Characterizing Flow, Application, and User Behavior in Mobile Networks: A Framework for Mobile Big Data

Yuanyuan Qiao, Zhizhuang Xing, Zubair Md. Fadlullah, Jie Yang, and Nei Kato

Abstract

The recent explosion of data traffic calls for specialized systems to monitor the status of networks. Traditionally, Internet service providers collect and analyze IP flow data as they present an aggregated view of traffic. In the era of mobile big data, new approaches are required to address new challenges regarding the flow characterization in the next generation wireless networks. In this article, we propose a framework for mobile big data, referred to as FMBD, which provides massive data traffic collection, storage, processing, analysis, and management functions, to cope with the tremendous amount of data traffic. In particular, by analyzing the specific characteristics of the mobile big data from flow, application, and user behavior, such as high volume, diversity of applications, and spatio-temporal distribution, our proposed FMBD demonstrates its capability to offer real data-based advice to address new challenges for future wireless networks from the viewpoints of both operators and individuals. Tested by real mobile big data, FMBD has been operational for more than five years, and can be generalized to other environments with massive data traffic or big data.

Introduction

In today's information-centric society, the highly integrated, interconnected, and internetworked people, devices, and objects produce huge amounts of data traffic. Currently, a 4G connection generates four times more traffic, on average, in contrast with that generated by a 3G connection. Combining more advanced device capabilities with faster, higher bandwidth and more intelligent networks leads to wide adoption of data-rich multimedia applications resulting in a phenomenal increase in mobile data traffic. As the monthly global mobile data traffic reached 7.2 exabytes at the end of 2016 [1], mobile networks have indeed become both generators and carriers of massive data.

Traditionally, in order to ensure the availability and smooth operations on a network, a network traffic monitoring system, which generally reviews each incoming/outgoing packet, is deployed to estimate the key indicators of the monitored network. The network traffic monitoring system typically uses indicators such as traffic volume, heavy users, bandwidth consumption, and high usage times. In general, the primary job of network traffic monitoring is to review, analyze, and manage network traffic for any abnormality or process that can affect network performance, availability, and/or security. Furthermore, recently, improving the quality of experience (QoE) of users has become the main goal for network optimization. With 5G, the next evolution of mobile technology, resources (channels) are anticipated to be allocated based on awareness of content, users, and locations to offer services to heterogeneous networks, technologies, and devices operating in different geographic regions to fulfill the high QoE expectation of customers [2].

The rapid development of mobile applications and explosive demands of mobile connectivity by end users present both challenges and opportunities for the future mobile network. On one hand, the future mobile network needs to have the ability to carry big data that has high volume and variety features. On the other hand, mining information hidden in the massive data traffic can improve users' QoE, provide personal service, and fully reveal the potential value of the future network [3, 4]. As a result, the analytics of mobile big data will become an indispensable tool for Internet service providers (ISPs) in terms of network deployment, resource management, and the design of future mobile network architectures.

It is well known that the characteristics of big data may be summarized as four Vs: volume (great volume), variety (various modalities), velocity (rapid generation), and value (huge value but very low density) [5]. Mobile big data is expected to be continuously generated by a versatile global infrastructure connecting users around the world. Since a lot of information regarding the network, application, and users can be extracted from the data traffic of mobile networks, including user ID, current time, location, uplink and downlink bytes, device type, protocol type, uniform resource locator (URL), and so forth, mobile big data has its own typical statistical characteristics, such as spatial-temporal distribution, data aggregation properties, and social correlations [6]. As a consequence, in order to measure and analyze mobile big data, the traffic monitoring and analysis system should have the capability to collect, store, and

Digital Object Identifier: 10.1109/MWC.2018.1700186

Yuanyuan Qiao (corresponding author), Zhizhuang Xing, and Jie Yang are with Beijing University of Posts and Telecommunications; Zubair Md. Fadlullah and Nei Kato are with Tohoku University.

1536-1284/18/$25.00 © 2018 IEEE

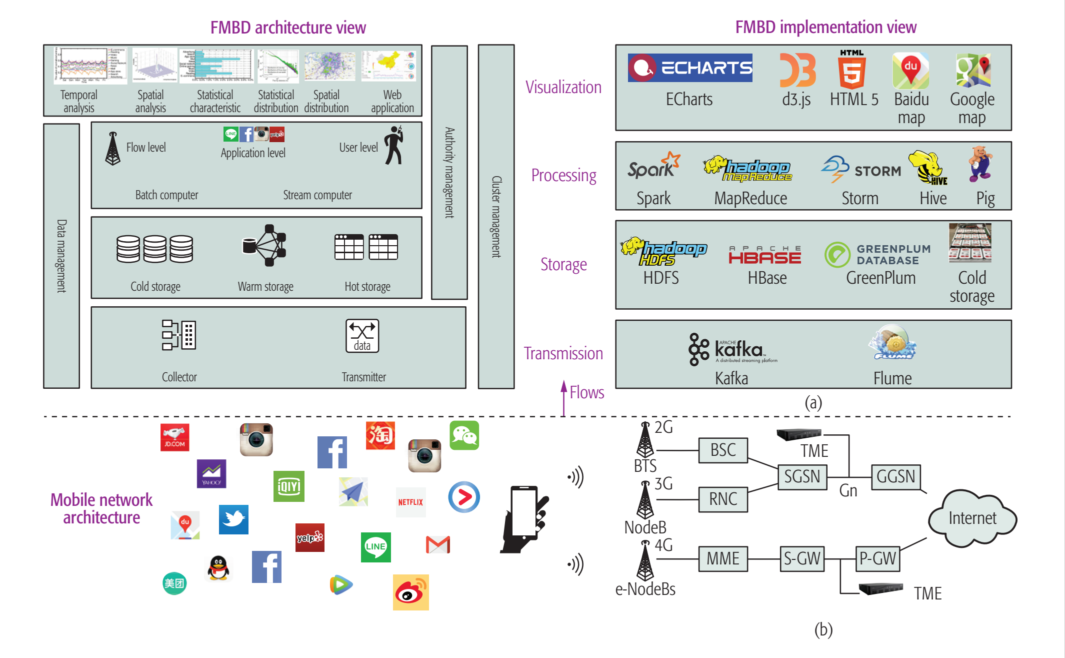


FIGURE 1. The overall architecture of FMBD and our considered mobile network architecture.

Unlike other frameworks for network traffic monitoring and analysis, in order to build a secure environment and easy-to-use platform for operators and data analysts, FMBD manages data traffic, machines, and users automatically. This allows FMBD to achieve data traffic analytics efficiency in a green, cost-effective fashion.

process massive data traffic efficiently and reliably. Furthermore, by taking into account the unique characteristics of mobile big data, it can provide useful recommendations to operators and users. For instance, in recent years, Apache open software tools [7] are widely applied to implement a highly parallel and distributed traffic monitoring and analysis system [8]. Thus, the technologies of big data continue to expand, which provides us with the techniques to gain deep insight into mobile big data.

In this article, we introduce the architecture of the framework for mobile big data, referred to as FMBD, which is designed to meet the needs of traffic monitoring and analysis in current and future mobile networks. FMBD has been operational for more than five years (since 2012) and deployed by ISPs in more than 10 provinces in China. FMBD is expanding and improving with practical needs, applications, and the development of big data technology. Unlike other frameworks for network traffic monitoring and analysis, in order to build a secure environment and easy-to-use platform for operators and data analysts, FMBD manages data traffic, machines, and users automatically. This allows FMBD to achieve data traffic analytics efficiency in a green, cost-effective fashion. In particular, according to the unique characteristics of mobile big data, we focus on flow-, application-, and user-level analysis, the results of which may demonstrate how to design a user- or even human- and data-centric architecture for current and future mobile networks. Based on real data traffic, FMBD continuously explores the patterns of data traffic consumption in mobile networks from operators' and individuals' viewpoints, which may provide potential benefits to improve QoE, optimize and balance resource usage, and hence make mobile networks effective and green.

The remainder of this article is organized as follows. We first present the architecture of FMBD. Then we discuss the mobile big data collection, storage, and processing. The FMBD system management is described in detail. Next, we present extensive analysis results on characterizing flows, applications, and user behavior based on real mobile big data, and introduce the potential value of mobile big data analytics for flexible deployment and optimal allocation of wireless resources. We then conclude the article.

Framework of Mobile Big Data

In this section, we first provide the details of our FMBD architecture and its implementation. The considered mobile network architecture is also described.

FMBD Architecture and Implementation

Figure 1a depicts the overall architecture of FMBD, which is designed to collect, store, process, present, and manage the data traffic of 2G/3G/4G networks. The implementations of most of the modules in the depicted framework are based on Apache open source software [7]. Furthermore, in order to provide a safe, stable, and efficient production environment to both users and operators, we define the following requirements for FMBD.

Performance: Based on Apache Spark and MapReduce, FMBD is able to process the analytics job of mobile big data in a streaming or batching way. It supports routine and also user-defined jobs that run periodically and complete users' analytics tasks, respectively. Computing resources are allocated to the running jobs according to their importance.

Energy efficiency: Currently, FMBD stores more than 600 TB of data traffic. It stores data in different storage systems according to the supported upper applications. FMBD consists of three types of data storage systems, namely, cold, warm, and hot storage systems. Data that are not accessed frequently, or may never be accessed, but which users wish to retain (e.g., logs or very old data), are stored in the cold storage system. On the other hand, data that need to be read

By 2021, a 5G connection is anticipated to generate 4.7 times more traffic, on average, than the state-of-the-art 4G connection. In order to ensure availability and smooth operations on a computer network, monitoring the network is indeed crucial to review, analyze, and efficiently manage the network traffic to combat any abnormality or process that can affect the network performance, availability, and/or security.

quickly at any time are stored in the hot storage system. In addition, rarely accessed data are stored in the warm storage system. Since the hot storage system consumes much more power and computing resources than that required by the cold storage system, in our considered environment, we allow only a limited amount of data to be stored in the hot storage system. In addition, the cluster management tool can monitor resource consumption and hence help us make the best use of cluster resources.

Portability: In our case, we mostly use FMBD to analyze data traffic. However, the architecture and application of FMBD are independent of the stored data and specific analysis jobs. Therefore, it could be used to provide big data analytical ability to other application scenarios.

Extensibility: The ability of storage and computing of FMBD could easily be increased or decreased by adding or removing machines or computing resources in clusters. The cluster management tool of FMBD analyzes the performance of each cluster, which exhibits the usage of the cluster resources.

Usability: Traditionally, for a given job, a developer writes the relevant code (in Java etc.), and submits the job by using command line in a Hadoop-based cluster. The data management tool of FMBD provides an interactive web-based interface in order to efficiently submit and manage the jobs. Furthermore, code packages of typical data traffic analysis jobs are made available through the web-based interface. Users with basic programming skills may select the input data and code package to initialize a job.

Security: The authority management tool of FMBD provides different authority roles to different groups of users. Identification, authentication, authorization, and access control functions are implemented to ensure the security of data and resources in the cluster without influencing users' daily usage.

Stability: The cluster management tool of FMBD monitors the status of machines, equipment, software, and everything running on the cluster at any time. In the case of a failure, alarm messages/emails are immediately issued to administrators to warn them about the failure. Thus, FMBD ensures that the system runs in an efficient and stable manner.

Next, we briefly describe our considered mobile network architecture in the remainder of this section.

Considered Mobile Network Architecture

Our considered mobile network architecture is also depicted in Fig. 1b. As shown in the figure, the data packets in the mobile networks are collected by our self-developed traffic monitoring equipment (TME), which monitors packets and aggregates them into flows in real time. Notice that we deploy TMEs at the core network edge connecting with the 2G/3G/4G network interfaces, which collect data traffic generated by user equipment (UE) (e.g., smartphones, tablets, laptops) equipped with mobile broadband adapters or any other device that can access the Internet through 2G/3G/4G networks. In 2G or 3G networks, each UE communicates with a base transceiver station (BTS) or Node B, which transmits its network traffic to the base station controller (BSC) or radio network controller (RNC). It is worth noting that the general packet radio service (GPRS) is considered in the 2G/3G networks. The BSC or RNC then delivers the network traffic to a serving GPRS support node (SGSN) that establishes a tunnel with the gateway GPRS support node (GGSN) on the Gn interface, through which data traffic enters the Internet. On the other hand, in case of the 4G network, an evolved Node B (eNodeB or eNB) establishes connection between the UE and the mobility management entity (MME). Data traffic flow to the Internet through the serving gateway (S-GW) and packet data network (PDN) gateway (P-GW).

In the following section, we discuss how our FMBD framework deals with the mobile big data generated by the aforementioned considered architecture. In particular, we focus on FMBD-based mobile big data acquisition and processing.

FMBD-Based Mobile Big Data Acquisition and Analytics

In this section, we discuss FMBD-based mobile big data acquisition and analytics. First, we describe how the large-scale mobile network traffic data are collected. Then the storage of mobile big data is delineated. Next, mobile big data processing is discussed.

The Collection of Large-Scale Mobile Network Traffic Data

By 2021, a 5G connection is anticipated to generate 4.7 times more traffic, on average, than the state-of-the-art 4G connection. In order to ensure availability and smooth operations on a computer network, monitoring the network (e.g., through network sniffing and packet capturing techniques) is indeed crucial to review, analyze, and efficiently manage the network traffic to combat any abnormality or process that can affect the network performance, availability, and/or security. In this vein, we develop a hardware- and software-based TME to collect and review each incoming/outgoing packet in the considered mobile networks described later. The functions and advantages of our envisioned software- and hardware-based TME are discussed below.

Hardware-based TME: The high-speed probe of TME can reach 200 Gb/s while mirroring the raw packages. TME can be deployed in the interface of the core network, IP backbone networks, exits of Internet data center (IDC) networks, provincial area networks, and metropolitan area networks. It can simultaneously monitor four 10G two-way links or a single 100G two-way link.

Software-based TME: TME mirrors uplink and downlink packets, and aggregates the packets into flows by using their five tuples {IP source address, IP destination address, source port number, destination port number, transport protocol}. In other words, a five-tuple flow is aggregated in terms of a sequence of packets that share the same five-tuple during a certain period (e.g., 64 s). It provides real-time protocol analysis, service identification, traffic statistics, and policy control functions for high-speed link traffic. Flow records generated by the deep packet inspection (DPI) processing engine of TME have information regarding the users, applications, and networks, which can be obtained from packets in real time. For example, the anonymized user ID, start and end time of each flow, IP addresses of the client and server sides, uplink and downlink bytes, protocol, specific service, and so forth can be extracted using our envisioned software-based TME.

The Storage of Large-Scale Mobile Network Traffic Data

Once a flow is collected by the hardware-/software-based TME, the collected flow record is uploaded to FMBD by utilizing a Kafka [7] based transmitter. Then the flow records are processed in real time with Spark streaming [7] or stored in the hot/warm/cold storage according to the needs of applications. The implementation approach and application scenarios are described below.

Hot storage: Hot data are frequently accessed on faster storage, for example, GreenPlum (greenplum.org) in our FMBD environment. Usually, hot data are processed or extracted from raw data, and directly accessed by interactive web applications. Based on the hot storage of FMBD, we develop the upper applications for analyzing the mobile network quality, understanding the user behavior, presenting the movements of a large population, and so forth.

Warm storage: Warm data are accessed less frequently and stored on a slightly slower storage device. We store our warm data in the Hadoop Distributed File System (HDFS) [7], including the flow records collected over a year, flow records used by ongoing projects, results of analysis jobs, and personal data uploaded by users.

Cold storage: This refers to the storage of inactive data that are not required to be accessed for months, years, decades, or even potentially ever. As a consequence, the data retrieval and response time for cold data storage systems are typically slower than services designed for active data manipulation. Therefore, cold data storage is much more economical than the high-performance primary storage used to support more active data. We design and develop our own cold data storage system, which stores rarely accessed flow records and logs collected over the span of several years.

The Processing of Large-Scale Mobile Network Traffic Data

After collecting and storing data flows, FMBD applies big data frameworks such as Apache Hadoop and Spark [7] for the processing of the mobile big data. Hadoop comprises a distributed data infrastructure, known as the HDFS, and a processing framework called MapReduce. On the other hand, Spark is a data processing tool that operates on those distributed data collections. While MapReduce operates in steps (batch-mode processing), Spark completes the full data analytics operations in-memory and in near real time (i.e., streaming-mode processing). In our considered environment, the analytics on streaming data, and using machine learning algorithms that require multiple operations, are based on Spark. On the other hand, the batch processing for massive data is carried out by MapReduce in order to avoid the out-of-memory failure.

We develop a set of MapReduce/Spark programs to analyze the mobile big data from the following three aspects.

Flow-level statistics: FMBD performs analytics regarding the traffic statistics in terms of the number of flows, flow duration, flow bytes, number of users, and flow packets with spatio-temporal dynamics. Regular patterns, specific fingerprints, or abnormal traffic can be identified to guide further analysis.

Application-level statistics: The explosion of mobile applications is driving the growth of global 4G deployments and adoption, soon to be followed by 5G growth [1]. FMBD investigates mobile applications running over HTTP to provide relevant insights into the current networks from a data traffic consumer perspective.

User-level analysis: Improving user experience is the ultimate goal of network operators. However, QoE is not taken into account in the contemporary mobile network ecosystem since it is mainly impacted by difficult-to-measure subjective factors (e.g., cost, reliability, efficiency, privacy, security, interface user friendliness, user confidence). To address this issue, FMBD provides analytics on the users' behavior, including online browsing and offline mobility behavior. Thus, the users' needs for mobile networks can be more accurately understood and even predicted.

Traditionally, the characterization of the access network traffic has been addressed from two diverse perspectives: a mobile operator viewpoint and a mobile user viewpoint [9]. Operators primarily focus on network characteristics and aggregated snapshots of many users within the coverage of specific areas. On the other hand, for a mobile user, researchers mainly aim at understanding how individual users consume mobile services. Unlike other radio signal services, 5G is anticipated to play a huge role in offering services to heterogeneous networks, technologies, and devices operating in different geographic regions to fulfill the high expectation of users with relatively low energy consumption. Satisfying the customers' requirements and providing green communication imply the necessity for moving from a system-centric design to a more user- or even human- and data-centric design paradigm. One of the overall guidelines for designing such an architecture should be "to keep the human in the loop" [10], which drives us to design our system providing the aforementioned flow-, application-, and user-level analysis.

In order to carry out the data acquisition, storage, and analytics in an efficient manner, it is therefore essential to understand how FMBD manages various functions. Therefore, in the next section, we provide the details of the FMBD system management.

System Management

As the clusters become bigger in FMBD, how to manage the clusters and keep them healthy appear to be an important issue. A cluster administrator may spend a lot of time installing new machines, changing the configurations, and solving the problems that stop any node in the clusters from running properly. Moreover, improving efficiency and ensuring data security are vital for managing the clusters in which we store mas-

Once a flow is collected by the hardware/software-based TME, the collected flow record is uploaded to FMBD by utilizing a Kafka based transmitter. Then, the flow records are processed in real time with Spark streaming or stored in the hot/warm/cold storages according to the needs of applications.

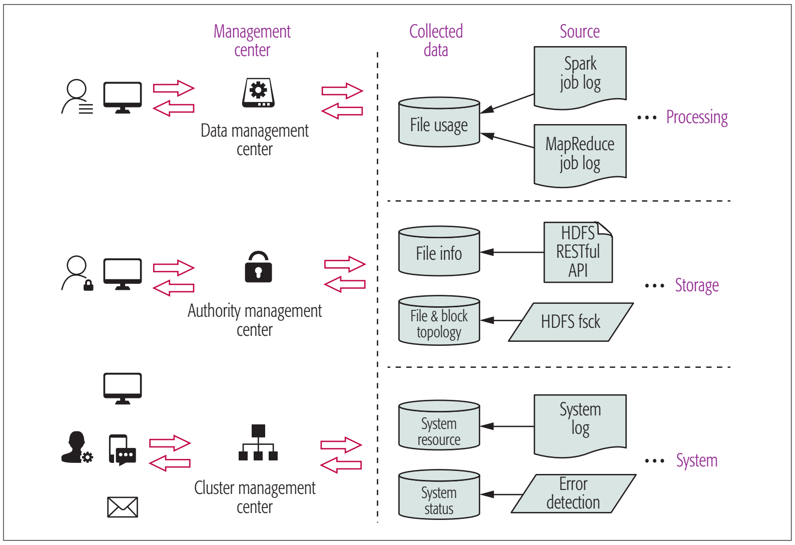


FIGURE 2. The architecture of data, authority, and cluster management tools for FMBD.

sive data and run jobs continuously. In order to understand how the clusters are performing while ensuring safe and efficient operation, FMBD provides cluster, data, and authority management functions as depicted in Fig. 2. To monitor, analyze, and manage machines, data, and users, FMBD collects source data from the operating system, storage system, and processing frameworks, and controls whole clusters by sending commands to each node.

Cluster Management

In a cluster, hundreds of machines are connected to one other through switches. Linux systems run on machines, while the software applications run on the operating system. To monitor the hardware and software in clusters, and then act quickly and effectively according to key indicators, are essential requirements in our considered environment. However, tracing and solving problems in such a highly distributed environment in time are indeed challenging, particularly when many users simultaneously execute different jobs on the clusters. Any human error, program bug, insufficient resources, hardware failure, network fault, and inappropriate configuration may result in serious operational risk for the clusters. Furthermore, in most cases, such major issues are difficult to identify because we can only observe a few failures that cause the issue. In order to address this challenge, we design a cluster management tool to collect information, metrics, and logs from machines in the clusters. Our cluster management tool is able to provide different levels of alarms if issues arise, and simultaneously send commands to all the machines in the clusters. Also, by using the cluster management tool, the administrators can check the status of the clusters, examine the results of statistical analysis for metrics, and make necessary modifications to the clusters through a user-friendly interface. In this way, we can identify any bottleneck and unbalanced load of cluster resources, and then optimize the resources accordingly.

The cluster management tool consists of three modules, namely, the collection, alarm, and configuration components. We briefly describe the design and function of each component as follows.

Collection Component: Flume [7] is used to collect performance metrics and alarm data in the collection component. Exec Source of Flume collects metrics from clusters, and collected information are listed below.

Hadoop/Spark/Storm/Hive/GreenPlum metrics: These metrics are collected through its logs, application programming interface (API), and the Java management extensions API.

Performance metrics (CPU, memory, disk Input/Output (I/O), and network I/O) of machines: These performance metrics are collected from files under the */proc* directory.

Machine/services/software status: These are collected by monitoring the heartbeat of the machine/services/software by periodically sending the Packet Internet Groper (PING).

The aforementioned collected data are transmitted to the master node through the Java Data Base Connectivity (JDBC) channel of Flume, and stored in the database through the self-developed JDBC and Alarm Sink of Flume.

Alarm component: Administrators receive an alarm massage/email/notification with the corresponding alarm level from the alarm component when the clusters encounter some anomalous event or operation. The alarm component reads the metric data stored in the database collected by the collection component as explained earlier. If the value of the monitored metric data is equal to the alarm value or larger than the alarm threshold value, an alarm is generated and sent to the administrators. It is worth noting that the administrators are able to set the alarm value/threshold, change the level of an alarm item, or even add a new alarm item. Common alarms include master nodes/services crash (first-level alarm), machine crash (middle-level alarm), CPU high load (low-level alarm), and so on.

Configuration component: Based on Zookeeper [7], the configuration component manages and configures machines, service, and software by sending Linux command lines or scripts. It maintains a queue on the producer/consumer mode, and a daemon on each machine runs the Linux command lines or scripts according to the received instructions. The web interface transmits the user's instructions and updates the name space of Zookeeper. On the other hand, Zookeeper sends the coordination data (user's instruction) from the server side to the client side. Finally, the daemons in each machine modify/delete/add metrics, start/stop service/software, and modify the configurations according to user instructions.

Data Management

In our clusters, data traffic is collected from 2G/3G/4G networks of several provinces in China. As the size of mobile data traffic continues to dramatically grow, it becomes increasingly difficult to trace the data. We usually are interested in the information on data usage in our clusters: What data are most frequently used? Which file has not been used for a significantly long time? Who always uses which files recently? Does the deletion of important data happen accidentally?

Who processes a lot of data recently? Which/how much data are uploaded to HDFS in the previous week? Are there some very huge or small files stored in the clusters? The answers to these questions help us understand the storage issue and usage of clusters. More importantly, monitoring the file usage is very important to ensure data security and keep the clusters healthy.

By tracing the life cycle of data stored in HDFS, we can improve the performance and reduce the resource consumption of clusters by putting the rarely used data into the cold data storage system, balance frequently used data, delete or merge small files, and catch abnormal behavior of users in the clusters. However, HDFS has no knowledge of (and does not take into consideration) what is stored inside the file. Furthermore, raw file is stored in accordance with rules that humans would not comprehend. In other words, the raw file is split into one or more blocks and these blocks are stored in various slave nodes in the cluster. Therefore, we design a data management tool to monitor uploading, storing, using, moving, changing, and deleting of data files by building a bridge between data blocks and files. Thus, we can understand the status of data from the file level (user view), job level (MapReduce view), and block level (HDFS/machine view) at the same time. In general, our data management tool provides data retrieval, statistics, and monitoring functions, as listed below:

- *Retrieval*: This provides the ability to retrieve files by using relevant keywords through the web interface.
- *Statistics*: The statistics show the number, size, and usage of files for each file owner, each directory, and each file type at different times. Also, they show the block (file) distribution among the machines to discover and reduce storage, and I/O hotspots.
- *Monitoring*: This means it is possible to monitor any operation to any file by any user, and also the changing and moving of data file blocks.

Authority Management

In our clusters, the flow records of data traffic are stored in the hot/warm/cold data storage systems according to the needs of the upper applications. The hot data storage is based on a relational database management system (RDBMS). On the other hand, the cold data storage is developed by our team. It can only be accessed by administrators and specific developers. However, the warm data storage system (i.e., HDFS) does not have authentication or authorization functions. The Kerberbos protocol provides the authentication function, which can be configured easily on Hadoop clusters. In order to ensure data security without affecting daily usage, we authorize different levels of authority to different groups of users, as delineated below:

- *Common user*: A common user is not allowed to access data that are not stored in his/her own user directory and owned by himself/herself.
- *Student user*: A student user is allowed to look through and use public data to run jobs on FMBD with a periodically updated password. However, he/she is not allowed to download public data.
- *Super user*: A super user is permitted to look through the public data and use it to run jobs on FMBD without a password. However, the super user is also not allowed to download public data.
- *Manager*: A manager is free to create, update, retrieve, and delete (CURD) any data on any directory in the HDFS.

The authorization function can be fulfilled by applying the Lightweight Directory Access Protocol (LDAP) to manage user accounts and users' operating authority. In addition, every action of a user is traced and recorded. Thus, the authority management tool provides identification, authentication, authorization, and access control functions.

In the following section, we evaluate the performance of FMBD through extensive experiments and real application deployment.

Experiments and Applications

In order to evaluate the performance of FMBD and present the results from real mobile big data, we conduct a series of experiments with real data traffic. In this section, we first introduce the workload and performance of FMBD in our environment. Then we present the analysis results from the flow, application, and user levels, which illustrate the theoretical and practical significance of the mobile big data analytics for optimizing network resources and improving user experience.

System Setup

FMBD runs on a cluster with 28 machines having the same hardware: 2 × Intel Xeon E5 2620 with 8 cores, 64 GB of RAM, and 3 TB × 12 serial attached small (SAS) computer system interface disks. All machines are connected by 1 Gb/s Ethernet optical fiber lines with 10Gbps switch. One machine has been dedicated to NameNode of Hadoop. The software environment of each of the nodes employs Linux CentOS v6.8 64 bit, Spark v1.3.1, and Hadoop v2.6.0 (CDH v5.4.8) with a block size of 256 MB and replication parameter of two.

For jobs running on the machines, four resources have big influence on its running time (i.e., CPU, memory, disk I/O, and network). If any of these four resources reach the limit, a bottleneck may occur, which causes the entire process to slow down or even stop. In order to keep our environment efficient and productive, it is crucial to adjust the resource allocation by changing the configurations of tools/software running on FMBD. Usually, different analysis jobs consume distinct resources; for example, WordCount is intended to be CPU intensive, while Terasort is disk I/O intensive. Our cluster management tool helps us to identify the resource bottleneck and adjust the resource of clusters accordingly.

By using the cluster management tool, the administrators can check the status of the clusters, examine the results of statistical analysis for metrics, and make necessary modifications to the clusters through a user friendly interface. In this way, we can identify the bottleneck and unbalance load of cluster resources, then optimize the resources accordingly.

Analysis and Visualization of Large-Scale Mobile Network Traffic Data

The explosion of mobile applications and phenomenal adoption of mobile connectivity by the end users lead to the need for optimized bandwidth management and network monetization. Traditionally, characterizing the traffic dynamics in a mobile network from the flow level is paramount in understanding how the access network resources are consumed by mobile users and

A mobile network must deliver data flows from the server to clients since hosts and applications work with flows but not packets. In addition, flows are the aggregation of packets, which exhibit not only the characteristics of packets, but also users' behavior in the higher layers.

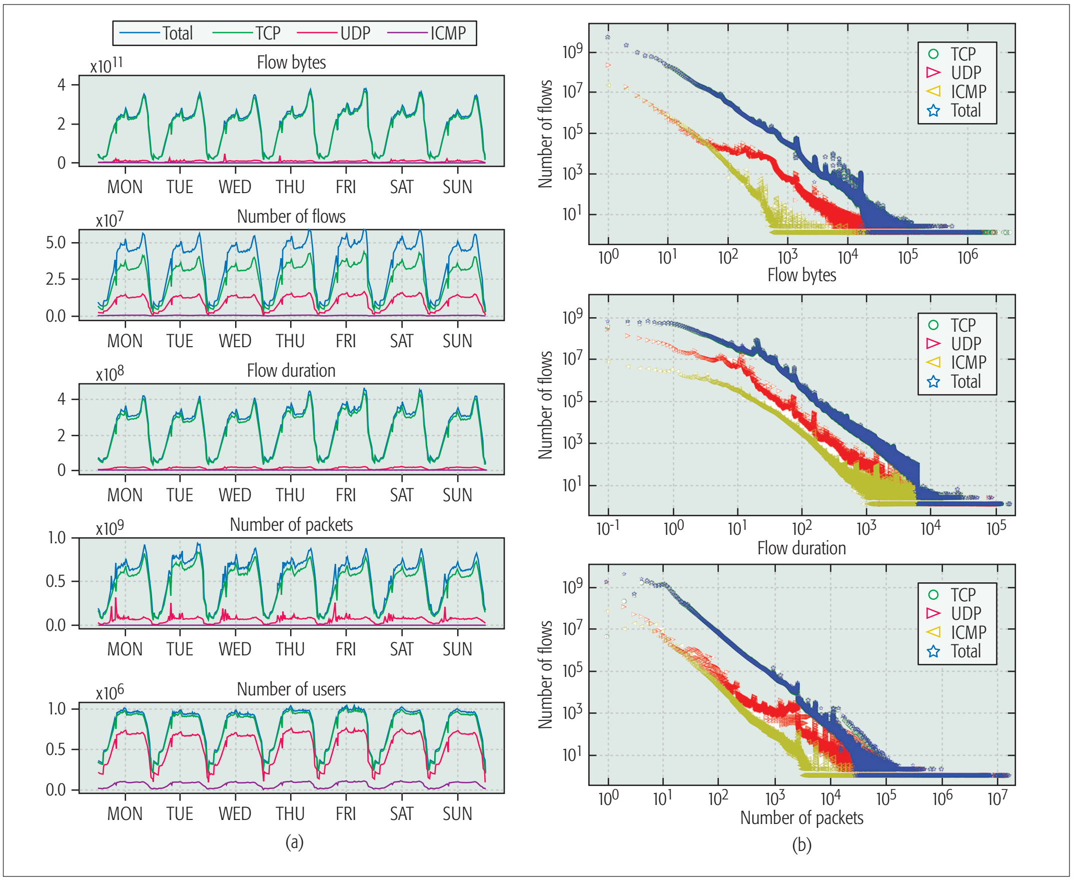


FIGURE 3. Flow level analysis: a) temporal characteristics of number of flows, flow duration (seconds), flow bytes (bytes), number of users, and flow packets (i.e., metrics changing with time); b) the frequency distribution of flow bytes (kilobits), flow duration (seconds), and number of flow packets are also shown.

improving a mobile network system. With 5G, resources (channels) will be allocated based on the awareness of content, users, and locations, which drives operators to analyze data traffic on the application and user levels. In recent years, mobile network users like to carry mobile phones and connect with the mobile network whenever and wherever possible. As mobile devices are becoming a sensor for sensing human behavior, studying users' spatio-temporal characteristics extracted from massive mobile network traffic data can reveal the traffic usage patterns in a city, which also has big potential to help operators to design the next generation network (NGN) with green communication and high QoE architecture. The real mobile big data used in our experiments were collected from a province in northern China on April 21 and 27, 2016. The dataset contains over 2.33 × 10^{10} flow records, which cover 12.5 million users and 0.15 million cell towers. Here, each user is identified by a distinct ID. According to the State Statistical Bureau of China, the total population of this typical province was about 38.1 million in 2015. In other words, our dataset covers nearly one-third of the total population in the province. In the remainder of this section, we present the analysis results on the flow, application, and user levels based on our FMBD. Unprecedented opportunity and big potential show up when we gain in-depth understanding of the large-scale mobile network traffic data using big data technology.

Flow Level Analysis: A mobile network must deliver data flows from the server to clients since hosts and applications work with flows but not packets. In addition, flows are the aggregation of packets, which exhibit not only the characteristics of packets, but also users' behavior in the higher layers. The existing research works demonstrate that mobile traffic tends to follow regular temporal patterns. Hence, in Fig. 3a, we examine the temporal dynamics of our data traffic by drawing the characteristics of several metrics: the number of flows, flow duration, flow bytes, number of users, and flow packets. We calculate each data point with 15 minutes time granularity. As depicted in Fig. 3a, during April 21 and 27, 2016, five time-series graphs of the total (i.e., flows for all protocols), Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and Internet Control Messages Protocol (ICMP) flows, demonstrate a clear daily pattern. TCP flows dominate the whole network on all the examined metrics. There is a clear peak for flow bytes, flow duration, and number of flow packets between 9 p.m. and 11 p.m. every day. This implies that the users tend to generate more traffic before bedtime.

In addition, in order to capture the statistical characteristics of the actual traffic in the mobile network, the frequency distribution of flow bytes, flow duration, and number of flow packets are also plotted in Fig. 3. It can be noticed from the results that 80 percent of the flows are less than 3 kb. This means that most users use very limited

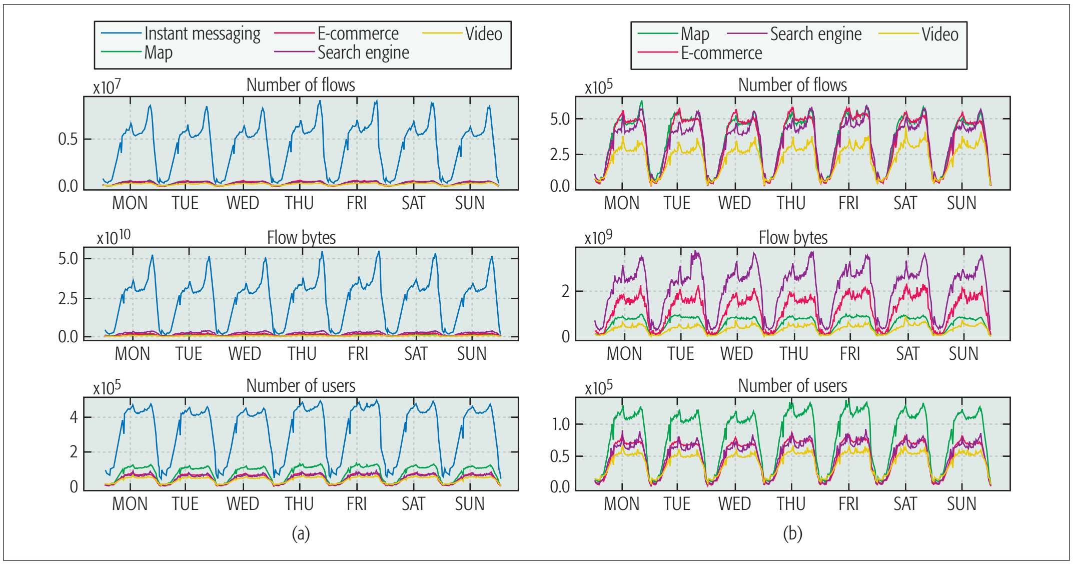


FIGURE 4. Application level analysis: temporal characteristics of the number of flows, flow bytes, and number of users for popular application categories, including: a) instant messaging, e-commerce, search engine, map, and video applications; b) e-commerce, search engine, map, and video applications.

The flow-level analysis demonstrates mobile traffic characteristics from the global user demand viewpoint. By considering together the temporal factor, spatial factor, and application usage behavior, we understand when, where, how, and why the access network resources are consumed by mobile users.

data traffic everyday, and a small group of users generates a great amount of data traffic in the network. It implies that by focusing on the heavy users, who consume much data traffic, we may be able to optimize network resources. In contrast with the fixed network, the distributions of flow bytes, duration, and number of packets for the mobile network are sharper with the increasing value of the above metrics. This is because the bandwidth of a 2G/3G network is much smaller than that of the fixed network.

Application Level Analysis: In this part, we present an application level analysis of the considered mobile network. In particular, we focus on applications running over HTTP, which contribute the major portion of the traffic in the mobile Internet. Each HTTP flow contains the following details: a user's anonymized ID, a timestamp, the accessed URL, and the cell ID. Therefore, the applications users use are distinguished by keywords in URLs, including *twitter*, *mail.google*, *map.google*, and so forth.

There are thousands of applications in the application store, so we filter out the top 100 popular applications, which contribute more than half of the total traffic, to study the traffic characteristics at the application level. Among the top 100 applications, five categories of applications generate more than 80 percent of the total traffic: instant messaging (WeChat, QQ, etc.), e-commerce (Taobao, JD, Meituan, etc.), search engine (Baidu search, Sogou search, and other search engines), maps (Baidu map, QQ map, and Amap), and video (QQ Video, Youku, and Iqiyi). We investigate the number of flows, flow bytes, and number of users of the above application categories, as demonstrated in Fig. 4a. By examining the temporal dynamics of these five application categories, it is evident that instant messaging has the largest number of users and consumes the most amount of data traffic. If we focus on the remaining four categories, as shown in Fig. 4b, we can find that the map applications have more than 0.1 million users online during peak hours (around 7:15 p.m. and 10:30 a.m.), but generate less than 6.87 kb data traffic for each user. However, for the search applications, the entrance for other websites, each user consumes, on average, 43.35 kb data in the peak hour (i.e., around 8 p.m.).

Researchers have proven that the dynamic urban mobile traffic usage exhibits only five basic time domain patterns among thousands of cellular towers, which implies that the spatial and temporal characteristics of traffic are correlated [11]. Here, we further investigate the spatial and temporal characteristics of the traffic amount for different application categories, as demonstrated in Fig. 5. In different functional areas (i.e., comprehensive areas, entertainment districts, residential zones, and work areas), instant messaging still contributes the maximum amount of data traffic, as shown in Fig. 5a. On the other hand, in Fig. 5b, the traffic patterns of the other categories tend to follow similar patterns, as discussed earlier. This means that the users' traffic usage patterns among thousands of applications in the mobile Internet are mainly impacted by the spatial and temporal characteristics. In terms of the traffic of the residential zones, there is a sharp peak at 8 p.m. However, the peak is around 10 a.m. in work areas during weekdays. Furthermore, the users in the entertainment districts always consume more traffic during 9 a.m. and 6 p.m., and consume even more traffic on weekends. The comprehensive areas are a mixture of other kinds of functional areas, and the traffic on them shows similar patterns with the total traffic. It exhibits two traffic peaks in a day, at around 11 a.m. and 8 p.m., respectively. Based on the above observations, while allocating network resources, the ISPs should consider the traffic usage patterns of users in both spatial and temporal dimensions.

User-Level Analysis: The flow-level analysis demonstrates mobile traffic characteristics from the global user demand viewpoint. By considering together the temporal factor, spatial factor, and application usage behavior, we understand when, With real mobile big data, the experimental results demonstrated that the recent advances of wireless technologies and ever-increasing mobile applications require a mobile network with a significantly larger data traffic capacity, and a traffic monitoring and analysis system that can deal with the mobile big data.

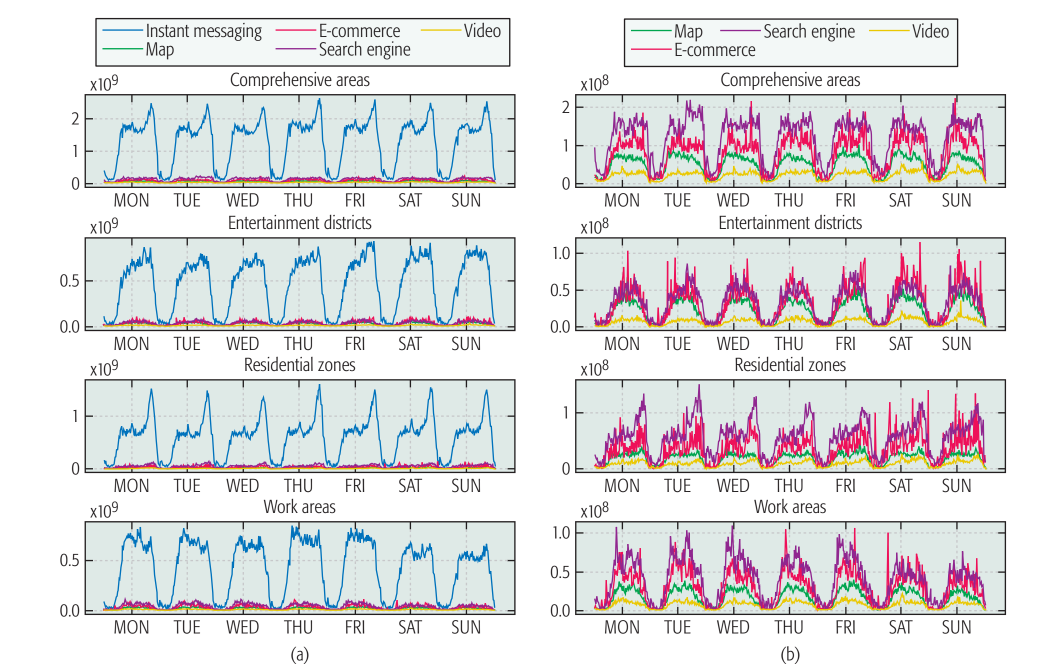


FIGURE 5. Application-level analysis: The amount of traffic (flow bytes) for popular application categories: a) instant messaging, e-commerce, search engine, map, and video applications; b) e-commerce, search engine, map, and video applications, in comprehensive areas, entertainment districts, residential zones, and work areas of city.

where, how, and why the access network resources are consumed by mobile users. Since the user behavior in the mobile network can reveal human dynamics [12], mobile big data analytics also refers to the discovery of previously unknown meaningful patterns and knowledge from data collected from mobile users [13]. In order to effectively take human factors into consideration in the next generation mobile networks, in the remainder of this section, we explore the relationship between users' application usage behavior and daily mobility behavior, the results of which may provide personalized service for users and help to allocate network resources in advance by predicting users' future behavior.

Recently, fingerprints/patterns of data traffic have been found to identify the function of physical locations. This means that the location characteristics can be discovered by data traffic patterns [11, 14, 15]. Based on previous studies, in our conducted experiment, we investigate the predictability of mobile app usage behavior and mobility behavior of users by observing users' mobility behavior/app usage behavior. As a user's mobility/app usage behavior can be characterized by a Markovian stochastic process, we build a hidden Markov model (HMM)-based predictor. Figures 6a and 6b demonstrate the example of instantiated HMM for app usage behavior and mobility behavior modeling that has five time slots. Prediction results are presented with a graph of CDF in Fig. 6c. If we know the current time and user's points of interest (POIs), for 23 percent of users, we can predict the app a user uses with 60 percent prediction accuracy. For more than 14 percent of users, predicting POIs that he/she may visit by only considering the current time and the apps he/she uses, we can achieve up to 60 percent accuracy. These results imply that there is a strong correlation between users' app usage behavior, mobility behavior, and data traffic patterns. As a consequence, it is possible to design a system that can actively learn, predict, adapt, and steer user behavior, so as to greatly improve the system efficiency and provide users with superior QoE levels.

Conclusion

In this article, we present FMBD, a framework for mobile big data, for operators and data analysts. FMBD provides data collection, storage, processing, analyzing, and management functions to monitor and analyze massive data traffic. With real mobile big data, the experimental results demonstrate that the recent advances of wireless technologies and ever ncreasing mobile applications require a mobile network with significantly larger data traffic capacity, and a traffic monitoring and analysis system that can deal with mobile big data. In addition, it is also revealed that since the difference between usage behaviors of different application categories is huge, designing different strategies of network resource allocation and management for specific applications can provide green communication for users. Furthermore, the results indicate that due to the highly spatio-temporal and non-homogeneous nature of data traffic, a reasonable allocation of network resources is challenging but indispensable to reduce resource consumption and improve QoE.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (61701031), Funds of Beijing Laboratory of Advanced Information Networks of BUPT, the Funds of Beijing Key Laboratory of Network System Architecture and Convergence of BUPT, the 111 Project of China

(B08004), and EU FP7 IRSES MobileCloud Project (612212). This work is an achievement of an academic exchange between Tohoku University and Beijing University of Posts and Telecommunications.

References

[1] C. V. N. Index, "Global Mobile Data Traffic Forecast Update, 2016–2020," Cisco white paper, 2017.

[2] E. Liotou *et al.*, "Shaping QoE in the 5G Ecosystem," *2015 7th Int'l. Wksp. Quality of Multimedia Experience*, 2015, pp. 1–6.

[3] P. Zhang *et al.*, "Opportunities and Challenges of Wireless Networks in the Era of Mobile Big Data," *Chinese Science Bulletin*, vol. 60, no. 5–6, 2015, pp. 433–38.

[4] M. Li *et al.*, "SPFM: Scalable and Privacy-Preserving Friend Matching in Mobile Cloud," *IEEE Internet of Things J.*, vol. 4, no. 2, 2017, pp. 583–91.

[5] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, vol. 19, no. 2, 2014, pp. 171–209.

[6] X. Zhang *et al.*, "Social Computing for Mobile Big Data," *Computer*, vol. 49, no. 9, 2016, pp. 86–90.

[7] Apache Software Foundation, 2017; http://apache.org.

[8] L. Qian, J. Zhu, and S. Zhang, "Survey of Wireless Big Data," *J. Commun. and Info. Networks*, vol. 2, no. 1, 2017, pp. 1–18.

[9] D. Naboulsi *et al.*, *Mobile Traffic Analysis: A Survey*, Ph.D. dissertation, Université de Lyon, INRIA, Grenoble-Rhône-Alpes, 2015.

[10] J. S. Silva *et al.*, "People-Centric Internet of Things," *IEEE Commun. Mag.*, vol. 55, no. 2, Feb. 2017, pp. 18–19.

[7] H. Wang *et al.*, "Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment," *Proc. 2015 ACM Internet Measurement Conf.*, 2015, pp. 225-–38.

[12] Y. Qiao *et al.*, "A Mobility Analytical Framework for Big Mobile Data in Densely Populated Area," *IEEE Trans. Vehic. Tech.*, vol. 66, no. 2, 2017, pp. 1443–55.

[13] D. Z. Yazti and S. Krishnaswamy, "Mobile Big Data Analytics: Research, Practice, and Opportunities," *2014 IEEE 15th Int'l. Conf. Mobile Data Management*, vol. 1, 2014, pp. 1–2.

[14] A. K. Das *et al.*, "Contextual Localization Through Network Traffic Analysis," *2014 Proc. IEEE INFOCOM*, 2014, pp. 925–33.

[15] Y. Qiao *et al.*, "Mobile Big-Data-Driven Rating Framework: Measuring the Relationship between Human Mobility and App Usage Behavior," *IEEE Network*, vol. 30, no. 3, 2016, pp. 14–21.

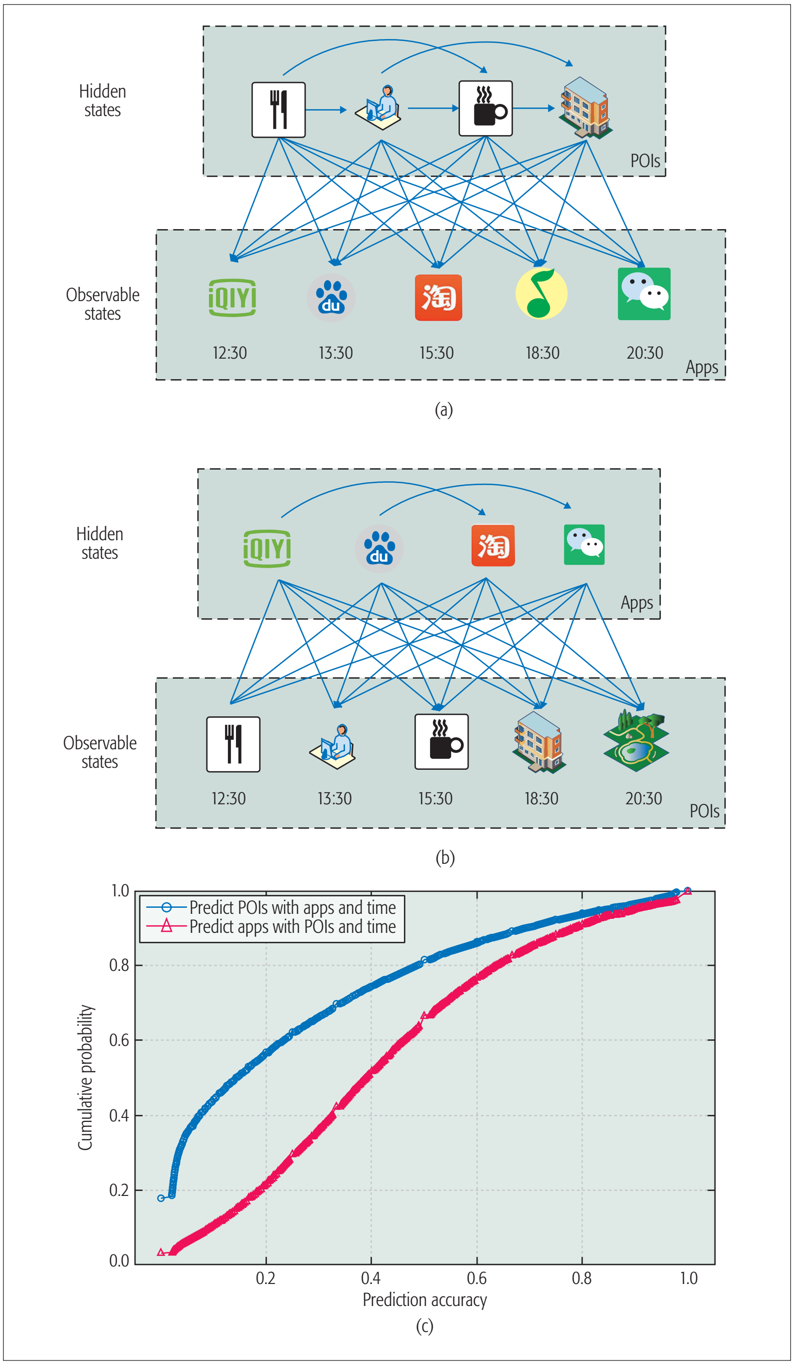
Biographies

Yuanyuan Qiao (yyqiao@bupt.edu.cn) [M'15] is an associate professor in the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), China. She received her B.E. degree from Xidian University in 2009, and received her Ph.D degree from BUPT in 2014. Her research focuses on mobile big data based urban computing, and big data analysis in the Internet and telecom.

Zhizhuang Xing is a graduate student in the School of Information and Communication Engineering, BUPT, and received his B.E. degree in information engineering from BUPT in 2016. He is engaged in the research of human mobility analysis and big data analytics.

Zubair Md. Fadlullah [M'11, SM'13] is an associate professor at the Graduate School of Information Sciences (GSIS), Tohoku University, Japan. His research interests are in the areas of 5G, smart grid, network security, deep learning, game theory, and quality of security service provisioning mechanisms. He was a recipient of the prestigious Dean's and President's Awards from Tohoku University in March 2011, the IEEE Asia Pacific Outstanding Researcher Award in 2015, and the NEC Tokin Award for research in 2016 for his outstanding contributions. He has also received several best paper awards at conferences including IWCMC '09 and IEEE GLOBECOM '14.

Jie Yang received her B.E., M.E., and Ph.D. degrees from BUPT in 1993, 1999, and 2007, respectively. She is now a professor and deputy dean of the School of Information and Communication Engineering, BUPT. Her current research interests include broadband network traffic monitoring, user behavior analysis, big data analysis in the Internet and telecom, and more. She was the Vice Program Committee Co-Chairs of IEEE IC-NIDC 2016, 2014, and 2012.



(a)

(b)

(c)

Figure 6. User-level analysis: HMM-based models for predicting users' future: a) mobility behaviors with app usage behavior and time; b) app usage behavior and time with mobility behaviors; c) their prediction accuracy.

Nei Kato [M'04, SM'05, F'13] is a full professor and the director of the Research Organization of Electrical Communication (ROEC), Tohoku University. He has been engaged in research on computer networking, wireless mobile communications, satellite communications, ad hoc and sensor and mesh networks, smart grid, IoT, big data, and pattern recognition. He is the Vice-President-Elect (Member & Global Activities) of the IEEE Communications Society (2018–2019) and Editor-in-Chief of *IEEE Transactions on Vehicular Technology* (2017–). He served as a Member-at-Large on the Board of Governors, IEEE Communications Society (2014–2016), and as the Chair of the Satellite and Space Communications Technical Committee (2010–2012) and Ad Hoc & Sensor Networks Technical Committee (2014–2015) of the IEEE Communications Society. He is a Distinguished Lecturer of the IEEE Communications Society and Vehicular Technology Society. He is a Fellow of IEICE.