Noname manuscript No.
(will be inserted by the editor)

A Human-in-the-loop Architecture for Mobile Network: from the View of Large Scale Mobile Data Traffic

Yuanyuan Qiao, Jianyang Yu · Wenhui Lin · and Jie Yang

Received: date / Accepted: date

Abstract 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, which implies the necessity for moving from a system-centric design to a more user- or even human- and data- centric design paradigm "to keep the human in the loop" in future network. It drives us to design a system with capacity to allocate network resource dynamically according to feedback from users. This paper presents a Human-In-The-Loop architecture for mobile network that discovers users' needs on network resource by understanding data traffic usage behavior of users. Based on real data traffic of mobile network, we analyze data traffic patterns of heavy and normal users from the view of online browsing behavior and urban functional area to explain how and why the data traffic is consumed. Then we propose a Latent Dirichlet Allocation model based solution to correlate data traffic, user behavior, and urban ecology to gain deep insights into spatio-temporal dynamic of data traffic usage behavior for different groups of users. Drawing upon results from a comprehensive study of users in a metropolitan city in China, we achieve a broad understanding about the difference of data traffic usage patterns of heavy and normal user: (1) besides the amount of generated data traffic, two groups of users can be easily distinguished by usage behavior of limited number of applications at midnight, (2) the functions of locations have huge impact on data usage patterns of users, which implies that urban ecology will shape users' online behavior. The results of this work can potentially be exploited to

Yuanyuan Qiao (Corresponding author), Jianyang Yu and Jie Yang
Research Center of Network Monitoring and Analysis, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
E-mail: {yyqiao}@bupt.edu.cn

Wenhui Lin
Technology Research Institute, Aisino Corporation, Beijing, China

help to allocate network resource, improve Quality of Experience according to users' needs, and even design the future network.

Keywords Human-in-the-Loop · Data Traffic Pattern · Data Traffic Usage Behavior · Mobile Networks

1 Introduction

With the pervasiveness of mobile Internet, smart phones are becoming ubiquitous providing the continuous Internet access, which brings extensive interaction between users and mobile networks. In 2G/3G/4G networks, how to provide high Quality of Service (QoS) and Quality of Experience (QoE) in the case of limited network bandwidth and link resources has always been a research hotspot [1]. Traditionally, QoS and network quality are measured by requirements on all the aspects of a connection, such as service response time, loss, signal-to-noise ratio, crosstalk, echo, interrupts, frequency response, loudness levels, and so on. However, QoE is not taken into account in current mobile networks ecosystem since it is mainly affected by subjective factors that are difficult to measure, such as, cost, reliability, efficiency, privacy, security, interface user-friendliness and user confidence [2].

Unlike other radio signal services, 5G would have a huge task to offer services to heterogeneous networks, technologies, and devices operating in different geographic regions to fulfill the high expectation of people. Satisfying the customers' requirements implies the necessity for moving from system-centric to more user- or even data-centric designs [3]. One of the overall guidelines for designing such an architecture should be "*to keep the human in the loop*", which makes human-in-the-loop mobile networks a promising system by considering human factors. In this case, the system can actively learn, predict, adapt, and steer user behavior, so as to greatly improve system efficiency and to provide superior user's QoE.

For fully taking human factors into consideration in the next generation mobile networks, an exhaustive understanding of user behavior becomes crucial for Internet Service Providers (ISP). In mobile networks, mobile traffic information is paramount in drawing a clear picture of how the access network resources are consumed by mobile users to design and evaluation of solutions concerning not only technological aspects of cellular systems, but also spatio-temporal dynamics of the global user demand. According to an authoritative prediction, monthly global mobile data traffic reached 7.2 exabytes at the end of 2016, and will surpass half a zettabyte by 2021 [4]. The explosion in data traffic consuming brings many opportunities to obtain data source with rich information, as well as great challenge for making fully use of it. In mobile networks, passively collecting data traffic and extracting human behavior (movements and app usage behavior) while he/she is accessing to mobile Internet has lots of advantages: high cost efficiency, low energy consumption, covering a wide range and a large number of people, and with fine

time granularity (people tend to surf mobile Internet frequently while moving, and many apps may send or receive network traffic packets periodically when running in the background), which can offer a comparable perspective on human activities.

In order to understand the dynamics of the mobile traffic demand of large scale people, and how to evolve the mobile network infrastructure to better accommodate it, a human-in-the-loop architecture for mobile network is presented in this paper. Within the main components related to all aspects of the human-in-the-loop system, we focus on analyzing passively collected mobile data traffic to study the temporal and spatial features of user behavior at traffic and service level. Our goal is to discover the traffic consuming patterns of different group of users, and to provide useful advices to ISP, such as "*How to set the initial parameter when deploy a new base station (what's the traffic consuming pattern in this area)?*", "*How to allocate the network resources to ensure QoE (what is the pattern of dynamic change for traffic consuming/service usage/human behavior)?*", and "*How to further discover the value of data traffic from user behavior point of view?*". In paper [5, 6], it has already been observed that data usage across mobile users is highly uneven. It is well know that the amount of data traffic for each flow follows heavy tail distribution, i.e., a small number of heavy users (nearly 20% of total users) consume a majority of data traffic (more than 80% of data traffic) in the mobile network. Above observation drives us focusing on traffic heavy user based on the assumption that revealing special properties of these users may guide resource allocation in mobile networks.

In recent years, combining more advanced device capabilities with faster, higher bandwidth, and more intelligent networks leads to a wide adoption of data-rich multimedia applications resulting in a phenomenal increase in the mobile traffic. It is necessary to update analysis results with latest data, even many researches characterize the data usage patterns of traffic heavy users [5–7] before. Furthermore, we extend and apply previous work to a "*human in the loop*" architecture, which may be helpful to design future mobile network. Overall, the contributions of our work are summarized as follow:

1. In order to provide superior users QoE, we present a human-in-the-loop architecture for mobile networks, which fully considers human behavior by analyzing the collected data traffic. Furthermore, we examine the spatio-temporal characteristics of heavy and normal users with real data traffic of mobile network that covering 5.4 million users, and 2.4×10^4 base stations for a month. The analysis results verify the necessity to optimize and balance the resource usage of mobile network dynamically based on users activities.

2. Based on real data, we group users into heavy and normal users according to the amount of data traffic they generated. We compare the difference of data traffic usage patterns between two groups of users from the aspects of temporal changing, urban ecology, and user behavior in the city. By understanding the highly spatio-temporal and non-homogeneous nature of the data traffic, interesting results, which may guide us to improve QoE, optimize and balance the resource usage timely and effectively, show up in our experiments.

3. In our analysis, we try to find out the influence factors that shape the data traffic usage patterns of heavy and normal users. Therefore, a topic model was trained to discover the correlated venues, locations, and behaviors extracted from data traffic. Several interpretable topics in space and time were identified, which help us understand how the network resource is consumed from the view of human behavior and the environment they are in.

This paper is structured as follows. Related work is discussed in Section 2. In Section 3, the architecture of human-in-the-loop system for mobile networks is presented. In Section 4, we provide details about our real dataset, and present some basic observations of data traffic spatio-temporal distributions for heavy and normal users. Then, we conduct a deep analysis on the data traffic usage pattern from the view of online browsing behavior, and urban functional regions in Section 5, and 6 respectively. In Section 7, we reveal the correlations among data traffic, urban functional regions, and human behaviors for heavy and normal users. At last, we summarize our discoveries and discuss potential investigations in Section 8.

2 Related Work

Human-in-the-loop is defined as a model that requires human interaction, which allows the user to change the outcome of an event or process [8, 9]. Human-in-the-loop has been widely studied in the area of energy management [10], health care [11], automobile systems [12], and evaluation tools [13], which allowing the human in the loop to handle the more challenging tasks of supervision, exception control, optimization tasks and maintenance duties. However, designing and implementing a Human-in-the-loop system poses tremendous challenges and is extremely time-consuming. Especially, determining how to incorporate human behavior models into the formal methodology of feedback control is the most difficult part [14, 15].

Although human users interact extensively with mobile communication networks, in current generation systems of mobile networks, human factors have not yet been well understood and fully taken into consideration. From mobile operator viewpoint, the traffic of mobile networks is considered from the network perspective and aggregated over many users, which usually focus on temporal dynamics [16–21], spatio-temporal dynamics [16,22–25] of the global user demand at aggregate level. Instead, from a mobile user viewpoint, the studies on activity distributions [18, 20, 26], mobile user categories [19, 23, 27], traffic-mobility correlations [18, 26, 28, 29], device and traffic types [20, 30] at per-user level are popular. In temporal dimension, it has been found that data traffic with an aggregate matter [16–20, 31], user's mobility behavior [32–35], and user's app usage behavior [5, 36] all tends to follow regular daily, weekly, monthly patterns. When considering the spatial dimension, the widely observed patterns in mobile networks including: data traffic usage usually varies with different location [16–18, 37], the usage of popular applications of crowds and web browsing preference of individuals is strongly depends on location [24,

28,38,39]. Recently, Aveek K. Das et al. observed that user's location category (such as cafe, university campus, residence etc.) can be identified just by passively monitoring and learning from aggregate network traffic from different categories of location [40]. Wang et al. found that only five basic traffic patterns exist among the 9,600 cellular towers. Each of the extracted traffic pattern maps to one type of geographical locations related to urban ecology, including residential area, business district, transport, entertainment, and comprehensive area [41]. All these research results strongly suggest the designer of future mobile networks pay much more attention to human behavior.

With the explosion in data traffic amount [4], increasing number of studies emerge in the area of mobile traffic analysis. It has been found that many factors may impact people's app usage behavior on smartphones, such as the devices people use, users' personalities, surrounding environments, and nearby users [8, 10–13, 42]. Recently, it has been proven that location has a strong influence on what kinds of apps we choose to use [14, 15]. For the purpose of further understanding human behavior and improving mobile networks, by using real data traffic, several efforts have been made from the aspect of social analysis, mobility analysis, and network analysis [43]. For drawing a clear picture of how the access network resources are consumed by mobile users, then offering useful advises for greatly improving system efficiency and providing superior users QoE, human-in-the-loop architecture based new techniques, which capture spatio-temporal characteristics of mobile networks from data traffic, are expected in mobile networks [43, 44]. As a critical step to addressing this challenge, it is imperative to understand how mobile data users access cellular data services and the difference of their demands on network resource between different users. As motivated by such observation, this paper aims to design a human-in-the-loop architecture for mobile networks with real data traffic, which could discover the difference of data traffic usage patterns between heavy and normal users, and interactions between user behavior and mobile networks from traffic pattern, application usage behavior, and urban ecology perspective. More specifically, different from the previous work, we examine the relationships between above perspective, which help to understand how to get feedback from human behavior in mobile networks.

3 A Human-in-the-loop Architecture for Mobile Network

In this section, we present a Human-in-the-Loop Architecture for Mobile Network (HLA-MN), which takes feedbacks of users into account dynamically and makes adjust according to real time data traffic, as shown in Fig. 1. Mobile traffic data generated by User Equipment (UE), such as smart phone, tablet, laptop, computer equipped with mobile broadband adapter or any other devices that access to Internet through mobile networks, is connected with Internet by Base Station (BS) and collected by deploying our self-developed Traffic Monitoring Equipment (TME) [45] in the Core Network (CN) edge. In 2G and 3G networks, a Base Transceiver Station (BTS) or Node B, which transmits

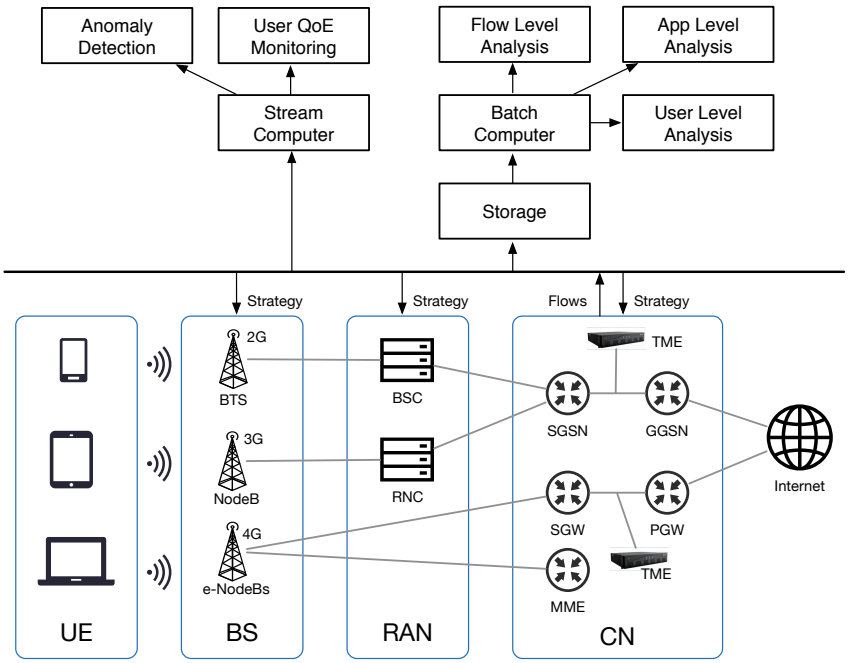


Fig. 1: The Human-in-the-loop architecture for mobile network

its network traffic to Base Station Controller (BSC) or Radio Network Controller (RNC). The controllers (BSC/RNC) then deliver the network traffic to a Serving GPRS Support Node (SGSN) that establishes a tunnel on Gn interface (Interface between the GGSN and the SGSN) with Gateway GPRS Support Node (GGSN) through which the data enters the Internet (GPRS represents "General Packet Radio Service"). In 4G network, evolved Node B (eNodeB) establishes the connection between UE and mobility management entity (MME). Users' data traffic goes into the Internet through the Serving GateWay (SGW) and Packet Data Network (PDN) GateWay (PGW).

The core idea of proposed HLA-MN is to understand data traffic usage pattern, then allocate network resource to provide superior users QoE. To achieve above goal, we deploy TME between SGSN and GGSN in 2G and 3G networks, which provides rich information on protocol-, service-, and user-level operations. Note that, the data captured by RNC probes, Mobile Switching Center (MSC) probes, and Charging Gateway Function (CGF) [43] also provide fine-grained (hard to collect) or coarse-grained (easy to collect) information. Here, in order to get insight on the type of data traffic generated by the devices, we focus on data collected in Gn interface. The proposed HLA-MN can be applied to current and future mobile network no matter what kinds of mobile traffic are collected.

Collected flows are processed in real time by stream computer or stored in distributed storage for offline batch processing. Real time processing is very useful to identify abnormal flows, which may be caused by big events or equipment failures. Also, monitoring the QoE of each user then react quickly is also vital for a human centric network. However, due to limited resource and huge amount of data, real time processing is only suitable for jobs with low computational and time complexity. Batch computer can deal with more complex analysis jobs with more data. In our proposed architecture, we focus on flow, application, and user level analysis, to discover the unique characteristics of mobile big data. For flow level analysis, we perform analytics regarding the traffic statistics in terms of the number of flow, flow duration, flow bytes, and number of users with spatio-temporal dynamics. Regular patterns, specific fingerprint, or abnormal traffic can be identified to guide further analysis. Combining more advanced device capabilities with faster, higher bandwidth, and more intelligent networks lead to a wide adoption of data-rich multimedia applications. Therefore, HLA-MN also investigates mobile applications running over the HyperText Transfer Protocol (HTTP) to provide us with relevant insights into the current networks from a data traffic consumer perspective. In addition, HLA-MN provides analytics on the users' behavior, including online browsing and offline mobility behavior. Thus, the users' needs on mobile networks could be more accurately understood, and even predicted. Based on above analysis results, strategies, such as "how to allocate the network resource?" or "how to provide super QoE to somebody", can be provided to real network.

Traditionally, characterizing the traffic dynamics in a mobile network from the flow and application level is paramount in understanding how the access network resources are consumed by mobile users and improving mobile network system. In the rest of this paper, in order to study the imbalance of the network resources from individuals' viewpoint, then give insights into a human centric network, we try to understand users' needs on data traffic at different time and places by discovering the spatio-temporal regularity of user's behavior.

4 Preliminaries

In this section, firstly, we illustrate the characteristics of dataset used in our experiment. Then, in order to focusing on users who generate a huge amount of traffic in mobile network, we define heavy user and normal user based on real mobile traffic data.

4.1 Characteristics Overview of Collected Mobile Traffic Data

Our analysis is based on aggregate 2G/3G mobile traffic data, supplied by an operator who provides services for more than 50% mobile Internet users in China. TMS mirrors all uplink and downlink packets, and then groups them

Table 1: Flow metrics of our dataset for 4 weeks

Duration	Number of users($\times 10^6$)	Number of flows($\times 10^9$)	Flow bytes(TB)	Online duration for each user(Min)
3/8/2015 9/8/2015	3.31	2.57	14.36	76.84
10/8/2015 16/8/2015	3.06	3.02	16.86	98.08
17/8/2015 23/8/2015	2.88	2.75	15.51	99.57
24/8/2015 30/8/2015	2.77	2.62	13.85	99.92
Total	5.42	10.96	60.58	206.19

into flows by 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). Usually, we can extract a lot of valuable information from records of traffic flows, such as, user ID, start time and end time of each flow, ID of base station user connected with, uplink and downlink bytes, number of uplink and downlink packets, and so forth. Especially, flows over the HTTP contain host and Uniform Resource Locator (URL) user visited, which give us the opportunity to understand users' browsing behavior by crawling and analyzing the webpages they visited. For the security reason, user privacy information in packets are replaced by a hashed number, which could be used for identifying users, without affecting the usefulness of our analysis.

In our experiment, we collected mobile traffic data between August 3rd (Monday) and August 30th (Sunday), 2015 at base station level from a northern city (covers 53068 square meters with more than 10 million population) in China. Collected dataset covers 5.4 million distinct UEs (in this paper, we assume that every UE is carried by a distinct user), and 2.4×10^4 base stations. The overall summary of flow metrics of our dataset is illustrated in Table 1. We collected flows of 5.4 million users, they generated 60.58 TB traffic and 10.96×10^9 flows in total. Each user was online for 3.43 hours average in 4 weeks (the duration between start and end time of flow is defined as online duration of user who generated this flow), generated 428.49 KB traffic and 72 flows in average per day.

In addition, in order to have an overview of daily usage for our dataset, we further draw time-series graph of flow metrics with 15 minutes time granularity in 4 weeks, as shown in Fig. 2. Daily pattern is very clear for all metrics, e.g., the amount of users who connect with mobile network begins to rise after 5 am, along with the amount of data traffic. The first peak for the number of generated flows/flow bytes appears around 10 am, followed with a drop during users' lunch time. As can be seen in Fig. 2, there is a peak around 8 pm in the evening for number of flows, flow bytes and online duration for each user, which implies that users tend to use mobile devices for much longer time and generate more data traffic before bed time than other time in a day. Since people's daily behavior tends to follow obvious pattern (e.g., get up at 7 am,

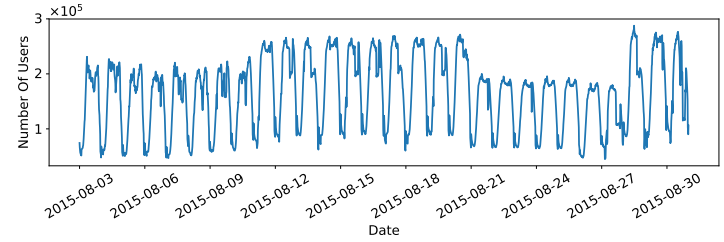
go to work place before 9 am, have lunch at 12 am, go home after 5 pm, go to bed after 10 pm, and so forth), the consuming of mobile network resource, i.e., mobile traffic data generated by users, also follows daily pattern. As a result, keeping human in the loop when managing the network is crucial to mobile networks for capacity planning and allocating appropriate network resources to cope with the growing mobile traffic data demands of users.

4.2 Heavy User and Normal User

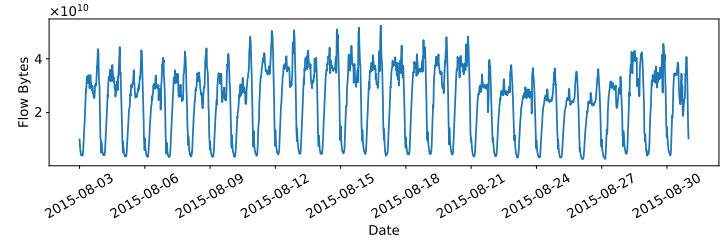
In order to understand data usage pattern of users in our dataset, then define heavy users accordingly, we draw the flow bytes distribution firstly, as shown in Fig. 3. In our dataset, we find that 85.72% of users only contribute 20% of total traffic, and 95.30%/74.90%/46.03% of users generate less than 100MB/10MB/1MB HTTP traffic in 4 weeks. Therefore, we define the heavy users as the mobile users contributing the 80% of user traffic, and others are normal users. Here, among the total 5.5×10^6 users, 7.7×10^5 (14.28%) heavy users can be identified in 4 weeks. In statistics, these traffic heavy users contribute to a significant portion of the total mobile data traffic in 4 weeks: 54.04 TB out of 67.56 TB. In the following sections, we analyze and compare the online browsing and offline mobility behavior of traffic heavy and normal users, which help us understand the difference between above two groups of users and the factors that heavy users emerge.

For heavy and normal users, we compare the value of flow metrics changing with time in 4 weeks, as shown in Fig. 4. As we can see from Fig. 4a, although traffic heavy user only occupies 14.28% of total users, it is very clearly that, in every day, for those who are online at least once, the number of traffic heavy users is larger than normal users. In every 15 minutes, there are 96917 traffic heavy users and 67804 traffic normal users online. More specifically, traffic heavy users generated 30 flows (201 KB) in 187 seconds, and traffic normal users generated 17 flows (70 KB) in 100 seconds averagely in every 15 minutes. If we focus on the ratio of flow metrics for two groups of users, it is very interesting that, for all flows metrics, the largest ratio value appears between 11 pm and 1 am, and the smallest ratio value shows up between 8 am and 10 am. Furthermore, for each individual, the greatest differences of generated flow bytes, number of generated flows, and online duration between traffic heavy user and normal user emerge between 11 pm and 1 am too, but minimal differences emerge around 4 am. It implies that the two groups of users may be distinguished by only paying attention to data usage patterns of users during midnight. In other words, we may quickly identify a traffic heavy or normal user by observing his/her data traffic usage behavior in the midnight. They tend to follow very similar data traffic usage pattern around 4 am, when most of people are sleeping.

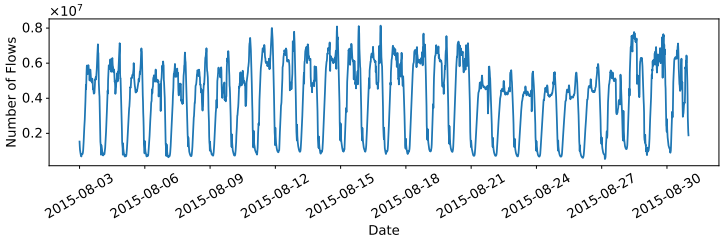
Since it is also informative whether the mobile user uses their data service consistently or only occasionally, we further examine (a) the number of days in 4 week that mobile users are active online and (b) the number of hours



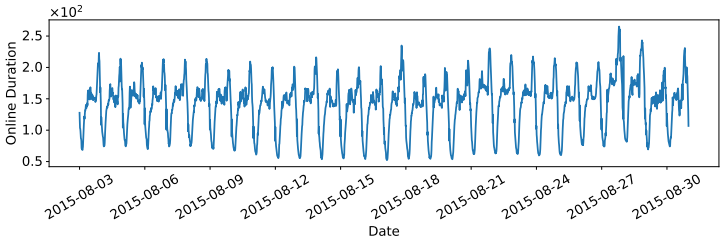
(a) Number of users



(b) Flow bytes



(c) The number of flows



(d) Online duration for each user

Fig. 2: The value of four typical flow metrics changing with time in 4 weeks

averagely in a day that mobile users use data service, as shown in Fig. 5. In 4 weeks, 6.18%/0.19% of traffic heavy/normal users generate traffic every day and 1.79%/44.59% of traffic heavy/normal users are active only for less than 3 days. Averagely, each traffic heavy user uses mobile data traffic for 17 days in 28 days, but other users only generate traffic in 5 days. It is interesting to note

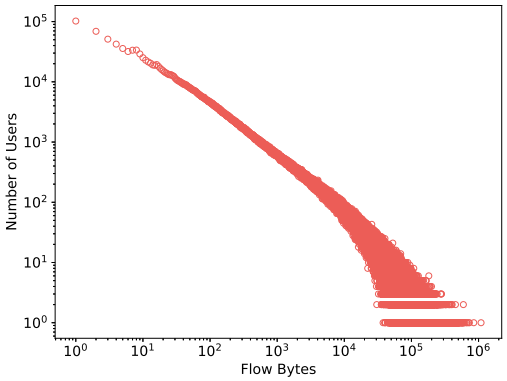
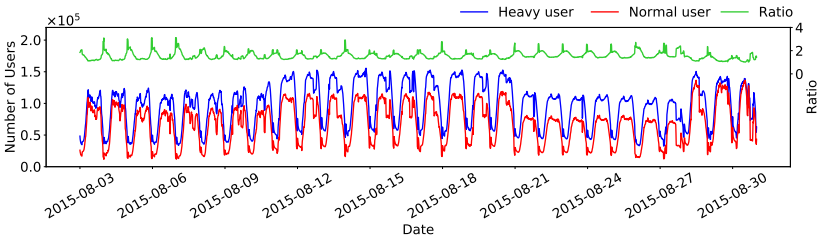


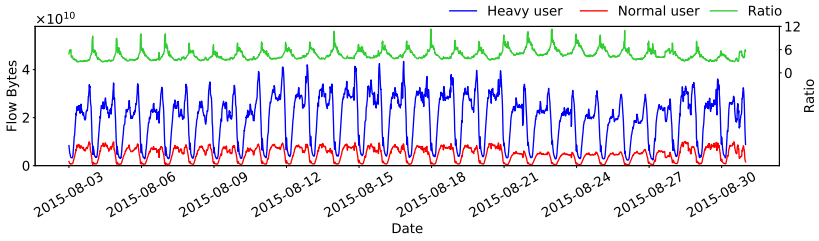
Fig. 3: The flow bytes distribution for 4 weeks

that, in 4 weeks, about 60.19% of traffic heavy users use the data service for more than 15 days, and 60.94% of traffic normal users only use it for less than 5 days. It suggests that connecting with mobile networks is a daily essential thing for traffic heavy users while most of traffic normal users rarely use the mobile data traffic. On the hourly activity, traffic heavy/normal users stay online for 4.5/0.705 hours per day averagely, 1.10%/0.015% of traffic heavy/normal users use mobile data traffic for more than 20 hours, and 41.57%/95.09% of traffic heavy/normal users use in less than 3 hours in a day. We can conclude from above analysis that two groups of users have very distinct data usage patterns. Traffic heavy users tend to use data traffic more frequently in a day or in a month, and other users only connect with mobile networks occasionally.

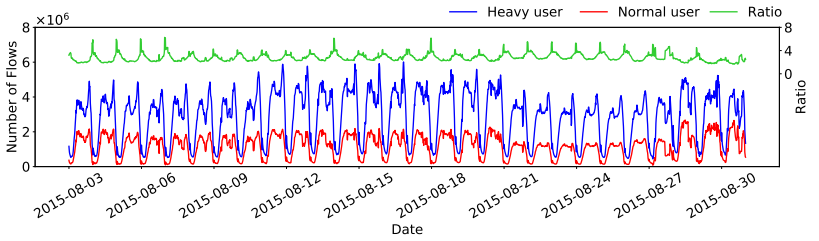
Then, we have two reasonable questions, "*What is the difference between traffic heavy and normal users when they use cellular data services?*" and "*Why traffic heavy and normal users have different demands on mobile network resource?*". The answers of above questions may help us design HILP-MN. Since human daily activities are influenced and restricted by time and space, at following sections, we aim to discover patterns of temporal and spatial changes of traffic heavy and normal users by studying the online browsing and offline mobility behavior of them. After understanding the usage pattern of cellular network data for traffic heavy users, ISP may allocate specific resource or limit resource consumption for a specific group of users, to ensure QoE for most of the users. Note that, the curves of flow metrics we draw do not always follow same trend in different days. For example, the number of distinct users online is a little larger during Aug. 11th 2015 and Aug. 20th 2015 than others days, and during Aug. 21st 2015 and Aug. 27th 2015, the value of flow metrics is smaller than other days. It is because some data is lost due to network problem. In this paper, since we tend to focus on the patterns and differences between traffic heavy and normal user, and do not present model or calculate specific parameters according to dataset, so, dataset that covers a part of mobile users



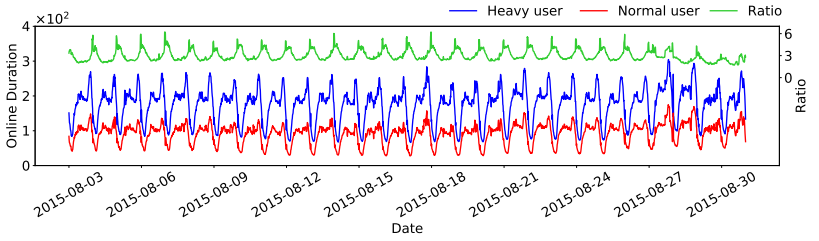
(a) Number of users



(b) Flow bytes



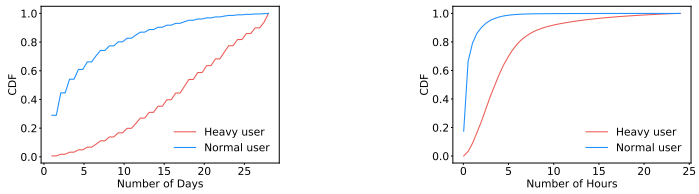
(c) The number of flows



(d) Online duration for each user

Fig. 4: The value of four typical flow metrics changing with time in 4 weeks for traffic heavy and normal users

is enough to carry out the experiment and analysis. Hence, in the following experiments, we use the mobile network data traffic in the first week in our dataset, i.e., from Aug. 3rd 2015 to Aug. 9th 2015, the daily pattern of which is relatively smooth in whole week, as experimental data.



(a) In 4 weeks (b) In a day

Fig. 5: The number of days/hours that traffic heavy and normal users are online. (a) CDF of the number of days in 4 weeks that mobile user generate traffic (b) CDF of the number of hours in a day that mobile user generate traffic averagely

5 Data Traffic Usage Patterns of Heavy and Normal User from the view of Online Browsing Behavior

In this section, we look into the online browsing behavior of heavy and normal traffic users. The analysis of HTTP-head traffic gives us the opportunity to deeply understand all kinds of applications, include web browsing, web video, web music and so on. Each HTTP flow contains the following details: a user's anonymized ID, start time and end time of each flow, the accessed URL, and the base station ID. Therefore, the applications users use can be distinguished by keywords in URLs, such as *twitter, mail.google, map.google*, and so forth. In this part, we only focus on 100 most popular applications that contribute more than 50% of total amount of HTTP flows. Since some applications may meet same needs of users, we further classify the URLs accessed by users into 23 applications interests via keyword mining over the URL. We focus on the following 12 categories which contribute more than 93% of total traffic among top 100 applications: ***instant messaging, search, news, maps, e-commerce, social network, data service, mobile assistant, game, video, cloud service,*** and ***music*** as shown in Table 2.

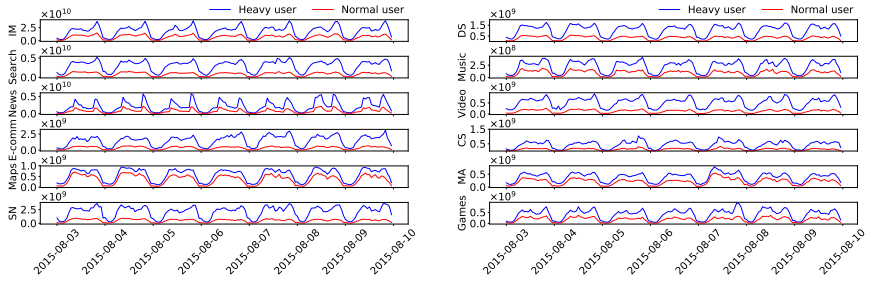
Fig. 6 shows the differences of generated data traffic between traffic heavy and normal users for different application categories. From Fig. 4 we know that, although the number of traffic heavy users only occupies 20% of total users, the number of online heavy users and the amount of generated data traffic are larger than normal users at any time. For each application category, the usage behavior of users is very different. For example, as shown in Fig. 6a, instant messaging, game, video, music, and social networking applications tend to consume large amount of data traffic before bedtime; users like to use map application during daytime; people usually read news around 9am and 7pm; data service, mobile assistant, and cloud service upload and download data during 6am and 10pm in the background. From the figure we can clearly see that, for the same application category, the amount of generated data traffic in every 15 minutes follows similar pattern for heavy and normal users in a week.

Table 2: Keywords in URLs for 12 application categories

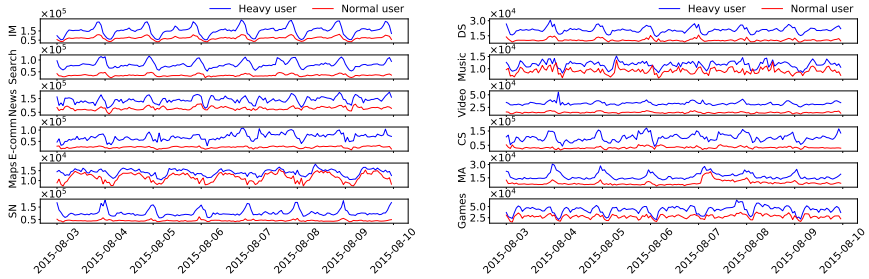
Interest	Keywords Example
Instant Messaging (IM)	*short.weixin.qq.com, chat.xiaomi.net, img.qq.com, ...* (30 in total)
Search	*m.baidu.com, www.baidu.com, get.sogou.com, ...* (19 in total)
News	*inews.gtimg.com, view.inews.qq.com, inews.gtimg.com, ...* (8 in total)
Maps	*map.baidu.com, map.qq.com, amap.com, ...* (11 in total)
E-commerce(E-comm)	*www.taobao.com, meituan.com, alipay.com, ...* (24 in total)
Social Networking(SN)	*mobile.qzone.qq.com, m.weibo.cn, api.weibo.cn, ...* (10 in total)
Data Service(DS)	*google-analytics.com, beacon.qq.com, talkingdata.net, ...* (7 in total)
Mobile Assistant(MA)	*zhushou.sogou.com, apple.com, mmarket.com, ...* (7 in total)
Games	*happyelements.com, mmocgame.qpic.cn, cmgame.com, ...* (4 in total)
Video	*video.qiyi.com, video.qq.com, youku.com, ... (7 in total)*
Cloud Service(CS)	*uu08.net, micloud.xiaomi.net, gxpan.cnhiido.com*
Music	*music.qq.com, kugou.com, kuwo.com, ...* (4 in total)

However, for some applications, such as news, map, mobile assistant, and game, the difference of data usage between heavy and normal uses is smaller than other applications. It implies that above applications are"daily necessities" for all users. If we look into the data usage behavior of individuals, as shown in Fig. 6b, it is very clear that the data usage patterns of e-commerce, social networking, video and cloud service are quite distinct for heavy and normal users. In other words, although traffic heavy users generate much more data traffic than traffic normal users on all kinds of applications, they seem to prefer to use *e-commerce, social networking, video* and *cloud service.*

It is also interesting to find out that "*how do heavy and normal users allocate their limited data traffic?*" and "*what's the difference of application usage preference between two groups of users?*". In order to further examine the difference of application usage preference between traffic heavy and normal users, we compare the data traffic allocation for two groups of users in each hour, as shown in Fig. 7. First, we calculate the proportion of consumed data traffic of each application category in each hour for traffic heavy and normal users respectively. Then we get the ratio of heavy users to normal users for above proportion value. Note that, since we only focus on the proportion of consumed data traffic for each application here, Fig. 7 illustrates the diversity



(a) The amount of data traffic



(b) Generated data traffic for each user

Fig. 6: The generated data traffic of (a) all traffic heavy and normal users, and (b) each traffic heavy and normal user, changing with time in a week for different application categories

of data traffic assigned to each application in each hour for two groups of users. Heavy users prefer to visit social networking, e-commerce, video, search and cloud service applications more, especially at midnight. In other words, heavy users spend much more data traffic on above five applications than others.

6 Data Traffic Usage Patterns of Heavy and Normal User from the View of Urban Functional Regions

Human activity shapes the spatial and temporal characteristics of the city, which influences the usage of network resource. It is well known that the usage of mobile network data traffic is highly uneven in spatio-temporal domain. In section 5, we have discovered the daily pattern and different preference of application usage for traffic heavy and normal users at different time. It is also interesting to know the spatial distribution of data traffic usage, which is caused by users' daily mobility behavior. According to our daily experiences, center area of city consumes much larger amount of data traffic compares to the edge of the city, mass data traffic is generated at business zone in work hours but decline sharply in non-work hours, and the opposite situation happens in residential area. In this section, we further compare spatial and temporal characteristics of data traffic usage for traffic heavy and normal users in the

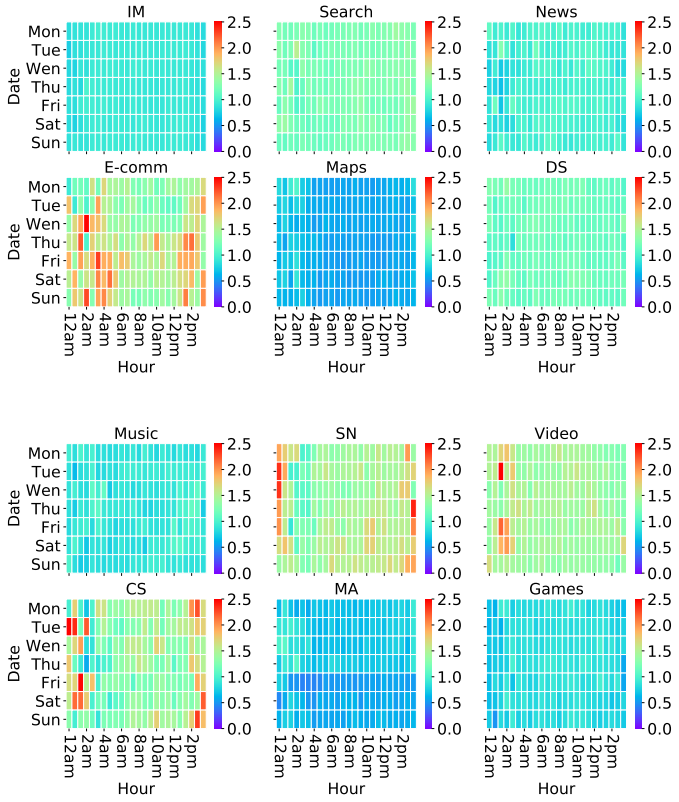


Fig. 7: The ratio of heavy users to normal users for proportion value of data traffic consumed by different applications in each hour of the day. The higher the ratio value is, the more heat the color will be.

city. The results may tell us whom we should focus on at a specific time and location to optimize the network resource.

In order to get an overview of relationship between users and data traffic for each base station, we count the number of distinct heavy or normal users and the amount of data traffic in each base station in a week, as shown in Fig. 8. Obviously, from base station's viewpoint, heavy users tend to generate much more data traffic than normal users. There exist a few base stations that consume a large amount of data traffic (around 10GB) with not many connected heavy users (around 2000 users). On the contrary, for normal users, nearly 17500 users connect with a base station but only consume 4.56G data traffic in a week. If we look into some extreme example in Fig. 8, we find that base stations in circle A are located at residential area (with high-rise apartments), and base stations in circle B are located at transport area (with train, subway stations, and airport) in the city. That is to say, some base

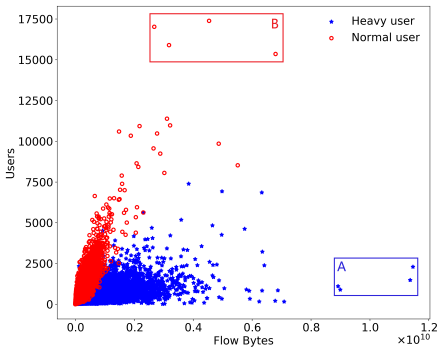


Fig. 8: The scatter plot of the amount of data traffic (x-axis) and the number of users (y-axis) for each base station

stations may need much more network resource due to the urban functions of area they are located. Therefore, we further study the data traffic usage pattern in different urban functional regions in the city.

Previous researches have illustrated that (1) the application usage in different locations has similar "signatures" at similar locations [38], (2) only five basic time-domain mobile traffic patterns of large scale base stations exist, which are related to urban ecology, including residential area, business district, transport, entertainment, and comprehensive area (a mixture of other four kinds of functional areas) [41]. Based on our dataset, we try to examine and compare the data usage patterns of traffic heavy and normal users at different urban functional regions in the city. In order to identify the urban functional regions in the city, we apply follow method.

First, for each base station j, we define a data traffic vector as

$$X_j = \{x_j[1], x_j[2], ..., x_j[N]\}$$

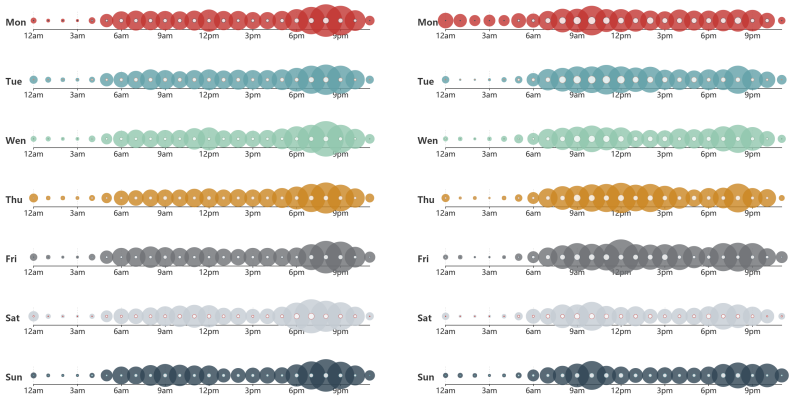
where $x_j[i]$ represents the normalized data traffic in the i-th hour interval, and $0 < x_j[i] < 1, N = 24 * 7 = 168, \{i, j | 1 \leq j \leq 2.4 \times 10^4; 1 \leq i \leq 168\}$. Second, we apply the unsupervised method, K-Means++ [46], with Euclidean distance as distance measure to cluster the base station, i.e., cluster the data traffic vectors. Here, Davies-Bouldin index [47] is used to evaluate the clusters under different value of k. Davies-Bouldin index can be denoted as

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq i} \{D_{i,j}\} \tag{1}$$

where $D_{i,j} = \frac{(\vec{d_i} + \vec{d_j})}{d_{i,j}}$, d_i and d_j are the average distance between every node to center in i-th and j-th clusters respectively, and $d_{i,j}$ is the Euclidean distance between the center of i-th and j-th clusters. At last, we have 7 clusters with

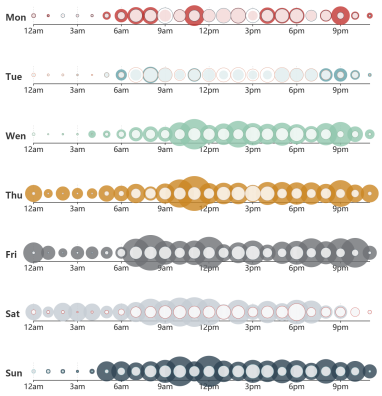
five kinds of data traffic patterns, which can be mapped to residential area, business district, transport, entertainment, and comprehensive area.

As shown in Fig. 9, different urban functional regions show distinct data usage patterns for traffic heavy and normal users. Colored circles and white circles on top of colored circles represent the amount of data traffic consumed by heavy users and normal users respectively. Here, the size of each circle presents the volume of data traffic at a specific hour, which is helpful to compare the difference between heavy and normal users in each sub-figure. Note that, for better visualization, the scale of each circle for different urban functional area is different. In residential area, data traffic peak appears at night for two groups of users, and heavy users consume much more data traffic at night, which implies that heavy users heavily rely on mobile Internet at night. In business district, a peak appears during work hours for two groups of users, but heavy users have another peak during 7pm and 9pm, means that heavy users may stay at work place longer than normal users, or work overtime makes users use more data traffic. As for transport area, i.e., airport, train, and subway station, the difference of amount of generated data traffic for heavy and normal users is very small. It implies that users may be busy at commuting and do not like to spend a lot of time browsing applications at transportation hub. At entertainment area, the amount of data traffic generated by heavy users has peaks in daytime or night, especially at night, but normal users do not have obvious peaks, which means heavy users like to consume a lot of data traffic at night in entertainment area. For comprehensive area (the combine of above four functional area), a significant peak of data traffic appears at night for heavy users, and normal users have two small peaks in daytime and at night. The analysis results show that urban functions have big influence on data traffic usage behavior of both heavy and normal users. In the next part, we will correlate the online browsing behavior of heavy and normal users with urban function, to further understand the data traffic usage patterns of users.

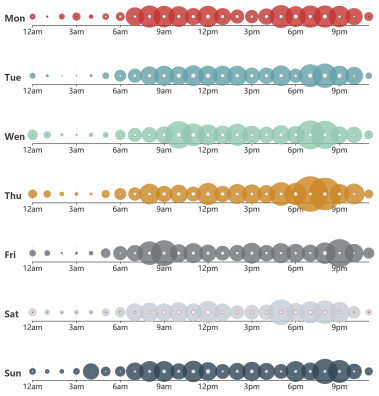


(a) Residential area

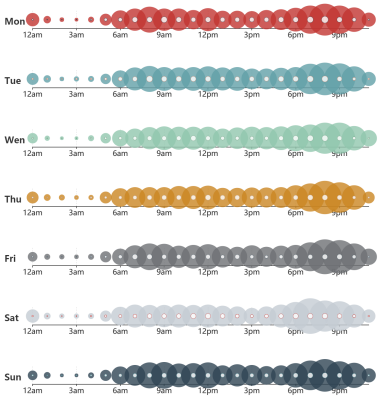
(b) Business district



(c) Transport area



(d) Entertainment area



(e) Comprehensive area

Fig. 9: The data traffic usage patterns of traffic heavy (colored circles below) and normal users (white circles on top) in every day at different urban functional areas

7 Spatio-temporal Dynamics of Heavy and Normal Users

In section 5 and section 6, we study the data traffic usage patterns of heavy and normal user from the view of online browsing (application usage) behavior and urban functional regions respectively. In this section, we investigate the spatio-temporal dynamics of heavy and normal users from the view of their interests to find out how and why the data traffic is consumed. The online browsing behavior and offline mobility behavior can reveal one's personal interests, for example, young people tend to spend a lot of time on online games, and entertain themselves at bar in the midnight; business man usually check mobile phone applications of trip, stock, and news, and they spend most of their time at workplaces; students like to shop online to buy some high quality and inexpensive stuffs, and they usually stay at campus. Based on above observations, we extract users' interests from their online browsing behavior, and visualize them in a spatio-temporal dynamic manner.

7.1 Extraction of users' online browsing behavior and location attributes

Fig. 10: Examples of categorization tags for items in Meituan and JD. Tags are marked by rectangle, and the left/right-most tag is high/low-level (general/specific) category.

The services that we study offer a large number of possible items that users can access. In order to understand users' online and offline preference, we extract tags of items that users browsed from ***Taobao, Tmall, Suning, JD (online shopping)*** and ***Meituan (Life service)***. For each flow, we also extract location (user can be located in the coverage of base station he/she connects with) and host (the application user visits) information, then we can

further get the area function information of each location as location tags, e.g., delicious food, fast food, snack bar, and so on, from map Application Programming Interface (API). For example, in Fig. 10, we show the tags extracted from webpages of Meituan and JD a user visited. Usually, the item ID, short for the distinct identifier of an item in a service, can be extracted from the related URL through the matching of regular expressions. By sending an item ID in a formatted HTTP GET request to the service's API, the item's tags can be retrieved as a JavaScript Object Notation (JSON) object. Using this method, the lower-level tags, which are most relevant to an item's features, are collected.

7.2 Topic Modeling

Under the idea that behavior of users that captures urban dynamics can be represented in a form of topics that are characterized by common preferences of users extracted from generated data traffic [48], we apply a probabilistic topic model, Latent Dirichlet Allocation (LDA), a baseline and powerful topic model, which provides a natural way to enhance the interpretability of our analysis for decision-making. It represents a time interval in an urban functional area with a distribution of online browsing activities and location attributes. We then analyze the topic distributions of various time intervals in order to study the correlation between data traffic, user behavior, and urban ecology.

1) Obtaining online behavior tags for each time interval

Based on the tags browsed by a user u_i, we remove duplicate tags and segment them to small words. According to Fig. 4 and 6, we divide four time intervals in a day, i.e., 12:00am - 6:00am, 6:00am - 12:00pm, 12:00pm - 18:00pm, and 18:00pm - 12:00am. Here, words extracted from tags browsed in each time interval are words in a document. Each time interval of the week is considered to be a document.

2) Calculating the distribution of words for each time interval

We calculate the Term Frequency-Inverse Document Frequency (TF-IDF) to measure the importance of a word in a time interval. For a given time interval t_i, v_{ij} is the TF-IDF value of the j-th word and n is the number of words. The TF-IDF value v_{ij} is given by:

$$v_{ij} = \frac{n_j}{N_i} \times \log \frac{T}{||\{w_i | word_j \in w_i\}||} \tag{2}$$

where n_i is the number of w_j and N_i is the number of words appeared in time interval t_i.

The IDF term is calculated by computing the quotient of the number of time interval T divided by the number of time interval containing the w_j, and taking the logarithm of that quotient.

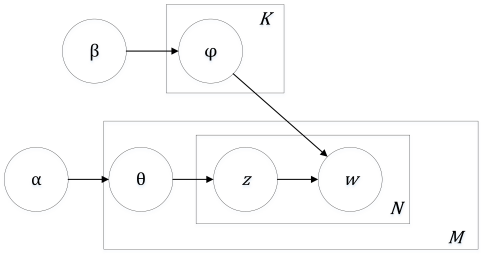


Fig. 11: Graphical models of Latent Dirichlet Allocation

In text mining, LDA has been successfully used to extract the hidden semantic structure in large archives of a document [49]. Fig. 11 is a generative model, introduced by Blei [49], in which each document of a corpus exhibits multiple topics and each word of a document support a certain topic. Given all the words $w = (w_{1 \cdot N})$ of each document in a corpus as observations, a topic model is trained to infer the hidden semantic structure behind the observation. In LDA, these inferred hidden semantic structure is a distribution over K topics $z = (z_{1 \cdot K})$.

LDA assumes a Dirichlet prior distribution on the topic mixture parameters θ and ϕ, to provide a complete generative model for documents. θ is an $M * K$ matrix of document-specific mixture weights for the K topics, each drawn from a Dirichlet(α) prior, with hyperparameter α. ϕ is an $V * K$ matrix of word-specific mixture weights over V vocabulary items for the K topics, each drawn from a Dirichlet (β) prior, with hyper parameter β.

The main objectives of LDA inference are to (1) find the probability of a word given each topic k, $p(w = t|z = k) = \phi_k^t$, and (2) find the probability of a topic given each document m, $p(z = k|d = m) = \theta_m^k$.

Several approximation techniques have been developed for inference and learning in the LDA model [49, 50]. In this work we adopt the Gibbs sampling approach [49]. For the LDA model visualized in Fig. 11, the following distributions hold:

$$p(\theta|\alpha) = p(\theta) \sim Dirichlet(\alpha) \tag{3}$$

$$p(\phi|\beta) = p(\phi) \sim Dirichlet(\beta) \tag{4}$$

$$p(z|\theta^{(d)}) \sim Multinomial(\theta^{(d)}) \tag{5}$$

$$p(w|z, \phi^{(z)}) \sim Multinomial(\phi^{(z)}) \tag{6}$$

Table 3: Analogy from online-behaviors to document topics

Time interval	→	Documents
Browsing behaviors	→	Topics of a document
Segmented words	→	Words

where $\phi^{(z)}$ represents the word distribution for topic z, and $\theta^{(d)}$ represents the topic distribution for document d. From the assumptions in equations (3)-(6), we obtain $p(w|z,\phi) = \prod_{k=1}^{K}\prod_{t=1}^{V}(\phi_k^t)^{n_k^t}$, where n_k^t is the number of times word t is assigned to topic k. n_k^t is also called the word-topic count and $n_k = \sum_{t=1}^{V} n_k^t$ is called the word-topic sum. We also obtain, $p(z|\theta) = \prod_{m=1}^{M}\prod_{k=1}^{K}(\phi_k^m)^{n_m^k}$, where n_m^k is the number of times topic k occurs in document m. n_m^k is also called the topic-document count, and $n_m = \sum_{k=1}^{K} n_m^k$ is called the topic-document sum.

7.3 Topic Models

The problem of identifying the latent online browsing behavior and location attribute in a time interval for an urban functional area can be analogized to the problem of discovering the latent topics of a document. As shown in Table 3, we regard a time interval as a document and a browsing behavior as a topic. The LDA model produces ϕ_t^k and θ_k^m , which represent the probability of words (w_i) for each topic k, and the probability of topics k for each time interval t_i, respectively. Given these probability distributions, we can rank words for each topic discovered and determine activities patterns, which are discovered as a set of topics.

7.4 Words Clouds

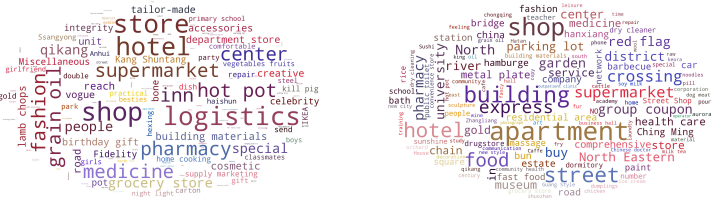
In Fig. 9, we find that the data usage pattern of heavy and normal user is very different in business district, residential, entertainment, and comprehensive area at night and weekends. In order to know why the difference emerges, after we getting the topic distribution per time interval, we set the number of topics to 3. It can be interpreted as decomposition of activities happened at daytime on weekdays, night on weekdays, and weekends. Then we apply LDA based model to data traffic consumed in business district, residential, entertainment, and comprehensive area, and visualize each topic as a word clouds, as shown in Fig. 12. The size of each word is proportional to the probability of it to belong to the given topic.

Fig. 12 shows the word cloud for a part of topics, which visualizes the most different online browsing preference of heavy and normal users. In business district, at weekday dusk, as shown in Fig. 12a, heavy users browse group coupon for dinner, normal users don't generate a lot of data traffic through

(a) Word cloud for topics in business district toward evening for heavy (left) at and normal users (right)

(b) Word cloud for topics in residential area at weekends for heavy (left) at and normal users (right)

(c) Word cloud for topics in entertainment area at evening for heavy (left) at and normal users (right)



(d) Word cloud for topics in comprehensive area at weekends for heavy (left) and normal users (right)

Fig. 12: Word cloud for topics in business district, residential, entertainment, and comprehensive area

mobile network, which implies that heavy users tend to and choose what to eat online for dinner at work place. Normal users seldom use mobile data traffic at business district, may be because they go home directly after work or they focus on work or they use computer during work hours. In residential area, normal users browse *shopping, delicious food*, and *group coupon* at weekends much more than workday, which implies a "stay at home" life style. In en-

tertainment area, we compare the work cloud of topics for heavy and normal user at evening. In Fig. 12c, the big size words in word cloud for heavy users include *inn, clothing, boutique, massage*, and so on, which reveals the interests for many kinds of entertainment. For normal users, they just browse *delicious food*, and *group coupon* for meal. In Fig. 12d, we can see that, at weekends, in comprehensive area, heavy users browse many consumption-related webpages, such as, *delicious food, fashion things*, and they usually stay near with *shop, supermarket, inn, hotel*, and so on. But, at the same time, normal users pay more attention to life related information, and they usually stay at residential areas. It means that users that don't stay at home like to consume much more traffic on shopping online at weekends in comprehensive area. Above figures and findings connect data traffic, user behavior, and urban ecology together in a visualization way. It provides a method to help us understand how and why the data traffic is consumed, which is very crucial for mobile network with a human-in-the-loop architecture.

8 Conclusion

In this work, we present HLA-MN, a Human-in-the-Loop Architecture for Mobile Network, which takes feedbacks of users into account dynamically by analyzing data traffic they generated. HLA-MN collects data traffic from mobile network and processes it online in real time to detect the needs of users, and abnormal traffic, or offline manually to discover data traffic usage patterns from flow, user, and application level. In order to understand how the network resource is consumed, we conduct several experiments on heavy and normal users from the view of online browsing behavior, urban functional regions, and spatio-temporal dynamics. We find that the two groups of users may be distinguished by only paying attention to data usage patterns of users during midnight. In other words, we may quickly identify a traffic heavy or normal user by observing his/her data traffic usage behavior at the midnight. The difference of application usage behavior between two groups of users appears mainly in e-commerce, social networking, video and cloud service applications, especially, the difference becomes bigger at midnight. It implies that heavy users strongly rely on above applications. When we focus on urban functional areas in the city, heavy users tend to consume much more data traffic at night in residential area, business region, and entertainment area compare with normal users, which means that the region function will influence the usage behavior of users. Furthermore, we apply topic model to correlate data traffic, users' interests, locations, urban functions, and temporal and spatial dynamics in the city. We find that daily activities of users, life style, and urban ecology have big influence on data traffic usage pattern. More specifically, interesting results includes: heavy users browse delicious food, and group coupon in business district at dusk, but normal users try to choose what to eat in residential area at weekends; in entertainment area, for online browsing behavior, heavy users are interested in inn, clothing, boutique, massage, but normal users only

find information about meal. Above analysis shows how the network resource is consumed by two groups of users from the view of temporal and spatial characteristics of data traffic, user behavior, and urban ecology, the results of which may guide ISP to pay attention to heavy users at specific time and locations for their usage behavior of some applications.

The rapid growth in mobile data requires mobile service providers to efficiently manage their resources and evolve their infrastructure so as to meet the growing demands and diverse expectations of mobile data users. For future work, we should build models to explain the correlation between user behavior, urban ecology, and data traffic, then establish the designed system in real production environment. In this way we can provide real data based strategies to allocate network resource, and further examine the methods we proposed according to feedbacks from real mobile networks.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (61671078), Funds of Beijing Laboratory of Advanced Information Networks of BUPT, Funds of Beijing Key Laboratory of Network System Architecture and Convergence of BUPT, 111 Project of China (B08004), and EU FP7 IRSES MobileCloud Project (612212).

References

1. Peter Brooks and Bjørn Hestnes. User measures of quality of experience: why being objective and quantitative is important. *IEEE network*, 24(2), 2010.
2. Alan Dix. Human-computer interaction. In *Encyclopedia of database systems*, pages 1327–1331. Springer, 2009.
3. Eirini Liotou, Hisham Elshaer, Raimund Schatz, Ralf Irmer, Mischa Dohler, Nikos Passas, and Lazaros Merakos. Shaping qoe in the 5g ecosystem. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, pages 1–6. IEEE, 2015.
4. Cisco Visual Networking Index. Global mobile data traffic forecast update, 2015–2020 white paper. *link: http://goo. gl/ylTuVx*, 2016.
5. Jie Yang, Yuanyuan Qiao, Xinyu Zhang, Haiyang He, Fang Liu, and Gang Cheng. Characterizing user behavior in mobile internet. *IEEE transactions on emerging topics in computing*, 3(1):95–106, 2015.
6. Yu Jin, Nick Duffield, Alexandre Gerber, Patrick Haffner, Wen-Ling Hsu, Guy Jacobson, Subhabrata Sen, Shobha Venkataraman, and Zhi-Li Zhang. Characterizing data usage patterns in a large cellular network. In *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, pages 7–12. ACM, 2012.
7. Sahar Hoteit, Stefano Secci, and Marco Premoli. Crowded spot estimator for urban cellular networks. *Annals of Telecommunications*, pages 1–12, 2017.
8. A t foaaamiGH KXTatart. Dod modeling and simulation (m&s) glossary.
9. Waldemar Karwowski. *International encyclopedia of ergonomics and human factors*, volume 3. Crc Press, 2001.
10. Jiakang Lu, Tamim Sookoor, Vijay Srinivasan, Ge Gao, Brian Holben, John Stankovic, Eric Field, and Kamin Whitehouse. The smart thermostat: using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 211–224. ACM, 2010.

11. Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel Watson, Sunny Consolvo, and Julie A Kientz. Lullaby: a capture & access system for understanding the sleep environment. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 226–234. ACM, 2012.
12. G Burnham, Jinbom Seo, and G Bekey. Identification of human driver models in car following. *IEEE transactions on Automatic Control*, 19(6):911–915, 1974.
13. Randy L Sollenberger, Ben Willems, Pamela S Della Rocco, Anton Koros, and Todd Truitt. Human-in-the-loop simulation evaluating the collocation of the user request evaluation tool, traffic management advisor, and controller-pilot data link communications: Experiment i–tool combinations. 2005.
14. Sirajum Munir, John A Stankovic, Chieh-Jan Mike Liang, and Shan Lin. Cyber physical system challenges for human-in-the-loop control. In *Feedback Computing*, 2013.
15. Gunar Schirner, Deniz Erdogmus, Kaushik Chowdhury, and Taskin Padir. The future of human-in-the-loop cyber-physical systems. *Computer*, 46(1):36–45, 2013.
16. Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.
17. Diala Naboulsi, Razvan Stanica, and Marco Fiore. Classifying call profiles in large-scale mobile traffic datasets. In *INFOCOM, 2014 Proceedings IEEE*, pages 1806–1814. IEEE, 2014.
18. Utpal Paul, Anand Prabhu Subramanian, Milind Madhav Buddhikot, and Samir R Das. Understanding traffic dynamics in cellular data networks. In *INFOCOM, 2011 Proceedings IEEE*, pages 882–890. IEEE, 2011.
19. Ram Keralapura, Antonio Nucci, Zhi-Li Zhang, and Lixin Gao. Profiling users in a 3g network using hourglass co-clustering. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, pages 341–352. ACM, 2010.
20. M Zubair Shafiq, Lusheng Ji, Alex X Liu, and Jia Wang. Characterizing and modeling internet traffic dynamics of cellular devices. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 305–316. ACM, 2011.
21. Zhanyu Ma, Jiyang Xie, Hailong Li, Qie Sun, Zhongwei Si, Jianhua Zhang, and Jun Guo. The role of data analysis in the development of intelligent energy networks. *arXiv preprint arXiv:1705.11132*, 2017.
22. Daniel Willkomm, Sridhar Machiraju, Jean Bolot, and Adam Wolisz. Primary users in cellular networks: A large-scale measurement study. In *New frontiers in dynamic spectrum access networks, 2008. DySPAN 2008. 3rd IEEE symposium on*, pages 1–11. IEEE, 2008.
23. M Cerinsek, J Bodlaj, and V Batagelj. Symbolic clustering of users and antennae. *Net-Mob D4D Challenge, Boston, MA, USA*, 2013.
24. M Zubair Shafiq, Lusheng Ji, Alex X Liu, Jeffrey Pang, and Jia Wang. Characterizing geospatial dynamics of application usage in a 3g cellular data network. In *INFOCOM, 2012 Proceedings IEEE*, pages 1341–1349. IEEE, 2012.
25. Zhanyu Ma, Jing-Hao Xue, Arne Leijon, Zheng-Hua Tan, Zhen Yang, and Jun Guo. Decorrelation of neutral vector variables: Theory and applications. *IEEE transactions on neural networks and learning systems*, 2016.
26. Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of physics A: mathematical and theoretical*, 41(22):224015, 2008.
27. Richard A Becker, Ramón Cáceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Clustering anonymized mobile call detail records to find usage groups. In *Workshop on Pervasive and Urban Applications (PURBA)*, 2011.
28. Ionut Trestian, Supranamaya Ranjan, Aleksandar Kuzmanovic, and Antonio Nucci. Measuring serendipity: connecting people, locations and interests in a mobile 3g network. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 267–279. ACM, 2009.

29. Zhanyu Ma and Arne Leijon. Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2160–2173, 2011.
30. Ying Zhang and Ake Årvidsson. Understanding the characteristics of cellular data traffic. *ACM SIGCOMM Computer Communication Review*, 42(4):461–466, 2012.
31. Zhanyu Ma, Pravin Kumar Rana, Jalil Taghia, Markus Flierl, and Arne Leijon. Bayesian estimation of dirichlet mixture model with variational inference. *Pattern Recognition*, 47(9):3143–3157, 2014.
32. Marta C Gonzalez, Cesar A Hidalgo, and A-L Barabasi. Understanding individual human mobility patterns. *arXiv preprint arXiv:0806.1256*, 2008.
33. Xuan Zhou, Zhifeng Zhao, Rongpeng Li, Yifan Zhou, Jacques Palicot, and Honggang Zhang. Human mobility patterns in cellular networks. *IEEE communications letters*, 17(10):1877–1880, 2013.
34. Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modeling the scaling properties of human mobility. *arXiv preprint arXiv:1010.0436*, 2010.
35. Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *arXiv preprint arXiv:1111.0586*, 2011.
36. Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. Diversity in smartphone usage. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 179–194. ACM, 2010.
37. Zhanyu Ma, Arne Leijon, and W Bastiaan Kleijn. Vector quantization of lsf parameters with a mixture of dirichlet distributions. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1777–1790, 2013.
38. Yuanyuan Qiao, Xiaoxing Zhao, Jie Yang, and Jiajia Liu. Mobile big-data-driven rating framework: measuring the relationship between human mobility and app usage behavior. *IEEE Network*, 30(3):14–21, 2016.
39. Keun-Woo Lim, Stefano Secci, Lionel Tabourier, and Badis Tebbani. Characterizing and predicting mobile application usage. *Computer Communications*, 95:82–94, 2016.
40. Aveek K Das, Parth H Pathak, Chen-Nee Chuah, and Prasant Mohapatra. Contextual localization through network traffic analysis. In *INFOCOM, 2014 Proceedings IEEE*, pages 925–933. IEEE, 2014.
41. Huandong Wang, Fengli Xu, Yong Li, Pengyu Zhang, and Depeng Jin. Understanding mobile traffic patterns of large scale cellular towers in urban environment. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 225–238. ACM, 2015.
42. Sahar Hoteit, Stefano Secci, Zhuochao He, Cezary Ziemlicki, Zbigniew Smoreda, Carlo Ratti, and Guy Pujolle. Content consumption cartography of the paris urban region using cellular probe data. In *Proceedings of the first workshop on Urban networking*, pages 43–48. ACM, 2012.
43. Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. Mobile traffic analysis: a survey. *Université de Lyon, Tech. Rep. hal-01132385*, 2015.
44. Telecomitalia. Big data challenge 2015, 2015.
45. Yuanyuan Qiao, Yihang Cheng, Jie Yang, Jiajia Liu, and Nei Kato. A mobility analytical framework for big mobile data in densely populated area. *IEEE transactions on vehicular technology*, 66(2):1443–1455, 2017.
46. David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.
47. David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.
48. Felix Kling and Alexei Pozdnoukhov. When a city tells a story: urban topic analysis. In *Proceedings of the 20th international conference on advances in geographic information systems*, pages 482–485. ACM, 2012.
49. Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*, pages 17–24, 2004.

50. Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):712–725, 2015.