Walelgne, Ermias; Asrese, Alemnew; Manner, Jukka; Bajpai, Vaibhav; Ott, Jörg

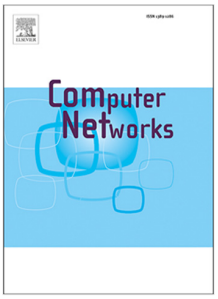
Clustering and predicting the data usage patterns of geographically diverse mobile users

Published in:
Computer Networks

DOI:
10.1016/j.comnet.2020.107737

Published: 14/03/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print



Clustering and predicting the data usage patterns of geographically diverse mobile users

Ermias Andargie Walelgne, Alemnew Sheferaw Asrese, Jukka Manner, Vaibhav Bajpai, Jörg Ott

PII: S1389-1286(20)31324-4
DOI: https://doi.org/10.1016/j.comnet.2020.107737
Reference: COMPNW 107737

To appear in: *Computer Networks*

Received date : 14 April 2020
Revised date : 25 November 2020
Accepted date : 10 December 2020

Please cite this article as: E.A. Walelgne, A.S. Asrese, J. Manner et al., Clustering and predicting the data usage patterns of geographically diverse mobile users, *Computer Networks* (2020), doi: https://doi.org/10.1016/j.comnet.2020.107737.

Clustering and Predicting the Data Usage Patterns of Geographically Diverse Mobile Users

Ermias Andargie Walelgne[a,*], Alemnew Sheferaw Asrese[a], Jukka Manner[a], Vaibhav Bajpai[b] and Jörg Ott[b]

[a]Aalto University, Finland
[b]Technical University of Munich, Germany

ARTICLE INFO

Keywords:
Mobile Networks
Data usage patterns
User behavior modeling
Clustering data usage

Abstract

Mobile users demand more and more data traffic, yet network resources are limited. This creates a challenge for network resource management. One way of addressing this challenge is by understanding the data usage patterns of mobile users so that resources can be optimally allocated based on user traffic demand and data usage behavior. However, understanding and characterizing the data usage patterns of mobile users is a complex task. In this work, we investigate and characterize users' data usage patterns and behavior in mobile networks. We leverage a dataset (∼113 million records) collected through a crowd-based mobile network measurement platform – Netradar – across five countries. Data usage behavior of users over a cellular network is primarily driven by user mobility, the type of subscription plan marketed by Mobile Network Operators (MNOs), network congestion, and network coverage. We apply an unsupervised machine learning approach to cluster mobile user types by considering different factors such as data consumption, network access type, the number of sessions created per user, throughput, and mobility. By defining data usage pattern of mobile users, we develop a user clustering model and identify three different mobile user groups (clusters). Our clustering model shows that the data usage patterns are unevenly distributed across the five countries studied, characterized by a small number of heavy users consuming the highest volume of data. We show how the types of applications installed by users correlate with data consumption patterns in some countries. Heavy users tend to install more traffic-demanding apps than users from the other two groups – regular and light users. Finally, we trained a classification model using the labeled dataset produced by our aforementioned user clustering method. The model helps classifying mobile users according to their usage patterns (i.e., heavy, regular, and light) with an accuracy of ∼80% in the test dataset.

1. Introduction

The demand for mobile data traffic is increasing. As mobile technology and network coverage improve over time, the use of data-intensive applications such as video streaming from mobile devices is growing rapidly. Studies show that a significant share of traffic generated from mobile devices increasingly consists of multimedia content [1, 2, 3, 4]. As reported by Ericsson [5] in 2018, video application content alone covered 60% of the mobile data traffic and it is projected to cover 74% of the traffic by 2024. Moreover, according to Cisco's global mobile data traffic report [6], there is an increasing demand for mobile data traffic where more than 75% of mobile data traffic will be multimedia contents by 2020.

Mobile networks are becoming more heterogeneous to keep up with the ever-increasing demand for mobile traffic [7]. As the traffic demand, the complexity of the network, and the number of users increase, managing the network resources and understanding the data usage patterns of mobile users becomes complex. As a result, service and content providers need to efficiently manage available resources based on the data usage behavior of their customers. Identifying the data usage patterns of mobile users can be useful in various scenarios, such as managing the increasing demand for mobile data usage [8], understanding urban dynamics [9] for improved urban planning [10], developing of data plan products[1], as well as enhancing communication and service quality. It has also been shown that data usage and transfer patterns of different applications have a significant impact on the energy consumption of mobile devices [2]. Therefore, understanding the data usage patterns of mobile users at various locations and market places is becoming increasingly important.

A number of research articles focus on mobile data usage patterns and behavior. The focus of previous work includes user location and mobility patterns [12, 13], temporal dynamics of mobile users [14], and quality of experience [15, 16, 17]. However, most of the previous studies are limited to a single operator [18]; target a specific city and location [19]; study usage behavior targeting application types accessed by users [20, 21, 22, 23]; or are based on limited measurement data and user spaces [24].

In this paper, we characterize data usage patterns and behaviors of mobile users across five different countries. We study whether clusters emerge from users' data usage patterns and user behavior. We also investigate whether it is possible to build a model that predicts (classifies) such type of data usage patterns using machine learning methods. This

*Corresponding author
ermias.walelgne@aalto.fi (E.A. Walelgne); alemnew.asrese@aalto.fi (A.S. Asrese); jukka.manner@aalto.fi (J. Manner); bajpaiv@in.tum.de (V. Bajpai); ott@in.tum.de (J. Ott)

[1]For instance, Telecom Italia introduces unlimited access on chat, streaming and social media but limited data usage on other services [11].

paper provides new insight into mobile users' data usage patterns and solid confirmation on the previous outcomes using real-world measurement datasets across countries. The **contributions** of this paper are summarized as follows:

First, we leverage a measurement dataset (∼113M records) collected using the Netradar [25] mobile measurement platform from five countries. The dataset covers a wide range of geographical areas, MNOs, and mobile users. We investigate data usage patterns of mobile users by considering both data traffic flows and the type of installed applications.

Second, we define the data usage pattern of mobile users based on user mobility, location, device model, network performance (e.g., throughput and latency), and network technology coverage. Using this definition, we present a mobile user clustering model by applying an unsupervised machine learning algorithm. Using the clustering model, we identify and study the commonality among groups of mobile users, including the trend of data consumption and interaction with their device. We observe that small share of heavy cellular network users (from 2% to 4%) have the highest data consumption. We also show that the type of apps installed by mobile users has a relationship with the users' data consumption patterns. For instance, heavy users often install more apps that generate high data traffic, such as photography, social media, and video players.

Third, we develop a prediction model that helps classify mobile users' data usage patterns and behavior. If the users' data usage pattern and behavior are predictable, MNOs can apply different pricing and traffic resource optimizing methods [26] based on their customer resource demand and data usage patterns. In-line with this, using the labeled dataset produced from the user clustering method as input features, we develop a classification model that helps classify users' data usage patterns (with 80% accuracy on the test dataset). As part of our contribution, we will make the dataset available to the community, upon publication.

The paper is structured as follows. Section 2 presents the measurement platform and the dataset that we have used for the analysis. Section 3 presents the unsupervised model of clustering mobile users based on their data usage behavior. Section 4 presents a supervised based user classification model. Finally, Section 5 covers related work, and Section 7 concludes the paper.

2. Methodology

In this section, we introduce the Netradar mobile measurement platform, which has been used to collect the dataset. We also describe the metrics that we have used from the dataset.

2.1. Measurement Platform

Netradar [25] is a crowdsourced mobile measurement platform. The platform estimates link capacity of cellular networks on smartphones, using probe-based measurement methodologies. The method is a hybrid of Probe Gap Model (PGM) and Probe Rate Model (PRM) [27]. PGM and PRM utilize packet pair [28] probes to estimate the available bandwidth. In this paper, we describe the parts of the measurement platform relevant to our study. The detailed description of the Netradar measurement platform and its validation are available [29].

The Netradar platform passively listens to the ingress and egress traffic of a device without imposing any synthetic traffic. The measurement application at the client device runs in the background until it triggers the measurement when a user starts sending or receiving data. If it observes incoming or outgoing traffic on the device, then the application starts sampling the traffic rate of the ingress and egress traffic (e.g., on Android using Android traffic Stat API [30]). Currently, the measurement platform runs on Android mobile devices.

The session starts if there is enough traffic (at least five IP packets) in either the uplink or downlink direction. The session ends if the link stays idle for two seconds. Session duration is defined as the interval between the starting time of the sampling phase until the traffic stops. The duration of the session can be in the range from less than a second to several minutes. The platform does not record sessions of only a few packets (< 5 IP packets).

The platform also records unconstrained and constrained speeds of the network. Unconstrained speed is the maximum speed recorded during the session when users were not limited by the network. It is the data rate that the user needs from the network to use mobile apps on his/her device suitably. In contrast, the constrained speed is recorded when the network is a limiting factor. It reflects the maximum data rate offered by the network to the user. It is inferred based on the queuing delay of packets, the available bandwidth and the latency information [29].

The constrained speed can be `null` if there is no latency information or the user never hit the network speed (*i.e.*, constrained did not happened at all; e.g., when the server did not send data fast enough that could possibly congest the network). A given session can have only uplink or downlink data recorded. For instance, if the user is watching video from YouTube, most of the sessions are downlink data. For such cases, there is no need to send much data on the uplink, or there are very few traffic which is not statistically sufficient to keep records related to the uplink information. On the other hand, if a user is uploading a picture to Facebook for instance, then most traffic flows are in the uplink direction.

Every session has a unique identifier with its own start and end time and metadata about the session. For each measurement session the system records the following meta information: device information, information about the subscriber's MNO, location and user velocity. The methadata also contains information about network type (WiFi or cellular) and accessed radio technologies (2G, 3G, 4G) with detailed radio quality of service (QoS) values. The constrained and unconstrained speeds for both downlink and uplink are also recorded. Besides, every session contains the average download and upload speed, total upload and download bytes, latency, signal strength, session length, and

information about the base station (e.g., cell ID, area code, radio frequency channel number). Every session has associated tile information (e.g., country, city, population density), where each tile is the area coverage of 100m by 100m times.

2.2. Dataset

Table 1
Number of users and sessions created in cellular networks by country.

Country	# of sessions (M)	# Users
Finland (FI)	35.1	22795
United Kingdom (UK)	34	20529
Japan (JP)	19.8	8081
Brazil (BR)	17.8	7164
Germany (DE)	6.3	6548

The dataset we use for the analysis is collected using the Netradar mobile network measurement platform. We use data from the mobile users devices in five different countries (Finland, Germany, the United Kingdom, Japan, and Brazil). Mobile users in the respective countries are identified based on the network and subscribers Mobile Country Code (MCC) value. In other words, a user in a given country has to be a subscriber to one of the Mobile Network Operators (MNO) in that country and accessing the network within the same country. For the sake of simplicity, roaming users are not included. For our analysis, we use a month-long (July 2018) dataset. The number of measurement sessions created per country over cellular networks are in the order of millions. Table 1 summarizes the number of users and sessions created per country in cellular networks. We performed a detailed analysis of different network factors and data usage patterns of mobile users across six countries [31]. We showed that the data usage behavior of mobile users depends on different factors such as user mobility, presence of network congestion, the accessed radio technology type, and network coverage. In this paper, we use our previous analysis as a base for feature selection to develop a clustering model of mobile users' data usage patterns.

3. Clustering Mobile Users

In this section, we apply the unsupervised K-means clustering method [32] to group mobile users based on their traffic consumption and activity level. First, we present data processing followed by similarity computation to create user clustering. Then we present the analysis on the types of users and the app categories they use per cluster.

Say for a set of m users and a maximum limit of a given time T, usage pattern UP is defined by the tuple $UP = (C, F, R, P, D, V, N, B)$, where C is the number of times user runs under congestion (constrained download speeds); F is the number of times the user runs without congestion (unconstrained download speeds); R is the percentage of accessing 4G or 3G network [2]; P is the average population size of the area where the user was accessing the network; V is the throughput (average constrained and unconstrained download speed); D is the average session duration; N is the number of sessions created by the user; B is the total data volume transferred (uplink, and downlink direction) per time interval t. For our case, the time interval t is set to one hour. The usage pattern for a given user ID i can be defined as: $up^i = up^i_k | 1 \leq k \leq T$. Therefore, the usage pattern for a given user ID i at time k can be defined as: $up^i_k = (c^i_k, f^i_k, r^i_k, p^i_k, d^i_k, v^i_k, n^i_k, b^i_k)$. User ID I is a combination of *installation id*, *device model*, *device brand name*, *subscriber MNO*, and *network country*. The aforementioned eight features are identified and selected based on the observation from our previous [31] and other related work [19, 33].

3.1. Data processing

For cluster analysis, we consider a one-month (July 2018) dataset collected from five countries over cellular networks. Note that, since September 2018, Android started halting background processes that stay for longer sessions. Therefore, we picked July 2018 as it has more datasets and the measurements were not interrupted by the OS.

The dataset is filtered and prepared as follows: First, for every user i, the measurement data is grouped by user ID. The time slot we choose is a one-hour interval. So, u^t_i contains a set of all sessions that lay within the time interval of t for a given user i. Here, the time t we considered is every hour across all days per user during the one-month. For every user, we calculate the hourly total traffic flows (downlink + uplink), average session duration, number of sessions, the number of times users get access to 4G and 3G network, the number of times users run into congestion and without congestion, the average download and upload speeds that the user gets when it run into congestion (constrained download speed). From our observations, the parameters mentioned above have small variation across different days in the same hour, as also observed in [19]. Hence a one-hour time slot for the whole month could represent the traffic of a given user with better time granularity.

For clustering, we consider only users that have at least seven days of active measurement sessions during the month. As a result, for every user, we get enough measurements and user activities for at least seven different days. This is important to capture the data usage behavior of users. Note that, since a user might not have measurements in all of the time intervals, we assume that missed values are created either because users were not using their devices to access the Internet or there was no active traffic flowing in both ingresses and egress directions. Finally, the eight features are reshaped and transformed into row-vector $V(f, h)$, where f represents the features and $h = \{1, \ldots, 24\}$. Every row representing a user have information on the aforementioned features at every one-hour interval.

[2] Note that 4G refers to all releases of Long Term Evolution (LTE) radio technology. The 3G refers to all other releases of radio technologies prior to LTE.

3.2. Similarity computation and clustering

User grouping and clustering are performed based on the filtered and prepared datasets using Python's scikit-learn library [34]. We use the K-means clustering [35] method, which uses the Euclidean distance to measure the similarity between the mobile users. For a set of m mobile users $U = \{u_1, u_2, u_3, ..., u_m\}$, as defined above, we apply the Euclidean distance [36] to compute the similarity between metrics for every user. As k-means clustering depends on the distance matrix to group data points, the algorithm works well when all features are in a common range. This is important so that features with large scale value do no dominate small scale features [37]. To ensure this, before computing the distances, the dataset is normalized so that the mean and the standard deviation is 0 and 1, respectively.

Since the measurement has multiple features, we applied dimensionality reduction using Principal Component Analysis (PCA), before performing K-means clustering. Applying PCA helps to reduce the dimensionality of the feature space without losing too much information. As shown in [38], applying PCA before clustering potentially improves the clustering quality. Especially for the K-means-based clustering method, the PCA can potentially improve the accuracy of the cluster. We also observe that the quality of the cluster improves when we apply PCA than without PCA. We choose to use the PCA component size that explains 99% of the variance.

Let $C = c_1, c_2, c_3, ..., c_k$ be a set of clusters where every c_i is a group of users with 'similar' traffic patterns and demand, and k is the number of clusters. In unsupervised learning, the number of clusters (i.e., the number of centroids) has to be specified before doing the clustering. To decide on the optimum number of clusters, we use within cluster sum of squares (WCSS) – elbow method and the dendrogram structure of hierarchical clustering. Note that, we have also tested hierarchical clustering and found that K-means is faster and produces more valid clusters. The validity of the clusters (as we will discuss in the next paragraph) is measured in terms of groups of users that have similar data usage patterns. Accordingly, we found that K-means has better cluster results on the dataset than the hierarchical clustering method.

To measure the quality of the separation of the clusters, we applied three different stopping criteria. The criteria are silhouette score (SH) [39], Calinski-Harabasz index (CH) [40], Davies Bouldin score (DB) [41] and Dunn index (DI) [42]. These scores are among the list of recommended metrics for choosing the number of clusters as surveyed in [43, 44]. The silhouette score is a measure of how close each data point is in a single cluster (cohesion) compared with the other clusters (separation). The silhouette coefficients measure is in the range between -1 and 1, where the optimal clustering is with the highest SH score. SH of negative value suggests a data point is wrongly assigned to the cluster. The Calinski-Harabasz index is another measure that is used to quantify how well the clusters at different groups are separated and how data points in a single cluster are closer to each other.

Table 2
The cluster size (K) with different cluster separation quality measure values.

K	CH	SH	DB	DI
2	1203.99	0.26	2.5	0.015833
3	968.91	0.12	2.58	0.013325
4	822.73	0.12	2.35	0.007417
5	755.61	0.1	2.25	0.013325
6	687.46	0.09	2.2	0.013325

The CH score is higher when clusters are dense and well separated. The DB index measures the dispersion of data points within a cluster (intra-cluster distance) in terms of the dissimilarity measure between two different clusters (the inter-cluster distance). The DB index values closer to zero indicate a better partition. The optimal number of clustering could be found by minimizing the DB index values. The Dunn index (DI) [42] is a metric is an internal cluster evaluation scheme, where the metrics is calculated based on the clustered data itself. Higher values of the DI score indicate better clustering. Table 2 shows the score of these values at different cluster sizes.

To decide the cluster size, we use the combination of domain knowledge (e.g., from a previous study [19], WCSS – elbow method, the dendrogram structure (plot not shown), and optimizing the aforementioned scores (CH, SH, DB, and DI). As a result, we choose the cluster size of $K = 3$ and have run the K-means clustering algorithm. The optimum number is chosen after testing each country dataset separately. We found that the cluster size of three is a more reasonable number of user groups based on the metrics mentioned above. The K-means clustering algorithm runs 1000 times independently with centroids selection based on K-means++ [45]. Running the algorithm several times with different initialization of the centroid is essential so that it does not converge to the local minimum. K-means++ based centroid selection randomly picks the centroid (of K size) for the first iteration. Then it assigns each data point to the nearest centroid based on the calculated distance. K-means++ based centroid selection chooses the centroid that minimizes the Sum of Square (SS) distance between every data point to the class centroid. The maximum number of iteration for every single run is set to 600 with the tolerance value of 0.0001.

3.3. Cluster analysis

We run the clustering algorithm over the dataset for each country separately and identify group of cellular users belonging together based on their data usage patterns. Accordingly, we found three distinct group of users and named them as 'Heavy', 'Regular', and 'Light' users. 'Heavy' users can be characterized as users that consume the highest volume of data (both in download and upload). They are also who frequently interact with their devices (based on the number of sessions created by users). 'Light' users are users with a small amount of download/upload bytes and session number. 'Regular' users are found in between the two user groups.

Table 3
The median data consumption and number of sessions per user group across countries. The three numbers inside the bracket next to each country name show the number of users (%) under Heavy, Regular, and Light clusters, respectively. The numbers under the total Bytes column separated by the pipe represent the download and upload size, respectively.

Country (%)	Total Bytes (MB)			# Sessions		
	Heavy (H)	Regular (R)	Light (L)	H	R	L
FI (3.5, 41.9, 54.6)	309.9 | 18.8	59.9 | 4.4	8.5 | 0.9	389	85	20
DE (2.2, 23, 74.8)	58.2 | 5.1	9.9 | 1.2	2.4 | 0.3	169	32	10
UK (3.1, 27.2, 69.7)	65.2 | 5.8	12.1 | 1.4	2.9 | 0.4	146	32	11
JP (3.1, 32.9, 64)	98.4 | 7.7	14.7 | 1.6	3.4 | 0.4	211	38	11
BR (4.6, 35.8, 59.6)	186.7 | 13.8	29.8 | 2.8	4.8 | 0.7	277	71	21

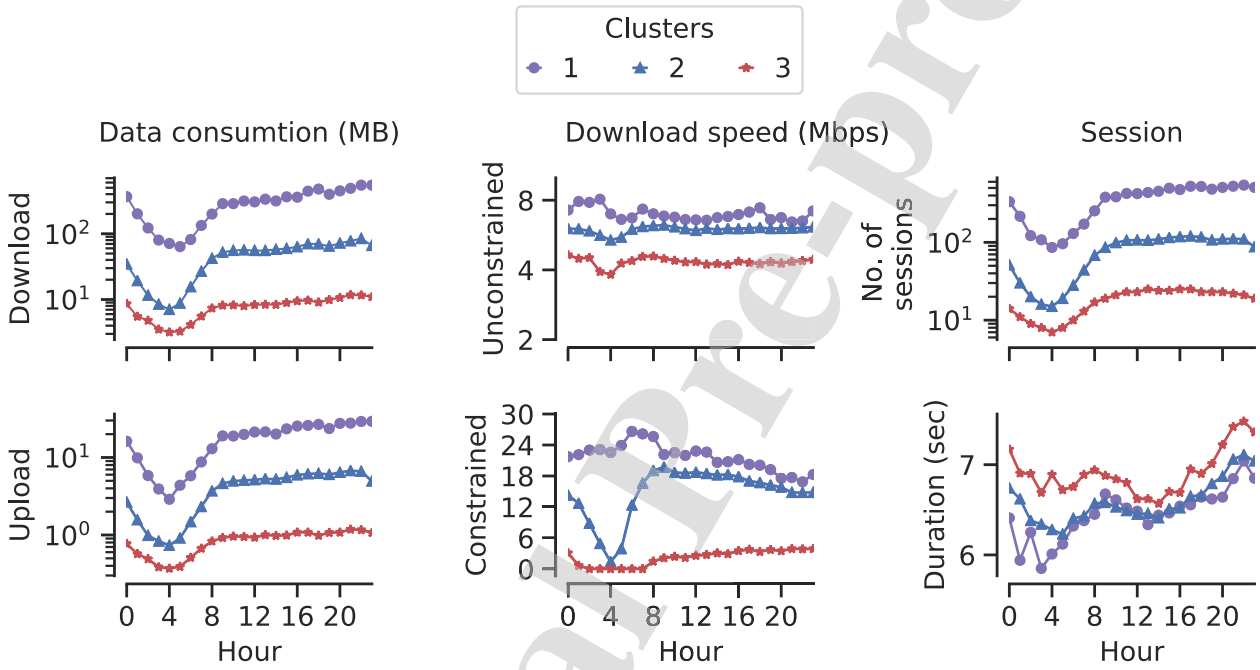


Figure 1: The median distribution of different metrics per cluster for mobile users in Finland. Note the scale difference in both X and Y axes.

Identifying a heavy user from light users can be, for instance, used as an input for optimal resources management based on demands.

Table 3 summarizes the number of users in each cluster along with total data consumption and sessions created per every country. It shows that both the number of users and the data consumption grouped together in one country varies compared with the similar group (cluster labels) in another country. For instance, considering the level of data consumption and session creation heavy users in Finland and Brazil consume two times more than heavy users in Germany and the UK. For brevity, we discuss the cluster of two countries (Finland and Brazil) in detail and opt out the detail discussion of other countries.

Fig. 1 depicts the median total download and upload values (left column); the unconstrained and constrained data rates (middle column); the number of sessions and session duration (the right column) per cluster for users in Finland. From the figure, we can observe that cluster 1 consists of a group of users that mostly consume the highest amount of download (309.9 MB) and upload (18.8 MB) values in all times of the day. We refer this group of users as heavy users. Heavy users cover only 3.5% of the total users from our measurement. Note that, the label assignment as 'heavy', 'regular', and 'light' user is primarily based on the median total download/upload bytes and the number of sessions created per clusters from highest to lowest. These group of users actively engage with their device as can be seen from the number of sessions created by the user (top right column). These are also users who get the highest download speed compared to the other user groups. Moreover, they constantly hit the maximum network speed of the network (constrained speed) at all times of the day. In the heavy users group, there are more than 3.8M measurements collected from 399 unique users.

Users in cluster 2 can be considered as regular users.

They cover ∼42% of the total users. Most of the time, regular users do not hit the maximum network speed. In a cluster 2, there are more than 12.3M measurements collected from more that 4.7K unique users. By considering the total measurement sessions from each cluster, we observe that 86% of the sessions were collected over LTE networks whereas the remaining 14% over 3G networks (i.e., 12% (HSPA+), and 2% of the time over the other 3G network families). We observed a similar distribution between heavy and regular users in terms of radio technology they accessed (i.e., 3G and 4G).

Cluster 3 consists of the majority of mobile users, which covers ∼55% of the users from the measurement dataset. These group of users can be considered as light users. Light users consume lower data volume (in both download and upload). They are less engaged with their device as can be seen from the median number of sessions created per user. There are more than 6.2M sessions generated from ∼4K unique users. Considering the total measurement sessions created by light users, we observe that 69% of the sessions are created over LTE network and the rest 31% over 3G networks (i.e., 27% (HSPA+), and 4% over the other 3G network families).

To understand the reason why light users have accessed 3G network more often than the other two groups, we study the user mobility and the availability of network coverage. Users in this group have visited the least number of unique tile-IDs (in the median case, 3). Users in Cluster 1 and 2 have visited 15 and 7 unique tile-IDs in the median case, respectively. This implies that the users accessed the 3G networks, either because their subscription plan was 3G networks, or they were living in an area where there is no 4G coverage. To investigate this further, we cross-checked the number of users that never get access to the LTE network at least once. We found that about 15% of users in the light user group have never gotten access to the LTE network throughout the measurement period, while the rest (85%) have accessed the LTE network at least once. Focusing on the 15% of the users, we cross-check the location of the base station. We use OpenCellID [46] service to map the location based on network MCC, mobile network code, cell ID, and location area code, where users were connected to. We observe that in the areas where these users moved around have 4G radio coverage. This implies that users were accessing the 3G network due to their data subscription plan, but not due to the lack of 4G network coverage.

Fig. 2 shows the three clusters for the measurement data from Brazil. It shows the median distribution of total download and upload, the number of constrained and unconstrained speeds, the number of sessions, and session duration per user at every hour for the three clusters. Each cluster from 1 to 3 has 4.6, 59.7, and 35.7% of users, respectively.

Cluster 1, covering the least number of mobile users (∼5% of the users), are 'heavy users' with the highest median total download and upload amount of ∼187 MB and ∼14 MB, respectively. This group of users mostly accesses 4G networks in 89.4% of the cases, and they visited six unique tiles on average, which is the highest from the other groups. This indicates that this group of users move more frequently from place to place than the other groups. They are also those who mostly hit the network maximum since they have the highest number of download constrained speed than the other groups. We study the type of radio technology accessed by this group of users. We observe that, from the total measurement sessions collected from these users, 63% of them were accessing the LTE network. This group has the highest percentage in terms of accessing to LTE network. The rest of the sessions were collected over 3G networks (28%).

Regular users (cluster 2) in Brazil cover ∼36% of the total users from our measurement. They have the second data consumption value in both total median download (∼30 MB) and upload (∼3 MB). Observing the radio technology distribution, from the total measurements collected from this group of users, 60%, and 35% of them have been measured over LTE and 3G networks, respectively. The majority of users (60%) are light users, as shown in the figure labeled with cluster 3. They have a total median download and upload of 4.8 MB and 0.7 MB, respectively. Compared to the other groups, users in this group have accessed the LTE network less frequently. From the total number of sessions created by this group, only 44.5% of them were over the LTE network. In more than 45.8% of the measurement sessions, there were 3G networks, and in 9% of the cases, the network technology was unknown due to different reasons (e.g., if the ITelephony interface is not up [47]).

Similar to users in Finland, the number of light users in Brazil also covers the highest percentage (60%), compared with the other clusters. The majority of the sessions (46%) created by this group of users have accessed 3G networks. We observe that only 33% of the sessions were accessing LTE networks. Note that here there are also some sessions with unknown radio technology. In both Finland and Brazil, we observe that a significant number of light users have accessed 3G networks. For instance, Light users in Brazil and Finland have accessed 3G networks 46% and 45.8%, respectively. When we consider the LTE, the users in Brazil got 33%), and in Finland 44.5%. This difference in the LTE network might be due to the penetration of 4G technologies in the developed region than the third world countries. We notice that access to the LTE network alone can not be a determining factor for users' data usage behavior. This is evident as a large number of users accessing LTE networks are observed in both light and regular user groups. Generally, we observe that though the percentage distribution of heavy, regular, and light users per the respective country has a similar trend, there is a significant variation on the data consumption, access to the radio technology type, and the number of sessions created under each the clusters.

3.4. Application category per cluster

Currently, there are more than 2.7 million applications only in the Google Play store [48]. From this plethora of applications, users need to make a selective installation based on their needs and preference. Different users have different

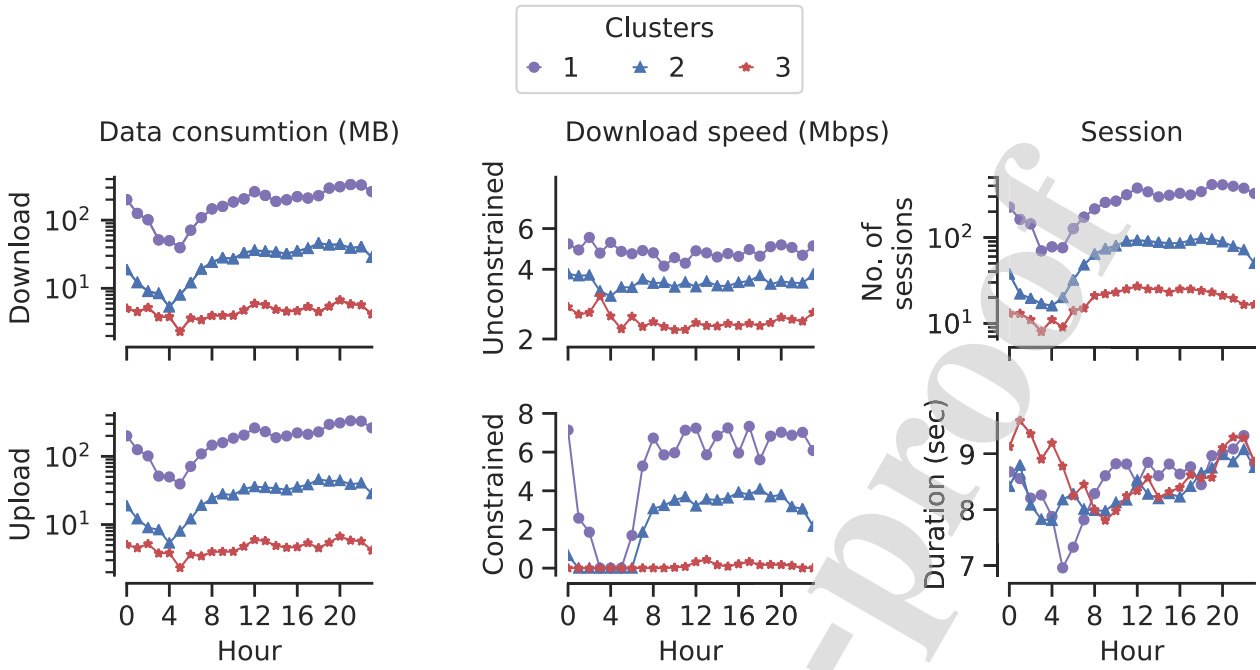


Figure 2: The median distribution of different metrics per cluster for mobile users in Brazil. Note the scale difference in both X and Y axis.

preferences on the type of app they want to access. Accordingly, users could spend a different amount of time while using various apps. The amount of traffic generated and consumed while accessing different types of apps also vary from application to application. Google Play has multiple app categories, where each category consists of numerous app types. For instance, apps designed for streaming video, movies, and TV contents are grouped with the entertainment category. We use the list of installed apps on the users' device to analyze which app categories contribute more to each cluster group. Note that the installed app list does not include apps that come with the device preinstalled by default. As a result, the list of installed apps we are focusing on is all types of app categories that are explicitly installed by users from the app store. We expect that focusing on the variety of app categories intentionally installed by the user reflects the interest and the data usage pattern of smartphone users.

To estimate the influence of the app category per user within each cluster, we calculate $p(a_i)$ – the proportion of a given app category a being installed by the user i as follows. $p(a_i) = \frac{a_i}{A_i}$ where, a_i is the number of apps installed by user i for the app category type a and A_i is the total number of apps installed by user i, irrespective of the app category.

Figures 3 and 4 show different application categories installed on each users' device and their proportion values per cluster for mobile users in Finland and Brazil, respectively. We can see that heavy users usually install app categories that consume much data volume including entertainment, video player, photography, and games. Regular users in Finland focus on entertainment, photography, social media, finance, music & audio, shopping, food & drinking apps than the light users' group. They frequently install apps related to weather, map & navigation, finance, sports, business, library & demo than heavy users. Differently, light users use apps related to education, tools, news & magazine, travel & local, and business app categories more frequently than the other two groups. However, when it comes to basic applications such as productivity, weather, books and references, which consume relatively smaller amounts of data volume, we do not observe a significant difference in the number of apps installed across the different user groups. Compared to Finnish users, Brazilian heavy user types have installed a few app categories that generate high data traffic. Despite this, Brazilian heavy users still install video players, photography, and productivity apps more frequently than the other groups. Regular users in Brazil use communication, shopping, sports, games, education, finance, professionalization, business, and weather apps more frequently than the other two groups. However, light users use social, map & navigation, lifestyle, entertainment, travel & local, books & reference app categories more often than the other two groups.

The proportion values of installed app categories per cluster in some countries (e.g., FI and DE) show that heavy users install more traffic demanding apps than regular and light users. Some of the app categories installed by heavy users in different countries include social, photography, video players, music and audio, gaming, and entertainment app categories. Productivity, news and magazine, tools, maps & navigation, and communication are app types often install by regular users than the other. In most cases, the app categories installed by light users are fewer than the other two groups. There are only a few cases, such as entertainment

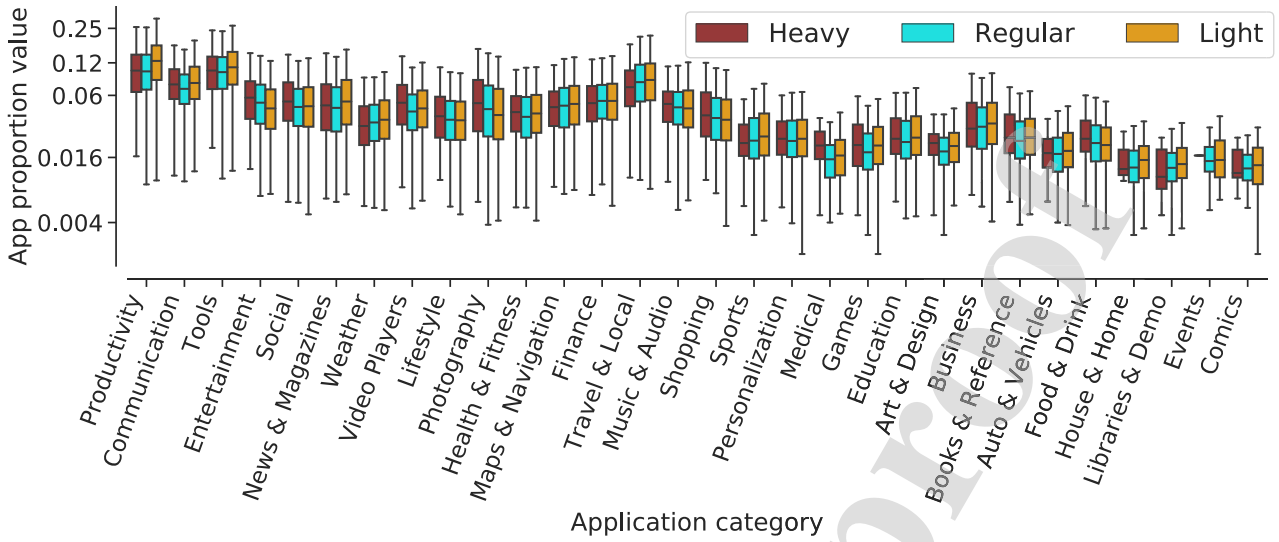


Figure 3: The proportion value of installed app categories per user within each cluster label for mobile users in Finland. Heavy users mostly installed apps that demand more traffic than other group of users. Note the log scale difference on the y axis ticks.

and photography app types, where light users installed more frequently than regular users. We observe that heavy users in Finland, the UK, and Germany have installed similar app types more than users in Japan and Brazil. Japan heavy users have installed apps categories such as photography, business, music and audio, gaming, book and reference, libraries and demo. Heavy users in five countries have photography as a common app category with the highest proportion value. Social and video player app categories are among the most commonly installed apps by heavy users in different countries (found in at least three countries). App categories installed in at least two different countries by heavy users include entertainment, social, video, photography, music and audio, food and drink, and personalization. For brevity, we opt out of the detail discussion of other countries.

Considering the type of apps installed by users and the cluster labels, we can observe that the installed app category by different user groups across countries is not necessarily related to the data usage patterns. This could be due to several reasons. For instance, users might install the apps and seldomly access them or never use them on their devices. We also noticed that the relation between the installed app types and users' data consumption patterns depends on the users' location. For instance, FI and DE heavy users commonly installed similar app categories (potentially generating more traffic) than the other countries. These app categories include Entertainment, Social, Video Player, and Photography.

Takeaway: User grouping (clustering) can be used to reveal different types of users based on their data traffic consumption and usage patterns. Generally, mobile users' data usage patterns and behavior can be categorized into three distinct user groups – heavy, regular, and light. We have seen that data usage patterns of mobile users are unevenly distributed, where few percentages of heavy users consume the highest volume of the data. The type of apps installed by mobile users could also be used as a hint and related with mobile users' data usage patterns. Although these depend on users' locations, we have observed that heavy mobile user groups in some countries have installed high data traffic demanding apps more often than the other two groups. We also observed that there is a significant variance in the amount of total data consumption and the number of sessions created across different countries of similar user groups. In other words, a group of mobile users that have been identified as heavy users in one country might not necessarily be categorized as heavy users in another country.

4. Mobile User Classification and Prediction

This section presents a classification model that we have applied to predict the mobile user type based on the labeled dataset generated from clustering (Section 3). Studies such as [49] suggest that customers are willing to pay flat-rate prices rather than being concerned with the detail cost analysis provided by operators. However, due to finite network resources, providing a flat-rate for all customers at all times is still challenging. Moreover, few heavy users might potentially create a bottleneck and become a cause for the poor experience to the other nearby users, especially during peak hours.

Usually, MNOs are applying different (both static and dynamic) pricing and traffic resource optimizing methods [50]. In addition to the flat-rate data plan, tiered-based data services and usage-based data plans are also the common types of data plan pricing schemes. For instance, operators such as AT&T propose to apply speed tiers in the upcoming 5G network [51]. These types of data plans can be used for optimal target pricing towards the users' traffic demand and data usage patterns. To achieve this, users' data usage patterns and behavior need to be predictable. If users' data usage behav-

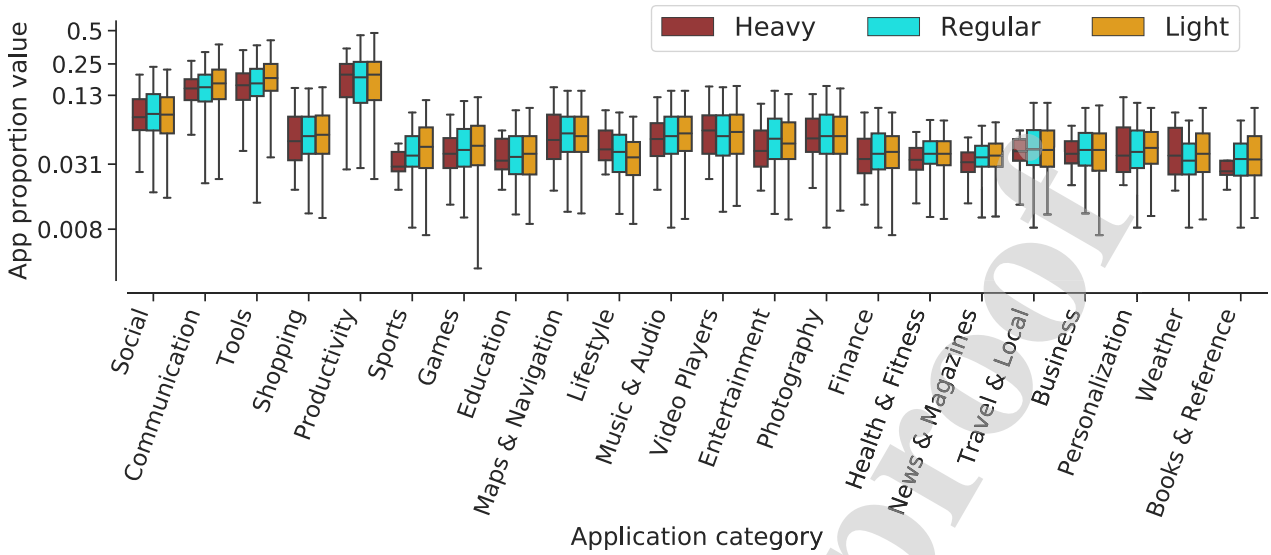


Figure 4: The proportion value of installed app categories per user within each cluster for users in Brazil. Compared with users in FI, the proportion of traffic demanding app types installed by Heavy users in BR are very few. Note the log scale difference on the y axis ticks.

ior is predictable, MNOs can apply an incentive mechanism targeting their customers. For instance, operators may lower the price for heavy users if users can shift their high traffic demanding tasks from peak hours to off-peak hours. We now develop a model that helps to predict the type of users in cellular networks based on users' data usage patterns and behavior.

Using labeled clustered dataset as an input, we apply a supervised machine learning algorithm to train a model that classifies user groups based on their group cluster labels. The prediction model does not consider the installed app category feature as detailed information about the app is not easily acquired (e.g., due to privacy), and the app category alone might not be precise enough. Since the variation between the number of users per cluster is very high, we have an imbalanced dataset. That is, the impact of majority class labels will overwhelm the classification model. Also, the minority cluster labels (e.g., clusters with a few numbers of users) could be missed out, even if the classification accuracy report is higher [52]. For instance, the cluster label imbalance ratio between heavy and regular users in Finland is ∼14. As a result, directly performing a general classification model on the existing clustered dataset could lead to a spurious mode that does not consider the impact of the minority class labels. To address this, we first apply the Synthetic Minority Over-Sampling Technique (SMOTE) [53] and Adaptive Synthetic (ADASYN) [54] methods. These techniques create new synthetic data points based on the observations of the actual dataset so that all the minority observations are oversampled and balanced. Note that, the accuracy we found by applying both techniques at a different time is closely similar. The reported accuracy is based on the SMOTE.

We take the clustered dataset from Finland as input; first, we split the dataset into two - testing and training datasets. The data split is performed with a ratio of 20% and 80% for the testing and training datasets, respectively. Then we train the classifier on the training dataset. Note that we perform the training with different classifier algorithms, such as the decision tree, gradient boosting classifiers, and random forest (RF). We found that the RF classifier method gives a better classification result than the other tested algorithms (a detailed analysis is omitted for brevity). By applying an exhaustive grid search along with ten-fold cross-validation, we found the optimal parameter values of the RF classifier. Accordingly, the model is trained by setting the depth of each tree in the forest (`max_depth`) to 32, the minimum number of data points allowed in a leaf node (`min_sample_leaf`) to 10, and the number of trees in the forest (`n_estimators`) to 500. The classification model of the RF model has accuracy of 87.5% and 79.9% on the training and the test, respectively.

Table 4 shows the confusion matrix of the three cluster labels for the RF classification model. The recall value of heavy, regular, and light class label are 91%, 77%, and 73%, respectively. The precision values are 90%, 66%, 84% for heavy, regular, and light, respectively. The F1- Score, which is the harmonic mean of precision and recall, is 90.5%, 71.1%, and 78.1%, for heavy, regular, and light class, respectively. The highest F1-score value, especially for heavy user type, shows that the RF model fairly clarifies the data usage pattern of mobile users with good precision and recall.

Fig. 5 shows the order of features based on their contribution to predicting the cluster labels. We observe that the number of times the user download contents over unconstrained speeds, the number of tiles visited by users, the number of sessions created per user, and total upload/download traffic contributes more to the classification model. These variables have been also played a significant role during our

Table 4
Confusion Matrix for RF classification in Finland.

Predicted \ Actual	Heavy	Regular	Light
Heavy	17531	1033	1004
Regular	1478	12851	5138
Light	192	2839	16259

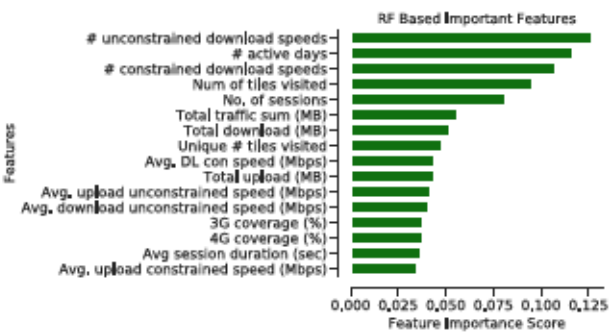


Figure 5: Rank of important features from RF classification.

clustering model as we have seen in Section 3.

Takeaway: Predicting data usage patterns and behavior of mobile users can be used to optimally target towards users' data traffic demand and usage patterns. In this section, using the labels of the clusters as input features, we develop a classification model that helps to classify mobile users' data usage patterns. From the rank of important features that have been used for classification, we observe that the presence or absence of congested network and user mobility play a role in mobile users' data usage patterns.

5. Related Work

Several studies analyzed the data usage patterns of mobile users in cellular networks (surveyed in [55, 56]). Previous work such as [57, 58, 59] studied mobile users' app usage behavior by considering different network usage and app related activities (e.g., installing, uninstalling, and updating). Shafiq *et al.* [4] model traffic dynamics on mobile devices using a week-long dataset collected from the operator's core network. They study traffic dynamics and characteristics of applications on three different cellular device brand families. They show that the type of device attributes to different traffic behavior.

Yang *et al.* [33] characterize user behavior in terms of mobility, data usage, and application usage pattern based on a dataset collected from 2G and 3G core networks in China. Similarly, the authors in [14] study the patterns of mobility and temporal activity as well as how the radio resources are utilized by different applications using a dataset collected from 3G core networks. The authors show that traffic distribution per subscribers is uneven, such that 90% of traffic load in the 3G network is generated by 10% of the subscribers. A study by Oliveira *et al.* [19], the closest study with our work, characterizes mobile users' data usage behavior based on the data collected in a 3G network in Mexico. The authors profile mobile users into three classes (light, medium, and heavy users) by using metrics such as the number of sessions, traffic volume, and inter-arrival time. Other studies such as [60, 61], and [62] study human mobility patterns and behavior in mobile networks. These studies show that the mobilities of mobile users have patterns overtime of a day and location.

Falaki *et al.* [24] study traffic generated from smartphone users along with user interaction with their devices from 255 users. The authors show that user interaction with the device contributes to higher battery consumption. Yu *et al.* [22] show that users' app usage behavior and dynamics in a given location can be predicted by using the point of interest (POI) information of that location. Authors in [63, 23] study application usage patterns of smartphone users and differentiate different groups of mobile users. Zhao *et al.* [23] apply K-means clustering to group Android mobile users application usage behavior and show that there are a more diverse set of mobile app usage patterns characterized by their age, income level, and demographics.

Canneyt *et al.* [64] study mobile users' app usage behavior by investigating users' engagement patterns with their apps. They use data collected using `Flurry` app analytics tool [65]. The authors show that mobile users' app usage activity and disruption patterns are correlated with major events such as sport and political events. Zhang *et al.* [66] study the characteristics of cellular data traffic based on HTTP-based traffic traces collected from cellular and fixed-line networks. They investigate different applications using packet, flow, and session-level traffic metrics in comparison with cellular and wire-line networks. They show that cellular networks have multiple short flows than wire-line networks. The authors cluster different applications based on similarity patterns observed using the inter-packet gap, flow, and session size as metrics. They indicate that the inter-packet gaps between different applications have a significant variation and suggest application dependent based optimization methods.

Oliveira *et al.* [67] characterize temporal data usage of mobile users using a dataset collected from a large group of people accessing the 3G networks. The authors classify the data usage into six patterns and show that mobile users can be profiled into daily peak and non-peak temporal data usage periods. Wei *et al.* [68] study the characteristics of network traffic generated from hand-held devices of users while accessing campus Wi-Fi networks. They show that the amount of traffic generated by users has a broader variation (from MB to several GBs), based on users' habit and demand. Qin *et al.* [69] and Wu *et al.* [70] use a dataset collected from cellular network operators in China, where the datasets are generated from mobile users while accessing different services. The authors characterized traffic patterns and application usage of mobile users to propose a model that predicts the traffic demand of mobile users.

Data usage patterns and user behavior in mobile networks have not been explored very well. Most previous studies are either limited to application usage pattern and identifi-

cation [71, 20, 21, 23]; focus on a single cellular core operator network and area [61, 72]; consider few number of users [24]; or is specific to application types [71]. Our work focuses on the clustering of mobile users' data usage patterns and behavior analysis based on ∼113 million data records collected from end-user devices. It covers the vast geographical and user-space in five different countries. We study by considering different network features that could determine data usage patterns of mobile users. The features include data traffic consumption, the number of sessions created per user, network congestion and coverage, user mobility, and the app types installed by users. Our study considers both data traffic flows and application types installed by mobile users.

6. Limitations

The dataset we have used has been collected from only Android mobile users as the current measurement platform does not run on other mobile operating systems (MOS). As a result, mobile users using other platforms such as on iOS and Windows are not considered in our study. As future work, it is important to study data usage patterns of mobile users accessing other than Android MOS and make a comparison among them. The demography of mobile users is not taken into consideration since we do not have the information. As another dimension of a study, it is possible to use the Netradar measurement platform augmented with demographic data. It can be possible to conduct data usage patterns with targeted mobile users of different user groups. For instance, whether mobile data usage varies by age, income level, and education status. Since it is a crowd-based measurement, the measurement might be limited to a certain group of mobile users who are curious and want to monitor their network performance. However, these groups of users are also essential as they are most likely active mobile users and could be affected by the quality of mobile network performance.

7. Conclusion

We studied mobile users' data usage patterns and behavior. We used a month-long dataset with more than 113 million measurement sessions collected from the crowd across five countries. We defined data usage patterns of mobile users by considering different factors such as network congestion, type of radio technology users have accessed (3G and 4G), user mobility, and the total bytes consumed per user. Using this definition, we applied an unsupervised clustering model to identify different types of mobile users based on their data traffic consumption and usage patterns. Our clustering model shows that there are three (heavy, regular, and light) different usage patterns across the five countries we studied, characterized by a small number of heavy users consuming the highest volume of data. We showed that there is a significant variance in the amount of data consumption and the number of sessions created across the different countries per similar group. We also showed that in some locations the type of apps installed by mobile users has a relationship with mobile users' data usage patterns and user groups. Finally, by using the clustered dataset as an input, we trained a classification model that helps classify the data usage patterns of mobile users with accuracy of ∼80% in the test dataset. The predictability of user behavior in mobile networks can be applied for optimal resource management based on users' data usage patterns.

References

[1] A. Gember, A. Anand, A. Akella, A Comparative Study of Handheld and Non-handheld Traffic in Campus Wi-Fi Networks, PAM.

[2] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, O. Spatscheck, An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance, SIGCOMM.

[3] G. Maier, F. Schneider, A. Feldmann, A First Look at Mobile Hand-Held Device Traffic, PAM.

[4] M. Z. Shafiq, L. Ji, A. X. Liu, J. Wang, Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices, ACM SIGMETRICS.

[5] Ericsson, Mobile traffic by application category., 2018.

[6] Cisco, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast, 2016.

[7] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, D. Malladi, A survey on 3GPP Heterogeneous Networks, IEEE Wireless Commun. 18 (2011).

[8] C. V. N. Index, Global Mobile Data Traffic Forecast Update, 2015–2020, Cisco white paper (2016) 9.

[9] T. Xia, Y. Li, Revealing Urban Dynamics by Learning Online and Offline Behaviours Together, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3 (2019).

[10] J. L. Toole, M. Ulm, M. C. González, D. Bauer, Inferring Land Use from Mobile Phone Activity, ACM SIGKDD, ACM, 2012, pp. 1–8.

[11] T. Italia, TIM Social and Chat, 2020.

[12] S. Bekhor, Y. Cohen, C. Solomon, Evaluating Long-distance Travel Patterns in Israel by Tracking Cellular Phone Positions, Journal of Advanced Transportation (2013).

[13] M. Lenormand, M. Picornell, O. G. Cantu-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frías-Martínez, J. J. Ramasco, Cross-checking Different Sources of Mobility Information, CoRR abs/1404.0333 (2014).

[14] U. Paul, A. P. Subramanian, M. M. Buddhikot, S. R. Das, Understanding Traffic Dynamics in Cellular Data Networks, IEEE INFOCOM.

[15] A. S. Asrese, E. A. Walelgne, V. Bajpai, A. Lutu, Ö. Alay, J. Ott, Measuring web quality of experience in cellular networks, volume 11419 of *Passive and Active Measurement - 20th International Conference, PAM* , Springer, 2019, pp. 18–33.

[16] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, H. Yan, Modeling Web Quality-of-Experience on Cellular Networks, ACM MobiCom.

[17] E. Boz, B. Finley, A. Oulasvirta, K. Kilkki, J. Manner, Mobile QoE Prediction in the Field, Pervasive and Mobile Computing 59 (2019) 101039.

[18] H. Shi, Y. Li, Discovering Periodic Patterns for Large Scale Mobile Traffic Data: Method and Applications, IEEE TMC 17 (2018).

[19] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, C. Sarraute, Measurement-driven Mobile Data Traffic Modeling in a Large Metropolitan Area, PerCom.

[20] C. Shin, J. Hong, A. K. Dey, Understanding and Prediction of Mobile Application Usage for Smart Phones, ACM Ubicomp.

[21] V. Srinivasan, S. Moghaddam, A. Mukherji, K. K. Rachuri, C. Xu, E. M. Tapia, MobileMiner: Mining your Frequent Patterns on Your Phone, ACM UbiComp.

[22] D. Yu, Y. Li, F. Xu, P. Zhang, V. Kostakos, Smartphone App Usage Prediction Using Points of Interest, IEEE IMWUT 1 (2017).

[23] S. Zhao, J. Ramos, J. Tao, Z. Jiang, S. Li, Z. Wu, G. Pan, A. K. Dey, Discovering Different Kinds of Smartphone Users Through Their Application Usage Behaviors, ACM UbiComp.

[24] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, D. Estrin, Diversity in Smartphone Usage, MobiSys.

[25] Netradar, Netradar , 2019.

[26] J. Huang, F. Qian, Z. M. Mao, S. Sen, O. Spatscheck, Screen-off Traffic Characterization and Optimization in 3G/4G Networks, ACM Internet Measurement Conference (IMC).

[27] D. Xu, D. Qian, A Bandwidth Adaptive Method for Estimating End-to-End Available Bandwidth, IEEE ICCS.

[28] S. S. Chaudhari, R. C. Biradar, Survey of Bandwidth Estimation Techniques in Communication Networks, Wireless Personal Communications 83 (2015).

[29] E. Boz, J. Manner, A hybrid Approach to QoS Measurements in Cellular Networks, Computer Networks 172 (2020) 107158.

[30] G. Developers, Android API - TrafficStats , 2019.

[31] E. A. Walelgne, A. S. Asres, J. Manner, V. Bajpai, J. Ott, Understanding Data Usage Patterns of Geographically Diverse Mobile Users, TNSM (2019).

[32] J. MacQueen, et al., Some Methods for Classification and Analysis of Multivariate Observations, in: Berkeley Symposium on Mathematical Statistics and Probability, volume 1.

[33] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, G. Cheng, Characterizing User Behavior in Mobile Internet, IEEE TETC 3 (2015).

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, CoRR abs/1201.0490 (2012).

[35] T. Hastie, R. Tibshirani, J. H. Friedman, The Elements of Statistical Learning: Data mining, Inference, and Prediction, 2nd Edition, Springer Series in Statistics, 2009.

[36] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On Clustering Validation Techniques, Springer Intell. Info. Syst. 17 (2001).

[37] A. K. Jain, M. N. Murty, P. J. Flynn, Data Clustering: A Review, ACM CSUR 31 (1999).

[38] C. H. Q. Ding, X. He, *K*-means Clustering via Principal Component Analysis, in: ACM ICML.

[39] P. J. Rousseeuw, Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, Springer CAM 20 (1987).

[40] T. Calinski, J. Harabasz, A Dendrite Method for Cluster Analysis, Communications in Statistics 3 (1974).

[41] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE TPAMI 1 (1979).

[42] J. C. Dunn, Well-separated Clusters and Optimal Fuzzy Partitions, Journal of cybernetics 4 (1974) 95–104.

[43] U. Maulik, S. Bandyopadhyay, Performance Evaluation of Some Clustering Algorithms and Validity Indices, IEEE TPAMI 24 (2002).

[44] G. W. Milligan, M. C. Cooper, An examination of Procedures for Determining the Number of Clusters in a Dataset, Psychometrika 50 (1985).

[45] D. Arthur, S. Vassilvitskii, k-means++: The Advantages of Careful Seeding, ACM SODA.

[46] opencellid, The World's Largest Open Database of Cell Towers, 2018.

[47] G. Developers, TelephonyManager: Android API, 2019. Accesed on April 09, 2019.

[48] Appbrain, Number of android apps on google play, 2019.

[49] D. A. Lyons, Internet Policy's Next Frontier: Usage-based Broadband Pricing, Federal Communications Law Journal 66 (2013) 1.

[50] S. Sen, C. Joe-Wong, S. Ha, M. Chiang, A Survey of Smart Data Pricing: Past Proposals, Current Plans, and Future Trends, ACM CSUR 46 (2013).

[51] C. Welch, AT&T CEO says 5g phone plans might be tiered and priced based on data speeds, 2019.

[52] P. Branco, L. Torgo, R. P. Ribeiro, A Survey of Predictive Modeling on Imbalanced Domains, ACM CSUR (2016).

[53] K. W. Bowyer, N. V. Chawla, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, CoRR abs/1106.1813 (2011).

[54] H. He, Y. Bai, E. A. Garcia, S. Li, ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, IEEE IJCNN.

[55] V. Bajpai, J. Schönwälder, A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts, IEEE COMST 17 (2015) 1313–1341.

[56] D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, Large-scale mobile traffic analysis: A survey, IEEE COMST 18 (2016).

[57] X. Liu, W. Ai, H. Li, J. Tang, G. Huang, F. Feng, Q. Mei, Deriving User Preferences of Mobile Apps from Their Management Activities, ACM TOIS 35 (2017).

[58] X. Liu, H. Li, X. Lu, T. Xie, Q. Mei, F. Feng, H. Mei, Understanding Diverse Usage Patterns from Large-Scale Appstore-Service Profiles, IEEE TOSE 44 (2018).

[59] E. A. Walelgne, A. S. Asrese, J. Manner, V. Bajpai, J. Ott, Understanding Data Usage Patterns of Geographically Diverse Mobile Users, IEEE Transactions on Network and Service Management (2020) 1–1.

[60] M. C. González, C. A. H. R., A. Barabási, Understanding Individual Human Mobility Patterns, Nature 453 (2008) 779–782.

[61] Y. Qiao, Y. Cheng, J. Yang, J. Liu, N. Kato, A mobility analytical framework for big mobile data in densely populated area, IEEE TOVT 66 (2017).

[62] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, H. Zhang, Human Mobility Patterns in Cellular Networks, IEEE COML 17 (2013).

[63] P. Welke, I. Andone, K. Blaszkiewicz, A. Markowetz, Differentiating Smartphone Users by App Usage, ACM UbiComp.

[64] S. V. Canneyt, M. Bron, A. Haines, M. Lalmas, Describing Patterns and Disruptions in Large Scale Mobile App Usage Data, World Wide Web Companion.

[65] Yahoo, Flurry App Analytics for iOS & Android, 2019.

[66] Y. Zhang, A. Arvidsson, Understanding the Characteristics of Cellular Data Traffic, ACM CCR 42 (2012).

[67] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, C. Sarraute, Mobile Data Traffic Modeling: Revealing Temporal Facets, Comput. Networks 112 (2017) 176–193.

[68] X. Wei, N. Valler, H. V. Madhyastha, I. Neamtiu, M. Faloutsos, Characterizing the Behavior of Handheld Devices and its Implications, Comput. Networks 114 (2017) 1–12.

[69] Z. Qin, F. Cao, Y. Yang, S. Wang, Y. Liu, C. Tan, D. Zhang, CellPred: A Behavior-Aware Scheme for Cellular Data Usage Prediction, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4 (2020).

[70] J. Wu, M. Zeng, X. Chen, Y. Li, D. Jin, Characterizing and Predicting Individual Traffic Usage of Mobile Application in Cellular Network, ACM Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, ACM, 2018, pp. 852–861.

[71] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, S. G. Rao, YouTube Everywhere: Impact of Device and Infrastructure Synergies on User Experience, ACM IMC.

[72] F. Xu, Y. Li, H. Wang, P. Zhang, D. Jin, Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment, IEEE/ACM TON 25 (2017).

1 List of Acronyms

Mobile Network Operators (MNOs)
Mobile Network Operators (MNO)
quality of service (QoS)
Mobile Country Code (MCC)
mobile operating systems (MOS)
Probe Gap Model (PGM)
Probe Rate Model (PRM)
Synthetic Minority Over-Sampling Technique (SMOTE)
Adaptive Synthetic (ADASYN)
Long Term Evolution (LTE)
Evolved High Speed Packet Access (HSPA+)
within cluster sum of squares (WCSS)
Sum of Square (SS)
Principal Component Analysis (PCA)
random forest (RF)

Ermias Andargie Walelgne is an experienced network and cloud developer at Ericsson and a doctoral student at Aalto University, Finland. He is advised by Prof. Jukka Manner and Prof. Jörg Ott. He received a Master's (2014) degree in Computer Science from the University of Trento, Italy, and a Bachelor's (2018) degree in Computer Science from Hawassa University, Ethiopia. He was a visiting Ph.D. student at Elisa (2017) and TU Munich ('17/'18/'19). His research focuses on understanding, characterizing, and modeling the performance of cellular data networks. He has been involved in the Marie Curie ITN project - METRICS research project.

Alemnew Sheferaw Asrese is a senior developer at Ericsson and a doctoral student at Aalto University, Finland. He is advised by Jörg Ott and Pasi Sarolahti. He received Master's (2014) degree in Computer Science from University of Trento, Italy and Bachelor's (2010) degree in Information Science from Adama University, Ethiopia. He was a visiting PhD student at TU Munich (2017), INRIA (2017), Simula Research Labs (2016). His research focus is on network measurement and QoE modelling. He has been involved in METRICS ITN, FP7 Leone and VENUS-C European research projects

Jukka Manner received his MSc. (1999) and PhD. (2004) degrees in computer science from the University of Helsinki, Finland. He is a full professor of networking technology at Aalto University, Department of Communications and Networking since 2008. His research and teaching focuses on networking, software and distributed systems, with a strong focus on wireless and mobile networks, transport protocols, energy efficient ICT and cyber security. He has been principal investigator and project manager for over 20 national and international research projects. He has authored over 150 publications, including eleven IETF RFCs.

Vaibhav Bajpai is a senior researcher at TUM, Germany. He received his PhD (2016) and Masters (2012) degrees in Computer Science from Jacobs University Bremen, Germany. He is the recipient of the best of CCR award (2019), ACM SIG- COMM best paper award (2018), and IEEE COM- SOC award (2017) for the best dissertation in net- work and service management. He is interested in future Internet protocols, web and video content delivery, network operations and management, and reproducibility of scientific Internet research.

Jörg Ott holds the Chair for Connected Mobility at Technische Universität München in the Faculty of Informatics since August 2015. He is also Adjunct Professor at Aalto University, where he was Professor for Networking Technology with a focus on Protocols, Services, and Software from 2005 until 2015. He is interested in understanding, designing, and building Internet-based (mobile) networked systems and services. His research focus is on network and system architectures, protocol design, and applications for mobile systems.

We have addressed all the comments raised by the reviewers and updated the manuscript accordingly. The diff of the paper and the final manuscript is attached to this response letter.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: