ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com


ISSN: 2321-21
IJMEC
International Journal of mode
electronics and communication eng
E-Mail
editor.ijmece@gmai
editor@ijmece.co
www.ijme

ISSN2321-2152www.ijmece .com

Vol 6, Issuse.2 May 2018

Massive Mobile Traffic Data: A Window In to Urban Dynamics

Ch Venkatesh, S Rajeshwar, B Bhavani, DASARI SHAILAJA

Abstract—

Understanding mobile data consumption patterns is crucial for learning about urban ecosystem and human activity. This task is challenging in the sense that

The complexity of mobile data usage in vast metropolitan environments, the disruption of unusual events, and the absence of prior understanding of urban traffic patterns are the three problems. We suggest a fresh method for creating a strong system that consists of three subsystems: time series decomposition of mobile traffic data, pattern extraction from various elements of the original traffic, and anomalous event detection from noises.

Three significant findings come from our examination into the mobile traffic data of 6,400 cell towers in Shanghai.

First, we find five daily patterns among the 6,400 cellular towers that correlate to various human daily activity patterns.

Introduction I.

Massive amounts of mobile traffic data are used as a result of the widespread availability of LTE and 4G networks.

In the last ten years, the amount of mobile data traffic has increased by 4,000 times, and our society is currently dealing with a remarkable acceleration in the expansion of cellular data traffic. In 2015, there were 3.7 exabytes of monthly mobile data traffic, and by 2020, that number is predicted to rise to 30.6 exabytes. As a result, studying mobile cellular traffic becomes a crucial method for comprehending human behaviour and urban environment. However, our understanding of how people's routine activities and unplanned occurrences impact the mobile traffic of cellular towers is relatively restricted [2]. Such information is quite important.

M. Zhang and Y. Li work for Tsinghua University's Department of Electronic Engineering, which is located at the Tsinghua National Laboratory for Information Science and Technology

ASSISTANT PROFESSOR[1,2,3], STUDENT[4]
Department of CSE
Arjun College Of Technology & Sciences
Approved by AICTE& Affiliated to JNTUH
SPONSORED BY BRILLIANT BELLS EDUCATIONAL SCOITEY

liyong07@tsinghua.edu.cn and mingyangzhang@tsinghua.edu.cn). The Princeton International School of Mathematics and Science employs H. Fu.
S. Chen is affiliated with King Abdulaziz University in Jeddah, Saudi Arabia, as well as the Department of Electronics and Computer Science at the University of Southampton in Southampton, UK (E-mail: sqc@ecs.soton.ac.uk).
This research was funded by the Tsinghua University Research Fund, the National Natural Science Foundation of China (Nos. 61301080, 91338203, 91338102, and 61321061), as well as the National Basic Research Program of China (973 Program) (No. 2013CB329001) (No 20161080099)
Determine the placements of the cellular towers based on traffic patterns and apply suitable tactics for the towers of various patterns to reduce traffic loads during peak hours. In
Additionally, if a technique can be created to precisely identify abnormalities in cellular traffic data, it will assist ISP in identifying equipment failures or unexpected crowd occurrences so that they may take appropriate action to minimise possible loss. Fundamentally, accurately recognising mobile traffic patterns is crucial for comprehending human behaviour, which may be used to improve infrastructure and living conditions.
An appropriate dataset for analysing urban people activity is cellular network record. Today's mobile lifestyle frequently involves using a cellular network to access the internet. In our metropolitan lives, we constantly use cellphone data to check out in. Using smartphone applications to access stores, hail a cab using taxi hailing services, connect with pals on social media, etc. Call description records (CDRs), which are more detailed,

Due to the increasing frequency of access, mobile traffic data records are also frequently employed to reveal human life patterns in cities. In contrast, there is only 200–300 m between two nearby cellular towers in metropolitan areas, so mobile traffic data likewise offers high spatial granularity. Another significant data source that is frequently generated in metropolitan areas is social media data.
Social media data are difficult to gather and exploit due to the variety of platforms and formats. For instance, when a route is busy, the nearby cellular towers' mobile traffic increases.
Three factors make it difficult to identify the traffic patterns of large-scale cellular towers. The first reason why the traffic around cellular towers is complicated is because

Towers vary greatly from one another. Furthermore, even the traffic from a single cell tower exhibits various patterns over a range of time periods. Finding a method that can examine it universally is challenging because of this. However, we must discover a way to describe these variations and create a model that can take into account varied contexts. Second, unintentional incidents have an impact on the traffic at cellular towers, which further complicates analysis by adding complication. For instance, a cellular tower's traffic volume will increase suddenly and depart substantially from its usual patterns when a parade happens surrounding it.

It is so challenging to figure out how to reduce the impact of these unintentional occurrences on pattern analysis and to identify anomalous events. Thirdly, it is challenging to choose the right number of patterns and to recognise their significance

since we have limited prior knowledge about the traffic patterns of cellular towers.

Although it hasn't been fully edited, this paper has been accepted for publication in a subsequent edition of this magazine. Before the final publishing, the content may change. TBDATA.2017.2778721, IEEE Transactions on Big Data, DOI for citation

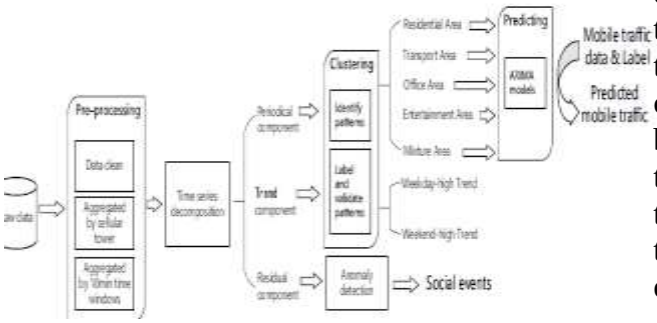


Fig. 1. Framework overview.

features. On the one hand, there may be an enormous number of patterns due to the high density of cellular towers, yet some of them are not particularly helpful in analysing human urban activity. In contrast,

However, some mobile phone towers exhibit a combination of traits from several designs. Due to the volume of communication between these cellular towers, it is difficult to discern the few key patterns that are buried.

We create a powerful system to decode and model data on mobile traffic from thousands of cellular towers in order to overcome these difficulties. Our suggested framework is shown in broad strokes in Fig. 1. Data preprocessing, time-series decomposition, pattern modelling, anomalous event detection, and traffic prediction make up our system's five components. It can handle massive amounts of data. component components include residual, seasonal, and trend. An irregular tendency across time is represented by the trend component.

The residual component is thought of as sounds or odd events on a larger time scale, while the seasonal component shows a periodic variation that typically corresponds to routine activity.

We create a model of human activity patterns based on seasonal factors and present a technique for predicting mobile data traffic. Additionally, we provide a technique to identify anomalous occurrences from traffic data and verify the outcomes using actual traces. In order to reduce the impact of unusual eventsWe create a model of human activity patterns based on seasonal factors and present a technique for predicting mobile data traffic. Additionally, we provide a technique to identify uncommon occurrences from

the traffic logs and verify the findings using actual traces. We breakdown the original mobile traffic data and extract the primary traffic patterns by utilising hierarchical clustering [8], which does not need a specified number of patterns, in order to remove the effect of anomalous occurrences and model the traffic pattern from several angles.

As a result, we may use the residual component to identify unexpected occurrences.

We obtain the following intriguing results by using our method to look at the mobile traffic records of 6,400 cellular towers, which were gathered by ISP from Shanghai.

According to the one-day seasonal component of their traffic consumptions, the cellular towers can be divided into five groups, and these groups correspond to five different types of urban function

areas: the residential area, the transportation area, the office area, the entertainment area, and the mixture area. This discovery demonstrates how human activity patterns affectWe introduce the ARIMA model, which forecasts mobile traffic with great accuracy, demonstrating the importance of this result.

In order to detect the two main trends—which alternately rise and decrease during the week—we may model the traffic patterns with the weekly trend component. These two weekly trend patterns are a reflection of actual, week-long human activity.
Our analysis demonstrates that the discovered anomalous events match the anomalies with actual occurrences, demonstrating that odd events may be identified from the residual component utilising our suggested technique of anomalous event identification.

The remainder of this essay is organised as follows. We discuss the used dataset and explain our rationale in Section II.
We outline each element of our system in Section III.

We specifically outline our decomposition and grouping methods, and by talking about the relationship between daily patterns and weekly trend patterns, we create a model for projecting mobile traffic consumption. Then, using the residual component, we provide a strategy to identify anomalous events. We analyse the results in Section IV, and after reviewing the related work in Section V, we wrap up this study in Section VI.

II. MOTIVATION AND DATA SET

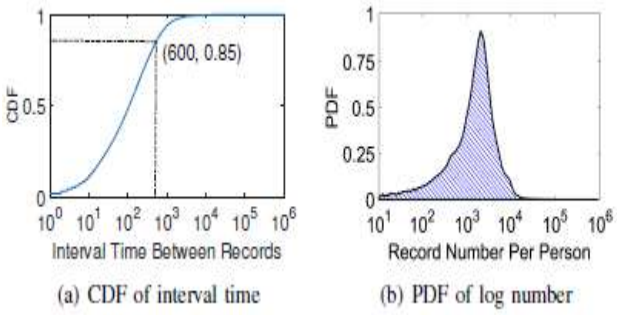
A. Data Description and First Findings

An anonymized cellular trace is the original dataset.

by an ISP from Shanghai, which spans the dates of August 1 and August 31, 2014, and comprises 2.4 petabytes (1015) of logs from over 6,400 base stations (BSs) located across Shanghai. The trace includes the (anonymized) device ID, start and finish times of each data connection, the BS ID, the BS address, and the amount of 3G or LTE data utilised during each connection for each entry. This massive dataset covers human activity in Shanghai and offers a solid physical foundation for our investigation in the actual world.
In Fig. 2, we display many straightforward representations of this dataset's properties. The Cumulative Distribution Function (CDF) of the time span between is displayed in Subplot (a). between two records in a row. The findings show that more than 85% of consecutive records occur in less than

60 minutes. Compared to an average inter-event time of 8.2 hours

Although it hasn't been fully edited, this article has been accepted for publication in a subsequent issue of this journal. Before the final publication, the content may change. TBDATA.2017.2778721, IEEE Transactions on Big Data, DOI for citation



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

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

The cellular data accessing logs are substantially more finely detailed during periods of consecutive mobile phone conversations [3]. The Probability Density Function (PDF) of the number of records is displayed in Fig. 2(b).

per user, in that order. The majority of mobile users have over 1,000 recordings overall. These findings show that our dataset has a large number of records of mobile users, and the precise temporal granularity ensures the validity of human activity modeling.

Our processing of the material into an easily consumable format allows for more effective use of the data. To be more specific, we begin by erasing logs that are duplicated or conflicted due to technological difficulties. Then,

We achieve this by averaging the traffic and user count of each BS across brief intervals of time. The data we've collected suggests that a time span of 10 minutes works well. This allows us to collect a set of data for the total traffic and user count at each BS. The array has a length of 4,032, with each member representing the amount of traffic or the number of users at a certain ten-minute period throughout 28 days. Finally, we use APIs from Baidu Map to translate BS addresses to longitudes and latitudes, allowing us to locate specific landmarks. Figures 3 and 4 depict typical cases from our dataset.

mobile phone users in a certain area served by a particular cell tower. Even if we can distinguish 28 distinct one-day periods from the four-week mobile traffic shown in Fig. 3 (a), it is still possible to see that

Mobile traffic data is challenging to examine because of the various variations it includes. You can see the fluctuation in the tower's user base over the course of four weeks in Fig. 3 (b). It is interesting to compare Fig. 3 (b) with Fig. 3 (a) and see that the mobile traffic series is definitely full of peaks, some of which correlate to the peak numbers of users, while others are generated by unknown sources. Naturally, the impact of out-of-the-ordinary occurrences and sounds further muddles interpretation.

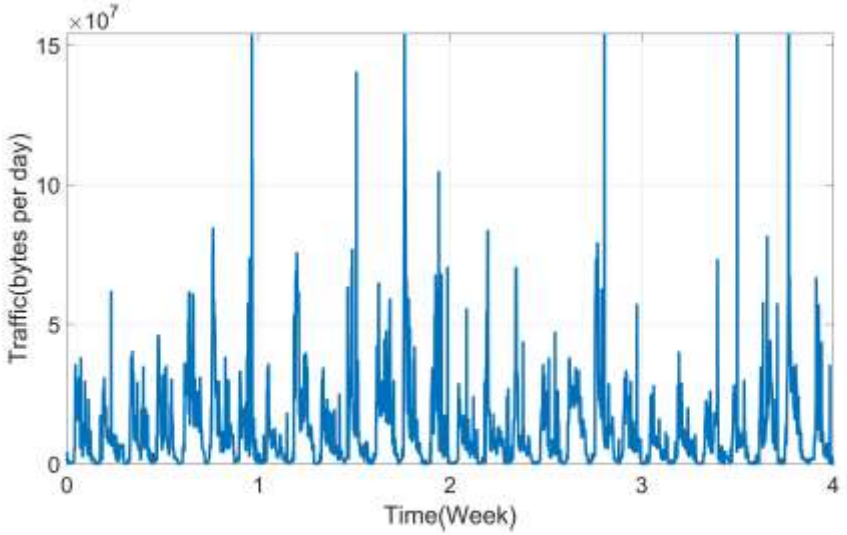
We may glean two insights from Fig. 4's depiction of the geophysical traffic volumes at 4 a.m., 10 a.m., 4 p.m., and 10 p.

To begin, diverse human activities at various times of the day result in varying traffic consumption. Since most individuals are still asleep at 4 in the morning, mobile data usage is low across the board. Due to most individuals being at

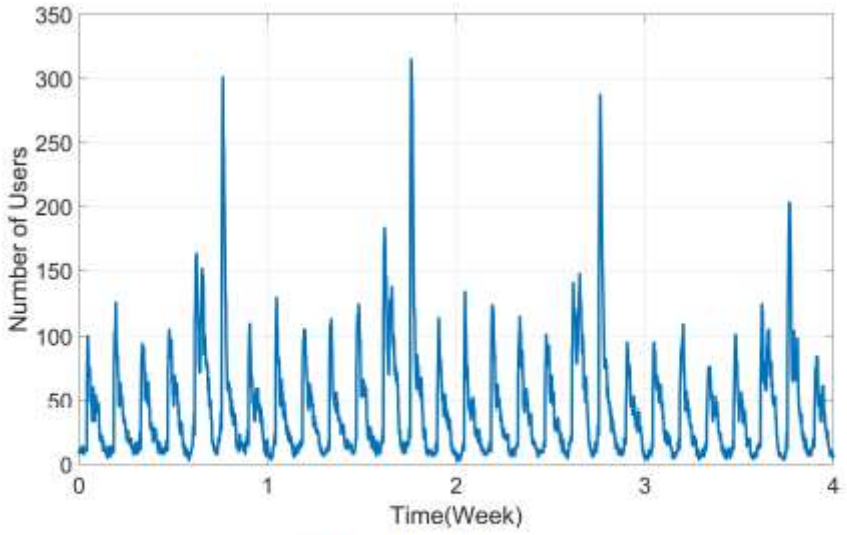
work, peak times for mobile data use occur around 10am and 4pm.

At 10 p.m., when most people have left work for the day and are starting to wind down at home, there is a surge in the volume of mobile data traffic.

Two, cellular towers' bandwidth is used in various ways and at different times depending on the locations' mobile traffic. Towers for wireless communication, such as mobile phones, are often installed at strategic locations across a city.



(a) Mobile traffic



(b) The number of users

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

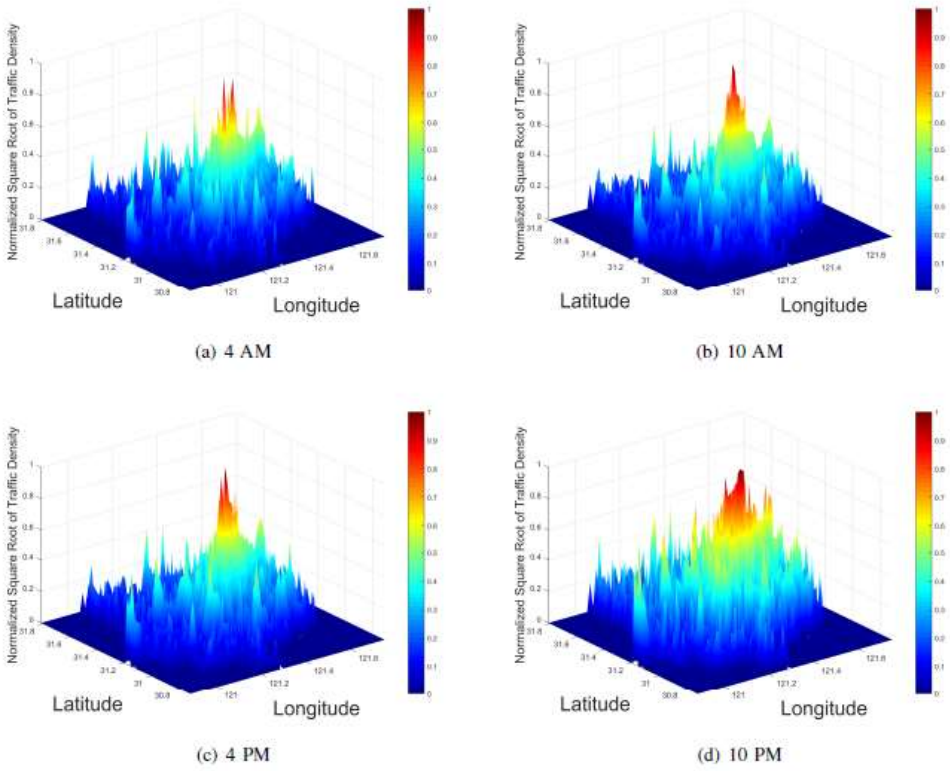
experience high traffic consumption in all time. The hottest areas covered by the darkest color also change over time, which suggests the movement of the crowd during one day.

B. What Drives Us

Data from cell towers may be used as a timely snapshot of urban life. The difficulty arises, however, from the intricacy of

Due to the dynamic nature of mobile traffic data, the presence of outliers, and the absence of previous knowledge, it is difficult to directly extract the information we need from the original traffic data.

The complexity of mobile traffic data, the disruption of such abnormal occurrences, and the absence of previous information for data patterns are all issues that we use a time series decomposition technique to solve by drawing on the theory of time series analysis [7]. This kind of mobile traffic analysis has two benefits. As a first step, if we think of the traffic data as a time series, we can see that the four-week time series has a natural period of one day and a discernible trend within one week. By breaking down the time series into its constituent parts, we may conduct our own analyses of these characteristics. So, the initial traffic complexity may be much reduced after this breakdown. In addition, a time series decomposition allows us to isolate the disruption caused by out-of-the-ordinary occurrences, laying the groundwork for future research in this area.



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

detection of unusual events Next, hierarchical clustering may be used to automatically find the most important patterns among the decomposed components.

hundreds of cell towers' worth of traffic data.

Figure 5 shows the cumulative traffic over all 6400 towers across various time intervals. The daily variations are shown in Fig. 5 (a), whereas the weekly variations are depicted in Fig. 5 (b), and they are quite comparable. In addition, the monthly traffic amounts shown in Fig. 5 (c) show a distinct weekly pattern. These commonsense conclusions push us to analyze mobile traffic data in order to draw out daily patterns and weekly trends. Thus, we recommended decomposing the traffic data into three components: the periodic components, the trend components, and the residuals, in order to study the traffic data in detail and to predict the traffic patterns at various scales.

Systems and algorithms, part III

1. Disintegration

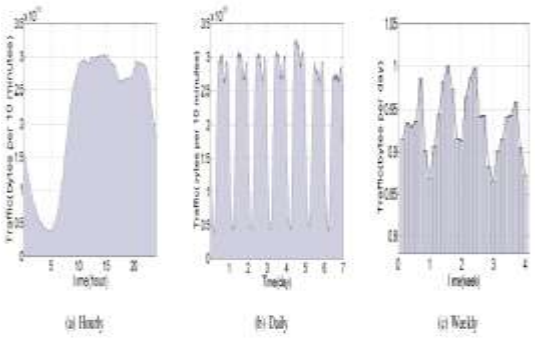
Following the format described in Section II-A, we have a mobile traffic record or observation fx1; x2; x4032g. The observation is a 28-day record of traffic data from a single tower, and each entry represents a 10-minute sum of that tower's traffic. Further, it has already been shown that this record displays both trend and periodicity. In light of this, we may use a time series technique [7] to break down this record as follows:

Where st is a daily periodic traffic component, satisfying st = st+d for t = 1; 2; ; nd with the period d = 144 that corresponds to one day, mt is a trend traffic component, and rt is the residual component comprising the noise and the impacts of exceptional occurrences. In the absence of seasonality, we may estimate the trend component from the remaining series to complete the decomposition.

First, we use a simple moving-average filter to estimate the trend traffic component, as follows: emt = 0:5xtq + xtq+1 + + xt+q1 + 0:5xt+q =d; where q = d=2 = 72 and q t n q. We then compute the average!k of the deviation series f(xk+jd emk+jd): q k + jd n qg for k = 1; 2; ; d.

The average daily traffic pattern at the cellular tower may be expressed as sk =!k d1 Xd i=1!i; k = 1; 2; ; d; sk = skd if k > d.

When the periodic component of the original series is subtracted from it, giving us the de-seasonality series, we get dt = xt st; t = 1; 2; ; n:



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

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

Take into account that the monthly series fdtg is used to determine the weekly trend of mobile traffic statistics. Therefore, we take the original four-week data fdtg and calculate its one-week average, b dt.

The weekly trend is then estimated using the moving average series b dt 1008 t=1. Time series trend estimation may be achieved in two broad ways: either by using a finite moving average filter to smooth the data or by using a function model to simulate the data. It is difficult to model the mobile traffic series using a

universal function due to the fact that its trends differ between cellular towers.

The former approach, then, is the most appropriate one for our data. An optimistic positive integer p is assumed here. Two-sided moving average

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

$j=-p$

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

Finally, the residual component is simply obtained as

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

Decomposition in one week is shown as an example in Fig. 6. A visual comparison of the original traffic data xt and the periodic component st is shown in Fig. 6 (a). Starting with the

Based on the findings shown in Fig. 6 (a), we can see that despite the fact that the original traffic statistics are quite noisy, there is a daily periodic component that displays a two-peak pattern.

The mt component of the weekly trend, shown in Fig. 6 (b), clearly shows that weekday traffic is greater than weekend traffic. The decomposition's one-week residual component rt is shown in Fig. 6(c). Unpredictable fluctuations in the volume of traffic may be accounted for by the residual component. For all four weeks, the autocorrelation of the residuals is shown in Fig. 7 to be less than 0.1, suggesting that rt is very similar to white noise.

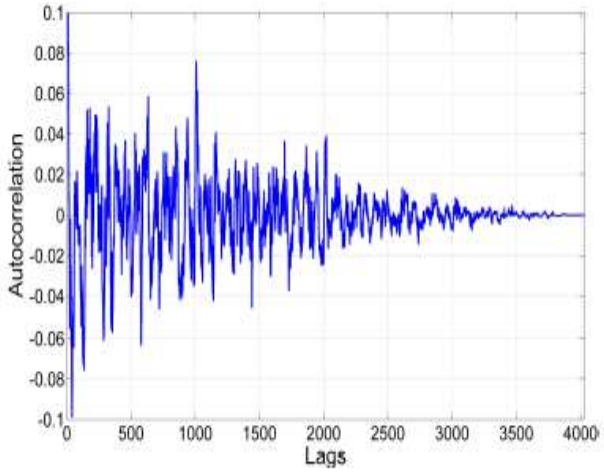
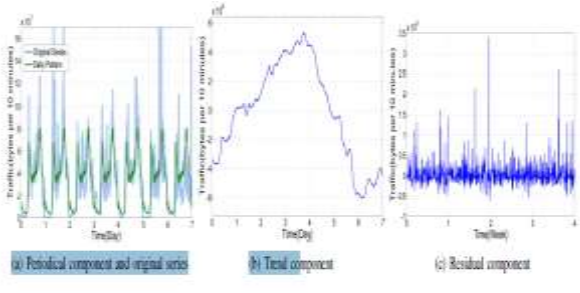


Fig. 7. Autocorrelation of the residuals.

B. 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.
Firstly, we generally have no idea how many main patterns

(a) Periodical component and original series (b) Trend co



(b)

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

(d)

Algorithm 1 Agglomerative Hierarchical Clustering.

1: Number of base stations N, threshold value D, mobile traffic $\{X_i[t]\}$ for $i = 1, 2, \cdots, N$

Output:

2: Number of clusters n_C, labels L_i for $i = 1, 2, \cdots, N_C$

Initialize:

3: Number of clusters: $n_C \leftarrow N$
4: Clusters: $c_i \leftarrow \{X_i[t]\}$ for $i = 1, 2, \cdots, n_C$
5: $stop \leftarrow false$
6: **while** *stop*==*false* **do**
7: $mindistance \leftarrow \infty$
8: **for** $i = 1$ **to** n_C **do**
9: **for** $j = i + 1$ **to** n_C **do**
10: $distance \leftarrow compute\_distance(c_i, c_j)$
11: **if** $mindistance > distance$ **then**
12: $mindistance \leftarrow distance$
13: $index \leftarrow [i, j]$
14: **end if**
15: **end for**
16: **end for**
17: $n_C \leftarrow n_C - 1$
18: $c_{index[1]} \leftarrow merge(c_{index[1]}, c_{index[2]})$
19: $c_{index[2]}.delete()$
20: **if** $n == 1$ **or** $mindistance > D$ **then**
21: $stop \leftarrow true$
22: **end if**
23: **for** $i = 1$ **to** n_C **do**
24: $\forall X_i[t] \in c_i, L_i \leftarrow i$
25: **end for**
26: **end while**
27: **return** n_C, and L_i for $i = 1, 2, \cdots, n_C$

should be determined for thousands of towers' worth of data.

In addition, cell towers are often situated in densely populated metropolitan areas, therefore tower traffic patterns may be very variable.

each other because to variations in user population and geographic dispersion. Furthermore, there are some "poor" towers with missing traffic data. It's difficult to know how to 'kick' these abnormal people out. We design a two-stage process to reliably recognize the most important patterns in the traffic logs: Two steps are involved here: 1) counting, 2) locating, and 3) confirming the most important patterns.

1) Recognize Commonalities: The Identifier is the heart of our mining infrastructure, which extracts useful information from network data.

Because it is not necessary to know the whole number of clusters beforehand, hierarchical clustering [8] was selected as our identifier. In hierarchical clustering, each input point is treated separately as a cluster, and then the closest clusters are merged into larger ones in an iterative process from the bottom up. In Algorithm 1, we see how this hierarchical grouping is accomplished.

Evidently, knowing when to stop the clustering process is a crucial technological difficulty for this kind of hierarchical clustering. To find the right amount of clusters, we use the Davies-Bouldin index (DBI) [9]. We'll use vector notation to express a series for ease of writing; for example, the ith mobile traffic sequence will be written as Xi[t].

may also be written as Xi. The DBI is specified using this notation style 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

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

where Ti is the total number of towers in the ith cluster, R is the total number of clusters, and Xi is the traffic data from the ith cellular tower. When

As a result, we can determine the optimal amount of patterns while still achieving the lowest possible DBI. Important patterns may be isolated from the whole set by applying a set of filters to the data.

a) Patterns of Periodic Components: Fig. 8 (a) displays the DBI as a function of the number of clusters, where the least DBI suggests that the optimal number of clusters is 1,040 for the periodic component of the mobile traffic.

We extract the five main patterns from all the clusters by treating the groups of more than 100 cellular towers as daily traffic patterns. These five main patterns, shown in Fig. 8 (b), are representative of daily traffic fluctuations and cover the hours 00:00 to 24:00. You'll see that all of them have a noticeable lull in activity between midnight and sunrise, when most people are asleep, but that their peak activity times vary widely.

In particular, Pattern #1 experiences its highest volume of mobile traffic in the late evening, Pattern #2 exhibits two rush hour peaks at around 8:00 and 18:00, Pattern #3 has a lasting stable high traffic from 8:00 to 16:00, Pattern #4 displays high traffic during the day, especially during lunch and dinner, and Pattern #5 appears to show the mixed features of the first four patterns.

FIGURE 8: The cumulative distribution function (CDF) of the correlation distance between towers in each cluster and the cluster centroid (c). Nearly all the towers in each cluster are located a safe distance from the cluster's center, since all the curves reach 100% at a distance less than 0.2. Thus, the clustering result may be trusted.

b) Trend Component Patterns: The weekly trend reveals the ebb and flow of mobile traffic on a weekly time scale. To examine the weekly trend part of the traffic, we use the same identifier we used before. However, in this case, we have some a priori knowledge about two obvious key patterns: during the weekdays, people go to their workplaces, for example in business districts, and the traffic reaches high values in weekdays at these places, while during the weekends, people go to entertainment places or stay in residential areas, and the traffic reaches peak values in weekends, at these places. Because of this, we can limit the number of key patterns to two, and Fig. 9 shows the two key patterns discovered using clustering for the weekly trend part of the traffic. It's clear that Pattern #1 has steady, high traffic throughout the week but

much lower numbers on the weekends, whereas Pattern #2 has low traffic during the week and high values during the weekend.

The two weekly trends shown here are direct reflections of human urban activity.

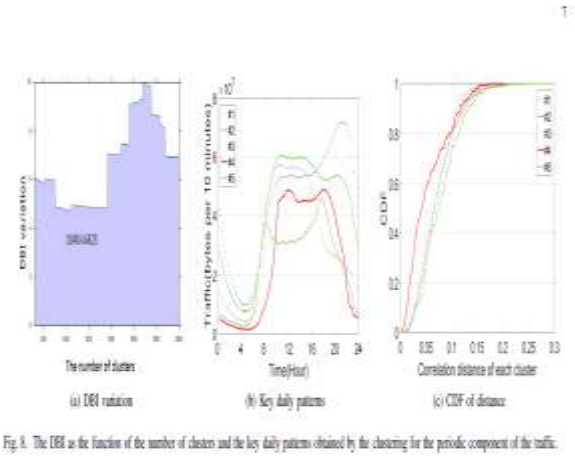


Fig. 8. The DBI as the function of the number of clusters and the key daily patterns obtained by the clustering for the periodic component of the traffic.

TABLE I

TABLE I
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

We then classify the daily patterns by the particular urban function in order to establish a connection between the detected patterns and normal human urban activities.

thematically connected domains. We link the everyday rhythms to the interplay of four essential city activities. The detected patterns are then further verified by examining the correlation between daily and weekly trend patterns.

1. Identify Recurring Activities and Name Them a) Figure 8 depicts five different daily traffic patterns, each of which has to be placed in its geographical context before we can make any meaningful connections between them and the regular activities of urban dwellers.

Typical rush hours in cities correspond to the two peaks in the daily pattern #2 shown in Fig. 8 (b), which occur at 8:00 and 17:00-18:00. So, we might speculate that the Shanghai transportation function region is involved in this everyday trend. Points of interest (PoI) distribution is used to characterize the geographical elements of each daily pattern, allowing for more precise labeling. One of the largest providers of online map services, Baidu Map, supplies us with 23 different types of Points of Interest (PoI) including 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, and overpasses.

We start by counting the PoIs within 200m of each cell tower for every daily pattern. Afterwards, we standardize the points of interest across all clusters. Table I presents a summary of the PoI distribution for all of the different patterns. To make the average number of PoI more legible, we increased it by 1000. All three of the most severe forms of PoI are indicated by shades of orange, with the most severe ones appearing at the top of the list for each daily pattern. It is evident that the distribution of points of interest (PoI) varies widely amongst patterns, and we may assign names to patterns based on the primary categories of PoI they include, such as Residential Area, Transport Area, Office Area, Entertainment Area, or Mixture Area.

Location of Households: Towers in cluster#1 mostly serve residential and commercial buildings, as shown in Table I. When taken in conjunction with the evening-high mobile traffic patterns shown in Fig. 8 (b), we can conclude that this cluster may be classified as a residential region, where people typically return after a day at the office.

The subway station point of interest (PoI) is a substantially larger transport area than the other PoIs in cluster #2. Figure 8 (b) depicts the two daily traffic peaks that occur while individuals are traveling to and from their homes and places of employment. As a result, we may classify this grouping as a transportation hub.

Corporate and financial services are the most popular points of interest in the third office cluster. Also, keep in mind that the volume of this daily pattern remains strong all the way through '9 to 5' This allows us to connect the dots between this cluster and Shanghai's commercial districts.

Cluster #4 is heavily skewed toward points of interest (PoIs) related to dining and retailing. As can be seen in Fig. 8, this cluster's peak hours for traffic are just around lunch and evening (b). From these indicators, it's safe to

classify this location as a tourist hotspot.

Cluster #5's Points of Interest (PoIs) are dispersed quite uniformly among several functional zones, and the daily traffic patterns in these areas demonstrate a curious admixture of

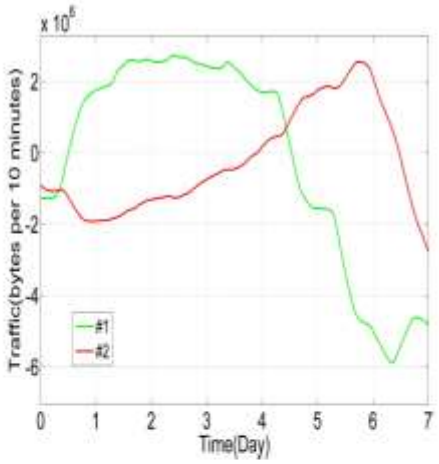
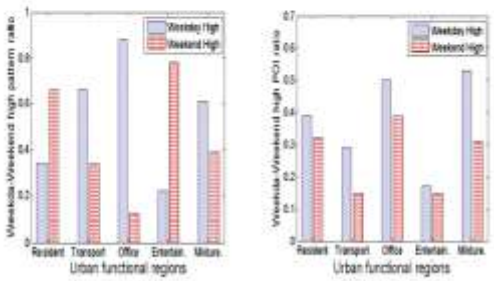


Fig. 9. Key weekly trend patterns.



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

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

from the first to the fourth cluster. Therefore, this cluster fits the criteria for a mixed region.

2. b) Identify the weekly pattern Patterns: Weekly tendencies

have been sorted up into the weekday-high and weekend-high categories, respectively. The former is more often linked with commercial and industrial settings, whereas the latter is more often connected with private homes.

c) The Connection Between Weekly and Daily Trend Patterns Here we calculate the ratios of the two weekly trend patterns to each daily pattern, as shown in Fig. 10, to investigate the connection between the two time scales (a). It is interesting to note that the weekend-high trend pattern accounts for 66% of towers in residential areas (daily pattern#1) and 78% of towers in entertainment areas (daily pattern#4), while the weekday-high trend pattern accounts for nearly 90% of office areas (daily pattern#3) and over 60% of transport areas (daily pattern#2). These findings are consistent with what we know about human behavior in urban settings, providing more support for our categorization.

As was previously discussed, there are 23 POI values distributed over five functional areas. Additionally, some of these 23 unique POIs (finance, offices, corporations) are common high POIs during the weekdays, while others (residential, live service, entertainment) are typical high POIs during the weekends. Figure 10 (b) displays, for each functional area, two ratios: the ratio of the sum of the weekdayhigh POI values to the total of all POI values, and the ratio of the sum of the weekendhigh POI values to the sum of all POI values. Figure 10(b) provides unambiguous confirmation that our pattern categorization is correct.

Predictions, C.

Now, we use the clustering result as the foundation for a mobile traffic forecasting system. The mobile traffic series is nonstationary and exhibits periodicity, hence we use the seasonal autoregressive integrated moving average (SARIMA) model [10] to accomplish this. ARIMA is the abbreviation for the overarching version of the SARIMA model (p; d; q) (P;D;Q) There are (P;D;Q) terms representing the seasonal portion of the model, where P is the number of seasonal autoregressive terms, D is the number of seasonal differences, and Q is the number of seasonal moving average terms, and there are (q; d; q) terms representing the non-seasonal portion of the model, where q is the number of non-seasonal autoregressive terms, d is the number of non-seasonal differences, and q is the number of non-season

From the mobile traffic series fxtg4032 t=1, we are able to reliably extract the SARIMA model ARIMA(1; 0; 1) (1; 1; 1)1008.

$$\left(1-ar\cdot B\right)\left(1-sar\cdot B^{1008}\right)\left(1-B^{1008}\right)x_t = \left(1+ma\cdot B\right)\left(1+sma\cdot N^{1008}\right)a_t,$$

where at is the white-noise series, m0 = 1008, B stands for the backward shift operator, and Bxt = xt1 establishes the parameters ar, sar, ma, and sma, respectively.

minimum arithmetic mean squared error

Using the aforementioned model, we can forecast the traffic for the next week using data from the past three weeks broken down into four daily patterns. The forecasting results are shown in Fig. 11, and the logarithmic traffic series is utilized to improve readability. Across all four configurations, it is evident that the predicted series closely matches the underlying actual logarithmic traffic series.

The reliability of the forecast is then measured objectively.

Let us define resident area (rsd) as the number of buildings designated as such, with the associated index set to be Nrsd. In this case, we get a genuine logarithmic series of traffic counts for each tower k: x(k) = 1; x(k) = 2;... = x(k) = 1008 median value (k). The median of the actual traffic counts for the residential areas is then determined by using the formula

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

Using the model built from the classified data, we obtai
forecasting series $\{\widetilde{x}_1^{(k)}, \widetilde{x}_2^{(k)}, \cdots, \widetilde{x}_{1008}^{(k)}\}$ for tower k, w
has the mean square error (MSE) of $\widetilde{\text{MSE}}^{(k)}$. Thus the
of the classified model prediction for resident area is give

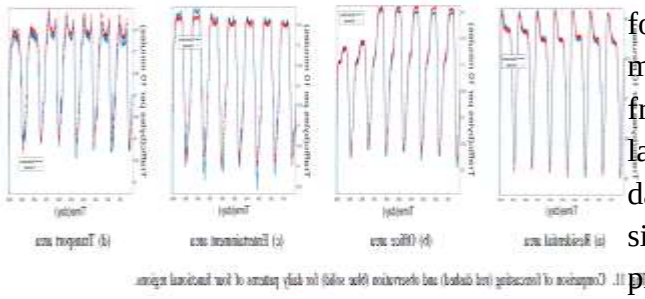
$$\text{MSE} = \frac{1}{N_{\text{rsd}}}\sum\nolimits_{k\in\mathcal{N}_{\text{rsd}}} \widetilde{\text{MSE}}^{(k)}.$$

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

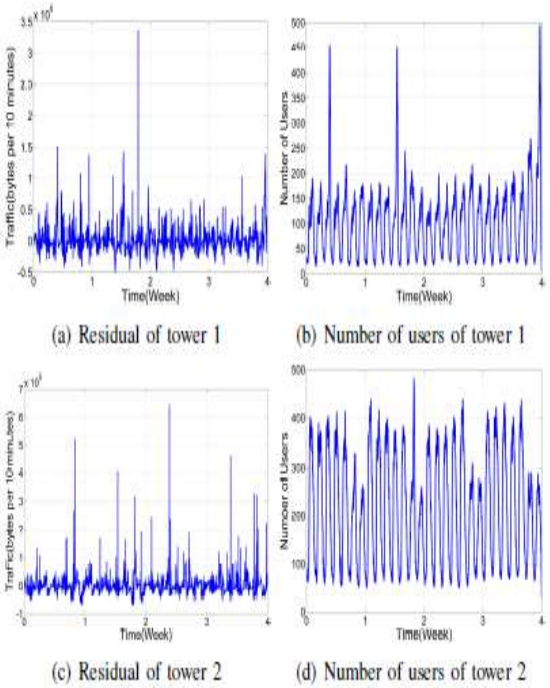
$$\widehat{\text{MSE}} = \frac{1}{N_{\text{rsd}}}\sum\nolimits_{k\in\mathcal{N}_{\text{rsd}}} \widehat{\text{MSE}}^{(k)}.$$

We summary the MSEs of the classified model predictions for

the four daily traffic patterns in the first row of Table II. 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 II, while the ratios of the four[MSE values over
the corresponding mean values of the four daily-pattern traffic
series are given in the third row of Table II.



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

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

As a last step, we add mobile traffic forecasting as an application to our system and use cellular label data to dramatically boost the performance of the ARIMA model.

towers accumulated in the Clustering phase. We do this by training four separate forecasting models, each tailored to one of four distinct daily trends. We choose a model for a given set of input traffic data from a single tower based on the tower's label. Our experimental results on our dataset demonstrate that this method yields significantly improved prediction precision.

D. Taking Note of the Leftovers

We now show our technique for determining, given a cellular tower's residual traffic data, whether or not an accident or unusual event occurred. The residual and user count for a chosen cell tower in the HongKou area are shown in Fig. 12. As can be seen in Fig. 12 (b), there are three distinct peaks in the number of users that correspond to distinct crowd events; however, due to the high level of noise in the residual component, these three events cannot be identified from the residual alone. A similar issue is seen when contrasting Figure 12 (c) and (d), where a surge in the number of users is observed on the second Saturday but is obscured by the residual component.

We establish a threshold of 4 standard deviations from the mean for each residual series to filter out the noise, which represents meaningless random occurrences in the residual component. When the residual component stays over the threshold for more than 30 minutes, we classify it as an abnormal occurrence. If we assume a sampling interval of 10 minutes, then 3 samples in a row would represent a sampling interval of half an hour.

Thus, given a residual series $\{r_1, r_2, \cdots, r_{4032}\}$,

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 $j = 1, 2, \cdots, 4030$, 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 $j+2$.

To put our cellular-based anomaly detection system to the test, we first identify three cellular towers near gymnasiums or other public venues where mass gatherings take place.

element of mobile traffic that doesn't contribute much. Table III's RD (residual detection) column shows if the crowd event can be deduced from its matching residual series. Tower No.1 is situated near Hongkou Stadium, Tower No.2 is close to the Mercedes Benz Cultural Center, and Tower No.3 is close to the Luwan Gymnasium. We can thank soccer for the odd occurrences.

Thi

TABLE II
FORECASTING ERRORS FOR DAILY PATTERNS.

	resident area	office area	entertainment area	transport area
MSE (classified model prediction)	0.086295	0.11703	0.17745	0.21524
MSE/Mean (classified model prediction)	0.49%	0.69%	1.08%	1.26%
MSE/Mean (unclassified model prediction)	2.26%	2.55%	6.42%	3.93%

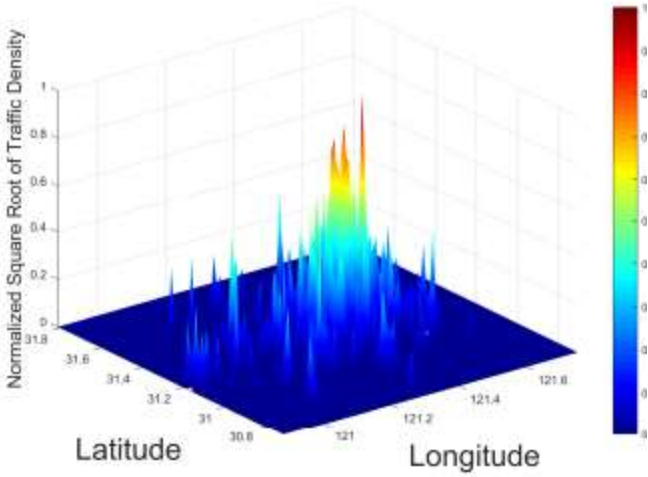


Fig. 13. The spatial distribution of anomalies.

events like concerts or sporting competitions that result in a spike in use. Table III demonstrates that our technique is able to accurately identify the vast majority of occurrences.

effective and trustworthy methodology

We then use our algorithm to identify unusual occurrences across all cell towers. Mapped out in Fig. 13 is the location of all the weird things that happened this month. It is clear from the heat map that most anomalous occurrences take place in and around the city's core, and that their distribution is highly correlated with the consumption of mobile traffic shown in Fig. 4. This suggests that anomalies tend to take place in areas with high levels of mobile traffic. Figure 14 displays the means and standard deviations of outliers throughout a four-week period. Weekends have higher averages than weekdays, which makes sense given that concerts and other special events tend to take place then. An further fact that has been noticed is that the standard deviation is highest on

Sunday. This finding suggests that the frequency of unusual occurrences varies throughout weeks. Both the temporal and geographical patterns of the identified abnormalities are consistent with human urban activity.

Our system's primary functions—traffic forecasting and anomaly detection—are also available as web-based products. With only a few inputs, our system can model and name every cell tower in a given area. Using input traffic data from a single tower, we may choose a prediction model in accordance with the label, and then update the model's input data to reflect traffic over the last three weeks (or whatever time period is of interest). Our system configures variables used for outlier detection.

TABLE III
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

use past information as the basis. Using an efficient decomposition strategy, our system is able to breakdown the most recent traffic data every few minutes (or, alternatively, take up the most recent one week's data as input) and identify patterns.

inconsistencies in the noise component in real time. Therefore, our system can provide live traffic forecasting and anomaly detection based on a previous mobile traffic dataset.

IV. CONCLUSION AND IMPRESSION

We have so far derived two natural weekly trend patterns from the trend components of our traffic data, and five daily patterns from the regular periodic components.

The human activities and surrounding environments that contribute to these traffic patterns are the focus of the next section.

1) Recognizing Daily Routines

1) Seasonal variation: Daily traffic patterns often show both high and low points throughout the day. Table IV details the daily patterns and when they peak and drop.

Table IV shows that, across all patterns, the lowest use occurs between 4 and 5 am, when most users are fast asleep. There is a peak period in the residential area at 21:00, when people return home from work, and two peaks in the transportation sector at 8:00 in the morning and around 5:00 to 6:00 in the evening, which correspond to the two rush hours of the day. There is constant, heavy foot traffic in the workplace, with no discernible "peak period."

There are two major periods of day when people congregate at entertainment venues; these are around noon and six o'clock, when people typically eat lunch and supper.

2) The ratio of daytime to nighttime traffic volumes: Fig. 15 (a) displays, for each daily pattern, the volume of traffic during the daytime hours of 7 a.m. to 7 p.m., and

the volume of traffic during the nighttime hours of 7 p.m. to 7 a.m. Daytime traffic volumes are clearly greater than nighttime volumes for all patterns except residential areas. This squares well with the way that people often go about their everyday lives.

Figure 15 (b) shows that the percentage of daytime to nighttime traffic varies significantly between patterns. More specifically, the ratio is around 0.8 in residential regions, which is much lower than the ratios in transportation hubs, business districts, and entertainment districts. Human everyday activities provide an excellent explanation for this phenomenon; during the day, individuals travel to areas like employment and amusement, and at night, they return home. The greatest ratios, up to 1.4, are seen in business districts, when most individuals are away at work.

Section B: Recognizing Weekly Trend Patterns

There are two distinct weekly trends discernible in the weekly trend components: a weekday-high trend and a weekend-high trend. In Fig. 16

We provide one week's worth of traffic based on these two typical weekly patterns. From Fig. 16 (a), we can see that the weekend-high pattern has lower traffic than the weekday-high pattern, and vice versa for the weekday-high pattern (b). As can be seen in Fig. 16 (a), the weekdays' traffic peaks in the late afternoon, suggesting that the residential pattern is the primary daily pattern contributing to the weekend's high pattern. In Fig. 16 (b),

TABLE IV
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

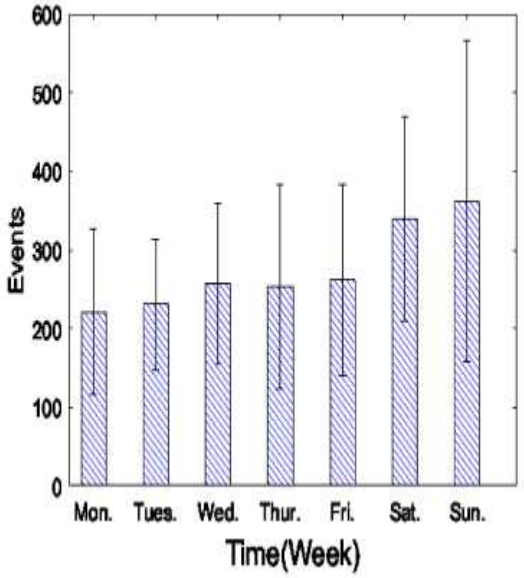
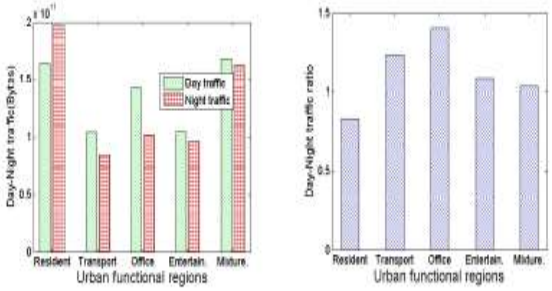


Fig. 14. Anomalous events in four weeks.



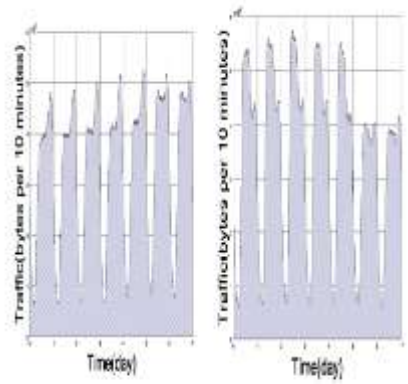
(a) Day-night traffic amount (b) Day-night traffic ratio

Fig. 15. Day-night traffic amount ratio.

Workday traffic is much greater than non-workday traffic, suggesting that the workplace schedule is the primary contributor to the weekday-high trend.

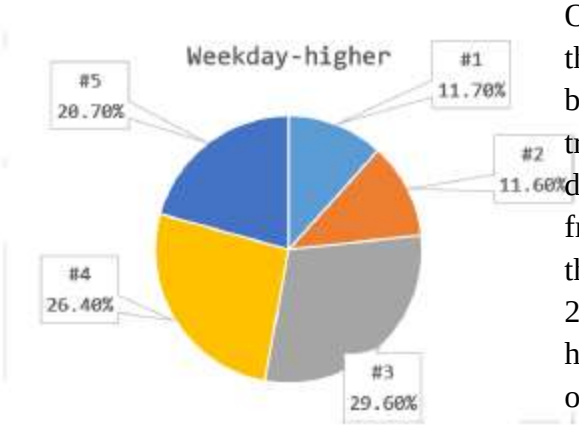
Figure 17 depicts the relative contributions of the two weekly trends' individual daily patterns. According to Fig. 17 (a), the most

common of the four common daily routines (1, 2, 3, and 4) is the

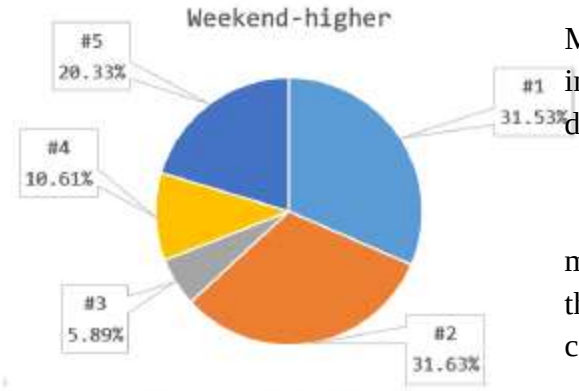


(a) Weekend-high pattern (b) Weekday-high pattern

Fig. 16. Weekly trend components in one week.



(a) Weekday-high pattern



(b) Weekend-high pattern

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

The office pattern (#3), which occupies 29.6% of the towers, is the single most important factor in the weekday-high pattern, followed by the entertainment pattern (#4).

whereby 26.4 percent of the buildings are used as storage space. It makes perfect sense that the office pattern is the primary driver of the weekday-high pattern, as this reflects the primary activities of most people during the week, while the entertainment pattern also contributes heavily to the weekday-high pattern, both because people need to eat at lunch and dinner every day and because there are a lot of towers in the entertainment district. On the other hand, as shown in Fig. 17 (b), the weekend-high pattern is driven mostly by the residential pattern (#1) and the transit pattern (#2). This is a very accurate depiction of how most people spend their free time on the weekends. Interestingly, the mixed pattern (#5) accounts for almost 20% of the towers in both the weekday-high and weekend-high patterns. Because of this, mixed areas are what they are.

FIVE: CONNECTED DOCUMENTS

Many studies have focused on extracting information about urban environment dynamics and social events from digital

marks left behind [15]. Here, we classify the relevant literature based on four criteria: Analyzing Traffic Data

enabled apps; digital footprints used to identify urban dynamics; time series techniques utilized to analyze mobile traffic data; and event detection from mobile traffic.

Data on cellular traffic flows has been put to use in several fields. Personal characteristics such as sexual orientation, race, religion, and political leanings may be inferred from the data [5]. Human movement patterns have been modeled using CDRs [3, 11], with the results indicating a high degree of temporal and geographical regularity [3] and a high degree of prospective predictability [11]. The research [4] examines 3G cellular networks with the intention of elucidating the habits of mobile data users and finds that a tiny percentage of very active users account for the vast bulk of the network's data consumption. Books [12], [13] use CDR statistics to infer and categorize land use. The utilization of cellular network traces enables further applications, such as the inference of friendship network structure [14], the comprehension of mobile user browsing activity [14], and the optimization of content delivery depending on user location [19].

CDRs, social media data, and mobile traffic data are the three most often utilized forms of digital footprints for uncovering human activity patterns in metropolitan areas. In order to simulate human activity patterns [3, 21] and to estimate population dispersion [21], CDRs are used. CDRs lack temporal density compared to mobile traffic data. Mobile social instant communication programs have largely supplanted phone calls as the major means of contact in metropolitan areas, thanks to the proliferation of mobile Internet. The telephone has been mostly replaced by instant messaging apps, which many prefer. In addition, the proliferation of mobile payment methods and the constant flow of mobile traffic that cities generate have cemented the Internet's dominance over all aspects of urban social life. As a result, there is a wealth of information about people's behaviors available in mobile traffic data. Based on a social activity dataset and GPS travel records, the study [22] offered a technique to recognize urban events. The information collected from social media sites provides firsthand evidence of the actions of specific users. Examples include the visual depiction of urban occurrences in Twitter posts. However, social media data are more challenging to analyse and mine than mobile traffic data since they are often in text, audio, or video format and include a great deal of duplicate information. Meanwhile, since they are a compilation of data from many different users, mobile traffic statistics are better at protecting citizens' privacy. A further use for mining the contexts and behavior information from mobile traffic data is shown by the study [20], which developed a system to categorize service usages using encrypted Internet traffic data of mobile messaging Apps. The disintegration of traces of human activity in urban areas has been the subject of several studies. To dissect a human endeavor, the authors [17, 22] offer a non-negative tensor factorization method.

tensor into more fundamental tensors of daily life. On the other hand, our system is able to pick the number of fundamental patterns on its own, which is an advantage over other methods that force you to choose an arbitrary number. An image segmentation method was used to mine social events from a 3D matrix [18] constructed to represent the time, location, and likelihood of a social event based on a probabilistic model. This algorithm did not clearly dissect the human traces and did

not take into account additional data like daily pattern or long-term trend, unlike our own. By splitting the original cell phone activity series into the seasonal communication series and the residual communication series, the research [2] is able to deduce aspects of urban ecology from spatial-temporal cell phone activity data. In contrast to the approach used here, the mobile phone traffic series is decomposed in [2] by first undergoing a frequency-domain transformation using FFT, from which the primary frequency components are then extracted. Our report proves that mobile traffic trends reflect long-term fluctuations, which are overlooked in this study. In order to analyze mobile traffic data, time series analysis is often used. This is particularly true for mobile traffic forecasting. While work [24] models and forecasts real wireless traffic, such as GSM traffic, using seasonal ARIMA models, work [23] proposes a technique for traffic forecasting based on multiple regression model for time-series. We also use an autoregressive integrated moving average (ARIMA) model, tailoring the model's parameters to actual daily trends. We also demonstrate that our model is capable of producing accurate predictions of mobile traffic usage. To break down a human activity tensor into fundamental life pattern tensors, [17] [22] present a non-negative tensor factorization technique. On the other hand, our system is able to pick the number of fundamental patterns on its own, which is an advantage over other methods that force you to choose an arbitrary number. An image segmentation method was used to mine social events from a 3D matrix [18] constructed to represent the time, location, and likelihood of a social event based on a probabilistic model. This algorithm did not clearly dissect the human traces and did not take into account additional data like daily pattern or long-term trend, unlike our own. By splitting the original cell phone activity series into the seasonal communication series and the residual communication series, the research [2] is able to deduce aspects of urban ecology from spatial-temporal cell phone activity data. In contrast to the approach used here, the mobile phone traffic series is decomposed in [2] by first undergoing a frequency-domain transformation using FFT, from which the primary frequency components are then extracted. Our report proves that mobile traffic trends reflect long-term fluctuations, which are overlooked in this study. In order to analyze mobile traffic data, time series analysis is often used. This is particularly true for mobile traffic forecasting. While work [24] models and forecasts real wireless traffic, such as GSM traffic, using seasonal ARIMA models, work [23] proposes a technique for traffic forecasting based on multiple regression model for time-series. We also use an autoregressive integrated moving average (ARIMA) model, tailoring the model's parameters to actual daily trends. We also demonstrate that our model is capable of producing accurate predictions of mobile traffic usage. A large amount of work has gone into using cellular traffic analysis to spot irregularities. Standard statistical methods are used to an analysis of phone data to discover how often and where unusual occurrences have occurred.

techniques from percolation theory to characterize these spatial and temporal oddities. There is a correlation between the

sort of event and the origins of those attending, as shown in [26], which analyzes over 1 million cell-phone traces to study crowd migration during special events. The research [27] examines how communities react to shocks from the outside world, like as terrorist bombings and natural disasters, by tracking how people's patterns of movement and communication alter in real time. Our study effectively extracts the root causes of anomalous events by deconstructing mobile traffic series, and our anomalous event detection approach has been used to identify out-of-the-ordinary activities, such concerts and matches, in mobile data.

In conclusion, we provide a novel framework for studying and modeling urban functional zones and human activities using data on massive amounts of cellular mobile traffic gathered by ISPs. There are three main ways in which our work is novel: At the outset, we use a massive dataset of mobile traffic. Our dataset better reflects urban dynamics in the mobile age, and it has finer temporal and geographical resolution than previous digital records. To further examine mobile traffic statistics, we use a novel approach and use a times series decomposition technique. On the one hand, we use periodic and trend components to represent human regular activity patterns over many time periods in urban environments.

Instead, we focus on the "residual" part of the model, where the impact of random occurrences is obscured by background noise. This approach gives a thorough understanding of interactions between human activities and network dynamics, and, to the best of our knowledge, has not been applied in mobile traffic analysis in the open literature. Finally, we present a methodology for investigating mobile traffic records that incorporates human activity pattern mining, mobile traffic forecasting, and anomaly detection.

INTERLUDE: Section VI. Conclusions

In this article, we investigate the link between urban dwellers' mobile data traffic and their day-to-day routines.

systemic and all-encompassing setting based on a massive collection of mobile traffic logs with a fine granularity. We have developed an effective system that combines the processes of traffic pattern clustering and labeling, mobile traffic forecasting, and event detection by using a generic time series decomposition approach. First, we do a natural decomposition of the traffic series into a daily periodic component, a weekly trend component, and a residual component. We then isolate five primary daily activity patterns that are intrinsically linked to many domains of human functioning.

Through the use of an ARIMA model and the training of several models based on historical data, we are able to considerably enhance the accuracy with which we can estimate mobile traffic.

In addition, the trend component's two weekly trend patterns are familiar to us, which again represent the underlying weekly human activities in the actual world. Finally, we utilize the residual component to identify outliers generated by abnormal human behavior, and we demonstrate that our anomalous event identification approach can effectively identify abnormal human behavior from

the noisy residual-component series that is characteristic of the actual world.

Our system is capable of handling large offline mobile traffic databases while simultaneously providing online traffic forecasts and event detection services. As a result, our research has laid the groundwork for efficiently managing large quantities of mobile traffic data and has offered a deep dive into the complex interplay between mobile data traffic consumption and human activities in the modern urban setting.

REFERENCES .

[1] Cisco Visual Networking Index: Global Mobile Data Traffic
Forecast Update, 2015–2016 White Paper. Available at:
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. MobiHoc 2015 (Hangzhou, China), Jun. 22-125, 2015,
pp. 317–326.

[3] M. C. Gonz´alez, C. A. Hidalgo, and A.-L. Barab´asi, “Understanding
individual human mobility patterns,” Nature, vol. 453, pp. 779–782,
2008.

[4] Y. Jin, N. Duffield, A. Gerber, P. Haffner, W.-L. Hsu, G. Jacobson,
S. Sen, S. Venkataraman, and Z.-L. Zhang, “Characterizing data usage
patterns in a large cellular network,” in Proc. CellNet Workshop 2012
(Helsinki, Finland), Aug. 13, 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. the
National Academy of Sciences of the United States of 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
Communications, vol. 21, no. 1, pp 80–88, Feb. 2014.

[7] P. J. Brockwell and R. A. Davis, Introduction to Time Series and
Forecasting (Second edition). Springer Science & Business Media, 2006.

[8] F. Corpet, “Multiple sequence alignment with hierarchical clustering,”
Nucleic Acids Research, 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 Analysis
and Machine Intelligence, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[10] R. S. Tsay, Analysis of Financial Time Series (2nd Edition). Wiley, 2005.

[11] C. Song, Z. Qu, N. Blumm, and A.-L. Barab´asi, “Limits of predictability
in human mobility,” Science, 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 Information Science, vol. 28, no. 9, pp. 1988–
2007, 2014.

[13] J. L. Toole, M. Ulm, M. C. Gonz´alez, and D. Bauer, “Inferring land use
from mobile phone activity,” In Proc. of ACM SIGKDD 2012 (Beijing,
China), Aug. 12, 2012, pp. 1–8.

[14] N. Eagle, A. S. Pentland, and D. Lazer, “Inferring friendship network
structure by using mobile phone data,” Proc. National Academy of
Sciences, vol. 106, no. 36, pp. 15274–15278, 2009.

[15] Daqing Zhang, Bin Guo, and Zhiwen Yu. 2011. The Emergence of Social
and Community Intelligence. Computer 44, 7 (July 2011), 21-28.

[16] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, “Contextual

localization through network traffic analysis," in Proc. INFOCOM 2014
(Toronto, Canada), Apr. 27-May 2, 2014, pp. 925–933.

[17] Fan Z, Song X, Shibasaki R. CitySpectrum: a non-negative tensor
factorization approach[C]//Proceedings of the 2014 ACM International
Joint Conference on Pervasive and Ubiquitous Computing. ACM, 2014:
213-223.

[18] Zhang, W., Qi, G., Pan, G., Lu, H., Li, S., Wu, Z., 2015. City-
Scale Social Event Detection and Evaluation with Taxi Traces. ACM
Transactions on Intelligent Systems and Technology 6, 40:140:20.

[19] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Contextual
localization through network traffic analysis," in Proc. INFOCOM 2014
(Toronto, Canada), Apr. 27-May 2, 2014, pp. 925–933.

[20] Fu Y, Xiong H, Lu X, et al. Service usage classification with encrypted
internet traffic in mobile messaging apps[J]. IEEE Transactions on
Mobile Computing, 2016, 15(11): 2851-2864
.
[21] Jiang, Shan and Ferreira, Joseph and Gonz´alez, Marta C. "Activity-based
human mobility patterns inferred from mobile phone data: A case study
of Singapore." IEEE Transactions on Big Data 3.2 (2017): 208-219.

[22] Chen, Longbiao, et al. "Fine-Grained Urban Event Detection and Characterization
Based on Tensor Cofactorization." IEEE Transactions on
Human-Machine Systems 47.3 (2017): 380-391.

[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. GLOBECOM 2005 (St Louis,
MO), Nov. 28-Dec. 2, 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. ICC 2003
(Anchorage, Alaska), May 11-15, 2003, pp. 1675–1679.

[25] J. Candia, M. C. Gonz´alez, P. Wang, T. Schoenharl, G. Madey, and
A.-L. Barab´asi, "Uncovering individual and collective human dynamics
from mobile phone records," J. Physics A: Mathematical and 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 Computing (Helsinki, Finland), May
17-20, 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,