Understanding Urban Dynamics From Massive Mobile Traffic Data

Mingyang Zhang, Haohao Fu, Yong Li, *Senior Member, IEEE*, and Sheng Chen, *Fellow, IEEE*

Abstract—Understanding the patterns of mobile data consumption is extremely valuable to reveal human activities and ecology in urban areas. This task is nontrivial in terms of three challenges: the complexity of mobile data consumption in large urban environment, the disturbance of abnormal events, and lack of prior knowledge for urban traffic patterns. We propose a novel approach to design a powerful system that consists of three subsystems: time series decomposing of mobile traffic data, extracting patterns from different components of the original traffic, and detecting anomalous events from noises. Our investigation involving the mobile traffic records of 6,400 cellular towers in Shanghai reveals three important observations. First, among all the 6,400 cellular towers, we identify five daily patterns corresponding to different human daily activity patterns. Second, we find that two natural patterns can be extracted from the weekly trend of mobile traffic consumption, which reflects modes of human activities. Last but not least, besides the regular patterns, we investigate how irregular activities affect mobile traffic consumption, and exploit this knowledge to successfully detect unusual events like concerts and soccer matches. Our proposed methodology therefore will aid a comprehensive understanding of large-scale mobile traffic consumption in urban areas.

Index Terms—Mobile data, data decomposition, urban human activities, urban computing

1 Introduction

As a result of the ubiquitous access to LTE and 4G networks, huge amount of mobile traffic data are consumed. The global mobile data traffic has grown 4,000 times in the past 10 years, and our society is facing a dramatic acceleration in the growth of cellular data traffic. According to [1], the global mobile data traffic has reached 3.7 exabytes per month in 2015 and it is expected to reach 30.6 exabytes per month by 2020. Consequently, analyzing mobile cellular traffic becomes a key approach to understand the human behaviors and ecology in urban areas. However, we still have limited knowledge about how people's regular activities and unexpected events affect the mobile traffic of cellular towers [2]. Such knowledge is extremely valuable to recognize the patterns of cellular traffic and to understand their geographical distribution [3], [4], [5], [6]. For example, the Internet service providers (ISPs) can determine the cellular towers' locations according to the traffic distributions and adopt appropriate strategies for the towers of different patterns to ease traffic loads in peak hours. In addition, if a method can be developed to accurately detect anomalies in cellular traffic data, it will help ISP to identify equipment failures or unusual crowd events to adopt actions to reduce potential loss. More fundamentally, correctly identifying mobile traffic patterns is extremely important in understanding human activities, which provides valuable insight for designing better infrastructure and living environments.

Cellular network record is an ideal dataset for investigating urban human activities. In mobile daily life today, Internet accessing through cellular network happens frenquently. We consume cellular data all the time throughout our urban life, checking out in stores by mobile phone, calling a taxi by taxi-hailing apps, sharing life with friends on social apps, etc. Compared to call description records (CDRs), which are also commomly used in revealing human life patterns in cities, mobile traffic data records are much denser in time domain due to the higher frequent accessing. Meanwhile, the distance between two neighboring cellular towers in urban area is only 200-300 m, which means mobile traffic data also provide a good geographical granularity. Social media data are another important data source widely produced in urban area. However, because of the varity of platforms and formats, social media data are hard to collect and mine. For example, when a road is under heavy traffic, the mobile traffic of cellular towers around it can grow sharpely as well and easy to be detected, while the information collected from social media is always vague and unreliable.

Recognizing the traffic patterns of large-scale cellular towers is challenging because of three reasons. First, the traffic of cellular towers is complex because the traffic patterns of different towers differ vastly. Moreover, even the traffic of a same cellular tower shows different patterns in different time scales. This makes it difficult to find a universal method to analyze it. Nevertheless, we need to find a

- *M. Zhang and Y. Li are with Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: {mingyangzhang, liyong07}@tsinghua.edu.cn.*
- *H. Fu is with Princeton International School of Mathematics and Science, Princeton, NJ 08540. E-mail: 18611608511@163.com.*
- *S. Chen is with Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, United Kingdom, and also with King Abdulaziz University, Jeddah 21589, Saudi Arabia. E-mail: sqc@ecs.soton.ac.uk.*

Manuscript received 1 Mar. 2017; revised 1 Sept. 2017; accepted 17 Oct. 2017. Date of publication 30 Nov. 2017; date of current version 7 June 2019. (Corresponding author: Yong Li.)
Recommended for acceptance by D. Zhang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TBDATA.2017.2778721

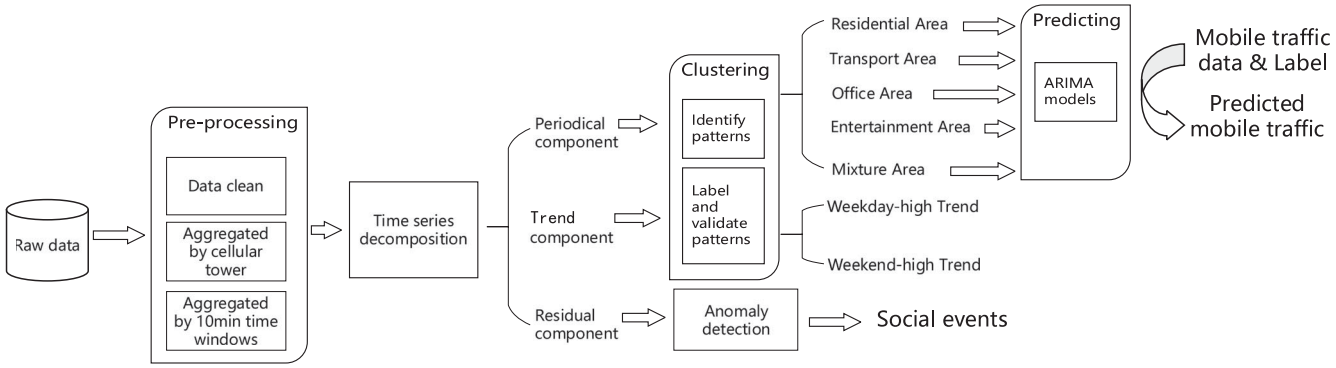


Fig. 1. Framework overview.

method to characterize these differences and to develop a model capable of considering various situations. Second, the traffic of cellular towers is affected by accidental events, which further introduces complexity into analysis. For example, when a parade occurs around a cellular tower, its traffic amount will rise sharply and deviate greatly from its normal patterns. Therefore, how to eliminate the influence of these accidental events to pattern analysis and to detect anomalous events are also difficult. Third, we have limited prior knowledge about the traffic patterns of cellular towers, and this makes it hard to decide an appropriate number of patterns and to identify their features. On one hand, there may exist vast number of patterns as number of cellular towers is large, but some of them are not very useful in studying human urban activities. On the other hand, some cellular towers show mixture features of more than one patterns. Therefore, it is challenging to identify a small number of principle patterns hidden in the vast amount of traffics among these cellular towers.

To address these challenges, we develop an effective system to analyze and model mobile traffic data from thousands of cellular towers. Fig. 1 gives an overview of our proposed framework. Our system is composed of five parts including data preprocessing, time-series decomposition, pattern modeling, anomaly event detecting and traffic predicting. It can deal with large-scale data efficiently and complete all the steps of analysis in a few minutes. More specifically, in our system, time series analysis theory [7] is utilized to decompose the traffic consumption series into three components: trend component, seasonal component, and residual component. The trend component represents a non-periodic tendency over the whole time scale, the seasonal component reflects a periodic change which usually corresponds to normal activities, and the residual component is considered as noises or unusual events. We model human activity patterns from seasonal component and introduce a method to forecast mobile data traffic. In addition, we propose a method to detect unusual events from the traffic records and validate the results with real world traces. In order to eliminate the influence of anomalous events and model the traffic pattern from different perspectives, we decompose the original mobile traffic data and extract the principal traffic patterns by exploiting hierarchical clustering [8], which does not require a predefined number of patterns. This enables us to detect unexpected events from the residual component.

By applying our system to investigate the mobile traffic records of 6,400 cellular towers, collected by ISP from Shanghai, we find the following interesting results.

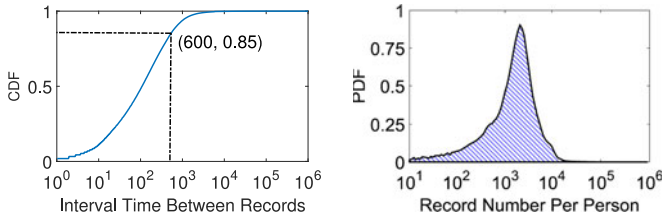
- The cellular towers can be divided into five groups according to the one-day seasonal component of their traffic consumptions, and these groups correspond to five types of urban function areas: residential area, transport area, office area, entertainment area, and mixture area. This finding reveals how human activity patterns influence the mobile traffic. Furthermore, based on this finding we introduce ARIMA model to forecast mobile traffic and achieve a high accuracy, which in turn prove the significance of this finding.
- We can model the traffic patterns with the weekly trend component to identify the two principal trends: which increases and decreases through the week, respectively. These two weekly trend patterns reflect week-long human activities in the real world.
- Our investigation confirms that unusual events can be detected from the residual component using our proposed method of anomalous event detection, and we show that the detected anomalous events match the anomalies with real-world events.

The rest of this paper is structured as follows. In Section 2, we describe the utilized dataset and provide our motivations. In Section 3, we present all the components of our system. Specifically, we present our decomposition and clustering techniques, and by discussing the correlation between daily patterns and weekly trend patterns, we develop a mobile traffic consumption forecasting model. Then, we propose a method to detect anomaly events from residual component. In Section 4, we evaluate the results, and Section 5 reviews the related work, while Section 6 concludes this paper.

2 Data Set and Motivation

2.1 Data Description and Initial Observations

The original dataset is an anonymized cellular trace collected by ISP from Shanghai, which contains 2.4 petabytes (10^{15}) logs from around 6,400 base stations (BSs) all over Shanghai between August 1st and August 31st of 2014. Each entry in the trace consists of ID of devices (anonymized), start-end time of data connection, BS ID, address of BS, and the amount of 3G or LTE data used in each connection. This large scale dataset represents human activities in Shanghai and it provides a sound real-world physical basis for our analysis. We present several basic visualizations



(a) CDF of interval time (b) PDF of log number

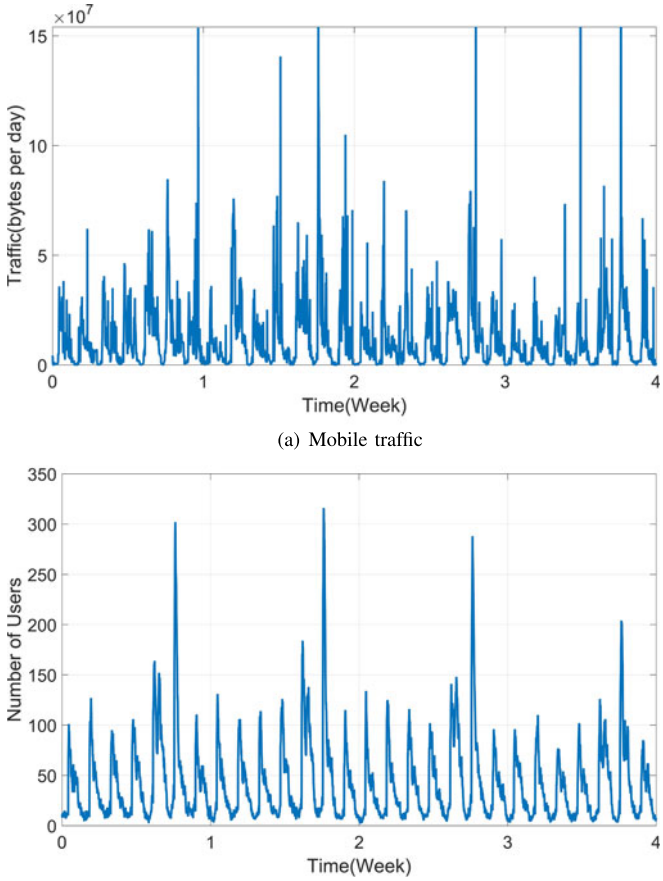
Fig. 2. Illustration of the quality of our dataset.

about characteristics of this dataset in Fig. 2. The subplot (a) shows the Cumulative Distribution Function (CDF) of interval time between two consecutive records. From the results, we can observe that more than 85 percent consecutive records happen in less than 600 seconds. Compared to a 8.2 hours average inter-event time of consecutive mobile phone calls [3], the cellular data accessing records are much more fine grained. Fig. 2b shows Probability Density Function (PDF) of the number of records per users, respectively. Most of mobile users have more than 1,000 records in total. These observations demonstrate that our dataset has extensive records of mobile users and the fine temporal granularity guarantees the credibility of human activity modeling.

In order to make better use of the dataset, we process it into a ready to use form. Specifically, we first eliminate the redundant and conflict logs caused by technical issues. Then, we aggregate the traffic and the number of users of each BS in small time windows. We find 10-minute interval is a suitable time window for our data. Thus, for each BS we obtain an array for both traffic volume and the number of users. The length of the array is 4,032 and each element is the traffic volume or the number of users in the corresponding ten-minute duration, which covers 28 days. Lastly, we convert the addresses of BSs to their geographical longitudes and latitudes through APIs provided by Baidu Map, which helps us to identify the geographical features. Typical examples of our dataset are illustrated in Figs. 3 and 4.

Fig. 3 shows the mobile traffic consumption and the number of users of a selected cellular tower. From the whole four-week mobile traffic shown in Fig. 3a, although we can identify 28 one-day periods clearly, it can be observed that the mobile traffic data contains complex fluctuations, which makes it difficult to analyze. Fig. 3b shows how the number of users changes during four weeks for this cellular tower. By aligning Fig. 3b with Fig. 3a, an intriguing observation is that the mobile traffic series is obviously full of peaks, and some of them correspond to the peak numbers of users, but others are caused by unknown factors. Clearly, the influence of abnormal events and noises further complicates analysis.

Fig. 4 shows the geophysical traffic densities at 4AM, 10AM, 4PM and 10PM, from which we can draw two observations. First, the traffic consumption varies in different times due to different human activities during one day. At 4AM, most areas are under low mobile traffic consumption because people sleep at this time. At 10AM and 4PM, the mobile traffic consumption increase since most people are working. At 10PM, a larger scale of high mobile traffic consumption arises because people go home and are relax at this time. Second, the mobile traffic consumption of cellular towers shows different temporal features at different locations. For example, cellular towers deployed at the center of the city experience



(a) Mobile traffic

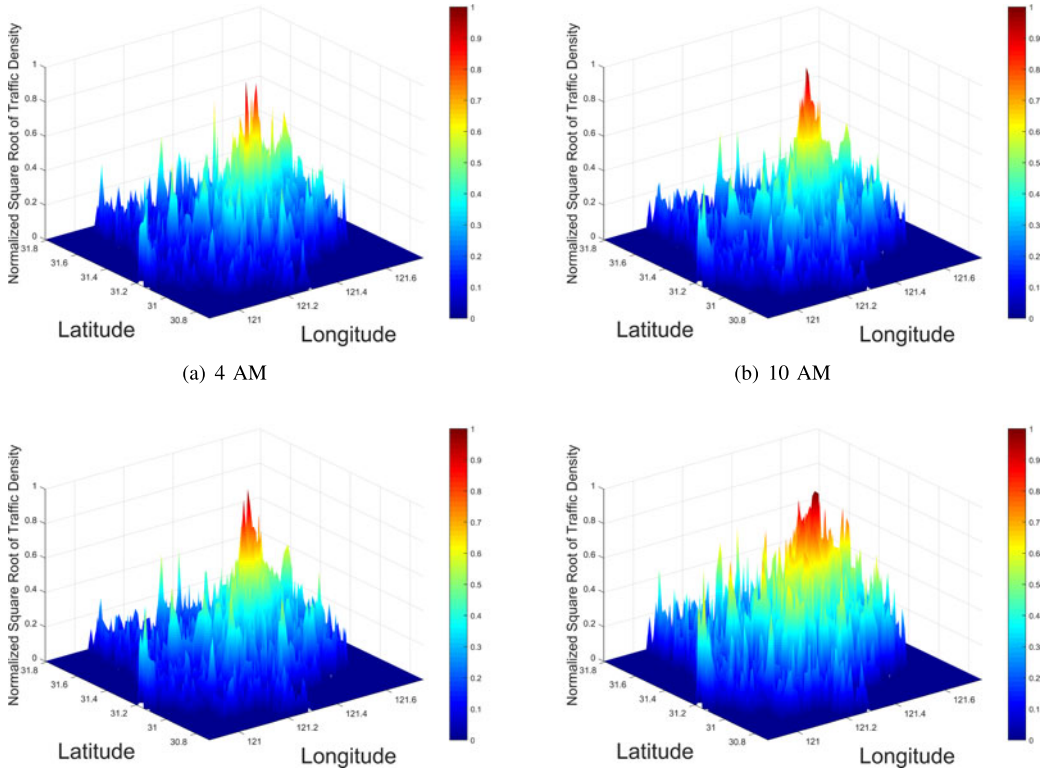
(b) The number of users

Fig. 3. Variations of mobile users and traffic of a month.

high traffic consumption in all time. The hottest areas covered by the darkest color also change over time, which suggests the movement of the crowd during one day.

2.2 Motivation

The cellular traffic data provide an up-to-date indicator on human urban activities. However, due to the complexity of mobile traffic data, the disturbance of unusual events and the lack of prior knowledge, it is challenging to extract the information we want directly from the original traffic data. By exploiting the theory of time series analysis [7], we adopt a time series decomposition strategy to address these three problems, specifically, complexity of mobile traffic data, the disturbance of of abnormal events, and the lack of prior knowledge for data patterns. There are two advantages of adopting this approach to analyze mobile traffic data. First, if we consider the traffic data as a time series, a basic observation is that the four-week time series have a natural period of one day and a clear trend within one week. We will be able to investigate these features independently by studying different time series components. Thus, this decomposition may reduce the complexity of the original traffic significantly. Moreover, we can extract the disturbance of irregular events by a time series decomposition which provides the basis for developing novel event detection. Second, based on the components produced by the decomposition, hierarchical clustering can be applied to automatically identify the principal patterns among the traffic data from thousands of cellular towers.



(a) 4 AM (b) 10 AM

(c) 4 PM (d) 10 PM

Fig. 4. The spatial distributions of cellular traffic at different times.

Fig. 5 depicted the aggregated traffic of the 6400 towers at different time scales. Specifically, Fig. 5a shows the daily fluctuations, and in Fig. 5b, the daily fluctuations among a week exhibit great similarity. Furthermore, Fig. 5c illustrate the traffic amount of everyday in a month, which indicates a clear weekly trend. These intuitive observations motivate us to extract the daily patterns and weekly trends from the mobile traffic time series. Therefore, in order to analyze the traffic data in detail and to model the traffic patterns in different scales, we proposed to decompose the traffic data into three components: the periodic components, the trend components and the residuals.

3 System and Algorithm

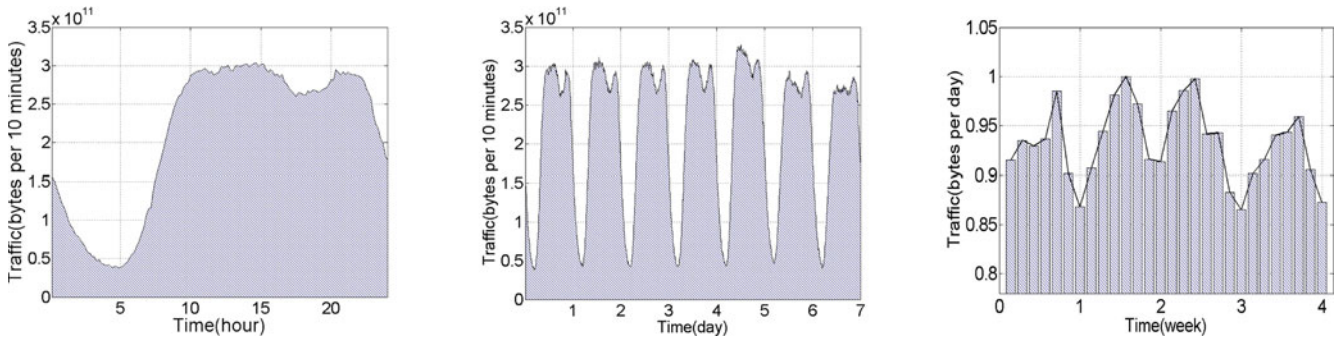
3.1 Decomposition

As explained in Section 2.1, we have a mobile traffic record or observation $\{x_1, x_2, \ldots, x_{4032}\}$. Each entry of the observation is a 10-minute traffic amount of one tower, and the observation is a traffic data record of the tower within 28 days. Furthermore, as shown previously this record exhibits both trend and periodicity. Hence, we can apply a time series approach [7] to decompose this record as follow:

$$x_t = s_t + m_t + r_t,\ t = 1, 2, \ldots, n,$$

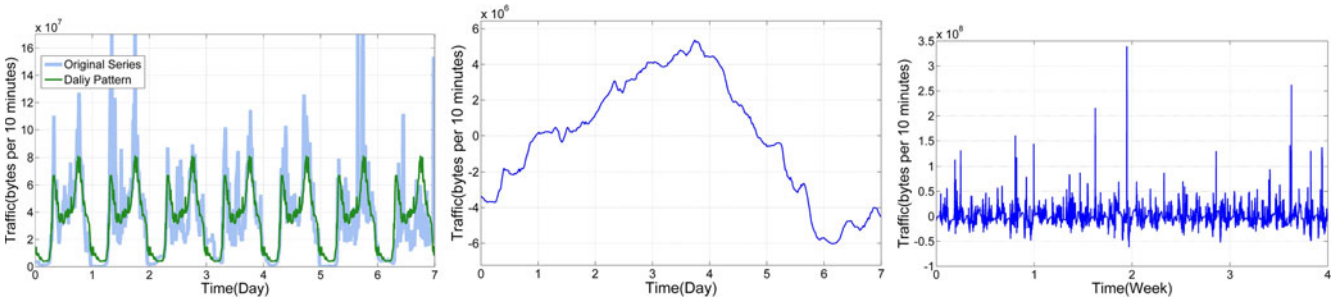
where s_t is a daily periodic traffic component, which satisfies $s_t = s_{t+d}$ for $t = 1, 2, \ldots, n-d$ with the period $d = 144$ that corresponds to one day, m_t is a trend traffic component, and r_t is the residual component containing the noise and the effects of unusual events. We can complete the decomposition in two steps: find the periodic component first and then estimate the trend component from the rest series in the absence of seasonality.

We first estimate the trend traffic component by adopting a simple moving average filter



(a) Hourly (b) Daily (c) Weekly

Fig. 5. The temporal distributions of the normalized cellular traffic at different time scales.



(a) Periodical component and original series (b) Trend component (c) Residual component

Fig. 6. Illustration of decomposition on the traffic patterns of one base station.

$$\widetilde{m}_t = \big(0.5x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + 0.5x_{t+q}\big)/d,$$

where $q = d/2 = 72$ and $q < t < n - q$. Then, we calculate the deviation series

$$\{(x_{k+jd} - \widetilde{m}_{k+jd}), q < k + jd < n - q\},$$

for $k = 1, 2, \ldots, d$, and compute the average ω_k of this series. We can express the one-day stable periodic component by

$$s_k = \omega_k - d^{-1}\sum_{i=1}^{d}\omega_i,\ k = 1, 2, \ldots, d,\ s_k = s_{k-d}\text{if}k > d,$$

which reflects the normal daily traffic pattern of the cellular tower.

The de-seasonality series is obtained by removing the periodic component from the original series, expressed as

$$d_t = x_t - s_t,\ t = 1, 2, \ldots, n.$$

Note that we aim to find the weekly tendency of the mobile traffic data from the monthly series $\{d_t\}$. Thus we compute the one-week average $\widehat{d}_t$ of the original four-week data $\{d_t\}$. Then, we estimate the weekly trend from the weekly average series $\{\widehat{d}_t\}_{t=1}^{1008}$. There are two general methods to estimate the trend component of a time series: smoothing with a finite moving average filter and function modeling. Since the tendencies of the mobile traffic series vary among cellular towers, they are hard to be modeled with a universal function. Thus, the former method fits our data better. Let p be a positive integer. The two-side moving average

$$m_t = (2p + 1)^{-1}\sum_{j=-p}^{p}\widehat{d}_{t-j},$$

provides a simple estimate of the weekly trend. Obviously, in the above expression $\widehat{d}_t$ is not defined if $t \leq 0$ or $t > 1008$, we solve this problem by defining $\widehat{d}_t = \widehat{d}_1$ for $t \leq 0$ and $\widehat{d}_t = \widehat{d}_{1008}$ for $t > 1008$. We empirically find that $p = 100$ is appropriate for our data.

Finally, the residual component is simply obtained as

$$r_t = x_t - s_t - m_t,\ t = 1, 2, \ldots, n,$$

Fig. 6 illustrates an example of the decomposition in one week. Specifically, Fig. 6a shows the original traffic data x_t in comparison with the periodic component s_t. From the results of Fig. 6a, we observe that although the original traffic data are very noisy, a one-day periodic component can be clearly seen which exhibits a two-peak daily pattern. Fig. 6b depicts the weekly trend component m_t, which clearly indicates that the traffic in weekdays is higher than the traffic in weekends. Fig. 6c shows the residual component r_t from the decomposition in one week. The residual component should reflects random events that suddenly increase or decrease the traffic data. In Fig. 7, We display the autocorrelation of the residuals over the whole four weeks, which is smaller than 0.1, indicating that r_t is close to a white noise.

3.2 Clustering

We aim to identify the key traffic patterns among 6,400 cellular towers according to both the seasonal and weekly trend components obtained from the above decomposition. As pointed out previously, this task is difficult for three reasons. First, we generally have no idea how many main patterns should be identified for the data from thousands of towers. Furthermore, the cellular towers are located in a large scale urban environment and the traffics of towers vary vastly from each other because of the differences in the numbers of users and their locations. Moreover, among all the towers, there are some 'bad' ones with incomplete traffic records. How to 'kick' out these outliers is also a challenge. We develop a system to accurately identify the key patterns from the traffic records in two steps: 1) determining the number of patterns, identifying the corresponding patterns, and finding the key patterns among them, and 2) labeling and validating these key patterns.

1) Identify the Patterns: Identifier is the key component of our mining system for recognizing the patterns from the traffic. We choose the hierarchical clustering [8] as our

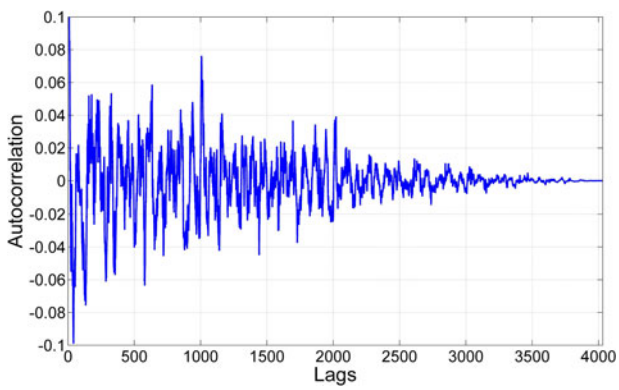
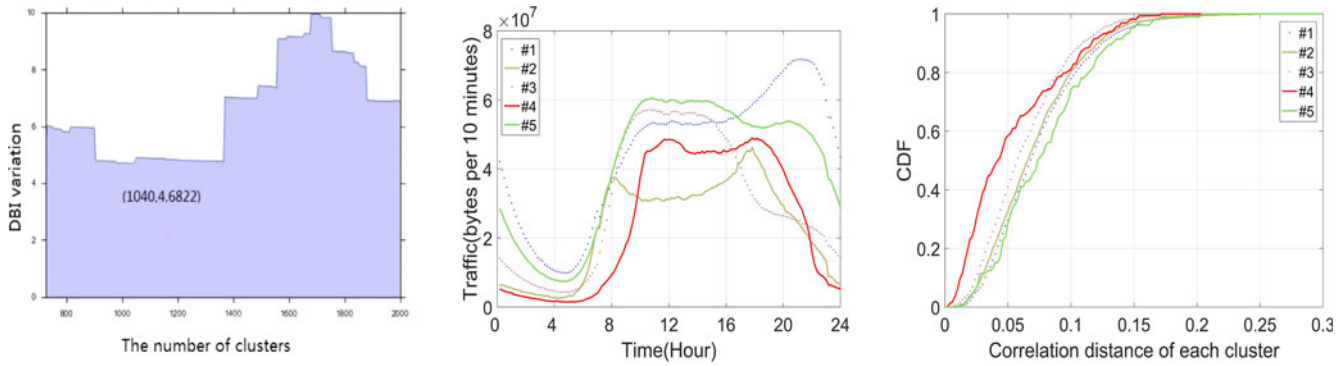


Fig. 7. Autocorrelation of the residuals.



(a) DBI variation (b) Key daily patterns (c) CDF of distance

Fig. 8. The DBI as the function of the number of clusters and the key daily patterns obtained by the clustering for the periodic component of the traffic.

identifier, because it does not require prior knowledge for the number of clusters. The basic idea of hierarchical clustering is first considering each input point as a cluster and then bottom-up iteratively merging the nearest two clusters until some terminating condition is met. The details of this hierarchical clustering are described in Algorithm 1.

Algorithm 1. Agglomerative Hierarchical Clustering

Input:
Number of base stations N, threshold value D, mobile traffic $\{X_i[t]\}$ for $i = 1, 2, \ldots, N$
Output:
Number of clusters n_C, labels L_i for $i = 1, 2, \ldots, N_C$
Initialize:
Number of clusters: $n_C \leftarrow N$
Clusters: $c_i \leftarrow \{X_i[t]\}$ for $i = 1, 2, \ldots, n_C$
stop $\leftarrow$ *false*
while *stop* == *false* **do**
  mindistance $\leftarrow \infty$
  for $i = 1$ **to** n_C **do**
    for $j = i + 1$ **to** n_C **do**
      distance $\leftarrow$ *compute_distance*(c_i, c_j)
      if $mindistance > distance$ **then**
        $mindistance \leftarrow distance$
        $index \leftarrow [i, j]$
      end if
    end for
  end for
  $n_C \leftarrow n_C - 1$
  $c_{index[1]} \leftarrow merge(c_{index[1]}, c_{index[2]})$
  $c_{index[2]}.delete()$
  if $n == 1$**or***mindistance* $> D$ **then**
    stop $\leftarrow$ *true*
  end if
  for $i = 1$ **to** n_C **do**
    $\forall X_i[t] \in c_i, L_i \leftarrow i$
  end for
end while
return n_C, and L_i for $i = 1, 2, \ldots, n_C$

It can be seen that when to terminate the clustering procedure is a key technical issue for this hierarchical clustering. We use Davies-Bouldin index (DBI) [9] to determine the optimum number of clusters. For the notational convenience, we will use vector notation to represent a sequence, e.g., the ith mobile traffic sequence $\{X_i[t]\}$ is also denoted as $\boldsymbol{X}_i$. With this notation convention, the DBI is defined as

$$\mathrm{DBI} = \frac{1}{R}\sum_{i=1}^{R} \max_{1\le j\le R, j\neq i} \frac{S_i + S_j}{M_{i,j}},$$

with

$$M_{i,j} = \|\boldsymbol{A}_i - \boldsymbol{A}_j\|_2 \quad \text{and} \quad S_i = \frac{1}{T_i}\sum_{k=1}^{T_i} \|\boldsymbol{X}_k - \boldsymbol{A}_i\|_2,$$

where $\boldsymbol{X}_i$ is the traffic data of the ith cellular tower, $\boldsymbol{A}_i$ is the centroid of the ith cluster, R is the number of clusters, and T_i is the numbers of towers within the ith cluster. When the minimum DBI is obtained, the optimum number of the patterns is identified. From all the patterns identified, we can determine the key patterns according to some specific criteria.

a) Periodic Component Patterns: For the periodic component of the mobile traffic, the DBI as the function of the number of clusters is shown in Fig. 8a, where the minimum DBI indicates that the optimum number of clusters is 1,040.

By considering the clusters containing more than 100 cellular towers as daily traffic patterns, we obtain the five key patterns from all the clusters. Fig. 8b displays these five key patterns identified by the clustering, which reflect typical traffic variations from 0:00 to 24:00 in one day. Observe that they all exhibit the low traffic hours around midnight to early morning when most people are sleep but they differ considerably in terms of the hours where peak traffic appears. Specifically, Pattern #1 reaches the peak mobile traffic in evening, Pattern #2 shows two rush hour peaks at around 8:00 and 18:00, and Pattern #3 has a lasting stable high traffic from 8:00 to 16:00, while Pattern #4 exhibits high traffic during daytime especially in lunch and dinner hours, and Pattern #5 appears to show the mixed features of the first four patterns. The cumulative distribution function of the correlation distance between the towers in each cluster and the cluster centroid is shown in Fig. 8c. All the curves reach almost 100 percent in a distance smaller than 0.2, indicating that almost all the towers of each cluster are sufficiently close to the cluster centroid. This confirms that the clustering result is reliable.

b) Trend Component Patterns: The weekly trend shows how mobile traffic fluctuates at the scale of one week. We use the same identifier discussed above to investigate the weekly trend component of the traffic. However, in this case, we have some priori knowledge regarding two obvious key patterns: people go to their workplaces at business districts for example in weekdays and the traffic reaches

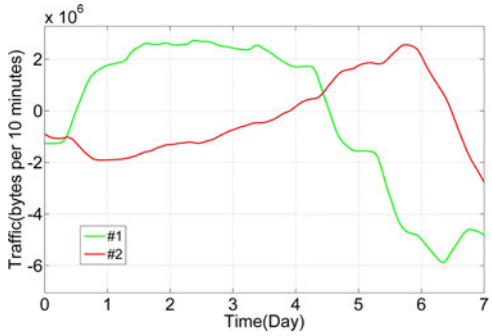


Fig. 9. Key weekly trend patterns.

high values in weekdays at these places, while in weekends, people go to entertainment places or stay in residential areas, and the traffic reaches peak values at these places in weekends. Therefore, we can set the number of the key patterns to two, and Fig. 9 depicts the two key patterns identified by the clustering for the weekly trend component of the traffic. It can be seen that Pattern #1 maintains high stable traffic during weekdays and goes down in weekends, while Patter#2 stays relatively low during weekdays and reaches the high peak values in weekends. These two weekly trend patterns clearly reflect human urban activities.

2) Label and Validate the Patterns: In order to relate the identified daily patterns with typical human urban activities, we first label the daily patterns by the specific urban function areas related to them. We find that the daily patterns can be related to four major urban functions and their mixture. Then we study the relationship between daily patterns and weekly trend patterns to further validate the daily patterns and weekly trend patterns identified.

a) Label the Daily Patterns: To understand the five daily traffic patterns of Fig. 8, we must obtain their geographical context so that we can relate them with typical human urban activities. For example, the daily pattern #2 in Fig. 8b has two peaks around 8:00 as well as between 17:00 and 18:00 that are typical rush hours in cities. Thus we may hypothesize that this daily pattern is related to the transportation function area in Shanghai. To label all the daily pattern accurately, we use points of interests (PoI) distribution to describe the geographical features of every pattern. Our PoI data are provided by one of the biggest online map service provider Baidu Map, which contains 23 kinds of PoI: restaurants, hotels, shopping centers, entertainment, sports, schools, tourist attractions, tourist development zone, finance areas, offices, corporates, factories, industrial areas, science park, economic development zone, high technology development zone, residential areas, living services, towns, villages, subways, overpasses.

For every daily pattern, we first compute the number of PoIs located within 200 m of each cell tower. Then we normalize the PoIs for every cluster. The PoI distribution of each pattern is summarized in Table 1. The average number of PoI is multiplied by 1000 to make it easy to read. The three highest types of PoI are marked by the orange color and ranked by the depth of the color for each daily pattern. It can be seen that the PoI distribution varies significantly between patterns and we can set the labels of patterns according to their main PoI types to: Residential Area, Transport Area, Office Area, Entertainment Area or Mixture Area.

TABLE 1
PoI Distribution

PoI	cluster1	cluster2	cluster3	cluster4	cluster5
Restaurant	4.44	16.1	7.88	30.7	5.65
Hotel	5.01	8.50	7.24	12.4	5.78
Shopping	3.47	14.0	5.67	32.6	4.235
Entertainment	4.82	14.1	7.93	24.6	6.19
Sports	3.2350	9.27	6.54	22.2	4.89
School	3.79	4.54	5.65	6.74	4.52
Tourism	1.44	10.4	5.26	13.8	2.75
Tourism Dev	0.50	0.00	0.00	0.00	0.70
Finance	2.75	13.6	10.5	18.235	5.21
Office	4.21	3.86	3.2350	4.88	2.80
Corporate	3.60	10.4	11.2	12.9	6.13
Science	1.2359	5.29	1.48	7.235	1.49
Factory	3.25	2.01	3.27	2.25	4.57
Industry	2.05	3.76	4.86	1.75	3.02
Tech Par	0.2359	1.51	3.01	0.48	1.79
Eco Dev	0.16	0.0	2.86	1.00	0.69
High Tech	0.09	0.0	2.76	0.0	0.13
Residential	7.46	4.2350	4.26	4.42	5.06
live Ser	6.51	12.4	9.46	18.4	7.44
Town	1.28	2.03	1.75	0.65	2.11
Village	4.53	1.26	2.85	4.03	4.03
Subway	1.88	25.0	4.21	10.4	3.67
Overpass	1.11	6.85	1.46	3.00	0.75

Residential Area. Table 1 shows that the main PoIs of towers in cluster#1 are resident and living service. Combining with the evening-high mobile traffic characteristics showing in Fig. 8b, we infer that this cluster can be labeled as residential area, where people go home from work in evening.

Transport Area. The subway station PoI is much bigger than others in cluster#2. As shown in Fig. 8b, the traffic of this cluster has two peaks around 8:00 and 18:00 when people are commuting between home and work places in morning and evening. Thus, this cluster can be labeled as transport area.

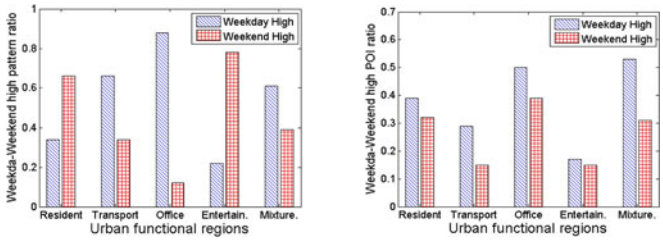
Office Area. The two highest PoIs for cluster#3 are corporates and finance. Further note that the traffic of this daily pattern stays high from '9 to 5'. Thus we can link this cluster with office areas in Shanghai.

Entertainment Area. The numbers of restaurant and shopping PoIs dominate cluster#4. The traffic of this cluster stays high between 10:00 and 20:00 with two peaks coinciding with lunch and dinner hours, as shown in Fig. 8b. These facts indicate that this cluster can be labeled as entertainment area.

Mixture Area. The PoIs for cluster#5 distribute much evenly among different functional regions, and the traffic of this daily pattern also exhibits the combined features of cluster#1 to cluster#4. Therefore we can consider this cluster as mixture area.

b) Label the weekly trend Patterns: The weekly trends have already been classified into two natural patterns: the weekday-high pattern and the weekend-high pattern. The former is related to business and work places, while the latter is associated with residential areas.

c) Relationship between Weekly Trend Patterns and Daily Patterns: To study the relationship between the daily patterns and weekly trend pattern, we compute the ratios of the two weekly trend patterns to each daily pattern, which are



(a) Weekly trend patterns ratio (b) PoI ratio

Fig. 10. Relationship between weekly trend patterns and daily patterns.

depicted in Fig. 10a. Observe that in residential area (daily pattern#1) and entertainment area (daily pattern#4), the weekend-high trend pattern occupies 66 and 78 percent of towers, respectively, while the weekday-high trend pattern occupies nearly 90 percent of office area (daily pattern#3) and over 60 percent of transport area (daily pattern#2). These results agree with our understanding of human urban activities, and thus they further validate our classification.

As explained previously, we have five functional regions, and for each region we have a list of 23 POI values. Furthermore, among these 23 different POIs, some are typical weekday-high POIs (Finance, Office, Corporate) and some are typical weekend-high POIs (Residential, Live Service, Entertainment). The two ratios plotted in Fig. 10b for each functional regions are the ratio of the sum of the weekday-high POI values over the sum of all the POI values and the ratio of the sum of the weekend-high POI values over the sum of all the POI values, respectively. The results of Fig. 10b clearly validate our pattern classification.

3.3 Forecasting

We now build a mobile traffic forecasting system based on the result of clustering. Considering the nonstationarity and periodicity of the mobile traffic series, we adopt the seasonal autoregressive integrated moving average (SARIMA) model [10] to achieve this goal. The general form of SARIMA model is denoted as $\mathrm{ARIMA}(p,d,q)\times(P,D,Q)_{m'}$, where m' is the period of the time series, (P,D,Q) represents the seasonal part of the model with P as the number of seasonal autoregressive terms, D as the number of seasonal differences and Q as the number of seasonal moving average terms, while (q,d,q) denotes the non-seasonal part of the model with q as the number of non-seasonal autoregressive terms, d as the number of non-seasonal differences and q as the the number of non-seasonal moving average terms.

In our study, we find that we can accurately identify an SARIMA model $\mathrm{ARIMA}(1,0,1)\times(1,1,1)_{1008}$ from the mobile traffic series $\{x_t\}_{t=1}^{4032}$ as

$$\begin{aligned}&\big(1-ar\cdot B\big)\big(1-sar\cdot B^{1008}\big)\big(1-B^{1008}\big)x_t\\&=\big(1+ma\cdot B\big)\big(1+sma\cdot N^{1008}\big)a_t,\end{aligned}$$

where a_t is the white noise series the period is $m'=1008$, and B denotes the backward shift operator, namely, $B\cdot x_t=x_{t-1}$, while the parameters ar, sar, ma and sma are determined by the least squares estimates.

By adopting the above model, we predict the forth week's traffic with the previous three weeks' traffic as input for the four daily patterns separately. For better viewing, the logarithmic traffic series is used, and the forecasting results are shown in Fig. 11. It is clear that the forecasting series fit the real logarithmic traffic series well in all the four patterns.

We next quantitatively evaluate the prediction accuracy. Using resident area (rsd) as an example, let N_{rsd} be the number of towers labeled as resident area and $\mathcal{N}_{\mathrm{rsd}}$ be the corresponding index set. For each tower k, we have a real logarithmic traffic data series $\{x_1^{(k)},x_2^{(k)},\ldots,x_{1008}^{(k)}\}$ with the mean value of $\mathrm{mean}^{(k)}$. Then the mean value of the real traffic data labeled as resident area is calculated as

$$\mathrm{Mean}=\frac{1}{N_{\mathrm{rsd}}}\sum_{k\in\mathcal{N}_{\mathrm{rsd}}}\mathrm{mean}^{(k)}.$$

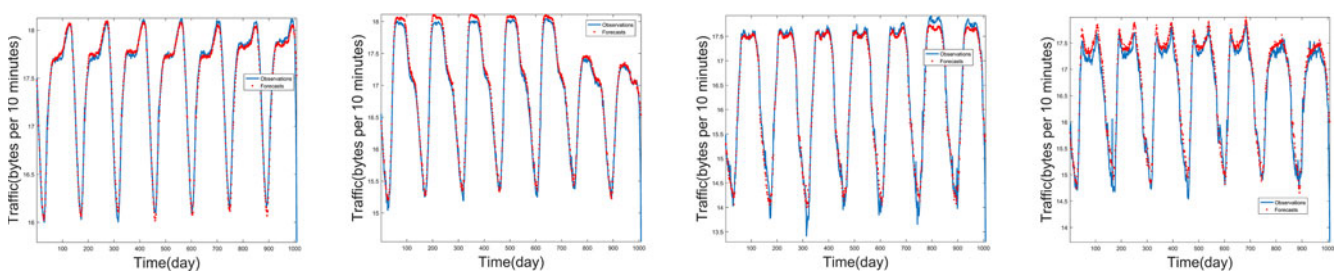
Using the model built from the classified data, we obtain the forecasting series $\{\widetilde{x}_1^{(k)},\widetilde{x}_2^{(k)},\ldots,\widetilde{x}_{1008}^{(k)}\}$ for tower k, which has the mean square error (MSE) of $\widetilde{\mathrm{MSE}}^{(k)}$. Thus the MSE of the classified model prediction for resident area is given by

$$\mathrm{MSE}=\frac{1}{N_{\mathrm{rsd}}}\sum_{k\in\mathcal{N}_{\mathrm{rsd}}}\widetilde{\mathrm{MSE}}^{(k)}.$$

Using the model built from all the traffic data (unclassified), we obtain the forecasting series $\{\widehat{x}_1^{(k)},\widehat{x}_2^{(k)},\ldots,\widehat{x}_{1008}^{(k)}\}$ for tier k, which has the MSE of $\widehat{\mathrm{MSE}}^{(k)}$. The MSE of the unclassified model prediction for resident area is given by

$$\widehat{\mathrm{MSE}}=\frac{1}{N_{\mathrm{rsd}}}\sum_{k\in\mathcal{N}_{\mathrm{rsd}}}\widehat{\mathrm{MSE}}^{(k)}.$$

We summary the MSEs of the classified model predictions for the four daily traffic patterns in the first row of Table 2. The ratios of these four MSEs over the corresponding mean values of the four daily-pattern traffic series are listed in the second row of Table 2, while the ratios of the four $\widehat{\mathrm{MSE}}$ values over the corresponding mean values of the four daily-pattern traffic series are given in the third row of Table 2.



(a) Residential area (b) Office area (c) Entertainment area (d) Transport area

Fig. 11. Comparison of forecasting (red dashed) and observation (blue solid) for daily patterns of four functional regions.

TABLE 2
Forecasting Errors for Daily Patterns

	resident area	office area	entertainment area	transport area
MSE (classified model prediction)	0.086293	0.11703	0.17743	0.21324
MSE/Mean (classified model prediction)	0.49%	0.69%	1.08%	1.28%
$\widehat{MSE}$/Mean (unclassified model prediction)	2.26%	2.55%	6.42%	3.93%

In conclusion, we integrate mobile traffic forecasting as an application to our system and significantly improve the performence of ARIMA model by utilizing the labels of cellular towers obtained in Clustering section. Based on four unique daily patterns, we train four different predicting models. For input traffic data from one tower, we choose model according to the label of this specific tower. Experiment on our dataset shows that by this means we get much better accuracy of forecasting.

3.4 Learning from the Residuals

We now present our system to detect whether an accident or anomalous event happened for a cellular tower from its residual component of the traffic data. In Fig. 12, we depict the residual and the user number of a selected cellular tower around HongKou gymnasium. Observe that the number of users in Fig. 12b has three obvious peaks corresponding to three crowd events, but the residual component in Fig. 12a is so noisy that we cannot identify these three events directly from the residual component. The same problem can be seen by comparing Fig. 12c with Fig. 12d, where the number of users exhibits a peak in the second Saturday but it cannot be clearly identified from the residual component.

To eliminate the effect of the noise, which reflects meaningless random events in the residual component, we set a threshold equaling to 4 standard deviations away from the mean value for each residual series. We consider that an unusual event is happening when the residual component exceeds the threshold and lasts for half an hour. Noting the sampling period of 10 minutes, half an hour corresponds to the three consecutive samples. Thus, given a residual series $\{r_1, r_2, \ldots, r_{4032}\}$, we set the threshold $r_{thr} = mean(r_i) + 4std(r_i)$, where $mean(r_i)$ denotes the mean value of $\{r_i\}$ and $std(r_i)$ stands for the standard deviation of $\{r_i\}$. For $j = 1, 2, \ldots, 4030$, if

$$\min\{r_j, r_{j+1}, r_{j+2}\} \geq r_{thr},$$

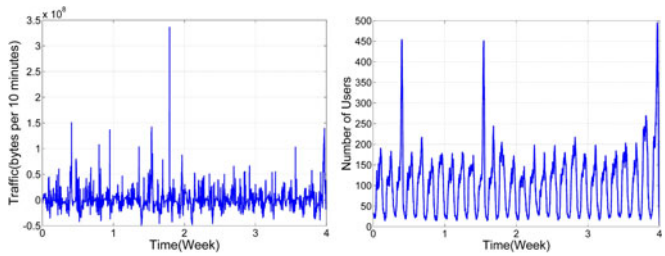
an anomalous event is considered happening during the time period between j and $j+2$.

We first find three cellular towers located around gymnasiums or other public places where crowd events occur to test our anomalous event detection method based on the residual component of mobile traffic. In Table 3, the RD (residual detection) column indicates whether the crowd event is detected from the corresponding residual series. NO.1 tower is located around Hongkou Stadium, NO.2 tower is located around Mercedes Benz Cultural Center, and NO.3 tower around Luwan Gymnasium. The anomalous events are soccer matches or concerts that cause a sudden increase in the number of users. It can be seen from Table 3 that most of the events are correctly detected by our method, which shows that our method is accurate and reliable.

We then apply our system to detect anomalous events for all the cellular towers. Fig. 13 shows the spatial distribution

TABLE 3
Event Detection

tower	crowd time	RD	event name
NO.1	08.6 19:00	no	CFA soccer match
	08.14 19:00	yes	CSL soccer match
	08.31 18:30	yes	CSL soccer match
NO.2	08.16 20:00	yes	We are family concert
	08.18 19:30	yes	unknown
	08.23 19:30	yes	Daphne concert
	08.30 18:30	yes	Michael Wong concert
NO.3	08.9 19:30	yes	Lee Min Woo concert
	08.16 19:20	yes	Shila Amzah concert



(a) Residual of tower 1 (b) Number of users of tower 1

(c) Residual of tower 2 (d) Number of users of tower 2

Fig. 12. Residual and user number of the selected cellular tower.

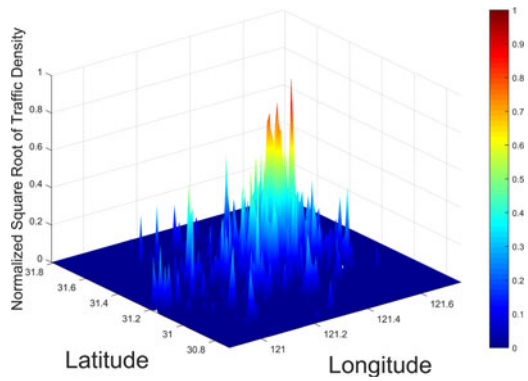


Fig. 13. The spatial distribution of anomalies.

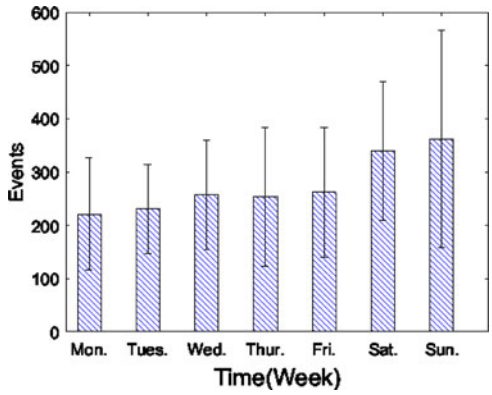


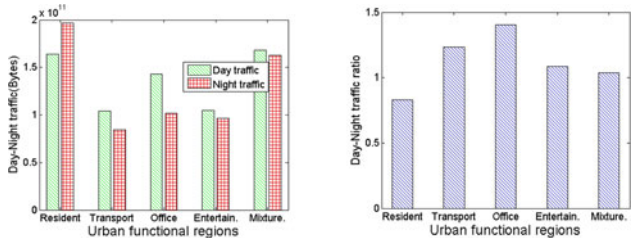
Fig. 14. Anomalous events in four weeks.

of unusual events in the whole month. As we can observe from the heat map, most anomalous events occur around the center of the city, and their distribution shows a great similarity with the distribution of mobile traffic consumption depicted in Fig. 4, which indicates that anomalies often occur where mobile traffic are heavy. The means and standard deviations of anomalies occurring in four weeks are shown in Fig. 14. The mean values in weekends are obviously higher than those in weekdays, which is reasonable because events such as concerts are more likely to happen in weekends. Another observation is that the standard deviation is the largest in Sunday. This result indicates that the occurring of anomalous events is not regular but varying in different weeks. It can be seen that both the spatial and temporal distributions of the anomalies detected agree with human urban activities.

As the major applications of our system, traffic forecasting and anomaly detecting can also work as online services. Given a set of cellular towers and initial data set, our system is able to model and label each cellular tower. Considering input traffic data from one specific tower, we can choose predicting model according to the label and shift model input data to latest three weeks (or other time span) traffic to forecast following sequences. For anomaly detecting, our system sets parameters based on historical data. With effective decomposition method, our system can decompose latest traffic data every few minutes (similarly pick up latest one week data as input) and detect anomalies in real time based on the noise component. Thus, based on a historical mobile traffic dataset, our system is able to provide online traffic predicting and anomaly detecting services.

4 Result and Observation

So far we have identified five daily patterns from the regular periodic components of our traffic data and extracted two natural weekly trend patterns from the trend components. We now explore the physical context or human activities associated with these traffic patterns.



(a) Day-night traffic amount (b) Day-night traffic ratio

Fig. 15. Day-night traffic amount ratio.

4.1 Understanding Daily Patterns

1) Peak-valley time: An obvious feature of a daily traffic pattern is its peak and valley times within a day. In Table 4, we list the peak and valley times of each daily pattern. According to Table 4, for all the patterns, the valley times appear between 4:00 am and 5:00 am when most users are soundly sleeping. Residential area has peak time at 21:00 when people come home in night, and in transport area, two peaks appear at 8:00 am and around 5:00 pm to 6:00 pm, reflect two rush hours in a day. Work area does not have a clear peak time but have a lasting high traffic level during the working time. In entertainment area there are two peaks around 12:00 am and 6:00 pm, which are regular times for lunch and dinner.

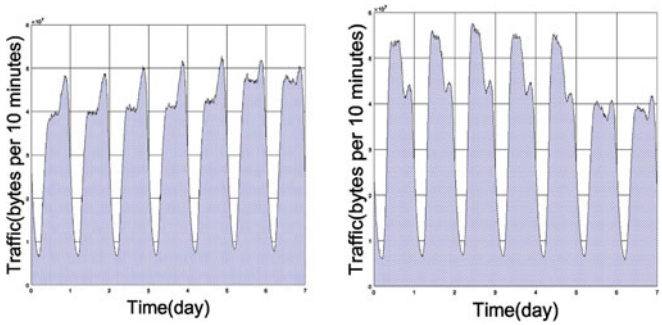
2) Day-night traffic amount ratio: For all the daily patterns, Fig. 15a shows the traffic amount between 7:00 am and 7:00 pm, which is considered as day-time traffic, and the traffic amount in the rest of time, which is regarded as night-time traffic. It is clear that the traffic amount in day time is higher than the traffic amount in night time for all the patterns except residential area. This agrees well with human daily activities. Observe from Fig. 15b that there are notable differences for the ratios of day-time traffic amount to night-time traffic amount among different patterns. Specifically, in residential area, the ratio is around 0.8 which is much lower than the ratios in transport, office and entertainment areas. This can be well explained by human daily activities–people go to work or to entertainment places in day time and return home in night time. In office area, the ratio is the highest up to 1.4, indicating that most people work during day time.

4.2 Understanding Weekly Trend Patterns

The weekly trend components show two weekly tendencies: the weekday-high one and the weekend-high one. In Fig. 16 we show the average full traffic of these two weekly patterns in one week. We can clearly observe from Fig. 16a that the traffic is higher in weekdays than in weekends for the weekend-high pattern, and the opposite is true for the weekday-high pattern as can be seen from Fig. 16b. We further notice from Fig. 16a that the traffic in weekdays shows a peak in evening, which indicates that the main daily pattern contributed to the weekend-high pattern is the residential pattern. In Fig. 16b, the traffic in weekdays is

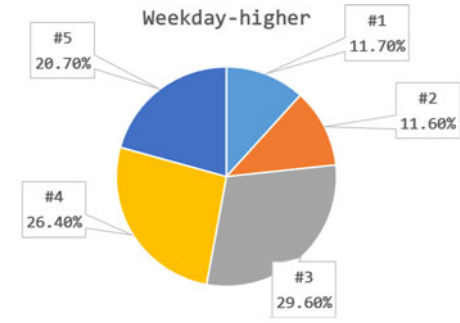
TABLE 4
Peak-Valley Time of Daily Patterns

–	resident area	transport area	office area	entertainment area	mixture area
peak time	21:00	8:00, 17:00-18:00	9:00-16:00	12:00,18:00	10:00
valley time	4:00-5:00	4:00	5:00	5:00	4:00-5:00

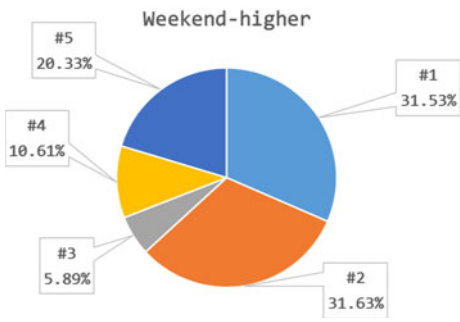


(a) Weekend-high pattern (b) Weekday-high pattern

Fig. 16. Weekly trend components in one week.



(a) Weekday-high pattern



(b) Weekend-high pattern

Fig. 17. Interaction between weekly trend patterns and daily patterns.

significantly higher in working hours, from which we can infer that the main daily pattern contributed to the weekday-high pattern is the office pattern.

In Fig. 17, we show the proportions of each daily pattern in the two weekly trend patterns. From Fig. 17a we can conclude that among the four basic daily patterns (#1 to #4), the largest contributor to the weekday-high pattern is the office pattern (#3), which occupies 29.6 percent of the towers, and the second largest contributor is the entertainment pattern (#4), which occupies 26.4 percent of the towers. The office pattern as the largest contributor to the weekday-high pattern completely makes sense as this reflects the main people's activities during weekdays, while the entertainment pattern contributes heavily to the weekday-high pattern partly because lunch and dinner are people's daily necessity and partly because there are large number of towers in entertainment area. By contrast, from Fig. 17b, we find that the transport pattern (#2) and the residential pattern (#1) are the two biggest contributors to the weekend-high pattern, which accounts for 31.6 and 31.5 percent of the towers, respectively. This clearly reflects people's main activities during the weekends. Interestingly, the mixture pattern (#5) contributes equally to the weekday-high and weekend-high patterns, accounting for about 20 percent of the towers in each case. This reflects the nature of mixture area.

5 Related Work

A large number of investigations have focused on revealing urban environment dynamics and social events from digital footprints [15]. In this section, we categorise the relevant works in the literature from four aspects: traffic data analysis enabled applications, digital footprints used for discovering urban dynamics, time series methods adopted for analyzing mobile traffic data, and event detection from mobile traffic.

Cellular traffic data have been utilized for a wide range of applications. The digital records can be used to predict personal attributes, including sexual orientation, ethnicity, religious and political views [5]. CDRs have been used to model human mobility patterns [3], [11], which reveal that human trajectories show a high degree of temporal and spatial regularity [3] and high potential predictability [11]. The study [4] aims to reveal the usage patterns of mobile data users by investigating 3G cellular networks and it finds that majority of data usage in the network are contributed by a small number of heavy users. The works [12], [13] infer and classify land usage from the CDR data. Other applications that are enabled by utilizing cellular network traces include inferring friendship network structure [14], understanding mobile user browsing behavior [14], and optimizing the content delivery based on user's location [19].

Three different kinds of digital footprints are commonly used for revealing urban human activity patterns: CDRs, social media data and mobile traffic data. CDRs are exploited to model individual human activity patterns [3], [21] and estimate population distribution[21]. However, compared to mobile traffic data, CDRs are sparse temporally. Due to the boom of mobile Internet mobile social instant communication applications have replaced phone calls to become the primary method of communication in urban areas. People rarely use phone call but chat on communication applications more instantaneously. Furthermore, mobile Internet has prevailed in every aspect of urban social life with the popularity of mobile payments and urban life produces mobile traffic consumption moment to moment. Thus, mobile traffic data contains much richer information regarding human activities. The work [22] proposed a method to detect urban events based on social activity dataset and GPS trip records. Social media data reveal individual activities directly. For example, Twitter messages literally show urban events. However, compared to mobile traffic data, social media data are hard to model and mine since they are mostly in word, audio or video format and contain massive redundant information. On the other hand, mobile traffic data also preserve citizen's privacy better because they are aggregation of individuals' user data. The work [20] built a system to classify service usages using encrypted Internet traffic data of mobile message Apps, which provides another application of mining the contexts and behavior information from mobile traffic data.

There have been several investigations of urban human activity trace decomposition. [17], [22] propose a non-negative tensor factorization approach to decompose a

human activity tensor into basic life pattern tensors. However, a disadvantage of this approach is that the number of basic patterns has to be set in advance while our system is able to decide the number of patterns automatically. [18] built a 3D matrix to represent the time, location and probability of a social event based on a probabilistic model, and then applied an image segmentation algorithm to mine social event from the matrix. Compared to our system, this work did not decompose the human traces explicitly and ignored other information such as daily pattern and long-term trend. The study [2] infers features of urban ecology from spatial-temporal cell phone activity data by decomposing the original cell phone activity series into the seasonal communication series and the residual communication series. Unlike the decomposition method adopted in this paper, the work [2] decomposes the cell phone traffic series by first performing the frequency-domain transformation on it using FFT, and then extracting the main frequency components. This work ignores the trend component of mobile traffic, which is proved reflecting a long-term fluctuation in our paper. Time series analysis is widely adopted for mobile traffic data analyzing in various situations, especially in mobile traffic forecasting. A method is proposed for forecasting traffic based on multiple regression model for time-series [23], while the work [24] models and predicts actual wireless traffic, such as GSM traffic, using seasonal ARIMA models. We also adopt a seasonal ARIMA model and we adjust the model parameters according to real daily patterns. Furthermore, we show that we can forecast mobile traffic consumption accurately using our model.

Cellular traffic has been extensively used for detecting anomalous events. The work [25] studies the occurrence of unexpected events using phone records and it applies standard percolation theory tools to describe these spatio-temporal anomalies. Crowd mobility during special events is studied in [26], which analyzes nearly 1 million cell-phone traces and it finds that the origins of people in an event are correlated to the type of event. The study [27] explores societal response to external perturbations, especially emergencies like bomb attacks and earthquakes, by identifying real-time changes in communication and mobility patterns. In our work, we extract the underlying causes of unusual events by decomposing the mobile traffic series and our anomalous event detection method has been shown to successfully detect irregular activities, such as concerts and matches, from the mobile traffic.

In summary, differing from the existing works, we study and model both urban functional regions and human activities based on large-scale cellular mobile traffic collected by ISP and propose an effective framework. The novelty of our work can be summarized from three aspects: First of all, we utilize a large-scale mobile traffic dataset. Compared to other digital records, our dataset has better temporal and spatial grain and also better reflect urban dynamics in a mobile era. Second, we creatively adopt a times series decomposition method to investigate mobile traffic data from different perspectives. On one hand, we model urban human regular activity patterns from different time scales using the periodic and trend components. On the other hand, we study unusual activities according to the residual component where the influence of unexpected events is buried in noise. To the best of our knowledge, this method has not been used in mobile traffic analysis in the open literature, and it provides a solid understanding of interactions between human activities and network dynamics. Lastly, we propose a framework combining human activity pattern mining, mobile traffic forecasting and anomaly detection, which provides a systematic workflow to investigate mobile traffic records data.

6 Conclusions

In this paper, we have studied the inherent correlation between the mobile data traffic and human activities in urban environment based on large-scale, well-grained mobile traffic records dataset in a systematic and comprehensive way. By adopting a generic time series decomposition method, we have designed a powerful system that integrates traffic pattern clustering and labeling, mobile traffic forecasting and event detection in one workflow. First, we decompose the traffic series naturally into three components: daily periodical component, weekly trend component and residual component. Then, we identify five fundamental daily activity patterns closely connected with different human activity functional areas. We introduce ARIMA model to forecast mobile traffic, and based on different patterns we train different models, which improves the performance of traffic predicting significantly. Furthermore, we recognize the two weekly trend patterns from the trend component, which again reflect the underlying week-long human activities in the real-world. Lastly, we use the residual component to detect anomalous events caused by irregular human activities, and we show our anomalous event detection method can accurately detect irregular human activities from the real-world noisy residual-component series. Meanwhile, our system can not only deal with massive offline mobile traffic database but also provide online traffic forecasting and event detecting services. Our study has thus provided an effective framework to deal with massive mobile traffic data and a comprehensive understanding of the intimate relationship between mobile data traffic consumption and human activities in urban environment.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China under 61861136003, 61621091 and 61673237, and research fund of Tsinghua University—Tencent Joint Laboratory for Internet Innovation Technology. A conference version of this paper was been presented in IEEE MASS 2016.

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, White Paper. 2015–2016. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11–520862.html

[2] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, "On the decomposition of cell phone activity patterns and their connection with urban ecology," in *Proc. 15th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jun. 2015, pp. 317–326.

[3] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.

[4] Y. Jin, et al., "Characterizing data usage patterns in a large cellular network," in *Proc. CellNet Workshop*, Aug. 2012, pp. 7–12.
[5] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Academy Sci. United States America.*, vol. 110, no. 15, pp. 5802–5805, 2013.
[6] 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.
[7] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Berlin, Germany: Springer Science & Business Media, 2006.
[8] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic Acids Res.*, vol. 16, no. 22, pp. 10881–10890, 1988.
[9] 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.
[10] R. S. Tsay, *Analysis of Financial Time Series*. Hoboken, NJ, USA: Wiley, 2005.
[11] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Sci.*, vol. 327, no. 5968, pp. 1018–1021, 2010.
[12] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," *Int. J. Geographical Inform. Sci.*, vol. 28, no. 9, pp. 1988–2007, 2014.
[13] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2012, pp. 1–8.
[14] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proc. Nat. Academy Sci.*, vol. 106, no. 36, pp. 15274–15278, 2009.
[15] D. Zhang, B. Guo, and Z. Yu, "The emergence of social and community intelligence," *IEEE Comput.* vol. 44, no. 7, pp. 21–28, Jul. 2011.
[16] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Contextual localization through network traffic analysis," in *Proc. IEEE Conf. Inf. Comput. Commun.*, Apr. 2014, pp. 925–933.
[17] Z. Fan, X. Song, and R. Shibasaki, "CitySpectrum: A non-negative tensor factorization approach[C]," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, 213–223.
[18] W. Zhang, G. Qi, G. Pan, H. Lu, S. Li, and Z. Wu, "City-scale social event detection and evaluation with taxi traces," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, Art. no. 40, 2015.
[19] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Contextual localization through network traffic analysis," in *Proc. IEEE Conf. Inf. Comput. Commun.*, 2014, pp. 925–933.
[20] Y. Fu, H. Xiong, X. Lu, J. Yang, and C. Chen, "Service usage classification with encrypted internet traffic in mobile messaging apps [J]," *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2851–2864, Nov. 2016.
[21] S. Jiang, J. Ferreira, and M. C. González, "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 208–219, Jun. 2017.
[22] L. Chen, J. Jakubowicz, D. Yang, D. Zhang, and G. Pan, "Fine-grained urban event detection and characterization based on tensor cofactorization," *IEEE Trans. Human-Mach. Syst.* vol. 47, no. 3, pp. 380–391, Jun. 2017.
[23] Y. Akinaga, S. Kaneda, N. Shinagawa, and A. Miura, "A proposal for a mobile communication traffic forecasting method using time-series analysis for multi-variate data," in *Proc. IEEE Global Commun. Conf.*, 2005, pp. 1119–1124.
[24] Y. Shu, M. Yu, J. Liu, and O. W. W. Yang, "Wireless traffic modeling and prediction using seasonal ARIMA models," in *Proc. IEEE Int. Conf. Commun.*, 2003, pp. 1675–1679.
[25] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *J. Phys. A: Math. Theoretical*, vol. 41, no. 22, pp. 1–11, 2008.
[26] F. Calabrese, F. C. Pereira, G. D. Lorenzo, L. Liu, and C. Ratti, "The geography of taste: Analyzing cell-phone mobility and social events" in *Proc. 8th Int. Conf. Pervasive Comput.*, May 2010, pp. 22–37.
[27] J. P. Bagrow, D. Wang, and A.-L. Barab asi, "Collective response of human populations to large-scale emergencies," *Plos One*, vol. 6, no. 3, pp. 589–589, 2011

Mingyang Zhang is currently working toward the BE degree with the Tsinghua University, Beijing, China. His research interests include mobile big data and data mining.

Yong Li (M'09-SM'16) received the BS degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, and the PhD degree in electronic engineering from the Tsinghua University, Beijing, China, in 2007 and 2012. He is currently a faculty member of the Department of Electronic Engineering, Tsinghua University. He has served as general chair, TPC chair, TPC member for several International Workshops and Conferences, and he is on the editorial board of three International Journals. His papers have total citations more than 2300 (six papers exceed 100 citations, Google Scholar). Among them, eight are ESI Highly Cited Papers in Computer Science, and four receive conference Best Paper (run-up) Awards. He received IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers and Young Talent Program of China Association for Science and Technology. He is a senior member of the IEEE.

Sheng Chen (M'1990-SM'1997-F'2008) received the BEng degree from the East China Petroleum Institute, China, and his PhD degree from the City University, London, in January 1982 and September 1986, both in control engineering. In 2005, he was awarded the higher doctorate degree, Doctor of Sciences (DSc), from the University of Southampton, Southampton, United Kingdom. From 1986 to 1999, he held research and academic appointments with the Universities of Sheffield, Edinburgh and Portsmouth, all in United Kingdom. Since 1999, he has been with Electronics and Computer Science, the University of Southampton, United Kingdom, where he holds the post of professor in Intelligent Systems and Signal Processing. His research interests include adaptive signal processing, wireless communications, modelling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods and optimisation. He has published more than 550 research papers. He is a fellow of the IET. He is a distinguished adjunct professor with the King Abdulaziz University, Jeddah, Saudi Arabia. He is an ISI highly cited researcher in the engineering category (March 2004). He was elected to a fellow of the United Kingdom Royal Academy of Engineering in 2014. He is a fellow of the IEEE.