Received February 10, 2016, accepted February 25, 2016, date of publication March 9, 2016, date of current version May 11, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2540520

Big Data Analytics in Mobile Cellular Networks

YING HE[1], FEI RICHARD YU[2], (Senior Member, IEEE), NAN ZHAO[1], (Member, IEEE), HONGXI YIN[1], HAIPENG YAO[3], AND ROBERT C. QIU[4,5], (Fellow, IEEE)

[1]Department of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[2]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada
[3]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[4]Research Center for Big Data Engineering and Technologies and the State Energy Smart Grid Research and Development Center, Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[5]Department of Electrical and Computer Engineering, Tennessee Technological University, Cookeville, TN 38505, USA

Corresponding author: H. Yao (yaohaipeng@bupt.edu.cn).

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61201224, Grant 61471056, and Grant 61372089, in part by the Fundamental Research Funds for the Central Universities under Grant DUT14QY44, in part by the China Jiangsu Future Internet Research Fund under Grant BY2013095-3-1 and Grant BY2013095-3-03, in part by the Beijing Advanced Innovation Center for Future Internet Technology, and in part by the Natural Sciences and Engineering Research Council of Canada. The work of R. C. Qiu was supported in part by NSFC under Grant 61571296 and in part by the National Science Foundation under Grant ECCS-0901420, Grant ECCS-0821658, and Grant CNS-1247778.

ABSTRACT Mobile cellular networks have become both the generators and carriers of massive data. Big data analytics can improve the performance of mobile cellular networks and maximize the revenue of operators. In this paper, we introduce a unified data model based on the random matrix theory and machine learning. Then, we present an architectural framework for applying the big data analytics in the mobile cellular networks. Moreover, we describe several illustrative examples, including big signaling data, big traffic data, big location data, big radio waveforms data, and big heterogeneous data, in mobile cellular networks. Finally, we discuss a number of open research challenges of the big data analytics in the mobile cellular networks.

INDEX TERMS Big data analytics, mobile cellular networks.

I. INTRODUCTION

Recent years have witnessed tremendous advances in wireless cellular networks [1], [2]. With recent advances of wireless technologies and ever-increasing mobile applications, mobile cellular networks have become both generators and carriers of massive data [3]. When geo-locating mobile devices, recording phone calls, and capturing mobile applications' activities, an enormous amount of data is generated and carried in mobile cellular networks.

Historically, the massive data in mobile cellular networks hasn't been paid much attentions. With data constantly accumulated in the database and the technologies of *big data analytics* rapidly developed, the great value hided behind data has gradually been revealed. It is desirable to make good use of this precious resource, big data, to improve the performance of mobile cellular networks and maximize the revenue of operators [4]. Traditional data analytics shows its in-adequateness when encountered with the big cellular data. First, traditional data analytics deals with structured data. The large amount of App-based data is, however, generally unstructured. Second, the implementation of data analysis is traditionally confined within a department, or a business unit. The final analytical conclusions come from very limited, local angles, rather than global perspectives. Third, the analytics mainly aims at transaction data, and pays less attention to the operational data, due to its incapability to make real-time decisions.

Big data analytics can extract much more insightful information than traditional data analytics, and can help improve the performance mobile cellular networks and maximize the revenue of operators [5]. For example, the complete data related to a subscriber is usually fragmented in different business departments. Big data analytics is capable of collecting the scattered data to understand the user behavior and preferences from multiple perspectives to portray an integrated picture. Moreover, subscribers' living habits and the timetable can be generally inferred from the usage of traffic over different time periods of a day; their surfing habits and interests can be roughly obtained from the logs; their frequently visited places or the range of activities can be approximately derived from home location register (HLR) databases. Another significant feature of big data analytics is real-time processing. With big data analytics, operators can monitor their infrastructure in real-time, and make autonomous and dynamic decisions.

Despite the potential vision of big data analytics in mobile cellular networks, many significant research challenges remain to be addressed before the widespread deployment of big data analytics in mobile cellular networks. In particular, there is data confidentiality issue for the purpose of subscribers' privacy-preserving and security-protection. Mobile cellular networks have large amount of sensitive personal information, such as subscriber's names, ID numbers, physical locations, images files, top contacts, passwords, etc. If operators fail to leverage big data in a proper way, big data analytics will bring privacy/security issues to mobile cellular networks. In addition, as mobile cellular networks have scarce bandwidth, how to filter out un-useful data and compress/transmit useful data presents significant challenges to the design of mobile cellular networks.

In this paper, we introduce an unified data model based on random matrix theory and machine learning. Then, we present a framework that enables big data analytics in mobile cellular networks. In addition, we discuss several case studies of big data analytics in mobile cellular networks, including big signaling data, big traffic data, big location data, big radio waveforms data, and big heterogeneous data. Moreover, we present a number of challenges that need to be addressed for the deployment of big data analytics in mobile cellular networks.

To the best of our knowledge, the interrelationship between big data and mobile cellular networks has not been well addressed in the existing works. In essence, it is the unique characteristics associated with big data and mobile cellular networks that present interesting challenges that have not been fully tackled in the literature. We believe that the works we have done here help understand how to make full use of big data analytics to improve the performance of mobile cellular networks.

The rest of the article is organized as follows. In Section II, we present an overview of big data analytics. A framework of big data analytics is then presented in Section III. Following that, we discuss several case studies that leverage big data analytics in cellular networks. Section IV discusses several open research challenges. We conclude this study in Section V.

II. OVERVIEW OF BIG DATA ANALYTICS BASED ON RANDOM MATRIX THEORY

In this section, we introduce an unified data model based on random matrix theory and machine learning. Then, we present big distributed data and deep learning, followed by big data analytics for mobile cellular networks.

A. REPRESENTATION OF BIG DATA WITH LARGE RANDOM MATRICES

The essence of big data analytics is to exploit the high-dimensionality of the spatial-temporal datasets. Following the arguments of [6], we introduce the basic ideas of big data analytics here. Denote the random vector by $\mathbf{x}$ of dimension n, where n is large in value. For example, for 100 antennas with each orthogonal frequency-division multiplexing (OFDM) symbol of 128 modulated tones, we reach $n = 100 \times 128 = 12,800$. These n data samples are corresponding to some spatial points (modulation tones are functions of spatial points). The n-dimensional vector space is natural for data modeling for this problem. High-dimensional statistics suggests that a great number N samples of n dimensions are jointly taken into account, to extract correlation from these massive datasets.

Naturally, one may wonder how the high dimensionality n affects the big data analytics? Answering this simple question is critical to big data research. In particular, we collect N independent realizations of the random vector $\mathbf{x}$, i.e., $\mathbf{x}_1, \ldots, \mathbf{x}_N$. A random (data) matrix $\mathbf{X}$ of size $n \times N$ is formed using

$$\mathbf{X} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T], \tag{1}$$

where T stands for the matrix transpose. In the big data era, many problems have comparable dimensionality and samples. In other words, both n and N are large, but their ratio $c = N/n$ is fixed. This large random matrix paradigm demands a fundamental deviation from the classical regime in statistics [6]: n is fixed (small), and N is very large.

It is believed that the foundation of big data analytics is based on three ingredients: (1) high-dimensional statistics, (2) matrix analysis, and (3) convex optimization. This belief is exhibited in the study of large random matrices for big data analytics.

From the viewpoint of analysis, the data matrix is the basic departure point for any big data analytics. The matrix can be decomposed into eigenvalues and eigenvectors using the eigenvalue decomposition. Let us say we obtain n eigenvalues of the $n \times n$ sample covariance matrix $\mathbf{S}$

$$\mathbf{S} = \frac{1}{N}\mathbf{X}\mathbf{X}^H, \tag{2}$$

where H represents the transpose and conjugate (Hermitian) of a matrix. These n eigenvalues are non-negative. Since the dimensionality n is large, we naturally turn to the cumulative distribution function (or probability density function) to study the n random values.

One may ask a simple question: Does the probability density function $f_n(x)$ converge to some deterministic function $f(x)$ as $n \to \infty$? We are lucky that the answer to this question is positive and the resulting distribution functions are the famous Marchenko-Pasture (MP) law (discovered in 1967). The conditions for the MP law to be valid are general enough for almost all engineering problems. The most fundamental condition is the requirement for the high dimensionality, or large n.

Naturally, the finite size n will be used to ''approximate'' the asymptotic limit $n \to \infty$. The penalty for this approximation is the error bound of $O(\frac{1}{n})$. This large deviation bound is very useful.

Given a data matrix X defined in (1), can we do something different from (2)? The answer for this basic question is the Single Ring Law that is discovered very recently in [7]. We define a new matrix transform

$$\mathbf{Y} = \sqrt{\mathbf{S}}\mathbf{U} = \sqrt{\frac{1}{N}\mathbf{X}\mathbf{X}^H}\mathbf{U}, \qquad (3)$$

where $\mathbf{U}$ of size $n \times n$ is the unitary Haar matrix, a random matrix. Let us compare the two paradigms. The n eigenvalues of the data matrix $\sqrt{\mathbf{S}}$ is supported on the non-negative real-axis. The n eigenvalues of the transformed matrix $\mathbf{Y}$ are, however, supported on the WHOLE complex plane!

An immediate consequence of this new paradigm is to study the anomalous behaviours of big data analytics, as illustrated in Fig. 1. We emphasize the observation that anomalous behaviors exhibit themselves in the form of outliers. Please refer to [8] for more details.

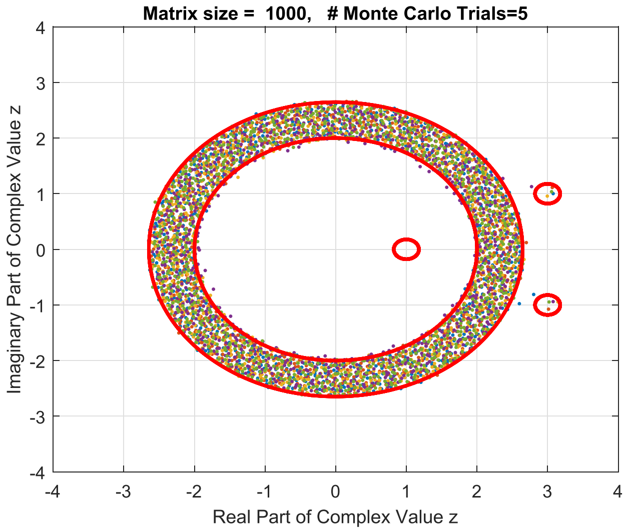

FIGURE 1. The eigenvalues of the transformed random matrix defined in (3) are distributed within the single ring, as predicted by the single ring theorem. Outliers are clear from the figure. Simulation parameters: $n = 1000$. Five Monte Carlo trials are overlapped in the figure.

B. CASE STUDIES FOR WIRELESS NETWORKS

For each wireless node in a wireless network [6], [8], we observe a sequence of T samples that is represented with a random (column) vector $\mathbf{x}_i$, for the i-th node. For N nodes, we represent N random vectors with a data matrix

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathbb{C}^{T\times N}.$$

Often we deal with Gaussian random matrix $\mathbf{G}$ with the entries G_{ij} Gaussian random variables. The entries are independent. The G is a non-Hermitian random matrix, which is way more difficult than the Hermitian case.

To mine the massive data, we usually study two types of Hermitian random matrices $\frac{1}{2}\left(\mathbf{G}+\mathbf{G}^H\right)$ and $\mathbf{G}\mathbf{G}^H$. Similarly, we can study the general case

$$\frac{1}{2}\left(\mathbf{X}+\mathbf{X}^H\right) \text{ and } \mathbf{X}\mathbf{X}^H.$$

Our fundamental technique arises from the simple observation that the two cases (hypotheses) for $\mathbf{G}$ and $\mathbf{X}$ are different! The key is to discover the statistic metrics that differentiate the two cases. We call this anomaly detection, where the case $\mathbf{G}$ is normal.

Modern research topics are of interest:

1) What is the *smallest size* of N such that the random matrix paradigm is still valid?
2) What happens if the entries of $\mathbf{X}$ are *non-Gaussian*?

Experiences indicate that we can replace general data matrix $\mathbf{X}$ with the Gaussian random matrix $\mathbf{G}$, when the sizes of matrices are sufficiently large! Gaussian random matrices are much easier to deal with analytically.

The phenomenological nature of random matrix models may be regarded as the drawback. But this feature provides, on the other hand, the model free frameworks, that make the approach valid for a wide variety of systems having different microscopic natures or origins. These frameworks have a certain amount of "robustness" of the random matrix models. It is believed that a "sufficiently large" number of those models should have no dependence or rather weak dependence on the random matrix ensemble used. This belief of universality partly explain the fact that the major ideas of the random matrix models are based on the Gaussian ensemble and the circular ensemble [9].

C. BIG DATA ANALYTICS WITH ASYMPTOTIC LIMIT DISTRIBUTIONS

In Section II-A, we represent the big data, in (1), using high-dimensional data matrix $\mathbf{X}$, which is regarded as a large random matrix. This viewpoint of random matrix theory connects the rich literature with the big data analytics. We deal with *matrix-valued* random variables, rather than the conventional vector-valued random variables [10]. There is a deep connection between large random matrices and big data analytics [11]. We only give a tutorial review here. Our goal is to make the link with the big data applications.

Modern probability theory is based on three fundamental results for *scalar-valued or vector-valued* random variables: (1) the law of large numbers; (2) the central limit theorem; (3) Berry-Esseen inequality, or a rate of convergence for the central limit theorem. In analogy with scalar-valued or vector-valued random variables, we obtain analogous fundamental results for matrix-valued random variables. A recent review is given in [11]. The key arises from the observation that the eigenvalues are *empirically* computed from the data matrices $\mathbf{X}$ defined in (1). These empirical eigenvalues are regarded as big data analytics. The central result was obtained by E. Wigner at Oak Ridge National Laboratory in 1950s; he discovers that the cumulative distribution function (CDF) $F_n(x)$ converges to a limit distribution $F(x)$ as the size of data matrix n goes to infinity, or $n \to \infty$. This milestone marks the birth of the so-called random matrix theory—See three monographs [6], [12], [13] together with a fourth one [14].

Our intuition is based on the following observation: For conventional multivariate statistics [15], our starting point is the data matrix $\mathbf{X}$ defined in (1). The conventional multivariate statistics is valid under the restricted condition that

$$n \text{ is fixed}, \quad N \to \infty. \tag{4}$$

Condition (4) is unfortunately invalid for the modern big data analytics. This observation is the very reason why one co-author (Qiu) dedicates four monographs on this subject. We are motivated by the unifying paradigm of using large random matrices for data representation—a vision that is for the first time spelled out in [6].

In classical probability, the Law of Large numbers and the Central Limit Theorems are the core. Theorems for Wigner's semicircle law and Marchenko-Pastur law can be viewed as random matrix analogues of the Law of Large Numbers from classical probability theory. Thus a Central Limit Theorem for or fluctuations of linear eigenvalue statistics is a natural second step in studies of the eigenvalue distribution of any ensemble of random matrices.

We define the iid matrix as follows. Let ξ be a random variable. We say $\mathbf{X}_N$ is an iid random matrix of size N with atom variable ξ if $\mathbf{X}_N$ is an $N \times N$ matrix whose entries are iid copies of ξ.

Theorem 1 (Circular Law [16]): Let ξ be a complex random variable with mean zero and unit variance. For each $N \geq 1$. Let $\mathbf{X}_N$ be an iid random matrix of size N with atom variable ξ. Then, for any bounded and continuous function $f : \mathbb{C} \to \mathbb{C}$,

$$\int_{\mathbb{C}} f(z) d\mu_{\frac{1}{\sqrt{N}}\mathbf{X}_N}(z) \to \frac{1}{\pi}\int_{\mathbb{U}} f(z) d^2 z$$

almost surely as $N \to \infty$ where $\mathbb{U}$ is the unit disk in the complex plane $|z| \leqslant 1$ and $d^2 z = dxdy$, with $z = x + iy$.

The Single Ring Theorem (or Law), by Guionnet, Krishnapur and Zeitouni (2011) [7], describes the empirical distribution of the eigenvalues of a large generic matrix with prescribed singular values, i.e. an $N \times N$ matrix of the form $\mathbf{A} = \mathbf{UTV}$, with $\mathbf{U}; \mathbf{V}$ some independent Haar-distributed unitary matrices and $\mathbf{T}$ a deterministic matrix whose singular values are the ones prescribed. More precisely, under some technical hypotheses, as the dimension N tends to infinity, if the empirical distribution of the singular values of $\mathbf{A}$ converges to a compactly supported limit measure Θ on the real line, then the empirical eigenvalues distribution of $\mathbf{A}$ converges to a limit measure μ on the complex plane which depends only on Θ. The limit measure μ is rotationally invariant in $\mathbb{C}$ and its support is the annulus $S := \{z \in \mathbb{C}; a \leqslant |z| \leqslant b\}$ with $a, b \geq 0$ such that

$$a^{-2} = \int x^{-2} d\Theta(x) \text{ and } \mathrm{b}^2 = \int x^2 d\Theta(x). \tag{5}$$

There is no eigenvalue outside the annulus (the support of the limiting spectral distribution). Fig. 1 illustrates the Single Ring Theorem. One application of the Single Ring Theorem is the anomaly detection for power grids. In recent years, the power grid has been experiencing a significant shift from the traditional electricity grid to the smart grid [17], [18]. When the power grid is normal, the Single Ring Theorem demands that all the experimentally obtained eigenvalues are distributed within a single ring; this is not true, on the other hand, when the power grid experiences some events. See [12], [19], [20] for more details.

The idea of using packet drops for a network of 200 radio nodes are studied, in order to locate strong interfere and measure its transmit power in a wireless data broadcasting system [21]. One advantage is that the packet drops data is free information and not particularly generated by consuming extra radio resource. Without strong interference, the eigenvalues fall within the single ring, as demanded by the Single Ring Theorem. This is not true in the presence of strong interference.

D. BIG DATA ANALYTICS WITH FLUCTUATIONS OF LINEAR EIGENVALUE STATISTICS

For a random matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ whose eigenvalues λ_i, $i = 1, \ldots, n$, are known as strong correlated random variables, the central object of statistical study is the so-called linear eigenvalue statistics defined as

$$\mathrm{Tr} f(\mathbf{A}) = \sum_{i=1}^{n} f(\lambda_i) \tag{6}$$

where Tr represents the trace of a square matrix and $f(\cdot)$ is a test function that assumes certain smoothness. Define the empirical eigenvalue distribution

$$\mu_n(dx) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - \lambda_i) dx,$$

where $\lambda_i, i = 1, 2, \ldots, n$ are eigenvalues of $\mathbf{X}_n$. The following functions are of special interest in the context of this paper.

1) Linear function $f(x) = x$.
2) Radius $f(z) = |z|, \quad z \in \mathbb{C}$.
3) Moments $f(x) = x^k, k \in \mathbb{N}$.
4) Generalized variance $f(x) = \log x$.
5) Mutual information $f(x) = \log(1 + x)$.
6) Likelihood ratio test (LRT) $f(x) = \log x - x - 1$.
7) Stieltjes transform $f(x) = \frac{1}{w-x}$, $\mathrm{Im}\, w \neq 0, w \in \mathbb{C}$.
8) Exponentials $f(x) = e^{jtx}, t \in \mathbb{R}$.
9) Analytical function $f(z) = \sum_{n=0}^{\infty} a_n z^n, \quad z \in \mathbb{C}$.

Theorems for Wigner's semicircle law and Marchenko-Pastur law can be viewed as random matrix analogues of the Law of Large Numbers from classical probability theory. Thus a Central Limit Theorem for or fluctuations of linear eigenvalue statistics is a natural second step in studies of the eigenvalue distribution of any ensemble of random matrices.

For each $n \geq 1$, let $\mathbf{A}_n = \frac{1}{n}\mathbf{X}_n^H \mathbf{X}_n$ be a real sample covariance matrix of size n, where $\mathbf{X}_n = \{X_{ij}\}_{1 \leqslant i,j \leqslant n}$, and $\{X_{ij} : 1 \leqslant i, j \leqslant n\}$ is a collection of real independent random variables with zero mean and unit variance. The eigenvalues are ordered such that $\lambda_1(\mathbf{A}_n) \leqslant \lambda_2(\mathbf{A}_n) \leqslant \cdots \leqslant \lambda_n(\mathbf{A}_n)$.

The test function f from the space $\mathcal{H}_s$ has the norm

$$\|f\|_s^2 = \int (1 + 2\,|\omega|)^{2s} |F\,(\omega)|^2 d\omega$$

for some $s > 3/2$, where $F\,(\omega)$ is the Fourier transform of f defined by

$$F\,(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{j\omega t} f(t)dt.$$

We note that if f is a real-valued function with $f \in \mathcal{H}_s$ for some $s > 3/2$, the both f and its derivative f' are continuous and bounded almost everywhere. In particular, this implies that f is Lipschitz.

Suppose that $\mathbb{E}\left[X_{ij}^4\right] = m_4$ for all $1 \leq i, j \leq n$ and all $n \geq 1$. Assume there exists $\varepsilon > 0$ such that

$$\sup_{n\geqslant 1} \sup_{1\leqslant i,j\leqslant n} \mathbb{E}\left|X_{ij}\right|^{4+\varepsilon} < \infty.$$

Let f be a real-valued function with $\|f\|_s < \infty$ for some $s > 3/2$. Then it follows from [22] that the linear eigenvalue statistics

$$\sum_{i=1}^{n} f\,(\lambda_i\,(\mathbf{A}_n)) - \mathbb{E} \sum_{i=1}^{n} f\,(\lambda_i\,(\mathbf{A}_n)) \to \mathcal{N}\left(0, v^2\,[f]\right) \quad (7)$$

in distribution as $n \to \infty$, where the variance $v^2\,[f]$ is a function of f defined by (8).

$$\begin{aligned} v^2\,[f] = &\frac{1}{2\pi^2} \int_0^4 \int_0^4 \left(\frac{f\,(x) - f\,(y)}{x - y}\right)^2 \\ &\times \frac{(4 - (x-2)\,(y-2))}{\sqrt{4 - (x-2)^2}\sqrt{4 - (y-2)^2}} dxdy \\ &+ \frac{m_4 - 3}{4\pi^2} \left(\int_0^4 \frac{x-2}{\sqrt{4 - (x-2)^2}} dx\right)^2. \end{aligned} \quad (8)$$

For big data, we are interested in the performance of algorithms at different scales of matrix sizes n. The variance of the linear eigenvalue statistics does not grow to infinity in the limit $n \to \infty$ for sufficiently smooth test functions. Finite size corrections are in the order of $O\left(1/n^2\right)$ for this variance. Massive datasets from power grids [23] indicate that the above expressions are very accurate.

E. BIG DATA ANALYTICS WITH FREE PROBABILITY

One central problem with big data is data fusion: how to merge a number of large random matrices? When the sizes of random matrices are sufficiently large, the *deterministic functional* relation exist through the free probability theory [6]. Finite size corrections for $n \times n$ random matrices are in the order of $O\left(1/n^2\right)$ for the variance of this functional.

For each $n \in \mathbb{N}$, let $\mathbf{X}_1^{(n)}, \ldots, \mathbf{X}_r^{(n)}$ be a system of r independent complex Hermitian random matrices from the class SGRM $\left(n, \frac{1}{n}\right)$, where SGRM $\left(n, \sigma^2\right)$ denotes the set of $n \times n$ self-adjoint Gaussian random matrices. Let $x_1, \ldots, x_r$ be a semicircular system. Then for any polynomial p in r noncommutative random variables the convergence

$$\lim_{n\to\infty} \left\| p\left(\mathbf{X}_1^{(n)}, \ldots, \mathbf{X}_r^{(n)}\right) \right\| = \|p\,(x_1, \ldots, x_r)\|$$

holds almost surely [24], [25]. Here $||\cdot||$ denotes the operator norm. The intuition behind using this polynomial lies in the following observation: the dimension of this polynomial is rn^2, rather than n^2. Higher dimensions leads to better properties to exploit. For example, we repeat the observation for the power grid or wireless network for r times, each of which gives one random matrix of size $n \times n$. When n is sufficiently large, we can use the polynomial to compute the big data analytics.

F. BIG DATA ANALYTICS WITH CONCENTRATION OF MEASURE

The non-asymptotic analysis of random matrices [26] says that the matrix dimensions n and N in (1) are large, but finite! The monograph [13] dedicates a large part to this topic. This non-asymptotic paradigm is much more practical—but much harder—than the asymptotic alternative. Spectrum sensing for cognitive radio can be viewed as big data analytics. A large family of algorithms are based on the eigenvalues of large random data matrices. Due to Concentration of Measure Phenomenon in high-dimensional (vector) space, it is shown that the individual eigenvalues exhibit a variance of $O\,(1/n)$, while the sum of eigenvalues (the trace) has $O\left(1/n^2\right)$. This fundamental difference suggest a new observation for hypothesis testing of alternative random matrices. The second-order statistics (variance) of the data matrices are basic to studying algorithms. Some applications are documented in [27]–[29]. This observation is similar to the second-order coding rate of the MIMO communications channel [30] and the well-known Polyanskiy-Poor-Verdu work [31]. The coding rate is within $O\left(1/\sqrt{L}\right)$ for finite code length L.

G. LINKING DEEP LEARNING WITH RANDOM MATRIX THEORY: A NOVEL VISION

Artificial intelligence and machine learning are main techniques for big data analytics. In the future, it is believed that they will account for 80-90% of the global computing resources [32]. Deep learning may account for 40-50% of this market [32]. Artificial intelligence is an area of computer science that deals with giving machines the ability to seem like they have human intelligence. One example is a robot. Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension – as is the case in data mining applications – machine learning uses that data to improve the program's own understanding.

Machine learning programs detect patterns in data and adjust program actions accordingly.

The algorithms for big data analytics need data and engineering to validate. Since 2006, deep learning [33] has been the most active new area of machine learning research. The major remaining challenges arise from the lack of explicit mathematical expressions: it is unclear that the designed algorithms are reproducible, extendable, theoretically provable, and interpretable. In the context of big data analystics for cellular networks, deep learning is of great interest. Deep learning—the mathematics in data science and a basic methodology—is the state-of-the-art for images and speech. The depth of images can be extracted.

On the other hand, the big data in cellular networks is distributed across different domains, e.g., space, time, codes, and antennas. Distributed machine learning is among the important areas in machine learning. Given the current explosion in the size and amount of data, a central challenge in machine learning is to design efficient algorithms for solving large-scale problem instances [34].

In [35], the authors propose matrix neural networks, which takes matrices directly as inputs. Therefore, the input layer neurons form a matrix, for example, each neuron corresponds to a pixel in a grey scale image. The upper layers are also but not limited to matrices. Matrices are passing through each layer without vectorization that is previous algorithms. To achieve this, each neuron senses summarized information through bilinear mapping from immediate previous layer units' outputs plus an offset term. The basic model of a layer is the following bilinear mapping

$$\mathbf{Y} = \sigma\left(\mathbf{U}\mathbf{X}\mathbf{V}^T + \mathbf{B}\right) + \mathbf{E} \tag{9}$$

where $\mathbf{U}$, $\mathbf{V}$, $\mathbf{B}$ and $\mathbf{E}$ are matrices with compatible dimensions, $\mathbf{U}$ and $\mathbf{V}$ are connection weights, $\mathbf{B}$ is the offset of current layer, $\sigma(\cdot)$ is the activation function acting on each element of matrix and $\mathbf{E}$ is the error. This bilinear mapping certainly connects matrix neural networks to matrix or tensor factorization type of algorithms such as principal component analysis (PCA). Random matrix theory is naturally linked with PCA.

H. BIG DATA ANALYTICS FOR MOBILE CELLULAR NETWORKS

Here, we highlight the connection between big data analytics and mobile cellular networks, and in a bigger picture, the link between data science and wireless networks. Big data analytics is not the replacement, but rather a supplement to fill in the gap between today's higher requirement for deciphering more potential information and the traditional data analytics.

In mobile cellular networks, the sources of traditional analytics are basically *centralized*, such as from charging and billing systems, operation systems, etc. In practice, however, the huge amount of data is scattered across the organization, like the device data, cell site data, network data, back office data, etc. Higher dimensionality of data implies better inference, as mentioned in Secction II-A, $O(1/n)$, the convergence rate of the empirical eigenvalue distribution to its limit. Big data analytics supports an global point of view, making it possible to integrate the distributed collected data to extract correlation, as mentioned in the discussion of big distributed data in Section II-G. Semi-structured, unstructured data cannot be processed until big data analytics comes to the scene. Speaking of data volume, besides its suitability for magnitude of TB, even PB, big data analytics has a remarkable feature, called scalability, i.e., the ability to analyze the data with ever-increasing scale and complexity. Mobile cellular networks' operational decisions have usually been made manually or depending on the hardware inside. With the advences of big data analytics, the operations of mobile cellular networks can be lower-error, higher-precision, dynamic, and importantly in real-time. Real-time reactions bring in better decisions, not only for the optimization of the network, but also the quality of user experience. Big data analytics has many striking advantages. Nevertheless, at this point, its analysis tools are generally complex and programming-intensive, not the same friendly as the traditional ones.

By effectively applying big data analytics, nearly every department involving sales and marketing, customer support, operation and maintenance, network construction, etc. can achieve significant benefits. Many opportunities are right over there, for instances, tailored marketing campaigns and recommendations can be carried out for a specific group of subscribers; more initiative concerns or solutions can be delivered to customers rather than waiting for the complains or claims; real-time monitoring of the operation and maintenance systems can prevent the fraudulent behaviours, or warn the congestion conditions; pinpointed network coverage analysis can facilitate the construction of network layout and further enhance subscribers' quality of experience. Through the various innovative schemes, mobile cellular operators can enhance customers' loyalty and thus lower the customer churn, simplify business operations, develop new services, reduce expenditure and increase revenue.

III. AN ARCHITECTURAL FRAMEWORK TO SUPPORT BIG DATA ANALYTICS IN MOBILE CELLULAR NETWORKS

In this section, we present an architectural framework to support big data analytics in mobile cellular networks.

A. DATA COLLECTION

Mathematically speaking, data collection is the process of forming data matrix $\mathbf{X}$, shown in (1). It is worthwhile to point out that the dimensionality of data vector is very high. We take a view of treating data matrix $\mathbf{X}$ as a large random matrix. This view has many consequences, leveraging new mathematical results such as free probability.

Big data in mobile cellular networks can be gathered from either internal or external sources. The external data comes from the state/local statistic bureau, market research agencies, customer complaint departments, and etc. The internal sources usually refer to the operational systems, business systems and other supporting systems. Operational systems have been but not well explored by the traditional data analytics, and enormous values are expected to be excavated by the powerful big data analytics. Therefore, the data collection in operational systems is the focus of this section.

Data collection methods can be divided into two categories: through data sources, and through auxiliary tools [36]. Mobile devices themselves are data collection tools. For examples, they can: (1) collect audio information through microphones; (2) collect pictures, videos, and other multimedia information through cameras; (3) collect geological locations through GPS, WiFi or bluetooth, etc. Network data can be acquired through some package capture technologies or certain specialized softwares, such as ComView, SmartSniff, etc. Furthermore, professional staff can put some probes into the network interfaces, such as air interfaces, A/Gn interfaces, and etc. to collect the signalling data.

B. BIG DATA ANALYSIS AND PREPROCESSING

From a statistical analysis point of view, we extract correlation contained in the data matrix **X**. We present two basic approaches. One is the sample covariance matrix. The other is, through a matrix transform, to use the fundamental Singe Ring Theorem [7].

The large-scale collected datasets reside at different locations with different formats. Thus, the gathered datasets are usually at a raw state with much redundancy, inconsistency or useless information. Meanwhile, the involved vast amount of semi-structured and unstructured data make it impossible to fit into the relational database with neat tables of columns and rows. NoSQL databases are becoming the alternative technology for big data. To avoid unnecessary storage space, and ensure the processing efficiency, the data should be preprocessed to be ready for data analysis, before it is transmitted to the storage systems. For example, the data collected from mobile APPs, geolocation sensors, video cameras, network logs, CDRs, weblogs will not be stored directly in a storage system until it is preprocessed to correlate and normalize the data. Three common data preprocessing techniques are: integration, cleaning and redundancy elimination [36].

C. BIG DATA ANALYTICS PLATFORMS AND TOOLS

When it comes to big data, the most frequently mentioned word is Hadoop [37]. Apache Hadoop is an open-source software framework for distributed storage and distributed processing of large-scale datasets. The power of clusters enforces Hadoop to store and process data at an amazing speed. Initially, Hadoop is developed for such routine functions as keyword classification on search engines. To cater for the requirements of various business applications, Hadoop gradually turns into a general-purpose big data operating platforms, where different data manipulations and data analytical operations can be plugged into. All the features make Hadoop peculiarly adapt to processing or analyzing the data in mobile cellular networks, such as CDRs, GPS data, web clickstream, network logs, etc. which is needed to be pulled out from storage systems for analysis frequently. Meanwhile, various suitable data analysis methods for mobile cellular networks can be integrated into Hadoop platforms.

Apache Spark is a popular open-source platform for large-scale data processing that is well-suited for iterative machine learning tasks [37]. By allowing user programs to load data into a cluster's memory and query it repeatedly, Spark is well-suited to machine learning algorithms. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's multi-stage in-memory primitives provides performance up to 100 times faster for certain applications [38].

D. BIG DATA ANALYTICS APPLICATIONS

The applications of big data analytics in mobile cellular networks can be divided into two categories: internal business supporting applications and external innovative business model development. The internal business supporting applications mainly include the operational efficiency, subscribers' experience enhancement, tailored marketing, etc. As mobile cellular networks have vast amount of useful data, innovative business models can be promoted, such as the third-party data providers for various enterprises without infringement of subscribers' privacy.

IV. CASE STUDIES OF BIG DATA ANALYTICS IN MOBILE CELLULAR NETWORKS

The following case studies provide an illustration of introducing big data analytics into mobile cellular networks. The focus is on improving network performance and deriving valuable insights. The application scope covers different scenarios, from current deployed mobile cellular networks to upcoming 5G, from network operational optimization to push for emerging research topics.

A. BIG SIGNALING DATA

In mobile cellular networks, the transmission of voice and data is accompanied by control messages, which are termed as signaling. The signaling works according to the predefined protocols and ensure the communication's security, reliability, regularity and efficiency. Signaling monitoring plays an important role in appropriate allocation of network resources, improving the quality of network services [39], real-time identifying network problems, and etc. With the rapid development of various mobile cellular networks, the volume of signaling data grows tremendously and the traditional signaling monitoring systems have too many problems to deal with.

In Fig. 2, we describe a signaling data monitoring and analyzing system architecture with big data analytics. This architecture mainly consists of three components: data collecting, data analyzing and applications. In data collection, various signaling protocols are copied from multiple network interfaces without interrupting normal operations. Afterwards, these copies are gathered and filtered through the protocol processor and then sent to the analyzer. In the analyzer,

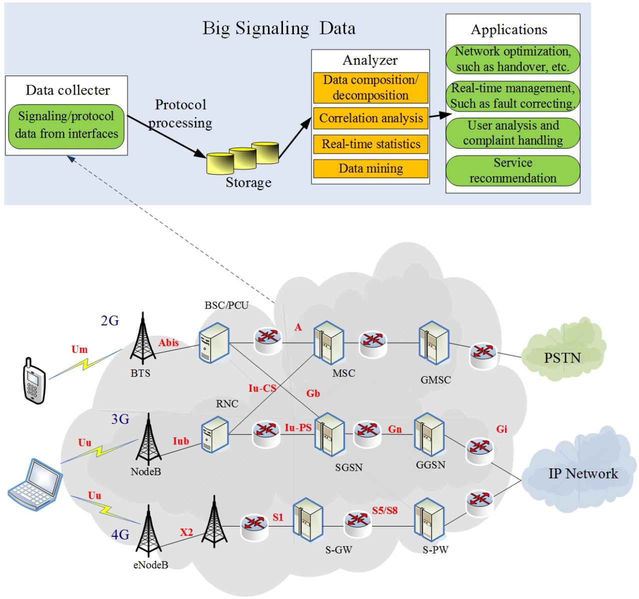

FIGURE 2. The big signaling data in cellular networks.

the data is processed using various algorithms, such as decomposition, correlation analysis, etc. Finally, the analysis results can be used by various applications. For example, Celibi *et al.* [40] analyzed the BSSAP messages from *A* interface in a Hadoop platform to identify handovers from 3G to 2G. The simulation results show that the identified 3G coverage holes are consistent with the drive test results.

B. BIG TRAFFIC DATA

With the widespread usage of mobile Internet, the volume of traffic data increases at an unprecedented rate. Acting as a carrier of the traffic data, cellular operators have to manage the network resource appropriately to balance network load and optimize network utilization. Traffic monitoring and analyzing is an elementary but essential part for network management, enabling performance analysis and prediction, failure detection, security management, etc. Traditional approaches to monitor and analyze the traffic data seem, however, straightforward and inadequate in the context of big traffic data, as illustrated in Fig. 3. In [41], the interrelationship between big data and softare-defined networking (SDN) [42]–[45] has been studied.

Liu *et al.* [3] proposed a novel large-scale network traffic monitoring and analysis system based on a Hadoop platform. The system is practically deployed in a commercial cellular network with 4.2 Tbytes input volume every day. The evaluation results indicate that the proposed system is capable of processing big network-generated data and revealing certain traffic and user behavior phenomenon.

Since understanding the traffic dynamics and usage condition is of significance for improving network performance, the topic of traffic characteristics becomes a hot focus. In [5], the authors investigated three features of network traffic, namely network access time, traffic volume,

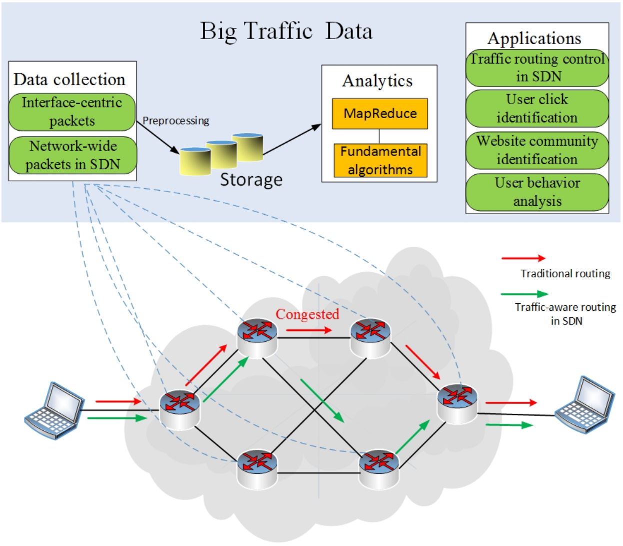
Big Traffic Data
Data collection
Interface-centric packets
Network-wide packets in SDN
Preprocessing
Storage
Analytics
MapReduce
Fundamental algorithms
Applications
Traffic routing control in SDN
User click identification
Website community identification
User behavior analysis
Traditional routing
Traffic-aware routing in SDN
Congested

FIGURE 3. Big traffic data in mobile cellular networks.

and diurnal patterns from the perspective of device models. The traffic characteristics from the perspective of service providers were revealed in [46]. Another angle from operating systems was introduced in [47]. All the above results are beneficial for cellular network operators to make corresponding adjustments for network capacity management and revenue growth.

C. BIG LOCATION DATA

Human activities are based on locations, and location data analysis is informative. As illustrated in Fig. 4, the location-based big data arising from GPS sensors, WiFi, bluetooth through mobile devices, have become precious strategic resources. These resources would provide support for government administration, such as public facility planning,

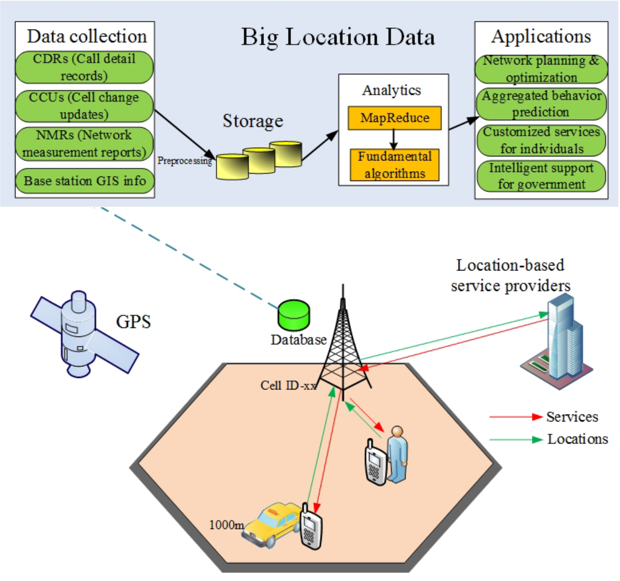
Data collection
CDRs (Call detail records)
CCUs (Cell change updates)
NMRs (Network measurement reports)
Base station GIS info
Big Location Data
Preprocessing
Storage
Analytics
MapReduce
Fundamental algorithms
Applications
Network planning & optimization
Aggregated behavior prediction
Customized services for individuals
Intelligent support for government
GPS
Database
Location-based service providers
Cell ID-xx
Services
Locations
1000m

FIGURE 4. Big location data in mobile cellular networks.

transportation system constructions, demographic trends, risk warnings for crowed people, rapid emergence responses, crime hot spots analysis, etc. It can also gain amazing business insights, such as mobile advertising and marketing [48]. In [49], an end-to-end Hadoop-based system was developed with a number of functional algorithms operated on call record details (CRDs). With the information about subscribers' habits and interests, it is capable of providing invaluable information about when, where and how a category of individuals (e.g., sports fans, music lover, et.) move.

D. BIG RADIO WAVEFORMS DATA

Zhang and Qiu [8] used large random matrices as building blocks to model the big data arising from a 5G massive MIMO system that is implemented using software-defined radios, as illustrated in Fig. 5. They exploited the fact that all data processing is done at CPU so all the modulated waveforms are stored at the RAMS or at the hard drives. On the other hand, big data analytics based on the random-matrix theory is applied to the collected data from their testbed, where a mobile user communicates with the massive MIMO base station while moving. The experimental results can estimate the user's moving speed, whether motionless, at a nearly constant speed, at a slow speed or at a higher speed. This analytics is also implemented to reflect the correlation residing in the transmitted signals. These applications validate the fact that the massive MIMO system is not only a communication system, but also a massive data platform which can brings tremendous values through big data analytics.

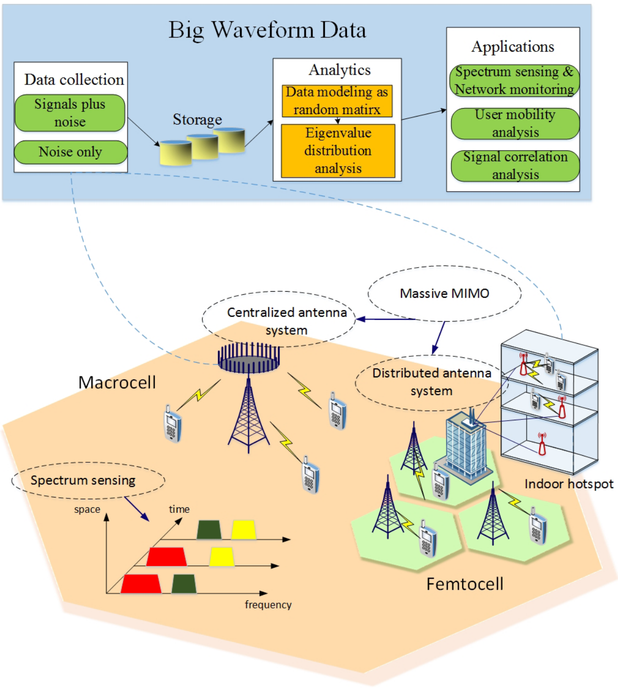

FIGURE 5. Big data captured for modulated radio waveforms in a massive MIMO cellular network.

E. BIG HETEROGENOUS DATA

One critical task of big data analytics in mobile cellular networks is the integration of very heterogenous data: correlation mining in massive database. Data sources are rich in types such as data rate, packet drop, mobility, etc. Different base stations host these data over time. They need be aggregated across space and time to obtain big data analytics. For example, for cyber security [50]–[54], there are many different heterogeneous sources, such as "numerous distributed packet sniffers, system log files, SNMP traps and queries, user profile databases, system messages, and operator commands." Essentially, data fusion is a technique to make overall sense of data from different sources that commonly have different data structures.

V. OPEN RESEARCH CHALLENGES

There are many open research challenges that are still not well studied and need to be tackled by future research efforts. We discuss some of these research challenges in this section.

From viewpoints of practice, privacy may be among the most important challenges. More advanced algorithms are needed to extract correlations from the data, while allowing different levels of privacy. For example, for mobile phones, data associated with bank transactions are highly private information, which should be carefully handled in big data analytics in mobile cellular networks. In addition, government and corporate regulations for privacy and data protection play a fundamental and necessary role in protecting the sensitive aspects of big data in mobile cellular networks.

How to filter out un-useful data is another significant challenge, due to the scarce bandwidth available in mobile cellular networks. Mobile cellular networks can produce staggering amounts of raw data, a lot of which are not of interest. It can be filtered out and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information.

Another challenge is to automatically generate the right metadata, to describe what data is recorded and how it is recorded and measured. This metadata is likely to be crucial to downstream analysis. Frequently, the information collected will not be in a format ready for analysis. We have to deal with erroneous data: Some quality reports for radio strengths, such as RSSI, frame error rate, and packet error rate, are inaccurate.

Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis, all of these have to happen in an *automated* manner. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big data computing environments. Today's analysts are impeded by a tedious process of exporting data from the database, performing a non-SQL process and bringing the data back. Having the ability to analyze big data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results.

VI. CONCLUSIONS AND FUTURE WORK

Big data analytics will be an indispensable part of the mobile cellular operators' consideration of network operation, business deployment, and even the design of the next-generation mobile cellular network architectures. In this paper, the connection between big data analytics and mobile cellular networks has been systematically explored. We provided a broad overview of big data analytics based on radom matrix theory. Next, an architectural framework for the applications of big data analytics in cellular networks was presented. Moreover, several illustrative examples were provided. Finally, we discussed some research challenges and big data analytics' prospects for next-generation cellular networks. Future work is in progress to address these challenges.

REFERENCES

[1] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5G mobile wireless networks," *IEEE Netw.*, vol. 29, no. 3, pp. 68–74, May/Jun. 2015.

[2] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, Mar. 2015.

[3] 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.

[4] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data," *IEEE Commun. Mag.*, vol. 53, no. 10, pp. 190–199, Oct. 2015.

[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] R. C. Qiu, Z. Hu, H. Li, and M. C. Wicks, *Cognitive Radio Communication and Networking: Principles and Practice*, (in Chinese). New York, NY, USA: Wiley, 2012.

[7] A. Guionnet, M. Krishnapur, and O. Zeitouni, "The single ring theorem," *Ann. Math.*, vol. 174, no. 2, pp. 1189–1217, 2011.

[8] C. Zhang and R. C. Qiu, "Massive MIMO as a big data system: Random matrix models and testbed," *IEEE Access*, vol. 3, no. 4, pp. 837–851, 2015.

[9] A. M. Khorunzhy, B. A. Khoruzhenko, and L. A. Pastur, "Asymptotic properties of large random matrices with independent entries," *J. Math. Phys.*, vol. 37, no. 10, pp. 5033–5060, 1996.

[10] J. Jacod and P. Protter, *Probability Essentials*, 2nd ed. New York, NY, USA: Springer, 2004.

[11] R. C. Qiu, "Large random matrices and big data analytics," in *Big Data of Complex Networks*. Boca Raton, FL, USA: CRC Press, 2016.

[12] R. C. Qiu and P. Antonik, *Smart Grid and Big Data*. New York, NY, USA: Wiley, May 2016.

[13] R. Qiu and M. Wicks, *Cognitive Networked Sensing and Big Data*. Berlin, Germany: Springer-Verlag, 2014.

[14] R. Qiu, *Principles of Massive Data Analysis: The Random Matrix Approach*, in preparation for publication.

[15] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, vol. 2. New York, NY, USA: Wiley, 1958.

[16] T. Tao, V. Vu, and M. Krishnapur, "Random matrices: Universality of ESDs and the circular law," *Ann. Probab.*, vol. 38, no. 5, pp. 2023–2065, 2010.

[17] S. Bu, F. R. Yu, and P. X. Liu, "A game-theoretical decision-making scheme for electricity retailers in the smart grid with demand-side management," in *Proc. IEEE SmartGridComm*, Brussels, Belgium, Oct. 2011, pp. 387–391.

[18] S. Bu, F. R. Yu, and P. X. Liu, "Dynamic pricing for demand-side management in the smart grid," in *Proc. IEEE Online Conf. Green Commun. (GreenCom)*, Sep. 2011, pp. 47–51.

[19] X. He, Q. Ai, R. C. Qiu, W. Huang, and L. Piao, "A big data architecture design for smart grids based on random matrix theory," *IEEE Trans. Smart Grid*, to be published. [Online]. Available: http://arxiv.org/pdf/1501.07329.pdf

[20] 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*, to be published, doi: 10.1109/TSG.2015.2508506.

[21] N. Guo, J. Aribido, and R. Qiu, "Estimation of interferer location and its transmit power using mobile user packet drop rates," *IEEE Trans. Veh. Technol.*, in preparation for publication.

[22] M. Shcherbina. (2011). "Central limit theorem for linear eigenvalue statistics of the Wigner and sample covariance random matrices." [Online]. Available: http://arxiv.org/abs/1101.3249.

[23] X. He, R. C. Qiu, L. Ai, Q. Ai, and X. Xu, "Linear eigenvalue statistics: An indicator set design for situation awareness of power systems," *IEEE Trans. Smart Grid*, to be published. [Online]. Available: http://arxiv.org/pdf/1512.07082.pdf

[24] U. Haagerup and S. Thorbjørnsen, "A new application of random matrices: Ext($C^*_{red}(F_2)$) is not a group," *Ann. Math.*, vol. 162, no. 2, pp. 711–775, Sep. 2005.

[25] M. Capitaine and C. Donati-Martin, "Strong asymptotic freeness for Wigner and Wishart matrices," *Indiana Univ. Math. J.*, vol. 56, no. 2, pp. 295–309, 2007.

[26] R. Vershynin. (Jul. 2011). "Introduction to the non-asymptotic analysis of random matrices." [Online]. Available: http://arxiv.org/abs/1011.3027.

[27] F. Lin, R. C. Qiu, Z. Hu, S. Hou, J. P. Browning, and M. C. Wicks, "Generalized FMD detection for spectrum sensing under low signal-to-noise ratio," *IEEE Commun. Lett.*, vol. 16, no. 5, pp. 604–607, May 2012.

[28] F. Lin, R. C. Qiu, and J. P. Browning, "Spectrum sensing with small-sized data sets in cognitive radio: Algorithms and analysis," *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 77–87, Jan. 2015.

[29] S. Hou and R. C. Qiu, "Kernel feature template matching for spectrum sensing," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2258–2271, Jun. 2014.

[30] J. Hoydis, R. Couillet, and P. Piantanida. (2013). "The second-order coding rate of the MIMO Rayleigh block-fading channel." [Online]. Available: http://arxiv.org/abs/1303.3400.

[31] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[32] E. Xing. (Dec. 2015). *CSDN Interview*, Carnegie Mellon University. [Online]. Available: http://www.cs.cmu.edu/~epxing/.

[33] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing [exploratory DSP]," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 145–154, Jan. 2011.

[34] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Comunication-efficient algorithms for statistical optimization," *J. Mach. Learn. Res.*, vol. 23, no. 1, p. 6792, 2012.

[35] J. Gao, Y. Guo, and Z. Wang. (2016). "Matrix neural networks." [Online]. Available: http://arxiv.org/abs/1601.03805.

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

[37] X. Meng *et al.* (2015). "MLlib: Machine learning in apache spark." [Online]. Available: http://arxiv.org/abs/1505.06807.

[38] *Wikipedia: The Free Encyclopedia*, Wikimedia Found., Inc., Jan. 2016.

[39] Y. Fei, V. W. S. Wong, and V. C. M. Leung, "Efficient QoS provisioning for adaptive multimedia in mobile communication networks by reinforcement learning," *Mobile Netw. Appl.*, vol. 11, no. 1, pp. 101–110, Feb. 2006.

[40] O. F. Celebi *et al.*, "On use of big data for enhancing network coverage analysis," in *Proc. ICT*, Casablanca, Morocco, May 2013, pp. 1–5.

[41] L. Cui, F. R. Yu, and Q. Yan, "When big data meets software-defined networking: SDN for big data and big data for SDN," *IEEE Netw.*, vol. 30, no. 1, pp. 58–65, Jan./Feb. 2016.

[42] D. Kreutz, F. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proc. IEEE*, vol. 103, no. 1, pp. 14–76, Jan. 2015.

[43] Q. Yan and F. Yu, "Distributed denial of service attacks in software-defined networking with cloud computing," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 52–59, Apr. 2015.

[44] Q. Yan, F. R. Yu, Q. Gong, and J. Li, "Software-defined networking (SDN) and distributed denial of service (DDoS) attacks in cloud computing environments: A survey, some research issues, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 602–622, Jan. 2016.

[45] Y. Cai, F. R. Yu, C. Liang, B. Sun, and Q. Yan, "Software defined device-to-device (D2D) communications in virtual wireless networks with imperfect network state information (NSI)," *IEEE Trans. Veh. Technol.*, to be published, doi: 10.1109/TVT.2015.2483558.

[46] L. Jun, L. Tingting, C. Gang, Y. Hua, and L. Zhenming, "Mining and modelling the dynamic patterns of service providers in cellular data network based on big data analysis," *China Commun.*, vol. 10, no. 12, pp. 25–36, Dec. 2013.

[47] J. Yang, S. Zhang, X. Zhang, J. Liu, and G. Cheng, "Characterizing smartphone traffic with MapReduce," in *Proc. IEEE WOCC*, Jun. 2013, pp. 1–5.

[48] L. Deng, J. Gao, and C. Vuppalapati, "Building a big data analytics service framework for mobile advertising and marketing," in *Proc. IEEE 1st Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar. 2015, pp. 256–266.

[49] V. Kolar *et al.*, "People in motion: Spatio-temporal analytics on call detail records," in *Proc. IEEE COMSNETS*, Jan. 2014, pp. 1–4.

[50] R. Zuech, T. Khoshgoftaar, and R. Wald, "Intrusion detection and big heterogeneous data: A survey," *J. Big Data*, vol. 2, no. 3, pp. 1–41, Dec. 2015.

[51] S. Bag, "A new key predistribution scheme for grid-group deployment of wireless sensor networks," *Ad Hoc Sensor Wireless Netw.*, vol. 27, nos. 3–4, pp. 313–329, 2015.

[52] F. R. Yu, H. Tang, P. C. Mason, and F. Wang, "A hierarchical identity based key management scheme in tactical mobile ad hoc networks," *IEEE Trans. Netw. Service Manage.*, vol. 7, no. 4, pp. 258–267, Dec. 2010.

[53] D. M. B. Ying and H. T. Mouftah, "Sink privacy protection with minimum network traffic in WSNs," *Ad Hoc Sensor Wireless Netw.*, vol. 25, nos. 1–2, pp. 69–87, 2015.

[54] S. Bu, F. R. Yu, X. P. Liu, and H. Tang, "Structural results for combined continuous user authentication and intrusion detection in high security mobile ad-hoc networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 3064–3073, Sep. 2011.

YING HE received the B.S. degree in communication and information systems from Dalian Ocean University, Dalian, China, and the M.S. degree in communication and information systems from the Dalian University of Technology, Dalian, in 2007 and 2012, respectively, where she is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering. She is a Visiting Student with Carleton University. Her current research interests include big data, cognitive radio networks, and machine learning.

FEI RICHARD YU (S'00–M'04–SM'08) received the Ph.D. degree in electrical engineering from the University of British Columbia (UBC), in 2003. From 2002 to 2006, he was with Ericsson, Lund, Sweden, and a start-up in CA, USA. He joined Carleton University in 2007, where he is currently an Associate Professor. He received the IEEE Outstanding Leadership Award in 2013, the Carleton Research Achievement Award in 2012, the Ontario Early Researcher Award (formerly Premiers Research Excellence Award) in 2011, the Excellent Contribution Award at the IEEE/IFIP TrustCom 2010, the Leadership Opportunity Fund Award from the Canada Foundation of Innovation in 2009, and the best paper awards at the IEEE ICC 2014, Globecom 2012, the IEEE/IFIP TrustCom 2009, and the International Conference on Networking 2005. His research interests include cross-layer/cross-system design, security, green IT, and QoS provisioning in wireless-based systems.

He serves on the Editorial Boards of several journals, including the Co-Editor-in-Chief of *Ad Hoc & Sensor Wireless Networks*, a Lead Series Editor of the IEEE Transactions on Vehicular Technology, the IEEE Communications Surveys & Tutorials, the *EURASIP Journal on Wireless Communications Networking*, the *Wiley Journal on Security and Communication Networks*, and the *International Journal of Wireless Communications and Networking*. He has served as the Technical Program Committee Co-Chair of numerous conferences. He is a Registered Professional Engineer in the province of Ontario, Canada.

NAN ZHAO (S'08–M'11) received the B.S. degree in electronics and information engineering, the M.E. degree in signal and information processing, and the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology, Harbin, China, in 2005, 2007, and 2011, respectively. From 2011 to 2013, he did post-doctoral research with the Dalian University of Technology, Dalian, China. He is currently an Associate Professor with the School of Information and Communication Engineering, Dalian University of Technology. He has authored nearly 70 papers in refereed journals and international conferences. His recent research interests include interference alignment, cognitive radio, wireless power transfer, and optical communications.

Dr. Zhao is a Senior Member of the Chinese Institute of Electronics. He serves as an Editor of *Wireless Networks*, the *AEU International Journal of Electronics and Communications*, *Ad Hoc & Sensor Wireless Networks*, and *KSII Transactions on Internet and Information Systems*. In addition, he served as a Technical Program Committee Member for many interferences, e.g., Globecom, VTC, and WCSP. He is also a Peer-Reviewer for a number of international journals, such as the IEEE IEEE Transactions on Communications, the IEEE Transactions on Wireless Communications, and the IEEE Transactions on Vehicular Technology.

HONGXI YIN received the B.Sc. degree from Shandong University, Jinan, China, in 1982, the M.Eng. degree from the Harbin Institute of Technology, Harbin, China, in 1988, and the Ph.D. degree from Zhongshan University, Guangzhou, China, in 1998. He finished his post-doctoral research work with Peking University, Beijing, China, in 2000. From 2005 to 2007, he was a Visiting Research Fellow with the Optoelectronics Research Centre, University of Southampton, Southampton, U.K. He was with the School of Electronics Engineering and Computer Science, Peking University, as an Associate Professor from 2000 to 2008 and then joined the School of Information and Communication Engineering, Dalian University of Technology, as a Professor. He has authored or co-authored about 80 national and international journal papers and has obtained several patents. His recent activities are in the field of optical fiber communication technologies, all-optical code-division multiple access networks and photonic signal processing, optical cross-connect and wavelength division multiplexing all-optical networking, optical packet switching, automatic switched optical network, and interference alignment in cognitive radio.

Prof. Yin is a Senior Member of the Chinese Institute of Electronics.

HAIPENG YAO received the Ph.D. degree with the Department of Telecommunication Engineering, University of Beijing University of Posts and Telecommunications, in 2008, and the M.S. degree in wireless communications from the Beijing University of Posts and Telecommunications, in 2006. He is a Lecturer with the Beijing University of Posts and Telecommunications. His main research interests are in the area of big data, future internet architecture, cognitive radio networks, and optimization of protocols and architectures for broadband wireless networks.

ROBERT C. QIU has nearly 20 years of teaching and research in academia, industry, and startup with diverse research experience in wireless communications and networks, wireless (and remote) sensing, big data, and smart grid. His professional experiences outside the academia include GTE Labs (now Verizon Wireless), Bell Labs (Lucent Technologies), and the startup. He served as the Founder, CEO, and President for Wiscom Technologies Inc., that grew to a total of over 30 staffs and whose assets were sold to Intel. He is a U.S. citizen (since 2001).

He joined Tennessee Technological University (TTU) in 2003. In 2008, he was tenured and promoted to full Professor at TTU. He was also the Principal Investigator for a Congressional Earmark Project that has the planned budget of U.S. $5 million in three phases, although only the first phase was actually funded due to the change of policy in U.S. Congress. He served as the Founding Coordinator for two college-level focus areas, such as smart grid and big data. He has attracted a total funding over U.S. $3 million from diverse sources, including NSF, ARO, ONR, AFOSR, and AFRL. He spent four summers in federal defense laboratories (AFRL and NRL).

Before coming to TTU, he spent three years to found his startup (as the CEO and President). Prior to his startup experience, he spent nearly five years in industrial labs (GTE Labs and Bell Labs), to conduct applied research in wireless cellular industry. In GTE Labs, he participated in the first CDMA field trial jointly conducted by AT & T, GTE, and Qualcomm. At Bell Labs, he was one of the pioneers in designing the core physical layer system algorithms for the 3G WCDMA base transceiver station that turns into multibillion dollars business later for Lucent.