Measurement-Driven Modeling for Connection Density and Traffic Distribution in Large-Scale Urban Mobile Networks

Jingtao Ding, Rui Xu, Yong Li, *Senior Member, IEEE*,
Pan Hui, *Senior Member, IEEE*, and Depeng Jin, *Member, IEEE*

Abstract—In the diverse usage scenarios of mobile networks, we have different performance requirements on *connection density* and *user experienced data rate*, and modeling such diversity is crucial to the strategy evaluation in addressing the problem of high traffic load and scalability of network resources. Therefore, it is necessary to build a network capability model in two dimensions of *connection density* and *user experienced data rate*. This paper aims at addressing this challenge based on an investigation of network capability in large-scale urban environments. First, our statistical study shows that the spatial distribution of these two parameters can be accurately modelled by the log-normal mixture distribution. Second, we find that only six basic capability patterns exist among the 9,000 cellular base stations, which indicates different levels of network capabilities. More importantly, these discoveries are similar in a cellular network deployed in a different city. Therefore, based on these two discoveries, we build a network capability model that can generate synthetic base stations with diverse *connection density* and *user experienced data rate*. We believe that this methodology of modeling network capability, with accuracy, generality, and flexibility, can help telecommunication operators to design and standardize mobile networks of the next generation.

Index Terms—Capability modeling, clustering, connection density, data rate, mobile network measurement

1 Introduction

With the tremendous growth in connectivity, density and volume of mobile traffic, both industry and academia are focusing on improving the performance and efficiency of mobile cellular networks. To meet the demands of a fully mobile and connected society, a broad range of usage scenarios for future mobile networks are expected and each of them has different network performance requirements. For example, according to the published white paper [1], in the scenario of *broadband access in dense areas* (e.g., pervasive video), person-to-person or person-to-group video communication with extremely high resolution should be available to every subscriber, where providing such large numbers of concurrently active connections and high data rate will be a challenge. When it comes to the scenario of *massive Internet of Things*, a single macrocell may need to support 10,000 or more low-rate devices with expected demands in machine-to-machine communication [2]. Therefore, *connection density* is a key performance parameter in the scenario of *massive Internet of Things*, while high *user experienced data rate* is vital in the scenario of *broadband access in dense areas*. Under these contexts, it is vital to achieve diverse network performances in terms of *connection density* and *user experienced data rate*.

The diverse mobile network usage scenarios discussed above require a model on the two-dimensional space of *connection density* and *user experienced data rate*. This model could be built by the granularity of base station, as different usage scenarios are achieved through the base stations with different *connection density* and *user experienced data rate* in the future mobile network. In this paper, we refer to the above two parameters as the network capabilities of a base station. Using this capability model, mobile operators can simulate the network capabilities in each cell, including number of connected devices and average access rate, which is extremely valuable in cellular network planning and performance evaluation.

However, there exists no previous work on modeling network capability from aspects of *connection density* and *user experienced data rate*. Mostly, cellular traffic dynamics are directly characterized by stochastic process theories, such as batch Markovian arrival process (BMAP) in [3], and multi-order Markov chain in [4]. If we want to build a capability model considering both *connection density* and *user experienced data rate*, there still remain three challenging problems to solve:

- How to obtain and analyze *connection density* and *user experienced data rate* of a real cellular network? A large-scale trace data containing these two parameters is vital in the analysis. Also, to build a capability model, we need to consider the spatial distribution of *connection density* and *user experienced data rate*. These tasks are challenging.
- How to extract the key patterns of *connection density* and *user experienced data rate* from the trace data? Note that we can combine the base stations with

- *J. Ding, R. Xu, Y. Li and D. Jin are with Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: dingjt15@mails.tsinghua.edu.cn, nbeiray@163.com, {liyong07, jindp}@tsinghua.edu.cn.*
- *P. Hui is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China, the Telekom Innovation Laboratories, Berlin, Germany, and Aalto University, Helsinki 02150, Finland. E-mail: panhui@cse.ust.hk.*

Manuscript received 17 Dec. 2016; revised 1 June 2017; accepted 2 Aug. 2017. Date of publication 14 Sept. 2017; date of current version 2 Apr. 2018. (Corresponding author: Yong Li.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2017.2752159

similar capability characteristics together and obtain a handful number of groups. Then we can build independent and also more accurate models for each of these groups. Therefore, a suitable clustering method is required, which helps us to understand the network capability in these two dimensions.

- How to build a model with generality? Since our analysis is measurement-driven, bias existed between different datasets may impact our built model. For example, if the dataset of cellular network is collected in one certain city, then the methodology of building model may only be workable in that dataset. Therefore, we can consider checking the generality of our modeling methodology in different mobile networks.

To address the first challenge, we carry out a base-station-level analysis of subscriber density and per-subscriber data-traffic demand in our fine-grained and large-scale trace data, which are collected from two mobile networks deployed in *Shanghai* and *Kunming* respectively. Subscriber density and per-subscriber demand correspond to the two key parameters of network capability, *connection density* and *user experienced data rate*. In our following study, we use a log-normal mixture model to characterize the spatial distribution of these two metrics. As for the second challenge, we adopt a 2-dimensional clustering method, which is based on the work by Mucelli et al. [5]. Then, for the third challenge, we show that our log-normal mixture model and clustering method are general to both *Shanghai* and *Kunming* data, which implies the generality of our modeling methodology. Moreover, the traffic patterns of cellular base stations do correspond to the urban functions of geographical locations [6]. Inspired by this, we introduce this urban function context information into our capability model. Our key contributions are threefold:

- First, we discover that the spatial distribution of subscriber density and per-subscriber demand can be accurately fitted by a log-normal mixture model. Our theoretical proof shows that the product of subscriber density and per-subscriber demand, i.e., traffic density, also follows a log-normal mixture distribution spatially, which is further validated by empirical data. In addition, the generality of this distribution model is also verified across different cities.
- Next, our extensive analysis provides a precise characterization of individual base station capability and clusters base stations into several types (6 types in *Shanghai* and 4 types in *Kunming*) according to subscriber density and per-subscriber demand. We also explore the relationship between the base station capability and urban functional regions where base stations are deployed.
- Finally, we build a base-station-level capability model as the function of subscriber density and per-subscriber demand. The highlight of our model is that we only need to input the urban function context information, and it can then generate synthetic base stations with realistic diverse capabilities in terms of the two key parameters. With an average error of 7 percent in aggregate level and 19 percent in individual level, our evaluation demonstrates that this model can reliably and accurately quantify network capability, which reduces the evaluation error by 57 percent. More importantly, our model provides an insight on how to improve the mobile network performance and efficiency in diverse usage scenarios.

This paper is structured as follows. In Section 2 we discuss the related work. In Section 3, we detail the utilized mobile network dataset and explain how we extract the useful information, i.e., subscriber density and per-subscriber demand in each cell. In Section 4, we analyze the spatial distribution of these two key parameters. In Section 5, using an unsupervised clustering algorithm, we identify the key patterns of network capability. Based on these discoveries, we build a capability model in Section 6. After discussing the strength of our proposed model in Section 7, we summarize our work and discuss future investigations in Section 8.

2 Related Work

Works related to our work can be divided into three topics: characterizations of the traffic distribution in a cellular network, investigations of the cellular traffic patterns, and models of the cellular traffic dynamics.

The spatial distribution of the cellular traffic has been studied in the literature [7], [8], [9]. Gotzner et al. [7] found that voice traffic in different cells of GSM networks can be described by a log-normal distribution. As for modeling traffic load in each cell, Lee et al. [8] demonstrated that the spatial distribution of the traffic density can be accurately modeled by a log-normal mixture distribution, while Wang et al. [9] found that the per-cell mobile traffic volume (not density) follows a trimodal distribution on both spatial and temporal dimensions. Unlike these works, we also discuss the reason behind the observed mixture of log-normal distribution. As for the temporal distribution of mobile traffic, Nan et al. [10] analyzed and statistically modeled the downlink throughput per cell distributions over time and over different cells based on a real network throughput dataset. Williamson et al. [11] identified power-law properties in the distribution of packet call activities, using the data collected from a CDMA2000 1x cellular data network. However, these analyses confine to statistical fittings, and they are not suitable for considering temporal correlations in traffic dynamics. Since the actual coverage area of a base station is difficult to measure, Voronoi cell is widely used in computing the traffic load density, where the edges of the Voronoi cells are considered as the boundary when computing the coverage area. Hoteit et al. [12] used the size of Voronoi cells when analyzing per-cell content consumptions. Paul et al. [13] investigated the correlation between size of Voronoi cell and corresponding traffic load. Lee et al. [8] calculated the traffic density in each Voronoi cell to represent the intensity of user traffic demand. All Inspired by these, we apply the method of Voronoi cells when computing the subscriber density.

Cellular traffic patterns have been extensively investigated for understanding various perspectives of cellular networks [6], [14]. Naboulsi et al. [14] defined categories of mobile call profiles based on Call Detail Records (CDRs) and then classified network usages accordingly. Unlike Naboulsi's work, our previous work [6] extracted the data traffic patterns of large-scale base station (BS) towers by combining three dimensional information (time, locations of towers and traffic frequency spectrum) together, where they observed a strong relationship between traffic patterns and spatial context information. In this area of linking land use with cellular traffic, many approaches were proposed [15], [16], [17], all using call detail records. Soto et al. [15] defined signatures as the activity aggregation in different

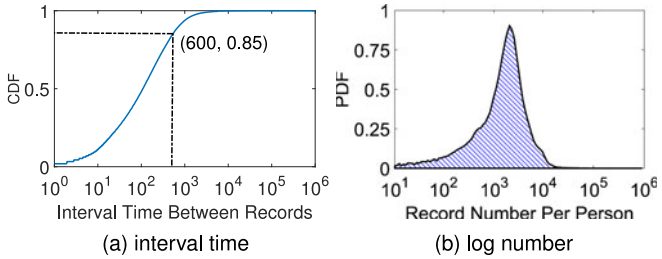


Fig. 1. Illustration of the quality of our dataset.

time-scales (weekly and daily scale) and clustered BSs by using these signatures, while Cici et al. [17] applied the spectral decomposition of original cell phone activity series before clustering. Unlike these works based on clustering, Toole et al. [16] classified groups of locations with similar zoned uses and mobile phone activity patterns. Moreover, data collected from cellular networks deployed in different cities or even countries are used to validate the generality of relationship between land use and cell phone activity patterns [18], [19]. Inspired by above works, when building our capability model, we consider the relationship between the network capability and urban functional regions where base stations are deployed.

Due to wide usage in the cellular network simulation, many models have been built to characterize cellular traffic dynamics [3], [4], [5], [20]. Klemm et al. [3] introduced an aggregated traffic model, batch Markovian arrival process (BMAP), for UMTS networks, which is analytically tractable. Shafiq et al. [4] proposed a Zipf-like model to capture the volume distribution of application traffic in celluar devices and then used a Markov model to characterize the dynamics. Moreover, they further analysed machine-to-machine traffic dynamics in a cellular network [20]. Mucelli et al. [5] classified subscribers into 4 profiles according to session number and traffic volume in a certain period. Then a traffic usage model was built for each profile of subscribers in peak and non-peak time periods, respectively. By contrast, our work considers higher-level capability modeling of mobile networks. For our problem of modeling network capability, we adopt Mucelli's methods of clustering.

Based on the original version of this work [21], following fields are substantially enhanced. With a newly introduced dataset covering a different city, we are able to validate the generality of our modeling methodology. In order to indicate the strength of our model, i.e., generating synthetic BSs, a per-BS-level validation is performed. In addition, we include more detailed analysis of network capability in this paper.

3 Dataset and Key Parameters

In this section, we provide details about our datasets. In addition, we also introduce the needed preprocessing to compute key parameters, i.e., subscriber density and per-subscriber data demand.

3.1 Dataset

In order to carry out a measurement driven study, we use two anonymous cellular traces collected in *Shanghai* and *Kunming* (capital of a Chinese southwest province, *Yunnan*) respectively, by one of the major operators in China.

The dataset *Shanghai* we investigated is collected from the Charging Gateway Function (CGF) of a commercial 3G cellular network deployed in *Shanghai*. It contains over 6.92×10^8 logs recording the detailed mobile data usage of 700,000 subscribers and 9,181 cellular base stations, within an interval of 31 days in August 2014. Each record of the trace collects devices ID (anonymized), start-end time of data consumption, base station (BS) ID, BS location and traffic volume (byte). The total consumed data in a month is 2.8×10^{15} Bytes and per-day consumed data of a BS is 9.8×10^9 Bytes on average. It is notable that there exists a lot of variation between cells, which we will detail in the subsequent analysis. As for the dataset *Kunming*, it is also a cellular trace with the similar format to the dataset *Shanghai*. With a time range of one month, this dataset contains about 6,000 BSs and 400,000 subscribers.

TABLE 1
BS Deployment Statistics

Dataset	BS Distance (km)		BS Density (km^{-2})	
	Mean	Deviation	Whole area	Downtown
Shanghai	0.88	1.20	0.65	17.65
Kunming	0.98	0.70	0.12	7.90

We present several visualizations about basic characteristics of the dataset *Shanghai* in Fig. 1. Subplots (a) and (b) show the empirical Cumulative Distribution Function (CDF) of interval time between two consecutive records and the empirical Probability Density Function (PDF) of the number of records per subscriber, respectively. From the results, we can observe that 85 percent consecutive records happen in less than 600 seconds and most of mobile users have more than 1,000 records in a month. In addition, we also provide statistics about BS deployment, including the distance between two neighboring BSs and the density of BSs, in Table 1. These fine-grained and large-scale datasets, including information on both subscriber number and data consuming volume, enable us to carry out a comprehensive study on the network capability.

Moreover, *Shanghai* and *Kunming* are significantly different from each other. The former is the largest Chinese metropolis, while the latter is a medium-sized city in southwestern China. Therefore, we are able to apply our modeling methodology in these two datasets collected from two different cities, and thus validate its generality.

3.2 Key Parameters

As mentioned previously, subscriber density and average data demand are the two key parameters to describe the network capability. Subscriber density, corresponding to *connection density*, can be computed by counting the number of access subscribers during a certain period of time. As for data traffic demand per subscriber, we define it as the consumed data volume divided by the number of subscribers of the cell during a certain period of time. Since *user experienced data rate* is the actual data rate required for the user to get a quality experience of the targeted application, it is highly correlated to the actual consumed data. Thus our defined per-subscriber data demand is a simple approximation for representing it.

Each BS delivers different coverage for cellular service. In order to understand the real spatial distribution of connection devices or traffic demand, it is vital to consider the different area of cell coverage. As the actual area is difficult to measure, Voronoi cell [22] is widely used in computing the

traffic load density [8], [12], [13]. Thus we obtain the area of Voronoi cells drawn by using the locations of BSs. Let X represent the whole network area. Further let K be the set of BS indices and $B = \{b_k, k \in K\}$ be the set of BSs. The Voronoi cell V_k, associated with the BS b_k, is the set of all the points in X whose distances to b_k are not greater than their distances to any other BS b_j with $j \neq k$. One limitation of this methodology is the accuracy of Voronoi cells in multi-tier cellular network. In Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TMC.2017.2752159, we demonstrate the reliability of using Voronoi cells, and then provide the possible solutions to the above limitation.

In this way, we obtain the subscriber density by dividing the number of subscribers with the area of the corresponding Voronoi cell, denoted as $S^{b_i}(t)$ (subscribers/km^2) for $b_i \in B$ and $1 \leq t \leq 744$, where t is the time sequence index. The length of the duration is 1 hour, which explains why the maximum is $744 = 24 \times 31$. Similarly, we denote the per-subscriber demand as $D^{b_i}(t)$ (bytes/subscriber) for $b_i \in B$ and $1 \leq t \leq 744$. It is worth noting that the product of subscriber density $S^{b_i}(t)$ and data demand $D^{b_i}(t)$ is the traffic density, denoted as $T^{b_i}(t)$ (bytes/km^2), which represents the degree of a per-cell traffic load.

4 Network Capability Analysis

In this section, we focus on three metrics of data traffic: traffic density, subscriber density and average demand per subscriber. The first two parameters are aggregated metrics of a BS, while the third one is a per-subscriber metric. By showing the heat maps, we provide a visual view on how they are geographically distributed in the urban area. Then we propose a model to describe the spatial distributions of the empirical data. The spatial distribution characterizes how these metrics change and distribute among different cells. Our main analyses are based on *Shanghai* data. After that, we build the similar distribution model on *Kunming* data.

4.1 Visualized Analysis

Fig. 2 shows the heat maps of the mean subscriber density, average demand per subscriber and traffic density in a month. Since the empirical data are highly right-skewed, the log-transformed data are used to draw heat maps. Subscriber density is high and concentrated in the city center, while it is relatively low in the rural area. However, the heat map of average demand per subscriber shows different characteristics: the peak values spread widely, from the city center to rural area.

To further reveal the relationships among the three parameters (traffic density, subscriber density and average demand per subscriber), correlation coefficients are used to test the correlations between them. The results show that traffic density and subscriber density are highly correlated, with the correlation coefficients greater than 0.9, which is reasonable. As a result, their fraction, i.e., the demand per subscriber, is expected to be fairly constant, which is confirmed by Fig. 2b where the monthly per-subscriber demand values range from 36 to 180 MB (with a logarithm between 17.4 and 19). Thus absolute values of correlation coefficients between per-subscriber demand and the other two parameters are less than 0.1, indicating the weak correlation between per-subscriber demand and subscriber density or traffic density. This observation explains why Fig. 2 shows the similarity between traffic density and subscriber density, but very different patterns for demand.

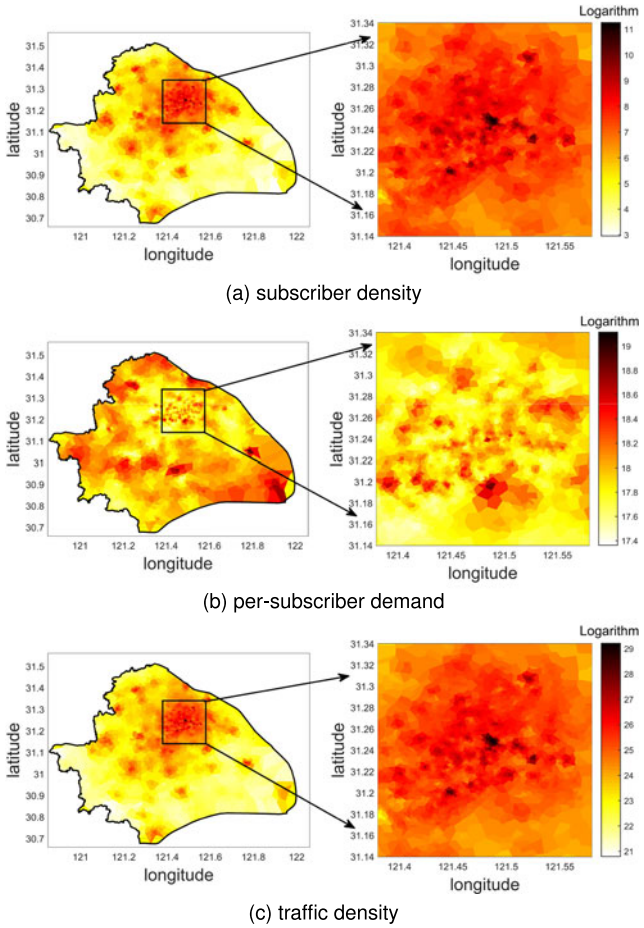


Fig. 2. Geographical distributions of subscriber density (a), per-subscriber demand (b), and traffic density (c).

4.2 Spatial Distributions

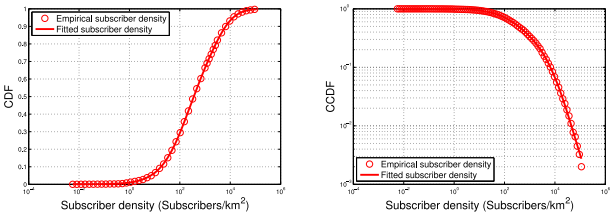
Our next step is to model the spatial distributions of subscriber density and average demand. Researchers [8] found that the spatial distribution of per-cell traffic density can be well fitted by a log-normal mixture distribution. We want to know *what is the reason behind this distribution*. We first start with investigating distributions of subscriber density and average demand. Then we further answer this question by a theoretical analysis of the traffic density distribution.

The probability density function (PDF) of the log-normal mixture distribution with l components is:

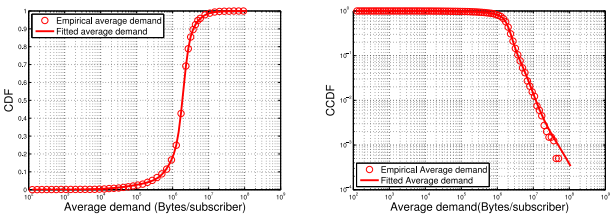
$$f_X(x) = \sum_{i=1}^{l} p_i \log \mathcal{N}(x; \mu_i, \sigma_i), \tag{1}$$

where $\log \mathcal{N}(x; \mu_i, \sigma_i)$ is the ith log-normal distribution with location parameter μ_i and scale parameter σ_i, while p_i is the mixture proportion of the ith component and the sum of all the mixture proportions is $\sum_{i=1}^{l} p_i = 1$. The parameters $\{\mu_i, \sigma_i, p_i\}_{i=1}^{l}$ can be obtained for example using the expectation maximization (EM) algorithm [23].

Considering the trade-off between accuracy and model complexity, a log-normal mixture with three components $l = 3$ is used to fit both hourly subscriber density and hourly demand. Fig. 3 shows both cumulative distribution function (CDF) and complementary CDF (CCDF) of



(a) CDF, subscriber density (b) CCDF, subscriber density



(c) CDF, average demand (d) CCDF, average demand

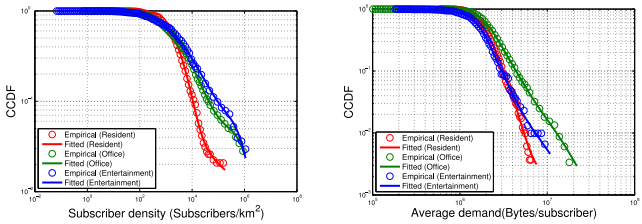
Fig. 3. Log-normal mixture fittings of the spatial distributions of subscriber density and average demand, in the timescale of one hour, in *Shanghai*. The circles represent the empirical data, and the solid lines represent the fitted log-normal mixture distribution.

TABLE 2
Parameters of the Log-Normal Mixture Models for Subscriber Density and Average Demand

Parameters		Subscriber density	Average demand
Location parameters	μ_1	5.4094	14.5001
	μ_2	5.1033	14.3824
	μ_3	8.2199	12.7958
Scale parameters	σ_1	1.5761	0.2798
	σ_2	2.6085	0.8331
	σ_3	1.2034	1.5647
Mixture proportions	p_1	0.4075	0.4543
	p_2	0.3950	0.4457
	p_3	0.1975	0.1000

empirical data and fitted model, which indicates that the proposed log-normal mixture distribution fits the empirical data very well. Quantitatively, the Kolmogorov-Smirnov (K-S) test is used to test the goodness of fit [24]. We test the distribution fitting of the cell traffic in every hour in a day at 5 percent significance level, and find that the log-normal mixture distribution is accepted all the time. The parameters of the models are listed in Table 2. The median of hourly subscriber density is about 300 (subscribers/km^2), while the highest value can be up to 10^4 (subscribers/km^2). By looking at exponential values of location parameters μ_i (i.e., the expectation of $\log X$), we observe group characteristics behind distributions of hourly subscriber density and hourly demand. Specifically, the three centroids of these two distributions are {164;221;4,447} (subscribers/km^2) and {0.36;1.76;1.98} (MB/subscriber) respectively. In order to show that this characteristic is maintained in different urban regions, we also use the log-normal mixture model to fit the empirical distributions in resident region, office region and entertainment region, as shown in Fig. 4.

Since the 3-component log-normal mixture distribution fits well to both subscriber density and average demand, we apply the similar distribution when fitting their joint distribution. The parameters are listed in Table 3. By comparing between exponential values of μ_i, we observe that there exist different levels of the network capability among thousands



(a) CCDF, subscriber density (b) CCDF, average demand

Fig. 4. Log-normal mixture fittings of the spatial distributions of subscriber density and average demand, in the timescale of one hour, in resident region, office region, and entertainment region, respectively. The circles represent the empirical data, and the solid lines represent the log-normal mixture distribution.

TABLE 3
Parameters of Joint Distribution

Component	p_i	μ_i	σ_i
Component 1	0.60	(4.74,14.38)	$(2.80, 0.17; 0.17, 0.53)^T$
Component 2	0.33	(7.70,14.43)	$(2.22, -0.01; -0.01, 0.05)^T$
Component 3	0.07	(1.59,12.93)	$(5.65, 2.77; 2.77, 7.05)^T$

of BSs. Some BSs serve more subscribers (near 2,200 subscribers/km^2) and these subscribers have higher data demand (near 2 MB/subscirber in an hour), while others only serve few subscribers (near 114 subscribers/km^2) or low-demand subscribers (near 0.4 MB/subscirber in an hour).

Moreover, it can be verified that the product of two independent log-normal mixture distributed random variables also follows a log-normal mixture distribution. We detail the proof of following proposition in Appendix B, available in the online supplemental material.

Proposition 1. *Assume that X and Y are independent log-normal mixture distributed variables with m and n components, respectively. Let $Z = XY$, then Z follows a log-normal mixture distribution with $m \times n$ components. The parameters of the distribution for Z are given by*

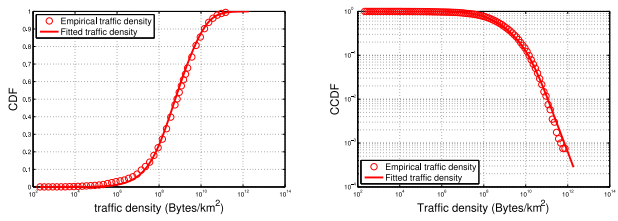
$$\begin{cases} p_{Z_{i,j}} = p_{X_i} p_{Y_j}, \\ \mu_{Z_{i,j}} = \mu_{X_i} + \mu_{Y_j}, \\ \sigma^2_{Z_{i,j}} = \sigma^2_{X_i} + \sigma^2_{Y_j}, \end{cases} \tag{2}$$

for $1 \le i \le m$ and $1 \le j \le n$.

As the correlation between subscriber density and average demand is weak, it can be assumed that they are independent. Thus the product of them, i.e., traffic density, follows a log-normal mixture distribution with 9 components. We can use the parameters given in Table 2 to compute the parameters of the distribution for traffic density, which are listed in Table 4. Fig. 5 shows the fitting of the empirical traffic density to the computed log-normal mixture model. Furthermore, the K-S test at 5 percent significance level also accepts this log-normal mixture distribution. In other words, the per-cell traffic density also follows a log-normal mixture distribution spatially, which is verified by both empirical data and theoretical proof. With a lot of variation between cells, it can be seen from Fig. 5 that the median of per-cell traffic density is 10^9 byte/(hour $\times$ km^2), and the highest value can be up to 10^{12} byte/(hour $\times$ km^2). As shown in Tables 2 and 3, the group characteristics of subscriber density and average demand are connected with their log-normal mixture distributions. By multiplying

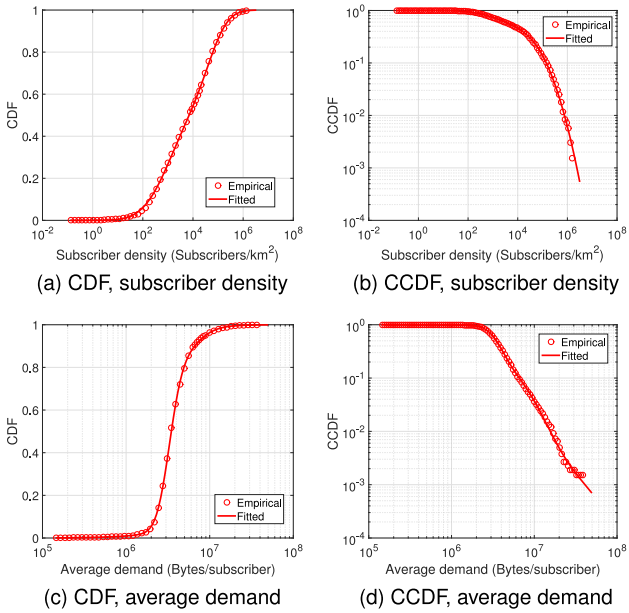
TABLE 4
Parameters of the Log-Normal Mixture Model for Traffic Density

Parameters	Location parameters	Scale parameters	Mixture proportions
Component 1	19.9095	1.6001	0.1851
Component 2	19.7918	1.7828	0.1816
Component 3	19.6034	2.6235	0.1795
Component 4	19.4857	2.7383	0.1761
Component 5	22.7200	1.2355	0.0897
Component 6	22.6023	1.4636	0.0880
Component 7	18.2052	2.6301	0.0408
Component 8	17.8990	3.3522	0.0395
Component 9	21.0156	2.4251	0.0198



(a) CDF, traffic density (b) CCDF, traffic density

Fig. 5. Log-normal mixture fitting of the spatial distribution for traffic density, in the timescale of one hour. The log-normal mixture model is calculated by (2). The circles represent the empirical data, and the solid lines represent the theoretical log-normal mixture distribution.



(a) CDF, subscriber density (b) CCDF, subscriber density
(c) CDF, average demand (d) CCDF, average demand

Fig. 6. Log-normal mixture fittings of the spatial distributions of subscriber density and average demand, in the timescale of one hour, in *Kunming*. The circles represent the empirical data, and the solid lines represent the fitted log-normal mixture distribution.

them together, these characteristics are maintained in the traffic density distribution.

4.3 Generality

To show that this log-normal mixture characteristic is general to cellular networks deployed in different cities, we analyze hourly subscriber density and per-subscriber demand of each BS in *Kunming* data similarly. The results of statistical fittings are shown in Fig. 6. Similar to Fig. 3, the proposed log-normal mixture distribution fits the empirical data very well. Also, using the K-S test, we test these two distribution fittings at 5 percent significance level, and find that the log-normal mixture distribution is accepted all the time.

Though there exist differences between distributions in *Shanghai* and *Kunming*, we can still observe some similarities. Due to the various capability requirements in different usage scenarios, both values of subscriber density and average data demand are widely distributed in each city. For example, majority of subscriber density values, about 90 percent in *Shanghai* and 60 percent in *Kunming*, are within $[1,10^4]$ (subscribers/km^2). Thus, in order to meet the requirements of different usage scenarios, it is vital to achieve the various network capability from aspects of subscriber density and average demand.

In a word, by using log-normal mixture distributions, we are able to accurately fit the spatial distributions of subscriber density, average data demand and traffic density. Moreover, we demonstrate that this distribution model is general to cellular networks in different cities, while the specific parameters are different. All the above advantages motivate us to choose this distribution model when generating synthetic BSs with the various network capability.

5 Network Capability Clustering

In this section, we provide insights into network service capability by extracting its key patterns. Various network capabilities are required for different cellular BSs, in terms of subscriber density and per-subscriber data demand. While some BSs only have a low subscriber density, others serve a large amount of subscribers or even high-demand subscribers. To analyze such different levels of the network capability, we first conduct a 2-round clustering process on thousands of BSs located in *Shanghai* based on the above two key parameters. In each round only one parameter is considered. Then, we investigate the generality of our clustering methodology by clustering BSs located in *Kunming*. Moreover, we reveal the relationship between clusters and functional regions of BSs located in *Shanghai*.

5.1 Clustering Methodology

When clustering BSs, we first need to define the peak hours because the network capability values during this period indicate actual intensity. Our previous work [6] exploited the spatial information embedded within mobile traffic by identifying key urban functional regions, such as resident region, transport region, office region, entertainment region and comprehensive region. More specifically, each BS was connected with the urban functional information of its deployed region. It was found that hourly dynamics of BSs in the same functional region follow the same pattern, with similar peak and non-peak hours. Based on this finding, we define the peak hour of BSs in a certain functional region as the hour when the average traffic density in this region is relatively higher, i.e., over 50 percent of its peak value. In this way, the averages of subscriber density and demand, i.e., $S^{b_i}(t)$ and $D^{b_i}(t)$, for each BS during peak hours are obtained, which are denoted by $Sp_a^{b_i}$ and $Dp_a^{b_i}$, respectively, and they are computed as

$$Sp_a^{b_i} = \frac{1}{|\mathcal{P}_i|}\sum_{t\in\mathcal{P}_i} S^{b_i}(t), \tag{3}$$

$$Dp_a^{b_i} = \frac{1}{|\mathcal{P}_i|}\sum_{t\in\mathcal{P}_i} D^{b_i}(t), \tag{4}$$

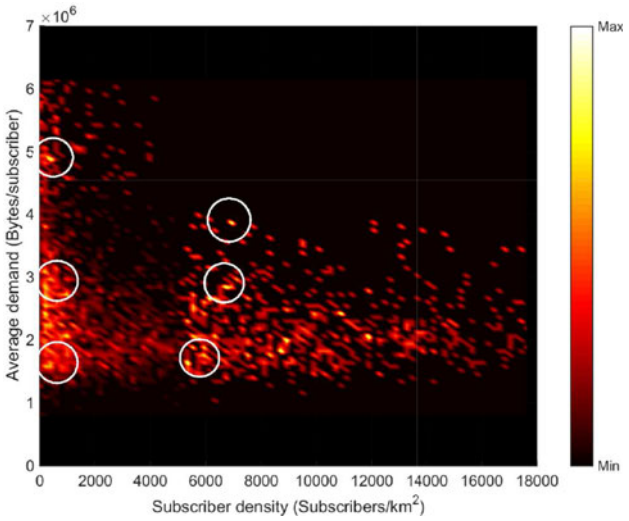


Fig. 7. Clustering result of *Shanghai* data in the space of subscriber density and average demand, in the view of density.

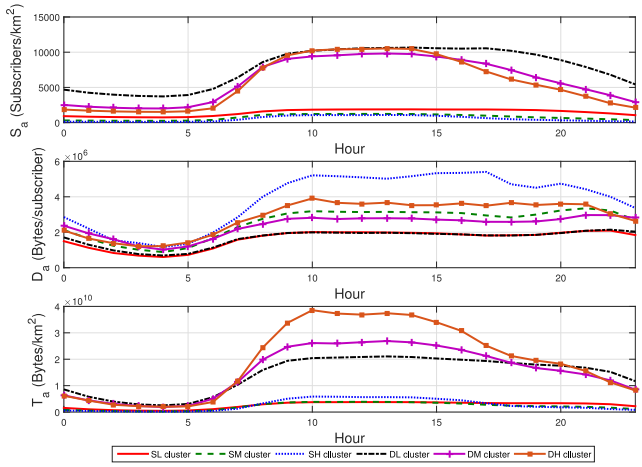


Fig. 8. Temporal dynamics of average subscriber density, data demand and traffic density in each cluster of *Shanghai* data. Each curve represents the average 24-hour dynamics of one BS cluster.

where $\mathcal{P}_i$ denotes the set of peak hours for BS b_i.

Since the best number of clusters is unknown, we choose the agglomerative hierarchical clustering algorithm [25]. Each BS is regarded as a vertex $\mathcal{V}_i$ with vertex value v_i (the values assigned to vertexes will be explained later). Each cluster is a set of vertexes (BSs), denoted by C_n. The distance between two vertexes $\mathcal{V}_i$ and $\mathcal{V}_j$ is $d(\mathcal{V}_i, \mathcal{V}_j) = |v_i - v_j|$. Using the average linkage criterion, the distance between two clusters C_m and C_n is measured as the average distance between vertexes in C_m and vertexes in C_n, i.e., $d(C_m, C_n) = \frac{1}{|C_m||C_n|}\sum_{V_i \in C_m, V_j \in C_n} d(V_i, V_j)$. Agglomerative hierarchical clustering starts by considering each vertex as a cluster. During each iteration, it calculates the distances between all pairs of clusters and merges the two clusters with the minimum distance into one cluster. The clusters continue merging until all vertexes are included in one cluster. In this way a hierarchical dendrogram is generated. In the next step, Silhouette criterion [26] is used to decide where to cut the hierarchical dendrogram in order to get the best separation among vertexes. Mathematically, silhouette $sil(i)$ of a vertex $\mathcal{V}_i$ is defined as

$$sil(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{5}$$

where $a(i)$ is the average of the distance from $\mathcal{V}_i$ to all vertexes in the same cluster C_m and $b(i)$ is the lowest average of the distance from $\mathcal{V}_i$ to other clusters. Normally $sil(i)$ ranges from -1 to 1, with a large value representing $\mathcal{V}_i$ is similar to its own cluster and dissimilar to others. Thus an average $sil(i)$ over all data is a measure of how appropriately the data have been clustered.

Clustering of BSs are based on $Sp_a^{b_i}$ and $Dp_a^{b_i}$ of each BS. However, due to lack of knowledge of relationship between $Sp_a^{b_i}$ and $Dp_a^{b_i}$, it is hard to define a suitable distance metric in the 2-dimensional space. In order to characterize different levels of the network capability in terms of $Sp_a^{b_i}$ and $Dp_a^{b_i}$, we perform the clustering process in two rounds, with only one parameter considered in each round. In the first round, each vertex is assigned the value $Sp_a^{b_i}$ of the BS. It divides the BSs into two subscriber-density-based clusters, C_1 and C_2, corresponding to low subscriber density and high subscriber density. The second round goes inside C_1 and C_2, respectively. Each vertex is assigned the value $Dp_a^{b_i}$ of the BS. In both C_1 and C_2, three average-demand-based subclusters are found, with low demand, medium demand, and high demand, respectively. Specifically, we give the details of deciding the best number of clusters in Appendix C, available in the online supplemental material. By varying the thresholds, we obtain the maximal Silhouette values of resulting clusters, i.e., 0.50 and (0.55,0.52), in the first and second round, respectively.

Finally, combining the two rounds of clustering process, we obtain six clusters. For the sake of representation, we name them as follows: SL (sparse low), SM (sparse medium), SH (sparse high), DL (dense low), DM (dense medium) and DH (dense high). $Sparse$ clusters contain BSs that serve under 4,100 subscribers per km^2, while $dense$ BSs of clusters serve more subscribers. As for average data demand, low clusters contain BSs that serve subscribers with a low demand of data, less than 1.5 MB per hour. Boundary between $medium$ and $high$ clusters is about 4 MB per hour. Moreover, we apply other criteria, i.e., Calinski-Harabasz index and Davies-Bouldin index [27], to decide the best number of clusters. Again we obtain 6 clusters. Due to space constraints, we will not show all these results. Next part better details our BS clusters.

5.2 Analysis of Clustering Results

In this part, we further analyze the clustering results. The clustering results of BSs located in *Shanghai* are shown in Fig. 7, where each BS is mapped onto the $\left(S^{b_i}, D^{b_i}\right)$ space. Rather than plotting in scatter form, we depict the number of BS samples in the unit area of the $\left(S^{b_i}, D^{b_i}\right)$ space, i.e., density, by the brightness of color bar. For better illustration, the plotted circles indicate 6 brightest areas, which are centroids of clusters. It is also noteworthy that some BS samples are far away from any clusters, like a bright area near (9000, 2), which represents a potential that there would be another cluster if more BSs should serve over 10,000 subscribers per km^2.

The six clusters exhibit different characteristics, in terms of the network capability. Fig. 8 presents the temporal dynamics of these capability parameters, i.e., $S_a(t)$, $D_a(t)$ and $T_a(t)$, in each cluster. There are significant differences on the time and duration of peak hours as well as the peak values among different clusters, indicating different levels

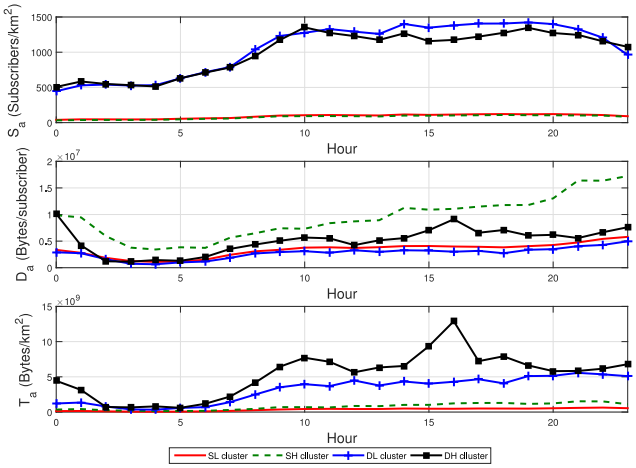


Fig. 9. Temporal dynamics of average subscriber density, data demand and traffic density in each cluster of *Kunming* data. Each curve represents the average 24-hour dynamics of one BS cluster.

of the network capability. Specifically, $dense$ clusters have higher $S_a(t)$, while $high$ clusters have higher $D_a(t)$. Another interesting observation is that subscriber density rises and reaches its peak hours during daytime, and then drops to low values at night, while average demand rises during daytime and does not drop to low values until midnight. This reveals that although subscriber density tends to be low at night, subscribers still consume much data traffic during night. These insights indicate that different peak and non-peak time periods are needed to characterize the hourly dynamics of subscriber density and average demand in each cluster, which is discussed in Section 6.

The advantage of the 2-round clustering is capturing the characteristics in terms of both subscriber density and average demand. To highlight this, we perform a comparative analysis of a one-shot clustering. The details are provided in Appendix C, available in the online supplemental material.

5.3 Generality

In this part, we repeat the above clustering process on *Kunming* data. Unlike those in *Shanghai*, BSs located in *Kunming* are clustered into four clusters, named as SL (sparse low), SH (sparse high), DL (dense low) and DH (dense high). The dynamics $\{S_a(t), D_a(t), T_a(t)\}$ of each cluster are plotted in Fig. 9.

Similarly, BSs in different clusters are with different network capabilities from aspects of subscriber density and average data demand. Thus we observe differences among these dynamics plotted in Fig. 9. BSs in DH cluster have the highest traffic density values, while traffic density values of BSs in SL cluster are the lowest. Comparing with those of *Shanghai* plotted in Fig. 8, different peak and non-peak periods between $S_a(t)$ and $D_a(t)$ are also observed. As for differences, as densely populated areas are common in a metropolis, $S_a(t)$ of BSs in *Kunming* are significantly lower than those in *Shanghai*, about 1,500 versus 10,000. Also, the peak of daily traffic dynamics in *Kunming* always appears at night, while that in *Shanghai* normally appears at about 12AM. To explain this, we look at both $S_a(t)$ and $D_a(t)$ in *Kunming*. We find that $D_a(t)$ slowly rises to its peak at night, even though $S_a(t)$ maintains a high level from daytime to night.

Though the number of cluster is not applicable to another city, the clustering process of mobile network capabilities is applicable among different urban areas, indicating the generality of this methodology.

TABLE 5
Distribution of BSs in Different Clusters and Functional Regions

Types of BS	SL	SM	SH	DL	DM	DH	Total
Resident	381	40	3	106	3	1	534
Transport	9	46	1	1	9	1	67
Office	481	340	103	145	85	23	1177
Entertainment	82	48	11	45	18	4	208
Comprehensive	359	109	13	170	30	4	685
Total	1312	583	131	467	145	33	2671

TABLE 6
Mapping Relation (%)

Types of BS	SL	SM	SH	DL	DM	DH
Resident	71.35	7.49	0.56	19.85	0.56	0.19
Transport	13.43	68.66	1.49	1.49	13.43	1.49
Office	40.87	28.89	8.75	12.32	7.22	1.95
Entertainment	39.42	23.08	5.29	21.63	8.65	1.92
Comprehensive	52.41	15.91	1.90	24.82	4.38	0.58

5.4 Differences among Urban Functional Regions

Our previous work [6] investigated the relationship between the traffic pattern and urban functional regions where BSs are deployed. More specifically, land use of regions can be detected by clustering traffic profiles of BSs. Applying the corresponding algorithm, each BS in our data is labeled with an urban functional region. Inspired by this, it is necessary to investigate differences of network capability among urban functional regions. Table 5 shows the numbers of BSs in each cluster and each functional region of *Shanghai*. A highly asymmetric characteristic is observed: almost 50 percent of BSs are in the SL cluster (low subscriber density and low average demand), which is consistent with the right-skewed log-normal mixture distribution we mentioned previously.

To further investigate the relationship between the (S^{bi}, D^{bi}) profile and the geographical location of BSs, several parameters are defined to measure the relationship between six clusters and five functional regions. Let $N_{m,n}$ denote the number of BSs in the mth functional region and the nth cluster. Then $\sum_{j=1}^{6} N_{m,j}$ is the total number of BSs in the mth functional region. The mapping relation, $P_{m,n}$, is defined to represent the proportion of the nth cluster of BSs in the mth region, which is given by

$$P_{m,n} = \frac{N_{m,n}}{\sum_{j=1}^{6} N_{m,j}}. \tag{6}$$

We will use this parameter in the next section when we use our model to generate synthetic BSs in different urban areas. Table 6 shows the mapping relation $P_{m,n}$. It is clear that the SL cluster is the main cluster in every functional regions, except for transport region, where the SM is the main cluster.

In order to reveal the differences among various functional regions, the relative proportion $P'_{m,n}$ is defined as follows. First compute

$$R_{m,n} = \frac{N_{m,n}}{\sum_{i=1}^{5} N_{i,n}}, \ \forall m, n. \tag{7}$$

TABLE 7
Relative Proportion (%)

Types of BS	SL	SM	SH	DL	DM	DH
Resident	44.01	10.40	3.47	3.14	34.40	4.59
Transport	3.65	41.99	4.06	33.03	1.14	16.13
Office	11.01	17.51	23.61	17.61	9.32	20.93
Entertainment	10.96	14.43	14.72	21.76	16.89	21.25
Comprehensive	21.86	14.93	7.93	16.53	29.08	9.68

Then

$$P'_{m,n} = \frac{R_{m,n}}{\sum_{j=1}^{6} R_{m,j}}. \tag{8}$$

The definition of relative proportion $P'_{m,n}$ eliminates the differences in the absolute numbers of BSs in different clusters, and thus it characterizes different patterns of subscriber density and demand in different functional regions. The values of $P'_{m,n}$ are listed in Table 7. Compared with Table 6, several different observations are made:

- In office regions, the relative proportion of SH cluster is the highest, followed by DH cluster. This indicates that office regions tend to handle heavier average traffic demand by each subscriber.
- In entertainment regions, the relative proportion of DL cluster is the highest, followed by DH cluster, indicating that entertainment regions handle larger numbers of subscribers. This is consistent with our common sense that entertainment regions tend to have denser population.
- In comprehensive regions, the relative proportion of DM cluster is the highest, followed by SL cluster, which indicates that the subscriber density in comprehensive regions is relatively higher than that in other regions.

Since different regions have different proportions for each cluster (measured by the mapping function $P_{i,j}$), it is obvious that these regions have different characteristics in subscriber density and average data demand, and thus they have different network capabilities. This observation will be utilized in the next section to model the network capability in large-scale environments. Moreover, since we already demonstrated that BSs located in *Kunming* can be similarly clustered into four clusters, we can repeat the above analysis and characterize the differences of network capability among urban functional regions, i.e., mapping relation. Therefore, the generality of our modeling methodology is still guaranteed.

6 Network Capability Modeling

To build an accurate model for the urban network capability, both capability parameters ($\{S^{b_i}(t), D^{b_i}(t), T^{b_i}(t)\}$) and numbers of synthetic BSs need to be consistent with those of real BSs. The schematic of our model is illustrated in Fig. 10. The previous clustering process extracts several network capability patterns among all BSs, each of which has unique characteristics in terms of the number of access subscribers during a certain period (subscriber density $S^{b_i}(t)$) and traffic volume they consumed (data demand $D^{b_i}(t)$). Thus the idea is to build a capability model which can generate different types of synthetic BSs in terms of subscriber density $S^{b_i}(t)$ and average data demand $D^{b_i}(t)$, i.e., DH, DM, DL, SH, SM and SL. With the input of urban function context information, the first step (Section 6.1) is to compute the proportion of each BS type in a given urban area, using mapping relation $P_{m,n}$, i.e., relationship between network capability and geographical context of BSs. Next for each type of BSs, a certain number of synthetic BSs are generated, as discussed in Section 6.2. After that, we conduct evaluations on the accuracy of our model compared to the original empirical data, in Section 6.3. Since our model is able to generate individual BSs, validations in both aggregate and individual level are considered. Finally, to investigate the performance gain of building independent models on different type of BSs, we compare it with a simplified model in Section 6.4.

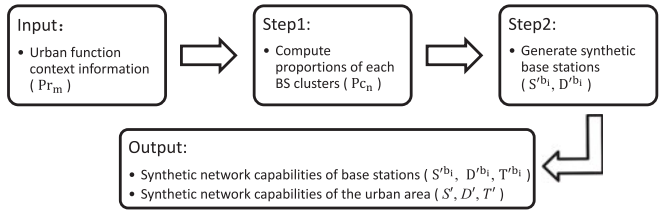


Fig. 10. Modeling Methodology.

In this section we focus on *Shanghai* data only. However, given that both log-normal mixture distribution of network capability and clustering method of BSs are workable in cellular networks of different cities, we can easily extend our modeling on other data.

6.1 Base Station Proportion Computation

Building a model of network capability in the given urban area first requires us to generate different BS types. We compute the proportion of each BS type, i.e., the probability used in synthetic BS generation, which is denoted as Pc_n for $1 \le n \le 6$, corresponding to DH, DM, DL, SH, SM and SL, respectively. The input is the proportion of the BSs deployed in different urban functional regions, denoted as Pr_m for $1 \le m \le 5$, corresponding to resident, transport, office, entertainment and comprehensive regions, respectively. Recall that we have already obtained the mapping relation $P_{m,n}$, i.e., the proportion of the nth type of BSs in the mth region, which are listed in Table 6. Pc_n can be computed as follows

$$Pc_n = \sum_{m=1}^{5} Pr_m \times P_{m,n}. \tag{9}$$

Then we generate a set S of synthetic BSs with $|S| = |B|$. Each type of synthetic BSs in S has the same proportion as that in the original BS set B. The details of synthetic BS generation are given next.

6.2 Synthetic Base Station Generation

The process of generating synthetic BSs can be divided into two parts, fitting ($S^{b_i}(t)$, $D^{b_i}(t)$) distributions and generating synthetic BSs correspondingly. In our model, we focus on the one-day dynamics of $S^{b_i}(t)$ and $D^{b_i}(t)$. Thus we compute the average one-day sequences in whole month, denoted as $S_a^{b_i}(t_h)$ and $D_a^{b_i}(t_h)$, as

$$S_a^{b_i}(t_h) = \frac{1}{31}\sum_{j=1}^{31} S^{b_i}(t_h + (j-1)\times 24), \tag{10}$$

$$D_a^{b_i}(t_h) = \frac{1}{31}\sum_{j=1}^{31} D^{b_i}(t_h + (j-1)\times 24), \tag{11}$$

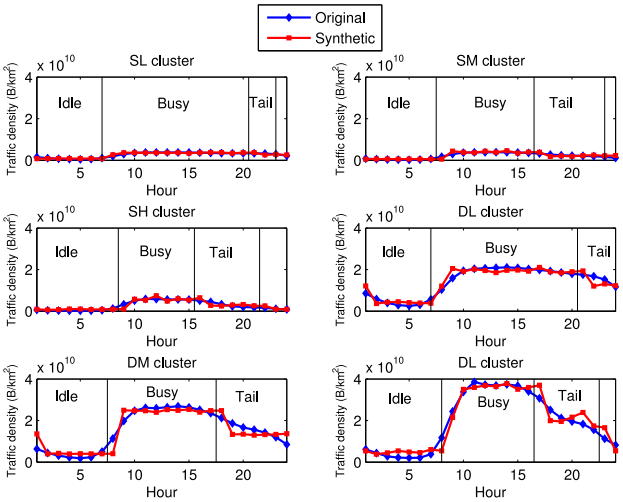


Fig. 11. Performance on modeling aggregated dynamics. Synthetic dynamics of average traffic density for each type of BSs are plotted, along with the original empirical ones.

for $1 \leq t_h \leq 24$, where $(t_h + (j-1) \times 24)$ represents the t_hth hour in the jth day. Inspired by our findings that the network capability of BSs differ significantly across different periods, we add this characteristic in our model and design a dividing mechanism. To ensure that each period is continuous, we set 80 percent of $\max\{S_a^{b_i}(t_h)\}$ and 60 percent of $\max\{D_a^{b_i}(t_h)\}$ as thresholds. As shown in Fig. 11, 3 continuous periods are obtained for each type of BSs. *Idle* periods represent periods when both $S_a^{b_i}(t_h)$ and $D_a^{b_i}(t_h)$ are under the related thresholds, while in *Busy* periods they are both above the thresholds. In *Tail* periods, $S_a^{b_i}(t_h)$ are under its threshold and $D_a^{b_i}(t_h)$ are above its threshold. With decreasing number of access subscribers, BSs in this period still transmit fairly large amounts of data, which is between *Busy* and *Idle*, because data demand per subscriber remains high. Note that the fourth case will not happen because the above-threshold periods of $D_a^{b_i}(t_h)$ cover those of $S_a^{b_i}(t_h)$.

In Section 4, we use the log-normal mixture distributions to fit the empirical distributions of subscriber density $S^{b_i}(t)$ and data demand per subscriber $D^{b_i}(t)$ in the original data. Besides high accuracy, we also show that this law holds true in different urban regions (Fig. 4) and even cities (Fig. 6). Thus when fitting the synthetic spatial distributions of $S^{b_i}(t)$ and $D^{b_i}(t)$ in each BS type, we naturally choose this log-normal mixture model. More specifically, we use the 2-dimensional log-normal mixture model to preserve the correlation between subscriber density and average data demand. In other words, we fit $S_a^{b_i}(t_h)$ and $D_a^{b_i}(t_h)$ together.

Given type of BS and urban function region where it is deployed, we obtain 3 CDFs for *Idle*, *Busy* and *Tail* periods. Since the log-normal mixture fitting is already detailed in Section 4, we skip this for space economy reason. Finally we obtain the fitting parameters needed in the network capability model.

We now briefly describe how to generate a synthetic BS using the capability model, in Algorithm 1. With inputs of BS number and fitted distributions, one-day dynamics of 3 parameters for each synthetic BS are generated. Since there exists a strong correlation in temporal dynamics of each individual BS, we need to carefully preserve this information. For each hour t_h, we first map it to the corresponding period n (line 5). Then we sample a possible 3-parameter set of the size N based on the fitted distribution $P_n^{s\times d}$ (lines 7-11). Considering that the traffic in current hour is similar to that one hour ago, i.e., temporal correlation, high-capability BS is always with higher capability in every hour. Thus we sort the set $\{(S'^{b_i}_a(t_h), D'^{b_i}_a(t_h), T'^{b_i}_a(t_h))\}$ by the key $T'^{b_i}_a(t_h)$ (line 12) and obtain 24 sorted sets corresponding to 24 hours. The one-day dynamics of 3 parameters for N BSs are constructed by iteratively connecting 24 elements, each of which is selected from the corresponding set (lines 15 to 19).

Algorithm 1. Synthetic Base Station Generation

Input: Base station number N, Fitted distributions $P_n^{s\times d}$ of period n, for $n = 1, 2, 3$

Output: Synthetic base stations b_i with capability parameters $(S'^{b_i}_a(t_h),\ D'^{b_i}_a(t_h),\ T'^{b_i}_a(t_h))$, for $i = 1, 2, ..., N$ and $t_h = 1, 2, ..., 24$

1: **Initialize :**
2: $c[l] \leftarrow set(),\ l = 1, 2, ..., 24$
3: **for** $t_h = 1$ to 24 **do**
4: //Map hour t_h to the corresponding period n.
5: $n \leftarrow period(t_h)$
6: //Sample possible set of 3 parameters for N BSs.
7: **for** $m = 1$ to N **do**
8: $S'^m_a(t_h),\ D'^m_a(t_h) \leftarrow sample(P_n^{s\times d})$
9: $T'^m_a(t_h) \leftarrow S'^m_a(t_h) \times D'^m_a(t_h)$
10: $c[t_h].add((S'^m_a(t_h), D'^m_a(t_h), T'^m_a(t_h)))$
11: **end for**
12: $sort(c[t_h], key = T'_a(t_h))$
13: **end for**
14: //Construct one-day dynamics of 3 parameters for N BSs.
15: **for** $i = 1$ to N **do**
16: **for** $t_h = 1$ to 24 **do**
17: $S'^{b_i}_a(t_h), D'^{b_i}_a(t_h), T'^{b_i}_a(t_h) \leftarrow c[t_h][i]$
18: **end for**
19: **end for**

6.3 Model Validation

6.3.1 Aggregate Level Validation

We now evaluate the accuracy of our method of building model, in the aggregate level. Though our model outputs 3 parameters, traffic density equals the product of subscriber density and average demand. Thus we only perform the validation of the traffic density, which is a good representation of validating other 2 parameters. For each type of *DH*, *DM*, *DL*, *SH*, *SM* and *SL*, we generate a set of 10,000 synthetic BSs based on the fitted log-normal mixture distribution. Each synthetic data set contains subscriber density and average demand of BSs in a certain cluster in 24 hours. In Fig. 11, we evaluate the model performance by comparing the synthetic traffic density dynamics for each type generated by the model with the original empirical dynamics. The one-day dynamics are plotted in 3 periods, i.e., *Idle*, *Busy* and *Tail*. It can be seen from Fig. 11 that the traffic density dynamics of the synthetic BSs have similar patterns to those of the original real BSs.

Next we evaluate how consistent the 6-type synthetic traffic density is by comparing its distribution with that of the original real BSs in the set B. To this aim, we use the Bhattacharyya (BH) measure or distance [5], which quantifies the similarity between two probability distributions

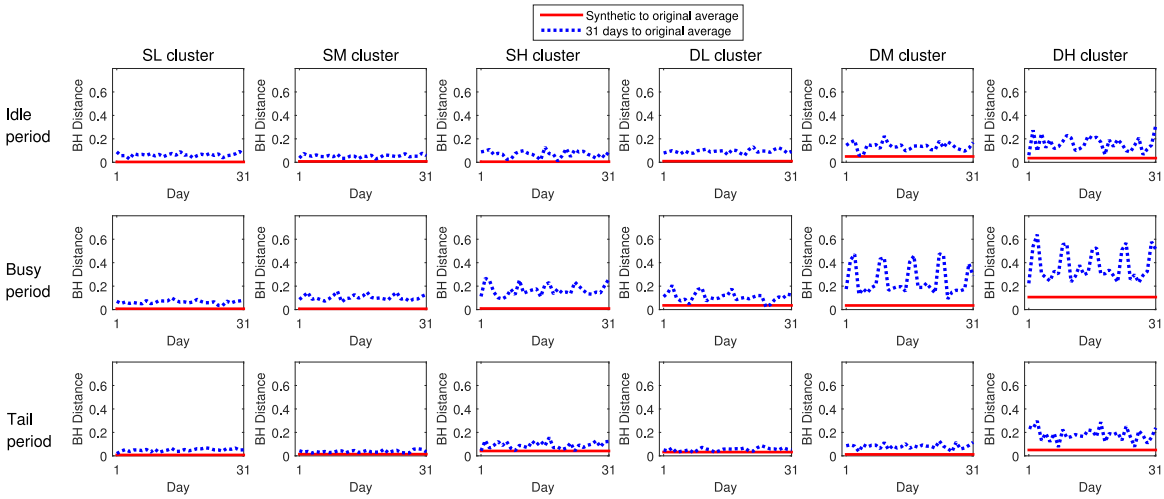


Fig. 12. Performance on modeling statistical distributions. Horizontal solid lines are Bhattacharyya distances between the traffic density distributions of original average and synthetic day, while dashed curves are Bhattacharyya distances between the traffic density distributions of original average and 31 days in the original trace.

$p(x)$ and $p'(x)$. For discrete probability distributions, the BH measure is defined by

$$\rho(p,p') = \sum_{x\in X} \sqrt{p(x)p'(x)}, \tag{12}$$

while for continuous probability distributions, it is given by

$$\rho(p,p') = \int \sqrt{p(x)p'(x)}\, dx. \tag{13}$$

In order to satisfy all the metric axioms, we use an alternative distance metric based on the BH measure defined as

$$d(p,p') = \sqrt{1-\rho(p,p')}. \tag{14}$$

Note that $d(p,p') = 0$ iff $p = p'$, indicating two identical distributions.

Let $\mathbb{M}$ denote the set of 31 days in the dataset (from 1st August to 31st August), and $\mathbb{H}$ denote the set 24 hours in a day. $\mathbb{H}$ is further divided into 3 subsets, denoted as $\mathbb{H}_m$ for $m = 1, 2, 3$, which correspond to *Idle*, *Busy* and *Tail* periods, respectively. In the sequel, the PDF of traffic density is denoted by $p_n^{X_m}(x)$, where m represents the time period $\mathbb{H}_m$, n denotes the BS type (DH, DM, DL, SH, SM or SL), and X represents the dataset.

For each BS in type n, a synthetic dataset of subscriber density and average demand for one day is generated, based on the log-normal mixture distribution model. We then obtain the synthetic traffic density of each BS by multiplying the subscriber density and average demand. The PDFs of traffic density in the synthetic dataset are denoted as $\{p_n^{S_m}\}$, while the PDFs of average traffic density over 31 days in the original dataset are denoted as $\{p_n^{A_m}\}$ and the PDFs of the original traffic density in a given day $D \in \mathbb{M}$ are denoted as $\{p_n^{D_m}\}$.

With distributions of both synthetic data $\{p_n^{S_m}\}$ and original data $\{p_n^{A_m}\}$, $\{p_n^{D_m}\}$, how to define a criterion of distance is still questionable. To evaluate our traffic density model, we first compute $d(p_n^{A_m}, p_n^{S_m})$, the distance between the traffic density distributions of averaged original data and synthetic day S, in terms of BS types n and time periods m. Then, we compute $d(p_n^{A_m}, p_n^{D_m})$ for $D \in \mathbb{M}$, the distance between the distributions of averaged original data and 31 days $D \in \mathbb{M}$ in the original trace. In this way, $d(p_n^{A_m}, p_n^{S_m}) < d(p_n^{A_m}, p_n^{D_m})$, for all possible D, indicates that the distance between synthetic and original data is always smaller than that within original data itself. Fig. 12 plots $d(p_n^{A_m}, p_n^{S_m})$ and $d(p_n^{A_m}, p_n^{D_m})$, where there are 18 figures for 6 BS types and 3 time periods. Finally, we also compute the probability of $d(p_n^{A_m}, p_n^{S_m}) < d(p_n^{A_m}, p_n^{D_m})$ for all possible $\{D, m, n\}$, which is over 0.95. Thus the error of our model, i.e., $d(p_n^{A_m}, p_n^{S_m})$, is sufficiently small.

Based on the above aggregated-level evaluations, we have demonstrated that our model performs well on characterizing the traffic density of BSs, in terms of both dynamics and statistical distribution.

6.3.2 Per-BS Level Validation

In order to further evaluate performance in per-BS level, we now provide a use case in a real-world cellular network deployed in *Shanghai*. The input is the distribution of urban functional regions and the number of BSs deployed in each region, i.e., Pr_m listed in Table 8. Using the mapping relation $P_{m,n}$ listed in Table 6, the probabilities of 6 BS types Pc_n can be computed by (9). Then subscriber density and per-subscriber data demand of synthetic BSs, i.e., $(S'^{b_i}_a(t_h), D'^{b_i}_a(t_h))$, are generated by applying Algorithm 1. The output is the network capability dynamics in the scales of BS and whole urban area, i.e., $\{(S'^{b_i}_a, D'^{b_i}_a, T'^{b_i}_a\})$ and (S'_a, D'_a, T'_a). Since our generated BSs are sorted according to their network capability values, we also process similarly on original BSs to match them with those synthetic BSs.

In Fig. 13, we first plot the subscriber density, average demand and traffic density of the whole urban area, i.e., averaging over all the BSs. More specifically, we compare the synthetic subscriber density, average demand and traffic density $\{S'_a(t_h), D'_a(t_h), T'_a(t_h)\}$, generated by our model,

TABLE 8
Percentage of BSs Deployed in Each Region

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

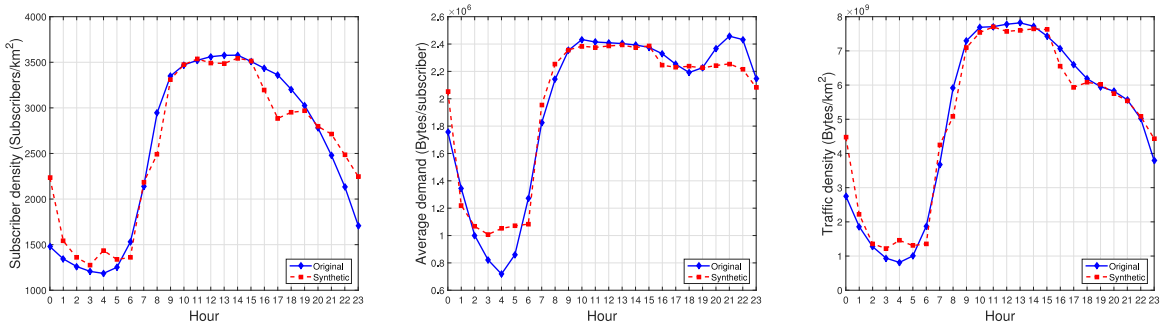
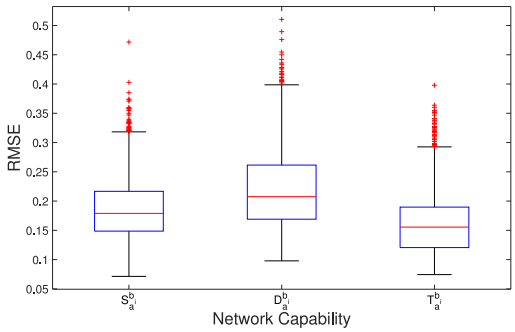


Fig. 13. Network capability modeling performance: (a) subscriber density, (b) average demand, and (c) traffic density. The dashed curve represents the result based on the synthetic BS set S, while the solid curve represents the result based on the original BS set B.

with the empirical $\{S_a(t_h), D_a(t_h), T_a(t_h)\}$, produced from the original trace data. It can be seen from Figs. 13a and 13b that our model accurately describes the one-day dynamics of subscriber density and average demand. By multiplying subscriber density and average demand, we obtain the similar result in traffic density, as can be seen from Fig. 13c.

One main advantage of our model is that it not only simulates the network capability of the whole urban area, but also can generate synthetic BSs with the various network capability. To evaluate this, we compute the Root Mean Square Error (RMSE) between synthetic capability and original ones, i.e., $\{S'^{b_i}_a(t_h), D'^{b_i}_a(t_h), T'^{b_i}_a(t_h)\}$ and $\{S^{b_i}_a(t_h), D^{b_i}_a(t_h), T^{b_i}_a(t_h)\}$, for each BS. Note that we normalize these three parameters by maximum values of their original data. The result is shown in Fig. 14a, using box-plot. Comparatively, we also list the mean values of these RMSE versus those in the aggregate level, corresponding to Fig. 13, in Fig. 14b. Though higher than aggregate RMSE values, RMSE values of BS individuals are acceptable considering that there are thousands of BSs. According to Fig. 14a, 25 percent of RMSE values for $S'^{b_i}_a(t_h)$, $D'^{b_i}_a(t_h)$ and $T'^{b_i}_a(t_h)\}$ are lower than 0.1487, 0.1690 and 0.1205 respectively, while a few of them are close to those of the aggregate RMSE. More importantly, our model not only generates synthetic traffic load in each cell, but also provides more detailed information on the various network capability among BSs, such as number of access subscribers and intensity of traffic demand.



Parameters	individual	aggregate
Subscriber density	0.1852	0.0757
Average demand	0.2201	0.0579
Traffic density	0.1603	0.0655

(b)

Fig. 14. (a) RMSE between synthetic network capability and original capability of BSs. (b) Mean values of individual RMSE versus aggregate RMSE.

6.4 Model Comparison

The key process in building our model is to extract typical patterns of subscriber density and average demand by clustering BSs. Then we build an independent capability model for each cluster of BSs. However, *how high is the gain by considering clustering strategies in building the model?* To indicate this, we use a simplified model, which ignores the capability differences among BS clusters, for comparison. More specifically, when generating synthetic BSs, we input all samples of subscriber density and average demand to obtain fitted distributions, rather than fitting an independent distribution for each BS cluster.

The results of model comparison are shown in Fig. 15. Similar to Fig. 13, here we also plot 3 dynamics of subscriber density, average demand and traffic density, with dashed curves and solid curves representing our proposed model and original data respectively. The added dotted dashed curves represent the comparative model, from which we observe the relatively worse performance. Quantitatively we compute the RMSE of both proposed and comparative model respectively, listed in Table 9. The RMSE value of traffic density in our proposed model is only 0.0655, which is reasonably lower than 0.1528 in comparative model and achieves a decrease of 57.1 percent.

In a word, our network capability model is more accurate, with the clustering process on subscriber density and average demand of BSs.

7 Discussion

In this section, we will further discuss the strength of our network capability model. The advantages are fourfold:

Accuracy. The accuracy of our proposed model is guaranteed because of two main reasons. First, our statistical analyses show that spatial distributions of both subscriber density and average demand can be well fitted by log-normal mixture models. We also validate the accuracy by the K-S test. Second, with the help of an clustering process, we are able to accurately characterize the network capability by building the independent model for each typical BS clusters. Therefore, we believe that this model can simulate accurate capability of cellular network, without the need of large-scale real traffic records.

Generality. Since the datasets of most previous works are collected from only one certain area, there exists an inevitable selection bias in their model. Thus the practical application of these models is questionable. With the help of two datasets collected from two different cities, *Shanghai* and *Kunming*, we verify that both techniques of fitting spatial

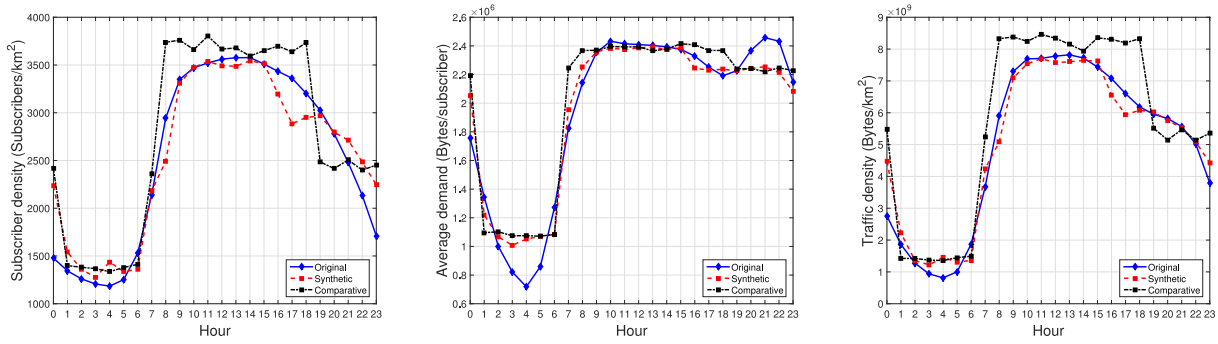


Fig. 15. Network capability model comparison: (a) subscriber density, (b) average demand, and (c) traffic density. Performance of our proposed model (dashed curve) and simplified model (dotted dashed curve) are plotted for comparison, along with the original trace (solid curve).

distribution and clustering BSs are workable in two datasets, indicating the generality of our modeling methodology. Here the generality means that we are able to build similar models, including log-normal mixture distribution of capability parameters and capability-based BS clusters, to characterize cellular networks across different cities, while the specific values of parameters are different across models. It is noteworthy that our proposed model is designed based on a single-tier architecture adopted in 3G networks. Accordingly, we discuss this limitation in Appendix A, available in the online supplemental material. We also provide some possible solutions which ensure the generality of our modeling methodology when applied in other cellular networks using the multi-tier architecture.

Flexibility. Our concepts of subscriber density and average data demand, and the corresponding modeling methodology guarantee the flexibility of this capability model. More specifically, it can flexibly generate synthetic BSs according to the specific capability requirement in a usage scenario. In future mobile networks, the capability requirement for BSs depends a lot on the specific usage scenario. For example, the IoT service requires a BS supporting the large subscriber density, while the high-revolution video service results in a huge data demand of subscribers. Thus, by using the established capability model, we are able to generate synthetic BSs with different levels of the network capability.

Extensive Applications. With the individual BS modeling, we are able to generate the synthetic network capability of each BS, i.e., nearly real-world subscriber density, average data demand and traffic density. Given the type of a cell, including network capability and urban functional information, we apply the corresponding established model. Finally we are able to reconstruct an urban mobile network with synthetic base stations in terms of access subscribers and their average demand.

Our network capability model has extensive applications in cellular network planning, operation and maintenance. Different from the aggregated model, our BS-level model can be helpful in the specific area, like a cell or an urban functional region. In terms of network planning, for example, facing the increasing demand of video content consumption in one area, the telecommunication operator can use our model to generate several synthetic BSs with capabilities of high data demand (such as *SH* and *DH* BSs) and then evaluate the network performance. As for network operation, similarly, by generating synthetic BSs, the operator can test the scheme of network resources allocation in a certain area. Once built, we can directly use the model in the above applications, where the cumbersome process of repeatedly collecting and preprocessing trace data can be skipped.

TABLE 9
Model Comparison (RMSE)

Parameters	Proposed model	Comparative model	Improvment (%)
Subscriber density	0.0757	0.1176	35.6
Average demand	0.0579	0.0838	30.9
Traffic density	0.0655	0.1528	57.1

8 Conclusion

In this paper, we investigate the capability of mobile cellular data networks in large-scale urban environments. Our investigation reveals two important discoveries. First, the spatial distribution of both subscriber density and average traffic demand in each cell can be accurately fitted by log-normal mixture models. Second, using an unsupervised clustering method, we find that large scale base stations can be clustered into several distinct types according to subscriber density and average traffic demand. Inspired by these two observations, we build a data network capability model and use this to generate real base stations with the diverse network capability. More importantly, we verify that the above two observations are general in cellular networks deployed in different cities, indicating the generality of our proposed model. Our evaluations show that the synthetic trace presents a consistent behavior with the original dataset, which demonstrates that our model is precise and flexible. With high accuracy, generality and flexibility, our network capability model has extensive applications in cellular network planning, operation and maintenance. In the future, we will focus more on practical applications of this network capability model. Besides, considering the limitation of using Voronoi cells, we will focus on applying more accurate method of estimating cell coverage in the network capability analysis.

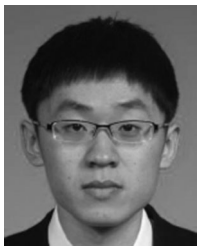
Acknowledgments

This work is supported by the National Nature Science Foundation of China under Grant 91338203, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

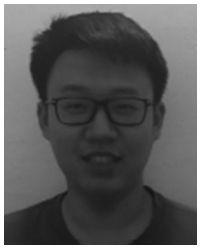
References

[1] NGMN 5G white paper, confirmed in Nov. 2015, [Online]. Available: https://www.ngmn.org/5g-white-paper.html
[2] J. G. Andrews, et al., "What will 5g be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

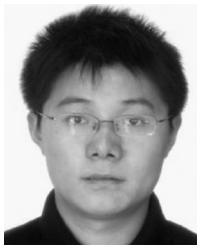
[3] A. Klemm, C. Lindemann, and M. Lohmann, "Traffic modeling and characterization for UMTS networks," in *Proc. IEEE GLOBECOM*, 2001, pp. 1741–1746.
[4] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proc. ACM SIGMETRICS Joint Int. Conf. Measur. Model. Comput. Syst.*, 2011, pp. 305–316.
[5] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2015, pp. 230–235.
[6] H. Wang, F. Xu, Y. Li, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," in *Proc. Internet Measurement Conf.*, 2015, pp. 225–238.
[7] U. Gotzner and R. Rathgeber, "Spatial traffic distribution in cellular networks," in *Proc. IEEE Vehicular Technol. Conf.*, 1998, pp. 1994–1998.
[8] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 80–88, Feb. 2014.
[9] 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. 7th Int. Workshop Hot Topics Planet-Scale Mobile Comput. Online Soc. Netw.*, 2015, pp. 19–24.
[10] 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*, 2013, pp. 468–472.
[11] C. Williamson, E. Halepovic, H. Sun, and Y. Wu, "Characterization of CDMA2000 cellular data network traffic," in *Proc. IEEE Conf. Local Comput. Netw.*, 2005, pp. Z000–719.
[12] S. Hoteit, et al., "Content consumption cartography of the paris urban region using cellular probe data," in *Proc. ACM 1st Workshop Urban Netw.*, 2012, pp. 43–48.
[13] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE INFOCOM*, 2011, pp. 882–890.
[14] D. Naboulsi, R. Stanica, and M. Fiore, "Classifying call profiles in large-scale mobile traffic datasets," in *Proc. IEEE INFOCOM*, 2014, pp. 1806–1814.
[15] V. Soto and E. Frías-Martínez, "Automated land use identification using cell-phone records," in *Proc. 3rd ACM Int. Workshop MobiArch*, 2011, pp. 17–22.
[16] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 1–8.
[17] 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. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2015, pp. 317–326.
[18] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti, "Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong," in *Computational Approaches for Urban Environments*. Berlin, Germany: Springer, 2015, pp. 363–387.
[19] M. Lenormand, et al., "Comparing and modelling land use organization in cities," *Roy. Soc. Open Sci.*, vol. 2, no. 12, 2015, Art. no. 150449.
[20] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-scale measurement and characterization of cellular machine-to-machine traffic," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1960–1973, Dec. 2013.
[21] J. Ding, X. Liu, Y. Li, D. Jin, and S. Chen, "Measurement-driven capability modeling for mobile network in large-scale urban environment," in *Proc. IEEE 13th Int. Conf. Mobile Ad Hoc Sensor Syst.*, 2016, pp. 92–100.
[22] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical Analysis and Modelling of Spatial Point Patterns*. Hoboken, NJ, USA: Wiley, 2008, vol. 70.
[23] G. McLachlan and D. Peel, *Finite Mixture Models*. Hoboken, NJ, USA: Wiley, 2004.
[24] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *J. Amer. Statistical Assoc.*, vol. 46, no. 253, pp. 68–78, 1951.
[25] R. R. Sokal, "A statistical method for evaluating systematic relationships," *Univ. Kans Sci. Bull.*, vol. 38, pp. 1409–1438, 1958.
[26] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[27] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.



Jingtao Ding received the BS degree in electronic engineering from Tsinghua University, Beijing, China, in 2015. He is currently working toward the PhD degree in the Department of Electronic Engineering, Tsinghua University. His research interests include mobile computing, mobile data mining, and user behavior modeling.



Rui Xu is working toward the undergraduate degree in the Electronic Engineering Department of Tsinghua University, Beijing, China. His research interests include mobile big data mining and user behavior modelling.



Yong Li (M'2009-SM'2016) received the BS degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007 and the PhD degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a faculty member in the Department of Electronic Engineering, Tsinghua University. Dr. Li has served as a general chair, TPC chair, TPC member for several international workshops and conferences, and he is on the editorial board of three international journals. His papers have total citations more than 2,300 (six papers exceed 100 citations, Google Scholar). Among them, eight are ESI Highly Cited Papers in Computer Science, and four received conference Best Paper (run-up) Awards. He received the IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers and Young Talent Program of China Association for Science and Technology.

Pan Hui received the BEng and MPhil degrees both from the Department of Electrical and Electronic Engineering, University of Hong Kong, and the PhD degree from the Computer Laboratory, University of Cambridge. He is currently a faculty member of the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology, where he directs the System and Media Lab. He also serves as a Distinguished Scientist of Telekom Innovation Laboratories (T-labs) Germany and an adjunct professor of social computing and networking with Aalto University Finland. Before returning to Hong Kong, he spent several years in T-labs and Intel Research Cambridge. He has published more than 100 research papers and has several granted and pending European patents. He has founded and chaired several IEEE/ACM conferences/workshops, and served on the technical program committee of numerous international conferences and workshops including IEEE Infocom, SECON, MASS, Globecom, WCNC, and ITC. He is a senior member of the IEEE.

Depeng Jin received the BS and PhD degrees from Tsinghua University, Beijing, China, in 1995 and 1999, respectively, both in electronics engineering. Now, he is an associate professor with Tsinghua University and vice chair of the Department of Electronic Engineering. He was awarded the National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design, and future internet architecture. He is a member of the IEEE.