

Social Computing for Mobile Big Data

Xing Zhang, Beijing University of Technology and Beijing University of Posts and Telecommunications

Zhenglei Yi, Beijing University of Posts and Telecommunications

Zhi Yan, Hunan University

Geyong Min, University of Exeter

Wenbo Wang, Beijing University of Posts and Telecommunications

Ahmed Elmokashfi and Sabita Maharjan, Simula Research Laboratory

Yan Zhang, Simula Research Laboratory and University of Oslo

Mobile big data contains vast statistical features in various dimensions, including spatial, temporal, and the underlying social domain. Understanding and exploiting the features of mobile data from a social network perspective will be extremely beneficial to wireless networks, from planning, operation, and maintenance to optimization and marketing.

Data services' exponential growth, and the constantly expanding wireless and mobile applications that use them, have ushered in an era of big data. This mobile big data—from applications such as planning, operations and maintenance, optimization, and marketing—poses many new challenges to conventional data analytics because of its large dimensionality, heterogeneity, and complex features. To help address this problem, we provide a classification structure for mobile big data and highlight several

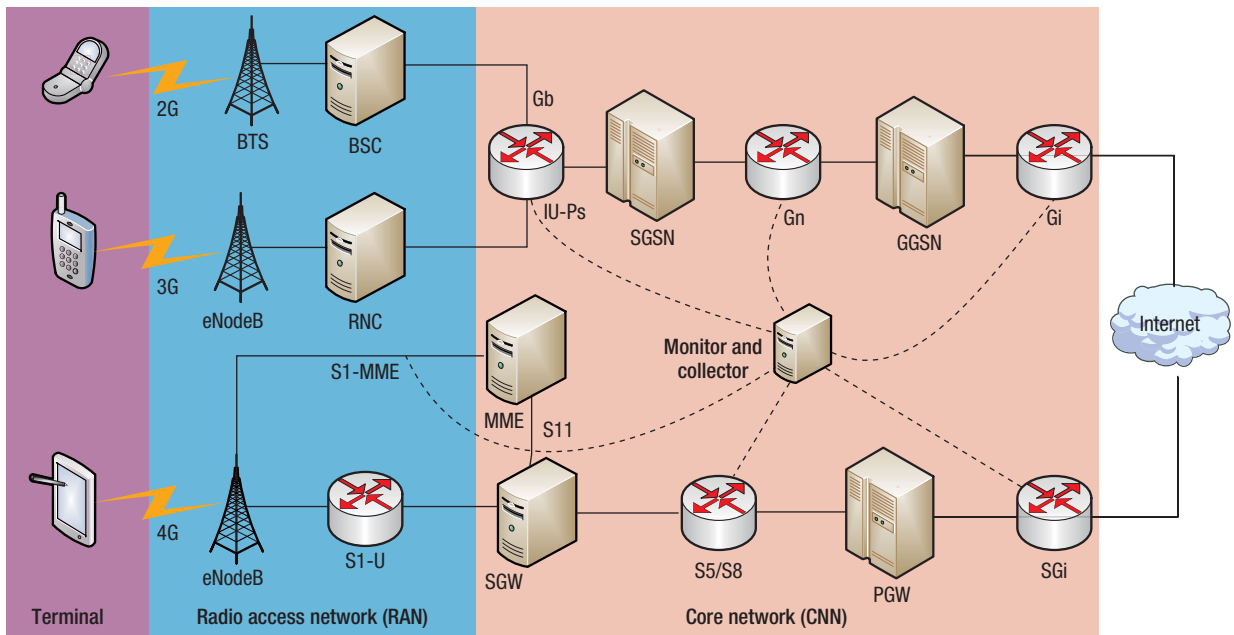


Figure 1. Mobile cellular network architecture and data-collecting points. BSC: base station controller; BTS: base transceiver station; eNodeB: Evolved Node B; GGSN: Gateway GPRS Support Node; Iu-PS: Packet Switched Core Network; MME: Mobility Management Entity; PGW: Packet Data Network Gateway; RNC: Radio Network Controller; SGi: Serving Gateway interface; SGSN: Serving GPRS Support Node; SGW: Serving Gateway.

research directions from the perspective of social computing using a significant volume of real data collected from mobile networks.

Conceptually, *big data* refers to not only a large volume of data but also other features that add complexity to it, thus giving it several unique characteristics and differentiating it from huge amounts of data.

DATA CATEGORIES IN MOBILE CELLULAR NETWORKS

Vast amounts of mobile data are collected and extracted from several key network interfaces, in both radio access networks and core networks, as shown in Figure 1. This data can be roughly classified into four categories: flow record data,

network performance data, mobile terminal data, and additional data information.

Flow record data

A *flow* is a collection of packets between two end nodes defined by specific attributes, such as the well-known TCP/IP five-tuples. The flow record data in cellular networks is typically obtained through deep packet inspection and deep flow inspection, which includes both data records and signaling records, in the form of XDR (call/transaction detail record). These data records contain the main attributes during a data-connection session. The flow record is perhaps the most important data for describing user behavior. For a medium-size city (1 million

subscribers), the total flow record data is about 70 Tbytes (approximately 10^{15} records) in one week.

Network performance data

Network performance data refers to aggregated datasets collected during a certain period (for example, 5 minutes, 15 minutes, or an hour). Network performance data mainly includes the key performance indicator (KPI) data and the measurement report (MR; a statistical data report that contains information about channel quality). KPIs are widely used in mobile cellular networks with the goal of evaluating the network performance and quality of service delivered to users. Network performance data can grow to several Tbytes in one week for a medium-sized city.

Mobile terminal data

With the development of the Internet of Things (IoT) and 4G cellular networks, a rapidly increasing number of connected devices are generating a large amount of data at terminals. Two typical categories of mobile terminal data are smartphones and the IoT. Terminal data can be collected through a

example, 5 minutes, 1 hour, or even 1 day). Similarly, in the spatial domain, mobile data can be studied from a whole city, a typical area, one base station (BS), or even one cellphone. Heterogeneity and fluctuation commonly exist in both spatial and temporal distributions. Mobile traffic in different locations such as stadiums, campuses,

utilized from various perspectives. We investigate the potential performance gain of analyzing mobile big data from the perspective of social network analysis (SNA). SNA is an interdisciplinary field that has been widely used to study the relationship among individuals, groups, or even entire societies.

Vast amounts of mobile data are collected and extracted from several key network interfaces, in both radio access networks and core networks.

mobile app. The data contains service characteristics, device information, and wireless network parameters such as international mobile subscriber identity, cell ID, signal strength, and download/upload rate.

Additional data

Some very important basic data information also exists in cellular networks, which can be summarized as subscriber profiles and geographic information data. The subscriber profile includes a user's billing information and data plan. The geographic information data contains the location of the cellphone or the point of interest information. Basic data information is relatively static compared with other data; however, it's vital for supporting data analysis.

STATISTICAL CHARACTERISTICS OF MOBILE BIG DATA

Unlike traditional big data in computer networks, mobile big data has its own typical statistical characteristics. In this article, we focus on three main statistical characteristics for mobile data: spatial-temporal distribution, data aggregation properties, and social correlations.

Spatial-temporal distribution

Mobile network data can be partitioned to different temporal granularities (for

and dense residential areas exhibits different patterns in different time scales. Full utilization of such characteristics has already been exploited in the deployment of mobile networks.

Data aggregation properties

Usually, the characteristics of aggregation properties for mobile data are more important for network performance optimization. We focused on group behavior rather than individual behavior.² For a given geographical area and a certain time period, group subscribers will likely request the same traffic, resulting in a similar traffic pattern. The aggregation features will be further used in predictive models to improve network performance.

Social correlations

Because of the social nature of human beings, mobile users in close proximity tend to exhibit similar habits, behavior, and mobility rules.³ For example, the social correlations of users leads to a larger scale of data-traffic correlations in both temporal and spatial domains, such as the autocorrelation and cross-correlation properties of mobile traffic. Thus, social correlations are a universal phenomenon in mobile data.

SOCIAL CHARACTERISTICS OF MOBILE BIG DATA

In current literature and applications, mobile big data has been studied and

User social networks

A user social network (USN) describes the social relationship among users and has recently received a great deal of attention in many fields. As for mobile data, several classical methods have been proposed to construct a USN by using call detail records. In such cases, social graph construction approaches use phone numbers (users) for the nodes and call connection for the graph edges. In recent years, as more people use mobile devices to access the Internet—including online social networks such as Facebook and Twitter—we see a convergence of social and mobile networks. People tend to highly value the content recommended by friends or people with similar interests; thus, the network edges can be easily defined in several ways, such as the content shared between two subscribers, their common interests, and the habits of each individual. Once established, the USN can be used to detect community structures, understand true communication behaviors, and build a user-centric wireless network.

Base station social networks

In mobile networks, it's difficult to obtain users' accurate positions. However, the location of a BS or cell tower can provide rough spatial information for wireless traffic, which is sufficient from the perspective of a large metropolitan area. Here, we present the concept of a BS social network (BSSN) and construct one with collected traffic data from Hong Kong's LTE network.

Each node in a BSSN is a BS. Unlike a USN, the edges in a BSSN represent

relationships rather than real social ties. To construct a BSSN, the relationship between BSs is quantified with the BSs' traffic traces by using the Pearson correlation coefficient. Then, a planar maximally filtered graph is applied to filter the relationship to obtain a BSSN.

Several interesting studies can be performed based on a BSSN, such as link prediction, community detection, and social influence analysis, but we mainly focus on a BSSN's community structure. "Community" refers to a subgraph structure in which nodes have a higher density of edges, whereas vertices between subgraphs have a lower density. Analytical results show that each community in a BSSN corresponds to a typical scenario with a common traffic pattern, which can be used to recognize typical traffic scenarios for mobile networks.

App social networks

With the popularity of smartphones and mobile applications, many unique mobile apps continue to emerge, such as location-based services, games, and shopping applications. Data traces collected from mobile networks reflect use of a large number of applications. Using a similar method to BSSN construction, an app social network (ASN) is established with the real spatial-temporal cellular network traffic data collected from Shenzhen's 3G network, as shown in Figure 2.

In an ASN, each node represents one typical app, and the edges indicate the strength of the relationships among various apps. In Figure 2, the node (app) with a higher degree has a larger size. This indicates that this app tends to be used with various other applications at a specific time; and apps of the same color belong to the same community, such as video, instant messaging, office applications, and others.

Interaction between social networks

With the vast amounts of data collected from mobile cellular networks,

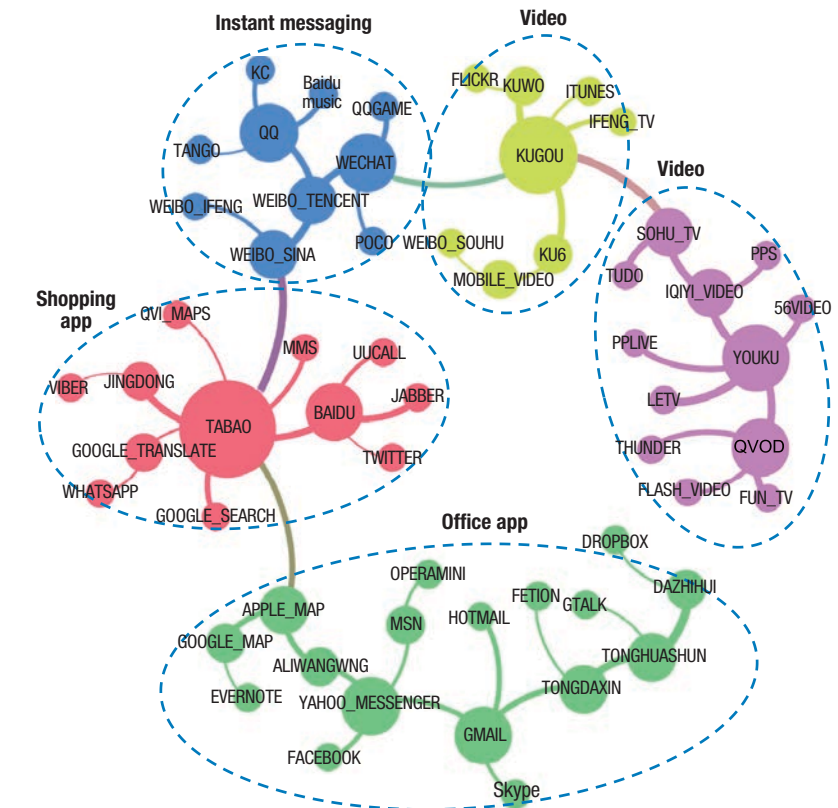



Figure 2. App social networks. Each node represents one typical app, and the edges indicate the strength of relationships among various apps. The node (app) with a higher degree has a larger size; apps of the same color belong to the same community, such as video, instant messaging, office applications, and others.

USNs, BSSNs, and ASNs are established from the SNA perspective: USNs mainly reflect on the social ties between users and model their behavior characteristics, BSSNs study the traffic relationship of each BS, and ASNs represent the usage patterns of diverse apps. With the emergence of new mobile applications, we see the convergence of these three networks. For example, *Pokemon Go* (a game in which people use real-time maps to search for *Pokemon* characters) recently became very popular all over the world. Players in the same vicinity (within one BS) share the same live map and can cooperate with each other. Hence, this social network can be exploited to provide a quality experience by utilizing the app's data.

Our results span a wide range of subjects in the field of social computing for mobile big data; however, there are many open questions. First, new social relationships (besides USN, BSSN, and ASN) from mobile big data need to be explored. Second, the dynamics and evolution of mobile big data's social characteristics should be investigated. Third, how do we exploit the social characteristics from mobile big data to design and optimize practical cellular networks?

Along with the evolution of mobile network architecture and other technologies like virtual and augmented reality, social computing will be incorporated in all aspects of cellular networks. Operators and

app developers face a great challenge in making full use of the available information. 

REFERENCES

1. W. Fan et al., "Sensing and Monitoring for Cellular Networks: A Crowdsourcing Platform from Mobile Smartphones," *Proc. IEEE Int'l Conf. Data Science and Data Intensive Systems (DSDIS 15)*, 2015, pp. 472–473.
2. X. Zhang et al., "Energy-Efficient Multimedia Transmissions through Base Station Cooperation over Heterogeneous Cellular Networks Exploiting User Behavior," *IEEE Wireless Comm.*, vol. 21, no. 4, 2014, pp. 54–61.
3. X. Zhang et al., "Enhancing Spectral-Energy Efficiency for LTE-Advanced Heterogeneous Networks: A User's Social Pattern Perspective," *IEEE Wireless Comm.*, vol. 21, no. 2, 2014, pp. 10–17.



Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

XING ZHANG is a full professor in the Beijing Advanced Innovation Center for Future Internet Technology at the Beijing University of Technology and the Key Laboratory of Universal Wireless Communication, Ministry of Education, School of Information and Communications Engineering at the Beijing University of Posts and Telecommunications. Contact him at zhangx@ieee.org.

ZHENGLEI YI is a master's student in the School of Information and Communications Engineering at the Beijing University of Posts and Telecommunications. Contact him at zhengleiyi@bupt.edu.cn.

ZHI YAN is an assistant professor in the School of Electrical and Information Engineering at Hunan University. Contact him at yanzhi@hnu.edu.cn.

GEYONG MIN is chair and director of the High-Performance Computing and Networking Research Group at the University of Exeter. Contact him at g.min@exeter.ac.uk.

WENBO WANG is a full professor in the Key Laboratory of Universal Wireless Communication, Ministry of Education, School of Information and Communications Engineering at the Beijing University of Posts and Telecommunications. Contact him at wbwang@bupt.edu.cn.

AHMED ELMOKASHFI is a Senior research scientist at Simula Research Laboratory. Contact him at ahmed@simula.no.

SABITA MAHARJAN is a postdoctoral fellow at Simula Research Laboratory. Contact her at sabita@simula.no.

YAN ZHANG is chief scientist at Simula Research Laboratory and a professor in the Department of Informatics at the University of Oslo. Contact him at yanzhang@ieee.org.