IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 28 July 2014; revised 31 October 2014; accepted 14 December 2014. Date of publication 12 January, 2015;
date of current version 4 September, 2015.

Digital Object Identifier 10.1109/TETC.2015.2389614

Computing on Base Station Behavior Using Erlang Measurement and Call Detail Record

SIHAI ZHANG, (Member, IEEE), DANDAN YIN, YANQIN ZHANG, AND WUYANG ZHOU

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui Province, 230026, China

CORRESPONDING AUTHOR: S. ZHANG (shzhang@ustc.edu.cn)

This work was supported in part by Natural Science Foundation of China under Grant (61461136002), in part by the National Programs for High Technology Research and Development under Grant 2012AA01A50603 and Grant 2014AA01A707, in part by the Key University Science Research Project of Anhui Province under Grant BJ2100060031, in part by the Fundamental Research Funds for the Central Universities under Grant under Grant WK2100060015, and in part by the Intel Collaborative Research Institute for Mobile Networking and Computing Sponsorship.

ABSTRACT With the impressive development of wireless devices and growth of mobile users, telecommunication operators are thirsty for understanding the characteristics of mobile network behavior. Based on the big data generated in the telecommunication networks, telecommunication operators are able to obtain substantial insights using big data analysis and computing techniques. This paper introduces the important aspects in this topic, including data set information, data analysis techniques, and two case studies. We categorize the data set in the telecommunication networks into two types, user-oriented and network-oriented, and discuss the potential application. Then, several important data analysis techniques are summarized and reviewed, from temporal and spatial analysis to data mining and statistical test. Finally, we present two case studies, using Erlang measurement and call detail record, respectively, to understand the base station behavior. Interestingly, the night burst phenomenon of college students is revealed by comparing the base stations location and real-world map, and we conclude that it is not proper to model the voice call arrivals as Poisson process.

INDEX TERMS Wireless big data, mobile communication networks, traffic analysis, spatial-temporal correlation.

I. INTRODUCTION

The world has witnessed the increasing popularity of mobile devices, which have become essential for common people in their daily life. Accordingly telecommunication operators are facing the tremendous challenge to provide satisfactory service to mobile users with varying QoS requirement, including high volume media transmission, huge amount of machine to machine (M2M) connectivity and etc. Wireless transmission technologies, like massive MIMO and cooperative communications, are being focused to deal with such challenge. On the other hand, the wireless network related techniques play even more significant role, like network management, energy saving, super dense cell deployment and etc.

The mobile communication networks have experienced network-oriented, user-oriented stages, and now the data-oriented stage is coming. So at this point, the hot topics about the wireless big data related promising research and applications are, How to understand the behavior of wireless/mobile networks? What benefit can such understanding bring to improve the network performance? What metrics and analysis tools are of vital importance to obtain these understandings?

As a matter of fact, the data generated by such huge amount of mobile devices has been verified to be of great social and economical value, thus can be utilized as low-cost but efficient tool in many aspects, such as mobile phone positioning for intelligent transportation systems [1], trajectory mapping for disease control [2] and individual detection [3]. In order to support the increasing demand, it's necessary to study the insight that mobile big data shows to us. Many studies have already pointed out that a good

knowledge of spatial and temporal dynamics in mobile networks can contribute to the understanding of subscriber behaviors and mobile networks. To be more specific, optimal pricing scheme, network and protocol design, spectrum allocation and energy saving plan can be promoted [4], so that mobile phone users are benefited with better service.

With the aid of modern information technology, collection and analysis of large-scale mobile communication traffic becomes possible and convenient, thus bring forth many empirical studies and significant outcomes. For example, as to the traffic modelling topic, many works have tried to answer the following most cared questions: what is the proper model for traffic arrivals in mobile networks? Or is there any long-range or short-range time-correlation of this traffic?

Here we take the call arrival modeling as example. In classic wired networks call arrival is modeled as Poisson process, which stimulates to model the call arrivals in wireless network also as Poisson process in previous works [5]–[8]. However unlike wired networks, correlation between users, congestions due to limited number of wireless channels and handover during calling will make call arrivals in cellular network more difficult to be modeled. So, based on the wireless big data, researchers may present reliable explanation on such kind of question.

Our contributions in this paper are three folds.

- We summarize the prevailing data analysis techniques for mobile communication big data, discuss their potential and possible usage or applications in mobile networks.
- We are pioneering, if not first, to investigate the traffic pattern and infer corresponding reason using large-scale aggregated voice data. The 'Night Burst' temporal pattern is revealed by performing K-means clustering and its reason is confirmed that they locate around university campuses.
- We investigate the call arrival issue using CDR data, and conclude that short-range call arrivals is independent, but long-range time-correlation exists. In addition, we find that it is improper to model the call arrivals in one hour as Poisson distribution, for almost in any time there are cases that the call arrivals fail the Chi-Square test for all stations.

The rest of this paper is organized as follows: Section II introduces the category of user-oriented and network-oriented, discuss popular and commonly used data analysis techniques. Section III presents our work based on Erlang measurement, including spatial and temporal analysis. Section IV investigates the time correlation issue using the CDR data. And conclusions are given in Section V.

II. DATA ANALYSIS TECHNIQUES

Because the telecommunication networks spread over large geographical areas and serve huge amount of mobile users in long time period, thus temporal and spatial analysis is commonly used to find out the interesting patterns in the time-space dimensions. And then statistical test and data mining techniques also play important role in discussing the rationality and validity. So in this section these topics will be presented briefly.

A. USER-ORIENTED AND NETWORK-ORIENTED DATA

Basically, in the telecommunication networks the data source can be categorized into user-oriented and network-oriented, corresponding to two fundamental components, mobile users/wireless devices who communicate with each other, and network devices which provide the wireless coverage, wireless transmission, positioning, data exchange and other functionalities.

Everyone has her/his own living habits, like when to sleep, where to live, whom to play with, therefore these habits will definitely be reflected in her/his communication information, since mobile phones have become part and parcel of our life and an integral part for every individual. The practical communication activities of mobile users include voice calls, SMS, and other data traffic through all kinds of Internet Applications, coined here as user-oriented data. We note here that, despite of traditional mobile devices, Machine Type Communication (MTC) devices will also contribute even larger amount of data traffic in the coming future. Communication networks also produce giant amount of data when serving the mobile users, like spectrum measurements [9], [10], device status report and etc, but this topic is beyond this paper's scope.

The most important user-oriented data is Call Detail Record (CDR), whose typical sample is shown in Table 1. Fine granularity study can rely on the detailed information CDR data provides. Many researches that concerning dynamic spectrum access and human dynamics are based on CDR data. By analyzing large-scale CDR data, temporal and spatial variation of primary usage can be characterized [11] and calling patterns can be understood [12]. Other related work includes [13]–[15].

TABLE 1. Data formats.

Field Name	Sample Data	Unit
calling number	123****1234	
called number	111****1111	
start time	1-Jun-2013 01:53:00	Second
duration	2	Minute
sector of each call	2048	

The voice traffic can also be aggregated, measured and reported mostly by base stations, which is easy to process and study due to its small amount. It's proportional to amount of people, thus revealing city dynamics is possible. But detailed studies on aggregated voice traffic are not that plentiful and relevant studies include hot spots detection based on Erlang data [16] and traffic analysis [17], [18].

Mobile telephone traffic data [4], [18], [19] can be used to characterize resource usage and subscriber behaviours in mobile networks, and some basic knowledge of spatial and temporal dynamics of data traffic has been captured [4], [20], [21].

The network-oriented data source may be categorized into KPI (Key Performance Indicator) measurement, abnormal event report and traffic behavior information. The KPI measurement include MO success rate, drop rate, call setup delay, data link dropping rate and etc. The abnormal event report will be activated when, for example, dropping call, MO failure, poor coverage or weak coverage, happens. The traffic behavior considers the information like APP name, user ID, time, location, duration. From the real time network-oriented data, the real status of mobile communication networks can be clearly revealed and demonstrated, although currently research work based these data are not so many due to the data collection difficulties and distribution obstacles.

B. TEMPORAL ANALYSIS

In temporal analysis, Allan Variance (AVAR), Modified Allan Variance (MAVAR) and Detrended Fluctuation Analysis (DFA) are important tools to analyze the time series data collected in the telecommunication networks, for example, to investigate the traffic dynamics.

1) AVAR AND MAVAR

Assume the one-sided power spectral density of a random process **x(t)** can be modeled as

$$S_x(f) = \begin{cases} h_\alpha f^\alpha, & 0 < f < f_h \\ 0, & f > f_h \end{cases} \tag{1}$$

Where α and h_α are parameters and f_h is the upper cut-off frequency. Such random processes are commonly referred to as power-law noise. The value of α equals -2(random walk frequency modulation), -1(flicker noise frequency modulation, 0(white noise frequency modulation), 1(flicker noise phase modulation), 2(white noise phase modulation) [22]. Since finite bandwidth and duration are measured in the real-world, the model that $\alpha \geq 1$ is common. It is useful to evaluate the variance of the M-th derivative of the process to address the problem such as infinite variance and even nonstationarity resulted from the value of a larger than 1. In particular, the Allan Variance (AVAR) is evaluated on the 2nd difference of phase samples.

But Allan Variance has disadvantage in evaluating the value of α especially when $\alpha \geq 1$. Thus David W. Allan proposed Modified Allan Variance as an improvement of AVAR in 1981. It converges on all power-law noise types with $\alpha < 5$ and has superior robustness against non-stationarity in data analyzed. Data offset and linear drift are canceled in MAVAR results. In addition, it has more accurate estimation of α and especially for $0 \leq \alpha \leq 1$, while AVAR fails.

MAVAR can be computed using the following formulation [13].

$$Mod\sigma_y^2(n\tau) = \frac{\sum_{j=1}^{N-3n+1}[\sum_{i=j}^{n+j-1}(x_{i+2n} - 2x_{i+n} + x_i)]^2}{2n^4\tau^2(N-3n+1)} \tag{2}$$

where n=1,2,...,$\lfloor N/3 \rfloor$, N represents the total number of samples, x_k denotes the number of new calls in the k-th sample, and τ represents the the sampling period. And observation interval $T = (N-1)\tau$.

If x_k is the sample of **x(t)**, the MAVAR formula turns to obey a power-law of the observation interval τ(t=nτ).

$$Mod\sigma_y^2(t) \sim A_u t^u \tag{3}$$

where $u = -3 - \alpha$ [13]. So if a log-log plot of MAVAR looks ideally as a line with its slope u$\cong$-3 and $\alpha \cong 0$, then **x(t)** is power law noise and x_k is uncorrelated.

2) DETRENDED FLUCTUATION ANALYSIS

Detrended fluctuation analysis (DFA) is a scaling analysis method used to accurately detect the long-range correlations embedded in a nonstationary time series, which was proposed in 1994 [23]. The output of the method, the scaling exponent α, can quantify the correlation properties of a time series, while other traditional approaches fails. The DFA method comprises the following steps:

i) Considering a time series u(i) (i=1,...,N). N is the length of the series. Integrating the time series and obtaining

$$y(k) = \sum_{i=1}^{k}[u(i) - \langle u \rangle] \tag{4}$$

where <u> is the mean of the series.

ii) The integrated series $y(k)$ is divided into N/l non-overlapping boxes of equal length l.

iii) In each box, we use a polynomial function $y_l(k)$ of order n that represents the trend of that box to fit $y(k)$. We denote the algorithm as $DFA - n$,

iv) $y(k)$ is detrended by subtracting the local trend $y_l(k)$ in each box and we obtain

$$Y(k) = y(k) - y_l(k) \tag{5}$$

v) For a given box size l, we calculate the root-mean-square function for integrated and detrended series

$$F(l) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[Y(k)]^2} \tag{6}$$

vi) The above computation is repeated for box size l, and we can obtain $F(l)$ as a function of l. The relationship between $F(l)$ and l can be represented as $F(l) \sim l^\alpha$, α represents the correlation properties of the signal which is in the range from 0 to 1. $\alpha > 0.5$ means the time series is correlated, $\alpha = 0.5$ means uncorrelated(white noise), and $\alpha < 0.5$ means anti-correlated.

C. SPATIAL ANALYSIS

In spatial econometrics or spatial statistics, a popularly used method to study spatial autocorrelation of a geographical feature is Moran's Indicator [24], as defined below:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2} \tag{7}$$

where, x_i denotes the value of variable $\{x_i\}$ in region i, with $\overline{x}$ being the mean. N denotes the total number of observations. w_{ij} measures the degree of influence between x_i and x_j.

The value of Moran's I ranges from -1 to 1. If variable $\{x_i\}$ is independent in space, the expected value of Moran's I is 0. When Moran's I is above 0, positive spatial correlation exists. Otherwise, negative spatial correlations exists.

Global Moran's I reveals the degree of spatial dependance in the whole area concerned, but we also need to investigate the local cases, to which, Local Indicators of Spatial Association (LISA) can be used. Popularly used LISA are local Moran' I [25] and Moran scatter plot [26].

Local Moran's I is the decomposition of global Moran's I into the contribution of each observation. It is defined as

$$I_i = \frac{(x_i - \overline{x}) \sum_j w_{ij}(x_j - \overline{x})}{\frac{1}{N} \sum_j (x_j - \overline{x})^2} \tag{8}$$

Local Moran's I serves to assess the influence of individual locations on the global statistic and to identify local non-stationarity, which will be used in subsection III-C.

D. DATA MINING AND STATISTICAL TEST

1) THE *K*-MEANS CLUSTERING

The k-means is the most well-known and commonly used clustering method in data mining.

Given a data set, $\boldsymbol{D}$, which contains n objects, k-means aims to partition the objects in $\boldsymbol{D}$ into k exclusive groups or clusters. Objects within a cluster are similar to one another (high intracluster similarity) and dissimilar to objects in other clusters (low intercluster similarity) [27].

Partitioning quality is assessed with with-cluster sum of squares (WCSS). In Euclidean space, the target is to minimize:

$$\arg\min_{\mathrm{s}} \sum_{i=1}^{k} \sum_{\boldsymbol{x_j} \in \boldsymbol{S_i}} \| \boldsymbol{x_j} - \boldsymbol{\mu_i} \|^2 \tag{9}$$

where $\boldsymbol{\mu_i}$ is the centroid of cluster $\boldsymbol{S_i}$, defined as the mean value of the objects within the cluster.

The k-means is a heuristic algorithm, which can not guarantee to converge to the global optimum. The choice of initial cluster centroid selection affect the results. Also, the proper choice of cluster number, k, is an important factor of the partitioning performance. The detailed process of k-means algorithm is shown in Algorithm 1:

Algorithm 1 Standard k-Means Algorithm

Require: k: the number of clusters; $\boldsymbol{D}$: a given data set.
Ensure: k clusters.

1: randomly select k objects from D as the initial cluster centers;
2: **repeat**
3: Assignment Step:
assign each object to the cluster to which the object is the most similar(least Euclidean distance between the object and the cluster mean);
4: Update Step:
calculate the mean value of the objects in each cluster to update the cluster means;
5: **until** no change;

2) KOLMOGOROV-SMIRNOV TEST

Kolmogorov-Smirnov (KS) test first applied by Kolmogorov in 1933 is a nonparametric test that quantifies the distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution (one-sample KS test), or between the empirical distribution function of two samples (two-sample KS test). The KS test is based on the following test statistic [28]:

$$K = \sup_x |F(x) - S(x)| \tag{10}$$

Where $F(x)$ is the hypothesized cumulative distribution function, $S(x)$ is the empirical distribution function based on the sampled data.

This maximum distance is then plugged into KS probability function to calculate the probability value, which ranges from 0 to 1. The lower the probability value is, the less likely the two distributions are similar.

3) CHI-SQUARE TEST

Chi-square test is a nonparametric test that is commonly used to 1) test whether there is a significant difference between the expected frequencies and the observed frequencies or not; 2) test the independence of two attributes; 3) test a null hypothesis on a specific value of the population hypothesis for single variance. The chi-square test is based on the following test statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{11}$$

That is, chi-square is the sum of the squared difference between observed (O) and the expected (E) data, divided by the expected data in all possible categories.

III. CASE STUDY I : ERLANG MEASUREMENT

In this section, we present one case study using Erlang measurement data from the base stations, which try to find some spatial and temporal characteristics in voice traffic. We conduct analysis on the characteristics of traffic distribution, distinguish the base stations and infer the differences

among clusters. Through this study, spatial and temporal characteristics of base stations have been revealed.

A. ERLANG MEASUREMENT DATA SET

In this section, we introduce the data set used in this paper, including Erlang measurement, temporal feature, spatial feature and data preprocess.

Our data set is provided by one telecommunication operator, which is collected from the mobile communication networks in one Southern China city with more than two million inhabitants. The Erlang measurement of only voice traffic data is collected from the base stations, not including the SMS or other data traffic. Totally, the traffic data are collected on an hourly basis from over 750 cellular base stations for continuous 18 days, from 2013-05-10 to 2013-05-27, in the temporal dimension. The maximum, minimum and mean Erlang of the data set is 143.1833, 0 and 6.3321 respectively.

In addition, the precise location(longitude and latitude) of each base station are also provided for spatial analysis. But for privacy protection, we linearly translate the position of each base station into a new coordinate system, thus the figures in the following parts do not indicate real positions of these base stations, although the relative positions of these base stations remain unchanged, which will also show us real and intuitive spatial patterns.

The base stations we use are mainly divided into two or three sectors, and the measurement data is collected on each individual sector. So we first sum the measurement of these sectors of the same base station together, which is used as the base station's Erlang measurement for the analysis in this paper.

Due to some unknown reasons, there exist about 4 percentage missing values in the collected data, among which, totally 14 hours of data are eliminated, when more than 50 base stations have void value. In addition, some base stations still have missing values. So after cleaning these data, finally there are 752 base stations and 418 hours remaining, where maximum and mean of the data is 116.4358 and 5.9564 Erlang now.

B. DAILY PATTERN OF GSM BASE STATION'S TRAFFIC

In this section, we will present the daily patterns of traffic, including temporal patterns clustering along with corresponding spatial analysis.

1) CLUSTERING OF DAILY TRAFFIC PATTERNS

The base stations of mobile networks being deployed in the same functional urban areas, like commercial center, research institute, citizen housing and etc, may have similar daily traffic patterns for weekday or weekend. So we first average the 18 days' traffic of each base station into daily 24 hours duration, thus our work is how to cluster these 752 base stations into several distinct and significant types.

K-means approach is utilized to automatically distinguish different daily patterns. But directly applying K-means algorithm to traffic data will cause base stations of similar traffic level to be clustered together while their contours may be not very alike. Hence, we normalize the traffic of each base station according to its maximum value. In K-means, how to choose the proper value of parameter k still keeps an open question. In this paper, gap statistic [29] is adopted to decide the value of k, which is defined as:

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}) - log(W_k) \tag{12}$$

where W_{kb} denotes the within-dispersion measure of a reference null distribution data set, W_k denotes intra-cluster dissimilarity. So the number of clusters is the smallest k that satisfies:

$$Gap(k) \geq Gap(k+1) - s_{k+1} \tag{13}$$

where s_{k+1} denotes the standard deviation of within-cluster dispersions in reference data sets.

Through this test, the best value of k is determined to be 4. So, by K-means approach we distinguish the normalized daily traffic patterns as 4 clusters, which is shown in figure 1.

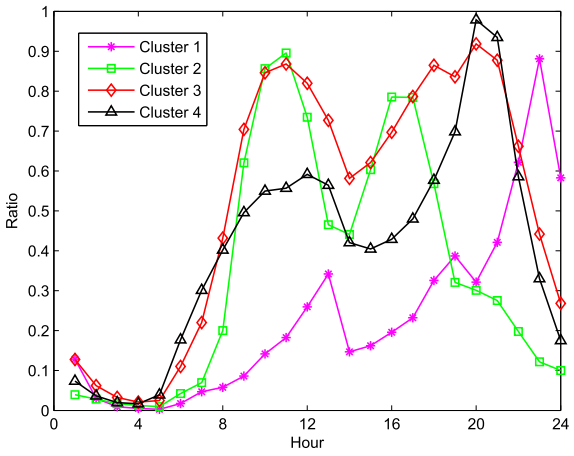

FIGURE 1. Centroid pattern of each cluster. The numbers of base stations belonging to cluster 1 to 4 are 35, 59, 369 and 289 respectively.

Cluster 2 shows typical pattern of non-housing estate, which might be commercial office, industrial factory and etc, because the traffic of base stations of this cluster exhibit two traffic peaks in the working hours, but decrease sharply after about 7pm. One more interesting finding is that, there are only 39 base stations in this cluster, which implies that in this southern China city, most areas are full of living inhabitants.

Cluster 3 and 4 stand for 369 and 289 base stations, respectively, which demonstrate very typical patterns for residential area, because both have a traffic peak at about 8pm and then decrease sharply. From the amount of such kind base stations, we can infer that most areas of this city contains residential houses, or mixture with residential houses.

But cluster 1 is quite different, due to its special and unique characteristics, although it just has 35 base stations.

Two traffic peaks appear at about 1pm and 11pm, and the traffic density of 11pm is quite higher than that of 1pm, so that such phenomenon is coined as 'Night Burst'. So in the next section we will investigate the reason of this phenomenon and discuss the spatial correlations of these base stations.

2) SPATIAL ANALYSIS OF TRAFFIC PATTERNS

The spatial locations of base stations in different clusters are plotted in figure 2. Base stations in cluster 1 are confirmed to mainly locate in three different parts, all in the neighbors of university campuses, by comparing the GPS location of base stations with real-world map. So the behavior of cluster 1 we believe may represent traffic distribution of students in university campuses. We think the reason of this 'Night Burst' pattern is that, after finishing their study in the evening or at noon, students begin to contact their parents, friends or mates, but in the morning or the afternoon they rarely use their mobile phones because of having courses.

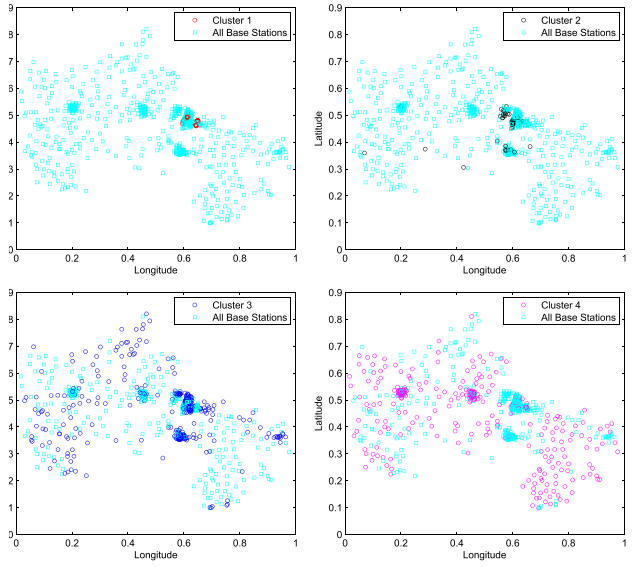

FIGURE 2. Real-World Position Recognition for All Base Stations in Four Clusters.

There are mainly 4 parts that base stations in cluster 2 locate at, but among them no direct relationship has been found. We can only infer that these parts to be official areas because the traffic is high only in working hours, and further verification will be our future work.

Finally, base stations of cluster 3 and 4 spread over the whole city, which deal with the communication requirement of major citizens.

C. SPATIAL AUTOCORRELATION ANALYSIS

In this part we will present the spatial autocorrelation analysis to investigate the spatial connection among the base station traffic distribution.

The question we are trying to answer is, how the spatial correlation forms? Thus we use local Moran's I, which is introduced in section III-C, to further study the local spatial correlation. We choose the data at one certain hour to plot the local Moran's I, which can help us get further insight visually.

In figure 3(a), we can see that there are four dense regions of base stations. In these regions, base stations of positive correlation are in the majority. The fact of interest is that in the positive correlation concentrated region, there are still some negative correlation base stations. So we choose the largest dense part for further study. Also, we compare the local spatial correlation shown in the figure 3(b), with the traffic density of this region at the same time(figure 3(c)).

From direct comparison, we can see that these positive correlation base stations are of the same low level traffic as their neighbors, while negative correlation base stations are mainly the ones of high level. We see that, around 6 am, the volume of voice traffic of most base stations is basically small, while there are a few base stations that have larger traffic. Thus, positive correlation exists among many proximate base stations, which will lead to larger global spatial correlation.

The fact that different correlations exist among base stations of different traffic volume gives us an idea that those 'abnormal' base stations that are different from their neighbors can be distinguished with spatial correlation detection. In cellular optimization, a direct purpose is to balance the loads of base stations, especially base stations nearby. The situation that one base station has high load while its neighbors are of small traffic brings inefficiency and resource wasting. The above study gives us an implication that local Moran's I may be a powerful tool to depict the imbalance and further used for cellular optimization and resource planning.

IV. CASE STUDY II : CALL DETAIL RECORD

In this section, we re-consider the time-correlation issue of call arrivals, using CDR data of large-scale base stations from a dense population district in one large northern China city, based on our previous work [30]. We first verify the long-range in (24 hours) time-correlation of call arrivals but find call arrivals in a minute is uncorrelated to the number of call arrivals in another minute in short-term. Second, we find the time-correlation of call arrivals is influenced by time and the location of stations.

There are several empirical studies on real GSM telephone traffic data attempting to modelling call arrivals in a cell. The answered call holding time and inter-arrival time were found to be modeled as lognormal-3 function [31]. Both [13] and [14] used MAVAR to analyse the time correlation of call arrivals of a base station over a day with τ=1s, and concluded that the number of call arrivals in a second is uncorrelated in short-range(about 500s). Then in long time period (24 hours) the number of call arrivals are confirmed to be uncorrelated [14] while [13] believed that non-negligible time-correlation may be found on long intervals. In [15], its dataset was collected from hundreds of cell stations contained voice call information over three weeks, and verified

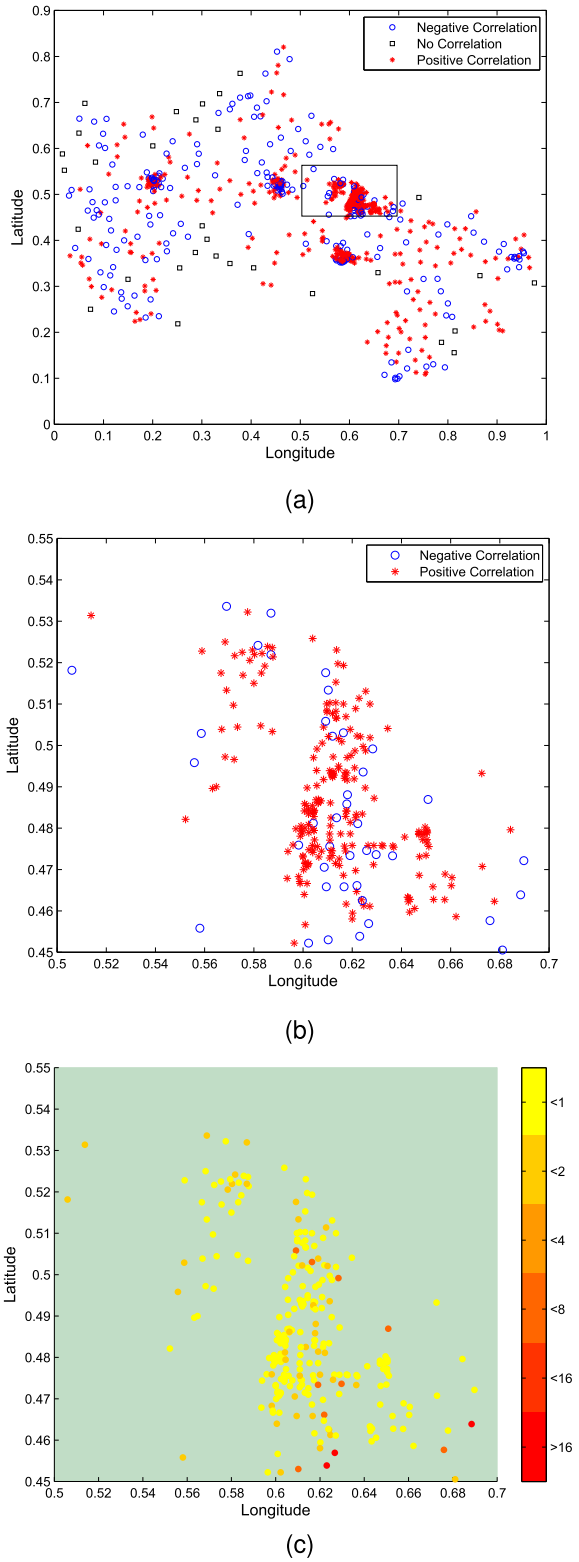

FIGURE 3. Local Moran's I at one certain hour and Zoom-in of one dense area. (a) Local spatial correlation of base stations. (b) Local Zoom-in. (c) Traffic Density of Local Zoom-in.

call arrivals can be modeled as Poisson process by using Maximum Likelihood Estimate (MLE) exponential fits for actual distribution of inter-call arrival time.

A. INFORMATION OF CDR

The 30 days(in June 2013) voice CDR data set of 10,000 specific users we use in this paper was collected from one GSM mobile communication operator, which was randomly sampled from those who satisfy following two constraints: (1) Aggregated call duration $>$ 100 minutes; (2) Their phone number registered in one specific high-tech industrial district. There are totally 2838 base stations, but we just choose 4 stations, represented as station A, B, C and D, with highest cell loads and users for further investigation, whose detailed basic information is shown in table 2. We note that the CDR data, like calling number, is preprocessed to keep anonymous due to privacy protection.

TABLE 2. Basic information of the chosen four cell stations.

Station	Location	Num. of Calls	Num. of Users
A	Commercial Building	39,889	655
B	Research Institute	26,755	1680
C	Commercial Building	23,902	396
D	Commercial Building	26,476	601

B. TIME-CORRELATION OF CALL ARRIVALS

In this section, the time-correlation analysis of call arrivals is presented, including both long-range and short-range cases.

1) TIME-CORRELATION OF CALL ARRIVALS

The time-correlation plot of call arrivals based on MAVAR is shown in figure 4, where x_k represents the number of calls in the k-th minute of the 30 days. Because the results are similar in all 30 days (maximum deviation of a is 0.18) for all four stations, so we choose just one day (16-JUN-2013)'s result for discussion.

MAVAR is almost perfectly linear(in the log-log plot) for $n\tau$ up to 50 minutes, with slop u $\cong$ -3.0

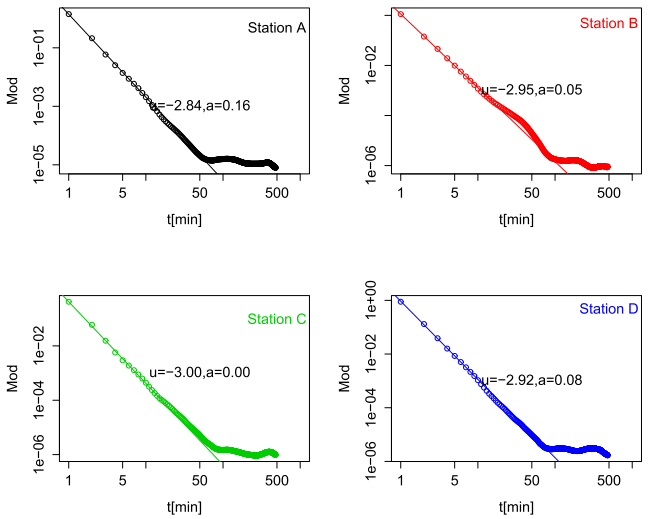

FIGURE 4. MAVAR of call arrivals of base station A, B, C and D (T=24h, N=1440, τ=1min, Date:16-JUN-2013).

corresponding $a \cong 0.0$, which indicates power spectral density of call arrivals is equal to white noise's power spectral density and means that in short term the number of call arrivals in a minute is independent. When $n\tau > 50$, MAVAR deviates from the straight line, which implies long-range time-correlation of the call arrivals.

The above discussion only considers time-correlation of call arrivals in granularity of one minute, and we then model call arrivals as Poisson process in short-term, which means the call arrivals is uncorrelated in any granularity.

2) SHORT-RANGE DEPENDENCY

One hour interval is chosen for discussion because the number of call arrivals in a minute is uncorrelated in short-term(about 50 minutes) through the observation of figure 4.

MAVAR of the sequence of call arrivals over one second slots of one hour for four base stations during 10:00-10:59 AM on 14-JUN-2013 is plotted in figure 5. The log-log plot of MAVAR of station A, B and D are found to be ideally straight lines with slopes u $\cong$ 3.0 for all time intervals, which confirms the non-correlation, but station C behaves differently, showing that when time interval > 800s, the time-correlation seems not true. Similar findings have been obtained in other one hour time when the time interval > 800s, thus the time-correlation seems not true for all base stations. We then take chi-square test and fitting with Poisson distribution for further verification.

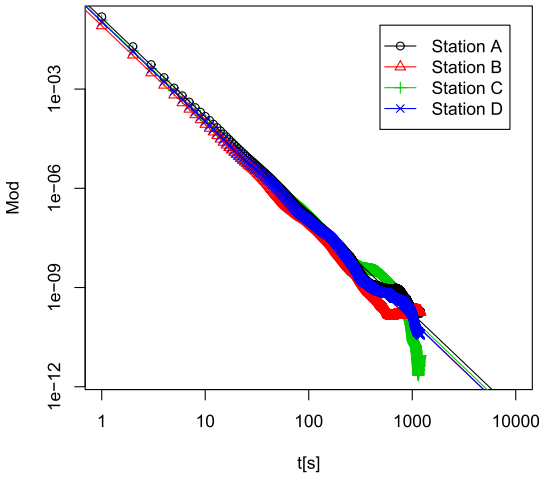

FIGURE 5. MAVAR of call arrivals on 10:00-10:59 AM of four base stations(T=1h, N=3600, τ=1s, Date:14-JUN-2013). The curves of black,red,green and blue represent station A, B, C, D respectively with u1=-3.00, u2=−2.98, u3=−3,02 and u4=−3.01.

Table 3 presents the mean, variance, variance-mean ratio and the value of chi-square test of call arrivals over one minute slots in the same hour for these four base stations, where the chi-square test is used to ascertain the fitting degree between the empirical distribution of the call arrivals and the Poisson distribution with same mean, where Chi-square test < 0.05 indicates that the distribution of call arrivals should not be modeled as Poisson distribution.

TABLE 3. Mean, Variance and χ^2-Test(to same-mean poisson distribution) of the number of new calls over 1 minute slots of 1 hour from 10:00 to 10:59 for 4 stations on 14-JUN-2013.

Station	m_x	x^2	x^2/m_x	χ^2-Test
A	3.067	2.640	0.861	0.957
B	1.700	1.569	0.923	0.706
C	2.350	3.316	1.411	0.087
D	2.200	2.129	0.968	0.996

Three stations A, B, D all pass the chi-square test, but C does not, because its chi-square test is too small and its variance-mean ratio is too large.

So we can conclude here that the call arrivals of station A, B, D can be modeled as Poisson process on 10:00-10:59 14-JUN-2013, while station C can not. Similar results hold true in other time intervals for these four stations(just the base stations which can not be modeled as possion process may be other base stations not base station C), which means although in most one hour interval the call arrivals can be modeled as Poisson process, it is NOT true for all cases.

C. CHI-SQUARE TEST FITTING

In the last section, we confirm that, in specific hour, call arrivals of station A, B and D satisfy Poisson process, but station C does not. So we take further and comprehensive test for all hours in this section.

1) DISTRIBUTION OF CHI-SQUARE TEST

Figure 6 illustrates the distribution of the chi-square test results of four stations for 27 days(10,11,12 JUN 2013 are ignored, because they are holidays so that call loads of four stations in these days are relatively small, which may influence the accuracy of chi-square test.) from 9:00-18:59, which brings us a thorough understanding and two important findings.

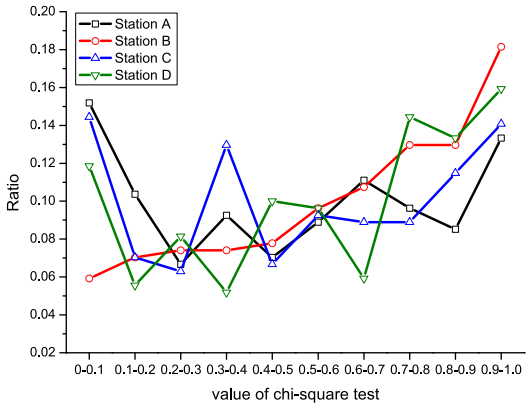

FIGURE 6. the probability of the value of chi-square test of four stations for 27 days from 9:00-18:59.

Firstly, all these four stations have a relatively high rate (A:41, B:16, C:39, D:32) that decline the chi-square test with 90% confidence level, which means that the distribution of

call arrivals in those specific hour interval can NOT be modeled as Poisson distribution. Our finding is not consistent with previous conclusion of call arrivals of base stations being able to be modeled as Poisson process, so that brings forward some insight on the traffic modeling issues in telecommunication networks.

Secondly, the decline rate of four stations vary, where station B is much smaller than that of other three stations. Because station B is located in a research institute while the other three stations are in commercial buildings, as previously depicted in table 2, so the variety of call arrivals of station B is more flatter. But, how the time-correlation of call arrivals is influenced by the location or properties of base stations needs further investigation.

2) TIME PERIODS FAILING THE CHI-SQUARE TEST

The characteristic of time periods that fail the chi-square test is also important. Figure 7 depicts the number of days where chi-square test value is smaller than 0.1 in an hour interval from 9:00-18:59 for four stations. We notice, on one hand, almost in any time there are cases that the call arrivals fail the chi-square test for all stations. But on the other hand, the number of days where chi-square test value is smaller than 0.1 is different in different time periods. First, the largest number of days where the chi-square test value is smaller than 0.1 for station A, C, D are at 18PM, 16PM, 17PM, corresponding to the time where the call arrivals begin to decline steeply respectively. Second, the pass rate of chi-square test in the afternoon from 12:00-15:59 is better for these four stations. This implies that the characteristic of call arrivals is changing over time. So it is improper to model the call arrivals in 1 hour as Poisson distribution in any time. In the future work, we should try to answer how to predict call arrival precisely.

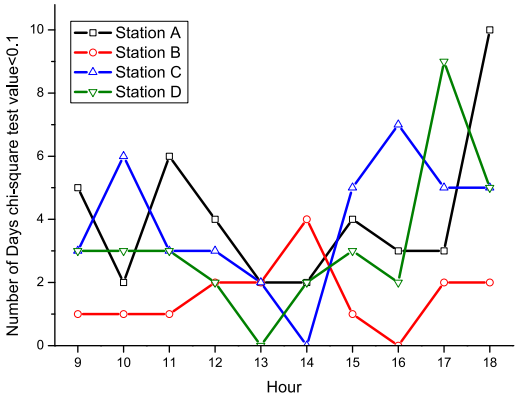

FIGURE 7. the number of days where chi-square test value is smaller than 0.1 in an hour interval for 27 days from 9:00-18:59 for four stations.

V. CONCLUSION

This paper concentrates on the data analysis issue in the telecommunication networks, covering the data set categorization, commonly used data analysis techniques and two general case studies using Erlang measurement and CDR. In addition, our work presents several interesting findings about base station behavior.

To be specific, we accomplish the spatial-temporal analysis to the GSM base station traffic of a city in Southern China, and comprehensively investigate the traffic pattern and spatial correlation, which brings some insight for future possible work. K-means method is adopted to help understand different patterns of base stations and finds that, different from common people, college students' special activity pattern, coined here as 'Night Burst', has been revealed through traffic of base stations near university campuses, based on spatial-temporal analysis. In spatial correlation part, we give our thinking and possible reason through the study of local Moran's I in the spatial dimension and find that local Moran's I can be a tool to discover 'abnormal' stations in a region.

As to the CDR data, we have analyzed the characteristics of the call arrivals based on real call detail records of large-scale GSM base stations in Beijing over 30 days using MAVAR and chi-square test. First, The preliminary observation reported in this paper shows that the call arrival patterns vary over time and the location of stations. Second, the number of call arrivals in a minute has been found uncorrelated in short-range but time-correlation exist in long-range because of the violent fluctuation of the call arrivals in the long-term(24 hours). Third, the call arrivals can be modeled as Poisson process in most cases, but the characteristic of call arrivals changes over time and space, so it is improper to model the call arrivals in one hour as Poisson distribution.

Our work is a first step for such data analysis in mobile communication networks, and our observations have potential applications, such as cellular optimization, resource planning and etc.

ACKNOWLEDGMENTS

Author Sihai Zhang would like to thank Mr. Yi Zheng's help for the CDR data and beneficial discussions.

REFERENCES

[1] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 141–151, Mar. 2011.

[2] A. Wesolowski *et al.*, "Quantifying the impact of human mobility on malaria," *Science*, vol. 338, no. 6104, pp. 267–270, Oct. 2012.

[3] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, Mar. 2013, Art. ID 1376.

[4] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 882–890.

[5] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Pers. Commun.*, vol. 3, no. 6, pp. 4–9, Dec. 1996.

[6] Z. Opechowski and L. M. Correia, "Analysis of traffic distributions in GSM," in *Proc. 13th Int. Conf. Microw., Radar Wireless Commun. (MIKON)*, vol. 2. May 2000, pp. 390–394.

[7] A. Al Daoud, M. Alanyali, and D. Starobinski, "Secondary pricing of spectrum in cellular CDMA networks," in *Proc. 2nd IEEE Int. Symp. New Frontiers Dyn. Spectr. Access Netw. (DySPAN)*, Apr. 2007, pp. 535–542.

[8] H. Mutlu, M. Alanyali, and D. Starobinski, "Spot pricing of secondary spectrum usage in wireless cellular networks," in *Proc. IEEE 27th Conf. Comput. Commun. INFOCOM*, Apr. 2008, pp. 1355–1363.

[9] A. Palaios, J. Riihijarvi, O. Holland, A. Achtzehn, and P. Mahonen, "Measurements of spectrum use in London: Exploratory data analysis and study of temporal, spatial and frequency-domain dynamics," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DYSPAN)*, Oct. 2012, pp. 154–165.
[10] A. Palaios, J. Riihijarvi, O. Holland, and P. Mahonen, "A week in London: Spectrum usage in metropolitan London," in *Proc. IEEE 24th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 2522–2527.
[11] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study," in *Proc. 3rd IEEE Symp. New Frontiers Dyn. Spectr. Access Netw. (DySPAN)*, Oct. 2008, pp. 1–11.
[12] Z.-Q. Jiang, W.-J. Xie, M.-X. Li, B. Podobnik, W.-X. Zhou, and H. E. Stanley, "Calling patterns in human communication dynamics," *Proc. Nat. Acad. Sci.*, vol. 110, no. 5, pp. 1600–1605, 2013.
[13] S. Bregni, R. Cioffi, and M. Decina, "An empirical study on time-correlation of GSM telephone traffic," *IEEE Trans. Wireless Commun.*, vol. 7, no. 9, pp. 3428–3435, Sep. 2008.
[14] B. Yuksel, M. Cingoz, G. Karabulut, and S. Oktug, "Call arrival model for GSM network including handover," in *Proc. IEEE 3rd Int. Symp. Adv. Netw. Telecommun. Syst. (ANTS)*, Dec. 2009, pp. 1–3.
[15] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Commun. Mag.*, vol. 47, no. 3, pp. 88–95, Mar. 2009.
[16] X. Zuo and Y. Zhang, "Detection and analysis of urban area hotspots based on cell phone traffic," *J. Comput.*, vol. 7, no. 7, pp. 1753–1760, Jul. 2012.
[17] M. Panda and S. P. Padhy, "Traffic analysis and optimization of GSM network," *Int. J. Comput. Sci. Issues*, 2011.
[18] M. Michalopoulou, J. Riihijarvi, and P. Mahonen, "Towards characterizing primary usage in cellular networks: A traffic-based study," in *Proc. IEEE Symp. New Frontiers Dyn. Spectr. Access Netw. (DySPAN)*, May 2011, pp. 652–655.
[19] E. Nan, X. Chu, W. Guo, and J. Zhang, "User data traffic analysis for 3G cellular networks," in *Proc. 8th Int. ICST Conf. Commun. Netw. China (CHINACOM)*, Aug. 2013, pp. 468–472.
[20] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding spatial relationships in resource usage in cellular data networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Mar. 2012, pp. 244–249.
[21] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1341–1349.
[22] D. W. Allan and J. A. Barnes, "A modified 'Allan variance' with increased oscillator characterization ability," in *Proc. 35th Annu. Freq. Control Symp.*, May 1981, pp. 470–475.
[23] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Phys. Rev. E*, vol. 49, no. 2, pp. 1685-1689, 1994.
[24] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, nos. 1–2, pp. 17–23, Jun. 1950.
[25] L. Anselin, "Local indicators of spatial association–LISA," *Geograph. Anal.*, vol. 27, no. 2, pp. 93–115, Apr. 1995.
[26] L. Anselin, "The Moran scatterplot as an ESDA tool to assess local instability in spatial association," *Spatial Anal. Perspect. GIS*, vol. 111, pp. 111–125, 1996.
[27] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 5. San Francisco, CA, USA: Morgan Kaufmann, 2001.
[28] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," *Proc. Eur. Phys. J. B, Condens. Matter Complex Syst.*, vol. 41, no. 2, pp. 255–258, 2004.
[29] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 63, no. 2, pp. 411–423, 2001.
[30] D. Yin, S. Zhang, W. Zhou, and Y. Zheng, "Time-correlation analysis of GSM telephone traffic in dense population district," in *Proc. 6th Int. Conf. Wireless Commun. Signal Process.*, Hefei, China, Oct. 2014, pp. 1–5.
[31] A. Pattavina and A. Parini, "Modelling voice call interarrival and holding time distributions in mobile networks," in *Proc. 19th Int. Teletraffic Congr. Perform. Challenges Efficient Next Generat. Netw.*, 2005, pp. 729–738.

Sihai Zhang (M'08) received the B.Sc. degree from the Department of Computer Science, Ocean University of China, Qingdao, China, in 1996, and the M.S. and Ph.D. degrees from the Department of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China, in 2002 and 2006, respectively. He is currently an Assistant Professor of Electronic Engineering with the Department of Electronic Engineering and Information Science, USTC, where he has been with the PCNSS Laboratory since 2009. He has participated in projects, including Innovative Wireless Campus Experimental Networks, Research on High Frequency Networking Technologies, and the Research on Transmission and Networking Technologies in Satellite Mobile Communications. His research interests include wireless networks, opportunistic networks, and intelligent algorithms.

Dandan Yin received the B.Sc. degree from the Hefei University of Technology, Hefei, China, in 2013. She is currently pursuing the master's degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei. Her research interests include wireless networks and wireless big data analysis.

Yanqin Zhang received the B.Sc. degree from the University of Science and Technology of China, Hefei, China, in 2013, where he is currently pursuing the master's degree with the Department of Electronic Engineering and Information Science. His research interests include wireless networks.

Wuyang Zhou received the B.Sc. and M.S. degrees from Xidian University, Xi'an, China, in 1993 and 1996, respectively, and the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2000. He is currently a Professor of Wireless Communication Network with the Department of Electronic Engineering and Information Science, USTC. He participated in the National 863 Research Project Beyond-3 Generation of Mobile System in China (FUTURE Plan), and has played the role of task director in projects, including Innovative Wireless Campus Experimental Networks, Research on High Frequency Networking Technologies, and the Research on Transmission and Networking Technologies in Satellite Mobile Communications. His current research interests include green technologies for communication systems, satellite mobile communications, and underwater acoustic communications.