Characterizing Data Usage Patterns in a Large Cellular Network

Yu Jin, Nick Duffield, Alexandre Gerber, Patrick Haffner, Wen-Ling Hsu,
Guy Jacobson, Subhabrata Sen, Shobha Venkataraman, Zhi-Li Zhang*
*Dept. of Computer Science, University of Minnesota, MN, USA AT&T Labs – Research,NJ, USA
{yjin,duffield,gerber,haffner,hsu,guy,sen,shvenk}@research.att.com,
*zhzhang@cs.umn.edu

ABSTRACT

Using heterogeneous data sources collected from one of the largest 3G cellular networks in the US over three months, in this paper we investigate the usage patterns of mobile data users. We observe that data usage across mobile users are highly uneven. Most of the users access data services occasionally, while a small number of heavy users contribute to a majority of data usage in the network. We apply statistical tools, such as Markov model and tri-nonnegative matrix factorization, to characterize data users. We find that the intensive usage from heavy users can be attributed to a small number of applications, mostly video/audio streaming, data-intensive mobile apps, and popular social media sites. Our analysis provides a fine-grained categorization of data users based on their usage patterns and sheds light on the potential impact of different users on the cellular data network.

Categories and Subject Descriptors

C.2.3 [**COMPUTER-COMMUNICATION NETWORKS**]: Network Operations

Keywords

Cellular Networks, Usage Characterization, Heavy users

1. INTRODUCTION

The increasing popularity of mobile devices (such as smartphones, tablet computers, and e-readers), coupled with established mobile applications (such as web browsing and streaming media) and emerging services (e.g. those based on location-awareness) catering to these devices, has led to a pronounced recent increase in *mobile cellular data traffic*. Indeed, the volume of mobile data traffic is projected to imminently surpass that of mobile voice traffic [1]. This rapid growth in mobile data requires mobile service providers to efficiently manage their radio spectrum resources and evolve their infrastructure so as to meet the growing demands and diverse expectations of mobile data users.

As a critical step to addressing this challenge, it is imperative to understand how mobile data users access cellular data services and what their demands are on different network resources. Building a detailed understanding of user behavior – in terms of the many factors that may influence or induce differing mobile user behaviors (e.g., apps and mobile devices) – is crucial to many cellular network management tasks, such as providing insights for capacity planning and allocating appropriate network resources to cope with the growing demands of mobile data users.

To this end, in this paper we take the first step to study mobile user behaviors with massive amounts of data collected from various locations in one of the largest cellular networks in the US from August 2010 to October 2010. Such data sources present a comprehensive view of the usage patterns of mobile data users in today's large cellular networks. We propose novel and scalable statistical tools for analyzing and mining such datasets. In particular, we characterize temporal data usage patterns with a 2-state Markov model, and discover two distinct user groups, *heavy users* and *normal users*, with distinct usage patterns and service and device preferences. To understand the intensive and continuous data usage from heavy users, we propose a scalable sampling based tri-nonnegative matrix factorization (tNMF) algorithm, which can extract dominant network activities (defined as combinations of applications and content providers) and thereby categorize different types of heavy users. Our main findings are below:

1) A few heavy users (top 3% in our study) contribute to nearly half of all the mobile data traffic, and exhibit very distinctive usage patterns in comparison to the rest of normal users. We identify that a majority of the heavy users display *continuous* and *intensive* usage patterns; while the normal users access data services mostly in a bursty or intermittent way. Such a difference in usage patterns is likely due to the general preference for different devices (e.g., smartphones, laptops with 3G mobile data cards) and mobile applications (e.g., video streaming) by heavy-data users.

2) A small number of dominant network activities define the heavy users. That is, even though heavy users are involved in millions of diverse network activities, the bulk of their data usage comes from a surprisingly small number of network activities, mainly associated with mobile video/audio sites, social networking sites and mobile apps. The high concentration of data usage on a few network activities enables a precise categorization of these heavy users.

Related work: There have been research to characterize data usage activities in cellular networks, mostly at a small scale, e.g., a few testing mobile devices. For example, [2] characterized the diversity of users' activities by tracking 255 smartphones. [3] characterized the relationship between users' application interests and their mobility properties. [4] applied a passive measurement of mobile devices and showed that mobile traffic is dominated by multimedia content and mobile downloading applications. The most related work to our paper is [5], where the authors studied mobile data us-

CellNet'12, August 13, 2012, Helsinki, Finland.
Copyright 2012 ACM 978-1-4503-1475-6/12/08 ...$15.00.

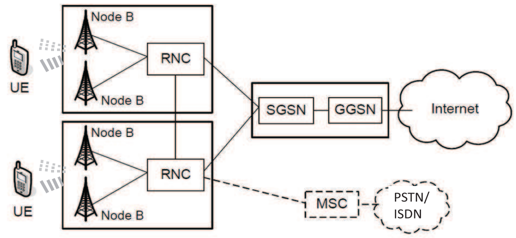


Figure 1: UMTS network architecture.

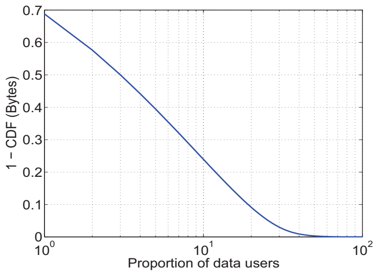


Figure 2: Proportion of bytes from the bottom $1 - x$% users

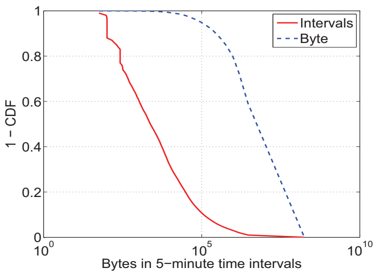


Figure 3: Distribution of bytes per 5-minute time intervals

age patterns and observed a similar heavy-user phenomenon in a cellular network dataset from 2007. Our work differs from existing works, especially [5] in that we proposes many novel statistical and data mining techniques carry out fine-grained characterization of the activities of smartphone data users. Our analysis also pinpoints the factors that lead to the heavy user phenomenon and reveals special properties of these users, which can potentially guide resource allocation in cellular networks to meet the demands from such users.

The remainder of this paper is organized as follows: Section 2 introduces the architecture of the cellular network and various data sources. In Section 3, we characterize different users' activities and propose a co-clustering based method to further analyze and categorize heavy users in Section 4. Section 5 concludes the paper.

2. BACKGROUND

Cellular Network Overview. The cellular network under study uses primarily UMTS (Universal Mobile Telecommunication System), whose key components are illustrated in Fig. 1. When making a voice call or accessing a data service, a mobile device directly communicates with a cell tower (also referred to as a node-B), which forwards the voice/data traffic to a Radio Network Controller (RNC). In case of mobile voice, the RNC delivers the voice traffic toward the PSTN or ISDN (Public Switched Telephone Network) or ISDN (Integrated Services Digital Network) telephone network, through a Mobile Switching Center (MSC) server. In case of mobile data, the RNC delivers the data service request to a Serving GPRS Support Node (SGSN), which establishes a tunnel with a Gateway GPRS Support Node (GGSN) using GPRS Tunneling Protocol (GTP)[1], through which the data enters the IP network (and the public Internet). The UMTS network has a hierarchical structure: where each RNC controls and communicates with multiple node-Bs, and one SGSN serves multiple RNCs.

Datasets. We collect data from a large UMTS network for a 3-month time period. These datasets include: (1) *Byte usage*, which contain the total number of bytes uploaded and downloaded by individual users in each day; and (2) *Application mix* which contain the total number of bytes uploaded and downloaded by individual users associated with each of the 14,000 applications. Traffic classification is conducted by matching TCP/IP headers and application headers against manual classification rules (See [6] for details). We also identify content providers from the top level domain names of application servers.

Privacy Measures: Given the sensitivity of the data, we took several steps to ensure the privacy of individuals represented in our datasets. First, only anonymous records were used in this study. In particular, customer anonymity was preserved during this study by hashing each customer phone number to a unique anonymous identifier prior to joining these datasets. Second, all our results are presented as aggregates to protect the privacy of individuals. No individual anonymous identifier was singled out for the study.

3. OVERVIEW OF DATA USERS

In this section, we overview the data usage patterns of mobile users, from which we identify two groups of users, *heavy users* and *normal users*, who are characterized with distinct usage patterns and device and network service preferences.

3.1 Byte Consumption

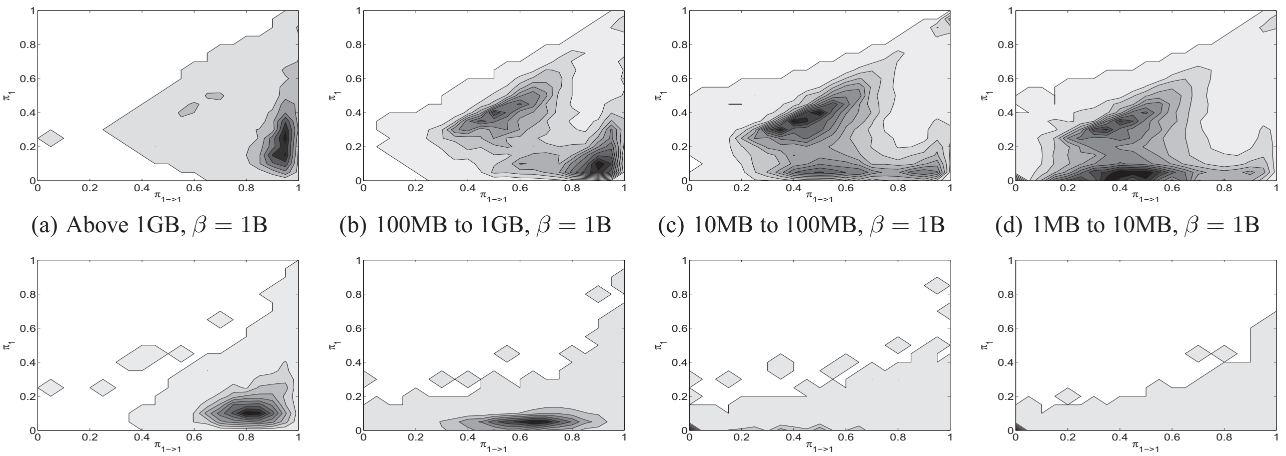
We start with the total byte usage, which is an indicator of the intensity of a user's data usage. Fig. 2 shows the CCDF of the total byte usage in two weeks, where the x-axis (in log scale) represents the percentage of the top users (ranked by their total byte usage), and the y-axis shows the total amount of bytes used by the *bottom* 1-x% of users. We observe that the total byte distribution is highly uneven, with a significant portion of bytes contributed by a relatively small number of intensive (heavy) users. For example, the top 1% of users consume more than 30% of the total bytes, and half of the bytes are contributed by the top 3% of users[2].

3.2 Data Usage Patterns

We next add a time dimension to study different data usage patterns. Do users tend to continuously access data services over time, or do they tend to have bursty usage? We first formally define usage patterns. We define the *usage pattern* of a user u as a time-series $X_u := \{x_{u,t}\}, 1 \leq t \leq T$, where $x_{u,t}$ represents the total amount of bytes from and towards u at t (In our study, t is bucketed into 5-minute time intervals, and T equals 864, the number of 5-minute time intervals in 3 consecutive days). For ease of analysis, we convert the time-series X_u into a binary sequence, $B_u := \{b_{u,t}\}$, where $b_{u,t} = 1$ if $x_{u,t} > \beta$ and $b_{u,t} = 0$ otherwise. Based on this formulation, the usage pattern of a data user can be represented by a two-state Markov model, where $b_{u,t} = 1$ represents the active state (with usage above β) and $b_{u,t} = 0$ stands for the dormant state (with usage no more than β). We use the following metrics to characterize B_u: (1) active state probability $\pi_1 := \sum_t I(b_{u,t} = 1)/T$, where I is the indicator function, and,

[1] GPRS stands for the General Packet Radio Service.

[2] A similar observation is made in [5], where the observed distribution is even more skewed, with the top 1% users accounting for 60% of the data usage. We suspect this is primarily because the data used in [5] is collected in 2007, when significantly fewer users had smartphone devices and access to 3G networks, and data-intensive applications like HD video/audio streaming, etc., were substantially less popular. As a consequence most users were incapable of heavy data usage, and hence the heavy users appear to consume a far greater fraction of the data of the network.



(a) Above 1GB, $\beta = 1$B (b) 100MB to 1GB, $\beta = 1$B (c) 10MB to 100MB, $\beta = 1$B (d) 1MB to 10MB, $\beta = 1$B

(e) Above 1GB, $\beta = 1$MB (f) 100MB to 1GB, $\beta = 1$MB (g) 10MB to 100MB, $\beta = 1$MB (h) 1MB to 10MB, $\beta = 1$MB

Figure 4: Data usage patterns: a larger π_1 implies longer usage time and a larger $\pi_{1\to 1}$ indicates more persistent usage.

similarly, dormant state probability $\pi_0 = 1 - \pi_1$; (2) active state transition probability $\pi_{1\to 1} := P(b_{u,t} = 1 | b_{u,t-1} = 1)$, where $1 < t \leq T$. A larger π_1 implies that a user accesses data actively more frequently, while a larger $\pi_{1\to 1}$ indicates that a user tends to remain in the active state, i.e., with contiguous data usage of the intensity defined by β.

We categorize users into 4 groups according to their (expected) total usage in a month: above 1GB, between 100MB and 1GB, between 10MB and 100MB, between 1MB and 10MB. The users with less than 1MB usage are removed from our analysis. Fig. 4 illustrates the usage patterns of the four groups of users in terms of the joint distribution of π_1 and $\pi_{1\to 1}$, where the darker/lighter colors represents a higher/lower density of users. We first choose $\beta = 1$ Byte (Fig. 4[a-d]), i.e., a user is considered to be active at a particular 5-minute time slot as long as traffic is observed from/towards that user in that time slot. Most users with more than 1GB usage (Fig. 4[a]) tend to use data services continuously (thereby forming a single cluster, with $\pi_1 \in [0.1, 0.7]$ and $\pi_{1\to 1} \approx 1$). In comparison, the usage patterns of data users with a usage between 100MB and 1GB appear bimodal (Fig. 4[b]), where a group of users show intermittent usage patterns (the cluster in the middle) and another group of users have continuous data usage (the cluster on bottom-right corner). However, compared to the cluster in Fig. 4[a], such continuous data usage is generally less frequent (i.e., $\pi_1 < 0.1$). As we further decrease the total usage (Fig. 4[c][d]), the cluster at the bottom-right corner disappears and a new cluster (with $\pi_1 < 0.1$ and $\pi_{1\to 1} \approx 0.5$) shows up. This new cluster represents the casual data users who occasionally access data service.

By increasing the threshold β, we can focus on time intervals with more intensive data usage. Fig. 3 shows the CCDF of bytes in 5-minute time intervals (note the x-axis is in log scale), where the solid curve represents the proportion of 5-minute intervals with bytes above x and the dotted curve stands for the proportion of total bytes contributed by these intervals. We choose $\beta = 1$ MB, which accounts for less than 5% of the 5-minute intervals. These intervals, however, together cover more than 90% of the total bytes.

Fig. 4[e-h] illustrates the intensive data usage patterns for the 4 user groups. We observe in Fig. 4[e] that users with more than 1GB usage also tend to have contiguous and intensive data usage, thereby forming a cluster with $\pi_1 \approx 0.1$ and $\pi_{1\to 1} \approx 0.8$. In comparison, the cluster moves to the middle for the users with usage between 100MB and 1GB (Fig. 4[f]), indicating that these users are more likely to have a burst of data-intensive activities, i.e., use data intensively for a short time period. Almost no data-intensive activities are observed for the remaining two user groups (Fig. 4[g][h]).

Table 1: Pct. of heavy data users in each device class

	Heavy user	Description
COMPUTER	14.61%	Laptop cards and netbooks
SMARTPHONE	5.78%	Phones with a data plan, e.g., Blackberry, Android phones, etc.
MODEM	0.46%	3G modems
PHONE	0.40%	Phones without a data plan
OTHER	0.19%	Security alarms, ebooks, terminals, electricity/water meters, etc.
VEHICLE	0.02%	GPS and vehicle tracking devices

Based on the distinct usage patterns observed for the different data users above, we use *1GB (expected) data usage per month* as the cut-off threshold to partition users into two groups: users with more than 1GB usage per month are referred to as *heavy users* and the remaining users are called *normal users*. Fig. 4 shows that heavy users tend to access data services in a continuous and intensive manner, while normal users are mostly bursty or intermittent in their access.

3.3 Device and Application Preference

In this section, we compare the difference between heavy users and normal users in terms of their favorite devices and applications. The comparison result explains the key factor that attributes to the large data consumption of heavy users.

Device: We identify the model of a device based on the first 8-digit Type Allocation Code (TAC) in the corresponding International Mobile Equipment Identity (IMEI) associated with the device. The remaining 6-digit serial number has been anonymized to protect users' privacy. Based on the functionality of mobile devices, we further categorize them into 6 classes (1st column in Table 1). We show the proportion of heavy users out of all data users with each particular device class in the second column of Table 1. The description of each device class is in the 3rd column.Not surprisingly, COMPUTER related devices (laptop cards and netbooks) and SMARTPHONE users, due to their capability of running many applications, have a much higher chance of becoming heavy users. In fact, even though both types of users only account for less than half of all the users, they together contribute to more than 90% of all heavy users. PHONE users, due their large population size, contribute to a majority of the remaining heavy users. Only a few heavy users are observed from VEHICLE and OTHER. we hence omit these two categorizes from our analysis in the following.

Application usage: In order to understand the difference between heavy users and normal users in terms of their dominant application usage, we employ a rule-based classification method to break down the traffic into a number of predefined *services*, e.g., `Email`, which includes *applications* like POP, IMAP, SMTP etc, and `Video`, which includes RTSP, FLV, QUICKTIME and other types of video applications (see [6] for the definitions of these services).

To characterize the dominant services accessed by different types of users, we define two metrics: the *service popularity* (Pop) and the *byte dominance* (Dom). Given an observation period T, let b_s^u be the total amount of bytes contributed by service s from user u, where $s \in \mathcal{S}$, represents the predefined services. Let $\pi_s^u := b_s^u / \sum_{j \in \mathcal{S}} b_j^u$ be the proportion of bytes associated with the service s. The popularity for service s is defined as $Pop(s) := E[I(\pi_s^u > 0)]_u$, where I is the indicator function and $E[x]_y$ stands for the expected value of x across all possible y's; and the byte dominance of a service s is defined as $Dom(s) := E[\pi_s^u]_u$. The metric $Pop(s)$ represents the likelihood that a user uses the application s during the measurement period T, while $Dom(s)$ indicates the average proportion of bytes that is contributed by the service s. We note that when u is conditioned on holding a certain device (e.g., in Table 2), $Pop(s)$ and $Dom(s)$ are defined specifically for that device.

Table 2 shows the Dom's and Pop's (Pop's are inside the parentheses) across various services and different device classes. We compute Dom's and Pop's for uploading bytes and downloading bytes separately. For simplicity, we only choose the most dominant 7 services. Services like DNS, due to their small traffic volume, are removed from the table.

Our first observation is that heavy users, no matter what kind of device they hold, use a much greater variety of services than the normal users do, i.e., their service Pop's are generally much larger. Nevertheless, despite the heavy users' general preference for a large number of different services, `Video` is far more overwhelmingly used by heavy users. For example, heavy users have 2 to 4 times higher chance of using `Video` than normal users, especially for COMPUTER, PHONE and SMARTPHONE users, the `Video` popularity is around 70% to 80%. In contrast, only 20% to 30% of the normal users use `Video`. Interestingly, we find that many users with large uploading bytes but small downloading bytes are often associated with uploading `Video` traffic from SMARTPHONE or MODEM devices, plausibly because they are using mobile devices as surveillance webcams for remote monitoring.

In contrast to `Video` which is more prevalent among heavy users, `Email` and `Web` are popular for both types of users (with heavy users having slightly higher Pop's)[3]. One thing to note is that a large portion of the bytes (above 80%) contributed by normal users are associated with `Email` and `Web`. This indicates that normal users mainly use their mobile devices for checking emails periodically and surfing the Internet occasionally. This explains the observation of bursty and intermittent usage patterns of normal users.

The `P2P` applications, such as eMule and BitTorrent, are not as dominant for mobile users compared to the users of DSL or other types of networks [6]. This is likely due to the incapability of running `P2P` on most of the mobile platforms. It could also because the customer billing has a tiered dependence on the total usage and hence they prefer to run `P2P` over Wi-Fi networks instead. We do observe some SMARTPHONE and PHONE users with `P2P` activities, presumably caused by users' tethering activities, i.e. when a mobile device is configured to become a wireless access point.

4. CATEGORIZING HEAVY USERS

We have just found in Section 3.3 that heavy users use a wide range of applications, and this naturally leads to the question: are there specific applications that these users use extensively so that they become heavy users? In this section, we propose the notion of *network activity matrix* as a means to characterizing data usage of heavy users, and classify them into clusters according to their dominant network activities.

Characterizing Users using Network Activity Matrix. The total byte usage of a user can be broken down into the byte usage of the various applications run by the user. Moreover, even within a single application, the byte usage depends on the specific content provider that the user accesses. Therefore, we define a *network activity* to be a combination of a specific application and the content provider serving that application. The network activities of all heavy users can be captured by a *network activity matrix*, which we formally define now.

Let $u_i \in \mathcal{U}$ be all the heavy users under study. We employ a rule-based classification method to break down the traffic associated with $\mathcal{U}$ from 08/23/2010 into over 14,000 applications. In addition, we identify content providers by the corresponding top-level domain names of the application servers. Let a_{ij} denote the amount of bytes contributed by the user u_i for participating in activity j. We define the *activity matrix* as $A := \{a_{ij} / \sum_j a_{ij}\}$, $1 \leq i \leq |U|, 1 \leq j \leq N^4$, where a_{ij} represents the proportion of total bytes contributed by u_i participating in activity j. In other words, $A_{i\cdot}$ represents the distribution of bytes corresponding to different types of network activities associated with u_i, which we refer to as the network *activity vector* of u_i.

Given the activity matrix, our objectives are two-fold. First, we want to mine from millions of network activities the ones that are dominant among the heavy data users. Second, we want to categorize heavy users according to these dominant network activities. We note that since the activity matrix A generally has very high dimensionality (recall we have observed millions of heavy users and activities), the first objective is equivalent to a feature selection problem on the columns of A. In comparison, the second objective can be considered as a clustering problem on the rows of A. Instead of solving two objectives independently which often yields suboptimal solutions, we apply a co-clustering algorithm based on the tri-nonnegative matrix factorization (tNMF in short). TNMF groups the rows and columns of the activity matrix simultaneously, and removes (by setting the corresponding coefficients to 0) insignificant activities automatically, which leads to both better heavy user clusters and more interpretable activity groups. We next describe how we apply tNMF to analyze and cluster heavy users.

Analyzing Heavy Users using tNMF Given an activity matrix A, the tNMF algorithm approximately *factorizes* A into three *low-rank nonnegative* matrices, $R_{|U| \times k}$, $H_{k \times l}$, and $C_{N \times l}$ in order to minimize the following objective function J, subject to orthogonality constraints on R and C:

$$\min_{R \geq 0, C \geq 0, H \geq 0, R^T R = I, C^T C = I} J(R, H, C) = ||A - RHC^T||_F^2$$

where $|| \cdot ||_F$ is the Frobenius norm, and $k, l << \min(m, n)$. Due to space limitations, we omit the algorithm for solving the tNMF problem (see [7] for details). After decomposition, the indicator matrix R provides categorization of heavy users. For example, a heavy user u is classified into group p where $p = argmax_i R_{ui}$. A large entry H_{pq} in matrix H indicates that the heavy user group p

[3] We note that `WebMail` is classified as `Web` instead of `Email`.

[4] There are in total over millions (N) of network activities observed in our dataset.

Table 2: Comparing of heavy users and normal users in terms of $Dom(Pop)$'s

		COMPUTER		MODEM		PHONE		SMARTPHONE	
Class	Type	up (%)	down (%)	up (%)	down (%)	up (%)	down (%)	up (%)	down (%)
Apps	heavy	0.2 (4.4)	0.4 (4.4)	0.0 (0.0)	0.0 (0.0)	5.7 (21.5)	5.9 (21.1)	26.5 (92.5)	28.7 (90.9)
	normal	0.1 (1.9)	0.1 (1.9)	0.5 (1.6)	0.8 (1.6)	1.2 (4.5)	1.2 (4.4)	19.4 (81.1)	19.6 (79.2)
Email	heavy	1.4 (12.3)	1.1 (11.5)	0.0 (0.0)	0.0 (0.0)	7.0 (33.4)	6.2 (32.7)	8.4 (80.6)	6.7 (78.6)
	normal	1.5 (9.1)	1.6 (8.6)	1.0 (5.8)	0.8 (5.3)	5.8 (14.4)	6.3 (14.2)	22.8 (80.2)	24.4 (78.7)
IM	heavy	1.1 (42.6)	0.8 (41.1)	0.1 (25.0)	0.0 (18.8)	0.2 (12.2)	0.1 (11.5)	0.0 (1.2)	0.0 (1.1)
	normal	1.2 (26.0)	1.0 (24.8)	0.0 (4.1)	0.0 (3.3)	0.0 (0.6)	0.0 (0.5)	0.0 (0.1)	0.0 (0.1)
P2P	heavy	0.8 (12.0)	0.6 (10.7)	0.0 (0.0)	0.0 (0.0)	0.8 (6.2)	0.6 (5.5)	0.1 (1.5)	0.1 (1.3)
	normal	0.2 (4.8)	0.1 (3.8)	0.0 (2.5)	0.0 (3.3)	0.0 (0.3)	0.0 (0.2)	0.0 (0.4)	0.0 (0.3)
Video	heavy	21.6 (75.7)	36.9 (74.3)	2.0 (37.5)	8.9 (37.5)	29.0 (81.5)	51.8 (80.2)	19.2 (69.7)	35.7 (67.4)
	normal	18.2 (31.3)	23.2 (29.7)	1.9 (9.5)	2.3 (10.3)	7.0 (30.1)	16.8 (28.9)	2.9 (20.9)	6.9 (19.4)
VoIP	heavy	0.5 (16.2)	0.4 (15.4)	0.0 (6.3)	0.0 (0.0)	0.2 (6.5)	0.2 (5.8)	0.1 (2.6)	0.1 (2.5)
	normal	0.1 (7.3)	0.1 (6.7)	0.0 (0.0)	0.0 (0.0)	0.0 (0.3)	0.0 (0.2)	0.0 (0.7)	0.0 (0.6)
Web	heavy	60.1 (98.2)	49.7 (97.1)	25.8 (81.3)	29.2 (50.0)	52.2 (99.5)	32.6 (99.2)	43.4 (99.9)	27.8 (99.8)
	normal	53.3 (93.6)	52.4 (88.7)	10.6 (80.7)	11.7 (33.3)	83.3 (98.9)	73.5 (98.8)	51.2 (98.1)	46.8 (97.8)

favors the activity group q. The column $C_{\cdot q}$ shows the dominance of different activities in the activity group q. Since $C_{\cdot q}$ is sparse, it only records the most dominant activities contributing to group q and the insignificant activities associated with zero coefficients. Fig. 6 illustrates four example activity groups (see Table 3 for interpretations), where the x-axis represents all activities and the y-axis stands for the coefficient (or importance) of each activity. We observe that only one or a few activities have high coefficients, which we can use to interpret or label each user behavior.

Algorithm 1 Clustering heavy users using tNMF

```
Parameters: activity matrix $A$, $\delta = 0.7$, $s = 1$, $k = l = 10$;
Output: Clustering heavy users into $k + 1$ groups;
Create $A_r$ by randomly sampling $s$% rows from $A$;
Run tNMF on $A_r$ and obtain factorization results $R$, $C$ and $H$;
Identify vector $\hat{C}_{\cdot j} := \{C_{ij} / \sum_i C_{ij}\}$ corresponding to activity group $j$;
for each $u_i \in \mathcal{U}$ do
    Compute the minimum Jenson-Shannon divergence $d_j := \min_j JS(A_{j\cdot}, \hat{C}^T_{\cdot j}), 1 \leq j \leq k$;
    if $d_j \leq \delta$ then
        Assign $u_i$ to user group $j$;
    else
        Assign $u_i$ to user group $k + 1$;
    end if
end for
```

How many heavy user groups? In order to apply tNMF on an activity matrix, we first need to determine First, we must provide the ranks k and l. Their optimal values correspond to the number of user groups and activity groups in the activity matrix. To solve this problem, we sample 1% of users (1% of rows in A) for multiple times, and run tNMF on these samples. Two criteria are employed in choosing appropriate ranks: (i) passing the goodness-of-fit tests introduced in [7]; (ii) reaching a stable clustering of activities across different user samples, i.e., a stable C.

Following this process, we select $k = l = 10$ for the activity matrix from the whole day data in 08/23/2010[5]. This implies that *a few network activities are responsible for the formation of a majority of heavy data users*, i.e., the activity matrix A can be factorized into much lower dimensional components, which is really surprising, given the huge number of heavy users and activities appeared in the data. In addition, we find from the group interaction matrix H that each user group mainly interacts with one particular activity group (i.e., there is only one non-zero entry in each row and each column of H). This implies that each heavy user cluster is primarily caused by one specific group of activities, and hence we can label user clusters using the corresponding activity groups.

Clustering heavy users. Since A is a high-dimensional matrix, directly applying the original tNMF on A is intractable. Instead, we propose a sampling based approach which scales up to large activity matrices. More specifically, we first randomly sample s% users (s% rows in A), which we denote as A_r. With the input ranks $k = l = 10$, we run tNMF on A_r and obtain the factorization results R, C and H (Eq. 1). The matrix C now contains the dominant activities, which is interpreted the same way as we have mentioned above. However, the indicator matrix R here only provides clustering results for the s% heavy user samples. Due to this reason, we do not rely on R for clustering heavy users. Instead, for each heavy user i, we measure the distance between the corresponding activity vector $A_{i\cdot}$ and each column of C to determine which activity group the user is associated with. We summarize the clustering process in Algorithm 1.

The Jensen-Shannon Divergence (JSD) is a symmetrized and smoothed version of the Kullback-Leibler divergence and is used to measure the distance between two distributions. Fig. 5 shows the histogram of the minimum JSD between the heavy user behaviors and the 10 centroids in C. We use $\delta = 0.7$ and classify all heavy users with a minimum JSD more than δ into a new group, representing the heavy data users whose behaviors differ significantly from all dominant behaviors. At the end, we cluster all heavy users into 11 groups, which are summarized in Table 3.

As expected, Table 3 shows mobile video/audio streaming is the main driving factor of heavy usage, which attributes to around 40% of all heavy users. Two popular online movie sites and one Internet radio site form three clusters, where users are expected to participate in long duration streaming with relatively stable speeds. In comparison, two video clip sites form 4 clusters (the second site forms three clusters due to three different domain names, where `Clip_2A` supports only one particular type of mobile devices). The video files hosted by these sites are generally smaller and shorter. Another cluster related to streaming activities is `CDN`, representing traffic associated with a content delivery network, which mainly hosts online videos.

In addition to streaming activities, one popular online social network (`OSN`) activity covers 26.8% of all users, and most data usage is attributed to uploading/downloading images and watching small video clips on the `OSN` website. Another large cluster, Apps, con-

[5] The activity groups remain stable in our two-week dataset. Tracking activity groups to identify significant user behavior changes will be our future work.

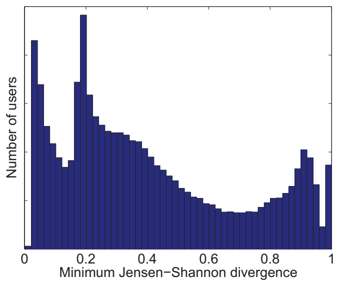


Figure 5: Minimum JSD.

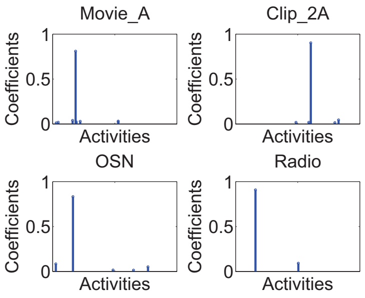


Figure 6: Activity groups.

Table 3: Heavy user clusters

Group	Pct.	Associated dominant network activities
Apps	11.2	Mobile apps
CDN	1.7	A content delivery network for videos
Clip_1	5.9	Video clip site 1
Clip_2A	1.2	Video clip site 2, with domain name A
Clip_2B	1.9	Video clip site 2, with domain name B
Clip_2C	8.9	Video clip site 2, with domain name C
Movie_1	0.7	Online movie site 1
Movie_2	1.3	Online movie site 2
OSN	26.8	An online social network site
Radio	14.7	An Internet radio site
Other	25.7	Other non-dominant activities

taining 11.2% of heavy users, is related to traffic associated with small mobile applications, e.g., online photo sharing, messaging, etc. The cluster with non-dominant activities (Other) account for 25.7% of heavy users.

Other findings. In addition to characterizing heavy user activities, we also investigate the temporal and spatial properties of heavy data users. We find that a heavy user tends to access data services at a single specific location over time, and therefore causes a persistent traffic increase at that location. More importantly, these heavy users exhibit strong temporal and spatial clustering. That is, they tend to concentrate in a few locations which are already crowded with a large subscriber population, e.g., large cities, and form many *hotspots*, where a significant increase of traffic volume is observed from these heavy users. For example, in a majority of the RNCs from 3 geographically dispersed markets, 50% to 70% of the total bytes during busy hours were from heavy users. At the SGSN level, as much as 50% of the bytes at an SGSN were due to heavy users at one single downstream RNC. Heavy users also become more active concurrently with normal users, during the busy hours of the day, and thereby impose additional pressure on the parts of cellular network which may be already overloaded or congested.

Implications. Despite the small number of heavy users, because each of them consumes data that is an order of magnitude more than a normal user does they are the main driving factor behind the variation of the traffic volume in the network. Due to their intensive data usage, the temporal/spatial clustering of heavy users can significantly increase the traffic volume at a particular network location and hence potentially impact the other normal users' experience at the nearby locations. For example, they can easily occupy all the high speed channels for a long time. As a consequence, the other normal users have to choose the relatively low speed channel and thus may not attain satisfactory network performance. Understanding such phenomenon can be very useful for traffic load balancing and troubleshooting network performance issues, such as congestion. In addition, the knowledge of the spatial and temporal distribution of different data users can provide useful guidance for many network management tasks, such as capacity planning and resource allocation, etc.

Our analysis also shows that heavy users tend to concentrate on a surprisingly small number of network activities, such as a few video/audio streaming sites and social network sites. This provides us with a good opportunity to design specific optimization strategies for these sites to reduce the traffic load from heavy users. For example, since most data usage from heavy users are associated with a few content providers, especially video/audio content providers, special data transmission schemes can be applied on these sites for improving the overall user experience. For example, these content providers can adaptively lower encoding rates during busy hours to reduce the network load. Cellular providers can also adopt specific strategies to remedy the potential adverse impact from different groups of heavy users. For example, more Wi-Fi hotspots can be deployed to offload a significant amount of traffic from heavy users. Also, we can encourage (with certain incentives) heavy users to watch videos during off-peak hours, when normal user activities are less dominant.

5. CONCLUSION AND FUTURE WORK

In this paper, we characterized usage patterns of mobile data users in a large 3G cellular network. By analyzing datasets collected from different locations of the network, we identified two user groups, heavy users and normal users, with distinct preference on mobile applications and devices. Using a novel co-clustering algorithm, we further categorized heavy users and extracted the network activities that triggered the intensive usage from heavy users.

We are continuing this research along a number of interesting dimensions, including understanding how heavy users collectively affect the voice calls and other services in a cellular network.

Acknowledgement

This work is supported in part by the NSF grants CNS-0905037, CNS-1017647, CNS-1117536 and the DTRA Grant HDTRA1-09-1-0050.

6. REFERENCES

[1] Mobile data traffic surpasses voice, 2010. http://www.ericsson.com/thecompany/press/releases/2010/03/1396928.

[2] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in smartphone usage. In *MobiSys '10*, 2010.

[3] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring serendipity: connecting people, locations and interests in a mobile 3g network. In *IMC '09*, 2009.

[4] G. Maier, F. Schneider, and A. Feldmann. A first look at mobile hand-held device traffic. In *PAM '10*, 2010.

[5] U. Paul, A. Subramanian, M. Buddhikot, and S. Das. Understanding traffic dynamics in cellular data networks. In *Infocom '11*, 2011.

[6] Y. Jin, N. Duffield, J. Erman, P. Haffner, S. Sen, and Z.-L. Zhang. A modular machine learning system for flow-level traffic classification in large networks. *ACM Trans. Knowl. Discov. Data*, 6(1), March 2012.

[7] Y. Jin, E. Sharafuddin, and Z-L. Zhang. Unveiling core network-wide communication patterns through application traffic activity graph decomposition. In *Proc. of SIGMETRICS '09*, pages 49–60, 2009.