IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 30 June 2014; revised 30 September 2014; accepted 2 December 2014. Date of publication 17 December, 2014;
date of current version 6 March, 2015.

Digital Object Identifier 10.1109/TETC.2014.2381512

Characterizing User Behavior in Mobile Internet

JIE YANG[1], YUANYUAN QIAO[1], XINYU ZHANG[1], HAIYANG HE[1], FANG LIU[1], AND GANG CHENG[2]

[1]Beijing Key Laboratory of Network System Architecture and Convergence, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Microsoft, Redmond, WA 98052 USA

CORRESPONDING AUTHOR: J. YANG (janeyang@bupt.edu.cn)

This work was supported in part by the Director Foundation Project under Grant 2014BKL-NSAC-ZJ-01, in part by the Important National Science and Technology Specific Projects under Grant 2012ZX03002008, in part by the European Union Seventh Framework Programme Marie Curie International Research Staff Exchange Scheme project MobileCloud under Grant 612212, in part by the 111 Project of China under Grant B08004, and in part by the National Natural Science Foundation of China under Grant 61072061.

ABSTRACT Smart devices bring us the ubiquitous mobile accessing to Internet, making mobile Internet grow rapidly. Using the mobile traffic data collected at core metropolitan 2G and 3G networks of China over a week, this paper studies the mobile user behavior from three aspects: 1) data usage; 2) mobility pattern; and 3) application usage. We classify mobile users into different groups to study the resource consumption in mobile Internet. We observe 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 ways to attract the users. Our analysis provides an comprehensive understanding of user behavior in mobile Internet, which may be used by network operators to design appropriate mechanisms in resource provision and mobility management for resource consumers based on different categories of applications.

INDEX TERMS Mobile Internet, network traffic, data usage, mobility pattern, user behavior.

I. INTRODUCTION

A. BACKGROUND AND PROBLEM STATEMENT

Smart phones are becoming ubiquitous providing the continuous Internet access. With the increasing popularity of Mobile Internet access, an exhaustive understanding of user behavior becomes crucial for Internet Service Providers to perform network management, capacity planning, network resource allocation and network planning.

In this paper, we try to understand the behavior of mobile users by investigating three important features – Data Usage, Mobility Pattern and Application Usage by using the data traffic collected at a core metropolitan 2G and 3G networks of China. These features are separately related to mobile data resource (e.g. Data Usage), radio data resource (e.g. Mobility Pattern), and resource allocation (e.g. Application Usage).

Data usage is a fundamental characteristic and is usually related to the specific users, mobile devices, applications and the type of access networks (2G or 3G). In 2013, the global Internet had 18,000 PB mobile data traffic, and according to an authoritative prediction, mobile data traffic will grow at a CAGR (Compound Annual Growth Rate) of 61 percent from 2013 to 2018. This brings a great challenge for operators and service providers to plan future network deployments [1], and they urgently need to know the user behavior on data usage.

Users have different mobility patterns. The mobility pattern of a specific user provides important information on the consumption of network resources of the user. But how to capture the mobility pattern remains open. In this paper, we use cell towers to define the user location, which is able to be collected from the data trace. The information of the cell towers users access can provide rough location data of the users, which, however, is sufficient for capturing people's daily movement patterns in a large metropolitan area. We suggest that if a user accesses a lot of cell towers during a specific time frame, it implies that user's moving range is large during the time frame and a great number of radio resources are consumed.

Application usage is another important feature on the user behavior of network resource consumption.

Unfortunately, no canonical method has been established for this kind of analysis. In this study, we collected one week long continuous HTTP traffic data from December 28, 2013 to January 3, 2014. The data provides more than 1 million accesses to ''applications'' in each day. With the massive data we find the relations between application usage and network resource consumption. The data was constructed by flows, which contain the following information: anonymized user identifier (phone number), the timestamp of flow begin and end, the total number of packets and bytes for this flow, URI (Uniform Resource Identifier), hostname and server IP. In practice a user might switch from cellular data service to Wi-Fi for Internet access, leading to that no unique user identifier (user phone number) can be obtained from the packets. In such a case we target the user behavior analysis based on their cellular data usage.

The goal of this paper is to provide an understanding and categorization of Mobile Internet traffic, in particular focusing on the user behavior in terms of data usage, mobility pattern and application usage. We try to provide answers to questions like ''Are there any significant differences in traffic patterns across different applications?'', ''What are the unique characteristics of traffic heavy user and high mobility user compared to normal user?'', ''Are the data usage pattern and mobility pattern different across applications?'' and ''What is the relationship between person's data usage pattern, mobility pattern and application usage?''

We classify the users into different groups based on their traffic volume, mobility patterns, and preferred application types. Essentially, we define two user groups (heavy traffic and normal traffic) by the amount of traffic they generated, and four user groups (high mobility, normal mobility, low mobility and non mobility) by the user mobility. With these group classifications, we study the user behavior inside and between the groups.

B. PAPER ORGANIZATION

The remaining of the paper is organized as follows. In section 2, past related works are introduced. Section 3 provides the preliminaries of this paper, including analysis method, data collection, an overview of the collected traffic data and the classification method of applications. Section 4 gives a detailed analysis of data usage, mobility pattern and application usage from the mobile user behavior point of view. Section 5 draws relations between data usage, mobility pattern and application usage. Finally, conclusions are presented in Section 6.

II. RELATED WORK

In this section, we provide an overview of the prior research relevant to traffic analysis in cellular data networks. Especially, we focus on user behavior in Mobile Internet.

Traffic analysis is a critical step to model network traffic. Traffic characteristics change with the increasing network usage demands from individual users as well as business communities. With the development of Internet, the flow based traffic analysis has always been an ''academic hotspot'' [2]–[7]. However, a large amount of work on traffic characteristics accomplished a decade ago might not be suitable for current networks. Early in the 19th century, the necessity of cellular network traffic analysis already grows dramatically [2], [8].

The authors in [9] studied the traffic composition, the transfer sizes, the performance of TCP transfers and the interaction with radio power management of smartphone traffic, providing valuable information for Internet service provider. Recently, more and more researchers pay attention to understanding the user behavior of smartphone users, and the following topics have been focused on [10]–[12], including mobile audience measurements, mobile-phone-based content, automatically uncover and quantify characteristic behavior patterns in users' daily lives.

A. TRAFFIC DATA USAGE PATTERN

[13] investigated the usage patterns of mobile data users by analyzing the characteristics of traffic heavy users and normal users. Research results suggestted that a small number of traffic heavy users contributed the majority of traffic in cellular network.

B. APPLICATIONS USAGE

In [14] and [15], the authors tried to find out the reason users chose and adopted an application in their daily lives. Xu Qiang et al. [16] investigated the usage patterns of smartphone apps in terms of physical location, time, user and device. Chad C. Tossell et al. [17] provided an empirical characterization of web use on smartphones, and studied how native applications and a browser were used on smartphones. The performance of smartphone applications was investigated in [18]. It quantified how application performance, in particular web browsing, was impacted by various factors.

C. SEARCH BEHAVIOR

Church Karen et al. [19] studied the Mobile Internet habits (mobile search especially) of more than 600,000 European Mobile Internet users. In [20], the authors investigated mobile Web access patterns. It focused on how, why, where and in what situations people used the Mobile Internet and mobile search.

D. MOBILITY PATTERN

The authors in [21] investigated the mobility patterns in mobile cellular networks. It found that both inter-arrival time and dwell time distributions could be well approximated by power-law distribution, no matter in daytime, night, rural and urban areas. The authors in [22] examined data service usage and mobility patterns from various perspectives including application breakdown, user roles, device types and diurnal characteristics. Paul Utpal et al. [23] analyzed the network resource usage and subscriber behavior in a large scale 3G data network. Traffic load, mobility and resource efficiency were used as traffic characteristics.

E. BROWSING BEHAVIOR PATTERN

Shafiq Muhammad Zubair et al. [24] characterized the geospatial dynamics of application usage in a 3G cellular data network. The authors in [25] proposed and developed a scalable co-clustering methodology, Phantom, to group both users and browsing profiles simultaneously in 3G networks. They found that there existed distinct "behavior patterns" among mobile users, and the behavior of most users could be classified as either homogeneous or heterogeneous.

Although these works provided some deep insights on certain aspects of mobile user behavior, further study on the user behavior defined by data usage, mobility pattern and application usage is expected.

III. PRELIMINARIES

In this section, we introduce the analysis method of this paper followed by a brief overview of 2G/3G cellular data network architecture and how the data is collected from network. The data set used in our study is also illustrated and the flow metrics extracted from the collected traffic data are shown for the purpose of providing an overview of the traffic characteristics.

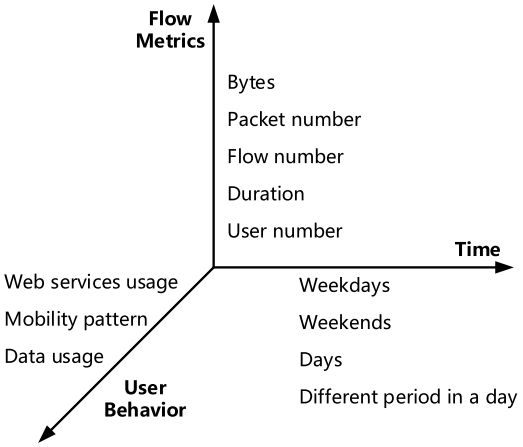


FIGURE 1. The analysis dimensions for mobile Internet.

A. ANALYSIS DIMENSION

We analyze the traffic data from three perspectives: flow metrics, time and user behavior, as shown in Fig. 1. The considered flow metrics include: bytes, packet number, flow number, duration and user number. Noting that the traffic characteristics always follows clear daily pattern [7], we focus on analyzing data ranging from a time slot in a day, weekdays and weekends. For user behavior analysis, we consider the visited applications, the mobility pattern and the data usage (traffic volume).

B. DATA COLLECTION

The collected traffic data come from a large Chinese 2G and 3G service provider. The high level view of a mobile network is shown as Fig. 2. There are three major components in one mobile network, including mobile devices, radio access network and core network.

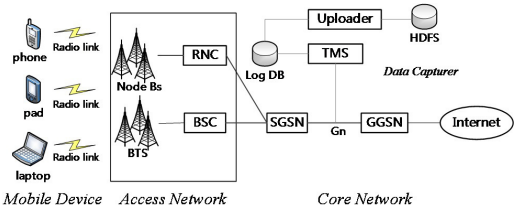


FIGURE 2. Mobile network architecture.

(1) Mobile device is the terminal connecting to the mobile network.
(2) 2G subscribers access the network via BTS (Base transceiver station) to BSC (Base Station Controller). In the case of 3G subscribers, request data is collected by Node Bs, and send to RNC (Radio Network Controller).
(3) Core Network is composed of SGSN (Serving GPRS Support Node) and GGSN (Gateway GPRS Support Node). Gn interface is between SGSN and GGSN. GGSN send the data to Internet through the Gi interface.

The data sets used in this study are collected by our Traffic Monitoring System (TMS) (this device has been placed in the production networks by several ISPs for traffic monitoring purposes), which is connected to the Gn interface. Mirrored packets with HTTP head have been collected from a large Chinese ISP that owns a large metropolitan area network in Southern China. This ISP has 4.5 million mobile subscribers in that area.

We group the packets into different flows by their 5-tuples {IP source address, IP destination address, source port number, destination port number, transport protocol}, i.e., a 5-tuple flow is a sequence of packets that share the same 5-tuple during a certain period (e.g. 64s). For the security reason, user privacy information in packets are replaced by a hashed number, which could be used for identifying subscribers, without affecting the usefulness of our analysis.

With the popularity of smart phones and the development of mobile applications (social network, e-commerce, video streaming and etc.), mobile traffic is undergoing a significant change over the last few years and the compositions of the mobile traffic have changed. However, web applications are still the dominant service in Mobile Internet [26], and mobile video tend to generate more and more traffic data in the next few years [1]. The analysis of HTTP-head traffic give us the opportunity to deeply understand all kinds of applications, include web browsing, web video, web music and so on.

C. TRAFFIC CHARACTERISTICS OVERVIEW

As mentioned above, the metrics we used to analyze the flow-related characteristics are bytes, packet number, flow number, flow duration and user number. We calculated each data point with ten minutes of time granularity.

Fig. 3 illustrates five time-series graphs of the flow records collected during the week from December 28, 2013 to January 1, 2014. We can see that each flow metrics show

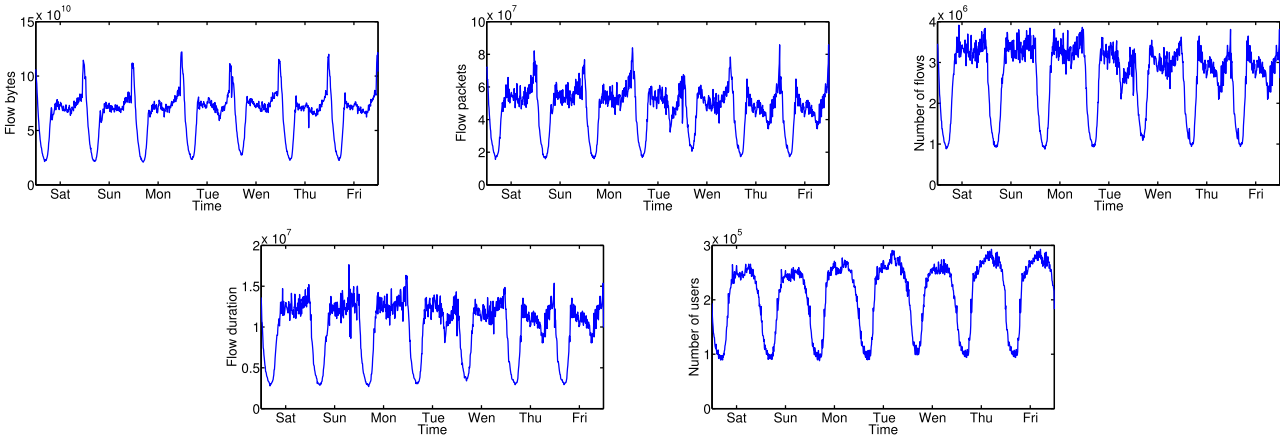


FIGURE 3. Time-series graphs of five flow metrics (bytes, packet number, flow number, flow duration and user number) from December 28, 2013 to January 3, 2014.

a clearly daily pattern. The overall summary of flow metrics is illustrated in Table 1. There were 4.5 million people accessed the mobile Internet during the time frame. They generated 60.81 TB HTTP traffic in total, 4.73×10^{10} HTTP packets and 2.68×10^{9} HTTP flows. Each user generated 3.15 MB traffic and 133 flows in average per day.

TABLE 1. Traffic characteristics overview.

Date	Traffic (TB)	Packet number ($\times 10^9$)	User number ($\times 10^6$)	Flow number ($\times 10^8$)	Duration for each user (second)
12/28	8.52	6.98	2.84	3.98	515
12/29	8.48	6.95	2.80	4.00	517
12/30	8.62	6.98	2.86	4.01	524
12/31	8.69	6.57	2.92	3.70	488
01/01	9.20	7.03	2.93	3.90	508
01/02	8.59	6.37	2.95	3.62	457
01/03	8.71	6.37	2.85	3.62	473
7 days	60.81	47.26	4.51	26.82	498

Notice that 1 January 2014 is the new year day and a national holiday. People generated more traffic during the daytime. However, around 10% less traffic is generated at the night than usual (from 11:30 p.m. December 31, 2013 to 0:30 a.m. January 1, 2014), which can be explained by the fact that people were having the celebration during the time frame.

IV. USER BEHAVIOR

In this section, we classify the mobile users based on their behavior of data usage, mobility pattern and application usage, and focus on the user groups that have significant impact on network resource (data resource, radio resource and resource allocation).

A. DATA USAGE

The intensity of a user's data usage could be indicated by the bytes in the flow record. Fig. 4 shows the CDF of bytes usage for each user in the week.

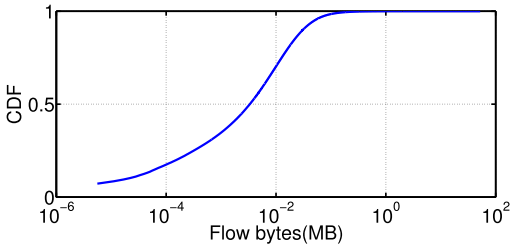


FIGURE 4. CDF of bytes usage of each user in the week.

In our dataset, we find that the total bytes distribution is highly uneven. 80% of users only contribute 0.21% of total HTTP traffic and 70% of users generate less than 10 MB HTTP traffic in the week. We define the heavy users as the mobile users contributing the top 1% of user traffic. Hence, among the total 4.50×10^6 users, 4.50×10^4 heavy users can be identified in the week. In statistics, these heavy users contribute to a significant portion of the total mobile data traffic: 9.98 TB out of 11.39 TB.

In addition to bytes usage, it is also informative whether the mobile user uses their data service consistently or only occasionally and whether the users with different data usage pattern have different temporal pattern. Hence, we study the temporal pattern of data usage in the time dimension.

Fig. 5(a) shows the statistics on the number of days users are active in the week, and Fig. 5(b) draws the number of hours users use their smart phones in a day. We also study the temporal activity (along the days in the week or along the hours in a day) of users' bytes usage. 32.36% of users generate traffic every day and 23.02% of users are active only for one day in the week. It is interesting to note that, in the week, about 43.35% of users use the data service for six or seven days, and 34.40% of users only use it for one or two days. It suggests that visiting the applications is a daily essential thing for a large fraction of users while a less fraction of users rarely use the applications. On the hourly activity, nearly 50% of users generated traffic in less than three hours.

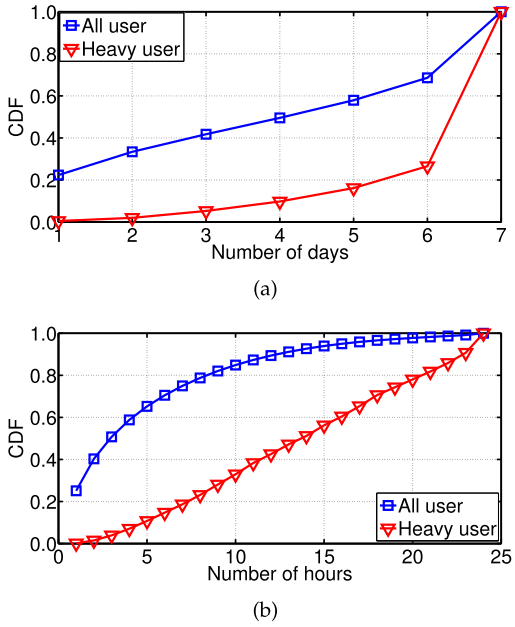


FIGURE 5. (a) CDF of the number of days in the week that mobile user generate traffic (b) CDF of the number of hours in a day that mobile user generate traffic.

In average, application users generate traffic for 5 hours in a day, and for 4 days in the week.

We further compare the difference between heavy users and normal users in terms of their daily and hourly activity. We can clearly see from Fig. 5(a) that most heavy users (83.84%) activate the data service more than six days in a week. As shown in Fig. 5(b), heavy users generate traffic in 14 hours averagely in a day, 70% and 31% of heavy users use applications for more than 10 and 20 hours respectively.

We can conclude from above analysis that different mobile users tend to have distinct data usage patterns. Top 1% heavy users contribute up to 88% of the total mobile data traffic, and use applications much more frequently.

B. MOBILITY PATTERN

Human mobility pattern is essential for a deep understanding of network dynamics and evolution. User mobility pattern in Mobile Internet also impact the network resource allocation and social network [27]. Here we investigate the users' mobility patterns when they are active in data service usage.

Due to the cell tower oscillation (a mobile device located at the boundary of two cells may access alternately one cell and then the other, even if the user location is not changed), we study the number of distinct cells that a user visit instead of the number of cells a user actually cross during a certain period of time, which represent the moving range of the user when he/she is connecting into Mobile Internet and how frequently the user moves across.

There are 1.81×10^5 cells in the area. In this paper, we define four groups of users according to their mobility to represent the user moving activities in the week. Define U_c the number of a cell user access in a week. Accordingly, we define user groups in terms of U_c:

Non mobility users, if

$$U_c = 1$$

Low mobility users, if

$$1 < U_c \leq 10$$

Normal mobility users, if

$$10 < U_c \leq 50$$

High mobility users, if

$$U_c > 50.$$

Fig. 6(a) shows CDF of the number of distinct cells users visited during a day. It suggests that CDF of the number of distinct cells for each day is nearly the same. On December 28, 2013, around 35% of the users visited only one cell (non mobility users) and 90% of users visit less than 10 cells in a day. In addition, only 1% of users visited more than 24 cells and users with the highest mobility visited 249 cells. Fig. 6(b) is the CDF of number of distinct cells users visited in a week. The proportion of non mobility users is about 12%, and 50% of users access more than 10 cells. The most active users visited 917 cells in the week. The proportion of high mobility users is about 10% in the whole week. Basically, the more people across different cells, the more hand off activities he will generate, and therefore the more radio resource he will consume.

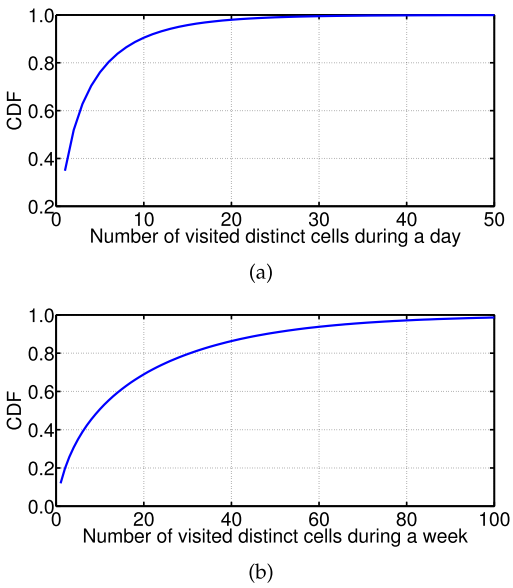


FIGURE 6. CDF of the number of distinct cells that user visit during a day (a)/a week (b).

Next, we investigate the traffic characteristics of these mobility user groups. Note that non mobility does not mean that the users don't move for the whole week. Instead, it just indicates that this kind of users likes to use data service of

cellular network in one place (cell) or they rarely use the data service. Therefore, we study the number of flows and traffic size the non mobility users generated in the week. We find that, with the increase of the number of flows the non mobility users generated, the traffic size for each user increases. Most non mobility users consume less than 2 flows and 3KB traffic in the week. However, there still exist very few non mobility users who generated more than 100 MB traffic in the week.

To evaluate the user movements in different time frames, Fig. 7 presents the average number of cells accessed by users in each group per hour of the week.

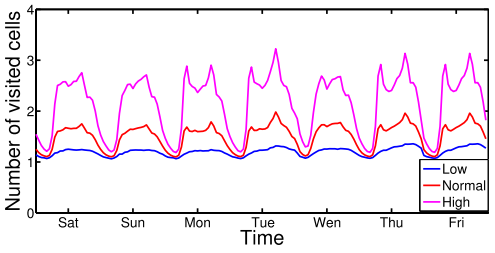


FIGURE 7. The number of different cells that user (three groups with different mobility pattern) moves across in a week.

An obvious daily pattern can be seen from Fig. 7 that highly mobile activities occur between 8 a.m. and 8 p.m. Moreover, high mobility users are much more active than the users in other groups in terms of their movement.

We can conclude from above that the moving ranges of the most mobile users are very limited: 90% and 50% of users visit less than 10 distinct cells in a day or in the week respectively. Mobile activities vary with time and the mobility characteristic of high mobility users is more vivid than other groups.

C. APPLICATION USAGE

The applications that user like to use reveal user's interest. For a better understanding of user interests, we build an application categoriy dictionary to classify the visited applications by examining the keyword of the URI field in flow records, e.g. if there is a "Facebook" in the URI field of a flow record, then this flow will be classified as "social network". We categorize the applications to 10 groups, each category represents one type of applications.

Note that flows generated by the application categories in Table 2 don't cover all the HTTP flows collected in our dataset. There are some flows not belonging to any of these application categories, and we exclude them from our study due to their minimum impact on the result presented in this paper.

The applications each user used are the direct indicator of user's interests in Mobile Internet. Different category of visited applications shows distinct preference of mobile users. In this part, we present the application usage pattern in Mobile Internet. We group users into different "interest cluster" according to their dominant application usage pattern and study the difference between these clusters. In addition,

TABLE 2. Application categories.

Category	User Interest
E-commerce	Online shopping, online payment
Reading	Online reading, EBook
Video	Online video
Music	Online music
Online gaming	Cyber games, network game
Social network	Social network, sharing site, blog, Micro-blog
News	Current news
Mail	E-mail
App store	Downloading the App or Android Applications
Search	Search engine, Google
Advertising	Advertisement

we further discover the diversity of application usage (user visit many applications or stick to very few application) for each application category.

Fig. 8 shows the CDF of visited application categories for users. In average, the number of applications one user visited in the week is 7 and each user visited 5 different application categories per day.

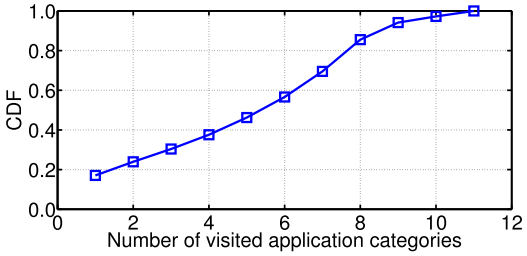


FIGURE 8. CDF of the number of visited application categories for each user.

Next, we find out which application attracts users to spend more time or generate more traffic using their mobile devices.

Table 3 shows the percentage of traffic, packets number, user number and flow number for each application categories. Clearly, we can see that social network is the dominant application in Mobile Internet: the top application categories are social network, search and e-commerce which contribute 55.65%, 14.27% and 7.53% of all application traffic respectively.

Among 11 application categories listed in Table 3, social network, e-commerce, advertising and search together contribute more than 80% of total HTTP traffic. The average flow duration of email in the week is 508 second, which is shorter than any type of applications. 10% of mail flow last more than 5.4 second, and 80% of mail flow last less than 2.4 second. E-commerce has the longest average duration (2453 second) for each user, implying that users like to spend more time on e-commerce.

For the purpose of understanding the browsing interests among users in Mobile Internet, we study the similarities

TABLE 3. The percentage of metrics for each application category.

	Traffic		Packets number		User number	Flow number	Duration
	Uplink	Downlink	Uplink	Downlink			
E-commerce	10.12%	7.11%	8.99%	8.22%	10.58%	11.16%	14.40%
Reading	1.35%	0.98%	1.04%	1.01%	1.18%	1.46%	1.34%
Video	2.07%	1.71%	2.22%	2.13%	2.69%	5.12%	2.18%
Music	4.76%	5.87%	5.51%	5.69%	4.85%	6.46%	5.58%
Online gaming	0.86%	0.62%	0.87%	0.77%	1.09%	2.09%	0.98%
Social network	41.85%	57.93%	50.44%	52.40%	45.82%	23.63%	42.88%
News	1.35%	0.95%	0.98%	1.00%	1.10%	2.77%	0.89%
Mail	0.37%	0.11%	0.22%	0.22%	0.40%	0.79%	0.21%
App store	5.05%	6.23%	5.62%	5.79%	5.20%	14.68%	6.18%
Search	17.56%	13.73%	14.48%	14.09%	13.12%	14.07%	12.04%
Advertising	14.65%	4.76%	9.63%	8.69%	13.96%	17.77%	13.30%

and differences between web users. To this end, we identify the "interest cluster" (a group of user who share the similar web browsing interest) by using the clustering method to classify users with similar behavior.

1) INTEREST CLUSTER

We see that 50% of users visit more than 5 different application categories per day, which suggests that most users have diverse interests. Yet users spend different amount of time in different applications, and generate different numbers of flows. In other words, users usually have different preference in terms of the applications they use. Identifying the application that users have most been interested in could help us predict user behavior. In this section, we classify users into different groups in terms of application preference (e.g. interest clusters) and study user behavior in these interest clusters.

Here we co-cluster the users and application categories using divisive hierarchical clustering [25] to investigate whether there exist distinct application usage patterns among mobile users. Divisive hierarchical clustering is an improvement of Spectral Graph-k-Part [28].

We group users using divisive hierarchical clustering in every hour to investigate the applications usage patterns with fine time granularity. The questions that we want to address are the following: (a) If user tends to stick to one category of applications all the time? (b) Does the browsing behavior change with time?

We use "entropy" to describe the diversity of user browsing behavior, which is defined as below:

$$H(X) = \sum_{i=1}^{n} p(x_i) I(x_i) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \qquad (1)$$

where n is the number of different application categories, each different i represents an application category, and $b = \mathrm{e}$ (constant value). $p(x_i)$ is the access probability for each application category.

$$p(x_i) = \frac{\textit{flow number for a application category}}{\textit{flow number for all applications}} \qquad (2)$$

Here, we use the normalized entropy as below to compare the different visited pattern between different application categories.

$$H_{norm}(X) = \frac{H(X)}{H_{max}(X)} \qquad (3)$$

$$H_{max}(X) = \log_b n \qquad (4)$$

After grouping users using divisive hierarchical clustering for each hour, users with the same application preference will be classified into the same "interest cluster" (there are 11 interest clusters in each hour, a user can only be classify into one interest cluster). Then the normalized entropy values are calculated for each cluster. The bigger normalized entropy value is, the more application categories the user visits in one hour, e.g. the user tends to visit more categories of applications in one hour. In contrary, a small normalized entropy value implies that user's interest is limited. Namely, users in an interest cluster which has small normalized entropy value tend to concentrate on very few application categories. Fig. 9 shows the normalized entropy of each interest cluster for every hour in the week.

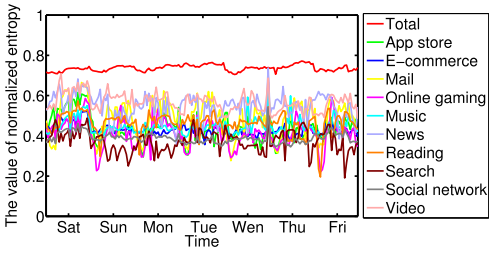


FIGURE 9. The normalized entropy value (the diversity of user's interest) of users for every hour.

The red line in Fig. 9 is the time curve of normalized entropy values for all users, and other lines are the normalized

entropy values of every interest clusters for every hour. The average value of normalized entropy of all users is 0.74, which is bigger than any interest cluster. This implies that users in the same interest cluster tend to concentrate on limit application categories and our cluster method works. In addition, different interest clusters with distinct application category preference have very different time curve. Namely, application usage behavior between distinct interest clusters is quite different. In order to see the normalized entropy values for each interest cluster more clearly, we further draw three graphs to show the difference between different interest clusters in Fig. 10, Fig. 11 and Fig. 12.

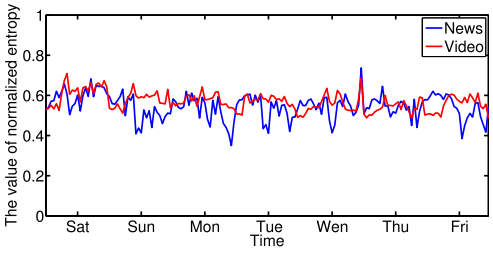


FIGURE 10. Users' normalized entropy values of interest clusters for news and video in every hour.

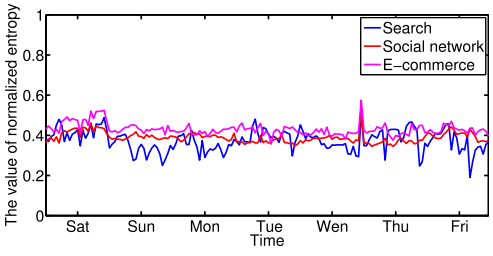


FIGURE 11. Users' normalized entropy values of interest clusters for search, social network and e-commerce in every hour.

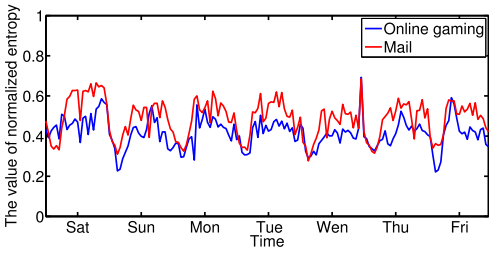


FIGURE 12. Users' normalized entropy values of interest clusters for online gaming and mail in every hour.

2) INTEREST CLUSTERS FOR NEWS & VIDEO

As shown in Fig. 10, the average values of normalized entropy of interest cluster for video (0.57) is the largest among all the application categories. It indicates that, even video is one's favorite application, he/she doesn't watch video using the mobile devices for a while. This observation can be explained by the expensive data service and poor watching quality, since that the traffic volume generated by video applications is usually very large, people might not prefer to watch mobile video online unless the bandwidth and QoS of network are improved. Similarly, the fact that interest cluster for news has the second largest average normalized entropy value (0.54), implies that users don't focus on news for a long time, and people usually browse news intermittently.

3) INTEREST CLUSTERS FOR SEARCH & SOCIAL & E-COMMERCE

The average value of normalized entropy of interest cluster for search and social network are 0.37 and 0.39 respectively, which are smaller than any other interest cluster, implying that if users prefer to use these two applications, they tend to stay a relatively long time with them.

Note that the normalized entropy values of interest clusters for social network and e-commerce do not vary much with time in Fig. 11. This means that although people in these interest clusters may access to social network or e-commerce at different time in a week, the attraction of these applications remains the same.

4) INTEREST CLUSTERS FOR ONLINE GAMING & MAIL

The periodic pattern is clear for online gaming and mail in Fig. 12. The minimum value appears between 2 a.m. and 6 a.m. every morning. It means that for those who play mobile online gaming or send email in the early morning, they tend to concentrate on online gaming or mail. The normalized entropy value for online gaming is smaller than mail, which means that the visiting behavior of mobile online gaming players is more concentrated than mail. However, the normalized entropy value for online gaming and mail become much larger during the day time, which indicates that, people get more distraction in this time frame than during the early morning. In addition, the visiting behavior of mail is more decentralized considering the larger normalized entropy during the day time.

5) OTHERS

A peak appears for all the interest clusters in the last hour of a year (during 11 p.m. December 31, 2013 and 0 a.m. January 1, 2014). This comes from the fact that when people celebrate the New Year eve, the visiting behavior of applications is much decentralized, and no obvious preference application category is shown.

6) APPLICATION USAGE IN EACH CATEGORY

Another way to explore user browsing interest is by looking at the visited applications for each category. Here, we further examine the diversity of application usage by calculating the normalized entropy values for each application category per hour. We answer the following questions, (1) Does there exist the popular applications in each application category which attract majority users? (2) Does the browsing behavior for each application in the same application category change with time?

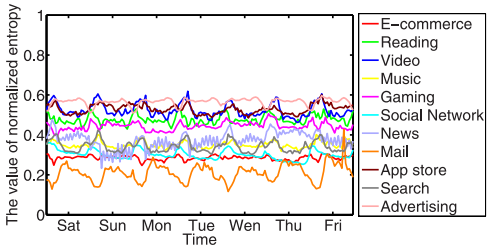


FIGURE 13. The normalized entropy values of each application category for each hour.

Here, we assume n as the number of different applications in a category, i represent the ith application, and $p(x_i)$ is the access probability for each application. Fig. 13 shows the normalized entropy value of each application category vary with time in one hour granularity.

In Fig. 13, a small normalized entropy value indicates the concentrated browsing behavior (user concentrate on one or a small number of popular applications in one application category), and a large normalized entropy value shows that users like to visit a variety of applications belonging to the same application category.

Also, in order to see the time curve more clearly, we further draw three graphs to show the difference between different application categories in Fig. 14(a), Fig. 14(b) and Fig. 14(c). As we can see from Fig. 14(a) that, users tend to stick to limited mail applications by using mobile phone (the average normalized entropy value is 0.21). The average normalized entropy value of app store, advertising and video (0.53, 0.57 and 0.53) are bigger, which shows there are many popular applications in these three application categories. In addition, clear daily pattern appears for all application categories except e-commerce, music and app store. The normalized entropy values of mail and advertising experience a bottom between 0 a.m. and 5 a.m. every day. However, a peak appears every morning for normalized entropy value of reading, online gaming, news, search, video and social network. By the entropy values at different time, we can see that the number of preferred mail and advertising applications vary greatly with different times in a day: users concentrate on limited types of mail and advertising applications in the morning but much more different applications are accessed during the other time of a day. In addition, it also shows that from 6 a.m. to 12 p.m., only few applications in each category of reading, online gaming, news, search, video, and social network are popular among users, e.g., weibo (the most popular Micro-blog in China). However, users show more diverse application interests in each category in the early morning. The reason behind this observation is that most people rest during the time frame (from 12 p.m. to 6 a.m.) and therefore the remaining users distribute more evenly on the different applications, which results in higher normalized entropy values.

V. RELATED CHARACTERISTICS FOR USER BEHAVIOR

To better understand the impact of user behavior on network resources, we further study the relations between the three features (Data Usage, Mobility Pattern and Application Usage) in this section. Our study will be performed within and between user groups, e.g. the groups with different data usage pattern and the groups with different mobility pattern. Heavy users generate the majority of the traffic in Mobile Internet, leading to a major consumption of the network bandwidth. High mobility users cause high frequency of hand off activities, which, in turn, consume much extra radio resources. We call these two user groups "Big Consumer of Resource". A deep understanding of behavior of the big consumers could help us plan network development and resource allocation better.

A. MOBILITY VS. DATA USAGE

It is interesting to find out how frequent users switch between different cells, and whether there is any difference between heavy and normal users in terms of their mobility.

We first draw a CDF of traffic generated by users in the week for each mobility group. We can see from Fig. 15 the higher users mobility is, the more traffic they generate. This means that people tend to use their mobile device when they travel or on commute.

More specifically, 90% of non mobility users generate less than 1 MB traffic in the week. 51.30%, 9.44% and 0.55% users in each group of low, normal and high mobility generate less than 1 MB traffic respectively. In average, one non mobility user generates 1.30 MB traffic and one high mobility user generates 28.11 MB traffic.

In addition, we further study the mobility of heavy user. Fig. 16 is CDF of distinct cells all users and heavy users visited in the week. 1.45% of the heavy users are non

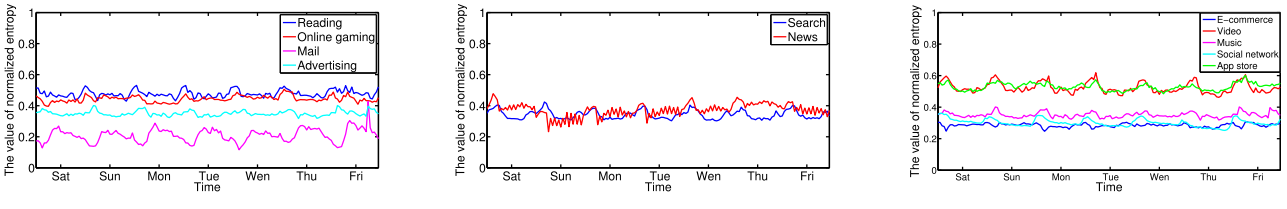


FIGURE 14. The normalized entropy values of some application categories for each hour. (a) Reading, online gaming, mail and advertising. (b) News and search. (c) E-commerce, video, music, social network and appstore.

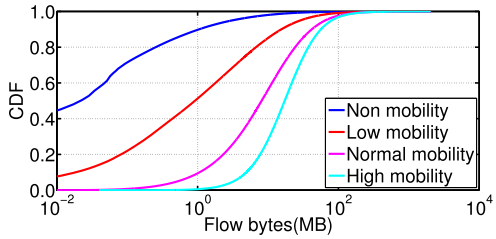


FIGURE 15. CDF of generated traffic for each mobility group (non mobility user, low mobility users, normal mobility users and high mobility users).

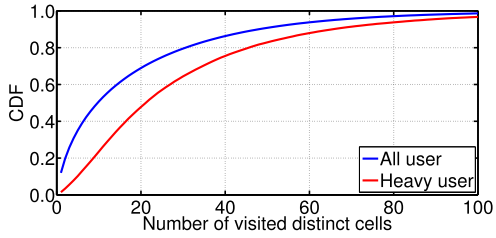


FIGURE 16. CDF of the number of visited distinct cells for all users and heavy users.

mobility users, and there are 22.04% low mobility users, 59.42% normal mobility users and 17.09% high mobility users. In the case of all users, the corresponding proportions are 11.97%, 38.77%, 40.06% and 9.58% respectively. We can conclude that heavy users tend to visit more distinct cells. In other words, heavy users in general also consume more radio resource in addition to the fact they generate more traffic than normal users.

B. APPLICATION USAGE VS. MOBILITY

In this section, we study the relationship between the user preferred application category and their mobility. First, the relationship between the number of visited application categories and user mobility pattern in the week is studied as shown in Fig. 17, which demonstrates the CDF of visited application categories from different mobility groups in the week. As we can see, generally speaking, the more the number of distinct cells a user visit, the more diverse applications user visited.

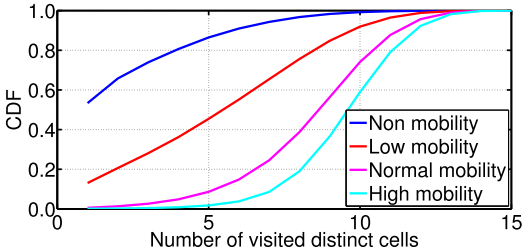


FIGURE 17. CDF of the number of visited distinct application categories for different mobility group.

In average, one non mobility user visits 3 application categories and one low mobility user visit 6 application categories in the week. For normal users and high mobility user, the average numbers of visited application categories are 9 and 10 respectively. The number of users who visit less than 5 application categories in the week accounts for 86.45%, 45.42%, 8.61% and 1.75% for non mobility, low mobility, normal mobility and high mobility users, respectively.

We now study the impact of the number of cells a user visit on the application categories the user accesses. To this end, we calculate the proportion of flow number of each application category for users with different mobility pattern.

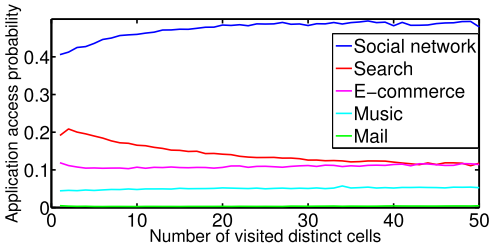


FIGURE 18. The number of distinct cell accessed by mobile user vs. the probability of mobile user that access different application categories.

Fig. 18 illustrates the probabilities that users with different mobility access the number of application categories. It shows the relationship between users' mobility and applications usage. We can see from Fig. 18 that among all the applications, social network is the leading application for users with different mobility span. And there is no clear correlation (and anti-correlation) between mobility and applications category that people access except social network and search engine. Indeed, for those who keep stationary, 40.58% of flows were generated by social network. Yet, for those who have a large mobility span, the percentage accesses for social network increase to 49%. This result is quite different from the result presented in [29], which concluded that email is the most popular application for those who have a large mobility span and social network shows highly intriguing behavior.

C. DATA USAGE VS. APPLICATION USAGE

We have presented several important observations regarding the relationship between application usage and flow metrics, data usage and mobility. Now we consider to correlate Data Usage with Web Usage to find out the preferences of application category for users with different data usage pattern. We focus on heavy users as they are the ones that generate the majority of traffic.

In Fig. 19, after a close examine of data usage pattern of heavy users, we find that heavy users tend to consume more data on app store, online gaming and music through mobile phone. Social network only takes 17.65% of total heavy user traffic, however, in the case of normal user traffic, the percentage is 55.65%.

Although heavy users only occupy 1% of total users, 3.22% mail users, 2.64% reading users and 3.22% news users were heavy users, and they contributed 21.91% mail traffic,

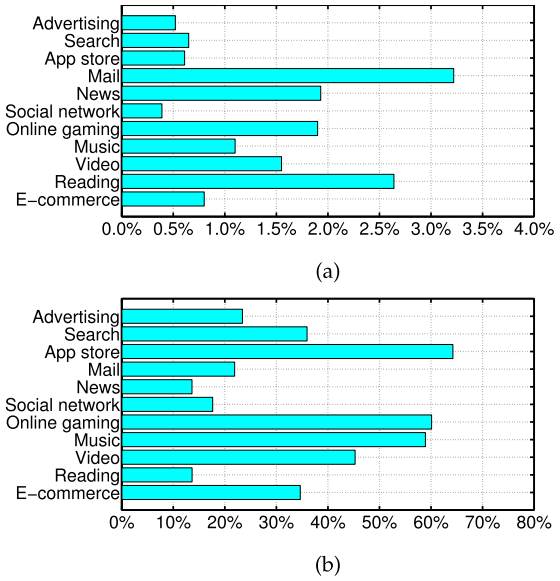


FIGURE 19. The percentage of heavy users for different application categories. (a) The percentage of the number of heavy users for different application categories. (b) The percentage of the traffic size generated by heavy users for different application categories.

13.65% reading traffic and 13.63% news traffic respectively. That means normal users don't check the mail, read books or news on the smartphone as much as the heavy users.

VI. CONCLUSIONS

In this paper, by using the real traffic data collected from mobile Internet in a large metropolitan area of China, we studied the mobile user behavior with detailed multi-dimension analysis by focusing on three features - data usage, mobility pattern and application usage.

A. FINDINGS

Firstly, we found that data usage pattern among mobile users is highly uneven. A few heavy users (top 1% in our study) contribute to 88.00% of all the mobile data traffic, and tend to use applications more frequently. In addition, about 43.35% of users use the applications for six or seven different days in a week, yet 34.40% of users only use the applications for one or two different days in a week.

Next, after grouping users with different mobility pattern, we conclude that moving range of most mobile users is very limited (90% and 50% of users visit less than 10 distinct cells in a day or in a week respectively). Mobile activities vary with time and the mobility characteristic of high mobility users is more vivid than other groups.

As for application usage, most users have a diversity of interests (50% of users visit more than 5 different application categories in a day), and users like to spend more time on e-commerce and less time on email. Besides, we identified the "interest cluster" to group the users with similar behavior. In one hour, users usually have a dominant application category to visit. If the dominant application category is online gaming or email, users tend to stick to these applications, but as for news or video, users don't like to stick to these application categories for a long time.

Finally, we further explored the relationship between Data Usage, Mobility Pattern and Application Usage. We found that big consumers of resource tend to largely consume data and radio resource at the same time. The more the number of distinct cells user visit, the more diverse the user browsing interest is. Lastly, heavy users tend to consume more data on app store, online gaming and music through mobile phone.

B. IMPLICATIONS

We found that the big consumers of resource are the main driving factor behind the variation of data usage and mobility pattern, and they tend to consume massive data and radio resources at the same time. In addition, both users' data usage and mobility pattern heavily impact their application access behavior. In order to improve the fairness and control the network congestion, network operators are advised to design appropriate mechanisms in resource provision and mobility management for the big consumers based on different categories of applications.

REFERENCES

[1] T. Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018," in *Proc. Cisco Public Inf.*, 2014.

[2] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. 18th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 1. Mar. 1999, pp. 126–134.

[3] G. Maier, A. Feldmann, V. Paxson, and M. Allman, "On dominant characteristics of residential broadband Internet traffic," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf.*, 2009, pp. 90–102.

[4] J. Yang, L. Yuan, C. Dong, G. Cheng, N. Ansari, and N. Kato, "On characterizing peer-to-peer streaming traffic," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 175–188, Sep. 2013.

[5] J. Yang, J. Ma, G. Cheng, Y. Wang, L. Yuan, and C. Dong, "An empirical investigation of filter attribute selection techniques for high-speed network traffic flow classification," *Wireless Pers. Commun.*, vol. 66, no. 3, pp. 541–558, 2012.

[6] J. Yang *et al.*, "Characterizing Internet backbone traffic based on deep packets inspection and deep flows inspection [j]," *China Commun.*, vol. 9, no. 5, pp. 42–54, 2012.

[7] M.-S. Kim, Y. J. Won, and J. W. Hong, "Characteristic analysis of Internet traffic from the perspective of flows," *Comput. Commun.*, vol. 29, no. 10, pp. 1639–1652, Jun. 2006.

[8] B. A. Mah, "An empirical model of HTTP network traffic," in *Proc. 16th Annu. Joint Conf. IEEE Comput. Commun. Soc. Driving Inf. Revolution (INFOCOM)*, vol. 2. Apr. 1997, pp. 592–600.

[9] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, "A first look at traffic on smartphones," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, 2010, pp. 281–287.

[10] H. Verkasalo, "Analysis of smartphone user behavior," in *Proc. 9th Int. Conf. Mobile Bus. 9th Global Mobility Roundtable (ICMB-GMR)*, Jun. 2010, pp. 258–263.

[11] A. Ghose and S. P. Han, "An empirical analysis of user content generation and usage behavior on the mobile Internet," *Manage. Sci.*, vol. 57, no. 9, pp. 1671–1691, 2011.

[12] J. Zheng and L. M. Ni, "An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 153–162.

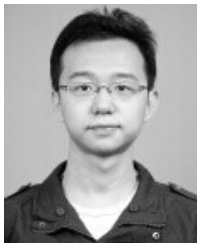
[13] Y. Jin *et al.*, "Characterizing data usage patterns in a large cellular network," in *Proc. ACM SIGCOMM Workshop Cellular Netw., Oper., Challenges, Future Design*, 2012, pp. 7–12.

[14] K. Church and R. de Oliveira, "What's up with Whatsapp?: Comparing mobile instant messaging behaviors with traditional SMS," in *Proc. 15th Int. Conf. Human-Comput. Interact. Mobile Devices Services*, 2013, pp. 352–361.
[15] H. Verkasalo, C. López-Nicolás, F. J. Molina-Castillo, and H. Bouwman, "Analysis of users and non-users of smartphone applications," *Telematics Inform.*, vol. 27, no. 3, pp. 242–255, Aug. 2010.
[16] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying diverse usage behaviors of smartphone apps," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 329–344.
[17] C. Tossell, P. Kortum, A. Rahmati, C. Shepard, and L. Zhong, "Characterizing web use on smartphones," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 2769–2778.
[18] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl, "Anatomizing application performance differences on smartphones," in *Proc. 8th Int. Conf. Mobile Syst., Appl., Services*, 2010, pp. 165–178.
[19] K. Church, B. Smyth, P. Cotter, and K. Bradley, "Mobile information access: A study of emerging search behavior on the mobile Internet," *ACM Trans. Web*, vol. 1, no. 1, p. 4, May 2007.
[20] K. Church and N. Oliver, "Understanding mobile web and mobile search use in today's dynamic mobile landscape," in *Proc. 13th Int. Conf. Human Comput. Interact. Mobile Devices Services*, 2011, pp. 67–76.
[21] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Human mobility patterns in cellular networks," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1877–1880, Oct. 2013.
[22] Z. Zhu, G. Cao, R. Keralapura, and A. Nucci, "Characterizing data services in a 3G network: Usage, mobility and access issues," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2011, pp. 1–6.
[23] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 882–890.
[24] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1341–1349.
[25] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3G network using hourglass co-clustering," in *Proc. 16th Annu. Int. Conf. Mobile Comput. Netw.*, 2010, pp. 341–352.
[26] Q. Yuan-Yuan, Y. Jie, and L. Zhen-Ming, "Structural analysis of complex networks from the mobile Internet," in *Proc. Nat. Doctoral Acad. Forum Inf. Commun. Technol.*, Aug. 2013, pp. 1–7.
[27] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1100–1108.
[28] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 269–274.
[29] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: Connecting people, locations and interests in a mobile 3G network," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas.*, 2009, pp. 267–279.

JIE YANG received the B.E., M.E., and Ph.D. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 1993, 1999, and 2007, respectively, where she is currently an Associate Professor and the Deputy Dean with the School of Information and Communication Engineering. Her research interests include broadband IP network, network traffic monitoring, mobile Internet data analysis, and big data mining.

YUANYUAN QIAO received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014, and the B.E. degree from Xidian University, Xi'an, China, in 2009. Her research focuses on traffic measurement and classification, mobile Internet traffic analysis, cloud computing, and big data mining.

XINYU ZHANG is currently pursuing the degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, where he received the B.E. degree in communication engineering in 2013. She is involved in the research of broadband IP network, and traffic identification and classification.



HAIYANG HE is currently pursuing the degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, where he received the B.E. degree in communication engineering in 2013. He is involved in the research of broadband IP network, and traffic identification and classification.

FANG LIU received the Ph.D. degree from Nankai University, Tianjin, China in 1997. She is currently an Associate Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include broadband IP network, network traffic monitoring, machine learning, and data mining.

GANG CHENG is currently a Senior Software Development Engineer with Microsoft, Redmond, WA, USA. He received the B.E. and M.E. degrees in information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree in electrical engineering from the New Jersey Institute of Technology, Newark, NJ, USA, in 2005. His research focuses on QoS guarantee issues in high speed networks, Internet routing protocols and service architectures, in particular, QoS routing, queue management, and congestion control in high speed networks.