IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 14, 2020, accepted March 24, 2020, date of publication March 30, 2020, date of current version April 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984276

A Real-Time Hidden Anomaly Detection of Correlated Data in Wireless Networks

TENGFEI SUI, (Student Member, IEEE), XIAOFENG TAO, (Senior Member, IEEE), SHIDA XIA, (Student Member, IEEE), HUI CHEN, (Student Member, IEEE), HUICI WU, (Member, IEEE), XUEFEI ZHANG, (Member, IEEE), AND KECHEN CHEN

National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Xiaofeng Tao (taoxf@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61932005, Grant 61941114, and Grant 61631005, and in part by the 111 Project of China under Grant B16006.

ABSTRACT Wireless networks have been generating a plethora of unstructured and highly-correlated big data with hidden anomalies. The anomalies may bring inaccurate predictions of network behaviors, which further lead to inefficient system designs such as proactive caching placement. Current Random Matrix Theory (RMT) approaches are unable to detect hidden anomalies with a satisfying tolerance of data correlation. We present a novel data Decomposition aided Random Matrix Theory (DC-RMT) framework, which enables a real-time anomaly detection of large scale multi-dimensional and highly-correlated data. The detection results show that the proposed DC-RMT methodology can detect anomalies with an accuracy of 28 times better than RMT applied without data decomposition. The prediction results present a 6 times higher accuracy than data with anomaly, which will facilitate the identification of regions of interests, and contribute to the improvement of resource allocation efficiency and user QoE.

INDEX TERMS Anomaly detection, random matrix theory, data decomposition, network traffic prediction, big data, 5G and beyond.

I. INTRODUCTION

In the past two decades, the data generated by devices, people, and their interactions has experienced an exponential expansion to an enormous scale. In particular, wireless networks produce big data way beyond their processing and storing capacities. Driven by the developing Internet of Things (IoT) and proliferation of mobile devices, the number of connected devices is predicted to reach 12.3 billion by 2022 [1]. With massive amount of devices (smart phones, M2M modules, etc.) connected, wireless network traffic will reach 77.5 exabytes per month by 2022 with an annual growth rate of 46 percent from 2017 to 2022 [1]. The pervasive and exponentially increasing wireless network data presents imminent challenges to almost all aspects of the wireless system design. Volume, variety, velocity, and veracity are known as the 4Vs characteristics of big data [2]. At present, mobile cellular networks are evolving towards 5G and beyond. To generate and carry the big data all along costs a huge amount of resources.

However, in the realm of wireless networks, instead of treating big system data as an undesirable burden, it shall be leveraged as a great opportunity for better understanding of user demands and system capabilities. Moreover, big data brings network planning abundant new information that can be inter-connected to achieve a better understanding of users and networks (e.g., location, user velocity, social geo-data, etc.) [3]. With the help of big data analytic techniques, values in hidden patterns of wireless networks can be revealed. Big data technology offers a variety of powerful data-driven tools for handling diverse phenomena [4]. In addition, big data technology can be congruent with classical analyses and pretreatments. So that big data analytic techniques can be integrated with optimized network resource allocation to improve the quality of user experience, and to help maximize the revenue of operators as well [5].

In this big data embraced era, soon 5G and beyond, an emergent issue is the efficiency of the utilization of big data in mobile wireless networks. We are faced with both

The associate editor coordinating the review of this manuscript and approving it for publication was Xujie Li.

challenges and opportunities in designing the next generation wireless networks [3], [6], [7]. Several big data aided architectural issues related to IoT and cloud technologies are discussed in the industrial context [8]. And the mobile traffic has been studied from the perspective of service providers [9], [10].

It is the value hidden in big data that we strive for the most, especially data that seem to involve abnormal, unusual behaviors or activity patterns. Anomaly detection is to identify unexpected behaviors and patterns. Take anomaly wireless network traffic load as an example, it may be caused by a sleeping cell or a sudden activity that takes place in certain region. The unusual behavior with hidden patterns will provide insightful and critical information about the wireless network big data.

Moreover, it has drawn tremendous attention to utilize big wireless network data for the purposes of network resource allocation. Researchers take the network key factors such as traffic load, spectrum usage, energy efficiency and computing resource usage, etc., as individual inputs to conduct network resource analyses [11]. However, the modeling and prediction of the network key factors concerning time and location as well as their correlation patterns have not been fully exploited.

This study contributes to resolving the aforementioned issues in the following aspects.

- With big data analytic techniques, network key factors are modeled as multi-dimensional data.
  We propose a novel data Decomposition aided Random Matrix Theory (DC-RMT) framework. It reveals values that are hidden in highly-correlated and multi-dimensional wireless network dataset with anomalies. And an effectiveness criterion is proposed to evaluate the detection accuracy.
- Moreover, a modified Recurrent Neural Network (RNN) traffic prediction model is trained to predict the network behavior. The prediction results can be used to demonstrate how the anomalies may affect the network traffic prediction.
- A step further, the connections between network anomalies and social activities are analyzed and investigated with ground truth. We try to show that with the ability to identify the regions of interests (ROIs) and accordingly to reallocate network resources more efficiently, a smarter and greener wireless network is within the reach of the upcoming 5G and beyond era.

The remainder of this paper is organized as follows. In Section II, we review the relevant work in the literature. The big data description and preprocessing are depicted in Section III. In Section IV, we propose a DC-RMT methodology with an overview of RMT and data decomposition from the perspective of mathematical foundation. Section V presents a case study of our proposed network traffic anomaly detection framework with simulation and comparison. And Section VI presents a performance evaluation of the anomaly detection framework by applying a prediction model to data with and without anomalies. Section VII concludes the paper.

II. RELATED WORK

In the literature, anomaly detection has been well investigated in traditional wireless communication networks. The proposed methods can be roughly classified into three categories, statistic based methods, machine learning based methods and clustering based methods [12], [13]. Considering the detection accuracy, time and computing consumption, it is suggested in [14] that clustering based technique performs better. The clustering tools, i.e. K-means clustering and Hierarchical clustering, have been employed to conduct anomaly detection, and variables with multiple dimensions are integrated into one single dimension [15]. But this method simply removes the multi-dimensional correlation between separate variables without preserving the data continuity.

Clustering is a technique of assigning a set of data with shared properties into separate clusters. When it comes to multi-dimensional data, the clustering techniques may overlook data correlation and regard some unusual data as isolated, thus eliminate them, which results in a data link discontinuity, and loses properties of the original data [16].

Deep learning has been widely applied as a machine learning based anomaly detection approach. For instance, [17]–[19] propose deep learning-based anomaly detection models in cloud computing environments, which have greatly improved the capability to deal with streaming data.

RMT aims to reveal the inter-relationship of the matrix with entries that follow various probability distributions, and the entries can be referred to as random matrix ensembles. RMT originates from physics, mathematical statistics and numerical analysis in the beginning of the 20th century. It has become one of the statistical foundations for big data analytics [20], and has been widely applied to different fields, such as quantum systems [21], [22], financial systems [23], biological systems [24], smart grid systems [25], [26] as well as wireless communication networks [27]–[30]. Benefited from the capacity to analyze multi-dimensional variates with hidden anomaly patterns, RMT enables us to analyze multi-dimensional variates with an overall good continuity.

RMT has achieved significant results in analyzing highly correlated data in the realm of smart grids. In [26], a RMT based architecture is designed to conduct multi-dimensional analysis of multivariate data in real power systems, which enables the architecture to conduct anomaly detections. Reference [31] applies an augmented matrix to analyze correlations between the factors and the system status data, making it possible to identify the causing factors of the anomaly.

However, the empirical data used in [26] and [31] are considered to follow Gaussian distribution, which is not practical to analyze highly time-correlated data. An exception is seen in [32], which proposes to apply RMT to power systems in a non-Gaussian environment, considering the power system data as a time series.

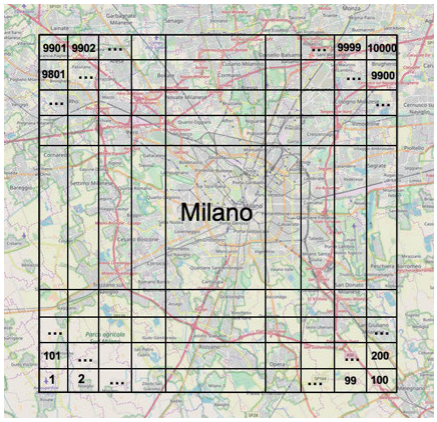


FIGURE 1. Approximate description of Milano grid.

Other works are devoted to the application of RMT to cellular networks from the perspectives of architecture and framework [29], [33], with all layers involved from the access network to the core network. Reference [34] utilizes RMT to analyze the abnormal user detection problem from the aspect of physical layer data analysis. To the best of our knowledge, the application of RMT in the realm of wireless network anomaly detection has not been addressed in the literature. Nevertheless, it is exactly the unique characteristics associated with RMT and wireless network anomaly detection that present interesting challenges.

Inspired by the above works, we propose a DC-RMT framework for the detection of anomalies in highly-correlated data. We consider the big data collected from real wireless network as a time series, and conduct data decomposition before anomaly detection in order to alleviate the impact of highly-correlated network traffic data. We then turn to DC-RMT anomaly detection framework, which contributes to the determination of ROIs. Finally, we apply a RNN based model to predict further network traffic as an evaluation of our proposed anomaly detection method.

III. CDR DATASET PREPROCESSING

Big data is an emergent paradigm for the analysis of network behaviors. In this section, we first describe the big data used for the understanding of user and network activities. A data preprocessing is then conducted by aggregation and a data description is also presented after aggregation.

A. DATASET DESCRIPTION

The Call Detail Records (CDRs) dataset consists of data collected from real LTE network of Telecom Italia at Milan, a major city located in the north of Italy, which is made public for the Big Data Challenge 2014 competition [35]. The data was collected from 3,450 base stations (BSs), providing information about the telecommunication activities of Milan. To facilitate the data analysis, the Milan region is divided into 100×100 grids named as Milan Grid, with each grid covering a square of 0.055 km^2, and all these BSs can be mapped into individual grids as shown in Fig. 1. When there are several

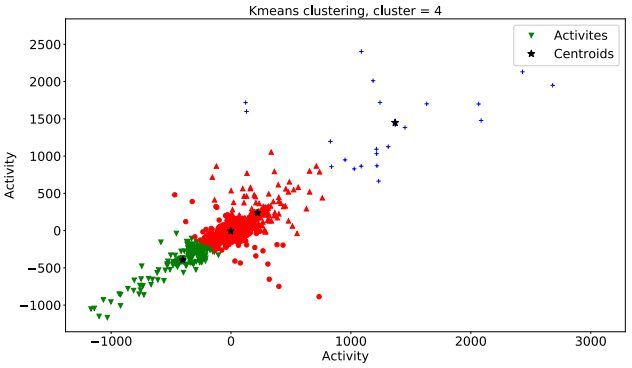


FIGURE 2. Clustering based anomaly detection.

TABLE 1. Dataset before processing.

Date	Time interval (min)	SMS_in activity	SMS_out activity	Call_in activity	Call_out activity	Internet traffic activity
11/1/2013	0	0.141864	0.156787	0.160978	0.052275	11.028366
11/1/2013	10	0.278452	0.119926	0.188777	0.133637	11.100963
11/1/2013	20	0.330641	0.170952	0.133637	0.054601	10.892771
...	...	...	...	...	...	...

BSs in one grid, all the traffic loads are aggregated into one traffic load [36].

The dataset is the result of computation over the CDRs generated by the Telecom Italia cellular network over the city of Milan. CDRs logs the user activities for billing purposes and network management. It contains various types of activities of each BS, including SMS activity, Call activity and Internet traffic activity with a 10 minutes time interval over two months, from November 2014 to December 2014. We are particularly interested in the correlation between network activities in the data. The network activities are shown in Table 1 in terms of all the recorded parameters inside the Square id within the Time interval (10 minutes). By means of clustering techniques, the hidden anomalies are clustered out of the CDRs dataset, as shown in Fig. 2.

B. DATASET PREPROCESSING

Data processing in wireless networks is first and foremost a priority challenge. The CDRs dataset provided by Telecom Italia is in a raw form which is not suitable for further analysis. Thus we perform a preprocessing, including cleaning and filtering. In order to enhance the readability, the CDRs dataset whose SMS, Call and Internet activities are recorded with a 10 minutes interval, is aggregated into a one-hour interval instead, as is shown in Table 2.

We observe that the network traffic runs with a routine, in both time scales of a month and a week, as is illustrated in Fig. 3. Although the traffic appears to be periodic in the time domain, some hidden values still remain in the irregular awaiting to be revealed.

TABLE 2. Dataset after aggregation.

Date	Time interval (hour)	SMS_in activity	SMS_out activity	Call_in activity	Call_out activity	Internet traffic activity
11/1/2013	0	0.732158	1.104748	0.591930	0.4019900	57.772872
11/1/2013	1	1.025872	0.770031	0.190564	0.1632640	11.100963
11/1/2013	2	0.3882782	0.3003906	0.0279247	0.1359635	41.179849
...	...	...	...	...	...	...

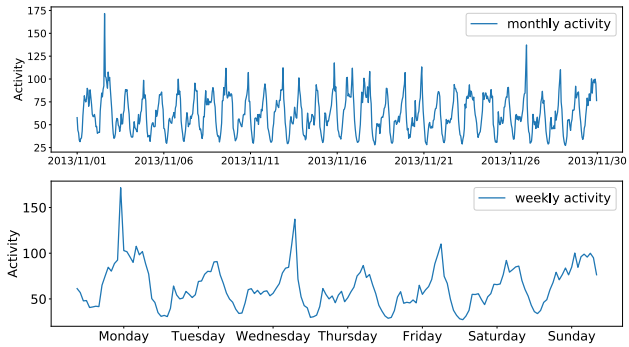


FIGURE 3. Traffic description after aggregation.

TABLE 3. Summary of notations.

Notations	Meaning
$\mathbf{X}, \mathbf{x}, x_{i,j}$	a matrix, a vector, an entry of a matrix
$\hat{\mathbf{X}}, \hat{\mathbf{W}}, \hat{\mathbf{x}}, \hat{y}$	raw data
$\bar{\mathbf{X}}, \bar{\mathbf{x}}, \tilde{x}, \tilde{\mathbf{Z}}$	transformed data, mostly by normalization
$\mathbf{X}_u$	singular value equivalent of $\bar{\mathbf{X}}$
$\mathbf{S}$	covariance matrix of $\mathbf{X}$
$\mathbf{U}$	Haar unitary matrix
$\bar{\tilde{x}}, \bar{\hat{x}}, \overline{r_{\lambda_{\tilde{Z}}}}$	averaged data
$\mathbb{C}^{N\times T}$	complex space
N, T	numbers of rows and columns
P, Q	numbers of $\kappa_{MSR} > \kappa_{0_{max}}$ and $\kappa_{MSR} < \kappa_{0_{min}}$
L	L independent matric product of $\mathbf{Z}$
c	ratio, $c = \frac{N}{T}$
r	radius of all eigenvalues of $\tilde{\mathbf{Z}}$ on the complex plane
κ_{MSR}	mean value of radius r
λ_S, λ_Z	eigenvalue of matrix $\mathbf{S}$, $\mathbf{Z}$
λ_{Z_i}	the i-th eigenvalue of matrix $\mathbf{Z}$
$\mu(x)$, $\sigma^2(x)$	mean, and variance
g_t, s_t, x_t	general trend, seasonality and stochastic component
$f_{ESD}(\cdot)$	empirical spectral distribution
$f_{ERMSE}(\cdot)$	effectiveness root mean square error
$K(\cdot)$	kernel function

IV. BIG DATA MODELING IN DC-RMT

This section presents our proposed novel DC-RMT methodology. A brief introduction of RMT is provided as a theoretical background before the detailed description of the DC-RMT based anomaly detection model for wireless networks. A criterion for effectiveness evaluation is also provided.

Table 3 summarizes the notations used in this paper.

A. RANDOM MATRIX THEORY

Massive multi-variate data can be structured by large random matrix naturally [25], [30]. We provide a general model as follows. We take N as variables that are sampled from the same grid at the same time. For each variable we take T times, thus a random matrix of $\mathbf{X} \in \mathbb{C}^{N\times T}$ can be obtained. According to RMT, when the dimension of a random matrix is sufficiently large, the empirical spectral distribution (ESD) of its eigenvalues always converges to some theoretical limits, which are Marchenko-Pastur Law (M-P Law) and Ring Law [37], presented respectively as follows.

1) MARCHENKO-PASTUR LAW

The M-P Law demonstrates that there exists an asymptotic behavior of singular values of large rectangular random matrices. Let $\mathbf{X} = \{x_{i,j}\}$ be an $N \times T$ random metric with independent identically distributed (i.i.d.) entries with the mean $\mu(x) = 0$ and the variance $\sigma^2 < \infty$. As $N, T \to \infty$ and the ratio $c = \frac{N}{T} \in (0, 1]$, the ESD of the corresponding sample covariance matrix $\mathbf{S} = \left(\frac{1}{N}\right)\mathbf{X}\mathbf{X}^H \in \mathbb{C}^{N\times N}$ converges to the M-P law. The density function is shown in Eq. (1),

$$f_{ESD}(\lambda_s) = \begin{cases} \dfrac{1}{2\pi\lambda c\sigma^2}\sqrt{(b-\lambda)\,(\text{lambda-a})}, & a \le \lambda \le b, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $a = \sigma^2\left(1-\sqrt{c}\right)^2$, $b = \sigma^2\left(1+\sqrt{c}\right)^2$.

2) KERNEL DENSITY ESTIMATION

The Kernel Density Estimation (KDE) is introduced as a comparison to the ESD of the eigenvalues of $\mathbf{S} \in \mathbb{C}^{N\times N}$. The KDE can be expressed as

$$f_{ESD}(\lambda_S) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-\lambda_{S,i}}{h}\right), \tag{2}$$

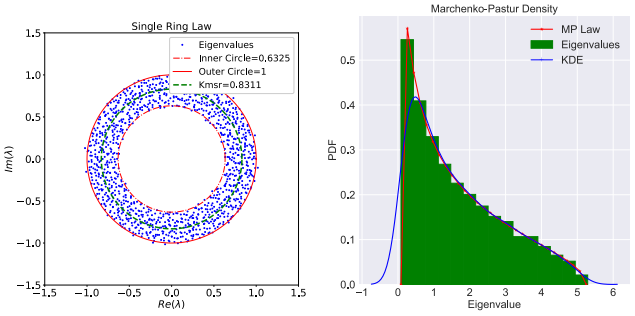
where $\lambda_{\mathbf{S},i}\,(i = 1, 2, \ldots, N)$ are the eigenvalues of $\mathbf{S}$, and $K(\cdot)$ denotes the kernel function.

3) THE RING LAW

The Ring Law has been one of the most remarkable modern probability achievements in the past decades. It acts as a useful tool to model massive multivariate datasets and extends RMT to large non-Hermitian random matrices [20], [38]. Consider the product of L non-Hermitian random matrices $\mathbf{Z} = \prod_{i=1}^{L}\mathbf{X}_{u,i}$, where $\mathbf{X}_u \in \mathbb{C}^{N\times N}$ is the singular value equivalent of the rectangular non-Hermitian random matrix $\tilde{\mathbf{X}} \in \mathbb{C}^{N\times N}$, whose entries are i.i.d. variables with the mean $\mu(x) = 0$ and the variance $\sigma^2 = 1$. The $\mathbf{Z}$ product operation effectuates the study of streaming datasets. For the sake of computational convenience and effectiveness, we set L to 1 in this paper, whereby the ESD of $\mathbf{Z}$ converges almost certainly to the limit given by Eq. (3),

$$f_{ESD}\left(\lambda_{\tilde{Z}}\right) = \begin{cases} \dfrac{1}{\pi cL}|\lambda|^{\frac{2}{L}-2}, & (1-c)^{\frac{L}{2}} \le |\lambda| \le 1, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

as $N, T \to \infty$, with the ratio $\frac{N}{T} = c \in (0, 1]$.



(a) Ring law and κ_{MSR}. (b) M-P law.

FIGURE 4. Ring law, M-P law, KDE and the κ_{MSR}.

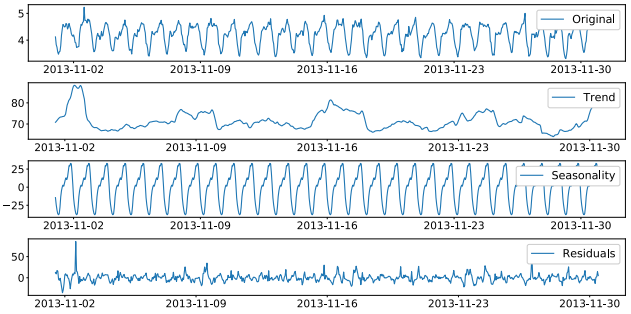


FIGURE 5. Data decomposition.

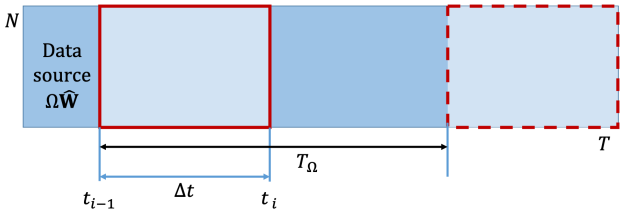


FIGURE 6. Real-time processing in a split-window.

The eigenvalues can then be distributed in a ring on the complex plane. The inner circle radius of the ring is $(1-c)^{\frac{L}{2}}$ and the outer circle radius is unity. Thus we are able to obtain $\mathbf{S} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^H = (\frac{1}{N})\mathbf{Y}\mathbf{Y}^H, (\mathbf{Y} = \sqrt{N}\tilde{\mathbf{Z}} \in \mathbb{C}^{N\times N}, \sigma^2(\tilde{z}) = \frac{1}{N}, \sigma^2(y) = \sigma^2(\sqrt{N}\tilde{z} = 1))$, and $\mathbf{S}$ conforms to the M-P law and the Ring law.

The mean spectral radius (MSR) is proposed to illustrate the distribution of the eigenvalues of $\tilde{\mathbf{Z}}$ [26], as shown in Fig. 4.

The mean value of the radius of all eigenvalues of $\tilde{\mathbf{Z}}$ is calculated and denoted as κ_{MSR} shown in Eq. (4),

$$\kappa_{MSR} = \frac{1}{N}\sum_{i=1}^{N}|\lambda_{\tilde{Z},i}|, \tag{4}$$

where $\lambda_{\tilde{Z},i}(i = 1, 2, \ldots, N)$ are the eigenvalues of $\tilde{\mathbf{Z}}$. The κ_{MSR} depicts the statistical distribution of all the eigenvalues, as the green line indicates in Fig. 4 (a).

With the aim to identify the anomaly behaviors of the wireless network, a real-time RMT analysis is conducted on the raw data collected from real cellular networks mentioned above in Section III.

B. DECOMPOSITION FOR REAL-TIME DATA PROCESSING

As shown in Fig. 3, the CDR dataset shows a clear temporal pattern that fluctuates similarly on a daily basis, which accords with the human periodic routines to work in the daytime and rest at night. However, at a closer look of the pattern, we can tell that it is composed of strong regular components and considerable stochastic components.

It is difficult to simply apply RMT to real wireless network data, because the data does not follow Gaussian distribution. More importantly, it is highly-correlated in the time domain, and may be correlated in the space domain as well, which could jeopardize our RMT analysis.

Considering the high time correlation characteristics of the data, and inspired by [9] and [32], in which a time series decomposition method is proved to be practical in modeling mobile traffic data, we consider each CDR parameter dataset as a time series, in particular, an Autoregressive Integrated Moving Average (ARIMA) time series. We leverage the Wavelet Transform decomposition method to separate stochastic components from regular components for further RMT analysis [39].

We denote the raw CDR parameter datum in a grid as $\hat{\mathbf{W}} = \{\hat{\mathbf{w}}_{t1}, \hat{\mathbf{w}}_{t2}, \hat{\mathbf{w}}_{t3}, \ldots, \hat{\mathbf{w}}_{tn}\}$, representing the activity of each CDR parameter at time t_i. Meanwhile, each parameter can be defined as in Eq. (5),

$$\hat{\mathbf{w}}_t = g_t + s_t + x_t, \quad t = 1, 2, \ldots, n, \tag{5}$$

where g_t represents the general trend of $\hat{\mathbf{w}}_t$, s_t the data seasonality, and x_t the stochastic data component, as illustrated in Fig. 5.

We are more concerned with the stochastic components, which contain hidden pattern information but have less time correlation. We use these stochastic components for a random matrix analysis.

C. DC-RMT ANALYSIS FOR WIRELESS NETWORK

As described above, highly correlated time-series data is not suitable for RMT analyses. Therefore, we propose a DC-RMT methodology for highly correlated time-series data collected from real networks.

For the raw data source $\Omega\hat{\mathbf{W}}$, we apply a split-window to obtain a sample data matrix $\tilde{\mathbf{W}}$ from it, as illustrated in Fig. 6. For a certain time $t_i = t_{i-1} + \Delta t$, each parameter can be arranged as a time series vector $\hat{\mathbf{w}}_{t_i}$. With the split-window sliding along T_Ω, a multi-dimensional matrix N can be formed as a raw data source $\Omega\hat{\mathbf{W}}$ that describes the network behaviors.

The decomposition of $\tilde{\mathbf{W}}$ results in a raw data matrix $\hat{\mathbf{X}} \in \mathbb{C}^{N\times T}$. The data after decomposition is then modeled as N-dimensional vectors $\hat{\mathbf{x}}_i$ at each sampling time. When the

sampling proceeds to a scale as large as T, $\hat{X} \in \mathbb{C}^{N\times T}$ can then be formed with $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_T$.

A normalization process is performed to obtain a non-Hermitian data matrix $\tilde{\mathbf{X}} \in \mathbb{C}^{N\times T}$ row-by-row in Eq. (6):

$$\tilde{x}_{i,j} = (\hat{x}_{i,j} - \overline{\hat{x}_i}) \times (\sigma(\tilde{x}_i)/\sigma(\hat{x}_i)) + \overline{\tilde{x}_i}, \ 1 \le i \le N; 1 \le j \le T, \quad (6)$$

where the raw data $\hat{\mathbf{x}}_i = (\hat{x}_{i,1}, \hat{x}_{i,2}, \ldots, \hat{x}_{i,T})$, the mean $\overline{\hat{\mathbf{x}}_i} = 0$, and the variance $\sigma^2(\tilde{\mathbf{x}}) = 1$.

The singular value equivalent matrix $\mathbf{X}_u \in \mathbb{C}^{N\times N}$ of $\tilde{\mathbf{X}} \in \mathbb{C}^{N\times T}$ can be obtained by

$$\mathbf{X}_u = \sqrt{\tilde{\mathbf{X}}\tilde{\mathbf{X}}^H}\mathbf{U}, \quad (7)$$

where $\mathbf{U} \in \mathbb{C}^{N\times N}$ is a Haar unitary matrix, $\mathbf{X}_u\mathbf{X}_u^H \equiv \tilde{\mathbf{X}}\tilde{\mathbf{X}}^H$.

For L arbitrarily assigned independent non-Hermitian matrices $\hat{\mathbf{X}}_i (i = 1, \ldots, L)$, which are sampled and decomposed from the raw data source $\Omega\hat{\mathbf{Y}}$, the matrices product $\mathbf{Z} = \prod_{i=1}^{L} \mathbf{X}_{u,i} \in \mathbb{C}^{N\times N}$ is obtained. The data matrix $\tilde{\mathbf{Z}}$ is then normalized row-by-row by procedure Eq. (8):

$$\tilde{z}_i = z_i/(\sqrt{N}\sigma(z_i)), \quad 1 \le i \le N \quad (8)$$

where $z_i = z_{i,1}, z_{i,2}, \ldots, z_{i,N}$, and $\tilde{z}_i = \tilde{z}_{i,1}, \tilde{z}_{i,2}, \ldots, \tilde{z}_{i,N}$.

The mean value of the radius of all eigenvalues of $\tilde{\mathbf{Z}}$ can be calculated and denoted as κ_{MSR}

$$\kappa_{MSR} = \overline{r_{\lambda_{\tilde{\mathbf{Z}}}}} = \frac{1}{N}\sum_{i=1}^{N} |\lambda_{\tilde{Z}}|, \quad (9)$$

where r is the radius of all eigenvalues of $\tilde{\mathbf{Z}}$ which are supported on the complex plane. The corresponding κ_{MSR} serves as an indicator of the anomalous dataset. Additionally, the sample covariance matrix $\mathbf{S} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^H$ can be obtained to perform the comparison of its histogram, KDE, and M-P law.

V. DETECTION AND VERIFICATION OF ANOMALIES

In this section, the performance of the proposed DC-RMT approach is compared to that of the RMT approach and the most commonly applied clustering approach. The effectiveness evaluation of the three approaches is conducted. A flow chart of the DC-RMT approach with a real-world application scenario is illustrated in Fig. 7.

A. DC-RMT DETECTION

Our big data based anomaly detection method through RMT has been demonstrated in Section IV (A). In the case study, we apply DC-RMT to analyze the highly-correlated data in a data decomposition architectural framework in mobile wireless network, and all data is obtained from real networks, hence the data contains noises. The specific functional grids are geographically accorded with the google map as depicted in Fig. 8, which is Grid 5848 with a convention center in it. It is characterized by a strong network traffic routine with occasional anomaly network behaviors when a conference takes place in the convention center.

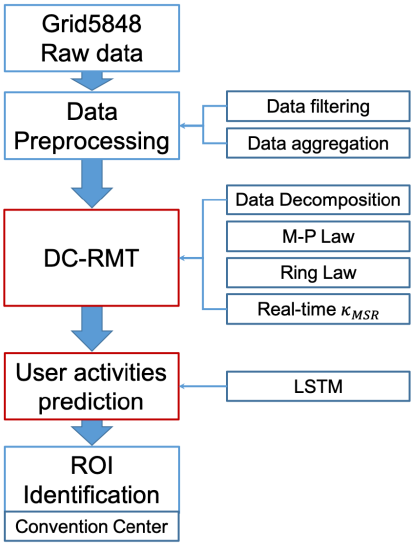


FIGURE 7. An illustrative flow of the DC-RMT approach with a real-world application scenario.

FIGURE 8. Selected grids in our case study.

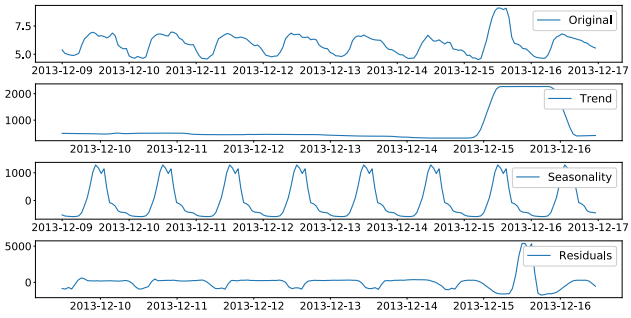
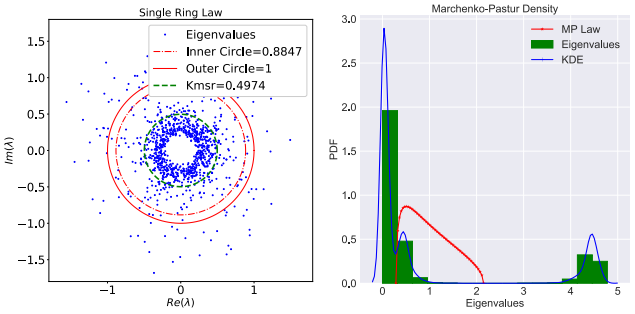


FIGURE 9. Grid 5848 after decomposition.

Fig. 9 plots the selected case with aggregated CDR data collected over a week, from December 10th to 16th, 2013. The selected case involves a convention center area on Google Map. The activity level shows regular seasonality in the first few days, but increases dramatically on December 15th,



(a) Ring law grid 5848. (b) M-P law grid 5848.

FIGURE 10. Ring law, comparison between KDE and M-P law.

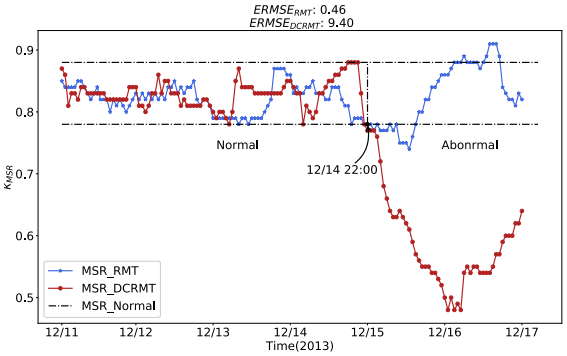


FIGURE 11. κ_{MSR} of grid 5848.

and returns to normal right after that day, which stands out ostentatiously in the Trend plot. The proliferating network demand on that day, is considered as a network anomalous behavior.

We use the residuals to conduct a DC-RMT analysis. In the simulation, the specific Δt is 24 hours, and we run it in one week $T_\Omega = 168$ hours. Fig. 10 (a) and (b) show the eigenvalues of the Ring Law and the histogram and KDE deviation from the M-P law in the presence of abnormal network behaviors. With a fixed $\Delta t = 24$ hours in the simulation, κ_{MSR} is obtained by hour with an average analysis and response time of 0.4057s, which indicates the hourly variation of the network traffic pattern.

It can be noted that the eigenvalues converge to κ_{MSR}, and the statistical histogram representation coincides with the theoretical distribution curve, which indicates that an abnormal network behavior occurs at this time.

B. MSR COMPARISON

For comparison, we apply both DC-RMT and the traditional RMT approach to analyze the same data. The comparison results of κ_{MSR} are shown in Fig. 11. The performance of DC-RMT is denoted with the red line, whose κ_{MSR} of the first five days fluctuates between 0.8 and 0.9. However, on the sixth day, it decreases dramatically down to below 0.55.

The verification results are investigated with Google Map geographically. We find that the results coincide with the anomalies occurred near Milano Convention Centre at grid 5848, where a big conference was held. The participants of the conference were gathered together, and their texting, calling and sharing of multimedia contents with mobile phones led to a surge of traffic flow. We consider the surge as an abnormal behavior, which corresponds to the significant change of κ_{MSR} exactly at 22:00 p.m. on December 14th. In other words, the anomalous data clearly impinge on the κ_{MSR}, and our proposed DC-RMT framework is effective and time-sensitive for anomaly detection.

The performance of the RMT approach is illustrated with the blue line, whose κ_{MSR} fluctuates between 0.7 and 1.0 for the whole week, which is unable to tell the anomalies.

Furthermore, we propose an effectiveness root mean square error (ERMSE) as an evaluation criterion. The anomaly κ_{MSR} that deviates from the normal values is calculated as f_{ERMSE}, which is given in Eq. (10),

$$f_{ERMSE} = \sum_{i=1}^{P} \sqrt{(\kappa_n - \kappa_{0_{max}})^2} + \sum_{j=1}^{Q} \sqrt{(\kappa_m - \kappa_{0_{min}})^2}, \quad (10)$$

where κ_n and κ_m are the κ_{MSR} of the data with anomaly, specifically on December 15th in this case, and P, Q are numbers of $\kappa_{MSR} > \kappa_{0_{max}}$ and $\kappa_{MSR} < \kappa_{0_{min}}$ respectively. The $\kappa_{0_{max}}$ and $\kappa_{0_{min}}$ are the maximum and minimum κ_{MSR} of the data without anomaly.

With the proposed DC-RMT methodology, the ERMSE can reach up to 9.40. Whereas with the RMT approach, the ERMSE shrinks to 0.46. Hence it is difficult for the latter to tell the anomalies from the normal condition.

The results suggest that DC-RMT outperforms RMT with respect to the same highly-correlated multi-dimensional wireless network traffic data. With the DC-RMT anomaly detection method, we are able to locate the anomalies with a much higher ERMSE, which is 28 times higher than that in the RMT analysis.

C. MDR COMPARISON

To evaluate the performance of DC-RMT and the clustering based anomaly detection approach proposed in [15], we adopt the miss detection rate P_{MDR} as the evaluation criterion, which is denoted by Eq. (11),

$$P_{MDR} = \frac{T_{AN} + T_{NA}}{T_\Omega}, \quad (11)$$

where T_{AN} is the detected times of abnormal behaviors in anomaly-free scenario, T_{NA} is the detected times of normal behaviors in anomaly-existing scenario, and T_Ω is the total detection times. $P_{MDR} = 0$ indicates that the anomaly detection approach detects all the anomalous network behaviors correctly with no miss detection and false detection. Note that the proposed DC-RMT can only detect the occurrence of anomaly entries, but cannot figure out the number of anomalous entries in the sampled matrix $\hat{\mathbf{W}}$. Hence, the miss detection rate P_{MDR} is adopted as the performance evaluation criterion.

Fig. 12 depicts the miss detection rate versus time epochs for the DC-RMT method and the clustering based anomaly

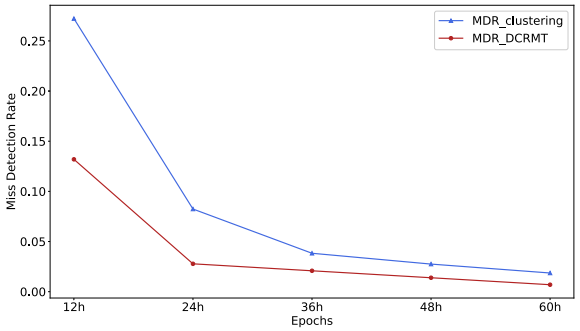


FIGURE 12. Simulation of miss detection rate versus time epochs.

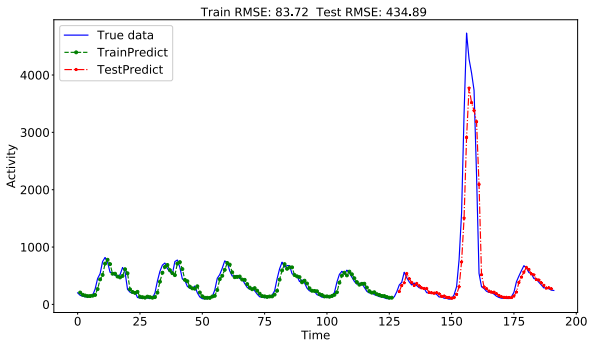


FIGURE 13. LSTM prediction RMSE of grid 5848 with anomaly.

detection approach proposed in [15]. The P_{MDR} declines substantially with the increment of time epochs, which coincides with the fact that with the more data analyzed, the lower miss detection rate we will get. With the characteristics of real-time anomaly detection on large scale multi-dimensional and highly-correlated data, the DC-RMT approach achieves a much lower miss detection rate than the clustering based approach.

The results suggest that DC-RMT outperforms clustering based approaches with respect to highly-correlated multi-dimensional wireless network traffic data.

A step further, ROIs can be determined with a higher level of identification accuracy and dynamics, and the network resources (including radio resource, power resource, cache resource, etc.) can be allocated more efficiently, which eventually will contribute to the improvement of user QoE.

VI. WIRELESS NETWORK TRAFFIC PREDICTION

In this section, a well trained network traffic prediction model is introduced to demonstrate the impact of the anomalies on network resource allocation.

The real-time anomaly detection procedure conducted for grid 5848 suggests that we are able to detect and locate the anomalies. After the abnormal data is replaced with average traffic data obtained from the two months' CDRs, we apply RNN for the network traffic prediction. However, since the highly-correlated time series, as described above in Fig. 3, is quite tricky for standard RNN predictive modeling, we turn to the Long Short-Term Memory (LSTM), which is known as a type of RNN, but powerful in handling time-relevant problems [40], [41]. With LSTM, we train a modified prediction model to conduct the performance evaluation.

We take the prediction Root Mean Square Error (RMSE) between the training, the test and the real data as an evaluation criterion. The first five days' data is used to train our LSTM prediction model, and the last three days' data serves as the test data. The prediction RMSEs of the data with and without anomaly are presented in Fig. 13 and 14 respectively. Fig. 13 shows that the LSTM prediction model, when acted upon the data with anomaly, results in a RMSE up to 434.89, which is almost 6 times higher than that of the data free of anomaly.

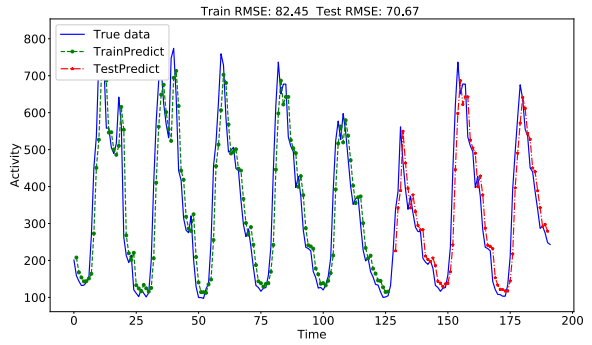


FIGURE 14. LSTM prediction RMSE of grid 5848 without anomaly.

As we can tell from the prediction results, even a well trained prediction model can result in poor performance when acting upon data containing anomalies, which may further lead to disappointing network performance. When it comes to 5G and beyond network with exponentially growing data to be generated and carried, a real-time and accurate anomaly detection method will become even more essential for the improvement of network performance.

VII. CONCLUSION

In this paper, we have proposed a novel big data based DC-RMT methodology for anomaly detection in wireless networks. The method is applied to analyze the wireless network CDRs data collected from real network with a data decomposition procedure. The successful real-time detection of anomaly patterns proves the effectiveness of DC-RMT in the analysis of highly-correlated and multi-dimensional wireless network traffic data.

Given that the network traffic data is inherently highly-correlated with user activities, unusual user activities may lead to abnormal network demands. In our case study, huge network traffic demands are caused by a conference held in that area. We categorize the unusual traffic surge as an anomaly, which is verified by ground truth. In addition, by testing with data with and without anomaly in a modified LSTM prediction model, we have shown that the model trained by anomaly-free data results in less RMSE.

The real-time anomaly detection capability of the proposed DC-RMT methodology enables us to predict and locate anomalies, whereby patterns of the ROIs can be dynamically identified, categorized and predicted. Owing to the capability of analyzing large scale multi-dimensional and highly-correlated data, DC-RMT can be further applied to the identification of the impact correlated factors that actually caused the anomalies in the ROIs. The network operations in ROIs can thus proceed more efficiently, in particular with respect to proactive caching and resource allocation.

Big data analytics aided network deployment and resource allocation will certainly become an indispensable technology for the upcoming 5G and beyond era. It will also insert its impact on the design of the next generation wireless network architectures.

REFERENCES

[1] C. V. N. Index, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper*. San Jose, CA, USA: Cisco, 2019.

[2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[3] 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. 2016.

[4] N. Zhang, P. Yang, J. Ren, D. Chen, L. Yu, and X. Shen, "Synergy of big data and 5G wireless networks: Opportunities, approaches, and challenges," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 12–18, Feb. 2018.

[5] J. Liu, N. Chang, S. Zhang, and Z. Lei, "Recognizing and characterizing dynamics of cellular devices in cellular data network through massive data analysis," *Int. J. Commun. Syst.*, vol. 28, no. 12, pp. 1884–1897, Aug. 2015.

[6] A. Imran and A. Zoha, "Challenges in 5G: How to empower SON with big data for enabling 5G," *IEEE Netw.*, vol. 28, no. 6, pp. 27–33, Nov. 2014.

[7] X. Cao, L. Liu, Y. Cheng, and X. Shen, "Towards energy-efficient wireless networking in the big data era: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 303–332, 1st Quart., 2018.

[8] B. Cheng, J. Zhang, G. P. Hancke, S. Karnouskos, and A. W. Colombo, "Industrial cyberphysical systems: Realizing cloud-based big data infrastructures," *IEEE Ind. Electron. Mag.*, vol. 12, no. 1, pp. 25–35, Mar. 2018.

[9] F. Xu, Y. Lin, J. Huang, D. Wu, H. Shi, J. Song, and Y. Li, "Big data driven mobile traffic understanding and forecasting: A time series approach," *IEEE Trans. Services Comput.*, vol. 9, no. 5, pp. 796–805, Sep. 2016.

[10] 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.

[11] Y. Li, Y. Zhang, K. Luo, T. Jiang, Z. Li, and W. Peng, "Ultra-dense HetNets meet big data: Green frameworks, techniques, and approaches," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 56–63, Jun. 2018.

[12] M. Ahmed, A. Anwar, A. N. Mahmood, Z. Shah, and M. J. Maher, "An investigation of performance analysis of anomaly detection techniques for big data in SCADA systems," *EAI Endorsed Trans. Ind. Netw. Intell. Syst.*, vol. 2, no. 3, p. e5, 2015.

[13] I. A. Karatepe and E. Zeydan, "Anomaly detection in cellular network data using big data analytics," in *Proc. Eur. Wireless 20th Eur. Wireless Conf.*, May 2014, pp. 1–5.

[14] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 303–336, 1st Quart., 2014.

[15] 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. Informat.*, vol. 13, no. 4, pp. 2058–2065, Aug. 2017.

[16] Y. Yan, G. Sheng, R. C. Qiu, and X. Jiang, "Big data modeling and analysis for power transmission equipment: A novel random matrix theoretical approach," *IEEE Access*, vol. 6, pp. 7148–7156, 2018.

[17] S. Garg, K. Kaur, N. Kumar, G. Kaddoum, A. Y. Zomaya, and R. Ranjan, "A hybrid deep learning-based model for anomaly detection in cloud datacenter networks," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 3, pp. 924–935, Sep. 2019.

[18] S. Garg, K. Kaur, N. Kumar, S. Batra, and M. S. Obaidat, "HyClass: Hybrid classification model for anomaly detection in cloud environment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[19] S. Garg, K. Kaur, S. Batra, G. S. Aujla, G. Morgan, N. Kumar, A. Y. Zomaya, and R. Ranjan, "En-ABC: An ensemble artificial bee colony based anomaly detection scheme for cloud environment," *J. Parallel Distrib. Comput.*, vol. 135, pp. 219–233, Jan. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0743731519304721

[20] T. Tao, *Topics Random Matrix Theory*, vol. 132. Providence, RI, USA: American Mathematical, 2012.

[21] T. A. Brody, J. Flores, J. B. French, P. A. Mello, A. Pandey, and S. S. M. Wong, "Random-matrix physics: Spectrum and strength fluctuations," *Rev. Modern Phys.*, vol. 53, no. 3, pp. 385–479, Jul. 1981.

[22] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, "Random-matrix theories in quantum physics: Common concepts," *Phys. Rep.*, vol. 299, nos. 4–6, pp. 189–425, Jun. 1998.

[23] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud, "Random matrix theory and financial correlations," *Int. J. Theor. Appl. Finance*, vol. 3, no. 3, pp. 391–397, Jul. 2000.

[24] F. Luo, J. Zhong, Y. Yang, R. H. Scheuermann, and J. Zhou, "Application of random matrix theory to biological networks," *Phys. Lett. A*, vol. 357, no. 6, pp. 420–423, Sep. 2006.

[25] R. C. Qiu and P. Antonik, *Smart Grid Using Big Data Analytics: A Random Matrix Theory Approach*. Hoboken, NJ, USA: Wiley, 2017.

[26] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 674–686, Mar. 2017.

[27] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[28] R. C. Qiu, Z. Hu, H. Li, and M. C. Wicks, *Cognitive Radio Communication and Networking: Principles and Practice*. Hoboken, NJ, USA: Wiley, 2012.

[29] 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.

[30] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Found. Trends Commun. Inf. Theory*, vol. 1, no. 1, pp. 1–182, 2004, doi: 10.1561/0100000001.

[31] X. Xu, X. He, Q. Ai, and R. C. Qiu, "A correlation analysis method for power systems based on random matrix theory," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1811–1820, Jul. 2017.

[32] B. Han, L. Luo, G. Sheng, G. Li, and X. Jiang, "Framework of random matrix theory for power system data mining in a non-Gaussian environment," *IEEE Access*, vol. 4, pp. 9969–9977, 2016.

[33] C. Zhang and R. C. Qiu, "Massive MIMO as a big data system: Random matrix models and testbed," *IEEE Access*, vol. 3, pp. 837–851, 2015.

[34] H. Chen, X. Tao, N. Li, S. Xia, and T. Sui, "Physical layer data analysis for abnormal user detecting: A random matrix theory perspective," *IEEE Access*, vol. 7, pp. 169508–169517, 2019.

[35] (2014). *First Edition of the Big Data Challenge*. [Online]. Available: https://dandelion.eu/datamine/open-big-data/

[36] (2014). *The Milano Grid Spatial Description*. [Online]. Available: https://dandelion.eu/datagems/SpazioDati/milano-grid/description/

[37] A. Guionnet, M. Krishnapur, and O. Zeitouni, "The single ring theorem," 2009, *arXiv:0909.2214*. [Online]. Available: http://arxiv.org/abs/0909.2214

[38] A. Edelman and Y. Wang, "Random matrix theory and its innovative applications," in *Advances in Applied Mathematics, Modeling, and Computational Science*. Boston, MA, USA: Springer, 2013, pp. 91–116, doi: 10.1007/978-1-4614-5389-5_5.

[39] M. West, "Time series decomposition," *Biometrika*, vol. 84, no. 2, pp. 489–494, Jun. 1997.

[40] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.* New York, NY, USA: Curran Associates, 2015, pp. 802–810.

[41] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," in *Proc. 9th Int. Conf. Artif. Neural Netw. ICANN* , Sep. 1999, pp. 850–855. [Online]. Available: https://digital-library.theiet.org/content/conferences/10.1049/cp_19991218

TENGFEI SUI (Student Member, IEEE) received the B.S. and M.S. degrees in electronic and communication engineering from Shandong Normal University, Jinan, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree in information and communication engineering with the Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interests are in the areas of wireless communications and networks, with current emphasis on big data aided proactive caching and network resource allocation.

XIAOFENG TAO (Senior Member, IEEE) received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 1993, and the M.S. and Ph.D. degrees in telecommunication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1999 and 2002, respectively. He is currently a Professor with BUPT. He has authored or coauthored over 200 articles and three books in wireless communication areas. His research interest focuses on 5G/B5G. He is also a Fellow of the Institution of Engineering and Technology and the Chair of the IEEE ComSoc Beijing Chapter.

SHIDA XIA (Student Member, IEEE) received the B.E. degree in communication engineering from the Harbin University of Industry, Harbin, China, in 2016. He is currently pursuing the Ph.D. degree in information and communication engineering with the Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interests are in the areas of wireless communication networks, with current emphasis on physical layer authentication for mobile communication systems, and machine learning in security mechanism.

HUI CHEN (Student Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in June 2013, and the Ph.D. degree in communications and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2019. From 2016 to 2017, he was a Visiting Student with the Department of Electrical and Computer Engineering, University of Houston. His research interests are in the areas of wireless communications and networks, physical layer security, random matrix, and data analysis.

HUICI WU (Member, IEEE) received the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2018. From 2016 to 2017, she was with the Broadband Communications Research (BBCR) Group, University of Waterloo, Waterloo, ON, Canada. She is currently an Assistant Professor with BUPT. Her research interests are in the areas of wireless communications and networks, with current emphasis on the collaborative air-to-ground communication and wireless access security. She served as the Publication Co-Chair for APCC 2018 and a Guest Editor for *Science China Information Sciences*.

XUEFEI ZHANG (Member, IEEE) received the B.S. and Ph.D. degrees in telecommunications engineering from the Beijing University of Posts and Telecommunications (BUPT), in 2010 and 2015, respectively. From September 2013 to August 2014, she was with the School of Electrical and Information Engineering, The University of Sydney, Australia. She is currently with the National Engineering Laboratory, BUPT. Her research interests include mobile edge computing, blockchain, reinforcement learning, and intelligent transportation systems.

KECHEN CHEN received the B.E. degree in information and communication engineering from the Communication University of China, Beijing, China, 2013. She is currently pursuing the M.B.A. degree with the Beijing University of Posts and Telecommunication, Beijing.