Journal of Communications and Information Networks, Vol.2, No.1, Mar. 2017
DOI: 10.1007/s41650-017-0001-2

Review paper

Special Issue on Wireless Big Data

Survey of wireless big data

Lijun Qian[1], Jinkang Zhu[2], Sihai Zhang[3]*

1. CREDIT Research Center, Prairie View A&M University, Texas A&M University System, Prairie View TX 77446, USA
2. PCNSS, University of Science and Technology of China, Hefei 230017, China
3. Key Laboratory of Wireless-Optical Communications, Chinese Academy of Sciences, University of Science and Technology of China, Hefei 230017, China

* Corresponding author, email: shzhang@ustc.edu.cn

Abstract: Wireless big data describes a wide range of massive data that is generated, collected and stored in wireless networks by wireless devices and users. While these data share some common properties with traditional big data, they have their own unique characteristics and provide numerous advantages for academic research and practical applications. This article reviews the recent advances and trends in the field of wireless big data. Due to space constraints, this survey is not intended to cover all aspects in this field, but to focus on the data aided transmission, data driven network optimization and novel applications. It is expected that the survey will help the readers to understand this exciting and emerging research field better. Moreover, open issues and promising future directions are also identified.

Keywords: wireless big data, data driven wireless networks, data aided network optimization

Citation: L. J. Qian, J. K. Zhu, S. H. Zhang. Survey of wireless big data [J]. Journal of communications and information networks, 2017, 2(1): 1-18.

1 Introduction

Wireless big data describes a wide variety of data sets of high technological and intellectual value, just as big data in other fields have demonstrated so far. For example, the wireless signaling data can describe the network deployment and service quality; the call detail records may reveal both the social network structure among the users and the behavior of the users; and the spatial-temporal location data could aid the potential commercial improvement.

The above examples represent just a few of the many cases that have recently prompted the research community to investigate the theory and methods in wireless big data. In the past several years, researchers all around the world have published many interesting works, including novel insights on big data for urban traffic analysis and planning, wireless network optimization using massive data sets, wireless user behavior modeling and so on. In addition,

Manuscript received Jan. 10, 2017; accepted Feb. 19, 2017
This research work is supported in part by the U.S. OASD (R&E) (Office of the Assistant Secretary of Defense for Research and Engineering) (No. FA8750-15-2-0119), and by the U.S. Army Research Office(No. W911NF-16-1-0496). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Office of the Assistant Secretary of Defense for Research and Engineering (OASD (R&E)), the Army Research Office, or the U.S. Government. Sihai Zhang received support from Key Program of National Natural Science Foundation of China(No. 61631018), the Fundamental Research Funds for the Central Universities and Huawei Technology Innovative Research on Wireless Big Data.

China NSFC has launched three five-year research projects on wireless big data, aiming to shed insight on some, if not all, topics in this field.

The rise of dramatic trends in this direction in the past few years is stimulated by several parallel developments. Firstly, the computerization of data acquisition in telecommunication operators led to the emergence of large databases on user behavior and network behavior. Secondly, the increased research success of big data in other fields encouraged us to investigate the wireless related big data analysis, to attempt to promote the transmission and to optimize the network performance. Finally, the challenges faced by current 5G and future wireless communications push us to seek innovative solutions, such as exploring the computation dimension or unified computation and communication.

Although the progress in the field of computer science especially in data mining and machine learning has led to many success stories in big data research, such as the new applications and services provided by companies like Google or iFLYTEK, many challenges still exist in wireless big data research. For instance, the stochastic nature of wireless channels and related modulation/demodulation, as well as the behavior of the wireless users, create very dynamic data sets.

There are already several review papers on specific topics in this field. The authors in Ref. [1] summarized the analysis based on mobile phone datasets, including the social networks that can be constructed with such data, the study of personal mobility, geographical partitioning, urban planning, facilitating development and security and privacy issues. The authors in Ref. [2] carried out a specific but comprehensive survey on GPS mining for mobility patterns, thus providing a general perspective for studies, by reviewing the methods and algorithms in detail and comparing the existing results on the same issues.

This survey aims to introduce and discuss the recent progress in wireless big data, ranging from fundamental concepts and notations, data collection and storage, transmission technology, and network layer related topics and applications. However, due to space constraints, this survey does not cover all the important topics, and we still try our best to provide the readers with an integrated research framework of wireless big data.

The contents of this survey are organized as depicted in Fig. 1. We separate the contents into four layers: data, transmission, network, and application layers, from bottom to top. In the data layer, we first introduce two existing concepts about wireless big data, and propose the purpose oriented notation. Thereafter we discuss the data collection techniques, data model and data analytics. In the transmission layer, the progress on spectrum big data and multiple user access using data analysis is reviewed. In the network layer, we select the three topics that have been focused on the most: network architecture design, traffic analysis and network planning. However we neglect topics such as data driven handover mechanism. There are quite a number of fruitful works in application layers, so we pick up two key aspects, user mobility analysis and social network analysis from the physical spatial domain and logical social domain, respectively. In addition, we also discuss three potential application areas: smart grids, IoT (Internet of Things) and Drones/UAV (Unmanned Aerial Vehicle).

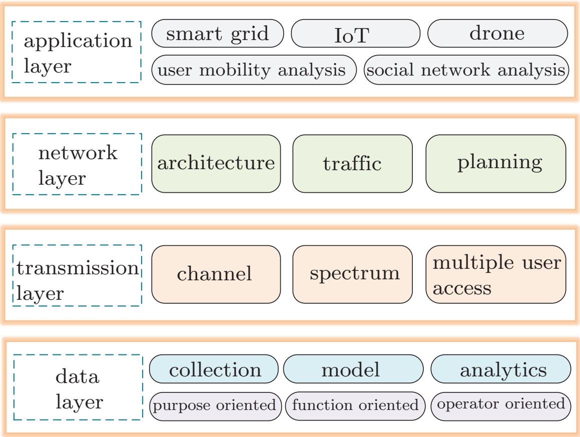


Figure 1 Research framework of wireless big data

The rest of this article is organized as follows. Section 2 introduces certain concepts and categories of wireless big data and then reviews the data collection, data model and data analytics. In section 3, two wireless transmission related topics are presented. Wireless network layer related architecture design, traffic analysis and network optimization are presented in section 4. Section 5 summarizes the five key aspects of wireless big data applications. The privacy and security of wireless big data is discussed in section 6. Finally, the conclusion and open issues are discussed in section 7.

2 Data layer related

In this section, we first discuss the concepts of wireless big data, and then introduce the data collection, model and analytics.

2.1 Concepts and categories

Concrete and precise definitions are always the first steps to understanding our world. However, as for the wireless big data, we have not yet established a well-acknowledged concept. Therefore, it is essential to begin by reviewing the definition of big data first. By now we are familiar with the 4Vs of big data: volume, velocity, variety and veracity, which evaluate the big data from the size of data set, speed of data in and out, range of data type and source and quality of data. However, despite all these features, wireless big data is usually considered as the data set that cannot be transmitted, accessed, processed and served in an endurable time period by existing communication and network systems.

From the perspective of telecommunication operators, the data is mainly derived from the following three aspects[3]: data in the IT system: user attributes, business consumption information, terminal information and etc. Moreover, these data are collected from CRM (Customer Relationship Management), billing system and terminal self-registration platform, respectively. Basic user portraits and characteristics can be described in accordance with these data. Data in the access network and core network: mobile signaling, DPI, M2M data, etc. These data are collected in wired/wireless networks when clients make use of voice, SMS or networking services. The underlying structure of the data is complex, therefore targeted analysis and processes are required for different types of data, in order to achieve scenario-based descriptions of users locations and preferences. Data in the Internet applications of operators: online business hall data, palm business office data, wing payment data, etc. All the data including the user access mode, address, time, business preference, investment and consumption habits are completely preserved in the background of the application which can be easily obtained.

Zhang et al.[4] divided the data in mobile cellular networks into flow record data, network performance data, mobile terminal data and additional data, mainly from the perspective of potential applications. Firstly, the flow record data in cellular networks may be the most important data for describing wireless user behavior, which includes both data records and signaling records, in the form of XDR (call/transaction Detail Records) and contains the main attributes during a data-connection session. Secondly, the network performance data, as coined, aims to evaluate the network performance and quality of service delivered to wireless users, which mainly includes the KPI (Key Performance Indicator) data and the MR (Measurement Report; a statistical data report that contains information about channel quality). Finally, mobile terminal data can be gathered through mobile app and contains device information, wireless parameters and so on.

In this survey, from a network socio-ecological perspective, we further categorize the wireless big data into three types: primordial wireless big data, derived wireless big data, developing wireless big data. Firstly, the primordial wireless big data denotes the data set generated by massive wireless users served by wireless/mobile communications, which includes wireless access behavior, wireless application requirements and etc. Secondly, derived wireless big data represents the spectrum, transmission, access, and network data developed and produced to provide effective communication service for wireless users. These data include the distribution of spectrum utilization, spatial statistics of ultra-densely deployed cells and resource allocation of transmission signals. Finally, developing wireless big data means the data set that are generated during the processes of testing and evaluating the performance of unknown spectrum, novel transmission techniques, innovative access and revolutionary network structure.

Here, we point out that the wireless big data can also be categorized according to their specific areas, which include cellular networks, Wi-Fi hotspots and smartphones D2D, smart grids, wireless sensor networks, IoT, etc.

2.2 Data collection

Data collection is in some sense an engineering oriented problem, which is mostly concerned with telecommunication operators, although their purposes are not for wireless big data research. However, several research works have been recently released on this topic.

As for the data gathering challenge in gathering real-time big data in a complex indoor industrial environment[5], a RTBDG (Real-Time Big Data Gathering) algorithm based on an indoor WSN is proposed, where sensor nodes can screen the data collected from the environment and equipment according to the requirements of risk analysis, which may be widely applied to risk analysis in different industrial operations.

Another interesting point in this topic is based on compressive sensing[6]. The authors attempted to deal with the shortage of energy in wireless sensor nodes, and proposed a compressive-sensing-based collection framework to minimize the amount of collection while maintaining data quality.

2.3 Data model

The random matrix theory model is applied to represent varying amounts of data collected from multiple sources. In Ref. [7], a big data analytic unified data model based on the random matrix theory and machine learning in mobile cellular networks is studied. Several examples of data types have been presented to clarify the performance of big data analytic based on random matrix theory, such as big signaling data, big traffic data, big location data, big radio waveforms data, and big heterogeneous data, in which the high dimensionality of the spatial-temporal datasets is exploited, and the interrelationship and unique characteristics between big data and mobile cellular networks is addressed. Moreover, in Ref. [8], the large-scale random matrices are introduced as building blocks to model the massive big data collected by the massive MIMO (Multiple Input Multiple Output) system, and forwarded to the base station for processing and storage. This model is applied to distributed spectrum sensing and network monitoring. The software defined radio platform, equipped with USRP (Universal Software Radio Peripheral), is used to emulate the antenna in the base station and demonstrate the data processing in the CPU.

Large-scale data and heterogeneous data may be the unique characteristics of wireless big data simply as variety and veracity, respectively. Based on these characteristics, various data types are proposed, such as unstructured data, semistructured data, and structured data. The authors in Ref. [9] introduced an unified tensor model to represent the data generated from multiple sources. Based on the tensor extension operator, different data types are represented in the forms of subtensors and processed to unified tensors. Using the aforementioned model, an incremental high order singular value decomposition method is described for reducing the dimensionality of the big data. Moreover, intelligent transportation is used as a case study to verify the performance of the data representation model and incremental dimensionality reduction method and then it can be seen that this model can be implemented as the big data system model for the data representation.

The authors in Ref. [10] introduced a mobility analytical framework for big mobile data, based on real data traffic collected from 2G/3G/4G networks covering nearly 7 000 000 people. In order to construct the history trajectories of users, the authors applied different rules to extract users locations from different data sources, and reduce oscillations between the cell towers.

Various formats of unstructured data are presented in Ref. [11], which are described as kind of big data representation, such as documents, multimedia, emails, blogs, websites, textual contents, etc. Analytics-as-a-Service tool, with NOSQ schema are proposed for data mining and to extract the information stored in the data. These tools can also be used on textual contents, such as tag-based files (e.g., HTML, XML, etc.) and non-tag based documents (e.g.,PDF). A sequence of pilot tests are then performed to validate the proposed tools.

2.4 Data analytics

Faced with massive data sets in spatial-temporal dimensions, more powerful analytical theories and methods are required in order to obtain novel insights. In this section, we will discuss several commonly adopted techniques, including time series analysis, machine learning and a game theoretical framework.

Wireless big data exhibits spatial-temporal dimensions, however, temporal analysis can also obtain significant findings in traffic pattern recognition or traffic modeling. The authors in Ref. [12] used the time series analysis to decompose the regular and random components, then used time series prediction to forecast the traffic patterns based on the regularity components, which exhibited high predictability. This work provides a novel approach at simplifying the time series data in wireless networks using time series analysis.

In recent decades, the development of machine learning[13], especially deep learning[14], has significantly improved the performance of modeling and prediction in many fields. It is developed from the artificial neural networks based mainly on the knowledge of the human brain, statistics, and applied mathematics. Deep learning[15], as a branch of machine learning algorithms, attempts to model high-level data representations by using multiple layers of neurons and multiple non-linear transformations[16] for big data analytics. It allows the computer to build complex concepts out of simpler concepts by constructing deeper neural networks.

Recently, with the increment of layers in the deep learning models, it has become the most popular and powerful when it is built on big data. Moreover, due to more powerful computers and larger datasets, the training of deeper networks becomes faster and easier. The deeper the network layer of the model is, the greater the ability of the model to represent raw data. This results in a better model performance. Although a lot of success has been achieved in recent years, particularly in computer vision and automatic speech recognition using deep learning, how to design deep learning models for analyzing wireless big data is yet to be fully studied .

Machine learning and deep learning has proven its power in many other fields, and the authors in Ref. [17] incorporated deep learning and Apache Spark into wireless communications areas. They proposed a scalable learning framework based on Apache Spark, which can support distributed deep learning. By using real-world datasets containing millions of records, this framework demonstrates its speedup effectiveness. In Ref. [18], the authors focused on the phone changing prediction problem, which is cared deeply by telecommunication operators, and verified the performance of four prediction models: Logistic Regression, Random Forest, SVM (Support Vector Machine) and E-BP (Enhanced Back Propagation) neural network, under three scenarios.

When considering the network management and control problem with the wireless big data, game theoretical analysis may be powerful tools for analyzing the interaction among multiple objects, no matter network nodes or terminal nodes. The authors in Ref. [19] presented a multiple cognitive agent-based divide-and-conquer network management and control architecture and proposed a Markovian game-theoretic modeling framework. In addition, they focused on the construction of state space, the state transition computation, and the convergence of the parallel Q-learning technique, which provides a suitable and effective modeling tool, as well as various learning techniques for wireless big data networks.

3 Transmission layer related

This section introduces the wireless transmission-related wireless big data.

3.1 Channel modelling

Wireless channels convey the information sent from the transmitter to the receiver, which serverly impacts the performance of wireless communication systems. The modeling methods for wireless channesl are from two perspectives: completely built environment can lead to precise CIR (Channel Impulse Response), coined as deterministic model, and stochastic simulation method can reconstruct CIR after predicting the received signal fading, coined as stochastic model. The deterministic model will require severe large spectrum resources, especially for MIMO. Moreover, the stochastic model is not accurate enough, therefore, how to achive a tradeoff between these two perspectives using big data analysis is now triggered to launch new research topics.

The authors in Ref. [20] adopted PCA (Principal Component Analysis) to extract the key information or features of the wireless signal, and to establish the relationships between large-scale parameters of wireless channels. This work shed light on the big data analysis in wireless channel modeling. However, the size and scale of measurement data in different scenarios, as well as the data dependency problem, lead to further challenges.

3.2 Spectrum big data

The authors in Ref. [21] proposed an architecture to analyze big spectrum data in DSA (Dynamic Spectrum Access) enabled LTE-A (Long Term Evolution Advanced) networks. The proposed architecture is based on the open source ELK stack (Elasticsearch, Logstash and Kibana stack). An experiment to validate the proposed architecture was conducted, which involved the generation of data sets of DSA enabled LTE-A networks, and the setup of the ELK stack for the spectrum analysis of LTE-A log data and sample visualizations of the spectral data analysis. The expected generated data for a radio environment with an area of 3 000 000 m^2, and an eNB coverage of 5km radius, will be in Giga bytes per day.

In Ref. [22], the authors proposed an auction based scheme to solve the wireless resources allocation problem, in terms of transmit power and wireless spectrum. The goal of the work is to ensure big data transmission in a wireless network environment with the effective capacity guaranteed. This achieved through wireless network virtualization, by enabling multiple VWNs (Virtual Wireless Networks) to be mapped onto one physical SWN (Substrate Wireless Network). The work does not involve big data collection, processing and visualization to test their proposed model.

In Ref. [23], various big data analysis methods are used to improve the cache node determination, allocation and distribution accuracy in a cluster. To minimize the delay in data transmission when a SU (Secondary User) is switching from a busy channel to an idle channel, creating cache of the SU signal at multiple nodes in a cluster aids in reducing the transmission delay if cache placement is performed accurately. The accurate placement of cache is possible if the data accumulated is accessed and processed quickly. The author discusses the different scenarios of sharing information in a cognitive radio network, as well as the possible big data solution.

In Ref. [24], the authors introduced a new concept of IoSDs (Internet of Spectrum Devices), and developed a cloud-based architecture for IoSDs over future wireless networks. The goal is to build a bridge network that links various Spectrum Monitoring Devices (SMDs) and massive SUDs (Spectrum Utilization Devices), thus enabling a highly efficient spectrum sharing and management in future wireless networks. Enabling techniques for the IoSDs such as big spectrum data analytics, hierarchal spectrum resource optimization, and quality of experience-oriented spectrum service evaluation, were discussed.

In Ref. [25], the authors proposed a grass-root based spectrum database architecture, which involves the design of a spectrum data visualization using the SpectrumMap, the development of a Hadoop based Web system for SpectrumMap to improve big data processing and provide spectrum data visualization in a user-friendly access interface, and a participating model to encourage users to join the SpectrumMap so as to enable grass-root based data collections.

3.3 Multiple user access

Kaddour et al.[26] proposed the opportunistic and efficient RB (Radio Block) allocation (OEA) algorithm and QoS (Quality of Service) based OEA, to allocate radio blocks for UEs (User Equipments) in LTE uplink networks. This was proposed with the aim of maximizing the aggregate throughput of UEs, while reducing the transmission power via power control adjustment with respect to the expected QoS of users and radio channel conditions. The RB allocation algorithms takes into account the SC-FDMA constraints specific to LTE release 8 networks, and updates the user transmission power for every RB based on the number of allocated RBs. The MCS (Modulation and Coding Scheme) mode is used in the throughput calculation for comparing performance with the optimal solution and relevant algorithm found in literature.

The authors in Ref. [27] presented the CR-WSN (Cognitive Radio-Wireless Sensor Network) as a solution to the problem of distributed spectrum access under unknown environment statistical information. The problem is modeled as an I.I.D (Independent and Identically Distributed) multi-Armed Bandit model, with the objective of maximizing throughput and minimizing regret, as well as predicting channel availability by exploration and learning. The proposed channel access scheme is divided into three parts: (1) A learning algorithm based on the tuned UCBT (Upper Confidence Bound Tuned) index introduces a variance factor and can select arbitrary channels with the kth largest index value; (2) channel grouping according to the water-filling principle based on the modified UCB-tuned index; (3) accessing channel grouping with fairness based on the grouping method to avoid collisions among cognitive users.

DSA (UCBT) framework is presented in Ref. [28] to optimize spectrum utilization and improve performance of small cell APs (Access Points) while maintaining a reliable system operation at all APs in multi-RAT, multi-operator deployment environments. The framework provides small cells with secondary access to spectrum that are primarily assigned to other RATS and operators. In the proposed framework, all APs send out SOI (Spectrum broadcast Occupancy Information) and system information for a predefined duration of time, to determine the accessible portion of RF blocks that are available for the APs. The SOI facilitates secondary small cell access to spectrum assigned to other RATs, as small cells with multi-mode transceivers can accurately determine the SOI of all employed RATs for all operators. Two modes were presented for the DSA: (1) SODSA(Single-Operator Dynamic Spectrum Access), where the set of primary APs have primary access to the RF block assigned to employed RATs and secondary access to the RF blocks of other RATs; and (2) MODSA (Multi-Operator Dynamic Spectrum Access), which employs SODSA, wherein small cells have primary access to small cell operator bands and secondary access to bands of other operators.

The work in Ref. [29] introduces the concept of GPS-OSDMA (Global Positioning System - Opportunistic Space Division Multiple Access), which realizes the position of MSs (Mobile Stations) using GPS technology. This practically eliminates the feedback bits used to transmit CSI information from the MS to the serving BS (Base Stations) in traditional systems. The serving BS calculates the angle of MS in order to form beams and subsequent data transmission without any known feedback CSI from MS.

4 Network layer related

Network layer related wireless big data research may be the most investigated area in wireless big data, including big data driven network architecture, network traffic analysis, network optimization and so on, which will all be presented in this section.

4.1 Network architecture

The approach towards the design of the network architecture of wireless networks aided/oriented wireless big data is essential for further potential data analysis, for which there are several perspectives we may consider, including scalability and flexibility, energy consumption, etc.

SDN (Software Defined Networ) has gained much more research focus since its proposal due to its ability to program the network, and its flexibility in the separation of the control plane and the forwarding plane. In Ref. na-sdn, the authors analyzed and expressed the bilateral advantages between SDN and big data applications, and showed that big data and SDN can benefit each other in many fields, including traffic engineering, cross-layer design, and defeating security attacks. However, this topic is just beginning, and many open issues need to be addressed in future research, although performing big data analysis based on SDN is promising.

In Ref. [31], the authors presented one communication network structure called DEINs (Data and Energy Integrated communication Networks), which aims to integrate WIT (Wireless Information Transfer) and WET (Wireless Energy Transfer), so that this network structure can achieve the trade-off between data processing efficiency and energy efficiency. Such an idea is implemented mostly using energy harvesting in the physical layer and energy efficient routing in the network layer. Two cases, the fair resource allocation algorithm in lower layer and data forwarding scheme in higher layer, respectively, exhibited the effectiveness of the proposed structure.

4.2 Network traffic analysis

Research works on traffic analysis and monitoring can be divided into two types: monitoring system design/construction and traffic analysis in temporal/spatial/spatio-temporal dimensions. In this section, we will introduce one work for each type, due to the space limitation.

In Ref. [32], the authors presented a traffic monitoring and analysis system for large-scale networks, based on the Hadoop platform. Moreover, this system has been deployed in the core network of a large cellular network and extensively evaluated. The authors claimed that the system can efficiently process 4.2 TB of traffic data from 123 Gbit/s links daily, with high performance and low cost, which can well support the emerging big data analysis. In this system, the traffic collector, data store, and traffic analysis algorithm along with the result presentation interface are provided.

As for traffic analysis, Zhang et al.[33] first categorized the data set in the telecommunication networks into user-oriented and network-oriented. Thereafter it presents the two case studies in the temporal dimension and spatial dimension. By examining the realistic data, it is concluded that it is not proper to model the voice call arrivals as a Poisson process in the temporal dimension. The authors then investigated the base station behavior in temporal dimension, and revealed the night burst phenomenon of college students by comparing the locations of the base stations with the real-world map.

Another work[34] studied the mobile user behavior from three aspects: 1) data usage; 2) mobility pattern; and 3) application usage. The authors classified mobile users into different groups to study the resource consumption in mobile Internet and observed that traffic heavy users and high mobility users tend to consume massive data and radio resources simultaneously. Both the data usage and the mobility pattern are closely related to the application access behavior of the users. Users can be clustered through their application usage behavior, and application categories can be identified by the methods utilized to attract the users. Such analysis may be helpful for network operators to design appropriate mechanisms in resource provision and mobility management for resource consumers, based on different categories of applications.

4.3 Network optimization/planning

Perhaps one of the most important applications of wireless big data is the improvement and optimization of the performance of wireless networks through data analysis based network planning or resource allocation/scheduling. This task is surely very tough and there are still many open issues concerning this topic. However, there are certain ongoing research efforts, which will be introduced in this paper.

In Ref. [35], the authors first proposed a framework of BDD (Big Data-Driven) mobile network optimization and presented the characteristics of big data that are collected not only from user equipment but also from mobile networks. In addition, several preliminary hints are discussed on the application of the proposed framework in order to improve the network performance. However, just as the authors pointed out, currently, how to use the big data to optimize the network still requires significant and solid progress in theory and practice.

5 Application layer related

In this section, application/user related works will be reviewed. We selected five topics: user mobility, social networking, smart grid, IoT and drones/UAVs. However, other topics, like urban planning[36], were neglected, due to the space limitation.

5.1 User mobility analysis

Understanding the human mobility is essential to wireless and mobile networks, such as resource allocation, network planning, content distribution, and hand-over strategies. Current mobile phone datasets allow for the analysis of human behavior on an unprecedented scale. From the data sources in the wireless big data area, existing works focus on extracting mobility patterns from CDRs (Call Detail Records), Wi-Fi traces, GPS traces and cellular networks traffic investigation. These data sources have their own pros and cons; i.e., CDR only captures movements during phone calls, and Wi-Fi traces can only provide the information on a relatively small scale. However, this topic may be the most investigated topic in the wireless big data area and there are many works worthy of discussion.

The first question concerning human mobility may be the prediction accuracy. With respect to this question, the most important work is the theoretical prediction limit performed in 2010, as discussed in Ref. [37]. The authors used the data of anonymized mobile phone users and measured three entropies defined on the temporal mobility trajectory. The result led to a 93% potential predictability in user mobility across the whole user base, which inspired the research in this direction. Thereafter, the authors in Ref. [38] analyzed the travel patterns in the mobile phone call data records of 500 000 individuals and verified that the theoretical maximum predictability is as high as 88%. A series of MC (Markov Chain) based models were then implemented to predict the actual locations visited by each user. The results showed that MC models can produce a prediction accuracy of 87% for stationary trajectories, and 95% for non-stationary trajectories. This indicates that human mobility is highly dependent on historical behaviors, and that the maximum predictability is not only a fundamental theoretical limit for potential predictive power, but also an approachable target for actual prediction accuracy.

Using mobile phone data sets, we can estimate or infer many difficult or impossible knowledge in sociology using other methods, such as human statistics and commuting analysis. The authors in Ref. [39] analyzed the communication data of 100 000 anonymized and randomly chosen individuals in Portugal and observed that the majority spend most of their time at only a few locations. The home and office locations can be robustly identified and compared, with official census data. This work may be useful for governments to carry out urban planning in a much more efficient and effective way.

Another perspective is to investigate the mobility using the data collected from cellular data networks. The authors in Ref. [40] asked if the mobility properties derived from cellular data traffic is different from the previous findings using other data sources, especially the commonly used CDR based approach. They discovered that the data network records can provide a finer granularity of location and movement information, and can also identify three different temporal movement patterns. This work can therefore be better utilized in the core network or mobile cloud center for mobility analysis. Moreover, the authors in Ref. [41] also investigated inter-arrival time, dwell time distributions and other mobility patterns in mobile cellular networks. Based on real cellular data measurements, the authors evaluated the fitness of various typical statistical distributions such as power-law, exponential, Weibull, lognormal and Rayleigh distributions, and found that a power-law distribution fits both the inter-arrival time and dwell time more precisely.

The authors in Ref. [42] investigated the relationships between cyberspace and the physical world, and revealed that human mobility and the consumed mobile traffic have strong correlations with distinct periodical patterns in the time domain. In addition, both human mobility and mobile traffic consumption are linked to social ecology, which in turn helps us to understand human behavior better. The proposed big data processing and modeling methodology, combined with the empirical analysis on mobile traffic, human mobility, and social ecology, paves the way toward a deep understanding of human behaviors in a large-scale metropolis.

Finally, researchers with physics background[43] are becoming increasingly interested in user mobility research, and they use empirical data on human mobility, captured by mobile-phone traces, to show that the predictions of the CTRW models are in systematic conflict with the empirical results, which brings us a novel insight on this topic. We believe that in the future, the collaboration of computer scientists and physicists will result in significant and remarkable conclusions in this field.

5.2 Social network/feature

The prediction of human behavior by the analysis of the behaviors from a large dataset is a challenging task. This paper[44] has discussed a community centric framework for predicting community activity by analyzing individual activities. In the proposed method, SVD (Singular Value Decomposition) and clustering are applied to co-relate the collected individual activities and communities and then provide individual activity prediction and customized recommendations. The proposed method is analyzed on a real data set collected by EPIC Lab Huazhong University of Science and Technology, over a 15 month period. For storage and processing of the datasets, the authors designed two models: IAM (Individual Activity Model) and CAM (Community Activity Model). Moreover, IAM is analyzed and then merged into CAM to reduce the complexity of the data set. In CAM, approximate tensors, i.e., approximate community activity rules, are found after tensor initialization and trucker decomposition. The evaluation performance of the proposed approach based on the dataset shows that it has a good performance prediction level, and that it reduces the complexity of data. The main limitation of this model is that it works only on offline modeling.

Web services have a great impact on the Web. However, due to the isolated characteristics and lack of communication among services, the progress and number of Web services are not satisfactory. This paper[45] has proposed a solution for constructing a global social service network by interlinking the isolated services using the linked social services approach. To improve the usability threshold for service consumers, a novel approach called link-as-you-go, which is a global social service network providing LSSaaS (Linked Social Services as a Services), is proposed. This service allows users to start browsing in one service and then navigate along social links into related services to explore service-to-service, so that users can explore the global social service network more deeply on the open Web. Here, SOAS (Service-Of-A-Service) is used as a platform to build a global social service network providing LSSaaS, which has 2 000 services from the OWL and SAWSDL test collection, approximately 3 000 services from the owls-mx20 test collection, around 5 000 services indexed by Seekda.com, and about 500 services from ProgrammableWeb.com. The evaluation of the effectiveness of LSSaaS in terms of the total service discovering time, is observed during changing the number of desired services and the participant group. In this experiment, 50 participants have participated for three months who got 60 000 transactions from their usage histories. The experimental results show that proposed approach can solve the quality of service discovery problem; improving not only the service discovering time, but also the success rate, by exploring service-to-service based on the global social service network.

Effective video recommendation using multimedia big data is a challenging at the same time painstaking task. The main reasons behind these complexities are the privacy of user context and video service vendor repositories. This paper[46] has discussed a cloud based video recommendation model and denoted it as geometric differentially private model. This proposed model concerned with the privacy issue of users as well as video service vendor repositories, and the dimensionality and sparsity issue of multimedia big data. Here, differential privacy is introduced into the distributed online learning model to ensure an efficient recommendation system. According to the model, video service vendors extract user context data and recommend videos to users from its repositories. Thereafter, user feedback is observed and the observation is used by the service vendors to learn and update their selection strategy for the next user. In the model, the user data context is converted to the context vector, which has an impact on the sparsity and dimensionality of big data. Moreover, the model becomes slow while learning from large, high dimensional data. To solve the problems mentioned above, the authors from the given paper have proposed to divide the entire context space into groups of multiple context subspaces, according to the number of arriving users by video service vendors. The model uses the e-differential privacy policy to ensure users and service vendors privacy policy. Furthermore, to reduce the performance loss, e is changed. Here, the authors have evaluated the performance of the proposed algorithm in terms of the converged regrets and accuracy level, by using the real dataset from Sina Microblog and public information from Youku; comparing the results with CAP (Centralized learning with Adaptive Partition), DUP(Distributed learning with Uniform Partition) and DAP. The experimental results show that the proposed model has low regret performance value with high privacy preserving level.

Online users work on cloud platform that contains high dimensional massive personal data. By using the personal information, the online websites predict the behaviors of users and make recommendations for them. Handling the large amount of data by securing privacy to make sensible recommendations is really a challenging task. In Ref. [47], a distributed online learning algorithm is proposed to handle the decentralized data. According to the algorithm, the authentic nodes have old parameters and after receiving the noise added parameters from cloud computing nodes, the nodes make a weight average value based on the received and old parameters and update the value. The scenario is different for the intruder. When the adverse parties attack the nodes to divulge the personal data, they fail to crack the private data due to its noise addition in data. However, in this model, sparsity is induced in the data by using Lasso which converts the irrelevant data into zero coefficient, thus the model reduces the computational complexity and enhances the accuracy of prediction. Moreover, the proposed algorithm has made feasible sense of recommendations for users due to its privacy policy and sparse solution. As an experiment, four simulation works are presented. In each case, the proposed algorithm exhibits better performance.

A good data management approach is required to handle the big data to enhance insight, process optimization, data mining and knowledge discovery. This paper[48] has proposed an approach of big data management by using big graph data to support big data mining. Here, distributed big data from different social network services are managed into four groups: directed graphs, bi-directed graphs, unidirected graphs, bi-partite graphs. The collections of the graphs can be represented by (u,v) key-pairs based on their relationships. Then these key pairs are defined into maps and then they are reduced. This paper has demonstrated the performance of this proposed data management method by using SNAP Facebook datasets and SNAP Twitter datasets, and encoding it into JAVA. The results show that in each case of the datasets, the evaluation time veries due to the managing data in different processor level, i.e., clustering is faster than that of single machine processor.

5.3 Smart grid

To deal with the missing or inaccurate smart meter data, the paper[49] has used the pseudo measurement generator that utilizes a weighted average combination of historic data and interpolated/extrapolated data from previous or future measurements. Positive sequence equivalent models are used for parameter estimation. Moreover, for visualization of the data, they have used contouring techniques, which can effectively display the summary of complex simulation results at a glance. They state that this is useful for the stakeholders to know the system performance.

In Ref. [50], a method for the clustering of electricity consumption behavior dynamics toward large data sets was proposed. It is different from traditional load profiling from a static prospective. Here, SAX and time-based Markov model are utilized to model the electricity consumption dynamic characteristics of each customer. The K-L distance is used to quantify the dissimilarity between two probabilistic distributions for classifying the customers into several clusters. A density-based clustering technique, CFSFDP, is performed to discover the typical dynamics of electricity consumption and segment customers into different groups. Finally, a time domain analysis and entropy evaluation are conducted on the result of the dynamic clustering to identify the demand response potential of each group's customers. Markov model predicts the trend or the level of electricity consumption for each customer. The challenges of massive high-dimensional electricity consumption data are addressed in three ways. First, SAX can reduce and discretize the numerical consumption data to ease the cost of data communication and storage. Second, Markov model are modeled to transform long-term data to several transition matrixes. Third, a distributed clustering algorithm is proposed for distributed big data sets. The number of states is a tradeoff between the information loss by SAX and the size in transition probability matrix.

A technique to analyze big smart metering data towards differentiated user services for electricity usage is presented in Ref. [51]. The model is based on an analysis of a real smart metering data trace, where various usage patterns among the power energy customers are observed. One key problem of a differentiating user service model is that the model computation faces a huge amount of data. There is a large number of customers, and for each customer, his/her electricity usage pattern is represented by a long period and multi-dimensional data. As a result, the complexity for a differentiate service model is not in the sense of computation, but in big data. For this, they developed a novel sublinear algorithm, where they use a sublinear amount of data and guarantees a small error bounds for a given confidence. Furthermore, they provided theoretical proofs and demonstrated that their algorithm can effectively reduce the amount of data to be processed to a range that is reasonable for the computing capability.

In Ref. [52], a methodology for extracting, classifying, and then verifying the reliability of the final clustering is presented for smart meter data. In particular, they analyzed different size demands and their distribution as a function of time-of-day and season. Then they identified four key time periods which described different peak demand behavior, coinciding with common intervals of the day: overnight, breakfast, daytime, and evening. They also found that demand in the different time periods changed as a function of seasonality and days of the week, thus identifying two major sources of variation. In addition, they also included a mean normalized standard deviation of the demand as a measure of the irregularity of a customer. The time periods not only helped to model peak time behaviors and their variation but also allowed to reduce the number of attributes in clustering implementation. They presented a clustering of their chosen attributes into ten groups using a finite mixture of Gaussian distributions. Such a method is commonly used in clustering but has not been explored in great detail within the power systems literature despite the advantages over more common methods. They showed that the final clusters identified many important behaviors of the customers. As well as time periods of greatest demand, they identified those customers with the largest variability according to seasonality and weekend versus weekday differences. Understanding such changes in a customer's seasonal behavior can aid network operators in longer term planning of the networks. Finally, using an existing bootstrap technique, they show that the clustering is reliable.

The objective of Ref. [53] is to appertain big data technology into smart grids. It suggests an architecture with two independent procedures, as a data-driven solution, to conduct anomaly detections. In addition, moving split-window technology was introduced for real-time analysis, and a new statistic MSR was proposed to indicate the data correlations, as well as to elucidate the parameters interchanged among the utilities. The algorithm of this architecture is based on RMT; it is a fixed objective procedure, which is lucid in logic and fast in speed. The non-asymptotic framework of RMT aids to conduct high-dimensional analysis for real systems, despite of relatively moderate datasets. It also provides a natural way to decouple the interconnected systems from data perspective. The group-work model of utilities in the systems, meanwhile, makes some data-driven functions possible, such as distributed calculation and comparative analysis.

5.4 Internet of Things

A network architecture combining SDN and message oriented publish/subscribe DDS (Data Ddistribution Service) middleware to define an abstract that is not dependent on a specific network protocol is presented in Ref. [54]. In the architecture, SDN controller integrates a DDS interface added to it and forms a DDS messaging layer that mediates between the IoT system and the network. Services in the DDS publish/subscribe messaging layer are packer handler service, packet forwarder service and flow programming service. It supports proactive and reactive flow programming, also allow anonymous, asynchronous, and many-to-many communication semantics.

Awais et al.[55] proposed a system architecture that gathers data from IoT devices and Social Network to form SIoT (Social IoT) data, analyze the Big data gathered use it to describe human behavior in a social environment using a smart city scenario in real time. The proposed architecture is made of three operational domains i.e. object domain where data objects are formed by the interconnected IoT devices and social network; SIoT server which handles initial processing such as handling redundancy, noise removal and data aggregation using relay nodes, aggregation classifiers to link SIoT information together. It further processes the data based on Web standard specifications such as Uniform resource identifier URI, HTTP REST (REpresentational State Transfer) and linked data and Application domain for user notification, security as well as social member profiles and their relationships. Using data set from water usage, parking lots, social network (Twitter), Vehicular mobility traces and highway vehicular traffic the system compares the historical data with the new data and provides feedback to the user.

Kafle et al.[56] proposed an identity-based heterogeneous mobile network for IoT and M2M communication called HIMALIS (Heterogeneity Inclusion and Mobility Adaptation through Location ID Separation). HIMALIS protocol stack has an identity layer inserted between the transport and network layer of the TCP/IP protocol stack. It also has a name registry for device name, IDs, locators and public key storage. The architecture introduces an alpha-numeric namespace independent of device location which is used for device discovery, authentication, management and a bit-string ID which do not change as the packet travel from source to destination used to uniquely identify the source and destination of the packets. The network component of HIMALIS consist of the access network; authentication agent, local name server and gateway, global transit network; high speed routers to connect access networks and name registers. HIMALIS also provide TCP/IP like socket APIs and system calls which application developers familiar with TCP/IP sockets can write IoT/M2M applications. It has a middleware so TCP/IP applications can also run on HIMALIS.

Authors in Ref. [57] proposed a method of wireless network optimization which aggregates external data sources at cache-enabled base stations. For a network formed by a set of small base stations and user terminal, a content library where the user terminal demands content is defined. In the network, each SBS (Small Base Station) has limited storage capacity, thus caches a subset of content from the library. The paper presents a joint optimization cache decision matrix and content popularity matrix estimation required to characterize the average backhaul load and user's average request satisfaction. To validate the concept, over four million data set was collected from a telecommunication company, relevant field of the data extracted and processed on the Hadoop platform using statistical machine learning tools.

5.5 Drone and UAV

In Ref. [58], the authors proposed a data collection protocol for optical UAV networks based on optical codewords. Large amounts of data generated by UAVs are delivered in real time and at high data rate connectivity using FSO (Free Space Optics) communication links. This was achieved using two trees; the identification tree, with hierarchical codewords allowing to globally identify nodes in the networks, and the collection tree to pick up data from drones and route the data to the root drone in order to be delivered to a collection node.

A middleware solution for ADDSEN (Adaptive Data processing and dissemination for Drone swarms in urban Sensing) was provided in Ref. [59]. A cyber-physical sensing framework, using partially ordered knowledge for distributed knowledge management in drone swarms, was proposed to efficiently process sensed data in the middleware. Here, a drone swarm is a collection of 3∼4 drones, carrying sensors to collect raw data about traffic speed, pollution and noise. The ADDSEN middleware on the drone processes and disseminates sensed data and knowledge items containing partially processed data. The proposed framework implements a reinforcement learning based dissemination strategy to better adapt to the link status for intra/inter-swarm data dissemination. Periodic intra/inter swarm broadcast dissemination consumes the storage and energy of each drone. To maximize the in-field time of the drone swarms as sensing grids, an optimal balancing method is designed in ADDSEN for the cooperative task of balancing storage and energy allocation.

The authors in Ref. [60] proposed a drone-based mechanism to relay traffic and improve network performance. The mechanism involves the deployment of drones to traffic hotspots and areas of capacity bottlenecks in a network. The drone acts as an active repeater in a way that users stop to transmit to any other cell when they start to communicate with the drone relay and a proportion of the macrocell backhaul resources are reserved. The relay behaves like a super mobile user transmitting to the macrocell on behalf of all the neighboring users. Users in the coverage of the drone relay will not be served by the macrocell and they will be aggregated to this drone relay. The mechanism activates the drone only when certain specific conditions that are mainly related to the cell congestion state and the location of traffic are verified.

A method for improving the reliability of a LSN (Linear Wireless Sensor Network) using UAVs was proposed in Ref. [61] to mitigate the impact of failed nodes within the network. The faulty nodes are detected using nodes signal transmission using multihop and UAVs. The UAVs temporarily perform the communication and sensing functions until the faulty nodes are fixed. A solution for isolating only defective nodes in LSNs was discussed.

6 Privacy and security

The privacy and security of wireless big data is a very important research area. Hua et al.[62] proposed a differentially private algorithm for location generalization in time-series trajectory data (human mobility traces) using exponential mechanism for processing. The algorithm combines nodes using the trajectory distances to merge sites at the same time points. Furthermore, a release algorithm is implemented to show trajectories after location generalization.

A pseudonym algorithm for developing location data set that conceals users' location but retains path information of the user is proposed in Ref. [63]. It dynamically changes the users' pseudonym to avert interference attacks, due to user home location knowledge. Based on the algorithm, numerous users' pseudonyms are exchanged only when users meet at the same site, so as to eliminate any form of connection between the pseudonyms before and after the exchange.

The authors in Ref. [64] present Promesse, a scheme for anonymizing mobile datasets. It conceals the point of interest and mobility habits of users by distorting time. The data anonymization is achieved by smoothing the speed of the users along a path and obscuring the end points of their paths so as to make them less recognizable, and thus, less vulnerable to attacks.

The authors in Ref. [65] present a method of securing transmitted data from sensor nodes and preservation of data in sensor node memory. The method is based on the use of symmetric cryptography using the TPM (Trusted Platform Module) for the implementation of the cryptographic procedures. For data transmission between Master nodes and CSN (Collecting Sensor Node), unicast transmission and symmetric cryptography (AES in CBC mode with individual NSK (Network Security Key) of CSN) is used. All CSNs in the cluster are equipped with TPM to support the cryptographic procedures.

7 Conclusion and open problems

The advancement of wireless communications systems, such as the 5G, ushers in a new era of wireless big data. This trend is amplified by the proliferation of reliable and low cost sensors, the development of social networks, and autonomous systems such as UAVs and intelligent transportation systems. In order to fully understand the emerging wireless big data, a survey is provided here in terms of the changes at different communication layers, and the effects on various important applications. Challenges and opportunities are identified and this survey could serve as a starting point for an exciting new research thrust in wireless big data.

There are many open problems in wireless big data research. Firstly, how to integrate the inherent nature of wireless big data from the electrical engineering perspective and the strength of machine learning and data mining from the computer science perspective is a future challenge. We may rely on information theory, random matrix theory or other theoretical tools to describe and model the corresponding nature brought on by the stochastic and nonstationary property of the wireless channel. Secondly, next generation communication systems can serve aggregated massive users, and the transmission contents among them may not be modeled as independent. Thereafter, how to model such dependency and use such dependency to improve the transmission efficiency could be a very interesting topic. Here, big data aided computation and prediction technique will play an important role in assisting the communications research.

References

[1] V. D. Blondel., A. Decuyper, G. Krings. A survey of results on mobile phone datasets analysis [J]. EPJ data science, 2015, 4(1): 1.

[2] M. Lin and W. Hsu. Mining GPS data for mobility patterns: A survey [J]. Pervasive and mobile computing, 2014, 12: 1-16.

[3] K. Chen, H. Zhou. Research on realization mode of telecom operators' big data resource and its strategy [J]. Mobile communications, 2016, 40(1): 63-67.

[4] X. Zhang, Z. Yi, Z. Yan, et al. Social computing for mobile big data [J]. Computer, 2016, 49(9): 86-90.

[5] X. Ding, Y. Tian, Y. Yu. A real-time big data gathering algorithm based on indoor wireless sensor networks for risk analysis of industrial operations [J]. IEEE transactions on industrial informatics, 2016, 12(3): 1232-1242.

[6] L. Kong, D. Zhang, Z. He, et al. Embracing big data with compressive sensing: a green approach in industrial wireless networks [J]. IEEE communications magazine, 2016, 54(10): 53-59.

[7] Y. He, F. R. Yu, N. Zhao, et al. Big data analytics in mobile cellular networks[J]. IEEE access, 2016, 4: 1985-1996.

[8] C. Zhang, R. C. Qiu. Massive mimo as a big data system: random matrix models and testbed [J]. IEEE access, 2015, 3: 837-851.

[9] L. Kuang, F. Hao, L. T. Yang, et al. A tensor-based approach for big data representation and dimensionality reduction [J]. IEEE transactions on emerging topics in computing, 2014, 2(3): 280-291.

[10] Y. Qiao, Y. Cheng, J. Yang, et al. A mobility analytical framework for big mobile data in densely populated area[J]. IEEE transactions on vehicular technology, 2016, PP(99): 1-13.

[11] R. K. Lomotey, R. Deters. Towards knowledge discovery in big data [C]//The 8th International Symposium on Service Oriented System Engineering (SOSE), 2014: 181-191.

[12] F. Xu, Y. Lin, J. Huang, et al. Big data driven mobile traffic understanding and forecasting: a time series approach[J]. IEEE transactions on services computing, 2016, 9(5): 796-805.

[13] K. Murphy. Machine Learning: A Probabilistic Perspective [M]. Cambridge: MIT Press, 2012.

[14] I. Goodfellow, Y. Bengio, A. Courville. Deep Learning [M]. Cambridge: MIT Press, 2016.

[15] J. Donahue, L. Hendricks, S. Guadarrama, et al. Long-term recurrent convolutional networks for visual recognition and description [C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 2625-2634.

[16] Y. Le Cun, Y. Bengio, G. Hinton. deep learning[J]. Nature, 2015, 521(7553): 436-444.

[17] M. A. Alsheikh, D. Niyato, S. Lin, et al. Mobile big data analytics using deep learning and apache spark [J]. IEEE network, 2016, 30(3): 22-29.

[18] Q. Ma, S. Zhang, W. Zhou, et al. When will you have a new mobile phone? an empirical answer from big data [J]. IEEE access, 2016.

[19] C. Yang. Learning methodologies for wireless big data networks: a Markovian game-theoretic perspective [J]. Neurocomputing, 2016, 174: 431-438.

[20] J. H. Zhang. The interdisciplinary research of big data and wireless channel: a cluster-nuclei based channel model [J](Accepted). China communication, 2016.

[21] S. GVK, S. R. Dasari. Big spectrum data analysis in dsa enabled lte-a networks: A system architecture [C]//The IEEE 6th International Conference on Advanced Computing (IACC), 2016: 655-660.

[22] Q. Zhu, X. Zhang. Effective-capacity based gaming for optimal power and spectrum allocations over big-data virtual wireless networks [C]//The IEEE Global Communications Conference (GLOBECOM), 2015: 1-6.

[23] A. Omar. Improving data extraction efficiency of cache nodes in cognitive radio networks using big data analysis [C]//The 9th International Conference on Next Generation Mobile Applications, Services and Technologies, 2015, 2015: 305-310.

[24] Q. Wu, G. Ding, Z. Du, et al. A cloud-based architecture for the internet of spectrum devices over future wireless networks [J]. IEEE access, 2016, 4: 2854-2862.

[25] Y. Li. Grass-root based spectrummap database for self-organized cognitive radio and heterogeneous networks: Spectrum measurement, data visualization, and user participating model [C]//The IEEE Wireless Communications and Networking Conference (WCNC), 2015: 117-122.

[26] F. Z. Kaddour, E. Vivier, L. Mroueh, et al. Green opportunistic and efficient resource block allocation algorithm for lte uplink networks [J]. IEEE transactions on vehicular technology, 2015, 64(10): 4537-4550.

[27] J. Zhu, Y. Song, D. Jiang, et al. Multi-armed bandit channel access scheme with cognitive radio technology in wireless sensor networks for the internet of things [J]. IEEE access, 2016, 4: 4609-4617.

[28] A. Alsohaily and E. S. Sousa. Dynamic spectrum access for multi-radio access technology, multi-operator autonomous small cell communication systems [C]// The IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), 2014: 1778-1782.

[29] P. Chaichana, P. Uthansakul, and M. Uthansakul. Gps-aided opportunistic space-division multiple access for 5g communications [C]//The 20th Asia-Pacific Conference on Communication (APCC2014), 2014: 468-472.

[30] L. Cui, F. R. Yu, Q. Yan. When big data meets software-defined networking: SDN for big data and big data for SDN [J]. IEEE network, 2016, 30(1): 58-65.

[31] K. Yang, Q. Yu, S. Leng, et al. Data and energy in-

tegrated communication networks for wireless big data [J]. IEEE access, 2016, 4: 713-723.

[32] J. Liu, F. Liu, N. Ansari. Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop [J]. IEEE network, 2014, 28(4): 32-39.

[33] S. H. Zhang, D. D. Yin, Y. Q. Zhang, et al. Computing on base station behavior using erlang measurement and call detail record [J]. IEEE transactions on emerging topics in computing, 2015, 3(3): 444-453.

[34] J. Yang, Y. Qiao, X. Zhang, et al. Characterizing user behavior in mobile internet [J]. IEEE transactions on emerging topics in computing, 2015, 3(1): 95-106.

[35] K. Zheng, Z. Yang, K. Zhang, et al. Big data-driven optimization for mobile networks toward 5G [J]. IEEE network, 2016, 30(1): 44-51.

[36] T. Louail, M. Lenormand, O. G. C. Ros, et al. From mobile phone data to the spatial structure of cities [J]. Scientific reports, 2014, 4(5276): 1-12.

[37] C. Song, Z. Qu, N. Blumm, et al. Limits of predictability in human mobility [J]. Science, 2010, 327(5968): 1018-1021.

[38] X. Lu, E. Wetter, N. Bharti, et al. Approaching the limit of predictability in human mobility [J]. Scientific reports, 2013, 3(2923): 1-9.

[39] B. C. Csi, A. Browet, V. A. Traag, et al. Exploring the mobility of mobile phone users [J]. Physica A: statistical mechanics and its applications, 2013, 392(6): 1459-1473.

[40] Y. Zhang. User mobility from the view of cellular data networks [C]//IEEE INFOCOM 2014-IEEE Conference on Computer Communications, Toronto, 2014: 1348-1356.

[41] X. Zhou, Z. Zhao, R. Li, et al. Human mobility patterns in cellular networks[J]. IEEE communications letters, 2013, 17(10): 1877-1880.

[42] F. Xu, Y. Li, M. Chen, et al. Mobile cellular big data: linking cyberspace and the physical world with social ecology [J]. IEEE network, 2016, 30(3): 6-12.

[43] C. Song, T. Koren, P. Wang, et al. Modelling the scaling properties of human mobility [J]. Nature physics, 2010, 6(10): 818-823.

[44] Y. Zhang, M. Chen, S. Mao, et al. Cap: community activity prediction based on big data analysis [J]. IEEE network, 2014, 28(4): 52-57.

[45] W. Chen, I. Paik, P. C. K. Hung. Constructing a global social service network for better quality of Web service discovery [J]. IEEE transactions on services computing, 2015, 8(2): 284-298.

[46] P. Zhou, Y. Zhou, D. Wu, et al. Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks [J]. IEEE transactions on multimedia, 2016, 18(6): 1217-1229.

[47] C. Li, P. Zhou, Y. Zhou, et al. Distributed private online learning for social big data computing over data center networks [C]//2016 IEEE International Conference on Communications (ICC), 2016: 1-6.

[48] C. K. Leung, H. Zhang.Management of distributed big data for social networks [C]//The 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2016: 639-648.

[49] J. Peppanen, M. J. Reno, M. Thakkar, et al. Leveraging ami data for distribution system model calibration and situational awareness [J]. IEEE transactions on smart grid, 2015, 6(4): 2050-2059.

[50] Y. Wang, Q. Chen, C. Kang, et al. Clustering of electricity consumption behavior dynamics toward big data applications [J]. IEEE transactions on smart grid, 2016, 7(5): 2437-2447.

[51] E. Pan, D. Wang, Z. Han. Analyzing big smart metering data towards differentiated user services: A sublinear approach [J]. IEEE transactions on big data, 2016, 2(3): 249-261.

[52] S. Haben, C. Singleton, P. Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data [J]. IEEE transactions on smart grid, 2016, 7(1): 136-144.

[53] X. He, Q. Ai, R. C. Qiu, et al. A big data architecture design for smart grids based on random matrix theory [J]. IEEE transactions on smart Grid, 2015.

[54] A. Hakiri, P. Berthou, A. Gokhale, et al. Publish/ subscribe-enabled software defined networking for efficient and scalable iot communications [J]. IEEE communications magazine, 2015, 53(9): 48-54

[55] A. Ahmad, M. M. Rathore, A. Paul, et al. Defining human behaviors using big data analytics in social internet of things [C]//The IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), 2016: 1101-1107.

[56] V. P. Kafle, Y. Fukushima, H. Harai. Id-based communication for realizing iot and m2m in future heterogeneous mobile networks [C]//2015 International Conference on Recent Advances in Internet of Things (RIoT), 2015: 1-6.

[57] M. A. Kader, E. Bastug, M. Bennis, et al. Leveraging big data analytics for cache-enabled wireless networks [C]//The IEEE Globecom Workshops (GC Wkshps), 2015: 1-6.

[58] N. Ramdhan, M. Sliti , N. Boudriga. Codeword-based data collection protocol for optical Unmanned Aerial Vehicle networks [C]//HONET-ICT IEEE, 2016: 35-39.

[59] D. Wu, D. I. Arkhipov, M. Kim, et al. Addsen: Adaptive data processing and dissemination for drone swarms in urban sensing [J]. IEEE transactions on computers, 2016.

[60] A. Jaziri, R. Nasri, T. Chahed. Congestion mitigation in 5g networks using drone relays [C]//The International Wireless Communications and Mobile Computing Conference (IWCMC), 2016: 233-238.

[61] N. Mohamed, H. AlDhaheri, K. Almurshidi, M. Al-Hammoudi, et al. Using uavs to secure linear wireless sensornetworks [C]//The IEEE 2nd International Con-

ference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016: 424-429.

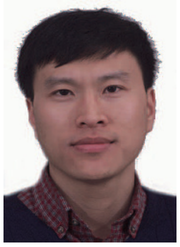
[62] J. Hua, Y. Gao, S. Zhong. Differentially private publication of general time-serial trajectory data [C]//The IEEE Conference on Computer Communications (INFOCOM), 2015: 549-557.

[63] K. Mano, K. Minami, H. Maruyama. Pseudonym exchange for privacy-preserving publishing of trajectory data set [C]//The IEEE 3rd Global Conference on Consumer Electronics (GCCE), 2014: 691-695.

[64] V. Primault, S. B. Mokhtar, C. Lauradoux, et al. Time distortion anonymization for the publication of mobility data with high utility [C]//The IEEE Trustcom/BigDataSE/ISPA, 2015, 1: 539-546.

[65] J. Furtak, Z. Zieliski, and J. Chudzikiewicz. Security techniques for the wsn link layer within military IoT [C]// The IEEE 3rd World Forum on Internet of Things (WF-IoT), 2016: 233-238.

About the authors



Lijun Qian is a professor in the Department of Electrical and Computer Engineering at Prairie View A&M University (PVAMU), a member of the Texas A&M University System located near Houston Texas, USA. He is also the director of the Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT Center) and the Wireless Communications Lab (WiComLab). Before joining PVAMU, he was a MTS in the Networks and Systems Research Department of Bell-Labs at Murray Hill, New Jersey, USA. He is a visiting professor of Aalto University, Finland. He received his B.E. from Tsinghua University in China, M.S.E.E. from Technion-Israel Institute of Technology, and Ph.D. from Rutgers University. His research interests are in big data analytics, wireless communications and mobile networks, network security and intrusion detection, and computational systems biology. His research is supported by NSF, DOE and DOD. (Email: lijunqian@ieee.org)

Jinkang Zhu has joined in University of Science and Technology of China (USTC) since 1966, and is a professor of USTC from 1992. He has been loyal to research on wireless mobile communications and networks, communication signal processing, and the future wireless technologies. Prof. Zhu was a member of the Expert Group of High Technology Communication Subject (863), director of the Expert Group of Wireless communications (863). He was director of the School of Information Science and Technology of USTC. He had been China delegate of Mobile Communication Forum of Asia-Pacific Region, the keynote speaker of IEEE ISSSC1992, the general chair of international conference of WCDMA technology, the general co-chair of international conference of WCSP2014, the general chair of Symposium of Green Wireless Communication Technologies, and the general chair of Symposium of 1st and 2nd Wireless Networks for Big Data. Recently, he studies with great interest in wireless big data, green wireless communications, and emerging technologies in wireless communications and networks. (Email: jkzhu@ustc.edu.cn)

Sihai Zhang [corresponding author] earned his B.E. in computer science from Ocean University of China, Qingdao, China, in 1996. He received M.S. and Ph.D. degree of Computer Science at University of Science and Technology of China (USTC) in 2002 and 2006, respectively. He worked as guest researcher in Department of Electronic Engineering of KAIST, South Korea during 2007∼2008. His main research interests focus on wireless networks, wireless big data and intelligent algorithm. He is now an associate professor in school of information science and technology at University of Science and Technology of China in Hefei, China, where he is leading the wireless big data research on wireless user behavior modeling and wireless network optimization technologies for future wireless systems. He is involved in China 5G mobile communication project, NSFC Sino-Finland MTC project, NSFC Key Program on wireless big data. He has published more than 60 research articles in high-impact IEEE international journals and conferences of wireless communication fields. He has also served as peer Reviewers of IEEE Wireless Communication Magazine, IEEE Transactions on Wireless Communication (TWC), IEEE Transactions on Vehicular Technology (TVT), IEEE Communications Letter (CL), and Technical Program Committee (TPC) Members and session chairs of some International Conferences including IEEE ICC/GC/PIMRC/WCSP/VTC, etc. In 2016, he co-organized two technical sessions, on mobile big data in WCSP2016 and machine type communications in WPMC 2016, respectively. He served as a guest editor of special issue on wireless big data in Journal of Communications and Information Networks in 2016. (Email: shzhang@ustc.edu.cn)