*International Journal of Research in Advent Technology, Vol.5, No.12, December 2017*
*E-ISSN: 2321-9637*
*Available online at www.ijrat.org*

# A Novel Approach for Big Data Analytics in Mobile Networks

Ms. Mrudula Gadkari, Prof. Vikas Nandgaonkar

**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 signalling 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.

**Keywords:** mobile, networks, big data, analytics

## 1. INTRODUCTION

Recent years have witnessed tremendous advances in wire-less cellular networks. With recent advances of wireless technologies and ever-increasing mobile applications, mobile cellular networks have become both generators and carriers of massive data. 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. Traditional data analytics shows its in-adequateness when encountered with the big cellular data. First, traditional data analytics deals with structured data.

## 2. LITERATURE SURVEY

Title: Information-centric network function virtualization over 5G mobile wireless networks, Year: 2013

Authors: Ming, M., G. Jing, and C. Jun- jie. Blast-Parallel

Wireless network virtualization and information-centric networking (ICN) are two promising techniques in software-defined 5G mobile wireless networks. Traditionally, these two technologies have been addressed separately. In this paper we show that integrating wireless network virtualization with ICN techniques can significantly improve the end-to-end network performance. In particular, we propose information-centric wireless network virtualization architecture for integrating wireless network virtualization with ICN. We develop the key components of this architecture: radio spectrum resource, wireless network infrastructure, virtual resources (including content-level slicing, network-level slicing, and flow-level slicing), and information-centric wireless virtualization controller. Then we formulate the virtual resource allocation and in-network caching strategy as an optimization problem, considering the gain of not only virtualization but also in-network caching in our proposed information-centric wireless network virtualization architecture. The obtained simulation results show that our proposed information-centric wireless network virtualization architecture and the related schemes significantly outperform the other existing schemes. Another new technology, called information-centric networking (ICN), has attracted great interests from both academia and industry [5]. The basic principle behind ICN is to promote the content to a first-class citizen in the network. A significant advantage of ICN is to provide native support for scalable and highly efficient content retrieval while enabling the enhanced capability for mobility and security. ICN can realize in-network caching to reduce the duplicate content transmission in networks. The ICN-based air caching technique has been recognized as one of the promising-candidate techniques to efficiently implement the SDN-based 5G wireless networks [6]. A number of research efforts have been dedicated to ICN, including the EU funded project Publish-Subscribe Internet Technology (PURSUIT) and the US funded project Named Data Networking (NDN). Although some excellent works have been done on wireless network virtualization and ICN, these two important areas have traditionally been addressed separately in the literature

**Advantages:** Big data is very difficult to process and store. Mainly Hadoop is used to process the big data.

Hadoop used HDFS to store the data efficiently and Map Reduce frame work for processing the data. MPI is also used to process the big data.

**Dis Advantages:** All data which is not structured and is in free format is unstructured. In fact, most individuals and organizations achieve their lives around free data.

Title: Cloud computing and the DNA data race, Year:2011

Authors: Schatz, M.C., B. Langmead, and S.L. Salzberg

Hadoop map/reduce is a parallel processing framework. Whenever any data is put on HDFS, data is divided into blocks with bock size of 128 MB. Name node stores the metadata for every data. The resources are managed by name node for data storage and resource manager manages processing on data nodes. After successful distribution of data on HDFS whenever any job is submitted by the user to process the stored data, job is submitted to the resource manager. Resource manager asks name node for the metadata of the data which is to be processed. And job is divided into tasks that are Mappers and Reducer. So the status of the whole job is monitored by Resource manager while status of the Mappers and Reducers is taken care by Node manager.

**Advantages:** This layer provides distributed storage for big data across the cluster of nodes. For reliable data storage it also provides replication of each block.

**Disadvantage:** It is a layer on top of HDFS which provides resource management and scheduling. On master node Resource manager is the daemon which is responsible for YARN and on worker nodes Node.

**Title :** Challenges and Opportunities with Big Data, **Year:** 2012

**Author:** Alexandros Labrinidis, H. V. Jagadish

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of "Big Data," including the recent announcement from the White House about new funding initiatives across different agencies, that target research for Big Data. This paper defines big data analytics and its characteristics, comments on its advantages and challenges in health care. Submissions purely focusing on the topics centered in some other sister IEEE Transactions, such as core machine learning theory, pattern recognition, image processing, computer vision, neural networks, and fuzzy systems, will not be considered. This transfer and transformation of problem-solving expertise from a knowledge source to a program is the heart of the expert-system development process. Building a KBS means building a computer model with the aim of realizing problem-solving capabilities comparable to a domain expert. While the promise of Big Data is real for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 – there is no clear consensus on what is Big Data. In fact, there have been many controversial statements about Big Data, such as "Size is the only thing that matters." In this panel we will try to explore the controversies and debunk the myths surrounding Big Data.

**Advantages**: Effective large-scale analysis all of this has to happen in a completely automated manner. Redundancy can be explored to compensate for missing data.

**Disadvantages:** Machine analyses algorithms expect homogeneous understand guidance.

Title: A probabilistic approach to mining mobile phone data sequences, Year: 2011

Authors: Farrahi, K. and D. Gatica – Perez

We present a new approach to address the problem of large sequence mining from big data. The particular problem of interest is the effective mining of long sequences from large-scale location data to be practical for Reality Mining applications, which suffer from large amounts of noise and lack of ground truth. To address this complex data, we propose an unsupervised probabilistic topic model called the distant n-gram topic model (DNTM). The DNTM is based on Latent Dirichlet Allocation (LDA), which is extended to integrate sequential information. We define the generative process for the model, derive the inference procedure, and evaluate our model on both synthetic data and real mobile phone data. We consider two different mobile phone datasets containing natural human mobility patterns obtained by location sensing, the first considering GPS /wifi locations and the second considering cell tower connections. With virtualization, physical cellular network infrastructure resources and physical radio resources can be abstracted and sliced into virtual cellular network resources holding certain corresponding functionalities, and shared by multiple parties through isolating each other. In other words, virtualizing mobile cellular networks is to realize the process of abstracting, slicing, isolating, and sharing mobile cellular networks. Generally speaking, the physical resources in cellular networks consist of licensed spectrum resource and infrastructure resources, including radio access networks (RANs), core networks (CNs), and transport networks. As shown in Fig. 1a, two logical roles can be identified after virtualization mobile network operator (MNO) and service provider (SP). MNOs own and operate infrastructures and radio resources of physical

substrate wireless networks, including licensed spectrum, RANs, backhaul, transmission networks, and CNs. MNOs implement the virtualization, and slice the physical mobile network resources into virtual mobile network resources. For brevity, we use virtual resources to indicate the virtual mobile network resources. SPs lease, operate, and program these virtual resources to offer end-to-end services to mobile users. The roles in the business model can be further decoupled into more specialized roles, including SP, infrastructure provider (InP), and mobile virtual network operator (MVNO) [7], as shown in Fig. 1b. Their functions in this model are detailed as follows.

**Advantage:** There are several difficulties to modelling human activities, including various types of uncertainty, lack of ground truth, complexity due to the size of the data, and diversity of phone users. One fundamental issue motivating this work is that we often do not know.

**Disadvantage:** We focus on probabilistic topic models as the basic tool for routine analysis for several reasons. Topic models are, first and foremost, unsupervised in nature.

**Title :** The single ring theorem, **Year:** 2012

**Author:** Chanchal Yadav, Shuliang Wang, Manoj Kumar

Data mining environment produces a large amount of data, that need to be analyzed, patterns have to be extracted from that to gain knowledge. In this new era with boom of data both structured and unstructured, in the field of genomics, meteorology, biology, environmental research and many others, it has become difficult to process, manage and analyze patterns using traditional databases and architectures. So, a proper architecture should be understood to gain knowledge about the Big Data. This paper presents a review of various algorithms from 1994-2013 necessary for handling such large data set. This paper defines big data analytics and its characteristics, comments on its advantages and challenges in health care. Submissions purely focusing on the topics centered in some other sister IEEE Transactions, such as core machine learning theory, pattern recognition, image processing, computer vision, neural networks, and fuzzy systems, will not be considered. This transfer and transformation of problem-solving expertise from a knowledge source to a program is the heart of the expert-system development process. Building a KBS means building a computer model with the aim of realizing problem-solving capabilities comparable to a domain expert These algorithms define various structures and methods implemented to handle Big Data, also in the paper are listed various tool that were developed for analyzing them.

**Advantages:** We find Associations, patterns and to analyze the large data sets. Different methodologies associated with different algorithms used to handle such large data sets Of the data can be used to classify it as high.

**Disadvantages:** It also describes about the various security issues, application and trends followed by a large data set.

**Title:** Matrix neural networks, **Year:** 2015
**Author**: Gemson Andrew Ebenezer J and Durga S

In some regression and classification problems (for example image, textual data, multi -dimensional time series analysis) we need to operate with matrices. One of standard general approaches is to decompose the input matrix into the vector and work with it. Such decomposition has two disadvantages. It can remove an important information about an inner structure of the input matrix. In most cases, when the inputs are matrices, they are relatively high dimensional. If dimensionality of the input is high and cardinality of the training set is relatively low, we can face a small training sample problem. In such cases we need other techniques to deal with the matrix inputs. The major contributions of this article are as follows: We propose an information-centric wireless network virtualization architecture that can enable both wireless network virtualization and ICN in 5G mobile wireless networks. We define and develop the key components of this architecture: radio spectrum resource, wireless network infrastructure, virtual resources (including content-level slicing, network-level slicing, and flow-level slicing), and an information-centric wireless virtualization controller. We formulate the virtual resource allocation and in-network caching strategies as a joint optimization problem, taking into account the gains of not only virtualization but also in-network caching in the proposed information-centric wireless network virtualization architecture. Simulation results are presented to validate and evaluate the performance of our proposed architecture and schemes. The rest of this article is organized as follows. The following section introduces wireless network virtualization and information-centric networking. Then we propose the architecture of information-centric wireless network virtualization. Following that we formulate the virtual resource allocation and in-network caching strategy. Then we evaluate our proposed scheme through simulations. The final section concludes this article and briefly discusses the future work. This paper defines big data analytics and its characteristics, comments on its advantages and challenges in health care. Submissions purely focusing on the topics

centered in some other sister IEEE Transactions, such as core machine learning theory, pattern recognition, image processing, computer vision, neural networks, and fuzzy systems, will not be considered. This transfer and transformation of problem-solving expertise from a knowledge source to a program is the heart of the expert-system development process.
**Advantage:** Building a KBS means building a computer model with the aim of realizing problem-solving capabilities comparable to a domain expert.
**Dis advantages:-**This paper also elaborates various platforms and algorithms for big data analytics and discussion on its advantages and challenges. This survey winds up with a discussion of challenges and future.

Title: The big challenges of big data, Year: 2013

Authors: Marx, V., Biology

Big data is defined as large amount of data which requires new technologies and architectures to make possible to extract value from it by capturing and analysis process. New sources of big data include location specific data arising from traffic management, and from the tracking of personal devices such as Smartphones. Big Data has emerged because we are living in a society which makes increasing use of data intensive technologies. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Since Big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are need to be understood. Big Data concept means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concepts & tools. The difficulties can be related to data capture, storage, search, sharing, analytics and visualization etc. Figure 1 : Example of Big Data Architecture Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges.

**Advantage:-** The various challenges faced in large data management include – scalability, unstructured data, accessibility, real time analytics, fault tolerance and many more.

**Disadvantage:-** In addition to variations in the amount of data stored in different sectors, the types of data generated and stored.

**Title :** Analysis of Diabetic Data Set Using Hive and R, **Year:** 2014
**Author:** Sadhana, Savitha Shetty

Modern medicine generates a great deal of information which is deserted in to the medical database. A proper analysis of such information may reveal some interesting facts, which may otherwise be hidden or go dissipate. Data mining is one such field which tries to extract some interesting facts from huge data set. This paper defines big data analytics and its characteristics, comments on its advantages and challenges in health care. Submissions purely focusing on the topics cantered in some other sister IEEE Transactions, such as core machine learning theory, pattern recognition, image processing, computer vision, neural networks, and fuzzy systems, will not be considered. This transfer and transformation of problem-solving expertise from a knowledge source to a program is the heart of the expert-system development process. Building a KBS means building a computer model with the aim of realizing problem-solving capabilities comparable to a domain expert In this paper an attempt is made to analyses the diabetic data set and derive some interesting facts from it which can be used to develop the prediction model.
**Advantages:** This is based on the security and efficiency analysis. Users share their attributes among a group of valid users
**Disadvantages**: Physicians cannot accept or utilize the records without an official cerification.

## 3. EXISTING SYSTEM

Hadoop is the developing & processing data with accessing easily. Here Big data is not fully developed and do not uncover to calculate performance and huge amount of data. 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. 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 Multiple attributes based Map and Reduce process not fully efficiently developed at system.
**Disadvantages:**
Not support big data at all applications. It can't handle Multiple Datasets.
Big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security.

## 4. PROPOSED SYSTEM

Hadoop data with Processing transactions and Enterprises needs tools to help them in understanding and analysing healthcare data easily and effectively. a unied data model based on the random matrix theory and machine learning. It is a form of distributed computing whereby resources and application platform are shared over the internet through on

demand and pay on utilization basis. Several companies have already built Internet consumer services such as search engine, use of some websites to communicate with other user in websites, E-mail services, and services to purchase items online that use cloud computing infrastructure. Here processing data transactions and certain methodologies based to be calculate the finalize the performance of Hadoopdata. A MapReduce job is an access and process- streaming job that splits the input dataset into independent chunks (blocks) and stores them in HDFS. Native Hadoop compiler processes MapReduce job by dividing the job into multiple tasks, then distributes these tasks to multiple nodes in the cluster. we will propose our enhanced Hadoop MapReduce workflow and compare the two architectures in terms of developing MapReduce performance.

**Advantages:-**

- Effective large-scale analysis all of this has to happen in a completely automated manner.
- Redundancy can be explored to compensate for missing data.

**4.1 System Architecture**

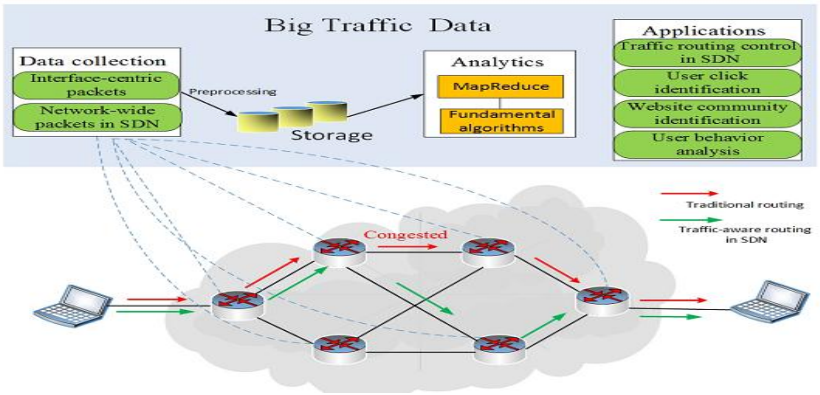



Figure 1 Syste Architecture

**Flow Chart**

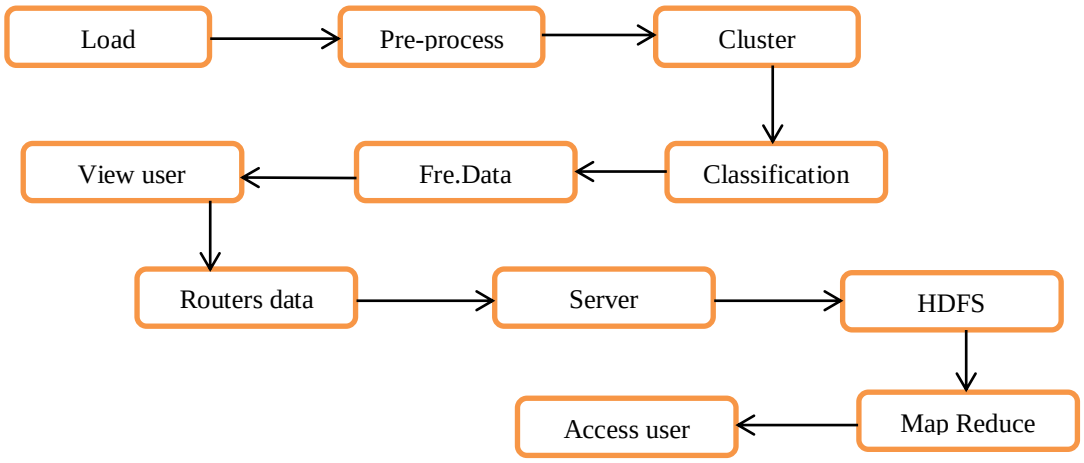



Figure 2 Flow chart of porposed system

**4.2 Modules**

1. Load Dataset
2. Data Clustering
3. Node Selection
4. Analysing Hadoop Performance

**Modules Description:**

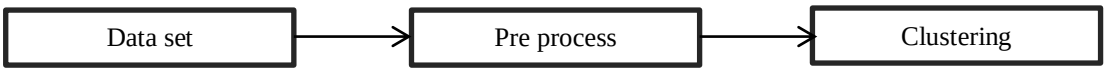
1. **Load Dataset:**

A data set or dataset, although this spelling is not present in many contemporary dictionaries is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in

question. Here Multiple attributes based create dataset and respective to the real world data transmissions with Processing data.

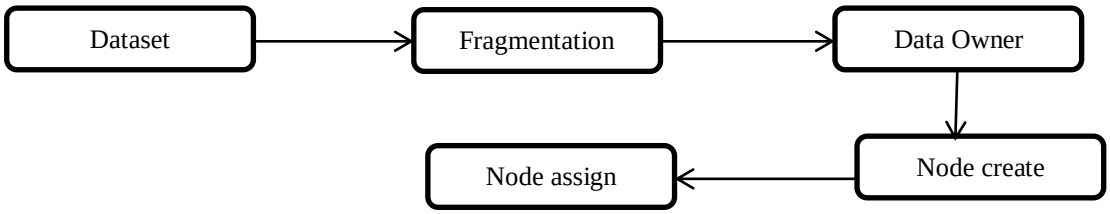

**2. Data Clustering:**

It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results.



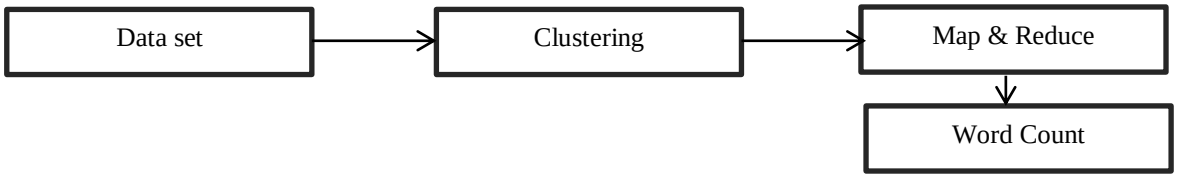

**3. Node Selection**

In the second iteration that node is selected that produces the lowest RC in combination with node already selected. The process is repeated for all of the file fragments. The centrality measure is the same for all of the nodes. This results in the selection of same node for storing the file fragment. Consequently, the performance showed the same value and all three lines are on the same points. However, this is not the case for the DCell architecture.




**4. Hadoop Performance:**

Hadoop is a framework which enables applications to work on petabytes of data on large clusters with thousands of nodes built of commodity hardware. Hadoop cluster in production is just half the battle won. It is extremely important for a Hadoop admin to tune the Hadoop cluster setup to gain maximum performance. During Hadoop installation, the cluster is configured with default configuration settings which are on par with the minimal hardware configuration. Hadoop with certain data as big data with processed on Data set creation based on performance analyzed.




5. SYSTEM IMPLEMENTATION

Implementation of software refers to the final installation of the package in its real environment, to the satisfaction of the intended users and the operation of the system. The people are not sure that the software is meant to make their job easier.

- The active user must be aware of the benefits of using the system
- Their confidence in the software built up
- Proper guidance is impaired to the user so that he is comfortable in using the application

Before going ahead and viewing the system, the user must know that for viewing the result, the server program should be running in the server. If the server object is not running on the server, the actual processes will not take place.

**5.1 User Training**

To achieve the objectives and benefits expected from the proposed system it is essential for the people who will be involved to be confident of their role in the new system. As system becomes more complex, the need for education and training is more and more important. Education is complementary to training. It brings life to formal training by explaining the background to the resources for them. Education involves creating the right atmosphere and motivating user staff. Education information can make training more interesting and more understandable.

**5.2 Training on the Application Software:**

After providing the necessary basic training on the computer awareness, the users will have to be trained on the new application software. This will give the underlying philosophy of the use of the new system such as the screen flow, screen design, type of help on the screen, type of errors while entering the data, the corresponding validation check at each entry and the ways to correct the data entered. This training may be different across different user groups and across different levels of hierarchy.

## 6. CONCLUSION & FUTURE SCOPE

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 analystics and mobile cellular net-works has been systematically explored.

**Future Scope:**

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. Here we combined different approaches and domains based improving performance and Security.

## 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. 6874, 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. 358380, 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. 3239, 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. 190199, 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. 18841897, 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] Guionnet, M. Krishnapur, and O. Zeitouni, ``The single ring theorem," Ann. Math., vol. 174, no. 2, pp. 11891217, 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. 837851, 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. 50335060, 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.