Understanding Data Usage Patterns of Geographically Diverse Mobile Users

Ermias Andargie Walelgne*, Alemnew Sheferaw Asrese*, Jukka Manner*, Vaibhav Bajpai† and Jörg Ott†
*Aalto University, Finland
(ermias.walelgne | alemnew.asrese | jukka.manner)@aalto.fi
†Technical University of Munich, Germany
(bajpaiv | ott)@in.tum.de

***Abstract*—The increasing trend of the traffic demand from mobile users and the presence of limited resources creates a challenge for network resource management. Understanding the data usage pattern and traffic demand of mobile users is a way forward to enable data-driven network resource management. However, due to the complex nature of mobile networks, understanding and characterizing data usage pattern of mobile users is a daunting task. In this work, we investigate and characterize data usage patterns and behavior of users in mobile networks. We leverage a dataset (~340M records) collected through a crowd-based mobile network measurement platform – Netradar – across six countries. We elucidate different network factors and study how they affect the data usage patterns by taking mobile users in Finland as a use case. We perform a comparison on data usage patterns of mobile users across six countries by considering total data consumption, network access, the number of sessions created per user, throughput, and user satisfaction level on services. We show that data usage behavior of users over a mobile network is primarily driven by user mobility, the type of data subscription plan marketed by Mobile Network Operators (MNOs), network congestion, and network coverage. Besides, the data usage patterns over different network technologies (e.g., preferring cellular over WiFi) and the percentage of users accessing congested networks vary by country; mostly due to the market pricing strategy and radio coverage. However, the overall data consumption (cellular and WiFi) is comparatively similar in most of the countries we studied.**

***Index Terms*—Network performance, Cellular network, Mobile user behavior, Data usage patterns, SLS, User satisfaction.**

I. Introduction

The advancements in mobile technologies and the need for ubiquitous communication by mobile users bring with them an increased growth in mobile data traffic. Mobile users regularly use data-intensive applications (video streaming and gaming) from their devices on the mobile network. Reports show that a significant share of Internet traffic generated from mobile devices increasingly consists of multimedia content [1], [2], [3], [4]. According to a report by Ericsson, in 2018, video content solely generated 60% of the mobile data traffic and is expected to cover 74% of the traffic by 2024 [5]. Furthermore, the projection shows that in 2022 the global mobile data traffic will be twelve times more than that of 2018 [6].

The mobile network is becoming a complex system to keep up with the ever-increasing demand for mobile traffic [7]. The increase in the traffic demand, the complexity of the network, and the number of connected users create a challenge for network resource management. The growth of mobile users and traffic demand also brings a challenge in understanding data usage patterns of mobile users for content providers. Service providers need to efficiently manage available resources based on data usage behavior of their customers. Studies also show that data usage patterns of different applications have a significant impact on energy consumption of mobile devices [4]. Understanding the data usage patterns and behavior of mobile users at different locations and market-places are paramount for service and content providers, and end-users. Mobile network operators can utilize the information to manage the increasing demand for mobile data usage [8], to plan and to optimize telecommunication resources [9], [10]. It can also be used to develop different data plan products [11] by targeting potential users. Similarly, policymakers and content providers will have more information to improve the quality of services, to understand urban dynamics [12] for improved urban planning [13].

In this paper, we characterize the data usage patterns and behavior of mobile users across six countries. We further investigate how data usage patterns vary by different factors including time of the day, user mobility, location, the device model, network performance (e.g., throughput and latency), and the coverage of radio technologies.

There are previous studies that investigate mobile user data usage behavior and have focussed on user location and user mobility patterns [14], [15], temporal dynamics [16], and Quality of Experience (QoE) [17], [18], [19]. However, most of the previous studies are either limited to a single operator [20], only target a specific city and location [21], study data usage behaviour targeting application types accessed by users [22], [23], [24], [25], or consider only a few measurement data and user spaces [26]. Unlike these studies, our work uses a large crowd-based dataset collected using Netradar [27] – a mobile network measurement platform. The dataset covers a wide range of geographical areas, Mobile Network Operators (MNOs), and mobile users.

The paper starts by laying the foundation for understanding the basic mobile network features related to data usage patterns of mobile users and their behavior. This includes investigating the relationship between data consumption (volume), session duration, user mobility, and the ratio of data consumption over WiFi versus cellular. The paper also provides a comprehensive view of the data usage pattern of mobile users in Finland in comparison with five other countries. Using a large-scale dataset (~340 million records) collected from six countries, this paper presents the following main **findings** –

First, we show that stationary users consume more data and are more likely to run into network congestion than users on the move. We observe that users with the latest device models have created more sessions, which is an indicator of active interaction of the users with their device. On the other hand, new devices with the latest Operating System (OS) version have relatively fewer total downloaded bytes when compared with devices that have older OS versions. This is due to frequently accessed content that is likely to be cached locally on newer devices with relatively better hardware specification such as higher memory capacity and processing power.

Second, we show that network throughput has a strong correlation with overall user data consumption. We also observe that different MNOs in the same country have an impact on the total data consumption of mobile users, as the network coverage and the maximum achievable speed could be limited.

Third, we investigate and compare user satisfaction levels on mobile network performance across six countries. If the data subscription plan is priced by data usage, then it is more likely that users reserve themselves from using applications that consume high data traffic. Service Level Score (SLS) is a method that measures satisfaction level of the mobile users. It considers the ratio of the number of times that users have received what they need from the network. The different value of SLS score across countries reveals that user satisfaction score can be higher if mobile users are conservative on how they use their mobile data. On the other hand, observing the total data consumption on both cellular and WiFi networks, we observe that mobile users across countries have a closely similar data consumption trend.

The goal of this paper is not to propose a new mining algorithm, but to study data usage patterns of mobile users across different locations using existing statistical algorithms. The paper is structured as follows: §II presents the measurement platform and the dataset used for the analysis. §III presents the study of individual network features contributing to the data usage pattern and behavior. §IV investigates the relationship between network performance and data usage patterns of mobile users. §V studies the data usage behavior of mobile users in six different countries. Finally, §VI discusses related work, and §VII concludes the paper. To encourage reproducibility [28], the dataset and scripts used in the analysis are publicly released to the community [29].

II. Methodology

A. Measurement Platform

Netradar [30] is a crowdsourced mobile measurement platform. It measures the link capacity of cellular networks using a hybrid of Probe Gap Model (PGM) and Probe Rate Model (PRM) [31] probe-based measurement methodologies. PGM and PRM utilize packet pair [32] probes to estimate the available bandwidth. For a detailed description of the measurement platform and its validation, we refer the reader to [33]. Going forward, we present a brief description of the measurement platform and the dataset relevant to this study.

The Netradar measurement platform passively listens on the ingress and egress traffic at the client side without imposing any synthetic traffic of its own. The application on the client device runs in the background until it is triggered when a user starts sending or receiving data. The application then starts sampling the traffic rate of the ingress and egress traffic (e.g., on Android using Android traffic Stat API [34]).

The measurement is recorded based on sessions. A session refers to the continuous traffic flow of content between a user device and a remote server. A given session has a duration, within which different ingress (and egress) traffic can flow, but the application does not have any visibility on the type of its content. A session duration is defined as the interval between the starting time of the sampling phase until the traffic stops. The session starts when there is enough traffic flow in either the uplink or downlink direction and ends when the traffic rate goes below half of the threshold for two seconds. The threshold for starting the session is 100 Kbps and 200 Kbps for uplink and downlink direction respectively. The duration of the session can be from milliseconds to minutes long, while sessions that have very few bytes transferred are not recorded.

The platform also records unconstrained and constrained speeds of the network. An unconstrained speed is the speed that the user needs from the network to use mobile apps, but at the same time also does not hit the maximum speed of the network during the session. On the contrary, the constrained speed is the maximum speed recorded when the network was a limiting factor. It is inferred based on the queuing delay of packets, the available bandwidth, and latency. The constrained speed is not recorded when there is no latency information, or when a user never hits the maximum speed of the network.

A given session may have only uplink or downlink data recorded. For instance, if the user is watching a video on YouTube, most of the sessions are on the downlink. As such, there are few traffic flows that are not statistically significant to keep records related to the uplink information. On the other hand, when a user is uploading a picture to Facebook, then most traffic flows are in the uplink direction instead.

Every session has a unique identifier with its own starting and ending time and other metadata information related to the session. Each measurement session consists of the following meta information: device information, MNO of the subscriber, location, user velocity, and installation (user) ID. The metadata also records information about network type (WiFi or cellular) and accessed radio technologies (2G, 3G, 4G) with radio Quality of Service (QoS) values such as Reference Signal Received Power (RSRP) that help to decide handover or cell re-selection. The constrained and unconstrained speeds for both download and upload with respective session length are also recorded. Besides, every session also contains the average download and upload speed, total uploaded bytes, total downloaded bytes, latency, battery level, signal strength, and information about the base station (e.g., cell ID, area code, radio frequency channel number). Every session has associated tile information (e.g., tile-ID, country, city, population density, postal code), where the tile-ID is the area coverage of 100 by 100 square meter.

TABLE I: *Number of users and sessions created in the cellular and WiFi network separated by country.*

Country	# of sessions (M)		# Users	
	Cellular	WiFi	Cellular	WiFi
Finland (FI)	35.1	26.0	22795	16200
Germany (DE)	6.3	17.8	6548	6381
United Kingdom (UK)	34	90.4	20569	20927
Japan (JP)	19.8	50.1	8081	8583
Brazil (BR)	17.8	42.9	7164	8368
India (IN)	2.5	5.8	1665	1786

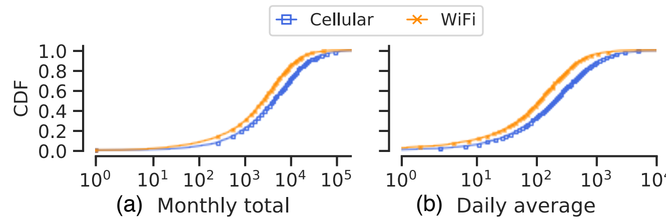


Fig. 1: *Monthly and daily download amount (MB) per user. Mobile users in Finland consume more data over cellular network than WiFi.*

B. Dataset

The dataset used in the analysis has been collected using Netradar from the devices of mobile users in six different countries (Finland, Germany, the United Kingdom, Japan, Brazil, and India) for one month (recorded in July 2018). Mobile users are assigned to the respective countries based on the Mobile Network Code (MNC) and Mobile Country Code (MCC) values. In other words, a user in a given country has to be a subscriber of one of the MNOs in that country and should access the network from within the same country. In this study we do not consider roaming [35] users. Table I summarizes the number of users and sessions created per respective countries in both cellular and WiFi networks.

Note, unless specified by the name, 4G refers to all releases of Long Term Evolution (LTE) radio technology. 3G refers to all other releases of radio technologies prior to LTE (including High Speed Packet Access (HSPA), Evolved High Speed Packet Access (HSPA+), and Universal Mobile Telecommunications System (UMTS)). 2G refers to releases before UMTS (such as General Packet Radio Service (GPRS), and Enhanced Data rates for GSM Evolution (EDGE)). As such, cellular network stands for all of the aforementioned cellular radio technologies. WiFi refers Internet connectivity through Wireless Local Area Network (WLAN) including the variants of IEEE 802.11 protocol standard.

In §III and §IV, we present analysis based on measurement data collected from Finland. We choose Finland as we have sufficient measurement samples and we understand the mobile market in Finland better than in other countries. As such, Finland is first presented as a case study to investigate features with respect to data usage patterns in mobile networks. Later, in §V, based on our previous observations, we compare the data usage patterns of mobile users across six countries by considering various features related to mobile networks.

III. Data usage behavior in Finland

We investigate data usage patterns of mobile users in terms of the number of created sessions, average session duration, and amount of traffic flows in both cellular and WiFi networks. We also study the association of temporal dynamics, users mobility, and device models to data usage behavior.

A. Data Volume and Sessions

As shown in Table I, during a month-long measurement period in Finland, there are more than 61M sessions created from ~22.7K users that were accessing cellular and WiFi networks. Out of these, ~35.1M and ~26M of the sessions are created when the users were accessing the Internet over cellular and WiFi networks respectively. From all the cellular-based dataset collected in Finland, 80% of the sessions were established over LTE; 16% over HSPA+, and the remaining 4% of the sessions were created over other radio technologies.

Data volume – The intensity of data consumption by mobile users can be indicated by considering the total traffic flow (of bytes) through individual devices. Fig. 1 shows the monthly (total) and the daily (average) consumed bytes per user in Finland over cellular and WiFi networks. Observing at the 95^{th} percentile, monthly data consumption of cellular and WiFi users is less than 42.5 GB and 19.6 GB, respectively. For mobile users in Finland, the total data consumption over cellular networks is about two times more than that over the WiFi network. For instance, considering the median daily case, cellular users consumed more data (3999 MB) than WiFi users (2406 MB). We compare the numbers of unique users that downloaded more than 10GB in a month under each cellular and WiFi network. We found that 27.5% and 14.5% of the users that have accessed cellular networks and WiFi networks, respectively have consumed more than 10 GB. The daily average download data consumptions per user over cellular and WiFi networks show a similar trend as the monthly total download consumption, as shown in Fig. 1 (b). For instance, the median daily consumption in a cellular and WiFi network is 195.2 MB and 110.6 MB, respectively.

The higher data consumption over a cellular network than over a WiFi network shows that mobile users in Finland mostly prefer to access the Internet over a cellular network and thereby tend to consume more data over a cellular network than over a WiFi network. A possible reason for this observation can be related to the availability of good mobile network coverage and the flat-rate data subscription plans at an affordable price [36]. This likely encourages mobile users to access the Internet over the cellular networks for most aspects and they tend to remain connected to the Internet over cellular for a longer time [37]. According to the survey taken from 2011 to 2017 in Finland [38], the number of mobile data subscriptions with unlimited data plans have increased while all other data subscription types have decreased during the given time frame. For instance, in 2017, the number of users with unlimited data was around 71.2%, while all other data subscription types covered only 28.8% of mobile users.

Sessions – The number of sessions created per user and

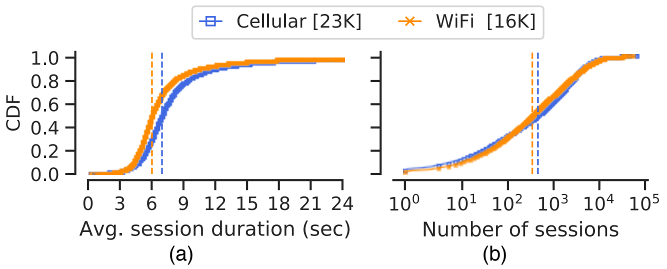


Fig. 2: *Distribution of session duration and number of sessions as observed per user. The vertical line refers median value. Both the session duration and the number of sessions created over cellular and WiFi network has only a marginal difference.*

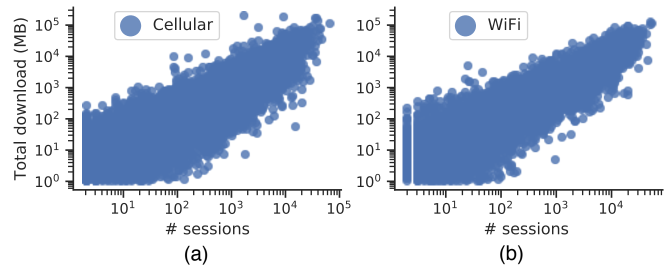


Fig. 3: *Total consumed bytes compared with the number of session per user per hour. There is a strong relationship between users' interaction with their device (the number of sessions created per user) and the total data consumption.*

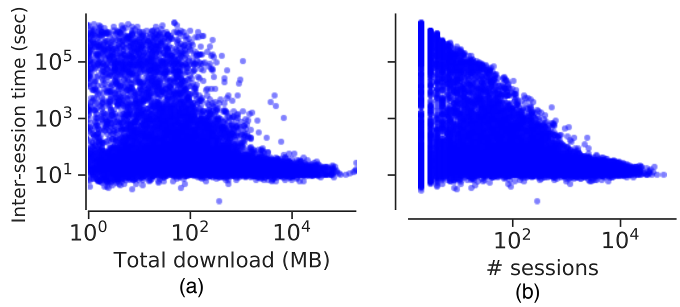


Fig. 4: *Inter-session gap compared with # of sessions (b) and total download bytes (a) per user. Users with lower inter-session gaps have higher number of sessions and data volume.*

their session duration can both be used as an indicator for data usage activity and interaction of users with their devices.

To observe the average session duration per user, we grouped every session by user and calculated the average session duration. Fig. 2 (a) shows the distribution of the average session duration per user over cellular and WiFi networks. We observe that in the median case, cellular and WiFi-based measurements have a marginal difference in the length of the sessions. The median session duration is ~7 seconds in cellular and 6.0 seconds in WiFi networks. The variance in session duration is higher over cellular network (variance (v) = 46.2 $seconds^2$, mean (m) = 8.4 seconds and standard deviations (Std. dev.) = 6.8 seconds) than WiFi (v = 31.4 $seconds^2$, m = 7.2 seconds and Std. dev. = 5.6 seconds). The higher variation in session duration of users in the cellular network reflects bursty arrival pattern of packets [39] and also the range of different (from heavy to light traffic demanding) application types that users could use. In both cellular and WiFi networks, the life span of majority of the sessions are relatively shorter. For instance, considering the 95^{th} percentile, the session duration in cellular and WiFi networks are 16.4 seconds and 14 seconds, respectively. This result is in line with the previous study [40]. The authors showed that in more than 90% of the cases, the sessions generated from Facebook and WhatsApp apps for both multimedia and text content are less than a minute.

We also study the number of sessions created per user. Fig. 2 (b) shows the distribution of the number of sessions created per user in a month for both cellular and WiFi network. The daily median number of sessions per user is 120 and 116 in cellular and WiFi networks, respectively. The total amount of bytes downloaded, the number of sessions created per user, and the session duration confirm that mobile users in Finland prefer a cellular network for most of their Internet activity over the WiFi network.

Furthermore, we study the relationship between the total download bytes and the number of sessions created per user. Fig. 3 shows the total downloaded bytes compared to the number of sessions created per user for both cellular and WiFi networks. We observe that there is a strong positive relationship between the total downloaded bytes and the number of sessions created per user (Pearson correlation coefficient (r) = 0.69 and r = 0.801 over cellular and WiFi networks, respectively. On the other hand, there is a moderately positive correlation between the total download bytes and the session duration (Pearson r = 0.312 and r = 0.401) in cellular and WiFi networks, respectively). Note that, in all of Pearson's correlation coefficient, the p-value was significantly lower than 0.05. We observe a similar trend in the relationship between total uploaded bytes and the number of sessions created per user (not shown in the plot).

When investigating the session duration and the total downloaded bytes per user, we observe that cellular users in Finland have bursty usage (median 21.3 Mbps) which lasts from 0.2 seconds to 6.5 seconds. About 4.7% of users continuously access the network (median 12 Mbps) for a session duration of more than a minute. Only 0.35% of the sessions have a duration longer than two minutes, of which the majority of these sessions (~68%) were accessing the cellular network.

Fig. 4 shows the scatter plot of inter-session time gaps compared to the number of sessions and total downloaded bytes per user. We observe that inter-session time has a strong negative relationship with both the number of sessions and total download bytes per user (Spearman correlation coefficient (r) = -0.524, p-value = 0). As defined earlier, a session creation starts when there is a flow of at least five IP packets and stops when there is no traffic flow for two seconds in either uplink or downlink direction. The inter-session gap per user indicates the regularity of such sessions and data transferring behavior of applications. Some video streaming applications download a chunk of content every few seconds, where the length of the download depends on how much content is needed to fill

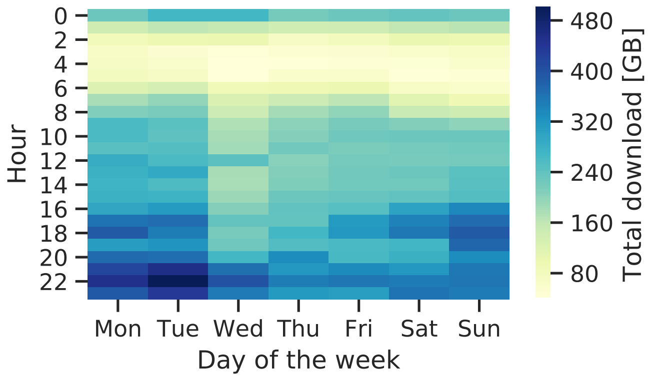


Fig. 5: *Heatmap of total downloaded bytes per user over cellular network. High volume of data consumption is observed during "off-duty" time for most of the days.*

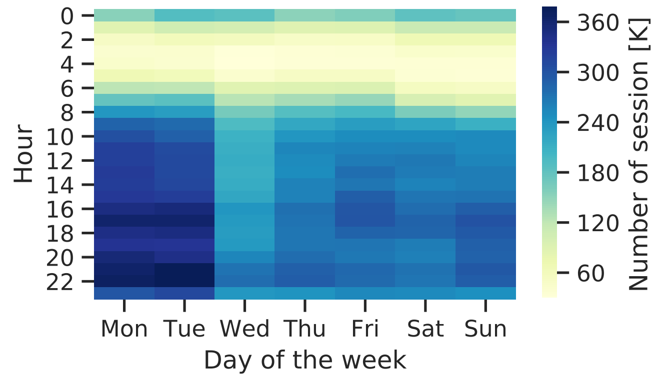


Fig. 6: *Number of sessions in a cellular network. Monday and Tuesday show the highest user activity, following a similar trend with data consumption of users.*

the players buffer size [41]. A study on Netflix streaming by Adhikari *et al.* [42] shows that once the player buffer has filled, the subsequent download events happen at about every four seconds interval. Siekkinen *et al.* [41] also identified that YouTube player in Galaxy SIII device has an "off" period of 60 seconds after it fills the players' buffer. Similarly, let us assume that a streaming application schedules the download for every x seconds, for x greater than the two seconds idle time. In that case, if there is no other active traffic during the x second scheduled time, the session will stop as it would pass the two seconds ideal time. Immediately after the scheduled time is over, new traffic will show up at the ingress, and a new session will start recording. As a result, such category of applications will have shorter inter-session gaps, but higher download bytes and number of sessions.

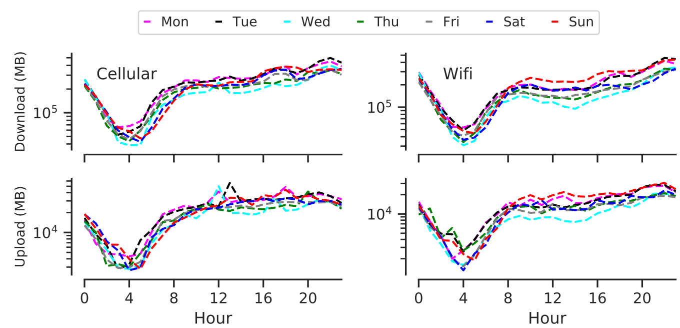


Fig. 7: *Data consumption by days of the week. Over the cellular network, weekdays appear similar to weekends, while data consumption during weekends is relatively higher (to weekdays) over the WiFi networks.*

B. Temporal Dynamics

We study how mobile data usage of individual users varies over time of the day. Understanding the temporal dynamics of data usage patterns can be a useful input for efficient network resource management such as for energy saving. For instance, MNOs can utilize temporal usage pattern of peak or off-peak hours and temporarily turn off some of the transceivers or even the entire base stations to save energy [43], [44]. In addition, application developers and maintainers can also adjust the time of app updates (e.g., for apps where the auto-update option is enabled) during a period of low traffic.

Fig. 5 shows the total downloaded bytes per hour by users every day of the week. The color bar on the right side shows the total bytes downloaded by all users at every hour in the given day. We can observe that especially during day time (from 8:00 to 22:00), there is more data consumption in most of the weekdays. Whereas, after midnight the total download bytes are scaled-down, as people go to bed. We study the hourly (total) downloaded bytes per user over 24 hours in a month for mobile users in Finland and observe that mobile users preferred cellular networks over WiFi in 100% and 83% of the time for uploading and downloading, respectively. For mobile users that prefer WiFi over cellular, the majority accessed WiFi networks during the working hours (14:00 to 17:00) than during off-working hours (at home).

Fig. 6 shows the total number of sessions created by users (as a heat map) per time of a day in every week day. The sessions were created while users were accessing a cellular network. The higher number of sessions during peak hours suggest that users were frequently interacting with their devices (especially on) Monday, and Tuesday.

The (total) downloaded bytes during the weekdays have a closer similarity to the weekends over cellular networks as shown in Fig. 7. However, downloaded data consumption per user during weekends is comparatively higher than during weekdays over the WiFi networks. For instance, Sundays have a higher (total) download consumption over WiFi at most of the times of the day. We suspect that during weekdays, users maybe spend most of their time indoors and tend to use WiFi more often than the cellular network.

We also explore whether users prefer downloading or uploading content over a certain time of the day. To do this end, for every user, we calculate the hourly average values of the total downloaded (and uploaded) bytes transferred during the month. We then calculate the difference between the total download and upload bytes over time of a day, *i.e.*, $\Delta Bytes(MB) = TotalDownload - TotalUpload$. We

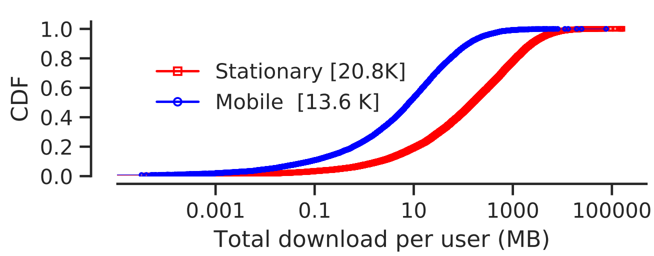


Fig. 8: *Distribution of total downloaded bytes per user. Stationary users consume more data than users on the move.*

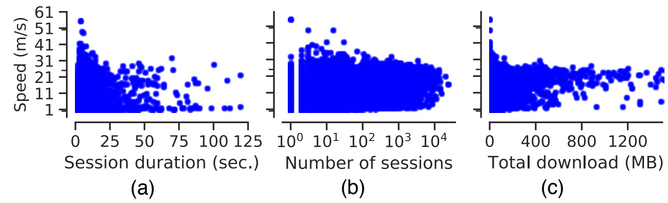


Fig. 9: *User velocity (m/s) compared with the number of sessions, the average session duration, and total download bytes. There is a moderately positive relationship between velocity and the number of sessions created by users.*

observe that about ~6% of the time, the total upload is greater than the total download bytes. During night time (from 3:00 AM to 5:00 AM) the difference between download and upload bytes are marginal (50% of the time, the difference is close to 0). On the other hand, during the day time, the difference between download to upload over the time of the day follows a similar trend. For instance, from 12:00 to 23:00, about 25% of the measurement have less than 1MB difference.

C. User Mobility and Data Usage

We further study under which mobility conditions (when stationary, walking or commuting in a train for instance) users usually consume data. We focus on the measurement sessions conducted only over cellular networks due to their increased availability in such mobility scenarios.

We use the velocity of a user (in m/s) over the ground inferred from the OS of the device. We filtered data that has a GPS location accuracy of less than 100 meters. As the location accuracy relates to the deviation in meters, the lower the number, the better the accuracy of the location and the velocity [45]. Since the current maximum commuting velocity in Finland is ~220 km/h, the data is filtered further such that the maximum velocity is not over 220 km/h. We then divide the dataset into mobile and stationary users based on the velocity value. In our case, mobile users are users whose velocity is greater than or equal to 1 m/s. Stationary users are users whose velocity is precisely 0 m/s. Here, to avoid variance, we do not consider measurements where the velocity is unknown. After filtering the data, users accessing their device on the move have the median velocity of 41.8 km/h, while only 5% of the values are above 88.2 km/h.

Fig. 8 shows the total downloaded bytes per user for stationary and mobile users. We observe that the ratios of cellular users at a stationary location are almost about 1.5 times greater than users that were moving from place to place (20K users at a stationary location vs. 13K users on the move). We also compared the median download speed that the users get when they are moving and when the are stationary. We found that stationary users get higher download speed (twice faster) than users on the move. Similarly, stationary users consume more data (median download is ~76MB) than moving users (median download is ~6MB).

Fig. 9 shows the relationship between user mobility (based on velocity) and the average session duration (a), the number of sessions created (b), and the total download bytes consumed per user (c). The number of sessions created by users and the velocity has a positive correlation (Spearman r = 0.57, p-value = 0.0). This indicates that users in a commuting train, tram, bus or as passengers in a moving vehicle, etc. spend most of their time interacting with their smartphone. Note that, the positive relation might not directly infer the quality of the network (e.g., throughput) users are getting while on the move. Instead, the relation is indicative of how often users are engaged with their device. We found that throughput and velocity are negatively related (Spearman r=-0.29, p-value=9.059e-204). This has also been noted in previous studies such that as velocity increases the throughput could degrade due to reasons such as frequent handovers, delays in connection establishment, packet loss, and signal interruption [46], [47]. The positive relationship between number of sessions and velocity values also indicate that users walking on foot or riding a bicycle are less likely to use their smartphone. On the other hand, as long-distance trains and metros move at high-speed, mobile users on such a commute will spend more time using the Internet with their devices; although the throughput might decline as mobility increases. As the number of sessions created per user is strongly related to the total download bytes (see: §III-A), the velocity should also have a positive relationship with total download bytes (Spearman r = 0.42, p-value = 0.0). When we consider measurements with a velocity greater than 0 m/s, users sitting in a commuting train or a bus (with relatively higher velocity but in a comfortable position) will interact with their mobile device more often than users walking or bicycling (with relatively with lower velocity). In terms of total download and data consumption, we have observed high data volume even in the lower tail of velocity. This could be related to the type of application users might have accessed; for instance, users might stream music while walking or bicycling.

On the other hand, session duration and velocity are negatively related (Spearman r = -0.17, p-value=8.69e-202). Similarly, the average session duration and the number of sessions created per user are also negatively related (Spearman r = -0.26, p-value = 8.69e-202). This can be associated with the presence of frequent handovers as users are on the move. At a higher velocity, users are most likely forced to leave the current serving cell and join the next target base station. As a result, it is possible to observe several number of sessions with a shorter life span.

User mobility can also be inferred from the number of base stations visited per user. The typical coverage of a cellular

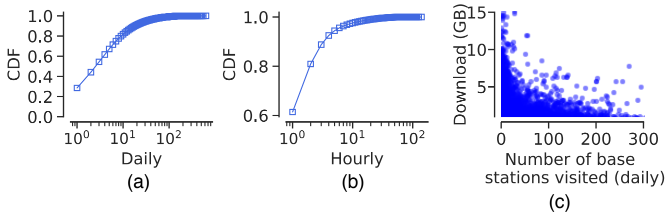


Fig. 10: *Unique number of base stations visited by users. Note the scale difference on y and x axis. About 28% of the users remain within a single base station coverage per day.*

base station tower is in the range from 200m to 1km in urban areas and from 1km to 5km outside of the city areas [48]. In our case, the base station is identified as a combination of MCC, MNC, Location Area Code (LAC), and Cell ID (CID) [49]. The number of different base stations visited by users in a given time interval can be used to infer how far users travel from one place to another location. For instance, if the user travels from home to the working place every day, there is a high probability that this user is going to visit more than one base stations. We consider the frequency of unique base stations visited by the mobile users per given time to estimate how often the users are moving from one place to another place. This is useful to understand how many times the mobile users move outside the coverage area of a given base station and to investigate the relation it has with data usage patterns. Knowing the patterns of the dynamics of mobile users and the associated peak/off-peak traffic hours can be useful input. For instance, operators could dynamically configure the coverage area of the cell based on the traffic demand and data consumption patterns. In addition, operators can utilize such type of information to adjust the transmission power of the base station during the off-peak hour to save energy [50].

Fig. 10 (a and b) shows the number of unique base stations visited by users at each day and hour, respectively. This shows the mobility of users per day while accessing cellular networks. Observing the total distribution of the measurement data, in most of the days, user mobility while accessing cellular networks is relatively limited to certain locations. For instance, in the median case, users visited not more than three different unique base stations per day. We observe that ~35.6% of users visited more than one base station per day, of which, 15.4% users had only a single location change. On the other hand, observing the daily and hourly change of base stations, ~28.5% and ~61.3% of the users, respectively, stay within the coverage of a single base station. Users who visited more than one base station per day have an average downlink speed of 130.2 KBps. Users who stayed at the same base station per day have an average downlink 133 KBps, which asserts that stationary users have a chance to get higher download speeds than movable users. Fig. 10 (c) shows the number of unique base stations visited by mobile users in Finland and its relation to the total downloaded bytes while accessing cellular networks. We observe that although there are few positively skewed distributions, majority of the mobile users who download content from their devices are stationary users.

TABLE II: *Device group used in FI based on market price*

Device group	Price range (€)	# Users	# Devices	# Sessions
Low-end	<=200	4748	112	6.2M
Mid-end	200 − 400	9316	147	11M
High-end	> 400	1200	32	2.1M

Previous studies, such as [51], have designed an algorithm to detect homes and working places using a dataset collected from an MNO. Tagging user location as "home" and "workplace" can also be achieved by observing frequent patterns of users' location and time of the day [14], [15]. Considering the nature of our dataset, we follow a similar approach to tag the location of mobile users as home and in the workplace. For each user working location is the most frequently visited address (base station) during the working hours. For each user, "working" location is the most frequently visited address (base station) during the working hours (9 AM to 4 PM). While, "home" location is the most frequently visited base station during the night time, as people would likely stay at their home during off working hours. We consider only measurements conducted at the stationary location. Accordingly, we found that in the median case, users at home consume the highest amount of data from their phone (51.7MB) than users at their working place (40.2MB).

D. Device Model Types and OS Versions

We further ask whether the choice of the device model impacts the data usage pattern of the users. Device manufacturers set the market price based on the device specification and their target user groups. However, the price of the devices might vary in different markets and usually decline over time. Knowing the release year of the devices to the market and their market-price value can be an indicator to identify potential user groups. By considering the device model, the OS version, and the year of release of the devices, we study whether the session length, the number of sessions, and the amount of data transferred vary per device used by the users.

We began by grouping device models by their brand name. Then, for every brand family, each specific model was mapped to the year of release and average market-price of the device. The device model, year of release and the average market-price of each device was extracted using the GSMA arena [52] service. Note that, the device model and the OS version of devices have already been inferred programmatically during the measurement. After filtering the dataset, we follow a heuristic approach to determine the price range. We found a total of 295 different device models in our measurement dataset from Finland. As shown in Table II, we categorized the device by price range. The table shows the number of users, the number of sessions, and the number of device models belonging to each category.

By considering the price range of the devices, we investigate three different groups of devices and the corresponding total downloaded bytes per user. We found that users with a high-end device model consume data twice more than mid-end and low-end device model users as shown in Fig. 11 (a).

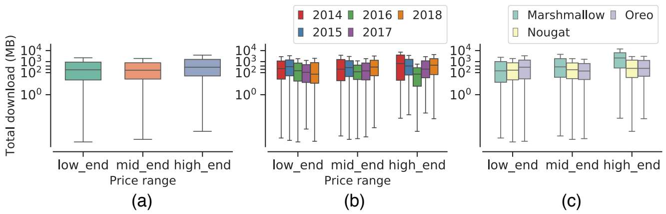


Fig. 11: *Total downloaded bytes per user categorized by price level (a), and price level per year of release of the devices (b) and OS version name (c). The data consumption over high-end device models is twice more than that of mid-end and low-end device models.*

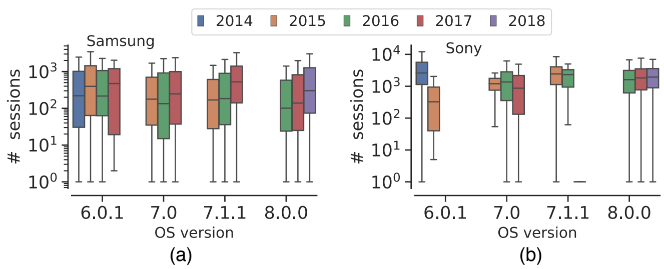


Fig. 12: *Number of sessions created by user per Android OS version for Samsung and Sony mobile phones. The color shows the year of release of the devices to the market. Users with recent device models have created more number of sessions.*

The total downloaded bytes for high-end, mid-end, and low-end devices in the median case is 50.3MB, 26.3MB, and 21.8MB, respectively. This is possibly explained by the type of applications installed and the data usage behavior of the user. For instance, users with a high-end device might tend to use "heavy apps" that consume a lot of data. One reason for this observation could be that the economic status of users might allow them to have a good data subscription plan. Besides, high-end devices usually have a larger screen size. As a result, users might use their smartphone for most of their daily activities on the Internet.

Fig. 12 shows the number of sessions created by users per Android OS version type for Samsung and Sony models. We observe that for Samsung models and under the same Android OS version, users who own recently released devices have more interaction with their smartphone than previously released device models. For instance, for Android version 8.0, device models released in 2018 have created the highest number of sessions compared with device models released in 2017 and 2016. The observation is consistent in most of the OS versions for Samsung devices but not for Sony, except for version 8.0. This inconsistency between the two devices models could be related to the background traffic generated by the device and the corresponding utilities.

We witness that the recent Android OS versions have relatively smaller total downloaded bytes per user than older OS versions. As shown in Fig. 13 (a), Oreo Android OS version has total downloaded bytes of 167.7 MB per user.

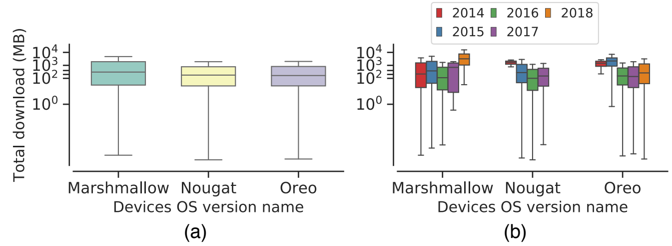


Fig. 13: *Total downloaded bytes per user categorized by Android OS version and year of release of the devices (LTE network). New devices with the latest Android OS have relatively less total downloaded bytes than similar device models (based on year of release) but with older OS version.*

Oreo[1] was the recent Android OS release compared to the other two versions in our measurement. On the other hand, devices based on Nougat and Marshmallow versions in the median case have 177.9 MB, 303.3 MB total download bytes, respectively. A possible reason for the latest OS version to have the least total downloaded bytes per user is that the OS might come with an improved and an optimized mechanism to avoid bulky downloads from the server for every content the user requests. Similarly, new devices come with higher memory size, storage capacity, and processing power. As a result, some frequently accessed pieces of information can be cached locally to minimize the size of the content to be downloaded directly from the servers. This can be observed in Fig. 13 (b), where latest devices (based on the year of release) with the latest Android OS version have relatively smaller total download bytes than older devices with the latest Android OS version. For instance, devices released from 2016 to 2018 and running Oreo version has smaller total downloaded bytes per user than models released before the year 2016.

Takeaway – Mobile users in Finland prefer cellular networks over WiFi due to good network coverage and flat-rate subscription prices. Stationary users and users at home consume more data than users on the move and users at work. Over the cellular network, data consumption in Finland over weekdays appears similar to weekends, while data consumption during weekends is relatively higher over the WiFi networks. Users with high-end devices (based on price range) have the highest total downloaded bytes and users with recent devices (based on year of release) have more interaction with their devices. High-end devices with the latest Android OS version have smaller total downloaded bytes than similar device models but with older Android OS version. Such high-end devices come with an improved hardware capacity to cache frequently accessed content reducing bulky downloads whenever possible.

[1]Using the associated name of Android OS release versions, we map the version number with the corresponding name based on the information [53]. Accordingly, release versions from 6.0 to 6.0.1 are known as Android Marshmallow, and versions from 7.0 to 7.1 named as Android Nougat and versions from 8.0 to Android 8.1 known as Android Oreo. There are other Android versions in our measurement. However, we select these three versions since we have more than 4M measurements for each of these OS versions.

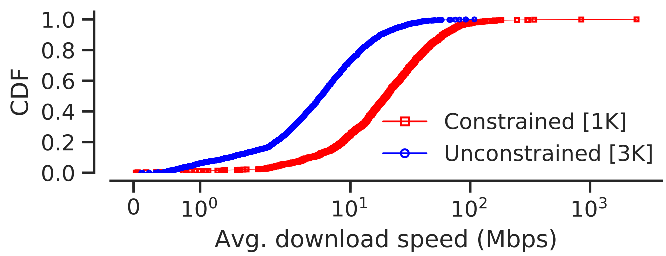


Fig. 14: *The constrained speed (which is related to the maximum network speed and created when the network is a limiting factor) is higher than unconstrained speed (when the network was not a limiting factor).*

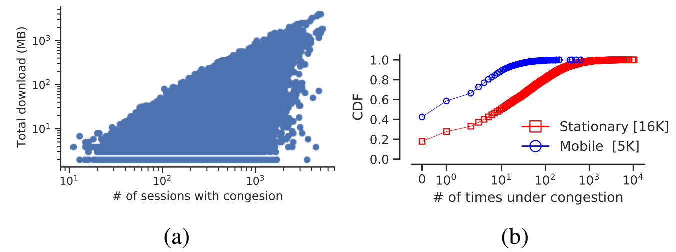


Fig. 15: *Number of congested sessions by stationary and mobile users. Stationary users are more likely to access the network under congestion than users on the move.*

IV. Network Performance and User Behavior

In this section, we investigate the impact of network congestion and performance such as throughput [54] and latency [55] on mobile data usage patterns.

A. *Network Congestion and Data Usage Patterns*

The platform records constrained and unconstrained speeds (see: §II for definitions) indicating presence of possible congestion [56] and non-congested networks, respectively. If there is a latency value recorded during the session, it is possible to isolate the congestion events (if any). As such, we only consider the sessions that have the latency data recorded. Fig. 14 shows the average daily download speed per user when they access the network with constrained and unconstrained conditions over the LTE network in the Helsinki area. The presence of constrained speed in a session is related to the existence of congestion in the network. Note that the idea of whether network congestion happened during measurement or not is loosely defined. In this case, the network was a limiting factor. Hence, the throughput demand of a given user is beyond what the current network can offer. Unconstrained speeds are all throughput values under normal network condition. Constrained speed is always recorded for the highest peak bitrate, where the network is a limiting factor. As a result, the average download speed recorded during constrained network is always higher than the unconstrained network.

On the contrary, unconstrained speed is related to the normal usage of mobile users without demanding more than or closer to the maximum network speed. This is because users often need lower download speed than what the network can provide, and they usually get the speed they require from the network (as shown in the blue line). For users accessing the network with unconstrained speed, the base station was not a limiting factor for poor user experience (if there is any). However, the user might be limited by servers or by the quality of the application in use. Users under the constrained speed network (red in the plot) hit the maximum speed of the network. This is because since such users are mostly transmitting more data traffic, there is a high probability that these users require more bandwidth than what the network can provide. We also observe that the constrained download speed (which indicates presence of congestion on the network) has higher variance than the unconstrained download speed (when users are not limited by the network).

For instance, the total downloaded bytes per user between three major MNOs in Finland have a median difference from 0.4 MB to 1 MB. The one-way ANOVA (F-statistic = 71.35 and the p-value = 1.03e-31) confirms that there is an overall significant difference among MNOs. The effect of MNOs on the total download (under the same radio technology) can be explained by the data subscription plan users have and the available network coverage (i.e., how good the infrastructure of a given MNO is). Since we do not have information regarding the data subscription plan of the mobile users, inferring a conclusion in this direction is difficult. We, therefore, study how network availability and coverage impacts the number of total downloaded bytes per user and the SLS values. Note that, SLS is a measure for user satisfaction level when they access the network (see: §V-C for more details). We notice that the availability of network coverage (e.g., 4G network) at different locations also varies by MNO. We observe that users subscribed to a given MNO with higher 4G availability, have higher SLS and total downloaded bytes than the MNOs with lower 4G coverage (plot not shown). This indicates that users subscribed to different network operators have different data usage experience and traffic demand.

Fig. 15 (a) shows the downloaded bytes versus the number of sessions with constrained speed per mobile user. The plot depicts that the presence of constrained speed (with possible congestion) limits the total downloaded bytes users could get, irrespective of the number of times the users are interacting with their device (the bottom of the x-axis). On the other hand, as the number of sessions increase the maximum amount of downloaded bytes are observed. This indicates that users are hitting the maximum network speed as they aggressively use the network. We also study the probability of users running under congestion when they move around versus when they are stationed at a single place during the measurement period. We observe that stationary users access a mobile network under congestion more than users on the move as shown in Fig. 15 (b). This result is in-line with the high amount of data consumption by stationary users (see: §III-C). In §V, we present more detailed analysis on the impact of congestion on the data usage pattern and its distribution by country.

B. Network Throughput

We study whether throughput affects the total data consumption (both download and upload) by taking the average throughput per user. Fig. 16 shows the correlation between total downloaded and uploaded bytes with download (a) and upload (b) speeds, respectively. The values represent the median value per user. We observe that both the total downloaded and uploaded bytes are positively correlated with download and upload speeds, respectively. For instance, evaluating the Pearson and Spearman correlation coefficient, the total download consumption per user and downloading speed (Pearson r=0.43, Spearman r=0.9 with p-value=0.0) and the total upload consumption per user and uploading speed (Pearson r = 0.66, Spearman r=0.8 with p-value=0.0) are strongly correlated. This indicates that network throughput contributes to higher data consumption trends, especially for countries such as Finland, where unlimited data plan subscription is widely adopted [38]. We also observe that the median throughput has a slightly positive relationship with the number of sessions created per user (Spearman r = 0.13, p-value = 1.48e-82), but does not have a meaningful relationship with the session duration.

Takeaway – The network throughput has a strong correlation with total downloaded bytes. Throughput is weakly correlated with the number of sessions created per user but has no meaningful relationship with the session duration. Stationary users have more probability of experiencing congestion in the network than users on the move. The constrained speed which is the maximum network speed recorded when the network was a limiting factor indicates possible presence of congestion. Such a condition indicates that users need more throughput than what the network can provide. When the network is constrained, the highest download speeds are recorded. However, as the mobile network is a limiting factor, the total downloaded bytes is limited to a certain extent, even if the users were able to reach the maximum network speed. This is because downloading with the highest download speed might sustain only for a short period of time.

V. Mobile Data Usage Patterns Across Countries

We further investigate the data usage patterns of geographically diverse mobile users across six countries: Finland (FI), Germany (DE), the United Kingdom (UK), Japan (JP), Brazil (BR), and India (IN). These countries are selected based on the geographical difference and the sufficient amount of measurement data collected from each country, as shown in Table I. Note that our focus is on highlighting similarities and differences in data usage and the traffic demand of mobile users in these six countries.

A. Session and Data Volume

Fig. 17 shows the distribution of download consumption per user over cellular and WiFi networks across the six countries. We observe that, except for users in Finland, mobile users in the other countries prefer WiFi networks for both uploading to (the plot not shown) and downloading content from the Internet. Moreover, we observe that the average daily download (unconstrained) speeds over cellular networks vary per country. For instance, in Japan, the average daily download speed over cellular is 5.6 Mbps, while in Finland it is ~8 Mbps. This indicates that mobile users in Finland mostly use heavy applications that demand high data traffic than mobile users in Japan. The flat-rate based data subscription plan in Finland (see: §III) could be one of the factors for users to use cellular networks for most of their activities on the Internet. The average daily download speed over cellular in the UK, BR, and IN are 6.3 Mbps, 4.2 Mbps, and 4.0 Mbps, respectively.

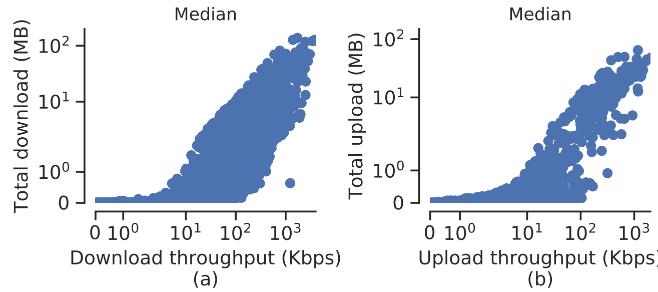


Fig. 16: *The relationships between download speed with the total download (a); and the median upload speed with the total upload bytes per user (b).*

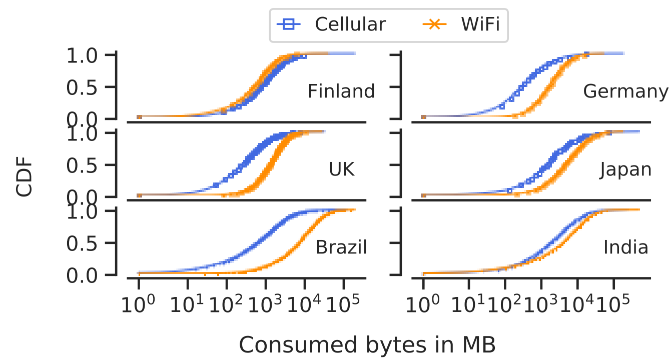


Fig. 17: *Total downloaded bytes per mobile users over WiFi and cellular networks in six countries. Mobile users (except in FI) prefer using WiFi for downloading content.*

We observe that the overall download consumption of users per country in the six countries (irrespective of whether users access cellular or WiFi networks) has a similar pattern. This indicates that mobile users in different countries mostly access the Internet from their devices in a similar fashion. However, in terms of the network technology type mobile users prefer to access the Internet on their device varies across countries. This might be due to different reasons such as network coverage and market price of subscription.

B. Ratio of Using WiFi and Cellular Networks

We also study the trend of mobile users on accessing content from cellular or WiFi networks over the time of the day in the aforementioned countries. To this end, we only focus on the

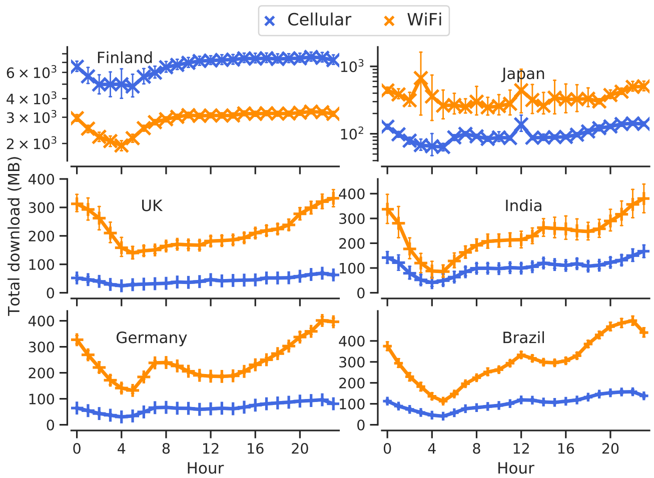


Fig. 18: *Median of hourly total download bytes per user between cellular and WiFi networks over the time of the day in six countries. Note the scale difference on Y-axis.*

set of users that have been accessing both cellular and WiFi networks during the measurement period. For every user, we calculated the total download bytes transferred over cellular and WiFi networks over the time of a day. Fig. 18 shows the trend of total download bytes per user at every hour during the month in the six countries. The trend of downloading content over WiFi and cellular network follows similar patterns. We observe that except for Japan, mobile users in other five countries follow relatively similar patterns. For instance, there is an increasing trend for downloading traffic volume during the evening hours (typically 16:00 to 20:00). In the afternoons, the number of users who start downloading content over WiFi increase more than the number of users downloading over cellular networks (e.g., in the UK and Germany, after 16:00). This might indicate that when users are at home or off-work time they prefer accessing Internet over WiFi. It might also indicate that users start accessing heavy traffic demanding apps over WiFi during that period. For instance, for the mobile user in Germany, hourly total download bytes per user over cellular and WiFi networks at 4:00 AM is 24 MB and 158 MB, respectively. Observing at 20:00 hours, it has 58 MB over cellular and 277 MB over the WiFi networks.

C. Network Congestion and Service Level Score (SLS)

We study how the network coverage, especially the availability of the 4G network varies across different countries and how this aspect is associated with the service satisfaction level (*i.e.*, SLS) of mobile users. Mobile users in Japan rank highest by reporting access to the 4G network in more than 97% of measurement samples. While, mobile users in Finland access the 4G network 83.4% of the time with users in India reporting 4G access in 72% of the measurement samples.

The measurement platform identifies whether the created sessions were constrained (see: §IV-A) in the mobile network. The network can be a bottleneck due to several reasons. For one, the session can be constrained in situations where congestion occurs at the base station. We study the number of such constrained sessions created by users per country. For instance, in India, ~22% of the user sessions were under constrained conditions, while in Japan only ~9.4% of the sessions were created as constrained sessions.

We consider the UK and Finland users as a sample to compare the detailed distribution of different radio technologies. In the UK, about 75% of the cellular data sessions were accessed over the LTE network and about 9.5% over HSPA+ networks. Other network technologies also take a share with High Speed Uplink Packet Access (HSUPA) 5.2%, IWLAN (3.3%), HSPA(3.3%), EDGE (1.4%). Note that multiple measurement sessions can be collected from a single user who accessed the LTE network. As a result, the highest percentage value per session of the LTE network does not directly reflect the number of users accessing the LTE network. Considering the unique number of users accessing the LTE and 3G networks, users accessing the 3G networks were higher than users accessing the LTE networks (about 3% difference). This implies that users in the UK have more frequently connected to 3G networks than the LTE networks. However, we also find that users accessing the LTE networks consume higher median and the total downloaded bytes.

While in Finland, about 80% of the sessions created by the users were over the LTE network. About 16% of the sessions were over the HSPA+ radio network. The rest were over HSUPA (1.6%), ~1% over HSPA, and only 0.3% over the EDGE network. In Finland, the number of sessions and the number of users accessing the LTE network is higher than that of users accessing the 3G network.

Service Level Score (SLS) – SLS measures user satisfaction level based on the number of times that the users have got what they ask from the network. It is the ratio of the difference between the number of unconstrained and constrained sessions to the number of unconstrained sessions, for every user, as shown in Eq. 1; where α is the number of unconstrained sessions and β is the number of constrained sessions.

$$SLS = \frac{\alpha - \beta}{\alpha} \times 100 \tag{1}$$

Based on user activity and application type, some times of the day tend to be more active in traffic flow than others. We observe the temporal traffic dynamics and the service level score of mobile users by time of the day. Fig. 19 shows the hourly SLS values in a cellular network across the six countries. The box plot in black shows the distribution of SLS over the time of a day for the whole dataset. The orange lines are median values per user for every hour. We observe that SLS values start to deteriorate during peak (e.g., starting from 4:00) hours of the day.

We also study the daily distribution of the SLS score per user across six countries. Fig. 20 depicts the daily distribution of SLS score per mobile users across the six countries. It can be seen that the daily average SLS score for each country is FI (66.52%), UK (75.75%), JP (76.78%), DE (74.22%), IN (54.73%), and BR (67.74%). Considering the daily median SLS value for every user, we found that mobile users in the UK have the highest SLS score (90%), while mobile users in India have the lowest SLS score (62%). Mobile users in the

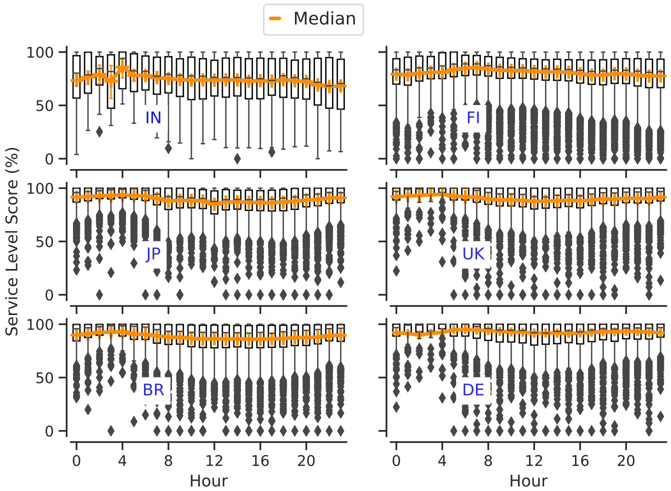


Fig. 19: *SLS distribution over time of the day in six countries. Note, the higher the percentage, the better the SLS score. The SLS score in most of the countries follows a diurnal pattern where poor SLS scores are observed during peak times.*

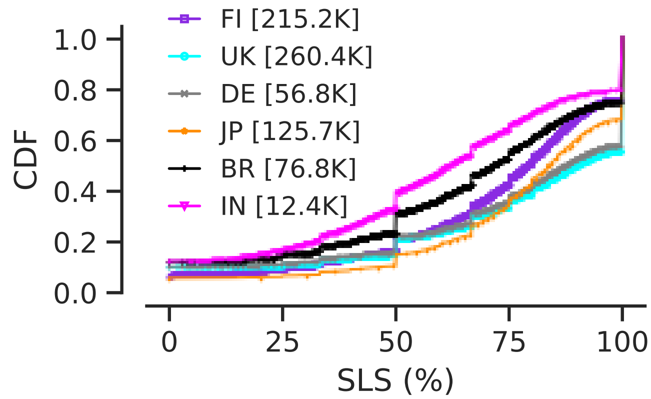


Fig. 20: *CDF of SLS values in cellular network for six countries. Mobile users in UK have higher SLS score.*

Germany and Japan have median SLS score of 87% and 84% per user per day, respectively. On the other hand mobile users in Brazil and India have relatively lower SLS score, 71% and 62%, respectively. Relatively, the daily median SLS score per user in Finland (FI) is also low (79%). This is because mobile users in Finland use the cellular network on their devices a lot for most of their Internet activities as shown in Fig. 17. On the other hand, mobile users in the UK and Germany (DE) have high SLS score since they might avoid consuming a lot of data over cellular networks. Note that, we understand that SLS can not be used as a universal metric for representing the Key Performance Indicators (KPIs) for various mobile applications and services. However, it can serve as a good first-hand approximation of the satisfaction level of mobile users as SLS takes into consideration the speed of the mobile network and its constraints and bottlenecks.

Takeaway – Most mobile users in the countries we have studied consume the highest data volume over WiFi networks than over cellular networks. Mobile users in Finland are an exception consuming the highest data volume over cellular networks instead. Mobile users in Finland also exhibit lower Service Level Score (SLS) than users in some other countries such as the UK and Germany. This indicates that the SLS score can be higher when mobile users are conservative on cellular data usage (e.g., if their subscription plan is priced per data usage). The availability of good network coverage and flat-rate market price are some of the reasons for mobile users to prefer using cellular over WiFi or vice versa. We also observe that the total data consumption across the countries (considering both cellular and WiFi networks together) of mobile users is comparable. This indicates that mobile users across countries have a similar trend of data usage, although there is a difference in network technology that is available for access.

VI. Related Work

There are several studies on data usage patterns and behavior of mobile users (see [57] for a recent survey) whereby the studies can generally be grouped into the following areas –

Mobility and location related patterns – Paul *et al.* [58], study the pattern of mobility and temporal activity as well as how the radio resources are utilized by different applications using a dataset collected at the core of a 3G network. The authors show that the distribution of traffic consumption across subscribers is uneven, such that 90% of traffic loads in the 3G network is generated by 10% of the subscribers. Similarly, Yang *et al.* [59], characterize behavior of mobile users in terms of mobility patterns, application usage based on the data collected at 2G and 3G core networks in China. They show that about 1% of users frequently use different applications and consume the highest data volume. They also indicate that the user mobility per day is limited to a few unique (not more than 10) cells of base stations. Other studies such as [60], [61], and [62] investigate human mobility and behavior in mobile networks using measurements. These studies show that user mobility exhibits patterns over time of the day and location.

Application usage patterns – Previous work [63], [64] has studied mobile application usage behavior by considering different application activities (e.g., installing, uninstalling, and updating) and network usage. Yu *et al.* [25] show that application usage behavior of users and dynamics at a given location can be predicted by using the point of interest (POI) information of that location. Authors in [65], [24] study application usage pattern of smartphone users and distinguish different groups of mobile users. Canneyt *et al.* [66] study application usage behavior of users by investigating engagement patterns of users with their applications. They use a dataset collected using `Flurry` [67]. The authors show that users application usage activity and disruption patterns are correlated with major events such as sports and political events.

Several other studies [68], [69], [70], [10] have focussed on mobile application usage behavior and retention patterns. Silva *et al.* [10] study mobile application usage patterns using an year-long dataset collected from devices of mobile users in Brazil. They investigate mobile application usage patterns in terms of temporal and location differences. They show that social networking applications such as Facebook and WhatsApp have the highest data exchange. The authors propose a model that predicts the next application based on synthetically generated datasets of mobile application usage. Wu *et al.* [70] study application usage of mobile users in combination with the temporal patterns and identified six group of users. The authors also propose a model (based on Wavelet-ARMA) that predicts the traffic demand of mobile users for different applications. Ben-Gal *et al.* [69] study mobility patterns of users and propose a clustering model that identifies group of users with a "shared lifestyle" irrespective of their location.

Patterns based on traffic size and flows – Oliveira *et al.* [21] study and characterize data traffic and usage behavior of mobile users based on a dataset collected from a 3G network in Mexico. The authors profile mobile users into three classes (light, medium, and heavy users) by using the number of sessions, traffic volume, and inter-arrival times. Zhang *et al.* [39] study the characteristics of cellular data 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 fixed-line networks. They show that cellular networks have multiple short flows [55] than fixed-line networks. The authors cluster different applications based on similarity patterns on the traffic metrics (flow and session size, and inter-packet gap). They show that the inter-packet gap between different applications have significant variations and suggest application-specific optimization methods.

Device access patterns and device model types – Shafiq *et al.* [2] model traffic dynamics on mobile devices using a week-long trace collected from the core network of a cellular operator. They study traffic dynamics and characteristics of applications on three different mobile device brand families. They show that the type of device attributes to different traffic behavior. Falaki *et al.* [26] studied traffic generated from 255 mobile users along with the interaction with their smartphones. The authors show that user interaction with the device contributes to a higher battery consumption.

Compared to related work, the data usage pattern and behavior of mobile users is not explored very well. This is because previous analysis on data usage patterns of mobile users is either limited to one city or country [21]; considers limited sample of users [26]; is based on data from a single cellular core operator and area [71] or focusses on specific application types [72]. Unlike the previous studies, our work instead focuses on an extensive analysis of data usage pattern and behavior analysis of mobile users based on ~340M records (measurement sessions) collected from the end-user devices. The dataset covers vast geography of users encompassing six countries. We consider several essential features that determine data usage patterns of mobile users. Some of the features we studied include device models, geographical location of users, application categories installed on mobile device, and the impact of different subscription markets on the data consumption of mobile users.

VII. Conclusion

We presented an analysis of data usage patterns and behavior of mobile users. To this end, we used a month-long dataset with more than 340 million measurement sessions (recorded in July 2018) from six countries. We investigated the behavioral pattern of mobile users by considering different factors such as time of the day, user mobility, location, and the frequency of users accessing the data traffic over cellular or WiFi network. We also studied how data usage patterns and the Service Level Score (SLS) of mobile users varies across the six countries.

We showed that data usage patterns of mobile users are correlated with multiple factors. The factors include user mobility, the accessed network type (cellular and WiFi), the choice of the device model, the type of radio technology (such as 3G and 4G), and user mobility. We showed that mobile users at a stationary position consume higher volumes of data than users on the move. Furthermore, pricing strategy of Mobile Network Operators (MNOs) and the availability of good network coverage can shape the data usage behavior of mobile users. Especially which network type (cellular or WiFi) users prefer to access the Internet using their mobile device is often subject to their data subscription plans. We showed that mobile users in Finland tend to use heavier applications over a cellular network than in the rest of the countries we studied as flat-rate pricing is dominant in Finland.

To further encourage reproducibility of our results, the measurement dataset and scripts used in the analysis are made publicly available [29].

References

[1] G. Maier, F. Schneider, and A. Feldmann, "A First Look at Mobile Hand-Held Device Traffic," ser. Passive and Active Measurement, 11th International Conference (PAM), 2010, pp. 161–170. [Online]. Available: https://doi.org/10.1007/978-3-642-12334-4_17

[2] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," ser. Measurement and Modeling of Computer Systems (SIGMETRICS), 2011, pp. 305–316. [Online]. Available: https://doi.org/10.1145/1993744.1993776

[3] A. Gember, A. Anand, and A. Akella, "A Comparative Study of Handheld and Non-handheld Traffic in Campus Wi-Fi Networks," ser. Passive and Active Measurement Conference, PAM, 2011, pp. 173–183. [Online]. Available: https://doi.org/10.1007/978-3-642-19260-9_18

[4] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck, "An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance," ser. ACM SIGCOMM, 2013, pp. 363–374. [Online]. Available: https://doi.org/10.1145/2486001.2486006

[5] Ericsson. (2018) Mobile Traffic by Application Category. [Online]. Available: https://www.ericsson.com/en/mobility-report/reports/november-2018/mobile-traffic-by-application-category

[6] ——. (2018) Future Mobile Data Usage and Traffic Growth. [Online]. Available: https://www.ericsson.com/en/mobility-report/future-mobile-data-usage-and-traffic-growth

[7] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A Survey on 3GPP Heterogeneous Networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, 2011. [Online]. Available: https://doi.org/10.1109/MWC.2011.5876496

[8] C. V. N. Index, "Global mobile data traffic forecast update, 2015–2020," *Cisco white paper*, p. 9, 2016.

[9] J. Guo, C. She, and C. Yang, "Predictive Resource Allocation with Coarse-Grained Mobility Pattern and Traffic Load Information," ser. IEEE ICC. IEEE, 2018, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICC.2018.8422705

[10] F. A. Silva, A. C. S. A. Domingues, and T. R. M. B. Silva, "Discovering Mobile Application Usage Patterns from a Large-Scale Dataset," *ACM Trans. Knowl. Discov. Data*.

[11] T. Italia. (2020) TIM Social and Chat. [Online]. Available: https://www.tim.it/offerte/mobile/internet-su-misura-te
[12] T. Xia and Y. Li, "Revealing Urban Dynamics by Learning Online and Offline Behaviours Together," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 1, Mar. 2019. [Online]. Available: https://doi.org/10.1145/3314417
[13] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring Land Use from Mobile Phone Activity," ser. ACM SIGKDD, O. E. Wolfson and Y. Zheng, Eds. ACM, 2012, pp. 1–8. [Online]. Available: https://doi.org/10.1145/2346496.2346498
[14] M. Lenormand, M. Picornell, O. G. Cantú-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frias-Martinez, and J. J. Ramasco, "Cross-checking Different Sources of Mobility Information," *PloS one*, vol. 9, no. 8, p. e105184, 2014. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105184
[15] S. Bekhor, Y. Cohen, and C. Solomon, "Evaluating Long-Distance Travel Patterns in Israel by Tracking Cellular Phone Positions," *Journal of Advanced Transportation*, vol. 47, no. 4, pp. 435–446, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/atr.170
[16] G. Miritello, R. Lara, M. Cebrian, and E. Moro, "Limited Communication Capacity Unveils Strategies for Human Interaction," *Scientific reports*, vol. 3, p. 1950, 2013. [Online]. Available: https://www.nature.com/articles/srep01950
[17] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, and H. Yan, "Modeling Web Quality-of-Experience on Cellular Networks," ser. Mobile Computing and Networking (MobiCom), 2014, pp. 213–224. [Online]. Available: https://doi.org/10.1145/2639108.2639137
[18] E. Boz, B. Finley, A. Oulasvirta, K. Kilkki, and J. Manner, "Mobile QoE Prediction in the Field," *Pervasive and Mobile Computing*, vol. 59, p. 101039, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574119218307673
[19] A. S. Asrese, E. A. Walelgne, V. Bajpai, A. Lutu, Ö. Alay, and J. Ott, "Measuring Web Quality of Experience in Cellular Networks," ser. Passive and Active Measurement - PAM, 2019, pp. 18–33. [Online]. Available: https://doi.org/10.1007/978-3-030-15986-3_2
[20] H. Shi and Y. Li, "Discovering Periodic Patterns for Large Scale Mobile Traffic Data: Method and Applications," *IEEE Trans. Mob. Comput.*, vol. 17, no. 10, pp. 2266–2278, 2018. [Online]. Available: https://doi.org/10.1109/TMC.2018.2799945
[21] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-Driven Mobile Data Traffic Modeling in a Large Metropolitan Area," ser. Pervasive Computing and Communications (PerCom), 2015, pp. 230–235. [Online]. Available: https://doi.org/10.1109/PERCOM.2015.7146533
[22] C. Shin, J. Hong, and A. K. Dey, "Understanding and Prediction of Mobile Application Usage for Smart Phones," ser. ACM Ubicomp, 2012, pp. 173–182. [Online]. Available: https://doi.org/10.1145/2370216.2370243
[23] V. Srinivasan, S. Moghaddam, A. Mukherji, K. K. Rachuri, C. Xu, and E. M. Tapia, "Mobileminer: mining your frequent patterns on your phone," ser. ACM UbiComp, 2014, pp. 389–400. [Online]. Available: https://doi.org/10.1145/2632048.2632052
[24] S. Zhao, J. Ramos, J. Tao, Z. Jiang, S. Li, Z. Wu, G. Pan, and A. K. Dey, "Discovering Different Kinds of Smartphone Users Through their Application Usage Behaviors," ser. ACM UbiComp, 2016, pp. 498–509. [Online]. Available: https://doi.org/10.1145/2971648.2971696
[25] D. Yu, Y. Li, F. Xu, P. Zhang, and V. Kostakos, "Smartphone app usage prediction using points of interest," *IMWUT*, vol. 1, no. 4, pp. 174:1–174:21, 2017. [Online]. Available: https://doi.org/10.1145/3161413
[26] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, "Diversity in smartphone usage," ser. Mobile Systems, Applications, and Services (MobiSys), 2010, pp. 179–194. [Online]. Available: https://doi.org/10.1145/1814433.1814453
[27] V. Bajpai and J. Schönwälder, "A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts," ser. IEEE Communications Surveys and Tutorials, vol. 17, no. 3, 2015. [Online]. Available: https://doi.org/10.1109/COMST.2015.2418435
[28] V. Bajpai, A. Brunström, A. Feldmann, W. Kellerer, A. Pras, H. Schulzrinne, G. Smaragdakis, M. Wählisch, and K. Wehrle, "The Dagstuhl Beginners Guide to Reproducibility for Experimental Networking Research," vol. 49, no. 1, pp. 24–30, 2019. [Online]. Available: https://doi.org/10.1145/3314212.3314217
[29] E. A. Walelgne, A. S. Asrese, J. Manner, V. Bajpai, and J. Ott. (2020) Understanding Data Usage Patterns of Geographically Diverse Mobile Users (Dataset). [Online]. Available: https://github.com/ermiaswalelgne/2020-TNSM-mobile-data-usage
[30] C. Ltd. (2019) Netradar. Last accessed October 10, 2019. [Online]. Available: https://www.netradar.com/
[31] D. Xu and D. Qian, "A Bandwidth Adaptive Method for Estimating End-to-End Available Bandwidth," ser. Communication Systems (ICCS). IEEE, 2008, pp. 543–548. [Online]. Available: https://doi.org/10.1109/ICCS.2008.4737243
[32] S. S. Chaudhari and R. C. Biradar, "Survey of Bandwidth Estimation Techniques in Communication Networks," *Wireless Personal Communications*, vol. 83, no. 2, pp. 1425–1476, 2015. [Online]. Available: https://doi.org/10.1007/s11277-015-2459-2
[33] E. Boz and J. Manner, "A hybrid Approach to QoS Measurements in Cellular Networks," *Computer Networks*, vol. 172, p. 107158, 2020.
[34] A. API. (2019) TrafficStats . [Online]. Available: https://developer.android.com/reference/android/net/TrafficStats
[35] A. M. Mandalari, A. Lutu, A. Custura, A. S. Khatouni, Ö. Alay, M. Bagnulo, V. Bajpai, A. Brunström, J. Ott, M. Mellia, and G. Fairhurst, "Experience: Implications of Roaming in Europe," ser. Conference on Mobile Computing and Networking, MobiCom, 2018, pp. 179–189. [Online]. Available: https://doi.org/10.1145/3241539.3241577
[36] T. Ertiö and P. Räsänen, "Mobile Online Proficiency and Mobile Internet Use - Findings from Finland," ser. Human Aspects of Information Security, Privacy and Trust HAS, 2017, pp. 45–56. [Online]. Available: https://doi.org/10.1007/978-3-319-58460-7_4
[37] R. Wei, "Motivations for Using the Mobile Phone for Mass Communications and Entertainment," *Telematics and Informatics*, vol. 25, no. 1, pp. 36–46, 2008. [Online]. Available: https://doi.org/10.1016/j.tele.2006.03.001
[38] S. R. Department. (2020) Mobile Phone Internet Data Subscriptions in Finland from 2011 to 2017, by Type. [Online]. Available: https://www.statista.com/statistics/551000/finland-mobile-phone-internet-data-subscription-by-type/
[39] Y. Zhang and Å. Arvidsson, "Understanding the Characteristics of Cellular Data Traffic," *CCR*, vol. 42, no. 4, pp. 461–466, 2012. [Online]. Available: https://doi.org/10.1145/2377677.2377764
[40] P. Fiadino, P. Casas, M. Schiavone, and A. D'Alconzo, "Online Social Networks anatomy: On the Analysis of Facebook and WhatsApp in Cellular Networks," ser. IFIP, 2015, pp. 1–9. [Online]. Available: https://doi.org/10.1109/IFIPNetworking.2015.7145326
[41] M. Siekkinen, M. A. Hoque, and J. K. Nurminen, "Using Viewing Statistics to Control Energy and Traffic Overhead in Mobile Video Streaming," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1489–1503, 2016. [Online]. Available: https://doi.org/10.1109/TNET.2015.2415873
[42] V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z. Zhang, "Unreeling Netflix: Understanding and Improving Multi-CDN Movie Delivery," ser. IEEE INFOCOM, 2012, pp. 1620–1628. [Online]. Available: https://doi.org/10.1109/INFCOM.2012.6195531
[43] F. Richter, A. J. Fehske, P. Marsch, and G. P. Fettweis, "Traffic Demand and Energy Efficiency in Heterogeneous Cellular Mobile Radio Networks," in *IEEE Vehicular Technology Conference*, 2010. [Online]. Available: https://doi.org/10.1109/VETECS.2010.5493683
[44] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "On the Effectiveness of Sngle and Multiple Base Station Sleep Modes in Cellular Networks," *Computer Networks*, vol. 57, no. 17, pp. 3276–3290, 2013. [Online]. Available: https://doi.org/10.1016/j.comnet.2013.07.016
[45] A. API. (2019) Location . [Online]. Available: https://developer.android.com/reference/android/location/Location
[46] Q. Liu, K. Xu, H. Wang, M. Shen, L. Li, and Q. Xiao, "Measurement, Modeling, and Analysis of TCP in High-Speed Mobility Scenarios," ser. International Conference on Distributed Computing Systems, ICDCS, 2016, pp. 629–638. [Online]. Available: https://doi.org/10.1109/ICDCS.2016.43
[47] L. Li, K. Xu, D. Wang, C. Peng, K. Zheng, R. Mijumbi, and Q. Xiao, "A Longitudinal Measurement Study of TCP Performance and Behavior in 3G/4G Networks Over High Speed Rails," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2195–2208, 2017. [Online]. Available: https://doi.org/10.1109/TNET.2017.2689824
[48] A. R. Mishra, "Fundamentals of Cellular network Planning and Optimisation," *Evolution to. Wiley-Interscience*, 2004.
[49] Johhny. (2015) CellFinder . [Online]. Available: https://cellidfinder.com/articles/how-to-find-cellid-location-with-mcc-mnc-lac-i-cellid-cid
[50] 3GPP. (2013) 3GPP Release 10. [Online]. Available: https://www.3gpp.org/ftp/tsg_sa/WG5_TM/TSGS5_81/_specs_for_checking/draft_32551-b00.doc
[51] H. Cao, J. Sankaranarayanan, J. Feng, Y. Li, and H. Samet, "Understanding Metropolitan Crowd Mobility via Mobile Cellular Accessing Data," *ACM Trans. Spatial Algorithms and Systems*,

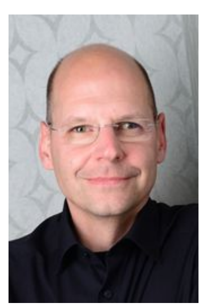
vol. 5, no. 2, pp. 8:1–8:18, 2019. [Online]. Available: https://doi.org/10.1145/3323345
[52] Gsmarena. (2019) GSMArena.com - Mobile Phone Reviews, News, and Specifications . [Online]. Available: https://www.gsmarena.com/
[53] AndroidDev. (2019) Android Developers Platform. [Online]. Available: https://developer.android.com/preview
[54] E. A. Walelgne, J. Manner, V. Bajpai, and J. Ott, "Analyzing Throughput and Stability in Cellular Networks," ser. IEEE/IFIP Network Operations and Management Symposium, NOMS, 2018. [Online]. Available: https://doi.org/10.1109/NOMS.2018.8406261
[55] E. A. Walelgne, S. Kim, V. Bajpai, S. Neumeier, J. Manner, and J. Ott, "Factors Affecting Performance of Web Flows in Cellular Networks," ser. IFIP Networking Conference, 2018. [Online]. Available: http://dl.ifip.org/db/conf/networking/networking2018/2A1-1570416544.pdf
[56] A. Dhamdhere, D. D. Clark, A. Gamero-Garrido, M. J. Luckie, R. K. P. Mok, G. Akiwate, K. Gogia, V. Bajpai, A. C. Snoeren, and kc claffy, "Inferring Persistent Interdomain Congestion," ser. SIGCOMM, 2018. [Online]. Available: https://doi.org/10.1145/3230543.3230549
[57] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-Scale Mobile Traffic Analysis: A Survey," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 1, pp. 124–161, 2016. [Online]. Available: https://doi.org/10.1109/COMST.2015.2491361
[58] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," ser. IEEE INFOCOM, 2011, pp. 882–890. [Online]. Available: https://doi.org/10.1109/INFCOM.2011.5935313
[59] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing User Behavior in Mobile Internet," *IEEE Trans. Emerging Topics Comput.*, vol. 3, no. 1, pp. 95–106, 2015. [Online]. Available: https://doi.org/10.1109/TETC.2014.2381512
[60] Y. Qiao, Y. Cheng, J. Yang, J. Liu, and N. Kato, "A Mobility Analytical Framework for Big Mobile Data in Densely Populated Area," *IEEE Trans. Vehicular Technology*, vol. 66, no. 2, pp. 1443–1455, 2017. [Online]. Available: https://doi.org/10.1109/TVT.2016.2553182
[61] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding Individual Human Mobility Patterns," *nature*, vol. 453, no. 7196, p. 779, 2008. [Online]. Available: https://www.nature.com/articles/nature06958
[62] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Human mobility patterns in cellular networks," ser. IEEE CL, 2013. [Online]. Available: https://doi.org/10.1109/LCOMM.2013.090213.130924
[63] X. Liu, W. Ai, H. Li, J. Tang, G. Huang, F. Feng, and Q. Mei, "Deriving User Preferences of Mobile Apps from Their Management Activities," *ACM Trans. Inf. Syst.*, vol. 35, no. 4, pp. 39:1–39:32, 2017. [Online]. Available: https://doi.org/10.1145/3015462
[64] X. Liu, H. Li, X. Lu, T. Xie, Q. Mei, F. Feng, and H. Mei, "Understanding Diverse Usage Patterns from Large-Scale Appstore-Service Profiles," ser. IEEE TSE, vol. 44, no. 4, 2018, pp. 384–411. [Online]. Available: https://doi.org/10.1109/TSE.2017.2685387
[65] P. Welke, I. Andone, K. Blaszkiewicz, and A. Markowetz, "Differentiating smartphone users by app usage," ser. ACM UbiCom, 2016, pp. 519–523. [Online]. Available: https://doi.org/10.1145/2971648.2971707
[66] S. V. Canneyt, M. Bron, A. Haines, and M. Lalmas, "Describing Patterns and Disruptions in Large Scale Mobile App Usage Data," ser. WWW, 2017, pp. 1579–1584. [Online]. Available: https://doi.org/10.1145/3041021.3051113
[67] Yahoo. (2019) Flurry App Analytics for iOS & Android. [Online]. Available: https://www.flurry.com/
[68] A. Zuniga, H. Flores, E. Lagerspetz, P. Nurmi, S. Tarkoma, P. Hui, and J. Manner, "Tortoise or Hare? Quantifying the Effects of Performance on Mobile App Retention," ser. WWW. ACM, 2019, pp. 2517–2528. [Online]. Available: https://doi.org/10.1145/3308558.3313428
[69] I. Ben-Gal, S. Weinstock, G. Singer, and N. Bambos, "Clustering Users by Their Mobility Behavioral Patterns," *ACM Trans. Knowl. Discov. Data*, vol. 13, no. 4, pp. 45:1–45:28, 2019. [Online]. Available: https://doi.org/10.1145/3322126
[70] J. Wu, M. Zeng, X. Chen, Y. Li, and D. Jin, "Characterizing and Predicting Individual Traffic Usage of Mobile Application in Cellular Network," ser. ACM UbiComp/ISWC . ACM, 2018, pp. 852–861. [Online]. Available: https://doi.org/10.1145/3267305.3274173
[71] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment," ser. IEEE/ACM TON, vol. 25, no. 2, 2017, pp. 1147–1161. [Online]. Available: https://doi.org/10.1109/TNET.2016.2623950
[72] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "YouTube Everywhere: Impact of Device and Infrastructure Synergies on User Experience," ser. ACM SIGCOMM, IMC, 2011, pp. 345–360. [Online]. Available: https://doi.org/10.1145/2068816.2068849

Ermias Andargie Walelgne is a doctoral student at Aalto University, Finland. 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 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. He is a full professor of networking technology at Aalto University 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 PI 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 SIGCOMM best paper award (2018), and IEEE COMSOC award (2017) for the best dissertation in network and service management. His current research focuses on the performance and management of next-generation networked systems.



Jörg Ott holds the Chair for Connected Mobility at the Technical University of Munich since Aug 2015. He is also an adjunct professor at Aalto University, where he was a professor from 2005. He is interested in understanding, designing, and building Internet-based communication systems and services. His research focus is on network and system architectures, protocols, and applications for mobile systems. This includes measuring, modeling, analyzing, and predicting network characteristics and application performance as well as preserving user privacy.