This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2016.2599878, IEEE Transactions on Services Computing

Big Data Driven Mobile Traffic Understanding and Forecasting: A Time Series Approach

Fengli Xu, Yuyun Lin, Jiaxin Huang, Di Wu, Hongzhi Shi, Jeungeun Song, and Yong Li, *Member, IEEE.*

Abstract—Understanding and forecasting mobile traffic of large scale cellular networks is extremely valuable for service provider to control and manage the explosive mobile data, such as network planning, load balancing, and data pricing mechanisms. This paper targets at extracting and modeling traffic patterns of 9000 cellular towers deployed in a metropolitan city. To achieve this goal, we design, implement, and evaluate a time series analysis approach that is able to decompose large scale mobile traffic into regularity and randomness components. Then, we use time series prediction to forecast the traffic patterns based on the regularity components. Our study verifies the effectiveness of our utilized time series decomposition method, and shows the geographical distribution of the regularity and randomness component. Moreover, we reveal that high predictability of the regularity component can be achieved, and demonstrate that the prediction of randomness component of mobile traffic data is impossible.

Index Terms—Mobile big data, mobile traffic, time series analysis, traffic forecasting

1 Introduction

In the past decade, a dramatic growth of mobile data traffic have been witnessed, which is mainly contributed by the prosperity of mobile devices, the first-class citizen of the mobile Internet. The worldwide mobile devices is expected to consume more than 24 exabytes (10^{18}) per month by 2019 [1], 9 times larger than the capacity of the existing mobile network. Cellular data networks, the main component of mobile Internet, have become ubiquitous in the past few years, which fuels the user-friendly mobile devices (e.g., cellphones, netbooks, and tablet devices) as well as the plethora of mobile applications. Indeed, mobile data traffic has surpassed voice traffic on a global basis and their pronounced growing trend continues. Given its pervasive usage, cellular data trace potentially provides more information about the network behaviors than the traditional voice traffic, and understanding the traffic distributions and modelling the traffic patterns becomes an important problem [2], [3]. However, despite the ambient cellular connectivity, we are facing a critical but challenging problem: the current understanding about the patterns of mobile traffic big data experienced in the urban areas is very limited, especially when 3G and LTE networks are widely deployed in modern metropolis. For example, we still don't know the key factors that affect the traffic variations of cellular towers deployed in urban downtown areas. Such limited knowledge significantly increases the cost of operating millions of cellular towers in big cities and reduces the quality of service provided.

Understanding mobile traffic patterns is extremely valuable for both service providers and mobile users, especially in the large scale urban environment[4], [5]. If the patterns of cellular data traffic can be identified and forecasted accurately, instead of applying the same strategy to operate all the towers, i.e., using the same load balancing and data pricing algorithms on each tower, the service provider can exploit the modeled traffic patterns to customize a strategy for each individual cellular tower. For example, a service provider can potentially have different pricing on individual cellular tower based on the traffic it experiences. In addition, mobile users will benefit from the traffic modeling as well because they can enjoy better service and reducing their cost by choosing base stations with predicted lower traffic. Therefore, utilizing the big data generated by cellular network, we are able to help service provider design better operation scheme as well as enhance the mobile users' experience.

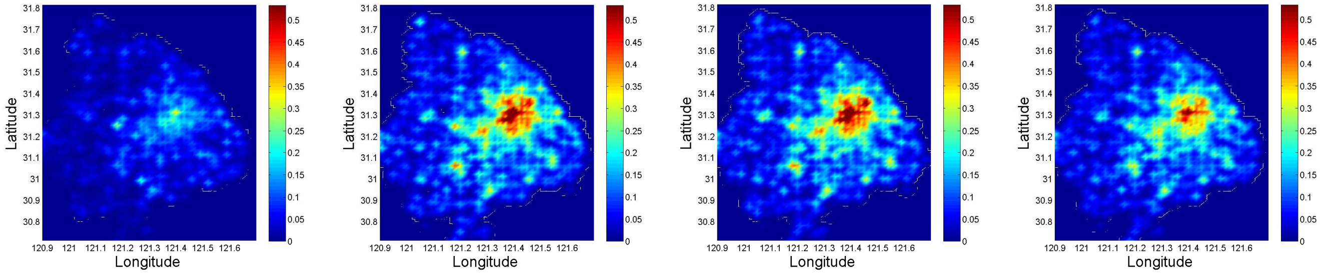
In order to answer these fundamental problems, in this paper, we carry out a study to understand and forecast the mobile traffic of large scale cellular data networks by a time series approach. Our deep and thorough investigation on large scale cellular data makes serval interesting observations about mobile traffic patterns. In addition, the contribution of this paper can be summarized as the following three parts.

- **Data:** We analyze a large scale of realistic mobile data traffic for duration of one month. The dataset is big in terms of tracking over 9,000 base stations and 150,000 mobile users. Such a big dataset allows us to better understand and model the cellular traffic in urban environment. To the best of our knowledge, we are the first to systematically study the mobile traffic patterns from large scale cellular data network.

F. Xu, Y. Lin, J. Huang, H. Shi, and Y. Li are with Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (E-mails: liyong07@tsinghua.edu.cn). D. Wu is with Hunan University (E-mails: d.wu@imperial.ac.uk). J. Song is with Sookmyung Women's University, Seoul, Korea.

This work is supported by National Key Basic Research Program of China (973 Program Grant Nos. 2013CB329001 and 2013CB329203), National Nature Science Foundation of China (Grant Nos. 61171065, 61021001 and 61133015), National High Technology Research and Development Program (Grant Nos. 2013AA010601 and 2013AA010605), and Program for Changjiang Scholars and Innovative Research Team in University.

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.



(a) 4AM (b) 10AM (c) 4PM (d) 10PM

Fig. 1. The spatial distribution of mobile traffic profile at different time.

- **Tools:** We design, implement, and evaluate a time series approach that is able to carry out large scale mobile traffic analysis. First, we use it to decompose the regular and random component of the mobile traffic. Then, we utilize the time series prediction and correlation to forecast the traffic patterns based on the regularity components.
- **Findings:** Our study reveals that mobile traffic patterns can be decomposed into regular and random components. We implement a time series approach to decompose the traffic patterns and we also verify the effectiveness of our method. In addition, we shows the geographical distribution of the regularity and randomness components. Moreover, we reveal a high predictability of the regularity component of the traffic, and demonstrate that the prediction of randomness component of mobile traffic data is impossible.

The rest of the paper is organized as follows. Section 2 introduces our motivation, data set, data preprocessing and basic observations from the data visualization. After discussing the main methodology of time series analysis based decomposition in Section 3, we present our main results of traffic forecasting for different components of the mobile traffic in Section 4 . Then, Section 5 shows the geographical distribution of the regular and random patterns of the mobile traffic. After that, we discuss related work in Section 6, and conclude this work in Section 7.

2 Data and Motivation

2.1 Dataset and Basic Observations

In this section, we provide details about the utilized dataset, and discuss how we extract the traffic information from the data. In addition, we also present some basic observations of mobile traffic characteristics obtained from the dataset.

2.1.1 Trace Description

We use an anonymized traffic consumption trace collected from the cellular networks of Shanghai, a large city of China, by one of the largest mobile operator during the whole month of August, 2014. The trace provides information about mobile traffic consumption by all mobile users, and each entry in the data set is corresponding to a continuous data communication, of which the device's ID, starting and ending time, base station (BS) ID, BS location and traffic volume are recorded. There are about 1.96 billion entries in the trace, including over 9,000 BSs and 150,000 subscribers. The total volume of traffic recorded in the dataset is over 2.8 PB, with over 7 GB per BS and 92 TB per day on average. Overall, this large-scale and fine-grained dataset guarantees the accuracy and credibility of our investigation.

2.1.2 Extracting Traffic Consumption

Based on the obtained dataset, a pre-processing procedure is conducted to sort traffic records by time and BS ID, and compute the data traffic consumption during a certain period within the same BS. Since the duration of one continuous data consumption varies from several minutes to several hours, we assume the traffic follows a uniform distribution in the interval and divide each duration into several ten-minute intervals with equivalent traffic, where intervals shorter than ten minutes are approximately regarded as ten-minute intervals. In addition, the start time of each interval is assumed to be located at specific moments, which are 10, 20, 30, 40 or 50 minutes away from the whole point of time. Thus, by this classification and summation, we obtain the traffic load of each BS with the minimum scale of 10 minutes for the whole 31 days. Moreover, to characterize the spatial distribution, we convert the address of BSs location to longitudes and latitudes through APIs provided by map service provider like Baidu. Adding up the traffic volume of BSs which are located at the same unit area, we obtain the spatial distribution of traffic.

2.1.3 Basic Observations

Our trace records the spatial and temporal distribution of the mobile traffic, and the spatial and temporal behaviors are interacted with each other. In order to show the characteristic of mobile traffic by combining the spatial and temporal features, we investigate the spatial distribution of mobile traffic profile at different time. Since the peak in the city center is too high, traffic at other place almost cannot be observed in the same figure with linear coordinate. Thus, we use the square root of the traffic density to represent its intensity. The cellular traffic density, which captures mobile users' behaviour in cyber space, should exhibit a strong periodicity due to human daily routine. Therefore, since the aggregated traffic load of different days do not change obviously, we plot the two-dimensional view of the traffic density of different time in one day in Fig. 1. The corresponding time is 4AM, 10AM, 4PM and 10PM,

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

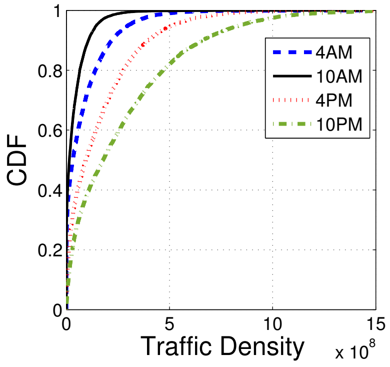


Fig. 2. The CDF of mobile traffic profile at different time.

respectively, whose cumulative distribution function(CDF) is shown in Fig. 2. From the results, we observe that the highest peak of traffic density located in the city center all over the day. When at 4AM, most areas are covered with relatively dark color, indicating that the traffic demand is small as most people are sleeping and their devices are inactive. Meanwhile, at 10PM, most areas of the city are covered with relatively light color, indicating that the traffic demand is large as most people and their devices are active. During this time, the peaks of traffic in some way showing the main concentrated areas of people, such as residential zone or central business district, spread over the city. There is one significant peak in the city center with many relatively low peaks all over the city. Then, the lightness of color shows a downtrend at 4PM and 10PM. Overall, there is a huge difference existing in the traffic distribution of these different hours, reflecting the extremely inhomogeneous traffic usages in terms of time and of different base stations. A more detailed traffic analysis and characterization can be found in our prior work [16].

2.2 Motivation

Our dataset records the information of mobile traffic consumption of about 9 thousand base stations spanning over one month, which quantitatively characterizes the spatio-temporal distribution of mobile traffic load and human traffic consumption behaviors. First, in order to demonstrate that it records the fundamental temporal patterns, we show the aggregated traffic load of the entire network at different time scales in Fig. 3, where hour scale shows the distribution of the aggregated traffic load for one day period and the week scale shows the distribution for the whole month. From the results, we can observe that the aggregated network load exhibits a stable periodic behavior tightly coupled with the sleeping and working routine of human. Overall, the traffic load is relatively high during the day and is lowest during midnight. The first sub-plot shows the daily traffic load of the entire network in the scale of weeks. Though not as much obvious as the periodicity of one day shown in the day scale, their daily traffic load exhibits a weekly periodic regularity. For example, the traffic load in weekends is obviously less than the traffic load in weekdays, which is resulted from the work schedule of humans. Overall, in terms of the temporal domain, our

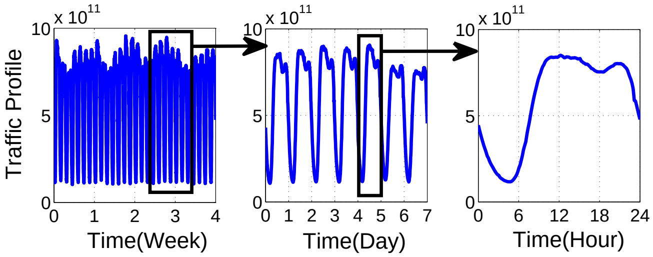


Fig. 3. Traffic profile of the whole network in three different time scales of a week, day and hour.

trace records different level of grains for traffic patterns at different time scales.

Above observed time-domain periodicity of traffic motivates our decomposition approach, i.e., we utilize the time series analysis to extract the regularity and randomness components of the traffic. A natural question to ask is what are the most discriminating and essential features that are able to present such traffic patterns. Motivated by answering this question, we conduct time series analysis on the obtained time-domain traffic series and reveal several important discoveries.

Another important aspect we investigate is the traffic prediction. As seen, the traffic pattern differs between weekday and weekend significantly. Although it appears to be irregular in weekend, the pattern also shows strong periodicity in weekday, which is contributed by human regular commute between home and working place. However, such pattern may not show up in other base stations where user number pattern varies significantly across days. Therefore, another key question we have to address is that how to predict and model the traffic from such complicate user number patterns despite of its regularity and stochastic components?

As for one base station, the traffic's stochastic component accounts for a significant proportion, which makes the prediction not accurate. However, for the base stations with the knowledge of the regularity and randomness, the random component is much less and it is more purposeful. Thus, with the decomposition results, we are able to carry out more precise mobile traffic prediction according to the variations and profiles type.

3 Mobile Traffic Decomposition

As we can observe from Fig. 3, the patterns of mobile traffic have strong regular components as well as considerable random components. In order to better model and forecast the covered traffic patterns, we are motivated to leverage a classic time series decomposition method to decompose the traffic patterns into regular components and random components [23].

3.1 Decomposition Method

We denote the original traffic pattern of one base station as $T = \{x_1, x_2, x_3, \cdots, x_n\}$ with x_n representing the traffic

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2016.2599878, IEEE Transactions on Services Computing

volume in the n_{th} time slot.Then, we define x_t as combination of u_t and y_t,

$$x_t = u_t + y_t, \;\; t = 1, 2, \cdots, n, \tag{1}$$

where the u_t is the seasonal component with regular patterns and y_t is the stochastic component that fluctuates around the seasonal components without periodicity. In classic time decomposition approach, time series can be decomposed into three components as follow,

$$x_t = m_t + s_t + r_t \;, \;\; t = 1, 2, \cdots, n, \tag{2}$$

where m_t shows a general trend of the series, s_t is the periodic patterns of the series and r_t represents the stochastic component of the series. In this paper, we define the seasonal component as the sum of m_t and s_t, and stochastic component as r_t,

$$\begin{aligned} u_t &= m_t + s_t \;, \;\; t = 1, 2, \cdots, n, \\ y_t &= r_t \;, \;\; t = 1, 2, \cdots, n. \end{aligned} \tag{3}$$

To decompose the traffic patterns, we first roughly estimate the trend by applying a moving average filter,

$$\hat{m}_t = \left(0.5x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + 0.5x_{t+q}\right)/d \;, \quad q < t \le n - q, \tag{4}$$

where d is a time window we set and $q = d/2$. Then we compute the average w_t of the deviations $\{(x_{k+jd} - \hat{m}_{k+jd}) \,,\, q < k + jd \le n - q\}$ for each $k = 1, \cdots, d$. Afterwards, we compute the periodic pattern s_k as follows

$$\begin{cases} \hat{s}_k = w_k - d^{-1} \displaystyle\sum_{i=1}^{d} w_i, \;\; k = 1, \cdots, d, \\ \hat{s}_k = \hat{s}_{k-d}, \;\; k > d. \end{cases} \tag{5}$$

With the obtained periodic pattern s_k, we define the remaining data as

$$d_t = x_t - \hat{s}_t, \;\; t = 1, \cdots, n. \tag{6}$$

Now, we can estimate the general trend better by applying moving average filter on d_t. Let q be a nonnegative integer. Since we want to observe the traffic data for $0 < t \le n$, we define $x_t = x_1$ for $t < 1$ and $x_t = x_n$ for $t > n$, and

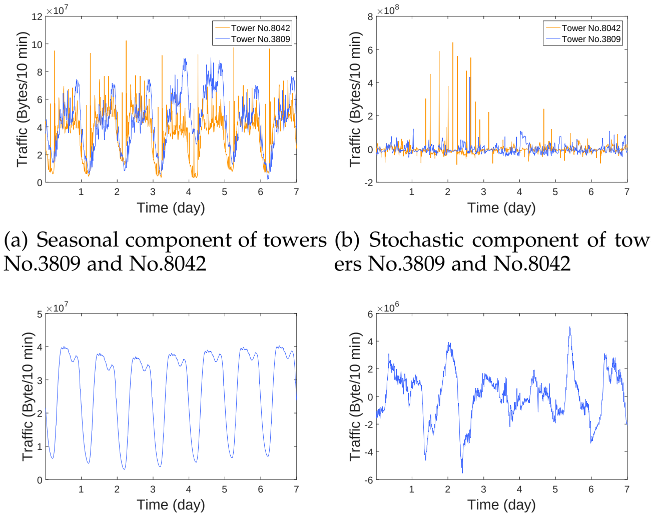
$$m_t = (2q + 1)^{-1} \sum_{i=-q}^{+q} x_{t-j}. \tag{7}$$

Finally, combining the x_t, m_t and s_t, we have

$$\begin{cases} u_t = s_t + m_t, \\ y_t = x_t - u_t. \end{cases} \tag{8}$$

3.2 Evaluation for Decomposition

To evaluate the performance of our decomposition method, we present the decomposition result of traffic series of both the overall and individual base stations. Fig. 4(a) and Fig. 4(b) show the seasonal component and stochastic component of two randomly selected cellular towers separately. Fig. 4(c) and (d) show these two components of all cellular towers. From the results, we can observe that, the seasonal component shows a strong periodicity both in overall and in specific base stations, while the stochastic component is much more irregular.



(a) Seasonal component of towers No.3809 and No.8042

(b) Stochastic component of towers No.3809 and No.8042

(c) Average of seasonal component of traffic of all towers

(d) Average of Stochastic component of traffic of all towers

Fig. 4. Cellular traffic decomposition result.

To further quantitatively evaluate the effectiveness of our time decomposition method, we carry out an analysis on the randomness of the stochastic component by adopting autocorrelation function to evaluate the periodicity of a time series. If a time series $Y_t = \{y_1, y_2, y_3, \cdots, y_n\}$ is an random sequence, about 95% of its autocorrelations should fall between the bounds $\pm 1.96/\sqrt{n}$.

We show the results in Fig. 5(a) and Fig 5(b), where the red curve represents the average of autocorrelation of raw data and stochastic component of overall traffic series, and the blackish green area covers corresponding standard deviation, while the blue line shows the bounds of $\pm 1.96/\sqrt{n}$. From the results, we can find that the autocorrelation of raw data goes up and down with a certain period. On the contrary, the autocorrelation of stochastic component are much lower and its shape resembles Dirac delta function. With more than 95% average autocorrelations fall between the bounds in Fig 5(b), we infer that the internal dependence among stochastic component of the same traffic pattern is weak, which means that the stochastic component is indeed an random sequence. Fig 5(c) describes the distribution of autocorrelations of the stochastic component of all towers at different lags. It indicates that the autocorrelation of stochastic component of different towers maintains low in general, no matter how long the lag is. It means that the stochastic components are quite random in both short-term and long-term. These observations demonstrate that decomposition method we utilized works well in decomposing the original traffic series into seasonal and stochastic component.

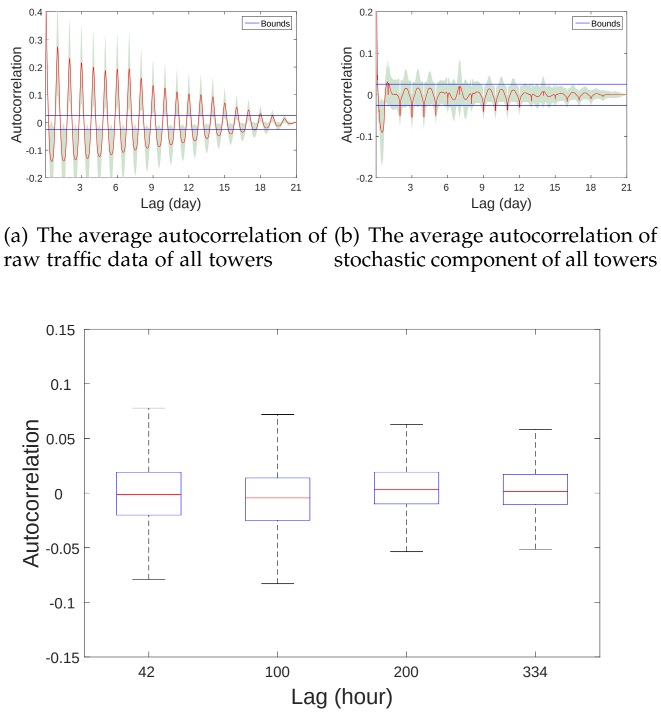
4 Prediction Analysis

4.1 Prediction of Regularity Component

In this subsection, we devote to model and forecast the traffic behaviour of cellular towers, by concentrating on in-

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2016.2599878, IEEE Transactions on Services Computing



(a) The average autocorrelation of raw traffic data of all towers

(b) The average autocorrelation of stochastic component of all towers

(c) The autocorrelation of stochastic component at certain lags

Fig. 5. The autocorrelation of cellular traffic decomposition.

vestigating the regular component of the traffic: the seasonal component. The regular component contains the trend and cyclical component of cellular traffic with the period of one day. Indeed, there are differences between the different days in one week. Thus, strictly speaking, it also has periodicity of a week.

In order to forecast regularity component, we utilize a Autoregressive Integrated Moving Average (ARIMA) model[6]. More specifically, to exploit the seasonal terms in the data, we utilize a seasonal ARIMA model with general form of ARIMA $(p,d,q) \times (P,D,Q)_m$, where m is the number of periods per season, (P,D,Q) are the seasonal parts of the model with P as the number of seasonal autoregressive terms, D as the number of seasonal differences, Q as the number of seasonal moving average terms, and the lowercase notations (p,d,q) are used for the corresponding non-seasonal parts of the model. The equation is shown as follow.

$$\overbrace{(1-ar_1 z)}^{\text{AR}}\overbrace{(1-sar_1 z^{24\times7})}^{\text{SAR}}\overbrace{(1-z^{24\times7})}^{\text{Seasonal Differenced}}\ y_t = \underbrace{(1+ma_1 z+ma_2 z^2+ma_3 z^3)}_{\text{MA}}\underbrace{(1+sma_1 z^{24\times7})}_{\text{SMA}}\,x_t \tag{9}$$

To evaluate our model's performance, we use the former three weeks' traffic to predict the forth week's traffic. To proceed, we need to see the ACF(autocorrelation function) and PACF(partial autocorrelation function) of the data. However, with the existence of an obvious period, the ACF and PACF of the data is meaningless, for they are defined for the so called stationary time series which means that the autocorrelation for any particular lag is the same regardless of where we are in time. The way out is to do the seasonal difference analysis. Based on the knowledge in classic time series analysis, we know that the behavior of the PACF decides the order of the AR terms and the behavior of the ACF decides the MA terms.

The parameters obtained from the above analysis reveal that the predicting result fits the true profile very well for different types of the BS. We take two randomly selected BS to demonstrate the effectiveness of our model. In Table 1, the trained error of the parameters is very small in both base stations, which suggests that our model predicts the traffic behaviour accurately. In addition, we show the comparison of the predicting results and their original profiles of the selected base stations in Fig. 6. We can observe that the predicting profile is very close to the original profile. Observing Fig. 7, we can know that 80% of the absolute value of relative error is less than 30% in both base stations, which suggests that the predicting results are accurate.

TABLE 1
Parameters of the ARIMA model of two different type BS.

(a)

Parameter	Value	Error	Statistic
Constant	0	Fixed	Fixed
AR1	0.972823	0.720408	1.35038
SAR1	0.972823	0.720408	1.35038
MA3	-0.017419	4.59487	-0.00379096
SMA1	-0.872813	3.11263	-0.28041
Variance	2.22149e+13	3.49061e-13	6.36418e+25

(b)

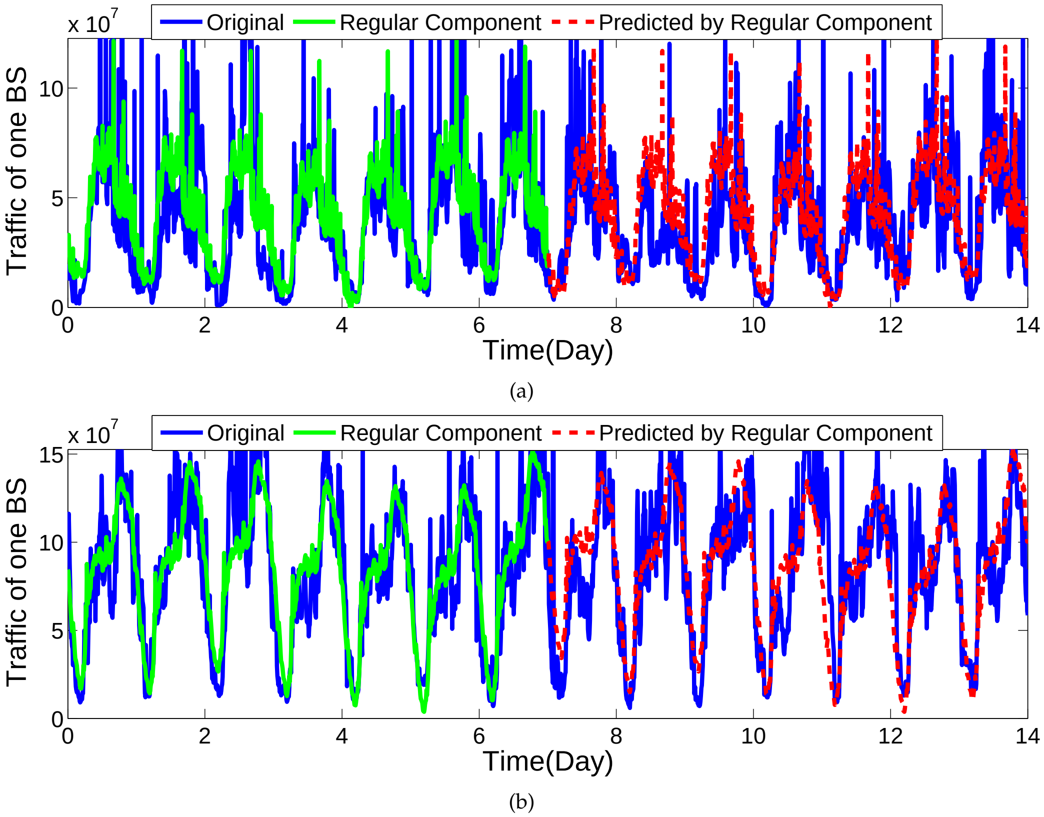
Parameter	Value	Error	Statistic
Constant	0	Fixed	Fixed
AR1	0.96773	5.74149	0.16855
SAR1	0.96773	5.74149	0.16855
MA3	-0.0590281	39.4714	-0.00149547
SMA1	-0.435804	16.7299	-0.0260494
Variance	5.07796e+13	2.19945e-12	2.30875e+25

4.2 Prediction of Stochastic Component

Besides relatively fixed daily activities, stochastic events also happen every day that affects the mobile traffic pattern. In order to investigate whether the stochastic component can be predicted by the stochastic component of other towers several minutes ahead of time, we turn to check whether there is a high cross correlation between the stochastic component of the mobile traffic data of a tower and that of the towers in its adjacent area. By applying Delaunay Triangulation[24] to the location of cellular towers, we obtain 23964 pairs of adjacent cells. We use the sample cross correlation function in equation(10) to estimate the cross correlation of the stochastic components of every pair.

$$c(k) = \begin{cases} \frac{1}{T}\sum\limits_{t=1}^{T-k}\left(y_{1,t}-\bar{y_1}\right)\left(y_{2,t+k}-\bar{y_2}\right), k = 0,1,2\cdots; \\ \\ \frac{1}{T}\sum\limits_{t=1}^{T+k}\left(y_{2,t}-\bar{y_2}\right)\left(y_{1,t-k}-\bar{y_1}\right), k = 0,-1,-2\cdots. \end{cases} \tag{10}$$

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.



(a)

(b)

Fig. 6. The Comparison of the Predicted Result and the original of two different type BS.

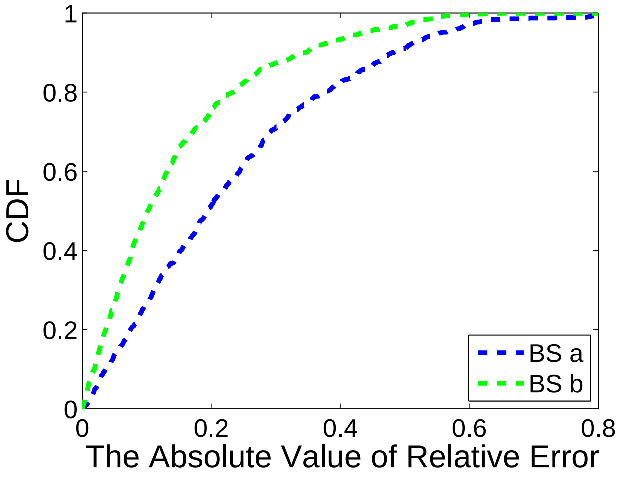


Fig. 7. The CDF of the Relative Error of the Predicted Results.

In the formula, c(k) is the cross correlation of sequence y_1 and y_2 at lag=k. T is the number of terms of the sequence, $y_{i,t}$ is the t^{th} term in sequence y_i, and $\bar{y_i}$ is the average of the term value in sequence y_i. This function reflects the cross correlation of mobile traffic sequence A with sequence B shifted k lags. Each lag represents a period of 10 minutes. Suppose a sudden event happens in the cell of tower A, spreads to places around and reaches the cell of tower B, then there is a possibility that the traffic sequence of B several time(lags) later has a high correlation with the traffic sequence of A.

4.2.1 Low Cross Correlation Between Towers

To obtain a global view of the correlation of stochastic component, we first calculate the average absolute cross correlation of the traffic sequence in a single day of 23964 pairs in Fig. 8, where the red line is the average of cross correlation at different lags. From the results, we can observe that there is a good symmetry and the peak value is at lag=0. However, the value of 0.14 is rather low, which implies that the stochastic component of adjacent cellular towers are nearly unrelated. In addition, the results show a large standard deviation between different tower pairs, which is painted grey. If a group of people pass through adjacent cellular towers, they might contribute to the cross correlation of the traffic data by linking their mobile phones to the Internet and use the Internet while traveling from one tower to another. This kind of events is the underlying behavior for predicting stochastic components. The low mean value of the cross correlation tells us, however, that either this kind of events happens in a very small amount, or they do not account for the main causes, i.e., some people might not have the habit of using the Internet while moving from place to place.

4.2.2 Spatial Variation of Cross Correlation

To investigate the distance at which the traffic of a tower can influence another nearby one, we study how the cross correlation change as the distance between two cellular towers increases. Fig. 9(a) shows the distribution of the

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2016.2599878, IEEE Transactions on Services Computing

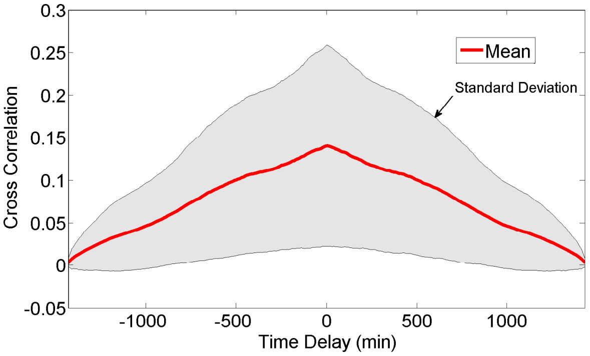


Fig. 8. The cross correlation of all pairs of towers in a single day

cross correlation between the stochastic component of each tower and the tower in its n-step neighborhood. An n-step neighborhood means that it takes a person to travel at least n adjacent towers to reach this tower. Each column is the data set of all cross correlation between the stochastic component of traffic of a tower and one of its n-step neighbor. We can imply from the graph that instead of descending as n grows, the correlation maintains at a rather low level.

We also choose two individual cellular towers to do case studies. Fig. 9(b) shows the cross correlation of an individual tower and its n-step nearby cellular towers which are all located in the city center. The result shows that even though the trend of the median of the correlation goes down as the step increases, there exists a huge gap between the maximum and minimum. An investigation on another cellular tower in the residence area gave us a totally different answer as is shown in Fig. 9(c). The cross correlation does not decrease as the spatial range increases. Instead, all steps of neighborhoods look almost the same.

Thus we conclude that the cross correlation of stochastic component does not necessarily decreases as the distance increases, which indicates that nearby cellular towers do not play a significant role in affecting its stochastic component, and the correlation between them is spatial-variant.

4.2.3 Time Variation of Cross Correlation

We investigate the time variance of cross correlation of adjacent cellular towers. Fig. 10(a) shows the cross correlation of the stochastic component of traffic data of two adjacent towers(No.5000 and No.5029) in three separate days(two days in the first week, and the third day in two weeks). The cross correlation in these three days differ from each other greatly, which shows that the influence between stochastic component of mobile traffic of adjacent towers is time-variant.

To more detailedly and comprehensively reflect both the time and spatial variation of cross correlation, we work on all 23964 pairs of adjacent cellular towers. We randomly selected the traffic data from two days and investigate on the cross correlation at lag=1. Fig. 10(b) shows the distribution of the cross correlation of tower pairs. Most of the pairs have a correlation of less than 0.3, which indicates a weak correlation. The number of towers decreases as the correlation goes up, and few tower pairs exists on the right part of the figure. To find out whether this correlation varies as time changes, we compared the correlation of this day with that of another randomly picked day and found that the cross correlation of 68% of the pairs appear in different correlation intervals, as is shown in the pie chart in Fig. 10(c). More importantly, from the pairs of towers that have correlation higher than 0.5 in either of the two days, only 5% of them remain higher than 0.5 in the other day, and 95% just fall back to lower than 0.5. From these observations, we can draw the conclusion that the cross correlation between adjacent pairs is time-variant.

In conclusion, by obtaining the result that the spatial and time cross correlation between the stochastic component is very low, we prove the unpredictability of stochastic component of mobile traffic data.

5 Geographical Distribution of Prediction

Since there is difficulty in predicting the stochastic component of the traffic, the larger the proportion of seasonal component covers, the more predictable the traffic series is. To measure the intensity of seasonal component u_t in a traffic series $X_t = \{x_1, x_2, x_3, \cdots, x_n\} = u_t + y_t$, we define the Power Ratio of the tower as follow,

$$R = \frac{\sum u_t^2}{\sum y_t^2} \tag{11}$$

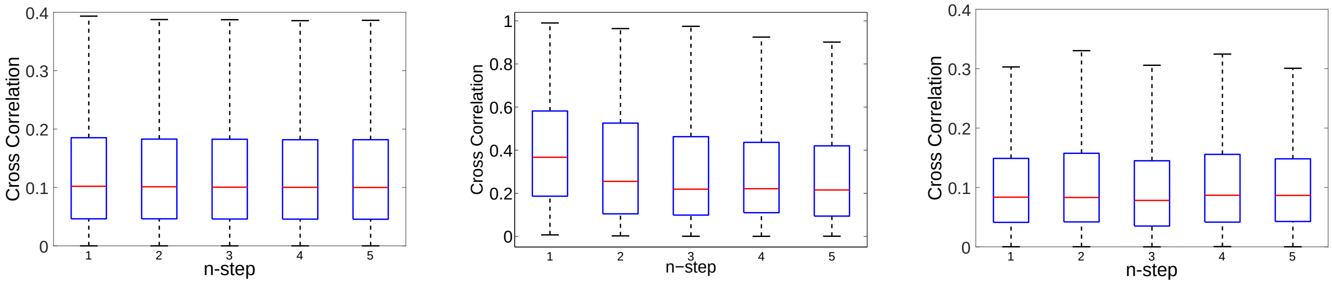
Therefore, the larger the power ratio of one cellular tower is, the stronger periodicity its pattern shows and consequently the more predictably its traffic series inhabits.

5.1 Frequency Distribution of Power Ratio

It is provoking to observe the frequency distribution of power ratio in a general perspective at first. A frequency distribution is a chart that displays the frequency of various outcomes in a sample. Fig. 11 describes frequency density in different intervals of power ratios of all cellular towers. In the figure, the peak of frequency distribution indicates the mode of power ratio and the area under the curve correspond to the frequency of interval. From the results, we can find that the cellular towers with power ratio $R \approx 0.45$ covers the highest proportion of all, while the average of all power ratio is 2.4392. Power ratio of half the towers disperse between 0.67 and 3.27. It is suggested that the traffic series of most cellular towers behave similarly and their predictability does not overweigh the randomness. However, there are some particularly high power ratio distributing sparsely from 10 to 45. Therefore, we have to inspect these ”special” cellular towers in combination with those towers with lowest power ratio. A crucial property of cellular tower is the geographical location, which may be related to its traffic pattern.

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2016.2599878, IEEE Transactions on Services Computing

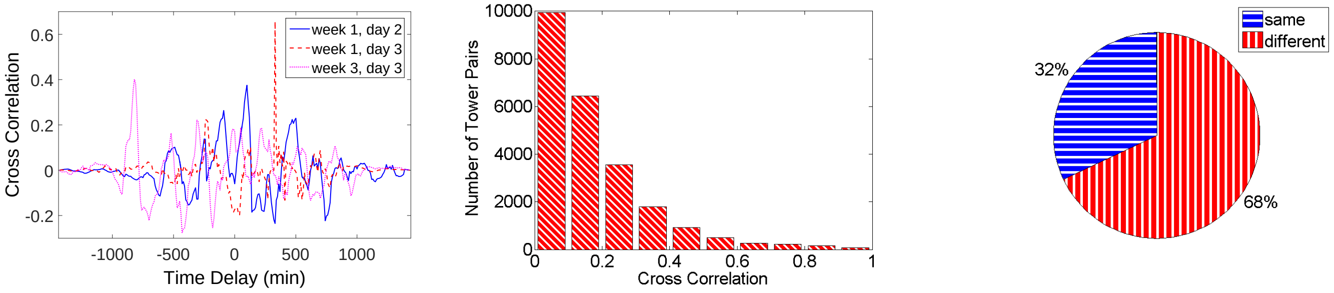


(a) All towers

(b) No.8398

(c) No.5000

Fig. 9. Cross correlation between a tower and its n-step neighbors

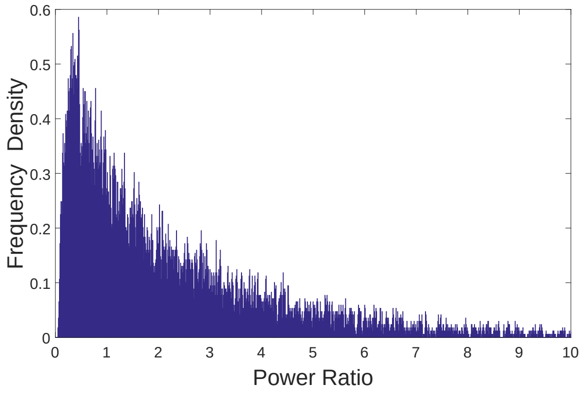


(a) cross correlation between two towers

(b) cross correlation distribution histogram

(c) the proportion of same and different correlation

Fig. 10. The time variation of cross correlation



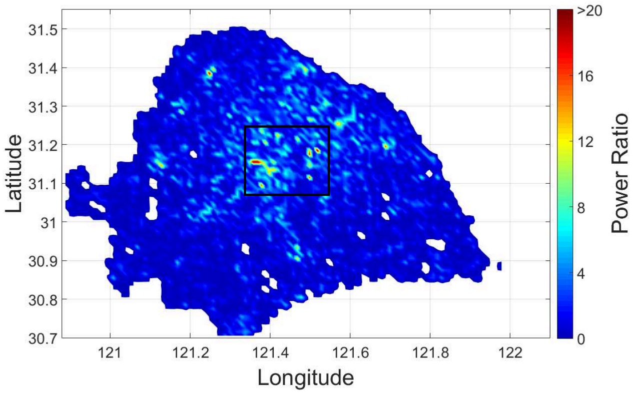


Fig. 11. Frequency distribution of power ratio of all cellular towers.

Fig. 12. The spatial distribution of power ratio of cellular tower.

5.2 Geographical Distribution of Power Ratio

We first investigate the geographical distribution of power ratio illustrated by Fig. 12. As shown in the color bar, the brown red means the extremely high power ratio, while the dark blue indicates the low power ratio and the green stands for the comparatively high power ratio. The results show that towers deployed at the center of the city are more likely to owns comparatively high power ratio while most areas of city are in the middle level.

In order to understand the characteristics of cellular towers with high power ratio deeply, we select those whose power ratio is higher than 99.5% towers and those that less than 99.5% to make comparison. Fig. 13 shows the locations of these two types of cellular towers. In Fig. 13, black boxed area is identical to the boxed area in Fig. 12, and the red points stand for the towers with power ratio higher than 99.5% towers while the purple points stand for the towers with power ratio lower than 99.5% towers . We can observe that the cellular towers marked by red color surround the center of city, and most of them are located at shopping malls and large parks. It is demonstrated that the towers with extremely high power ratio is highly concentrated in the place with greater population density and mobility. On the contrary, the towers marked by purple color are decentralized and gradually away from the city center.

To obtain intern relationship between location and power ratio of cellular tower, we adopt points of interest (POI) distribution to approach an accurate analysis. POI is a

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2016.2599878, IEEE Transactions on Services Computing

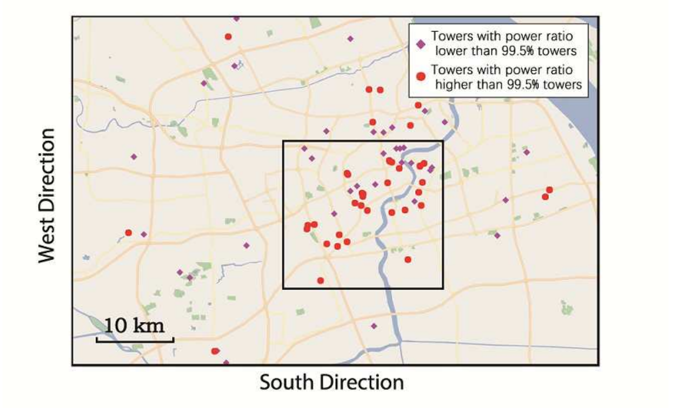

Towers with power ratio lower than 99.5% towers
Towers with power ratio higher than 99.5% towers
West Direction
South Direction
10 km

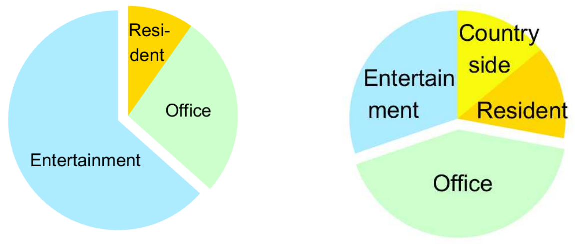
Fig. 13. The location of towers with power ratio higher than 99.5% towers and lower than 99.5% towers.

Location Types	High Power Ratio		Low Power Ratio	
	Amount	Proportion	Amount	Proportion
Entertainment	26	63.41%	13	30.23%
Office	11	26.83%	18	41.86%
Resident	4	9.76%	6	13.95%
Countryside	0	0%	6	13.95%

TABLE 2
Amount and Proportion of POI types in two types of cellular towers

specific location of a certain function such as business center. An area's POI distribution is in accord with its function. The POI data we use is provided by Baidu Map. We categorize all POI kinds into four main types, which are Entertainment, Office, Resident, Countryside. Consider a circular area with a radius of 200 meter, as the location of the tower, we regard its function as a representative of function of tower's location. We calculate POI distribution of those locations shown in Fig. 13 and summarized in Table 2 as well as Fig. 14.

From Table 2 and Fig. 14, we can obtain several observations. First, 63.41% cellular towers with extremely high power ratio lie in the entertainment area while the percentage is approximately halved when it comes to the cellular towers with extremely low power ratio. Taking their locations shown in Fig. 13 into account, the cellular towers which locates in the entertainment area at center of city are more likely to own high power ratio, which suggests that they are more predictable. The potential reasons are that the population and density of urban central area are 1.40 to 2.48 times larger than the average level of whole city, so that the population mobility exerts slighter influence on general traffic usage of one certain cellular tower. Besides, the regular opening hours of shops and stores in the entertainment area contributes to the periodicity of cellular traffic. On the other hand, the percentage of cellular towers with extremely low power ratio in the office areas arises, compared with that of towers with extremely high power ratio. Finally, the cellular towers in countryside occurs and covers nearly 14% towers with extremely low power ratio. This shows that cellular towers in countryside area behave more unpredictably. Lower population density and scattered functional regions in suburbs, may increase the traffic fluctuations influenced by population mobility and individual behavior.



(a) Pie chart of POI distribution of cellular towers with high power ratio

(b) Pie chart of POI distribution of cellular towers with low power ratio

Fig. 14. Pie chart of POI distribution of two types of cellular towers.

6 Related Works

The digital footprints of human activities in cellular network contributed by mobile devices have enabled various investigations range from human behaviour to network dynamics [8], [9]. In terms of human behaviour, cellular traces significantly benefited the study in human mobility in the past decades. Deville [10] developed a mechanism that is able to estimate the population density at national scales mobile phone call records. Ficek [11] proposed a model based on statistical characteristics of intervals between phone calls to obtain fine grained user mobility. Trestian [12] modeled the correlation between popular apps used by people and the mobility patterns of users based on investigating 3G network access fingerprints of over 280,000 users. In addition, other studies about the mobility of individual person have been conducted based on taxicab location logs [14], WiFi access fingerprints [13], fusion of multiple sources [15], etc. Barabási [19], [21], [20], [22] studies the long-term mobility of individuals based on a six months' phone call record across 100k users. Besides human mobility, cellular dataset has also been utilized to enable many other study in human behaviours. Cici [3] studied urban ecology by investigating cell phone activity patterns. Dong [4] utilized mobile dataset to infer user demographics and social strategies. In this paper, we investigate a large-scale cellular data traces to study the regularity and randomness of cellular data traffic patterns, which provides insight to understand and forecast cellular traffic.

In terms of network dynamics, the cellular traces also benefit a lot of study. Cellular traffic patterns have been extensively studied to model and understand its characteristics in the past few years. For example, Lee [17] revealed that in spatial domain the traffic density can be best approximated by a log-normal or Weibull distribution. Wang et al. [18] found that mobile traffic obeyed a trimodal distribution on both spatial and temporal dimensions. Zhang et al. [7] tried to understand the characteristics of cellular data traffic by comparing it to wireline data traffic. In addition, Shafiq et al. [5] characterized and modelled the internet traffic dynamics of cellular traces. Gao et al. [25] proposed a hourglass co-clustering model to profile 3G network dynamics. Unlike the previous work, in this paper we focus on investigating the regularity and randomness of cellular traffic

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2016.2599878, IEEE Transactions on Services Computing

patterns, which provides us a new angle to understand and analyze the cellular network dynamics.

In conclusion, we carry out a study focusing on understanding and analyzing the regularity and randomness of mobile traffic based on a large scale fine-grained cellular network traces in this paper. We believe it provides an angle to understand the characteristics of mobile traffic as well as a valid method to forecast mobile traffic consumption.

7 Conclusions

In this paper we carry out, to the best of our knowledge, the first large-scale study to investigate the 3G/LTE mobile traffic patterns in terms of regularity and randomness. Our study provides a powerful time series analysis approach to investigate traffic patterns of millions of cellular towers, and reveals that the dynamic urban mobile traffic consumption exhibits highly predictable regularity patterns and non-predictable stochastic component. These findings provide a systematic, comprehensive, while simple angle to understand the dynamic and complicated mobile traffic. Thus, we believe our analysis demonstrate significant promise to further research in this area, and open a new research angle in understanding mobile data consumption behaviors. In the future work, we are going to exploit the understanding we have on cellular data traffic to study the optimization of mobile network resources allocation. We believe that a next-generation efficient mobile system can be build based on a deeper understanding of mobile data traffic pattern.

Acknowledgment

This work is supported by National Key Basic Research Program of China (973 Program Grant Nos. 2013CB329001 and 2013CB329203), National Nature Science Foundation of China (Grant Nos. 61171065, 61021001 and 61133015), National High Technology Research and Development Program (Grant Nos. 2013AA010601 and 2013AA010605), and Program for Changjiang Scholars and Innovative Research Team in University.

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019 White Paper, Feb. 3, 2015.

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

[3] 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. ACM MobiHoc* (Hangzhou, Chin), Jun. 22-25, 2015, pp. 317–326.

[4] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, “Inferring user demographics and social strategies in mobile social networks,” in *Proc. ACM SIGKDD* (New York, USA), Aug. 24-27, 2014, pp. 15–24.

[5] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, “Characterizing geospatial dynamics of application usage in a 3G cellular data network,” in *Proc. IEEE INFOCOM* (Orlando, FL), Mar. 25-30, 2012, pp. 1341–1349.

[6] G. P. Zhang, “Time series forecasting using a hybrid arima and neural network model,” *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[7] k. A. Ying Zhang. Understanding the characteristics of cellular data traffic. In *ACM SIGCOMM CellNet Workshop*, 42(4):13–18, 2012.

[8] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proc. of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[9] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, et al. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, number EPFL-CONF-192489, 2012.

[10] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, “Dynamic population mapping using mobile phone data,” *Proc. National Academy of Sciences*, vol. 111, no. 45, pp. 888-893, 2014.

[11] M. Ficek and L. Kencl, “Inter-call mobility model: a spatio-temporal refinement of call data records using a gaussian mixture model,” in *Proc. IEEE INFOCOM*, 2012, pp. 469-477.

[12] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, “Measuring serendipity: connecting people, locations and interests in a mobile 3g network,” in *Proc. ACM SIGCOMM*, 2009, pp. 267–279.

[13] M. Kim, D. Kotz, and S. Kim, “Extracting a mobility model from real user traces,” in *Proc. IEEE INFOCOM*, 2006, pp. 1-13.

[14] R. Ganti, M. Srivatsa, A. Ranganathan, and J. Han, “Inferring human mobility patterns from taxicab location traces,” in *Proc. ACM UbiComp*, 2013, pp. 459-468.

[15] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, “Exploring human mobility with multi-source data at extremely large metropolitan scales,” in *Proc. ACM MOBICOM*, 2014, pp. 201-212.

[16] H. Wang, F. Xu, Yong Li, P. Zhang, D. Jin, “Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment”, in Proc. *ACM IMC 2015*.

[17] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang. Spatial modeling of the traffic density in cellular networks. *Wireless Communications, IEEE*, 21(1):80–88, 2014.

[18] H. Wang, J. Ding, Y. Li, P. Hui, J. Yuan, and D. Jin. Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks. In *Proc. of ACM HOTPOST*, pp. 19-24, 2015.

[19] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, “A universal model for mobility and migration patterns,” *Nature*, vol. 484, no. 7392, pp. 96-100, 2012.

[20] C. Song, T. Koren, P. Wang, and A.-L. Barabási, “Modelling the scaling properties of human mobility,” *Nature Physics*, vol. 6, no. 10, pp. 818-823, 2010.

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

[22] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding individual human mobility patterns,” *Nature*, vol. 453, no. 7196, pp. 779-782, 2008.

[23] M. West, “Time series decomposition,” *Biometrika*, vol. 84, no. 2, pp. 489-494, 1997.

[24] Y. Jiang, Y. Liu, and F. Zhang. An efficient algorithm for constructing Delaunay triangulation. *International Conference on Information Management and Engineering, IEEE*, vol. 3, pp. 600-603, 2010.

[25] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao. Profiling users in a 3g network using hourglass co-clustering. In *Proc. of ACM MobiCom*, pages 341C352, 2010.

1939-1374 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.