by considering two factors — the distance between clusters $M_{i,j}$ and S_i, which are the average distance from points to their cluster's centroid. When the minimum Davies-Bouldin index is obtained, the optimum number of patterns is identified. The variation of DBI is shown in Figure 7(a), according to which we set the stop condition—threshold value at 16.33 to achieve optimal clustering result.

Figure 7 shows the five time-domain patterns identified by our system from the 9,600 cellular towers((c) to (g)) and each cluster's CDF of points' distance to its centroid(b). The five clusters differ in terms of the time where peak traffic appears as well as the amount of traffic experienced during weekday and weekends. Figure 7(b) shows that the distance CDF curves of clusters are similar and all of them increase rapidly as distance increases. 80% of points' distance to their clusters' centroid are less than 10, which implicates the clustering result is good. The percentage of each cluster's cell towers is shown in Table I, which indicates the third cluster has most cell towers



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 7. Patterns of the five identified clusters and CDF of clustering distance. (a) DBI variation. (b) CDF of distance. (c) #1:Resident area. (d) #2:Transport area. (e) #3:Office area. (f) #4:Entertainment area. (g) #5:Comprehensive area. (h) Proportion of each cluster.

TABLE I
PERCENTAGE OF CELL TOWERS CLASSIFIED IN EACH CLUSTER

Cluster Index	Functional Regions	Percentage
1	Resident	17.55%
2	Transport	2.58%
3	Office	45.72%
4	Entertainment	9.35%
5	Comprehensive	24.81%

and second cluster the least. We also present the proportion of each cluster's cell towers explicitly in Figure 7(h).

In conclusion, we implement a system that is able to identify the key traffic patterns among thousands of cellular towers in this subsection. Since the five clusters are given by the hierarchical classifier, an interesting question to ask is what are the geographical locations where these five types of towers are deployed?

C. Geographical Context of Traffic Patterns

To understand the geographical locations of cell towers of the five clusters, we first manually label typical towers in the five patterns with urban functional regions and then validate the labels of all towers in each pattern with ground truth.

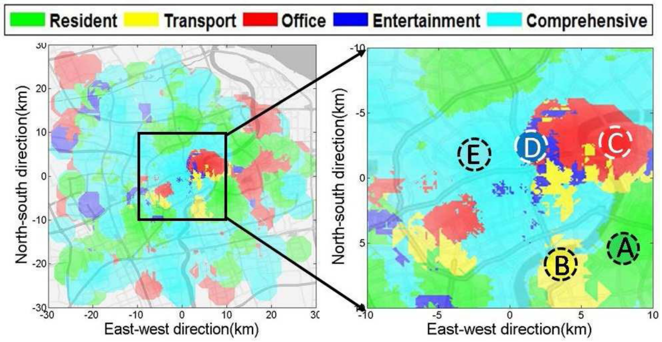


Fig. 8. Geographical distribution of base stations from the five identified patterns.

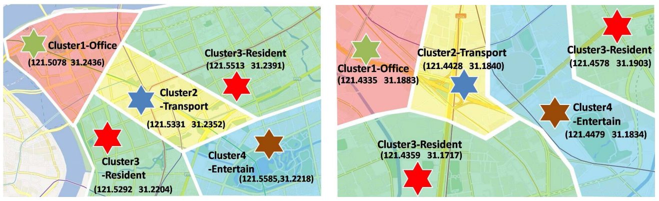
TABLE II
DISTRIBUTION OF POI AT CHOSEN POINT

Point	Points of Interest			
	Resident	Transport	Office	Entertain
A	195	0	19	51
B	68	2	56	36
C	151	1	1016	157
D	16	0	108	2165
E	59	0	179	26

1) Label Patterns With Urban Functional Regions: To understand the geographical context of traffic patterns, we label the five traffic patterns using urban functional regions. This process is nontrivial because given thousands of cellular towers, labelling cannot be done one by one manually. To address this challenge, we use a few human-labeled areas and combine with points of interests (POI) distribution to achieve accurate labelling. POI is a specific point location of a certain function such as restaurant and shopping mall. An area's POI distribution reflects its function. Therefore, studying POI distribution of one location can help us to accurately identify patterns' labels. The POI data we study is collected via APIs provided by Baidu Map introduced before. For calculating the POI distribution, we measure the number of four main types of POI, which are resident, transport, office and entertainment, within 200m of each cell towers. Figure 8 shows the geographical density map of towers in each cluster where deep color stands for higher density. Zooming in the urban area, for each cluster we pick the point with the highest tower density and calculate their POI distribution as summarized in Table II. Then, we infer the urban function region of each cluster by checking the geographical location information in Figure 8 and POI distribution in Table II. We obtain the following geographical labels for the five clusters.

Resident Area: Figure 8 shows that the towers in this cluster (green color) are mainly distributed on the surrounding areas of the city. In addition, the highest density point, ***A***, is located in a large resident neighborhood. Table II also shows that the number of residential points in ***A*** is more than others. Therefore,we label the area covered by this cluster's cell towers as residential area.

Transport Area: In Figure 8, the second cluster's highest density point ***B*** is close to an area with three subway stations



(a) (b)

Fig. 9. Two case studies for validating the geographical context of the five identified patterns. (a) Area A. (b) Area B.

and one overpass. In addition, Table II shows that around location ***B*** the number of transport POI is higher than the rest even though its absolute number is small. Therefore, we label this cluster as transport area.

Office Area: Figure 8 shows that the highest density point ***C*** is a well-known business district in Shanghai. This location mark is also verified by the third row of Table II where the number of office POI is dominant for the area 200m from ***C***. As a result, we label this cluster as office area.

Entertainment Area: The highest density point ***D*** in Figure 8 is a large shopping mall and entertainment park in Shanghai. Table II also shows that its number of entertainment POI is more than the rest. Therefore, we label this cluster as entertainment area.

Comprehensive Area: Figure 8 shows the tower density map of the last cluster, where we observe uniform distribution of towers across the city. In addition, the highest density point, ***E***, is a comprehensive area, which includes all kinds of urban functions, including residential area, offices, etc. However, there is no obvious dominant POI type. Therefore, it is labeled as comprehensive area.

2) Validate the Labels: In this section, we validate the labels of the five patterns in both micro and macro scale. Our labels are obtained by checking the geographical locations of a few towers in each cluster and verifying with the corresponding POI distribution. However, the correctness of labelling across all 9,600 cellular towers remains unknown. Therefore, we perform further analysis to validate our labels with POI data from micro and macro two perspectives.

Validate With Case Study: To validate our labels in micro scale, we randomly choose two areas shown in Figure 9. According to the POI data, we first color different functional regions in the area with different colors. Green represents residential area, yellow represents transport area, red represents office area, and blue represents entertainment area. After that, we investigate the labels of cell towers locating in the area. Observing both Figure 9(a) and (b), we find that the labels attached to the cell towers exactly match with the functional regions, which justifies our labels' correctness.

Validate With 9,600 Towers' POI: To validate our labels in macro scale, we perform further analysis on all 9,600 towers' POI. However, different types POI vary in magnitude significantly because of their different nature. To eliminate this interference, we first perform min-max normalization on each type's POI and then average them by clusters, which is summarized in Table III. The maximum of each row and column is marked with color, which shows the dominant urban function in each cluster. Figure 10 explicitly shows each POI's percentage in five clusters. According to Table III and Figure 10, transport type POI dominates the region labeled as transport area, with 44% of this area's POI, while entertainment area is dominated by entertainment type POI for 39%. These measurements validate the labels obtained from the sampled towers of each cluster.

TABLE III
AVERAGED NORMALIZED POINTS OF INTEREST OF FIVE CLUSTERS

Cluster	Points of Interest			
	Resident	Transport	Office	Entertain
#1	0.0528	0.0285	0.0232	0.0269
#2	0.0473	0.2000	0.1012	0.1020
#3	0.0439	0.0813	0.1034	0.0515
#4	0.0474	0.1201	0.0976	0.1674
#5	0.0508	0.0373	0.0453	0.0403

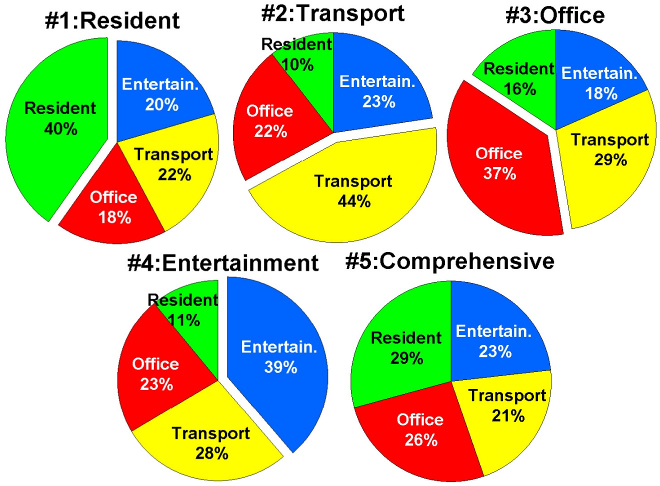


Fig. 10. Pie chart of averaged normalized points of interest of five clusters.

To conclude, in this subsection we verify our identified key traffic patterns as well as establish their relationships with urban functional regions.

IV. UNDERSTANDING MODELED TRAFFIC PATTERNS: TIME DOMAIN ASPECT

Understanding the hidden physical meanings of traffic patterns is important for exploiting them to solve practical problems, such as traffic load balancing or land usage identification. Although we have identified key traffic patterns and linked them to corresponding urban functional regions, we still have little knowledge of the hidden physical meaning of these patterns. In this section, we conduct an analysis to reveal the time and geographical characteristics of modeled traffic patterns.

A. Quantify Time-Domain Characteristics

It is obvious that traffic patterns of different urban functional regions possess different characteristics in time-domain. In this subsection, we dedicate to quantify these characteristics

TABLE IV

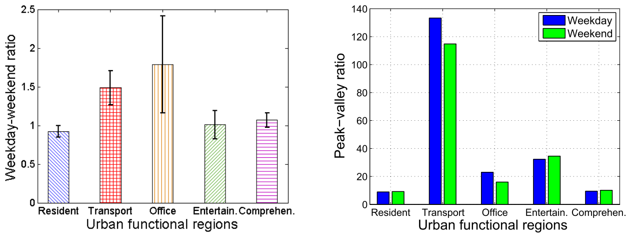
PEAK-VALLEY FEATURES

Features Regions	resident area		transport area		office area		entertainment area		comprehensive area	
	weekday	weekend	weekday	weekend	weekday	weekend	weekday	weekend	weekday	weekend
maximum traffic	7.77×10^8	7.99×10^8	2.76×10^8	1.55×10^8	4.69×10^8	2.78×10^8	4.55×10^8	4.90×10^8	7.36×10^8	7.38×10^8
minimum traffic	8.70×10^7	8.71×10^7	2.07×10^6	1.35×10^6	2.04×10^7	1.74×10^7	1.41×10^7	1.42×10^7	7.77×10^7	7.29×10^7
peak-valley ratio	8.93	9.17	133.33	114.81	22.99	15.98	32.27	34.51	9.47	10.12

TABLE V

TIME OF TRAFFIC PEAK AND VALLEY

Features Regions	resident area		transport area		office area		entertainment area		comprehensive area	
	weekday	weekend	weekday	weekend	weekday	weekend	weekday	weekend	weekday	weekend
time of peak	21:30	21:30	8:00 18:00		10:30	12:00	18:00	12:30		
time of valley	5:00	5:00	4:00	4:30	5:00	5:00	5:00	5:0	5:00	5:00



(a) (b)

Fig. 11. Time-domain characteristics of the five identified patterns. (a) Weekday-weekend traffic amount ratio. (b) Weekday and weekend's peak-valley ratio.

and provide insights of traffic behaviours in different urban functional regions.

Weekday-Weekend Traffic Amount Ratio: Observing Figure 7, traffic amount during weekday is significantly different from weekend in transport area and office area. We quantify this characteristic by computing the ratio between weekday's traffic amount and weekend's, which is presented in Figure 11(a). According to Figure 11(a), one day's traffic amount in resident area, entertainment area and comprehensive area is almost identical between weekday and weekend. However, weekday-weekend traffic amount ratio in transport area is 1.49 and the ratio in office area is 1.79, which suggests weekday's traffic amount of those two regions is much more than weekend. This phenomenon makes sense because people typically go to work in weekday while they do not in weekend.

Peak-Valley Features: Observing Figure 7, all traffic patterns experience periodic peaks and valleys. However, the traffic patterns are significantly different in peak value, valley value and peak-valley ratio. We quantify these characteristics and summarize them in Table IV. According to Table IV, in transport area and office area weekend's maximum traffic and minimum traffic is much less than weekday, which is consistent with last paragraph's finding. What's more, the transport's peak-valley ratio is much higher than other regions, which is explicitly presented in Figure 11(b). However, transport area's maximum traffic is less than other regions both in weekday and weekend. It suggests that transport area has the least traffic amount and the largest peak-valley traffic difference, while resident area and comprehensive area are the opposite.

Time of Traffic Peak and Valley: Different urban functional regions' traffic patterns differ not only in peak volume, but also in peak time. We quantify this characteristic and present it in Table V. We leave the blank unfilled, if there is not a periodic peak or valley. Observing Table V, we find that traffic valley always takes place in 4:00∼5:00. In weekday, transport area has two peaks in 8:00 and 18:00, which are probably caused by rush hour. In entertainment area, weekday's traffic peak time is 18:00 while weekend's traffic peak time is 12:30. It suggests that people go for entertainment later in weekday because of work.

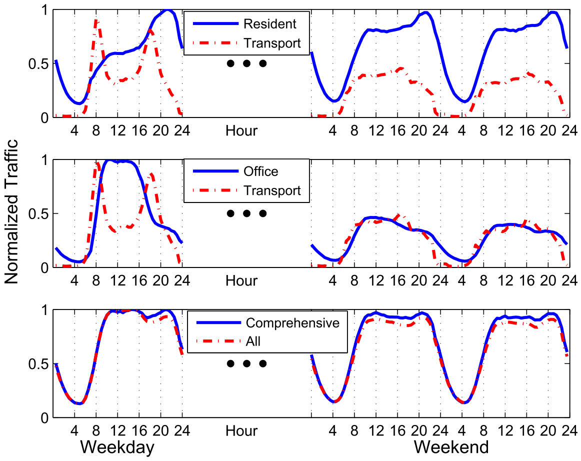
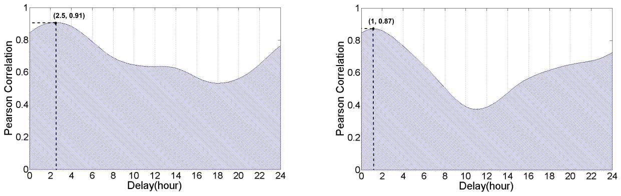


Fig. 12. Understanding the interrelationships between traffic patterns.

To conclude, we quantify the time-domain characteristics of each identified traffic pattern, which paves the way towards a deep understanding of cellular traffic patterns.

B. Interrelationships Between Traffic Patterns

We compare the interrelationships between normalized modeled traffic patterns in Figure 12. The first row of Figure 12 compares the modeled traffic patterns of residential areas and transport hot spots. The peak of residential area is about 3 hours later than the second peak of transport, and the slope of these two peaks is almost identical. In addition, when we compare traffic patterns of transport hot spots and business district shown in Figure 12, we find that the peak in business district takes place in the time period between the two peaks of transport hot spots. In order to better quantify the interrelationships, we compute the correlation



(a) (b)

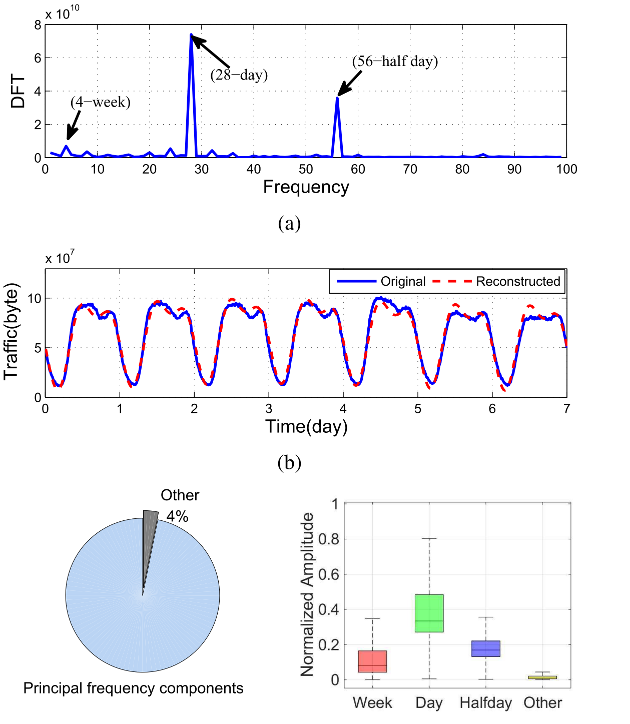
Fig. 13. The correlation between traffic patterns in different area. (a) Resident area and transport area. (b) Office area and transport area.

of these area's traffic with variant delay and present it in Figure 13. From the result, we can observe that the correlation between these area are high with the maximum value up to 0.91, which means the traffic in these area is strongly related. In addition, the correlation between resident area and transport area reaches its peak with delay of 2.5 hours, and the correlation between office area and transport area reach its peak with delay of 1 hour. These observations suggest that these three traffic patterns probably depict the daily routine of working populations, for them rush through heavy traffic area to work in morning and rush back home in evening.

In the third row of Figure 12, blue line stands for the traffic pattern in comprehensive area, and red line stands for the average traffic pattern of all cell towers. In fact, we find that these two patterns are of great similarity, which suggests that comprehensive area really is a mixture of other four kinds of functional areas. In conclusion, we analyze the interrelationships between the traffic patterns of different urban functional regions, which provides insightful understanding.

V. Frequency-Domain Representation for Traffic Modeling

In this section, we conduct frequency-domain analysis. Such frequency-domain analysis is motivated by observing the inherent time-domain periodicity of traffic and the disadvantages of pure time-domain traffic analysis, where time-domain traffic identification is not easy, especially when cellular towers are deployed in the comprehensive areas with couples of behaviors. For example, we know that traffic of cellular tower in the office area reaches the valley in weekends, and traffic of cellular tower in transport area has two peaks in one day, but for an arbitrary cellular tower which has both characteristics, we do not know which of the two will predominate. On the other hand, in frequency domain, we can quantify these characteristics by using the amplitude and phase of frequency corresponding to one day and one week. Thus, we can grasp the key points and compare the strength of different characteristics of traffic for one cellular tower, which is not intuitive in time domain. Here, a natural question to ask is what are the most discriminating and essential features to present traffic patterns of cellular towers. Motivated by answering this question, we conduct frequency domain analysis on the five extracted patterns and reveal several important discoveries.



(a)

(b)

(c) (d)

Fig. 14. Performance of reconstructing time-domain traffic with principal frequency-domain components. (a) Frequency spectrum of aggregated traffic. (b) Time-domain traffic reconstructed by the three principal frequency-domain components (k=4, 28, 56). (c) Energy lost of reconstructing traffic patterns. (d) Boxplot of normalized amplitude of different frequency components.

A. Frequency Transform

In order to analyze the strong periodicity existing in time domain, we first carry out discrete fourier transform (DFT) on the time-domain traffic vector $X = (x[1], ..., x[N])^T$. X can be either the time-domain traffic vector of one cellular tower, $i.e.$, X_j for cellular tower j, or the aggregate traffic vector of a cluster, $i.e.$, $\sum_{j \in C} X_j$ for the cluster C. The process can be formulated as the following:

$$\hat{X}[k] = \sum_{n=1}^{N} x[n] e^{-2\pi ikn/N},$$

where N is the number of traffic samples, that is 28 days' 10-minutes segmentation, $i.e.$, 4032 as discussed before in our analysis. $\hat{X}[k]$ is the frequency spectrum of time-domain traffic X. Figure 14(a) shows the DFT of the aggregate traffic of all cellular towers, where three peaks are observed, $i.e.$ k =4, 28, 56. Since the duration of our series is 4 weeks, the $4th$ point is corresponding to time-domain periodic patterns of one week. Similarly, the $28th$ and $56th$ points stand for the time-domain periodic patterns of one day and half a day, respectively. The absolute values of the three components are much higher than the rest of points, which suggests that most information of the time-domain traffic could be retained by the three components. Motivated by this hint, we use the three components for presenting the time-domain traffic. To evaluate the information loss of ignoring the rest of frequency components, we reconstruct the time-domain traf-

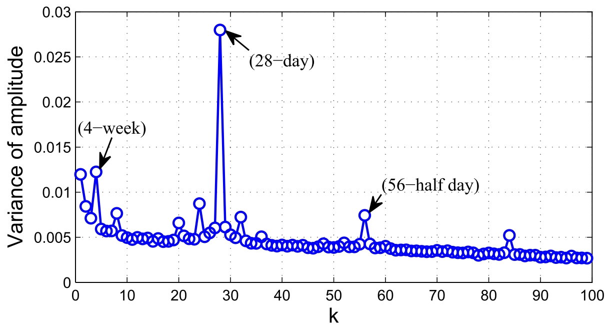


Fig. 15. Variance of the frequency components across the five identified patterns.

fic using the three main frequency components, which is expressed as follows:

$$\begin{cases}\hat{X}^r[k]=\begin{cases}\hat{X}[k], & \text{if } k=0,\ 4,\ 28,\ 56,\ N\text{-}4,\ N\text{-}28,\ N\text{-}56,\\ 0, & \text{otherwise},\end{cases}\\ x^r[n]=\frac{1}{N}\sum_{k=0}^{N-1}\hat{X}^r[k]e^{2\pi ikn/N},\end{cases}$$

where $x^r[n]$ is the reconstructed time-domain traffic. The reconstructed time-domain traffic of the aggregate traffic of all cellular towers is also shown in Figure 14(b). From the result, we can observe that the reconstructed curve is very close to the original curve. Specifically, as shown in Figure 14(c), the lost energy, $\sum_{k=1}^{N-1}\hat{X}[k]^2-\sum_{k=1}^{N-1}\hat{X}^r[k]^2$, is less than 4% relative to the total energy of all frequency components $\sum_{k=1}^{N-1}\hat{X}[k]^2$, which suggests the negligible energy contributed by frequency components beyond the three main components. In addition, the box plot of the normalized amplitude of the three principal frequency components compared with other components for all BSs is shown in Figure 14(d), in which their medians, 25th and 75th percentiles, most extreme points are marked with different lines. We can also observe that compared with other components, amplitude of the principal ones is much larger, indicating their uncomparable importance.

To further understand the capability of signal reconstruction using the three points, we analyze the variance of amplitude of DFT at each frequency component for different cellular tower, and the result is shown in Figure 15. We can observe that the DFT variances of the three frequency components are larger compared to the rest. In addition, we use the DFT to analyse the aggregate traffic for cellular towers of the four primary traffic patterns in Figure 16. We can find that the reconstructed curves are also very close to the original curves, and their DFT spectrum varies most significantly at the three frequency components, which suggests that these three frequencies are the most important components in distinguishing towers of different traffic patterns as well as constructing a time-domain traffic.

B. Visualized Analysis in Frequency Domain

In order to better understand the five traffic patterns of towers in frequency domain, we now provide visualized frequency analysis of them. In addition, based on our earlier observation in Section 5.1, we only analyze the three frequencies corresponding to one week, one day, and half a day. Since each DFT point is a complex number, we analyze the distribution of its amplitude and phase extracted by the following expressions:

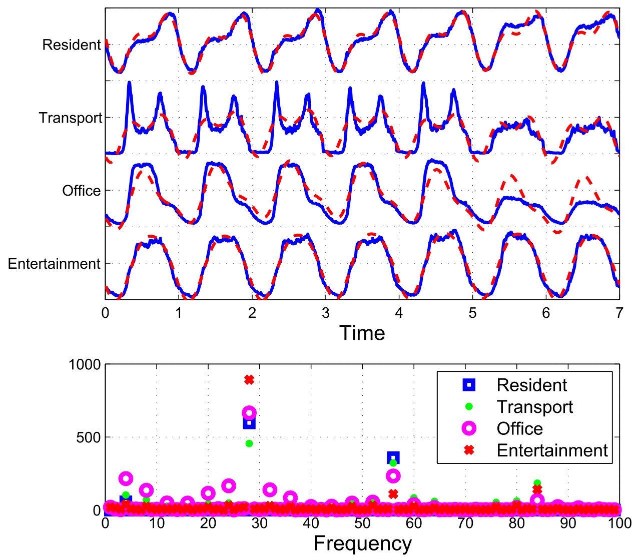


Fig. 16. Reconstructed time-domain traffic of the four patterns using the three principal frequency domain components.

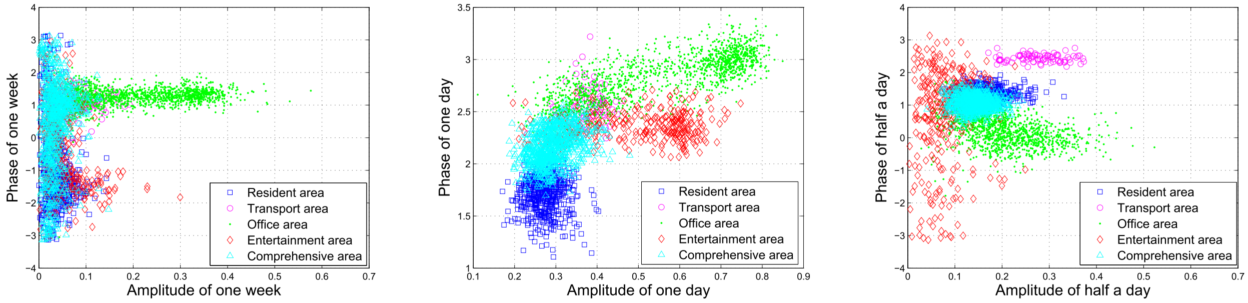
$$\begin{cases}A_k^m=||X^m[k]||,\\ P_k^m=arg\ X^m[k],\end{cases}$$

where A_k^m and P_k^m are the amplitude and phase of DFT for tower m at the k_{th} frequency component. The larger amplitude reflects the stronger periodicity at corresponding frequency, while different phases of DFT indicate different peak time or valley time. Intuitively, for example, larger A_{28}^m indicates the cellular tower m is located at the area that is significantly influenced by the holiday at the weekend, such as office and entertainment area. On the other hand, since the traffic peak at office area tends to be reached at weekdays, while it at entertainment area tends to be reached at weekends, their P_{28}^m will have much difference. Thus, by frequency analysis, we can quantify the inherent time-domain periodicity of traffic, which is difficult to achieve by the time domain analysis.

Figure 17 shows the distribution of the amplitude and phase of towers deployed in the comprehensive, residential, office, transport, and entertainment areas. Meanwhile, means and standard deviations of the amplitude and phase for towers at the three frequency components of towers in the 5 types of areas are presented in Figure 18.

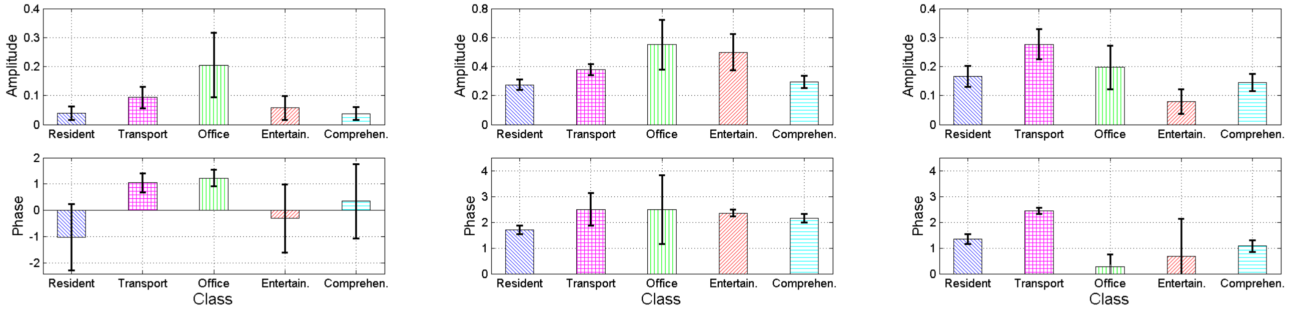
From Figure 17(a) and Figure 18(a), we can observe that towers in office area have the strongest periodicity of one week. Their phases mainly concentrate around 1.35, while the phase of towers in residential and entertainment area centers around -1.65, about π away from 1.35. This π separation suggests that towers in residential and entertainment area have reverse traffic characteristics as that in the office area in the scale of one week.

As we can observe in Figure 17(b), the distribution of towers is continuous with respect to the phase of one day. Moreover, it shows a smooth traffic transition from residential area to comprehensive and transport area, and finally to office area.



(a) (b) (c)

Fig. 17. Phase and amplitude distribution of the three principal frequency components in the frequency domain. (a) $k = 4$. (b) $k = 28$. (c) $k = 56$.



(a) (b) (c)

Fig. 18. Means and standard deviations of amplitude and phase for cellular towers from the five identified patterns. (a) Week. (b) One day. (c) Half a day.

On the other hand, according to our priori knowledge, the human migration flow usually leads to the peaks of traffic of areas appear sequentially with the same order that the flow passes through, which coincides with our observed phenomenon. Thus, such transition suggests the human migration flow from home to office via transport during rush hours. In Figure 18(b), we can also observe that the means of their phase are incremental with the same order.

Figure 17(c) and Figure 18(c) show characteristics of the amplitude and phase of the frequency component which stands for half a day. The amplitude of this frequency component indicates the strength of double-hump characteristic. In Figure 18(c), we can observe that the amplitude of towers in transport area is the largest, indicating their strongest double-hump characteristic. This result coincides with our priori knowledge that there are two rush hours of transport area in the morning and evening, respectively. In Figure 17(c), we find that traffic of residential and office area are not separated by traffic of transport area. This observation is not contradictory to our pervious analysis because the directions of people commute in the morning and afternoon are reversed.

Overall, the amplitude and phase of the three frequency components show a strong capability of differentiating towers with different traffic patterns. Based on the observations, we make the following statements. First, the most representative tower in each cluster is not the centroid. In fact, it is the farthest non-noise point from the hyperplanes, which separate clusters. To understand this problem, let us think about the points around a hyperplane, where we observe similar traffic patterns of points even though they belong to different clusters. In geographical context, these towers are deployed in areas of mixed urban functions. In contrast, the points far from the separating hyperplane are located at areas of a single urban function. Although perhaps not the most representative points, cluster centroids can well characterize the traffic patterns since they are distant from others clusters.

Second, the frequency-domain features of towers are distributed in a polygon. Such polygon is formed because the profile of each cluster in Figure 17 has a cigar shape. Thus, different features of towers can be regarded as being linear relevant or piecewise linear relevant approximately, which overlayed with a Gaussian noise can form the cluster with the cigar shape. As a result, a point in the frequency domain can be seen as a linear combination of the four vertex of the polygon, $i.e.$, the four most representative points, which we call as the four primary components.

To illustrate these two statements, we plot the distribution of towers and corresponding polygon in Figure 19. For better understanding, we only show three features, including amplitude and phase of one day, and amplitude of half day. According to our first statement, the most representative tower in each cluster is the furthest one from the hyperplane. Specifically, we do not calculate the hyperplanes, and only search for points with largest distance from points of other clusters. In addition, we use the density of the towers, i.e., the number of towers within a fixed distance away from it in the

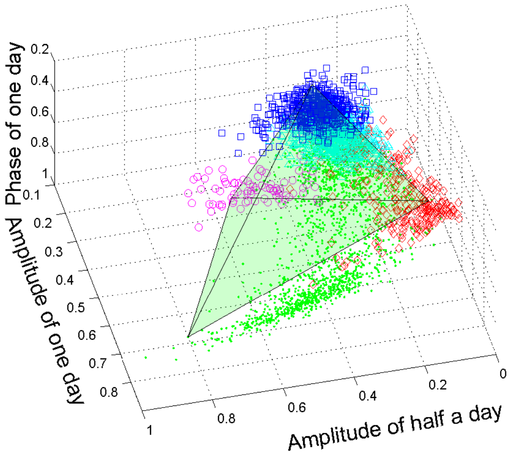


Fig. 19. Three-dimensional view of the distribution of cellular towers in the frequency domain.

TABLE VI

CONVEX COMBINATION COEFFICIENTS AND NTF-IDF

	Coefficient				NTF-IDF			
	1	2	3	4	1	2	3	4
F1	1.00	0.00	0.00	0.00	1	0	0	0
F2	0.00	1.00	0.00	0.00	0.00	0.81	0.05	0.14
F3	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
F4	0.00	0.00	0.00	1.00	0.00	0.00	0.28	0.72
P1	0.79	0.13	0.08	0.00	0.44	0.36	0.04	0.16
P2	0.09	0.07	0.00	0.84	0.36	0.39	0.03	0.22
P3	0.23	0.00	0.15	0.62	0.68	0.00	0.07	0.25
P4	0.00	0.29	0.42	0.29	0.00	0.68	0.07	0.24
P5	0.35	0.18	0.22	0.25	0.12	0.55	0.05	0.28

feature space, as a decision function to ensure that the tower is not a noise point. Figure 19 shows that all the towers are distributed in or along the edge and plane of the polygon, as we discussed above.

C. Component Analysis of Cellular Towers in Comprehensive Area

Based on the statements above, we may use a linear combination of the four most representative cellular towers to present each point in the polygon. By looking at the coefficient of each primary component, we can obtain the percentage of corresponding urban function of the area where an arbitrary cellular tower is deployed. We formulate the process of obtaining the coefficients as a quadratic programming problem, which is shown below:

$$\begin{aligned} &\text{minimize} \quad ||F - F^r||^2 \\ &\text{subject to} \quad \begin{cases} \sum_{i=1}^{4} F_i^0 x_i = F^r, \\ \sum_{i=1}^{4} x_i = 1, \\ x_i \geq 0, \quad i = 1, ..., 4, \end{cases} \end{aligned}$$

where $||\cdot||$ is the 2-norm of a vector, F is the feature of the target tower, F_i^0 is the feature of the most representative tower for cluster i in the frequency domain, and x_i is the obtained coefficient for cluster i. In this example, the feature of tower m, F^m, is $(A_{28}^m, P_{28}^m, A_{56}^m)$, where A_{28}^m ,P_{28}^m, A_{56}^m are the amplitude of one day, phase of one day, and amplitude of half a day for tower m, respectively. We use the quadratic programming to solve the problem because the traffic of an actual tower is usually overlayed with various noises, such that these points close to the plane of the the polygon may be driven out of the polygon. By solving this quadratic programming, for points inside the polygon, we can find their exact convex combinations, while for some point outside the polygon, we can find the point in the polygon with the smallest distance to the target point, which is a good approximation.

We dedicatedly select a list of towers in the comprehensive area. Then, we use the method presented above to solve the convex combinatorial coefficients of them. We compare these coefficients with a transform of the previously introduced POI, *i.e.*, the term frequency-inverse document frequency (TF-IDF) of the corresponding types and locations. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document. Similarly, it is used to reflect how important the POI of a specific type is in our analysis, which has been proposed in existing works, i.e., Yuan *et al.* [11] provided a TF-IDF-based method to cluster regions of different functions, which solely uses the POI data. Specifically, TF-IDF can be calculated as the following:

$$\begin{cases} \text{IDF}_i = \log(M/M_i), \\ \text{TF-IDF}_i^m = \text{IDF}_i \cdot \log(1 + \text{POI}_i^m), \end{cases}$$

where M is the total number of towers, and M_i is the number of towers of which the POI of type i appears within a specific distance, and POI_i^m is the times that the POI of type i appears within a fixed distance of tower m. To be better compared with, we normalize the TF-IDF of each tower by the sum of TF-IDF of all the four types for this tower, which is called as the normalized TF-IDF (NTF-IDF). This process can be formulated as the following:

$$\text{NTF-IDF}_i^m = \text{TF-IDF}_i^m / \sum_{j=1}^{4} \text{TF-IDF}_j^m.$$

The obtained NTF-IDF is proportional to the POI for each type, which roughly represents the density of the corresponding function in the corresponding area. Specifically, NTF-IDF close to 0 indicates this area do not have the corresponding function. However, the largest NTF-IDF do not completely indicate the corresponding function is dominant in the area, since it is also influenced by the size of related points and corresponding distance. For example, a large and close subway station has more influence than a small and far residential building on a cellular tower.

Then, the result is shown in Table VI. Expect for the towers in the comprehensive area, the NTF-IDF of the four most representative towers is also provided in the table. We can observe that their NTF-IDF of corresponding types is much larger than others, which is very close to 1, indicating the areas where they are located have a single type of function. As for towers in the comprehensive area, There are multiple relative large NTF-IDF for a cellular tower. As discussed earlier, this may lead to inaccuracy because of the influence of the size of related points and corresponding

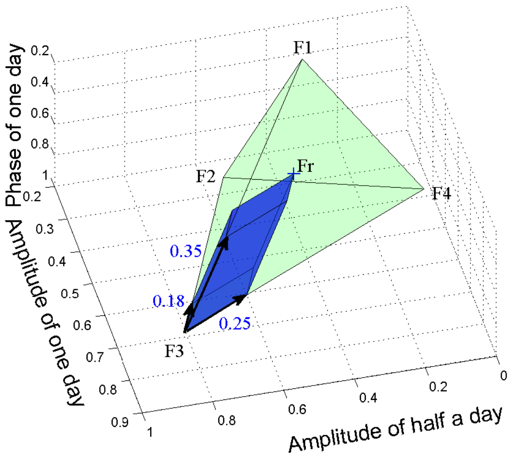


Fig. 20. Convex combination for P5 in Table VI in frequency domain.

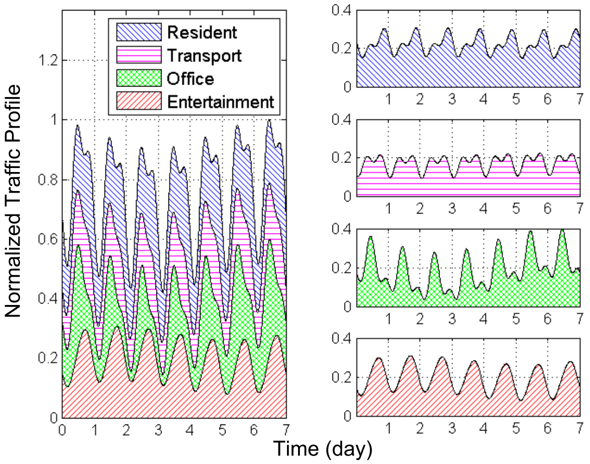


Fig. 21. Convex combination for P5 in time domain.

distance. Thus, we only consider the consistency of the small NTF-IDF and combination coefficients. We can observe that the majority of the smallest NTF-IDF$_i^m$ in all m for some fix i corresponds to the smallest coefficient in all m for the same i, respectively. For example, NTF-IDF$_2^{P3}$ and NTF-IDF$_1^{P4}$ are 0, and their corresponding coefficients are also 0. Thus, the obtained convex combination coefficients coincide with the POI distribution, indicating the correctness of our theory.

To further illustrate the convex combination, we take the tower P5 in Table VI as an example, and show its combination of frequency and time domain, respectively in Figure 20 and Figure 21. For a point inside the polygon, we can find its exact convex combination, that is:

$$F = F^r = \sum_{i=1}^{4} F_i^0 x_i = F_3^0 + \sum_{i=1, i\neq 3}^{4} x_i (F_i^0 - F_3^0).$$

As shown in Figure 20, in the feature space, the vector $(0, Fr)$ can be divided to the vector $(0, F3)$ and the weighted sum of the vector $(F3, F1), (F3, F2), (F3, F4)$. For P5, the weights are 0.35, 0.18 and 0.25, respectively, which is just as coefficients of cluster 1, 2, 4 in Table VI of P5.

On the other hand, in Figure 21, we show the components of traffic corresponding to four primary clusters for the comprehensive tower P5. Areas of different colors in the left figure represent components of different primary traffic patterns. To be better distinguished, each component is added with a static bias. In addition, we plot each component individually in the right figure. The result indicates that traffic patterns of an arbitrarily cellular tower can be approximated by a convex combination of four primary traffic patterns. The size of each component is highly related to density of corresponding function around the tower. It further demonstrates the correctness and usefulness of our frequency analysis method.

VI. Related Work

The digital footprints of human activities and network behaviors contributed by mobile devices have led to a plethora of investigations on the intersection between human and network dynamics [7], [12]. This section summarizes relevant research from three perspectives — data sources, types of collected data and targeted applications.

Dataset collected from mobile devices for investigating human behaviors and network performance can be divided into two broad categories: (1) data collected from mobile devices and (2) traces collected by mobile operators [2]. For the first categories, users or experimenters report their semantically annotated data about the locations, phone usages and network performance by installing some Apps in their devices [13], [14]. The limitation of this approach comes from the limited number of users sampled, which cannot stand for the global characteristics of a large scale cellular network. On the other hand, in the dataset collected by cellular operators, users are passively monitored and the operators decide which information to collect [15], [16]. As a result, the collected data is continuous as long as devices are connected, and includes detailed information of users behaviors, such as duration of each Internet connection. As a result, data collected via the second approach enables the study of overall network behaviors, such as large scale of human mobility and call activities analysis. In this paper, we use the data collected by an ISP for investigating the traffic patterns of large scale cellular towers.

Extensive studies have used various types of cellular data for understanding the characteristics of large scale cellular towers. For example, cell phone activities, commonly know as Call Description Records (CDR), are used for capturing human communication activities [17]. In addition, it is also used for recovering the human mobility trajectory [15], inferring demographics [4], and uncovering urban ecology [2]. Another type of data is the device-level metric obtained from mobile devices, such as device and application usage [6], [18], network access bandwidth [14], energy computation [19], personal GPS locations [20], etc. With the popularity of 3G and LTE access, mobile and application data traces become available as well. Cici et al. [2] characterizes the relationship between people's application interests and mobility patterns based on a population of over 280, 000 users of a 3G mobile network. Lee et al. [5] demonstrated that the spatial distribution of the traffic density can be approximated by the log-normal or Weibull distribution. However, mobile data traffic across a city-wide range with different time scale and variations contains complicated interaction between the space and time, which requires a deep and comprehensive understanding. The analysis and models in this work provide such insights.

Cellular network traces have been used for enabling a set of applications. One of the most important applications is investigating and modelling human mobility. With the mobile devices served as an ideal tool to monitor individual's location, human mobility has been extensively studied at different time-scale as well as different spatial-scale in the past few years. Barabási [15], [21]–[23] studies the long-term mobility of individuals based on a six months' phone call record across 100k users. They find that the long-term mobility of an individual is not consistent with the previously proposed levy flight model, and achieve up to 93% accuracy in predicting the mobility of individuals [23]. In addition, [21] proposes a radiation model to predict the movement of a group of people. Lee et al. [24] propose a mobility model to simulate the daily individual person mobility. Chaintreau et al. [25] study the human mobility across various time scale. They find that the inter-contact time can be approximated by power law.

The cellular network traces have also been used for characterizing and modelling the cellular data traffic patterns. Zubair and Lusheng [26] modelled the internet traffic dynamics of cellular devices. Jin et al. [27] characterized data usage patterns in large cellular network. And Zhang [28] tried to understand the characteristics of cellular data traffic by comparing it to wireline data traffic.Other studies combine the CDR, GPS locations, and application traces to investigate the land usage [16], [29], social interactions [30], location-based patterns [3], and web and data access patterns [27], [31]. In this paper, we focus on investigating the mobile data traffic patterns from different domains, including time, location and frequency, which provides a comprehensive understanding of the traffic patterns of large scale cellular towers with a simple but deep model that is able to characterize the city geographical features and human communication regularity.

In conclusion, we study a large scale urban mobile data access traces collected by the commercial mobile operators involving over 9600 towers and 150,000 subscribers. We first design an analysis framework for processing large scale cellular traffic data. Then, we reveals the basic but fundamental patterns embedded in thousands of cellular towers, which paves a way toward a comprehensive understanding of the connection among mobile data traffic, urban ecology and human behaviors.

VII. Conclusions

In this paper, we carry out, to the best of our knowledge, the first study of traffic patterns embedded in large scale 3G and LTE towers deployed in the urban environment. We propose a powerful model which combines time, location and frequency information for analyzing the traffic patterns of thousands of cellular towers. Our analysis reveals that the dynamic urban mobile traffic usage exhibits only five basic time domain patterns. In addition, the traffic of any tower can be reconstructed accurately using a linear combination of four primary components corresponding to human activity behaviors. Our analysis provides a systematic and comprehensive understanding of dynamic and complicated mobile traffic, and opens a set of new research directions.

References

[1] *Cisco C V N I. Global Mobile Data Traffic Forecast update, 2013-2018*. White Paper, Feb 2014, pp. 1–40.

[2] B. Cici, M. Gjoka, A. Markopoulou, and C. Butts, "On the decomposition of cell phone activity patterns and their connection with urban ecology," in *Proc. ACM MOBIHOC*, 2015, pp. 317–326.

[3] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Contextual localization through network traffic analysis," in *Proc. IEEE INFOCOM*, Apr. 2014, pp. 925–933.

[4] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social," in *Proc. ACM SIGKDD*, 2014, pp. 15–24.

[5] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 80–88, Feb. 2014.

[6] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3g cellular data network," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1341–1349.

[7] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 15, pp. 5802–5805, 2013.

[8] H. Wang *et al.*, "Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks," in *Proc. ACM HOTPOST*, 2015, pp. 19–24.

[9] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic Acids Res.*, vol. 16, no. 22, pp. 881–890, 1988.

[10] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[11] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proc. ACM SIGKDD*, 2012, pp. 186–194.

[12] J. K. Laurila *et al.*, "The mobile data challenge: Big data for mobile computing research," in *Proc. Pervasive Comput.*, 2012.

[13] A. Noulas C. Mascolo, and E. Frias-Martinez, "Exploiting foursquare and cellular data to infer user activity in urban environments," in *Proc. IEEE MDM*, Jun. 2013, pp. 167–176.

[14] W. Hu and G. Cao, "Quality-aware traffic offloading in wireless networks," in *Proc. ACM MobiHoc*, 2014, pp. 277–286.

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

[16] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proc. ACM SIGKDD*, 2012, pp. 1–8.

[17] B. Cici, A. Markopoulou, E. Frías-Martínez, and N. Laoutaris, "Quantifying the potential of ride-sharing using call description records," in *Proc. ACM HotMobile*, 2013, Art. no. 17.

[18] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, "A first look at traffic on smartphones," in *Proc. ACM IMC*, 2010, pp. 281–287.

[19] W. Hu and G. Cao, "Energy-aware video streaming on smartphones," in *Proc. IEEE INFOCOM*, May 2015, pp. 1185–1193.

[20] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proc. ACM UbiComp*, 2008, pp. 312–321.

[21] 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, 2012.

[22] 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, 2010.

[23] 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.

[24] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: A new mobility model for human walks," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 855–863.

[25] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Trans. Mobile Comput.*, vol. 6, no. 6, pp. 606–620, Jun. 2007.

[26] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling Internet traffic dynamics of cellular devices," *Perform. Eval. Rev.*, vol. 39, no. 1, pp. 305–316, 2011.

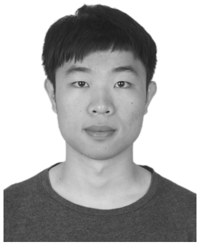
[27] Y. Jin *et al.*, "Characterizing data usage patterns in a large cellular network," in *Proc. ACM CellNet Workshop*, 2012, pp. 7–12.

[28] Y. Zhang and A. Årvidsson, "Understanding the characteristics of cellular data traffic," in *Proc. ACM SIGCOMM CellNet Workshop*, 2012, vol. 42. no. 4, pp. 13–18.

[29] T. Pei *et al.*, "A new insight into land use classification based on aggregated mobile phone data," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 9, pp. 1988–2007, 2014.

[30] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 36, pp. 15274–15278, 2009.

[31] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3g network using hourglass co-clustering," in *Proc. ACM MobiCom*, 2010, pp. 341–352.



Fengli Xu received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2015. He is currently pursuing the Ph.D. degree with the Electronic Engineering Department, Tsinghua University, Beijing, China. His research interests include human mobility, mobile big data mining, and user behavior modeling.

Yong Li (M'09–SM'16) received the B.S. degree from the Huazhong University of Science and Technology in 2007, and the Ph.D. degree from Tsinghua University in 2012. From 2012 to 2013, he was a Visiting Research Associate with Telekom Innovation Laboratories and the Hong Kong University of Science and Technology. From 2013 to 2014, he was a Visiting Scientist with the University of Miami. He is currently a Faculty Member with the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of mobile computing and social networks, urban computing and vehicular networks, and network science and future internet. He has served as the General Chair, the Technical Program Committee (TPC) Chair, and a TPC Member of several international workshops and conferences. He is currently the Associate Editor of the *Journal of Communications and Networking* and the *EURASIP Journal of Wireless Communications and Networking*.

Huandong Wang received the B.S. degrees in electronic engineering and mathematical sciences from Tsinghua University, Beijing, China, in 2014 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Tsinghua University. His research interests include software-defined networks, wireless ad hoc network, and mobile big data.

Pengyu Zhang received the bachelor's and master's degrees from Tsinghua University in 2007 and 2010, respectively, and the Ph.D. degree from the University of Massachusetts Amherst in 2015. He is currently a Post-Doctoral Researcher with Stanford University. His research interests are embedded systems, sensing, networking, and wireless Communication. He is the winner of the 2016 School of Computer Science Outstanding Dissertation Award from the University of Massachusetts Amherst, the UbiComp 2016 Honorable Mention Award, and the Mobicom 2014 best paper award runner up.

Depeng Jin (M'09) received the B.S. and Ph.D. degrees in electronics engineering from Tsinghua University, Beijing, China, in 1995 and 1999, respectively. He is currently an Associate Professor with Tsinghua University. His research fields include telecommunications, high-speed networks, ASIC design, and future internet architecture. He is a Vice Chair of the Department of Electronic Engineering. He received the National Scientific and Technological Innovation Prize (Second Class) in 2002.