Big Data Analysis Based Network Behavior Insight of Cellular Networks for Industry 4.0 Applications

Dingde Jiang, *Member, IEEE*, Yuqing Wang, Zhihan Lv, Sheng Qi, and Surjit Singh

***Abstract*—In this article, we propose a big data based analysis framework to analyze and extract network behaviors in cellular networks for Industry 4.0 applications from a big data perspective, using Hadoop, Hive, HBase, and so on. The data prehandling and traffic flow extraction approaches are presented to construct effective traffic matrices. Accordingly, we can capture network behaviors in cellular networks from a networkwide perspective. Although there have been a number of prior studies on cellular network usage, to the best of our knowledge, this article is a first study that characterizes network behaviors using the big data analytics to analyze a network big data of call detail records over a longer duration (five months), with more users (five million), more records (several hundred million lines) and nationwide coverage. The call pattern analysis and network behavior extraction approaches are designed to perform big data analysis and feature extractions. Then, the corresponding algorithms are proposed to characterize network behaviors, i.e., cellular call patterns and network resource usage. The detailed evaluation is proposed to validate our method. For example, we find that some unpopular calls can last longer time and thus consume more network resources.**

***Index Terms*—Big data analysis, call detail records, cellular networks, network behaviors, usage patterns.**

Manuscript received May 20, 2019; revised July 3, 2019; accepted July 13, 2019. Date of publication July 22, 2019; date of current version January 14, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61571104, in part by the Sichuan Science and Technology Program under Grant 2018JY0539, in part by the Key projects of the Sichuan Provincial Education Department under Grant 18ZA0219, in part by the Shandong Provincial Natural Science Foundation under Grant ZR2017QF015, in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2017KYQD170, and in part by the Innovation Funding under Grant 2018510007000134. Paper no. TII-19-1939. *(Corresponding author: Zhihan Lv.)*

D. Jiang, Y. Wang, and S. Qi are with the School of Astronautics and Aeronautic, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: jiangdd@uestc.edu.cn; yuqingwang1998@qq.com; 2504864728@qq.com).

Z. Lv is with the School of Data Science and Software Engineering, Qingdao University, Qingdao 266071, China (e-mail: lvzhihan@gmail.com).

S. Singh is with the Department of Computer Engineering, National Institute of Technology Kurukshetra, Haryana 136119, India (e-mail: surjitmehla@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2019.2930226

I. Introduction

With the advances of communication technologies and fifth-generation networks, it is significantly important to perform the big data analysis and feature extraction on network behaviors in cellular networks for Industry 4.0 applications from a big data perspective. Wide adoption of smart phones and other mobile devices has led to rapid growth in mobile traffic and places a huge demand on cellular network infrastructure, such as resources on cell towers,[1] radio network controllers, and so forth. Gaining a deeper understanding of cellular usage patterns and how they are affected by user behaviors and mobility is critical to effective management of cellular network resources and to meet user quality of experience expectation. Originally designed for billing purpose, the call detail records (CDRs) collected by cellular network operators provide a useful and rich data source for obtaining insights into network usage patterns and user behaviors. Since CDRs are collected at either an initiating cell tower or a terminating cell tower or both (when both caller and callee belong to the same cellular service provider), they allow for more detailed studies of cellular network usage patterns at the (finer-grained) *cell tower* level. Also, since CDRs are less voluminous and are typically stored by cellular network operators for longer periods of time, using CDRs one can also conduct studies of cellular usage patterns over a longer time horizon.

Traffic data measured [1], [2] are extensively used to study and characterize cellular networks in recent several years. These studies mainly include user mobility [3]–[5], user behaviors [6], [7], network traffic [8], [9], usage patterns [10]–[12], and network utility [13], [14]. In the last few years, both CDRs and Internet traffic data are often exploited to characterize usage patterns [15] and network resource utility in cellular networks. The big data technologies [16]–[20] were used to analyze mobile networks for finding popular service [21], studying mobile computing [22] and dynamic data analysis [23], [24], and analyzing big traffic data [25]. We use the larger CDRs dataset with longer time (five months), more users (five million), more records (several hundred million lines), and more global data

[1]In this article, we use the term "cell towers" loosely to refer to radio base stations in a radio access network of a cellular network infrastructure, although we know in reality the radio antenna (typically sitting on a "tower") may not be colocated with the actual base station that is attached to, which is the "active" entity that processes user "calls" – including voice, SMS, and data.

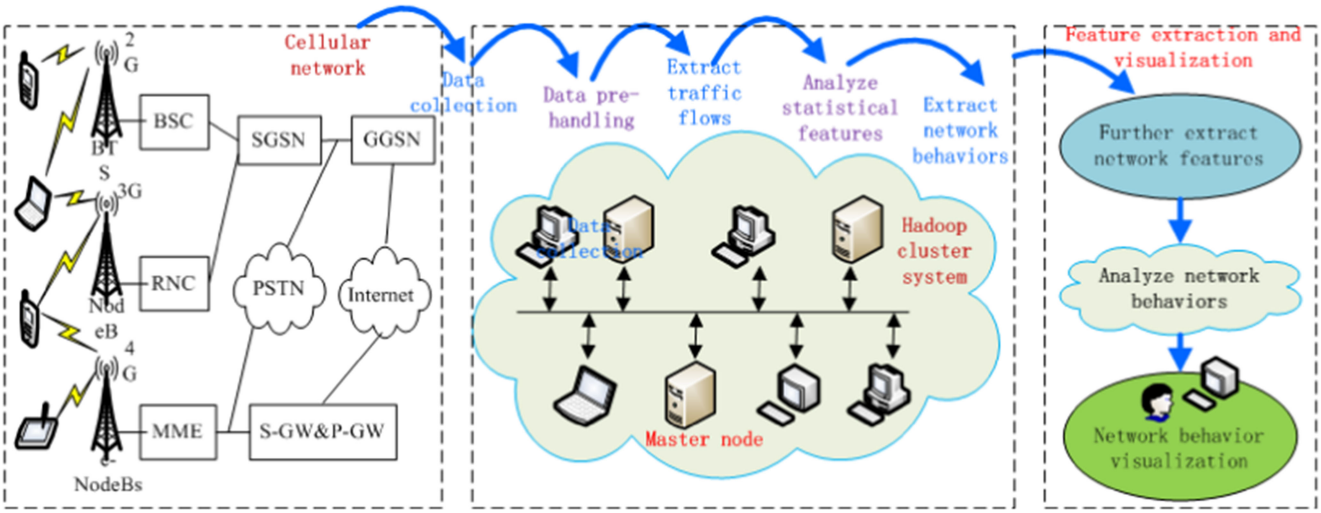


Fig. 1. Overview of cellular network behavior analysis framework with big data analysis technologies, including HDFS, MapReduce, HBase, and Hive.

(nationwide) to characterize network behaviors and network resource usage. Different form these methods, we used big data analysis technologies to extract network behavior features from big CDRs data. Our method can obtain networkwide call behavior patterns and network resource usage via CDRs data.

This article presents some advanced data modeling analysis methods about network behaviors in cellular networks for Industry 4.0 from a big data perspective. Although a number of studies have utilized CDRs and traffic data to study a variety of problems related to cellular networks, it is significantly difficult to attain the accurate patterns of network behaviors from a network big data in a cellular network. We present a cellular CDRs big data-driven analysis framework and method to capture network behaviors in mobile cellular networks via the national tower-level granularity, with several thousand hours and several million lines of records. This article mainly focuses on network behaviors and activities in mobile cellular networks from a nationwide tower-level perspective. We use big data technologies to revisit them. And we dissect network behavior and activity patterns, distributions, and predictability. We construct tower-level (networkwide) and time-varying traffic matrices at the cell tower level over time. From a *networkwide* perspective, we leverage tower-level traffic matrices and matrix analysis theory to analyze usage patterns of individual cell towers and tower-pairs, including distributions, temporal patterns, and geographic popularity of calls. Furthermore, we also study cell towers' activities and their relations to network call behaviors in cellular networks. We propose a new metric, Duration of Each Call (DEC), to more accurately describe network utility. Several algorithms are presented to dig out and extract network behaviors from the CDRs big data. We perform the detailed evaluation to validate our method.

II. Analysis Methodology

In this section, we propose a cellular network behavior analysis framework with big data analytics. Then, the feature extraction methods and algorithms for network behaviors in cellular networks are presented to characterize and describe complex network behaviors.

A. Analysis Framework

We exploit big data analytics and architecture to construct the analysis framework for cellular network behaviors. The overview of the analysis framework is shown in Fig. 1, which includes: Collection of cellular network big data, big data analysis and handling, and feature extraction and visualization. The collection module collects the cellular network big data, including saving the collected data to the database. The big data analysis and handling module prehandles the network big data, extracts traffic flows from network big data, analyzes statistical features of network big data according to traffic flows, and extracts network behavior patterns from statistical features, based on Hadoop cluster system. The feature extraction and visualization module uses the feature extraction algorithms to further refine network features, analyze and dig network behaviors, and visualize network behavior.

In big data analysis, the MapReduce process performs the computation and analysis on $< key, value >$ pairs. Our problem is to attain networkwide behaviors from cellular network big data. Thereby, the key is constructed through origin and destination cell tower IDs, while the value is the other fields of CDRs, such as call number, call duration, starting time, ending time, and so on. In such a case, we can implement the following MapReduce process:

$$
\begin{aligned}
&'input\ data' \Rightarrow < k_1, v_1 > \Rightarrow Map \Rightarrow < k_2, v_2 > \\
&\Rightarrow pre-handling\ algorithm \Rightarrow < k_3, v_3 > \\
&\Rightarrow Reduce \Rightarrow < k_4, v_4 > \\
&\Rightarrow statistical\ analysis\ algorithm \Rightarrow < k_5, v_5 > \\
&\Rightarrow feature\ extract\ algorithm \Rightarrow < k_6, v_6 > \\
&\Rightarrow' output\ result'
\end{aligned}
$$

According to the abovementioned process, Master submits MR jobs by Hive to perform data analysis.

B. Data Prehandling and Traffic Flow Extraction

Data collections in cellular networks provide us with a record line of CDRs—including timestamp, origin cell tower ID, destination cell tower ID, the number of calls, and the number of durations. However, these data not only have huge volume, but also can hold singular or abnormal data points. We need first to clean these record data. To the end, if any field in each record line is missing, delete the record line. If there exist the singular value in each record line, such as negative or the wrong origin (or destination) cell tower ID not being in the given numbers, delete the record line. If the duration field in each record line is singular, such as larger than 24 h, delete the record line. Finally, we delete cell tower IDs in the collected dataset, which do not generate any call during the collection duration.

Through such a handling process, we can obtain a clean dataset of CDRs. However, it is difficult to directly handle these records due to the huge data volume generate by the data collection process. In this article, we exploit the matrix perspective to describe and characterize these data. To attain the global points of view for call behaviors in cellular networks, call traffic matrices between cell towers are constructed correctly according to these CDRs to extract global traffic flows.

To conveniently formulate, we list the notations and variables used in this article as follows: $paircall$, $incall$, and $outcall$ are the calls between cell towers, flowing, and departing from a cell tower, respectively. $allcall$ represents the total calls in a cell tower, including $incall$ and $outcall$. $selfcall$ denotes the calls flowing into and out from same cell towers. $pairdura$, $indura$, $outdura$, $alldura$, and $selfdura$ hold the similar meaning for the duration. $pairdec$, $indec$, $outdec$, $alldec$, and $selfdec$ have the same meaning for the DEC. i and j denote the cell tower ID, t is the time stamp, n stands for the number of cell towers, and s is the number of time slots observed. In addition, $A_{\text{call}} = \{a_{\text{call}}(i,j,t)\}_{n\times n\times s}$ denotes the matrix about $paircall$ and the calls from cell tower i to cell tower j; $D_{\text{call}} = \{d_{\text{call}}(j,t)\}_{n\times s}$ represents the matrix about $incall$ and the calls flowing into all destination cell towers, where $d_{\text{call}}(j,t) = \sum_{i=1}^{n} a_{\text{call}}(i,j,t)$; $O_{\text{call}} = \{o_{\text{call}}(i,t)\}_{n\times s}$ denotes the matrix about $outcall$ and the calls generated by all origin cell towers where $o_{\text{call}}(i,t) = \sum_{j=1}^{n} a_{\text{call}}(i,j,t)$; $R_{\text{call}} = \{r_{\text{call}}(i,t)\}_{n\times s}$ is the matrix about $allcall$ and the total calls flowing into and out from cell towers; $S_{\text{call}} = \{s_{\text{call}}(i,t)\}_{n\times s} = \{a_{\text{call}}(i,i,t)\}_{n\times s}$ is the matrix about $selfcall$ and the calls flowing into and out from same cell towers. $A_{\text{dura}} = \{a_{\text{dura}}(i,j,t)\}_{n\times n\times s}$, $D_{\text{dura}} = \{d_{\text{dura}}(j,t)\}_{n\times s}$, $O_{\text{dura}} = \{o_{\text{dura}}(i,t)\}_{n\times s}$, $R_{\text{dura}} = \{r_{\text{dura}}(i,t)\}_{n\times s}$, and $S_{\text{dura}} = \{s_{\text{dura}}(i,t)\}_{n\times s} = \{a_{\text{dura}}(i,i,t)\}_{n\times s}$ denote the same meaning for $pairdura$, $indura$, $outdura$, $alldura$, and $selfdura$. $A_{\text{dec}} = \{a_{\text{dec}}(i,j,t)\}_{n\times n\times s}$ denotes the matrix about $pairdec$ and the DEC of the calls from cell tower i to cell tower j, where $a_{\text{dec}}(i,j,t) = a_{\text{dura}}(i,j,t)/a_{\text{call}}(i,j,t)$ and if $a_{\text{call}}(i,j,t) = 0$, $a_{\text{dec}}(i,j,t) = 0$. $D_{\text{dec}} = \{d_{\text{dec}}(j,t)\}_{n\times s}$, $O_{\text{dec}} = \{o_{\text{dec}}(i,t)\}_{n\times s}$, $R_{\text{dec}} = \{r_{\text{dec}}(i,t)\}_{n\times s}$, and $S_{\text{dec}} = \{s_{\text{dec}}(i,t)\}_{n\times s}$ have the same meaning for $indec$, $outdec$, $alldec$, and $selfdec$, respectively. Moreover, $incall$ and $outcall$ are aggregated by all $paircall$ flowing into and departing from a cell tower, respectively, while $allcall$ is aggregated by all $paircall$ generating at or flowing into a cell tower. Likewise, $indura$, $outdura$, and $alldura$ are aggregated by $pairdura$.

The tower-pair call traffic matrix at time t is built as follows:

$$A_{\text{call}}(t) = \begin{bmatrix} a_{\text{call}}(1,1,t), a_{\text{call}}(1,2,t), \ \dots, a_{\text{call}}(1,n,t) \\ a_{\text{call}}(2,1,t), a_{\text{call}}(2,2,t), \ \dots, a_{\text{call}}(2,n,t) \\ \dots \\ a_{\text{call}}(n,1,t), a_{\text{call}}(n,2,t), \ \dots, a_{\text{call}}(n,n,t) \end{bmatrix} \quad (1)$$

where the row and column of $A_{\text{call}}(t)$ denote the origin and destination cell towers, n represents the number of effective cell towers; $a_{\text{call}}(i,j,t)$ is the call number from cell tower i to cell tower j at time t. If there does not exist the calls from cell tower i to cell tower j at time t, then $a_{\text{call}}(i,j,t) = 0$.

We build the tower-pair call traffic matrix of $paircall$ as follows:

$$A_{\text{call}} = \{a_{\text{call}}(i,j,t)\}_{n\times n\times s} \quad (2)$$

where A_{call} holds the hour granularity of time. Note that here the unit of time t is hour.

Likewise, we construct the tower-pair duration matrix of $pairdura$ as A_{dura} as follows:

$$A_{\text{dura}} = \{a_{\text{dura}}(i,j,t)\}_{n\times n\times s} \quad (3)$$

where $a_{\text{dura}}(i,j,t)$ is the duration of the call from cell tower i to cell tower j at time t. If there does not exist the calls from cell tower i to cell tower j at time t, then $a_{\text{dura}}(i,j,t) = 0$.

Matrix A_{call} describes the number of calls of all tower-pairs at any time from the networkwide perspective for the whole cellular network, while A_{dura} states the DEC between tower pairs. They can effectively indicate the behaviors of tower-pairs' calls. We propose a metric, DEC, to measure the DEC on average in 1 h for network resources, such as radio channels. DEC enables us to characterize usage efficiency of each call for network resources. In contrast to durations, it more precisely gives the DEC in each hour on average. Likewise, we construct tower-pair DEC matrix $pairdec$ as A_{dec}. Matrices A_{call}, A_{dura}, and A_{dec} describe the behaviors of tower-pairs. To dig deeper, we aggregate them at cell towers to further construct matrices D_{call}, O_{call}, and R_{call} about $incall$, $outcall$, and $allcall$.

Similarly, for durations, we obtain matrices D_{dura}, O_{dura}, and R_{dura} about $indura$, $outdura$, and $alldura$, and attain matrices D_{dec}, O_{dec}, and R_{dec} about $indec$, $outdec$, and $alldec$, respectively. Likewise, to characterize the behaviors of calls originating and terminating at same cell towers, we construct matrices S_{call}, S_{dura}, and S_{dec} about $selfcall$, $selfdura$, and $selfdec$, respectively.

These matrices describe the data information about tower-pairs and cell towers. They can be regarded as a stochastic process over time. The detailed steps of traffic flow extractions are shown in Algorithm 1. What's following, based on these matrices constructed, we use matrix analysis theory to propose the statistical analysis and feature extraction methods to characterize cellular network behaviors from a global perspective.

Algorithm 1: Traffic Flow Extractions.

Input: CDRs file f_{cdr}, cellular network topology file f_{topo}, cell tower geography coordinate file f_{coor};
Output: The matrices about $paircall$, $incall$, $outcall$, $allcall$, $pairdura$, $indura$, $outdura$, $alldura$, $pairdec$, $indec$, $outdec$, $alldec$, $selfcall$, $selfdura$, andselfdec: A_{call}, O_{call}, D_{call}, R_{call}, A_{dura}, O_{dura}, D_{dura}, R_{dura}, A_{dec}, D_{dec}, O_{dec}, R_{dec}, S_{call}, S_{dura}, and S_{dec}, respectively; files $\hat{f}_{\text{topo}}$ and $\hat{f}_{\text{coor}}$; available cell tower ID array a_{twr}; new topology file $\hat{f}_{\text{topo}}$ and coordinate file $\hat{f}_{\text{coor}}$;
Procedure:
By CDRs file f_{cdr}, perform data pre-handling and obtain new CDRs file $\hat{c}_{\text{cdr}}$;
Find cell tower IDs with calls from file $\hat{f}_{\text{cdr}}$ to build the available cell tower ID array a_{twr};
By array a_{twr}, delete from files f_{topo} and f_{coor} the cell tower IDs which does not belong to array a_{twr}, and attain new topology file $\hat{f}_{\text{topo}}$ and coordinate file $\hat{f}_{\text{coor}}$;
Open and read file $\hat{f}_{\text{cdr}}$;
While (not be last line)
Extract the call and duration of each record line into A_{call} and A_{dura} according to (1)–(3);
End
$A_{\text{dec}} = \{a_{\text{dec}}(i,j,t)\}_{n\times n\times s}$ where $a_{\text{dec}}(i,j,t) = a_{\text{dura}}(i,j,t)/a_{\text{call}}(i,j,t)$ and if $a_{\text{call}}(i,j,t) = 0$, $a_{\text{dec}}(i,j,t) = 0$;
According to A_{call}, A_{dura}, and A_{dec}, obtain O_{call}, D_{call}, R_{call}, O_{dura}, D_{dura}, R_{dura}, D_{dec}, O_{dec}, R_{dec}, S_{call}, S_{dura}, and S_{dec};
Save the results to the database;
Exit the procedure;

C. Call Pattern Analysis

Now, we perform the statistical analysis for traffic flow matrices in cellular networks to characterize the distributions, temporal patterns, geographic popularity, and cell tower activities of calls, as well as relations of calls, and cell tower activities. First, we calculate the cumulative distribution function (CDF) of the calls in CDRs, where the calls of each tower-pair and each cell tower are aggregated over space.

We construct a set as follows:

$$X_{\text{call}} = \{x_k | x_k = \hat{a}_{\text{call}}(i,j), k = i + (j-1)\times n \quad i,j = 1,2,...,n\} \tag{4}$$

where $\hat{a}_{\text{call}}$ denotes the aggregated $paircalls$ over space. Due to the randomicity of the calls, X_{call} can be regarded as a random variable. The distribution of X_{call} can be denoted as follows:

$$F_{\text{call}}(x) = P(X_{\text{call}} \le x). \tag{5}$$

$F_{\text{call}}(x)$ denotes the distribution of the tower-pair calls $paircall$ aggregated over space, which describes the distribution of the calls between cell towers.

Similarly, for $incall$, $outcall$, and $allcall$, we hold

$$\begin{cases} Y_{\text{call}} = \{y_i | y_i = \hat{d}_{\text{call}}(i), i = 1,2,...,n\} \\ Z_{\text{call}} = \{z_i | z_i = \hat{o}_{\text{call}}(i), i = 1,2,...,n\} \\ V_{\text{call}} = \{v_i | v_i = \hat{r}_{\text{call}}(i), i = 1,2,...,n\} \end{cases} \tag{6}$$

where $\hat{d}_{\text{call}}(i)$, $\hat{o}_{\text{call}}(i)$, and $\hat{r}_{\text{call}}(i)$ represent the aggregated calls of cell towers over space. Y_{call}, Z_{call}, and V_{call} are the random variables. Their distributions can be expressed as follows:

$$\begin{cases} F_{\text{call}}(y) = P(Y_{\text{call}} \le y) \\ F_{\text{call}}(z) = P(Z_{\text{call}} \le z) \\ F_{\text{call}}(v) = P(V_{\text{call}} \le v) \end{cases} \tag{7}$$

$F_{\text{call}}(y)$, $F_{\text{call}}(z)$, and $F_{\text{call}}(v)$ denote the distributions of cell tower calls aggregated over space, which describes the distribution of the calls of cell towers.

To further characterize the distributions of calls, we capture the relations of tower-pairs and cell towers as well as their calls. To the end, attain the following equations:

$$\begin{cases} R_{\text{pair}} = [1,2,...,n^2]/n^2 \\ R_{\text{tower}} = [1,2,...,n]/n \end{cases}. \tag{8}$$

The distributions of $\tilde{A}_{\text{call}}$, $\tilde{D}_{\text{call}}$, $\tilde{O}_{\text{call}}$, and $\tilde{R}_{\text{call}}$ are as follows:

$$\begin{cases} F_{\text{call}}(x) &= P(\tilde{A}_{\text{call}} \le x | R_{\text{pair}}) \\ &= \{csum(\tilde{A}_{\text{call}}) | R_{\text{pair}}\} \\ F_{\text{incall}}(y) &= P(\tilde{D}_{\text{call}} \le y | R_{\text{tower}}) \\ &= \{csum(\tilde{D}_{\text{call}}) | R_{\text{tower}}\} \\ F_{\text{outcall}}(z) &= P(\tilde{O}_{\text{call}} \le z | R_{\text{tower}}) \\ &= \{csum(\tilde{O}_{\text{call}}) | R_{\text{tower}}\} \\ F_{\text{allcall}}(v) &= P(\tilde{R}_{\text{call}} \le v | R_{\text{tower}}) \\ &= \{csum(\tilde{R}_{\text{call}}) | R_{\text{tower}}\} \end{cases} \tag{9}$$

where $\tilde{A}_{\text{call}}$, $\tilde{D}_{\text{call}}$, $\tilde{O}_{\text{call}}$, and $\tilde{R}_{\text{call}}$ are the ratios of $paircalls$, $incalls$, $outcalls$, and $allcalls$, respectively, to total network calls, from biggest to smallest, and $csum(.)$ denotes the cumulative sum. Equation (9) captures the relations of tower-pairs and cell towers as well as their normalized calls.

Next, we discuss the temporal characteristics of calls in the cellular network from the tower-pair and cell tower perspective. We study the temporal properties of calls, including the total calls of tower-pairs and cell towers over time, the calls of top tower-pairs and cell towers, and the calls of a typical tower-pair and cell tower. To the end, we use the statistical analysis theory to calculate their rankings and medians. Additionally, the auto correlation function (ACF) about calls is computed to further

capture their temporal patterns

$$\begin{cases} \eta_{\text{call}}(k) = E((\bar{a}_{\text{call}}(t) - E(\tilde{a}_{\text{call}}(t)))(\bar{a}_{\text{call}}(t+k) \\ \quad -E(\tilde{a}_{\text{call}}(t+k))))/E(\bar{A}_{\text{call}}) \\ \eta_{\text{incall}}(k) = E((\bar{d}_{\text{call}}(t) - E(\tilde{d}_{\text{call}}(t)))(\bar{d}_{\text{call}}(t+k) \\ \quad (k) - E(\tilde{d}_{\text{call}}(t+k))))/E(\bar{D}_{\text{call}}) \\ \eta_{\text{outcall}}(k) = E((\bar{o}_{\text{call}}(t) - E(\tilde{o}_{\text{call}}(t)))(\bar{o}_{\text{call}}(t+k) \\ \quad -E(\tilde{o}_{\text{call}}(t+k))))/E(\bar{O}_{\text{call}}) \\ \eta_{\text{allcall}}(k) = E((\bar{r}_{\text{call}}(t) - E(\tilde{r}_{\text{call}}(t)))(\bar{r}_{\text{call}}(t+k) \\ \quad -E(\tilde{r}_{\text{call}}(t+k))))/E(\bar{R}_{\text{call}}) \end{cases} \tag{10}$$

where $\eta_{\text{call}}(k)$, $\eta_{\text{incall}}(k)$, $\eta_{\text{outcall}}(k)$, and $\eta_{\text{allcall}}(k)$, respectively, denote the ACFs about $paircall$, $incall$, $outcall$, and $allcall$. Therefore, (10) can be used to characterize the typical distribution and period features over time of call patterns in cellular networks.

To understand the geographic correlations, we aggregate all the calls of tower-pairs over time and perform the normalization with the maximum calls aggregated. According to Section II-B, the rows and columns of the call traffic matrix of tower-pairs describe the origin and destination cell towers deployed at the different locations, respectively. It indicates the geographic relations of calls at the different moments. The diagonal of the call traffic matrix of tower-pairs at any time t denotes the usage of calls originating and terminating at same cell towers. Hence, the following equation is obtained:

$$\vec{A}_{\text{call}} = \hat{A}_{\text{call}}/c_{\text{total}} = \{\hat{a}_{\text{call}}(i,j)/c_{\text{total}}\}_{n\times n} \tag{11}$$

where $c_{\text{total}} = \sum_{i=1}^{n}\sum_{j=1}^{n}\hat{a}_{\text{call}}(i,j)$ stands for the total network calls. Equation (11) states the geographic distribution and popularity of calls generating by different geographic distribution cell towers. To further analyze the geographic popularity of calls, we calculate the total calls between cell towers not considering the direction of calls.

Then, we calculate the ratio of the total calls for a given distance to the total calls of networks to find the relations between calls and distance. The below distribution is attained as follows:

$$\begin{cases} F_{\text{dist}}(d) &= P(\tilde{D}_{\text{tower}} \le d|\tilde{D}_{\text{tower}}) \\ &= \{csum(\tilde{D}_{\text{tower}})|\tilde{D}_{\text{tower}})\} \\ F_{\text{pcall}}(x) &= P(\tilde{P}_{\text{call}} \le x|\tilde{D}_{\text{tower}}) \\ &= \{csum(\tilde{P}_{\text{call}})|\tilde{D}_{\text{tower}})\} \end{cases} \tag{12}$$

where $F_{\text{dist}}(d)$ and $F_{\text{pcall}}(x)$, respectively, denote the distribution of the distance and calls between cell towers.

To dig deeper, we analyze the locality behaviors of calls originating and terminating at same cell towers. Although the aggregated information can exhibit the locality of tower pairs' calls, it cannot clearly illustrate the locality characteristics of different cell towers generating the calls at the different moments. To this end, the ratio of the total calls originating and terminating at same cell towers to the total calls of networks at the different times is computed.

To analyze the fluctuation patterns, we further calculate the distribution of the ratio, namely

$$F_{\text{rself}}(x) = P(R_{\text{selfcall}} \le x). \tag{13}$$

To characterize the locality fluctuation, we also discuss the locality properties of cell towers' calls during 24 h (namely each day). Therefore, the $paircalls$ and $selfcalls$ are aggregated in the 24-h interval, respectively, and their ratio is calculated to capture these features.

Now, we characterize the activity patterns of cell towers by calculating the distributions of the hours for generating calls. According to the statistical theory, the corresponding medians are computed to characterize the activities of the typical cell tower. Their distributions are denoted as follows:

$$\begin{cases} F_{\overleftarrow{A}_{\text{call}}}(x) = P(\overleftarrow{A}_{\text{call}} \le x) \\ F_{\overleftarrow{D}_{\text{call}}}(x) = P(\overleftarrow{D}_{\text{call}} \le x) \\ F_{\overleftarrow{O}_{\text{call}}}(x) = P(\overleftarrow{O}_{\text{call}} \le x) \\ f_{\overleftarrow{R}_{\text{call}}}(x) = P(\overleftarrow{R}_{\text{call}} \le x) \end{cases}. \tag{14}$$

Likewise, the day activity distributions of calls can be denoted as follows:

$$\begin{cases} F_{\hat{A}_{\text{day}}}(x) = P(\hat{A}_{\text{day}} \le x) \\ F_{\hat{D}_{\text{day}}}(x) = P(\hat{D}_{\text{day}} \le x) \\ F_{\hat{O}_{\text{day}}}(x) = P(\hat{O}_{\text{day}} \le x) \\ F_{\hat{R}_{\text{day}}}(x) = P(\hat{R}_{\text{day}} \le x) \end{cases} \tag{15}$$

where $\hat{A}_{\text{day}}$, $\hat{D}_{\text{day}}$, $\hat{O}_{\text{day}}$, and $\hat{R}_{\text{day}}$ denote $paircalls$, $incalls$, $outcalls$, and $allcalls$ per day, respectively.

To better understand the activity characteristics, we analyze the successive activity patterns of cell towers in successive hours of generating calls. Likewise, we also analyze the activity behaviors of cell towers in the day granularity. Similar to (15), we obtain the distributions of successive hour and day activity patterns as shown in (16) and (17)

$$\begin{cases} F_{\overleftarrow{A}_{\text{scall}}}(x) = P(\overleftarrow{A}_{\text{scall}} \le x) \\ F_{\overleftarrow{D}_{\text{scall}}}(x) = P(\overleftarrow{D}_{\text{scall}} \le x) \\ F_{\overleftarrow{O}_{\text{scall}}}(x) = P(\overleftarrow{O}_{\text{scall}} \le x) \\ f_{\overleftarrow{R}_{\text{scall}}}(x) = P(\overleftarrow{R}_{\text{scall}} \le x) \end{cases} \tag{16}$$

$$\begin{cases} F_{\hat{A}_{\text{sday}}}(x) = P(\hat{A}_{\text{sday}} \le x) \\ F_{\hat{D}_{\text{sday}}}(x) = P(\hat{D}_{\text{sday}} \le x) \\ F_{\hat{O}_{\text{sday}}}(x) = P(\hat{O}_{\text{sday}} \le x) \\ F_{\hat{R}_{\text{sday}}}(x) = P(\hat{R}_{\text{sday}} \le x) \end{cases}. \tag{17}$$

To further characterize the call patterns, we study the correlations between cell tower activities and corresponding calls that they generate. These corresponding calls are normalized by the total calls in the cellular network.

Algorithm 2: Statistical Analysis of Flow Features.

Input: The matrices about $paircall$, $incall$, $outcall$, $allcall$, $pairdura$, $indura$, $outdura$, $alldura$, and $selfcall$: A_{call}, D_{call}, O_{call}, R_{call}, A_{dura}, D_{dura}, O_{dura}, R_{dura}, and S_{call}, respectively; available cell tower ID array a_{twr}; new topology file $\hat{f}_{\text{topo}}$ and coordinate file $\hat{f}_{\text{coor}}$;
Output: Calls and durations's distributions $call_{\text{dist}}$, $dura_{\text{dist}}$; temporal patterns $temp_{\text{paircall}}$'s, $temp_{\text{allcall}}$; geographic popularity $call_{\text{geo}}$, $call_{\text{dist}}$, $call_{\text{local}}$; cell tower activity patterns $acti_{\text{hour}}$, $acti_{\text{day}}$; relations of cell tower activities and calls $c2a_{\text{hour}}$, $c2a_{\text{day}}$, $dist_{\text{hour}}$;
Procedure:
Use A_{call}, D_{call}, O_{call}, R_{call}, A_{dura}, D_{dura}, O_{dura}, and R_{dura} to calculate calls' and durations' CDF and correlations between cell towers according to (4)–(9), and save to $call_{\text{dist}}$ and $dura_{\text{dist}}$;
Exploit A_{call} and R_{call} to compute $paircall$'s and $allcall$'s aggregated, dominate, typical, and ACF features in terms of (10), and save to $temp_{\text{paircall}}$ and $temp_{\text{allcall}}$;
According to (11) and A_{call}, obtain total $paircall$'s geographic distribution $call_{\text{geo}}$;
Based on A_{call}, new topology file $\hat{f}_{\text{topo}}$ and coordinate file $\hat{f}_{\text{coor}}$, attain relations between calls and distances $call_{\text{dist}}$ according to (12);
Utilize A_{call} and S_{call} to calculate call locality $call_{\text{local}}$ according to (13);
Analyze A_{call}, D_{call}, O_{call}, and R_{call}, and then attain cell tower activity distributions $acti_{\text{hour}}$ and $acti_{\text{day}}$ according to (14)–(17);
Use A_{call}, D_{call}, O_{call}, R_{call} to compute relations of calls and the number of (successive) hours and days $c2a_{\text{hour}}$ and $c2a_{\text{day}}$, and obtain the distribution of the number of (successive) hours $dist_{\text{hour}}$ in terms of the same methods;
Save the results to the database;
Exit the procedure;

Next, we study the activity characteristics of the different categories of tower-pairs and cell towers. According to the call distributions, we classify tower-pairs into different categories in terms of the total calls of tower-pair itself. Similarly, we classify cell towers into different categories. Then, the corresponding distributions are calculated to capture the correlations. Due to limit of space, the detailed discussion and analysis are not provided here. We have derive our statistical analysis method for capturing flow features of network behaviors in cellular networks. The corresponding algorithm steps are listed in Algorithm 2.

D. Network Resource Usage Behavior Extraction

Here, we characterize network resource usage behaviors in cellular networks, including the duration distribution, usage diversity, usage dynamics, and usage popularity. The parameter duration in CDRs is very important, which describes the duration of calls in the observed process. Now, we compute the distributions of the durations of the calls in cellular networks. As discussed in (4)–(9), the duration distribution of calls can be expressed as follows:

$$\begin{cases} F_{\text{dura}}(x) = P(X_{\text{dura}} \le x) \\ F_{\text{indura}}(y) = P(Y_{\text{indura}} \le y) \\ F_{\text{outdura}}(z) = P(Z_{\text{outdura}} \le z) \\ F_{\text{alldura}}(v) = P(V_{\text{alldura}} \le v) \end{cases} \tag{18}$$

$$\begin{cases} F_{\text{dura}}(x) &= P(\tilde{A}_{\text{dura}} \le x | R_{\text{pair}}) \\ &= \{csum(\tilde{A}_{\text{dura}}) | R_{\text{pair}}\} \\ F_{\text{indura}}(y) &= P(\tilde{D}_{\text{dura}} \le y | R_{\text{tower}}) \\ &= \{csum(\tilde{D}_{\text{dura}}) | R_{\text{tower}}\} \\ F_{\text{outdura}}(z) &= P(\tilde{O}_{\text{dura}} \le z | R_{\text{tower}}) \\ &= \{csum(\tilde{O}_{\text{dura}}) | R_{\text{tower}}\} \\ F_{\text{alldura}}(v) &= P(\tilde{R}_{\text{dura}} \le v | R_{\text{tower}}) \\ &= \{csum(\tilde{R}_{\text{dura}}) | R_{\text{tower}}\} \end{cases}. \tag{19}$$

Next, we analyze the diversity of network resource usage in cellular networks using tower-pairs' calls. To characterize the usage diversity, we study the relations between the calls aggregated over time and the corresponding durations aggregated. It is very useful that each tower-pair uses the calls to predict their durations for network management. The entropy can be used to characterize the diversity of a variable. We use the temporal entropy of each tower-pair to describe the diversity of network resource usage. As in [7], we define our temporal entropy of tower-pair z as follows:

$$H(z) = -\frac{\sum_{t=1}^{n} \left(\frac{v_z(t)}{v_z} \times \log_2 \frac{v_z(t)}{v_z}\right)}{\log_2(n)} \tag{20}$$

where n denotes the total number of hours observed, $0 \le H(z) \le 1$, $v_z(\text{t})$ represents the calls or durations of tower-pair z at time t, and v_z is the total calls or durations of tower-pair z. $H(z)$ describes usage diversity of tower-pair z over time for network resources.

Now, we study the dynamics of network resource usage in cellular networks. $selfcalls$, $selfduras$, $selfdecs$, $paircalls$, $pairduras$, and $pairdecs$ are aggregated in the 24 h and are normalized by their corresponding maximums to characterize the distribution of network resource usage each day. The daily usage patterns of $selfcall$ and $pairall$ for network resources are obtained. And the daily usage patterns of $selfdura$, $selfdec$, $pairdura$, and $pairdec$ are obtained.

Now, we analyze the popularity of network resource usage. To this end, the calls of tower-pairs and cell towers are ranked from larger to smaller according to their total calls. The number of hours for calls to appear is analyzed. Then, we can obtain the relations of calls and the number of hours of calls to appear as

Algorithm 3: Resource Usage Behavior Extraction.

Input: The matrices about $pairdura$, $indura$, $outdura$, $alldura$, $selfcall$, $selfdura$, $selfdec$, $paircall$, and $pairdec$: A_{dura}, D_{dura}, O_{dura}, R_{dura}, S_{call}, S_{dura}, S_{dec}, A_{call}, and A_{dec}.
Output: Duration distribution $dura_{\text{dist}}$, usage diversity $usag_{\text{divers}}$, usage dynamics $usag_{\text{dynam}}$, usage popularity $usag_{\text{popu}}$.
Procedure:
According to A_{dura}, D_{dura}, O_{dura}, and R_{dura}, use Equations (18)–(19) to calculate durations' CDF and correlations between cell towers, and save to $dura_{\text{dist}}$;
Utilize S_{call}, S_{dura}, S_{dec}, A_{call}, A_dura, and A_{dec} to compute $selfcall$'s, $selfduras$', $selfdecs$', $paircalls$', $pairduras$', and $pairdecs$' aggregated features over day, and save to $usag_{\text{dynam}}$;
Exploit A_{dura}, D_{dura}, O_{dura}, and R_{dura} to compute usage popularity patterns $usag_{\text{popu}}$ in terms of Equations (21)–(22);
Exit the procedure;

follows:

$$\begin{cases} \vec{A}_{\text{hour}} = f_{\text{call}}(\overleftarrow{A}_{\text{call}}, \hat{\hat{A}}_{\text{hcall}}) \\ \vec{D}_{\text{hour}} = f_{\text{call}}(\overleftarrow{D}_{\text{call}}, \hat{\hat{D}}_{\text{hcall}}) \\ \vec{O}_{\text{hour}} = f_{\text{call}}(\overleftarrow{O}_{\text{call}}, \hat{\hat{O}}_{\text{hcall}}) \\ \vec{R}_{\text{hour}} = f_{\text{call}}(\overleftarrow{R}_{\text{call}}, \hat{\hat{R}}_{\text{hcall}}) \end{cases} \tag{21}$$

where $f_{\text{call}}(A, B)$ represents the correlation extraction between calls and network resource usage behaviors from A according to the ranked B.

To dig deeper, we study the relations of the popularity of calls and durations and DECs. Likewise, the popularity relations of calls and durations for tower-pairs and cell towers can be denoted as follows:

$$\begin{cases} \vec{A}_{\text{dura}} = f_{\text{call}}(A_{\text{call}}, \hat{\hat{A}}_{\text{call}}) \\ \vec{A}_{\text{dec}} = f_{\text{call}}(A_{\text{dec}}, \hat{\hat{A}}_{\text{call}}) \\ \vec{D}_{\text{dura}} = f_{\text{call}}(D_{\text{dura}}, \hat{\hat{R}}_{\text{call}}) \\ \vec{O}_{\text{dura}} = f_{\text{call}}(O_{\text{dura}}, \hat{\hat{R}}_{\text{call}}) \\ \vec{R}_{\text{dura}} = f_{\text{call}}(R_{\text{dura}}, \hat{\hat{R}}_{\text{call}}) \\ \vec{D}_{\text{dec}} = f_{\text{call}}(D_{\text{dec}}, \hat{\hat{R}}_{\text{call}}) \\ \vec{O}_{\text{dec}} = f_{\text{call}}(O_{\text{dec}}, \hat{\hat{R}}_{\text{call}}) \\ \vec{R}_{\text{dec}} = f_{\text{call}}(R_{\text{dec}}, \hat{\hat{R}}_{\text{call}}) \end{cases}. \tag{22}$$

We have derive our statistical analysis method for capturing flow features of network behaviors in cellular networks. The detailed steps of network behavior extraction algorithm are listed in Algorithm 3.

To visualize the analysis and extraction results, we use R and $Python$ to show the curves and figures. Fig. 2 plots the feature extraction process of cellular network behaviors using big data analytics, which includes three layers: collection layer, analysis layer, and application layer. The collection layer finishes the network big data capturing and collection through capturing and saving to the CDRs database. The analysis layer is in charge of big data handling and analysis, and outputs network behavior patterns. The application layer performs the data analysis and visualization.

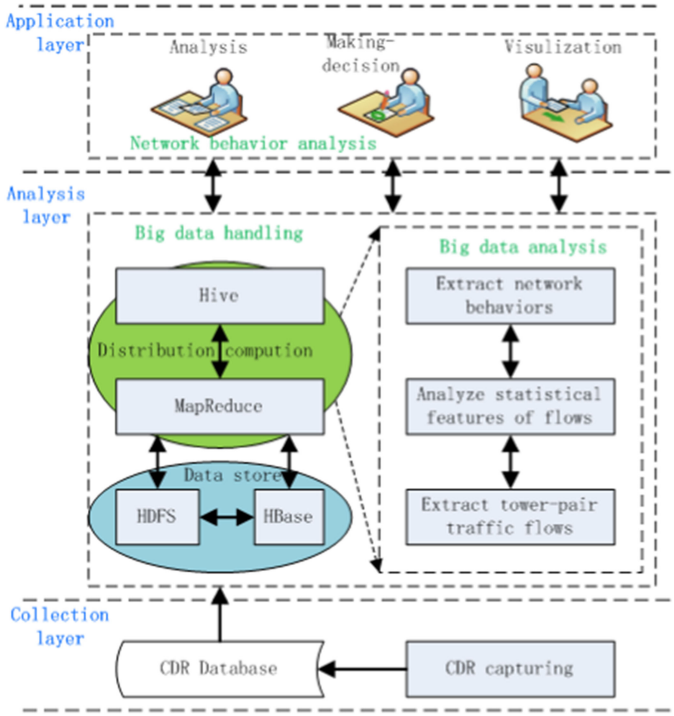


Fig. 2. Feature extraction process of cellular network behaviors using big data analytics, with three layers.

III. Evaluation

In this section, we conduct a series of tests to validate our method. The analyzed mobile call datasets are from the real national cell network, where there are over 1000 of cell towers and 140-day call data that collected with the hour granularity. The call dataset includes the CDRs of five million users. We use the whole dataset to analyze and study the behavior patterns of calls generated by all tower pairs and towers and network resource usage.

Fig. 3(a) denotes the distributions of total calls of each tower-pair and each cell tower, where the calls of each tower-pair and each cell tower are aggregated over space. Fig. 3(b) shows the relations of tower-pairs and cell towers as well as their calls normalized by total calls of networks, where "X" denotes tower-pairs or cell towers. Fig. 3 indicates that the distributions of calls of tower-pairs and cell towers are very different. Fig. 4(a) shows that network resources are not sufficiently used by tower pairs. Different from tower-pairs, cell towers' total $allcalls$ and the $allcalls$ of top three cell towers and a typical cell tower show the continuous characteristics in Fig. 4(b). The $outcalls$

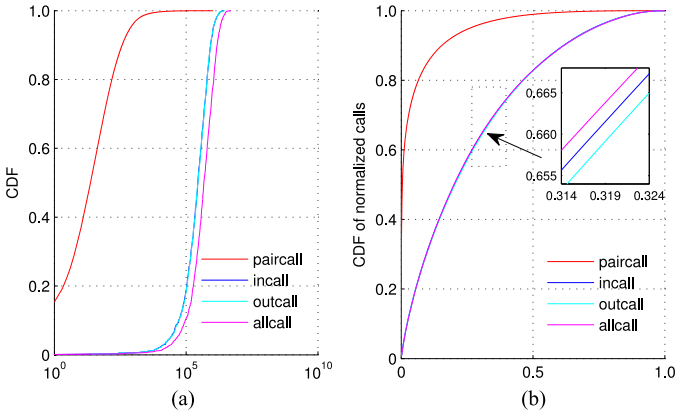


Fig. 3. Distributions of call popularity for tower-pairs and cell towers, with X denoting tower-pairs or cell towers. (a) Aggregated calls. (b) Percentage of X (ranked).

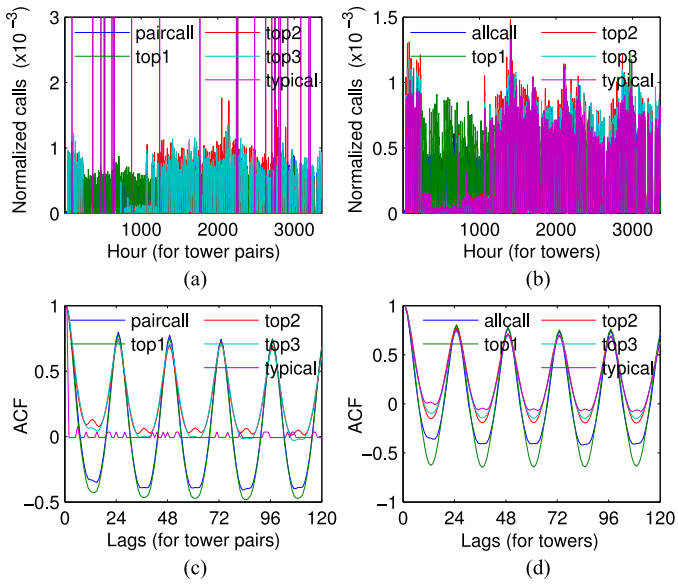


Fig. 4. Normalized total calls, three top calls, typical call and their ACF for tower-pairs and cell towers.

and $incalls$ hold the same temporal distributions but we do not plot them due to space limit. Fig. 4(c) and (d) shows that

1) the temporal patterns of tower-pairs' calls are very different from those of cell towers' calls;
2) the calls of most of tower-pairs are not predictable while those of most of cell towers embody the strong period behaviors.

Fig. 5(a) shows that the calls of tower-pairs holds the strong locality. Fig. 5(b) indicates that

1) the calls of towers themselves generating and flowing into themselves occupy about 38.5% of the total calls of networks;
2) the distance of 10% of tower-pairs is less than 30 km while they generate the 75% of the total calls, and 20% of tower-pairs with the distance less than 95 km construct about 85% of the total calls.

This further shows that the calls of tower-pairs hold the strong locality. Fig. 6(a) demonstrates that cell towers hold the very

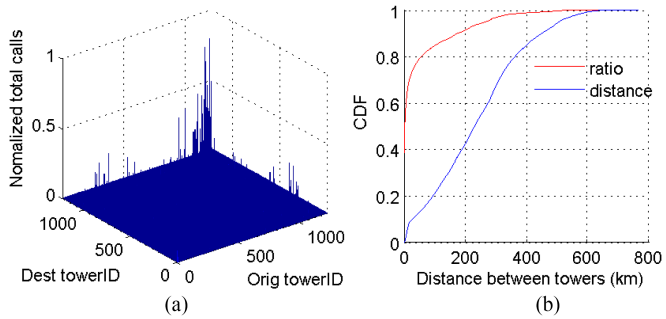


Fig. 5. (a) Geographic distribution of total calls for tower-pairs. (b) Relations of tower-pairs' calls and distance.

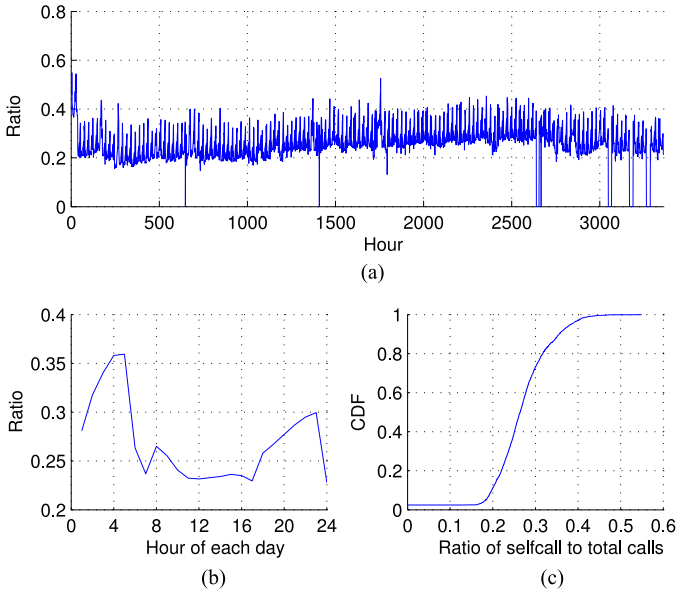


Fig. 6. Locality characteristics of tower-pairs' calls.

strong locality behavior at the different times and this locality also exhibits the stable fluctuation nature. Fig. 6(b) shows that the locality behaviors of cell towers are very different at the different hours each day. Fig. 6(c) indicates that the locality changes are fairly stable at most of moments.

Fig. 7(a) shows that most of cell towers keep active at most of hours. Fig. 7(b) indicates that most of tower-pairs do not generate calls in the successive hours while most of cell towers are more active but they do not create calls in the successive 24 h. Thus, there exist a large number of successive hours without calls. This successive noncall activity behaviors is helpful to let cell towers appropriately sleep for energy saving. Fig. 7(c) illustrates the activity behaviors of cell towers in the day granularity. Fig. 7(d) indicates that tower-pairs do not keep active while cell towers are very active.

The more the number of hours (days) for tower-pairs and cell towers to generate calls is, the larger the corresponding calls created are [see Fig. 8(a) and (c)]. For the number of successive hours (days) for generating calls, many peaks in the call traffic appear at the number of hours (days) [see Fig. 8(b) and (d)]. This indicates that for the peak points, cell towers are to create

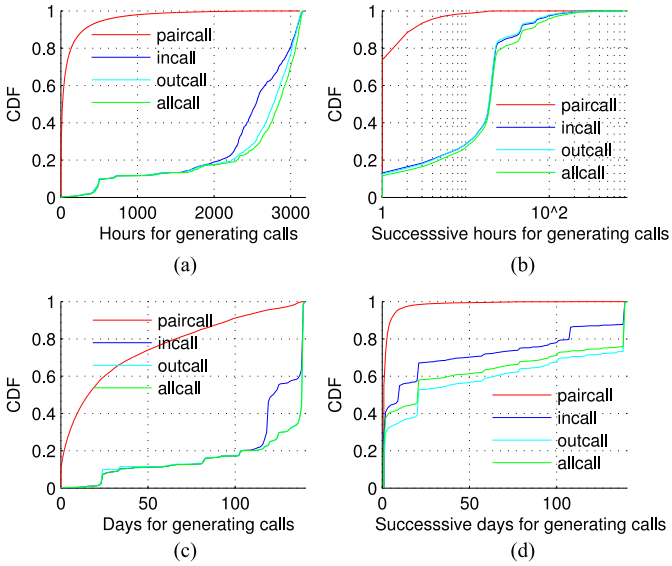


Fig. 7. Call activity distributions of cell towers.

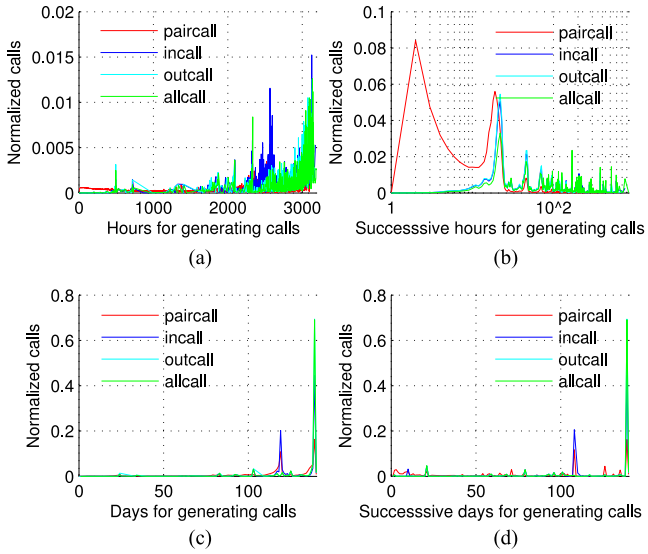


Fig. 8. Relations of calls and the number of (successive) hours (days) of calls to appear.

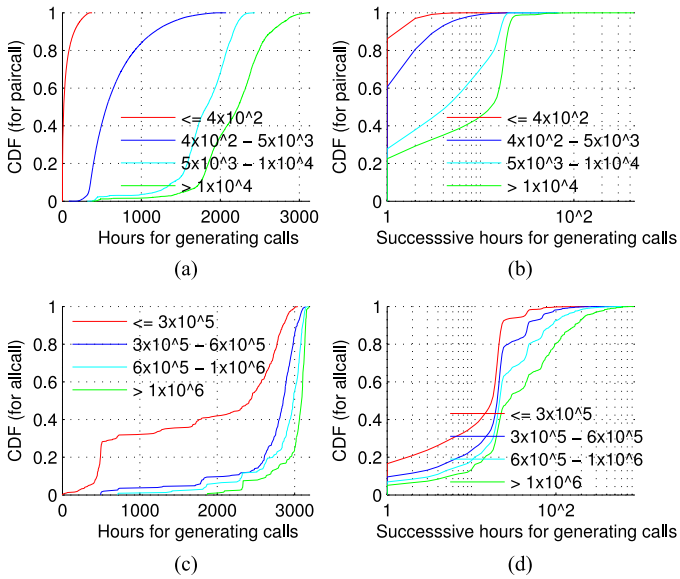


Fig. 9. Distributions of the number of (successive) hours (days) of calls to appear for different tower-pairs and cell towers.

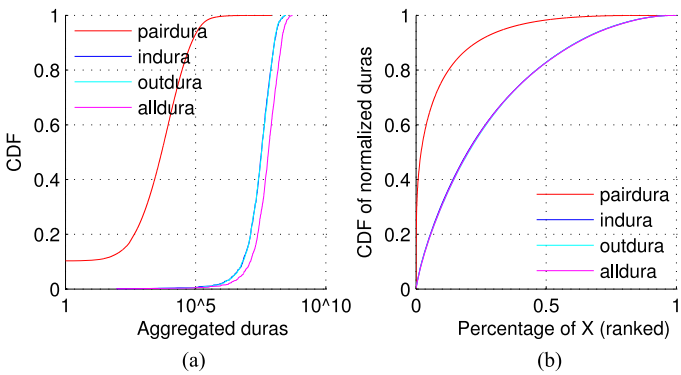


Fig. 10. Distributions of duration popularity for tower-pairs and cell towers, with *X* denoting tower-pairs or cell towers.

more calls and consume more network resources. Thus, when optimizing networks, these locations should be particularly taken into consideration. Fig. 9 shows that the more the calls of tower-pairs and cell towers are, the longer the (successive) time [the number of hours (days)] for generating calls is. This implies that the heavy tower-pairs and cell towers keep more frequent activities of calls.

Fig. 10(a) shows that cell towers keep idle for most of times and thus network resources are not used sufficiently. Fig. 10(b) indicates that the distribution of $pairduras$ is very skewed and the distributions of $induras$, $outduras$, and $allduras$ are consistent. Fig. 11(a) denotes that there exists a linear relation between the aggregated calls and durations. We can clearly find that the larger the calls and durations of each tower-pair aggregated over time is, the larger their temporal entropy in Fig. 11(b) and (c) (where x- and y-axes are log coordinate) is. More importantly, there exists a stronger power law relations between temporal entropy of calls and temporal entropy of durations in Fig. 11(d). Fig. 11 shows that we can use the temporal entropy of calls to predict those of durations. This predictability is very helpful to estimate the probable usage of network resources.

Fig. 12 shows that at night some users spend more times to talk by the call. This is probably because there is a attractive price for a call. These important distribution characteristics in Fig. 12 can be used to guide how to design the cellular network more effectively. Fig. 13 shows that tower-pairs and cell towers with more popular calls do not always spend more time to use network resources, that is, they are popular but are not necessary to be significantly active. Fig. 14 further indicates that tower-pairs and cell towers with unpopular calls can consume more network resources.

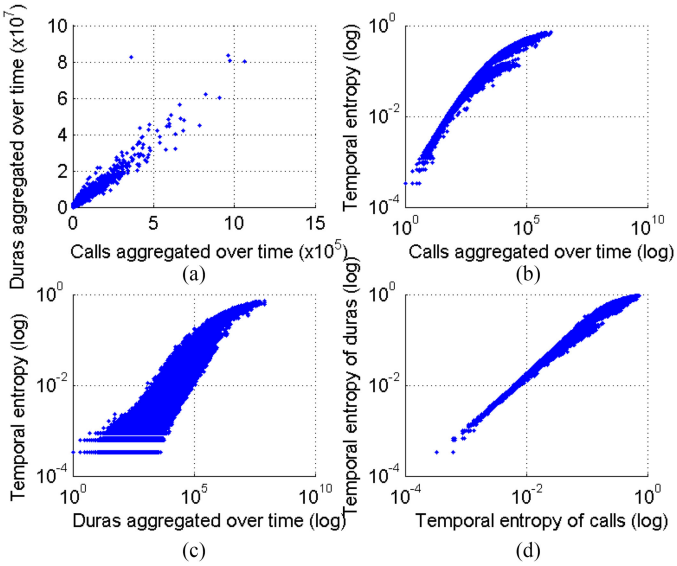


Fig. 11. Diversity of network resource usage, with duras denoting durations.

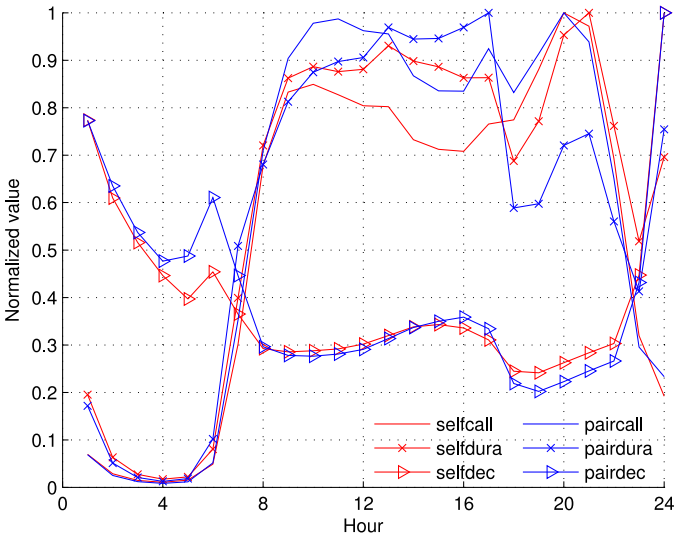


Fig. 12. Distributions of network resource usage each day.

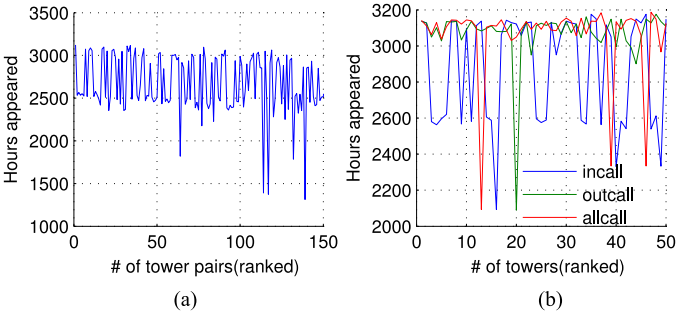


Fig. 13. Relations of the number of hours appeared and the corresponding calls of tower-pairs and cell towers.

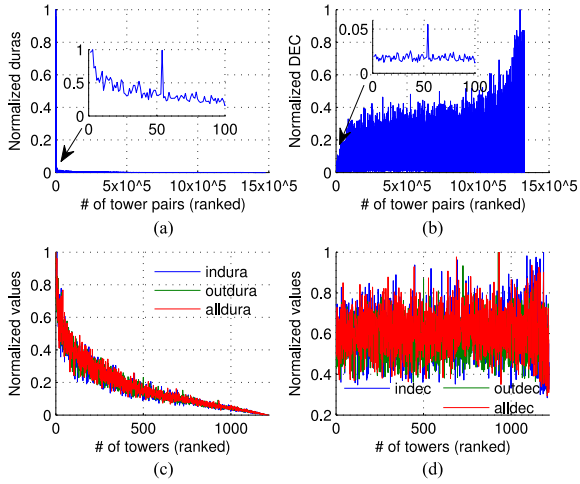


Fig. 14. (a) Relations of calls and durations for tower-pairs, with duras denoting durations. (b) Relations of calls and DECs for tower-pairs. (c) Relations of calls and durations for cell towers. (d) Relations of calls and DECs for cell towers.

IV. Conclusion

In this article, we proposed a big data-based analysis framework to analyze and extract network behavior patterns in cellular networks for Industry 4.0 applications from a big data perspective. The data prehandling and traffic flow extraction approaches were presented to build the networkwide big data. As a result, network behaviors could accurately characterize from a networkwide perspective. We also designed the call pattern analysis and network behavior extraction approaches to perform big data analysis and feature extractions. Then, the corresponding algorithms were proposed to characterize all kinds of network behaviors. We made a first step to network behavior patterns in the cellular network from a networkwide perspective using the longer, larger, and nationwide CDRs dataset, based on network big data analysis. The detailed evaluation was performed to validate our method. Using our method, we obtained some novel insights into call usage and network utility. Our future work is to use our proposed approach and these findings to explore network big data applications, such as effective network designs, energy savings, and optimization methods in the cellular network.

References

[1] M. S. Parwez, D. B. Rawat, and M. Garuba, "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network," *IEEE Trans. Ind. Inform.*, vol. 13, no. 4, pp. 2058–2065, Aug. 2017.

[2] 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.

[3] S. Jiang, J. Ferreira, and M. Gonzalez, "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 208–219, Jun. 2017.

[4] Y. Zhang, "User mobility from the view of cellular data networks," in *Proc. Infocom*, 2014, pp. 1348–1356.

[5] B. Hussain, Q. Du, and P. Ren, "Semi-supervised learning based big data-driven anomaly detection in mobile wireless networks," *China Commun.*, vol. 15, no. 4, pp. 41–57, 2018.

[6] S Qin *et al.*, "Applying big data analytics to monitor tourist flow for the scenic area operation management," *Discrete Dyn. Nature Soc.*, vol. 2019, no. 1, pp. 1–11, 2019.

[7] X. Wang *et al.*, "Travel distance characteristics analysis using call detail record data," in *Proc. 29th Chin. Control Decis. Conf.*, 2017, pp. 3485–3489.

[8] S. Bothe, H. N. Qureshi, and A. Imran, "Which statistical distribution best characterizes modern cellular traffic and what factors could predict its spatiotemporal variability?" *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 810–813, May 2019.
[9] N. Mizumura, K. Nkurikiyeyezu, H. Ishizuka, G. Lopez, and Y. Tobe, "Smartphone application usage prediction using cellular network traffic," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2018, pp. 753–758.
[10] Z. Aziz and R. Bestak, "Analysis of call detail records of international voice traffic in mobile networks," in *Proc. Int. Conf. Ubiquitous Future Netw.*, 2018, pp. 475–480.
[11] B. Hussain, Q. Du, and P. Ren, "Deep learning-based big data-assisted anomaly detection in cellular networks," in *Proc. IEEE Global Commun. Conf.*, 2018, pp. 1–6.
[12] K. Sultan, H. Ali, and Z. Zhang, "Call detail records driven anomaly detection and traffic prediction in mobile cellular networks," *IEEE Access*, vol. 6, pp. 41728–41737, 2018.
[13] N. C. Chen, W. Xie, R. E. Welsch, K. Larson, and J. Xie, "Comprehensive predictions of tourists' next visit location based on call detail records using machine learning and deep learning methods," in *Proc. IEEE Int. Congress Big Data*, 2017, pp. 1–6.
[14] M. Shafiq *et al.*, "A first look at cellular network performance during crowded events," in *Proc. Int. Conf. Meas. Model. Comput. Syst.*, 2013, pp. 17–28.
[15] Q. Ma *et al.*, "Factor analysis on call detail record," in *Proc. 27th Wireless Opt. Commun. Conf.*, 2018, pp. 1–5.
[16] E. J. Khatib, R. Barco, P. Munoz, I. D. L. Bandera, and I. Serrano, "Self-healing in mobile networks with big data," *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 114–120, Jan. 2016.
[17] J. Wen and V. O. K. Li, "Big-data-enabled software-defined cellular network management," in *Proc. Int. Conf. Softw. Netw.*, 2016, pp. 1–5.
[18] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Netw.*, vol. 30, no. 1, pp. 44–51, Jan./Feb. 2016.
[19] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu, "Big data analytics in mobile cellular networks," *IEEE Access*, vol. 4, pp. 1985–1996, 2016.
[20] S. Wang, X. Wang, J. Huang, R. Bie, and X. Cheng, "Analyzing the potential of mobile opportunistic networks for big data applications," *IEEE Netw.*, vol. 29, no. 5, pp. 57–63, Sep./Oct. 2015.
[21] P. Fiadino, P. Casas, A. D'Alconzo, M. Schiavone, and A. Baer, "Grasping popular applications in cellular networks with big data analytics platforms," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 3, pp. 681–695, Sep. 2016.
[22] E. Baccarelli, N. Cordeschi, A. Mei, M. Panella, M. Shojafar, and J. Stefa, "Energy-efficient dynamic traffic offloading and reconfiguration of networked data centers for big data stream mobile computing: Review, challenges, and a case study," *IEEE Netw.*, vol. 30, no. 2, pp. 54–61, Mar./Apr. 2016.
[23] J. Holub, M. Wallbaum, N. Smith, and H. Avetisyan, "Analysis of the dependency of call duration on the quality of VoIP calls," *IEEE Wireless Commun.*, vol. 7, no. 4, pp. 638–641, Aug. 2018.
[24] P. Chopade, J. Zhan, K. Roy, and K. Flurchick, "Real-time large-scale big data networks analytics and visualization architecture," in *Proc. the 12th Int. Conf. Expo Emerg. Technol. Smarter World*, 2015, pp. 1–6.
[25] J. Liu, F. Liu, and N. Ansari, "Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop," *IEEE Netw.*, vol. 28, no. 4, pp. 32–39, Jul./Aug. 2014.

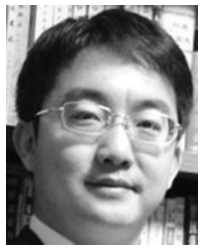
Dingde Jiang (S'08–M'09) received the Ph.D. degree in communication and information systems from the University of Electronic Science and Technology of China, Chengdu, China, in 2009.

He is currently a Professor in Communication and Information Systems with the School of Astronautics and Aeronautic, University of Electronic Science and Technology of China. His research is supported by National Science Foundation of China, and Program for New Century Excellent Talents with the University of Ministry of Education of China. His research interests include network measurement, modeling and optimization, performance analysis, network management, network security in communication networks, particularly in software-defined networks, information-centric networking, energy-efficient networks, and cognitive networks.

Dr. Jiang served as a Technical Program Committee Member for several international conferences. He was a recipient of best paper awards at several international conferences. He has been serving as an Editor for one international journal.

Yuqing Wang received the bachelor's degree in electronic information engineering from Southwest Jiaotong University, Chengdu, China in 2018. She is currently working toward the master's degree in control science and engineering with the University of Electronic Science and Technology of China, Sichuan, China.

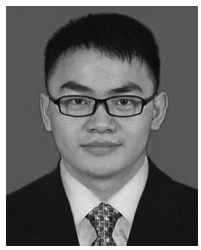
Her research interests include software defined networking and network management.



Zhihan Lv received the Ph.D. degree in computer applied technology from Paris7 University, Paris, France and Ocean University of China, Qingdao, China, in 2012.

He is currently an Associate Professor in Computer Science with Qingdao University, Qingdao, China. He was a Research Associate with the University College London, London, U.K. He worked with CNRS (France) as a Research Engineer; Umea University, Umea, Sweden, as a Postdoctoral Research Fellow; Fundacion FIVAN (Spain) as an Experienced Researcher. He was a Marie Curie Fellow in the European Union's Seventh Framework Program LANPERCEPT. He has authored/coauthored more than 200 papers in the related fields on journals such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and conferences, such as ACM Multimedia, ACM Conference on Human Factors in Computing Systems, ACM Special Interest Group on Computer Graphics, International Conference on Computer Vision, IEEE Virtual Reality. His research interests include Internet-of-Things, multimedia, augmented reality, virtual reality, computer vision, three-dimensional (3-D) visualization and graphics, serious game, HCI, big data, and GIS.

Dr. Lv is the Guest Editor for the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, FUTURE GENERATION COMPUTER SYSTEMS, *Neurocomputing*, *Neural Computing and Applications,* etc.



Sheng Qi received the bachelor's degree in communication and information systems from the College of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China, in 2018. He is currently working toward the master's degree in electronic and communication engineering with the University of Electronic Science and Technology of China, Sichuan, China.

His research interests include network measurement and traffic engineering.

Surjit Singh received the B.Tech. degree in computer engineering, the M.Tech. degree in computer science & engineering, and Ph.D. degree in computer engineering from the National Institute of Technology (NIT) Kurukshetra, Haryana, India, in 2008, 2012, and 2019, respectively.

He is currently working as an Assistant Professor with the Department of Computer Engineering, National Institute of Technology, Kurukshetra. He has several years of experience in teaching and research. He has authored/coauthored research papers with IEEE, IET, Elsevier, Springer, Wiley, and Taylor & Francis journals. His current research interests are Information Security, Internet of Things, Cloud Computing, Wireless Sensor and Ad hoc Networks, and Computational Intelligence.

Dr. Singh is also serving as a Guest Editor/Associate Editor for IEEE Transactions and other journals of high repute. He is also serving as a Reviewer and International Advisory Board of several journals including IEEE Transactions, Elsevier, Springer, Taylor & Francis, Inderscience and IGI Global. He has three books to his credit.