Understanding Mobile Internet Usage Behavior

Xueli An and Gerald Kunzmann
DOCOMO Communications Labs Europe GmbH, Munich, Germany
Email: {an_de_luca, kunzmann}@docomolab-euro.com

***Abstract*—Exploring the semantic contents of mobile traffic has significant meaning for Telco operators to gain a better understanding of the traffic generated by their subscribers. The mobile traffic dynamics do not only provide guidance for operators in terms of network planning and resource management, they also help operators to innovate new business models by providing value-added services. The mobile traffic dynamics are not static features. They are developing with the evolution of radio access technologies, subscribers' adoption of mobile services, etc. In this paper, we analyze the mobile traffic from multiple dimensions based on a large-scale dataset collected by a major European mobile operator in 2012. Within this work, we investigate the traffic dynamics and the application features used by different mobile devices. Moreover, for the first time, we also present a methodology about how to link these metrics to the operating system of mobile devices. We observed a noticeable impact of the operating system on some of the evaluated features.**

I. Introduction

Due to the promising evolution of mobile terminals in terms of processing power together with the advanced innovation of their user interface design, mobile devices like smart phones and tablets become the common ways for data access other than personal computers. The third generation technologies of mobile telecommunications (3G), like UMTS and CDMA2000, have been deployed worldwide, and the commercial 4G LTE networks are also deployed on a stunning speed, which promises mobile devices with an ultra-broadband access. Cisco has predicted that global mobile data traffic will increase 13-fold between 2012 and 2017, and the number of mobile devices will be more than the entire global population by the end of 2013 [1]. Since the mobile Internet usage becomes main stream, it changes the conventional voice service based traffic model and also the corresponding resource usage patterns. For instance, mobile devices break the bearer of Internet access at fixed locations, e.g. home, office, Internet-cafe, etc. The usage of web-based services becomes more dynamic without the space and temporal constraints. Due to the compelling billing plans as well as the service performance, mobile services like VoIP (e.g. Skype) are preferred instead of Telco conventional voice call services, and instant messaging software (e.g. WhatsApp) is preferred to SMS. This brings new challenges for Telco operators and leads them to re-think or even re-design their network and resource deployments, for instance reducing the CAPEX and OPEX via optimized resource allocation and core network planning. On the other hand, it also offers new opportunities upon conventional Telco services by innovative business models, for instance, via smart service bundling or promotion. To address such challenges and turn them into opportunities, it is essential for the operators to gain better understandings about the traffic carried by their infrastructure and also the mobile users generating this traffic. For instance, how do customers interact with mobile services and what are the corresponding impacts for the mobile networks; how could Telco operators target the right user groups to provide value-added services other than dump pipes? Seeking the answers for the above mentioned questions is the major motivation of this work.

ISBN 978-3-901882-58-6 © 2014 IFIP

Data analytics have been widely used for voice call based traffic dynamics modeling [2], [3], [4], [5] and human mobility trajectory modeling [6], [7], [8] in cellular networks. Mobile data plane traffic dynamics draw attention since the emergence of mobile broadband services [9], [10], [11], [12]. Compared to the above mentioned related work, the major contributions of our work are as follows: Our analysis is based on a unique dataset, which contains a tremendous amount of mobile Internet usage data in terms of HTTP requests collected by a major European mobile operator in 2012. Such large scale dataset could provide a general and critical view of mobile Internet usage. Based on this dataset, we investigate the mobile Internet usage behavior from multiple dimensions: traffic dynamics, mobile user profiling, application category usage features and also the operating systems (OS) running on the mobile devices. It is the first attempt in the field that links all these metrics together to obtain the insights of mobile traffic.

The rest of the paper is organized as follows. In Section II, we introduce our dataset, data aggregation, application category mapping mechanism, and the OS information extraction methodology. The traffic dynamics and the temporal features of our dataset are investigated in Section III. Section IV introduces the taxonomy for application usage behavior. In Section V, the device feature extraction methodology is explained. We analyze the influence of the OS in Section VI. Related work is discussed in Section VII and we discuss the major contributions and conclude this work in Section VIII.

II. Dataset Features and Activity Categorization

A. Dataset Features

The dataset used as the basis of our investigation is from a commercial mobile cellular network. It was collected by Orange in France as part of the European project SAIL [13]. We used a subset containing around 400 TB of traces comprising 7 days in the mid of January 2012. It covers roughly 6.7 million mobile devices (including mobile phones, tablets and 2G/3G dongles). The trace contains around 820 billion HTTP

requests seen during this collecting period. For each request we know, amongst others, the target IP address, host, and URL, as well as anonymous customer ID for each mobile device and their own IP addresses. However, it does not directly contain information about content volume and device specific parameters like operating system. Therefore, we try to derive those information in an indirect way. We use the amount of HTTP requests as a reflection of the amount of traffic going through the mobile network. We are aware that, the amount of HTTP requests is not equal to the traffic load generated by using a web service, however, for instance, if device x has more requests to stream video than device y, we could assume that x generates more video traffic than y. In a similar way, we try to estimate a device's OS by analyzing its HTTP requests, further details will be explained in section II-D).

B. Application Category Mapping

The major advantage of a dataset with HTTP request records is that it provides the possibility to obtain the semantic meaning of the traffic flowing through the mobile network. We first categorize the URLs of the requests into 76 URL categories following Cisco's IronPort Web Usage Controls [14]. These 76 URL categories are then further grouped into 12 general application categories as listed in Table I. URLs that cannot be categorized by this process are marked as category Null, which is around 11.2% over all the HTTP request load. In this way, we map each HTTP request into one of the application categories. Different application categories have different impacts on featuring a mobile device. For instance, categories like Video/Audio, Communication, Shopping, Gaming, etc. are considered as strong features to profile mobile Internet usage activities. This is because Video/Audio and Communication have high implications on the used traffic volume, and Shopping and Gaming have high implications on business models for mobile services. In contrast to this, categories like Search, Web and Advertisement are web browsing based services, which are considered as commonly used categories, and thus, they have few implications to profile a mobile device. Therefore, within these 12 categories, a Featuring Priority (FP), which is denoted as f_p and $f_p \in [1, 3]$, is assigned for each category as shown in Table I. A category j with $f_p(j) = 1$ is considered having a strong feature in our following evaluation. In comparison, $f_p(j) = 3$ is considered having a weak feature for mobile device profiling. For category Work and Education (WE), even though it is also a web browsing based category, the web services falling into this category are working, education and business related, which provides implications on how people use their mobile devices for professional activities. Therefore, we consider WE has a stronger feature ($f_p = 2$) than other web browsing services.

C. Data Aggregation

In order to capture the overall mobile Internet usage feature based on the application category, we keep the data analysis granularity at the application category level. Therefore, we aggregate HTTP requests generated by a mobile device to time slots of 5 minutes and grouped by application category. For instance, if a mobile device utilizes any service from a certain category in a time slot, it generates one *data sample*, the value of which is the aggregated amount of HTTP requests from this category in this time slot. As a consequence, the maximum number of data samples that could be generated by one device within one time slot is 12, i.e. the total amount of categories.

TABLE I
APPLICATION CATEGORY DEFINITION

ID	Category (Abbr.)	Description	FP
1	Video and Audio (V)	Video, audio, streaming media	1
2	Communication (C)	Chat, mail services, Internet telephony, SMS services/ring tones	1
3	Search (SR)	Search engines and portals	3
4	Web (W)	General web content, like news, sports, finance, health	3
5	Shopping (S)	Shopping, online brokerages	1
6	Gaming (G)	Online games like card, board	1
7	File Sharing (F)	p2p file sharing,storage/backup,update	1
8	Adult/Porn (P)	Adult and pornographic sites	1
9	Advertisements (A)	Banner and pop-up ads, advertising sites	3
10	Social Network (SN)	Social networks, dating sites	1
11	Work/Edu.(WE)	Business,edu.,career,science,technology	2
12	Null (N)	Uncategorized URLs	-

D. OS Information Extraction Methodology

The OS is an important feature to profile mobile devices. The OS feature could have direct influence on the target customer groups in terms of demographic and usage purposes, e.g. private users or enterprise business users. For instance, Android is an open-source software stack based on the Linux kernel. Its major business model is based on advertisement. In comparison, Apple iOS, which is developed and distributed by Apple Inc., focuses more on product purchases. Therefore, the usage behavior for mobile devices with different OS could also be different. In order to learn about user behavior based on the OS types, we use a reverse engineering method to retrieve OS information from the HTTP requests. For instance, we filter the data for websites specific to certain OS, e.g. a user with several requests to `android.clients.google.com` is likely to run Android OS, whereas requests for `iTunes.com` are more likely from iOS users. In our evaluation, for each user we estimate the OS, where most HTTP requests for selected web pages are observed from this user. The more hits for a specific OS-related website, the higher the probability that the guess is correct. Using this method, we obtain 137242 devices with OS information, i.e. Windows (including PC and Windows phones), iOS (including MAC, iPhone and iPad), Android (including Linux and Android phones), Blackberry and Symbian. Based on our observation, these 137242 devices generate more requests than most devices that we could not retrieve any OS information. There are also cases that some devices generate more requests than certain devices which could retrieve OS information, however, we observed that this is a very rare case, and thus, we do not consider such devices in the following analysis.

III. OVERALL CHARACTERISTIC OF THE DATASET

A. Traffic Dynamics

We first identify the overall traffic characteristic of our measured mobile devices. Traffic dynamics we mention here

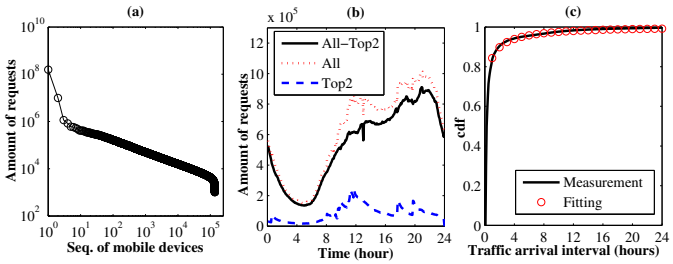

Fig. 1. Overall traffic statistics

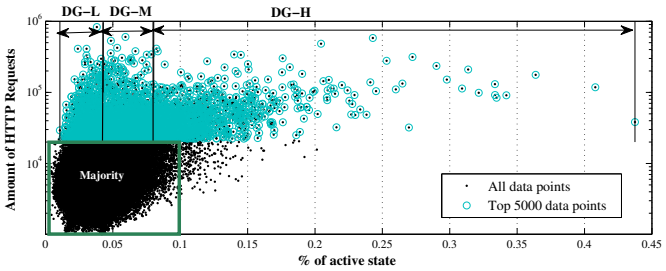

Fig. 2. Relationship between request load and η

refer to the characteristics of HTTP requests. As we mentioned before, even though our dataset does not contain the exact amount of traffic volume on the bit-level, there is an implicit relationship between the number of HTTP requests and the traffic load. By assigning a certain average load of requests in a certain application type (e.g. high for video and low for web browsing), the traffic volume can be estimated.

Figure 1 depicts the overall traffic statistics over seven days. Subplot (a) is a log-log plot showing the total amount of HTTP requests generated by each mobile device and sorted in descending order. The curve consists of three segments: (1) The total amount of requests from the majority of devices shows a linear trend in the log-log plot; (2) a small amount of devices forms a short tail in the plot; (3) the top two devices generate a significant amount of requests (13.31% over the total amount of requests) compared to the rest of devices. These two devices could be the operator's in-house testing devices or devices with flat-rate contract generating tremendous amounts of traffic in the mobile networks. In any case, such devices are considered as abnormal devices, a few amounts of which could totally mess up the performance of a mobile cellular network. In [11], the author reported that 1% of users in the network create 60% of the traffic load according to the dataset collected from 2007. In [9], the authors reported that the top 1% users create 30% of the traffic load according to the dataset collected from 2010. If we assume that traffic load is proportional to the request load, our result shows that the top 3.6% (5000 devices) generate 30% of the request load. This observation indicates that *the impact from the top users on the overall traffic becomes smaller and smaller over the years*. On the other hand, this also indicates that *subscribers are adopting the usage of mobile devices like smartphone and tablet for Internet access*, which creates a higher demand for radio and network resource.

We plot the averaged diurnal request load in Figure 1 (b) according to three scenarios: over all the mobile devices, over all the mobile devices without the top two devices, and only the top two devices. The results further indicate the significant influence of the abnormal devices. Moreover, the results also show that the request load peak shows up in the night around 21:00. This is an interesting observation, because it depicts a different diurnal traffic pattern compared to conventional voice-oriented mobile services [15] or fixed services [16].

For each device, we record the time interval between each two discontinuous data samples, which is called as request inter-arrival time. We accumulate the request inter-arrival times collected from all devices and plot their cumulative distribution function (cdf) in Figure 1 (c). This cdf curve can be fit by a Weibull distribution (with $\lambda = 0.07$ and $k = 0.3$), the cdf of which is given as $F(x;k,\lambda) = 1 - e^{-(x/\lambda)^k}$.

B. Relations between Request Load and Amount of Samples

The HTTP request load can be linked to the traffic load and the amount of data samples represents how often a device uses Internet services. In order to gain a better understanding of the dataset, it is essential to know what is the relationship between these two metrics. The maximum amount of data samples that a device could generate is the multiplication of the total amount of time slots and the number of categories. For each device, we divide the amount of generated data samples by the maximum number. The obtained ratio is defined as the percentage of active state η, which represents how active a mobile device is over the entire measuring period among all the categories. The relationship between the request load and η is plotted in Figure 2, in which the top 2 devices are excluded from the total 137242 devices due to their abnormal request loads (i.e. their η is close to 1). In general, the amount of HTTP requests increases once η increases, but this is not a very strong correlation, because for the same value of η, the amount of HTTP requests varies in a large range. Once η is increased, the range of HTTP requests variance however is decreased. This indicates an *asymmetric relation between the number of HTTP requests and the usage frequency*, i.e., on one hand, many HTTP requests do not necessarily imply a frequent usage of the mobile Internet. However, on the other hand, someone frequently using the mobile Internet normally also generates a high amount of HTTP requests. In Figure 2, we emphasis the 5000 data points with the top amount of requests using green circles. As shown in the figure, these 5000 data points cover a wide range of request load and active state. Another observation of this plot is that the majority of data points remains in a similar area as indicated in the figure, for which the percentage of active state is smaller than 0.1 and the request load is less than of 2×10^4.

C. Temporal Features of the Dataset

As indicated in the previous section, the data points from the top 5000 devices cover a wide range of request load and also usage frequency. In this section, we would like to profile these 5000 devices in terms of temporal features. We further classify these 5000 devices in three different Device Groups (DG) according to their active states: top 25% devices as

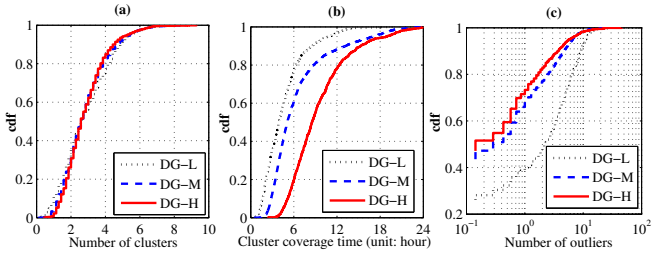

Fig. 3. Cluster performance, ϵ for DBSCAN is set to 30 mins

high usage devices (DG-H), lower 25% devices as **low usage** devices (DG-L), and the devices in the middle as **medium usage** devices (DG-M). The range of η for each DG is shown in Figure 2. We expect that the data samples generated by mobile devices show a clustering effect. For instance, a user may have intensive usage of his smart phone on the way to work by train or tram in the morning, he/she may check email or social networks after lunch break, etc. Therefore, we apply clustering techniques to obtain the temporal features of the mobile Internet usage for individual mobile devices. We use the DBSCAN [17] algorithm to cluster the daily data samples collected from different DGs: DG-L, DG-M and DG-H. Even though the selected devices are top 5000 in terms of generated request load in the entire dataset, we still observe that in average around 10% of the top users have less than 10 samples in a certain day. Hence, we set two rules to form clusters. If the amount of daily data samples is less than 10, the number of data samples in one cluster should be at least 1 data sample in a cluster; otherwise, a cluster is formed when there are more than 5 data samples in it.

The average number of obtained clusters over the entire collecting week is shown in Figure 3 (a). It is very interesting to observe that devices from different DGs show similar performance in terms of the number of generated clusters, which means that a finite number of clusters can be used to describe a mobile user's daily activities and *the number of formed temporal clusters is not sensitive with respect to the number of data samples generated by the mobile devices*. For a certain cluster i of device j, which contains x data samples generated at time $[t^1_{i,j}, ...t^x_{i,j}]$, we define the cluster coverage time as the time difference between the samples with the biggest and smallest time values as $\max(t^{x'}_{i,j}) - \min(t^{x'}_{i,j})$, where $\forall x' \in [1, x]$. The entire cluster coverage time within a day is the summation of the coverage times of all the clusters. The cdf of the entire cluster coverage time is plotted in Figure 3 (b). It indicates that, *even though the number of clusters is similar for the devices in different DGs, the clusters with higher amount of data samples tend to be bigger in terms of cluster coverage time.* In Figure 3 (c), we show the clustering performance by using the percentage of outliers after forming the clusters. Although the amount of data samples for DG-H spreads in a much wider range than the data samples for DG-M, the clustering performance for DG-M and DG-H are very close to each other, and they are better than the clustering performance of DG-L. Around 70% devices in DG-M and DG-H have an average number of outliers less than 1.

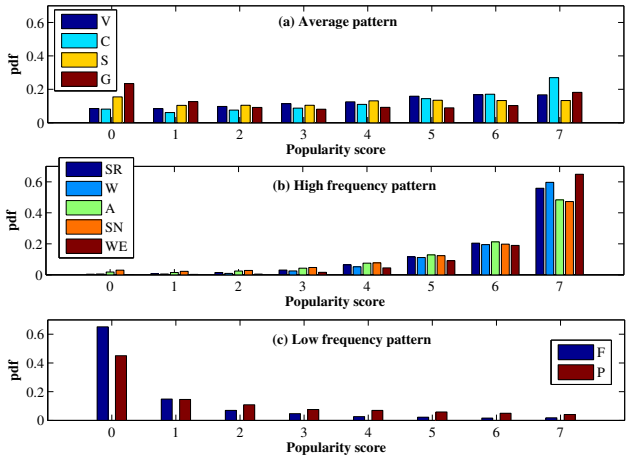

Fig. 4. Application popularity comparison

IV. Application Category Usage Behavior

As we mentioned in Section II-B, we categorized all HTTP requests into 12 high-level application categories as listed in Table I. In this section, we focus on the analysis of application category usage behavior from four different aspects.

A. Category Popularities

The application category popularity can be revealed from the aspect of usage frequency. We use a binary indicator $\kappa^c_x(i)$ to represent the daily application category usage. If device x uses any service from application category c at day i, $\kappa^c_x(i)$ is set to 1, otherwise it is set to zero. We define the popularity score of category c of device x over n_m measuring days as $\sum_{i=1}^{n_m} \kappa^c_x(i)$. The pdf of the popularity score for each category over the top 5000 mobile devices is plotted in Figure 4. The x-axis is the popularity score varying from 0 (not popular at all) to 7 (devices use services from a certain category every day). Within this figure, we further bundle the application categories into three groups according to the similarity of their popularity score distributions:

1) Average pattern - AP: As shown in Figure 4 (a), there is no obvious usage preference observed from the categories that are listed in this group, which means that the popularity score is rather equally distributed within the range $[0, n_m]$. There is a certain amount of devices that do not use the services from this group at all, and there are also some devices that use them on a daily basis. The categories falling in this pattern are [V, C, S, G], which can be considered as the categories that are in the adopting process, meaning that they have been accepted by a certain share of users, but not the majority.

2) High frequency pattern - HP: As shown in Figure 4 (b), all devices have obvious preference to frequently use the services classified into this group. For instance, more than half of the devices use the services in this group in each measuring day. The categories falling into this pattern are [SR, W, A, SN, WE]. It is interesting to see that Advertisement is in this group, which indicates that the mobile advertisement market has shown a great business potential. Also other web browsing based services and social networks have been well adopted by mobile users.

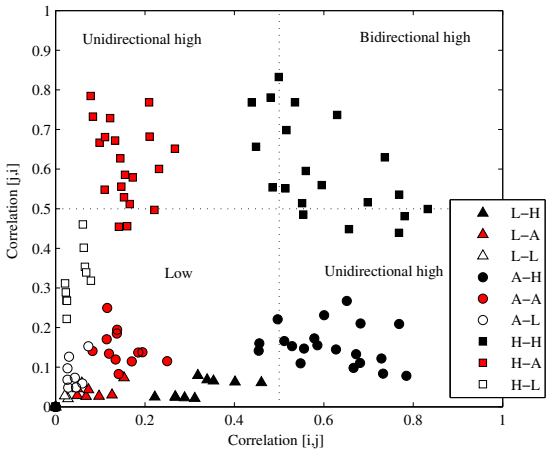

Fig. 5. Application category correlations

3) Low frequency pattern - LP: As shown in Figure 4 (c), services provided from this group are only targeting specific users. Most of the devices do not use them during the entire measuring period. The categories in this pattern are [F,P].

B. Category Correlations

On PCs, users tend to engage with multiple web activities simultaneously, which is called as the concurrency of application usage. Compared to PCs, mobile devices are used at different locations with different purposes, which might result in a different web service usage behavior. In this section, we aim to reveal the correlations between different application categories for mobile devices.

Due to the user interface limitations of the mobile devices, it is difficult to use multiple applications at the same time. However, it is possible for a mobile user to switch from one application to another application within a certain time period. For instance, a user could write an email, use a dictionary application to check a word and then switch back to the email service, etc. Therefore, we define the mobile application correlation as the current usage of any two applications in a certain time period. For device x, we use $l_x^i(t)$ to represent the time series of the amount of HTTP requests of category i at time slot t. Parameter a_x^i represents the set of the sequence number for the time slots in which $l_x^i(t)|_{\forall t} > 0$. We use $\nu_x(i,j)$ to express the dependency coefficient between category i and j. It is defined by the probability for category i to happen when category j is present. Hence, we have $\nu_x(i,j) = \frac{|a_x^i \cap a_x^j|}{|a_x^i|}$ for $i \neq j$, and $\nu_x(i,j) = 0$ for $i = j$. $|A|$ represents the cardinality of set A. To obtain the correlation relationship between any two application categories $\nu(i,j)$, we average the correlation values among the top 5000 devices. The correlation $\nu(i,j)$ is not necessary equal to $\nu(j,i)$. Three kinds of correlation relationship are defined as follows:

- Bidirectional high: $\nu(i,j) \geq .5$ & $\nu(j,i) \geq .5$, which means that two categories have high correlation with each other.
- Unidirectional high: $\nu(i,j) \geq .5$ & $\nu(j,i) < .5$, or $\nu(i,j) < .5$ & $\nu(j,i) \geq .5$, which means that only one category has high correlation with the other one, but not vice versa.

TABLE II
APPLICATION CATEGORY TEMPORAL SIMILARITY (WITH SMOOTHING FACTOR 0.9)

	V	C	S	G	F	P	SN	WE
V	0	0,44	0,32	0,39	0,34	0,39	0,41	0,32
C	0,44	0	0,37	0,68	0,23	0,41	0,64	0,63
S	0,32	0,37	0	0,34	0,21	0,25	0,28	0,33
G	0,39	0,68	0,34	0	0,16	0,36	0,58	0,72
F	0,34	0,23	0,21	0,16	0	0,33	0,31	0,04
P	0,39	0,41	0,25	0,36	0,33	0	0,35	0,33
SN	0,41	0,64	0,28	0,58	0,31	0,35	0	0,43
WE	0,32	0,63	0,33	0,72	0,04	0,33	0,43	0

- Low: $\nu(i,j) < .5$ & $\nu(j,i) < .5$, which means that two categories do not correlate with each other.

As we explained in the previous section, according to the category usage preference of mobile devices, three category groups are defined: AP, HP and LP. Figure 5 depicts the correlations among the different categories $[i,j]$. We further marked the plot with different shape and color according to the combination of $[i,j]$. For instance, if category i belongs to AP and j belongs to HP, their correlation plot is marked as A-H with a black filled-in circle. The results shown in this figure can be interpreted from the following aspects. First, the categories from the LP group have low correlation with all other categories (as indicated by the triangle samples) and vice versa (as indicated by all markers with white filling color), which means that, *the categories from the LP group do not have high temporal dependency with any other service*. Second, the correlations for the most categories within HP are in the bidirectional high region, which means that, *a category from the HP group has a high probability to be used at the same time with other HP categories*. Third, the concurrent effect is not high among the AP categories as indicated by the circle samples with red filling.

C. Request Load Temporal Similarity

The request load temporal similarity is defined to study the application category usage trend, i.e., the percentage of increase or decrease in terms of request load over time. We aggregate the request load over all devices for each time slot and use the Exponential Smoothing Function to smooth the obtained time series data as $\widehat{l}^i(t-1) = \alpha \sum_x l_x^i(t-1) + (1-\alpha)\widehat{l}_x^i(t-1)$, where α is the smoothing factor which is used to remove the impact of irregular data samples, and the initial value is $\widehat{l}^i(1) = \sum_x l_x^i(1)$. In the next step, we normalize the time series data with the maximum value within the samples as $\widetilde{l}^i(t) = \widehat{l}^i(t)/\max(\widehat{l}^i)$. We obtain the Euclidian distance between any two categories i and j as $d_{ij} = \left(\sum_{t=1}^{n_t}(\widetilde{l}^i(t) - \widetilde{l}^j(t))^2\right)^{\frac{1}{2}}$, where n_t is the total number of time slots. The temporal similarity between categories i and j is defined as $s_{ij} = 1 - \frac{d_{ij}}{\max d_{ij}}$, where $\max d_{ij}$ is the maximum distance for all the values of d_{ij}. Table II lists the similarity values among different categories with featuring priority higher than 3, $f_p > 3$. Similarity values higher than 0.5 are highlighted in the table. It is observed that Communication has a high temporal similarity with Gaming, Social Network has a high similarity with Communication, but Gaming and WE have the highest similarity with each other.

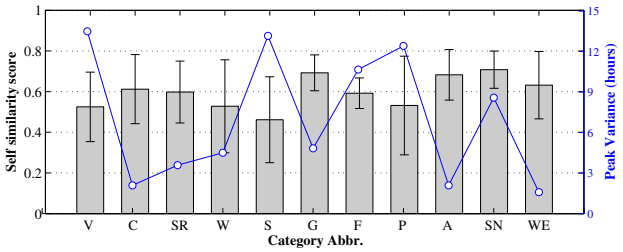

Fig. 6. Application self-similarity

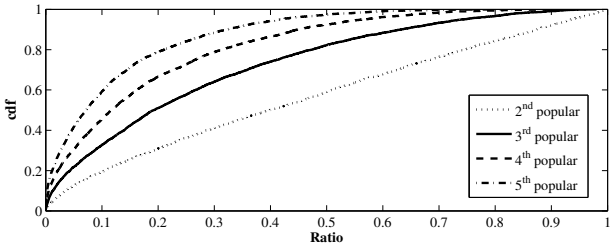

Fig. 7. Application popularity

Hence, *similarity among categories does not always have a bidirectional relationship*. On the other hand, for categories V, S, F, P, there does not exist a high temporal similarity pattern with all the other categories.

D. Request Load Temporal Self-similarity

Temporal self-similarity is used to study the stability of the daily request load from a certain category. For category i, its temporal self-similarity of request load between day u and day w is measured by $s^i_{wu} = 1 - \frac{d^i_{wu}}{\max d^i_{wu}}$, where d^i_{wu} is given by $d^i_{wu} = \left(\sum_{t=1}^{\frac{n_t}{n_m}} (\widetilde{l}^i_w(t) - \widetilde{l}^i_u(t))^2\right)^{\frac{1}{2}}$, and $\max d^i_{wu}$ is the maximum for all possible values of d^i_{wu}, $\forall u, w \in [1, n_m]$. We plot the mean and standard deviation of the self-similarity score as the bar chart in Figure 6, which indicates how similar a category is compared to itself in terms of temporal traffic load changes. Moreover, we also record the peak time for the number of requests for each day and depict the peak variance range (unit: hour) among the measuring days as shown in the right axis in Figure 6. This value indicates the stability of the traffic pattern from a certain category. We could observe from the figure that *Gaming and Social Network have the highest self-temporal similarity among others. Communication, Advertisements and Working/Education have small peak time variances and correspondingly high self-similarities*, which indicate that, they are stable categories in terms of the temporal traffic load feature. In contrast, the top three categories with high peak time variance are Video, Shopping, and Adult/Porn. For these three categories, they also have correspondingly low similarity scores and high standard deviations, which indicate unstable traffic load patterns.

V. Device Feature Extraction

A. Methodology

In order to extract application category depended features for mobile devices, we first need to understand how to represent these features. For each device, we sort its application

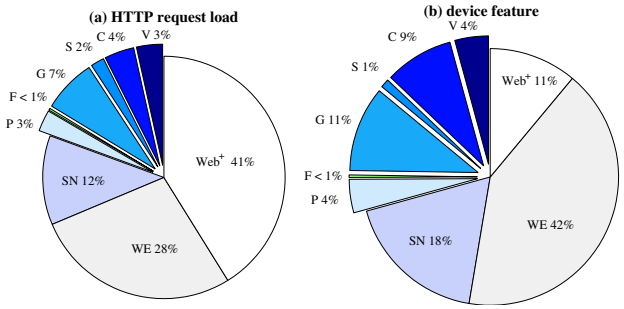

Fig. 8. Overall category popularity analysis

categories in decreasing order according to the total amount of generated requests from each category and normalize the obtained array with the maximum amount of the request load as $\psi = [p_{j_1}, p_{j_2}, ..., p_{j_{11}}]$, where $p_{j_i} = \frac{n_{j_i}}{n_{j_1}}$ and n_{j_i} is the total request load from category j_i. Hence, j_1 is considered as the most popular category for this device with $p_{j_1} = 1$. Category NULL is not considered as a feature for mobile devices, so the total amount of categories is 11. The element p_{j_i} represents the request ratio between the i^{th} category and the first category in ψ when $i \neq 1$. If the value of p_{j_i} is close to 1, it means that the i^{th} category in ψ generates a similar amount of requests compared with the most popular category. As we discussed in section III-C, the top 5000 devices generate most of the HTTP requests and also cover a wide range of usage frequency. We focus on these devices to investigate application category usage features. We plot the cdf of p_{j_2}, p_{j_3}, p_{j_4} of the top 5000 devices in Figure 7. As shown in the figure, the cdf of the second popular category is close to linear. The ratios from the third, fourth and fifth popular categories tend to have a value smaller than 0.5. For instance, there are only around 28% devices that have a ratio from the third popular category which is higher than 0.5. This observation means that *for a single mobile device, it is not common to use more than two dominant categories*. Therefore, we select the first two popular categories $[j_1, j_2]$ of each device to represent its feature, which are called as featuring categories.

The device feature is extracted based on the following rules. All the categories with FP value 3 have lowest priority to feature mobile device, hence they are combined together and called as Web$^+$. If one of the featuring categories has $f_p = 3$, then the category with $f_p < 3$ is used to feature this device. If both of the featuring categories have $f_p = 3$, this device is featured as Web$^+$ device. If $f_p(j_1)$ is 1, the device is featured by j_1. If $f_p(j_1) = 2$ and $f_p(j_2) = 1$ with $p_{j_2} > p_{th}$, then j_2 is used to feature the device, otherwise, j_1 is used. The reason to define a threshold p_{th} here is because the devices using categories with $f_p = 1$ may not generate as many requests as WE and Web$^+$, but the impact from such requests are much higher than the requests from WE and Web$^+$. For instance, requests from Video or File Sharing result in high traffic load, whereas requests from Shopping or Gaming generate commercial benefits for operators, etc. For the featuring categories $[j_1, j_2]$, we accumulate the request load over the top devices for category j_1 and j_2, which are denoted as R_{j_1} and R_{j_2}. We have $p_{th} = R_{j_2}/R_{j_1}$. The request

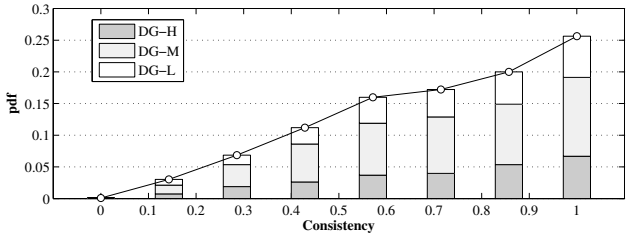

Fig. 9. Category popularity consistency

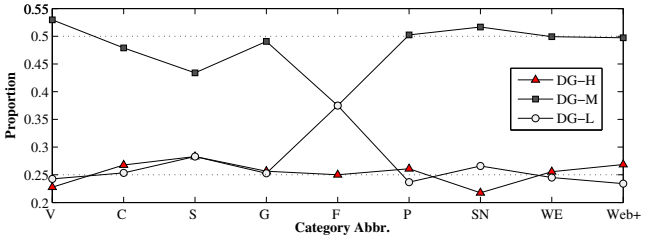

Fig. 10. Overall category popularity within different device group

load proportions of each category over the top devices are shown in Figure 8 (a), in which, except for the WE and Web+, SN generates most of the requests among the top devices.

B. Device Feature Distribution

The percentages of mobile devices featured by different categories are shown in Figure 8 (b). The major observations are as follows. Among the top devices, 40% are featured by WE, which contains the services related to professional activities or education. This is a very surprising result, because it indicates that mobile devices provide more convenient access to assist professional activities related information searching or knowledge update. There are 20% devices featured by SN. This observation should be understood from two aspects. First, it is obvious evidence that *mobile devices provide a convenient platform for social networks*. Second, due to the embedded message push notification features of social networks, end users prefer to use social networks as another platform to exchange messages instead of using paid Telco service like SMS. After SN, there are 11% of devices featured by Gaming. There are 8.5% mobile devices mainly used for Communication, which refers to IP services like Skype, etc. *Even though video has become the main stream of Internet usage, the percentage of devices featured by category Video is only* 4%. However, we could foresee a dramatic growth of this number when the bandwidth and link quality of the cellular networks is increased once 4G or 5G cellular networks are widely deployed. However, it is also surprising to observe that the devices featured by Video and Adult/Porn have similar proportion. There are 1% devices featured by Shopping. According to our collected data, there are less than 1% of users featured by File Sharing, which might be due to the limited bandwidth of current mobile networks as well as volume based charging plans.

C. Device Feature Consistency

The featuring category of each mobile device depends on its user's Internet content preference, usage behavior, and also their personal life pattern. To understand how consistent our defined device feature is over a certain period of time, we extract the device features for each day and calculate the percentage of days which have the same device features based on the collected data of the entire measuring week. The result is shown in Figure 9. It indicates that 78.84% of the devices have a consistency higher than 50%, within which 25.64% of the devices have 100% consistency. This proves a high consistency of the application category defined by the device features. The reason for device feature inconsistency is mainly due to spontaneous used categories, which generate a bursty traffic with high request load, but are not frequently used categories. More investigations about device feature consistency will be addressed in our future work, requireing a longer data collecting period.

We break down the bar plots using the device groups defined in section III: DG-H, DG-M, and DG-L. For each consistency value, we indicate the proportion of devices from each device group as shown in Figure 8 (b). It is very close to the proportion used to define the device groups for each consistency value, which indicates that *the device group feature defined by the amount of data samples does not influence the temporal consistency of the device feature defined by categories*.

D. Device Group Influence

To further understand the relationship between of the device groups and the category-defined device features, we plot the percentages of devices from DG-H, DG-M and DG-L for the devices which are featured by the listed categories as shown in Figure 10. There are two dashed lines plotted in the figure indicating the percentage for the high (25%), medium (50%) and low usage (25%) devices. If the device group feature does not have any influence on the category-defined device features, all plots should overlap with the two dashed lines. However as shown in the figure, DG-M has a higher percentage than the expected value ($52.97\% > 50\%$) for the Video category, which means that the devices from DG-M show higher interests on video services. DG-H has higher interests on Communication category compared to DG-M. It is very interesting to observe that for the Shopping category, both DG-H and DG-L have higher interests than devices in DG-M. It is surprisingly to observe that devices from DG-L show a very high interest on File Sharing category, which has the same percentage with devices from DG-M. It is also unexpected to observe that, DG-M and DG-L both have a percentage higher than the expected proportion for the Social Network category, but DG-H is lower than the expectation. For the categories Gaming and Working/Education, the amount of devices from each DG overlaps with the dashed lines, which means that these device groups do not show a certain preference.

VI. Operating System Characteristics

A. OS Popularity

In this section, we investigate the operating system (OS) features of the mobile devices, and examine the correlations

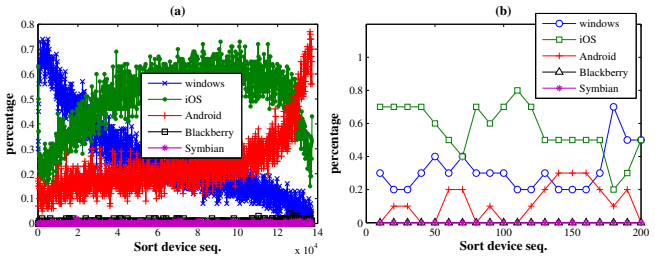

Fig. 11. Operating system characteristics

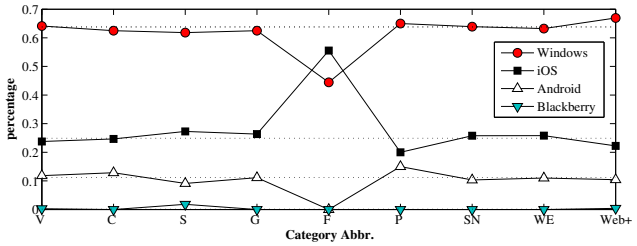

Fig. 12. Operating system influence on category-defined device features

between OS features and application categories. Based on the methodology presented in section II-D, we first investigate the OS popularity among all the mobile devices in the dataset that can be extracted with OS information. Within all 137242 devices, we detect 50.53% iOS, 25.6% Windows, 23.53% Android, 0.32% Blackberry and 0.3% Symbian devices. All the devices are sorted according to their request load in descending order. We segment every 100 devices and calculate the percentage of each OS within these 100 devices. The results are shown in Figure 11. Subplot (a) is the OS distribution among all 137242 devices. The percentage of Windows devices is reduced with a decreasing amount of request load, and the trend for Android devices is just the opposite of the Windows devices, whereas iOS dominates the devices in the middle range of the request load. However, if we zoom in to the top 200 devices in subplot (b), the iOS devices have a higher percentage than the Windows and Android devices, which indicates that *extremely heavy users prefer to use iOS; Windows devices generate most of the requests; compared to iOS and Windows devices, Android devices generate correspondingly low request loads; Blackberry and Symbian devices do not show any obvious impact within the entire range.*

B. OS Influence on Category-defined Device Feature

Based on the above results, we group the devices together that have the same category-defined features. For each category feature, we further track back the OS information of the devices. The results are shown in Figure 12. Within the top 5000 devices, devices that are estimated to use Windows system have the highest share with 63.79%, iOS and Android devices have a share of 24.89% and 11.19%, respectively. Moreover, there are also 0.12% Blackberry devices. Compared to the proportion in all 137242 devices, Windows devices have much higher share among the top users. We plot these four proportions using dashed lines as guidelines. As depicted in the figure, most of the data points are very close to the dashed lines. This indicates that, in general, the OS does not play an important role in terms of category usage preferences. However, the category File Sharing has an exceptional behavior: the iOS devices show more interest in FS than expected, whereas the Windows devices have less interest than their expectation and Android devices show very low interest in FS. Moreover, the Windows devices have slightly higher proportion on Web$^+$, which is the opposite for iOS devices. Android devices have higher proportion on Adult/Porn activities and iOS devices are just on the opposite.

VII. Related Work

1) Data Analytics for Cellular Networks: Data analytic has been widely used for cellular network traffic analysis and management. Empirical studies based on measurements of commercial cellular network, for instance [2], [3], have been used to investigate in terms of data throughput, latency, etc. The dataset reported in [4] is mainly used for modeling call arrival and duration. The inter-call duration is studied in [5] and the authors confirmed a power law distribution with an exponential cutoff at the population level. Research work in this field also covers mobile user profiling via human dynamic features extracted from cellular networks, for instance, user mobility modeling, handover, trajectory analysis, etc. Interested readers could for instance refer to [6]. The human mobility pattern is especially investigated in [7] and [8] based on voice call records. Moreover in [11], the authors draw relations between traffic dynamics to the mobility pattern and activity level for subscribers. The traffic dynamic is also linked to device types, which is identified from the type allocation code [12] and running applications [22]. In [9], the authors use a large scale dataset collected from a UMTS network for data analytics, but only heavy user behavior is revealed. Zhang and Arvidsson reported findings based on a dataset collected on a Gn interface between a GGSN and SGSN in a cellular network in [10]. The major focus of their work is to compare the traffic pattern difference between wireless and wireline communication. In [25], the authors analyzed the traffic at the radio network controller level. However, their work does not analyze at the content and user behavior level.

2) Web Service Usage Behavior: There are two major research directions about web service usage behavior. First is the web service usage behavior based on wired networks [23], [24] and second is the web service usage based on wireless networks. The authors in [23] analyzed user browsing sessions based on a one-week long dataset collected from the Yahoo! toolbar to discover topical and temporal patterns of user's browsing behavior. Research work like [24] investigated web browsing behavior in the context of demographic distribution, which includes age, sex, race, education, and income.

There is a number of research works focusing on characterizing mobile traffic pattern and usage behavior in cellular networks [21], [18], [19]. However, these works are limited by either a small scale data collecting period or a limited amount of devices. Plissonneau and Vu-Brugier [21] analyzed mobile Internet usage based on data collected from real field, but their data collection period is only 1 hour 39 mins, only

providing a limit view of the mobile device usage within a certain time period of a day. Moreover, their study is mainly focused on video usage behavior. In [18], the authors reported a method for profiling users based on their browsing behavior, and their dataset is based on one day collected on April 16, 2008. The characteristics of smart phone usage are reported in [19], however, their results are obtained only based on 255 users. Another research methodology for investigation of 3G mobile Internet usage behavior is based on survey or interview with mobile Internet users, e.g. [20], however this method is limited by the detailed usage information and the amount of interviewed users.

VIII. Discussion and Summary

In this paper, we conducted an in-depth analysis of mobile Internet service usage in a large scale mobile cellular network. There are several important observations in terms of traffic dynamics, application category usage and the operating system of mobile devices, which are summarized as follows:

- For the mobile Internet traffic dynamics, we identified the asymmetric relationship between the HTTP request load and mobile Internet usage frequency. High frequency of mobile Internet usage normally generates high HTTP request load, however, it is not true vice versa.
- Mobile users tend to use mobile Internet in certain time periods of a day, which means, the time points for mobile Internet access have a cluster effect. The distribution of the number of generated clusters within a day from different mobile devices tends to be the same, but the cluster coverage time tends to be bigger if a device generates more request load.
- The majority of mobile users does not have more than two preferred application categories, which generate most of the HTTP requests in their daily usage.
- The category-defined device feature has a high consistency over time and it is not influenced by the Internet usage frequency of mobile devices. However, the Internet usage frequency indicates an influence on the mobile devices featured by the same category. For instance, heavy usage users indicate more interest in communication and shopping, medium usage users indicate more interest in video and social networks.
- Users preferring Windows devices normally generate high HTTP request load, and users preferring Android devices generate correspondingly low requests. In comparison, users preferring iOS devices cover the middle range. However, most of the users that generate extremely high request load are iOS users.

These results reveal valuable insides in the characteristics of mobile data in cellular networks, which can provide Telco operators meaningful advices in terms of network resource management and planning, smart service bundling and provision, etc. This work is our first attempt to analyze this large scale dataset. Additional dimensions of observations, which can lead to further insights, will be considered as our next step.

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017", White Paper, Feb. 6, 2013.

[2] K. Pentikousis, M. Palola, M. Jurvansuu, and P. Perl, "Active goodput measurements from a public 3G/UMTS network", IEEE Communications Letters, vol. 9, pp. 802-804, 2005.

[3] W. Tan, F. Lam and W. Lau, "An Empirical Study on the Capacity and Performance of 3G Networks", IEEE Transactions on Mobile Computing, vol. 7, no. 6, pp. 737-750, June 2008.

[4] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary Users in Cellular Networks: A Large-Scale Measurement Study", in Proc. DySPAN, 2008.

[5] Z. Jiang, W. Xie, M. Li, B. Podobnik, W. Zhou, and H. Stanley, "Calling patterns in human communication dynamics", Proc. Natl. Acad. Sci. U.S.A. 110, 2013.

[6] P. Pathirana, A. Savkin, and S. Jha "Mobility modelling and trajectory prediction for cellular networks with mobile base stations", ACM MobiHoc 2003.

[7] J. Candia, M. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A. Barabasi, "Uncovering individual and collective human dynamics from mobile phone records," Journal of Physics A: Mathematical and Theoretical, vol. 41, pp. 1-11, 2008.

[8] M. Gonzalez, C. Hidalgo, and A. Barabasi, "Understanding individual human mobility patterns," Nature, vol. 453, pp. 779-782, 2008.

[9] Y. Jin, N. Duffield, A. Gerber, P. Haffner, W. Hsu, G. Jacobson, S. Sen, S. Venkataraman and Z. Zhang, "Characterizing Data Usage Patterns in a Large Cellular Network", In Proc. of the ACM SIGCOMM workshop on CellNet, 2012, pp. 7-12.

[10] Y. Zhang and A. Arvidsson, "Understanding the Characteristics of Cellular Data Traffic", In Proc. of the ACM SIGCOMM workshop on CellNet, 2012, pp. 13-18.

[11] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding traffic dynamics in cellular data networks", IEEE Infocom, 2011.

[12] M. Shafiq, L. Ji, A. Liu and J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices", ACM SIGMETRICS 2011, pp. 305-316.

[13] "Final NetInf Architecture", FP7-ICT-2009-5-257448-SAILD.B.3, SAIL - Scalable and Adaptable Internet Solutions, 2013.

[14] Cisco Systems, "Cisco IronPort AsyncOS 7.1 for Web-User Guide", 2010.

[15] M. Boulmalf, J. Abrach, H. Harroud, and T. Aouam, "Traffic Analysis for GSM Networks", International Conference on Computer Systems and Applications IEEE/ACS AICCSA 2009.

[16] N. Janssens, X. An, K. Daenen, C. Forlivesi, "Dynamic scaling of call-stateful SIP services in the cloud", IFIP Networking 2012, page 175-189.

[17] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", second International Conf. on Knowledge Discovery and Data Mining, 1996.

[18] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling Users in a 3G Network Using Hourglass Co-Clustering", ACM MobiCom, 2010.

[19] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, "Diversity in smartphone usage", MobiSys, 2010.

[20] C. Taylor, et al. "A Framework for Understanding Mobile Internet Motivations and Behaviors", CHI'08 Extended Abstracts on Human Factors in Computing Systems, pp. 2679-2684.

[21] L. Plissonneau and G. Vu-Brugier, "Mobile Data Traffic Analysis: How do you Prefer Watching Videos", International Teletraffic Congress (ITC), 2010.

[22] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang and S. Venkataraman, "Identifying Diverse Usage Behaviors of Smartphone Apps", in Proc. of the ACM SIGCOMM IMC 2011.

[23] R. Kumar and A. Tomkins, "A Characterization of Online Browsing Behavior", 19th International Conference on World Wide Web (WWW), pp. 561-570, 2010.

[24] S. Goel, J. M. Hofman, and M. I. Sirer, "Who does what on the web: A large-scale study of browsing behavior," Sixth International AAAI Conference on Weblogs and Social Media, 2010.

[25] Y. Chen, et al."Understanding the Complexity of 3G UMTS Network Performance-Modeling,Prediction,and Diagnosis", IFIP Networking 2013.