Taming QoE in Cellular Networks: from Subjective Lab Studies to Measurements in the Field

Pedro Casas*, Bruno Gardlo*, Michael Seufert†, Florian Wamser†, Raimund Schatz*

*FTW – The Telecommunications Research Center Vienna
{casas | gardlo | schatz}@ftw.at

†University of Würzburg
{seufert | florian.wamser}@informatik.uni-wuerzburg.de

***Abstract*—A quarter of the world population will be using smartphones to access the Internet in the near future. In this context, understanding the Quality of Experience (QoE) of popular apps in such devices becomes paramount to cellular network operators, who need to offer high quality levels to reduce the risks of customers churning for quality dissatisfaction. In this paper we address the problem of QoE provisioning in smartphones from a double perspective, combining the results obtained from subjective lab tests with end-device passive measurements and QoE crowd-sourced feedback obtained in operational cellular networks. The study addresses the impact of the downlink bandwidth on the QoE of three popular smartphone apps: YouTube, Facebook and Google Maps. As a main contribution, we show that the results obtained in the lab are highly applicable in the live scenario, as mappings track the QoE provided by users in real networks. We additionally provide hints and bandwidth thresholds for good QoE levels on such apps, as well as discussion on end-device passive measurements and analysis. The results presented in this paper provide a sound basis to better understand the QoE requirements of popular mobile apps, as well as for monitoring the underlying provisioning network. To the best of our knowledge, this is the first paper providing such a comprehensive analysis of QoE in mobile devices, combining network measurements with users QoE feedback in lab tests and operational networks.**

***Keywords*—QoE; Smartphones; End-device Measurements; Field Trial; Subjective Lab Tests; Mobile Apps; Crowdsourcing**

I. Introduction

Smartphones are becoming the most typical mobile device to access Internet today. Recent projections [2] show that by 2016, a quarter of the world population will be using smartphones to access the most popular services such as YouTube, Facebook, WhatsApp, etc. According to Cisco's global mobile data traffic forecast [1], smartphones will be responsible for more than three-quarters of the mobile data traffic generated by 2019. In the light of these trends, cellular network operators are becoming more and more interested in understanding how to dimension their access networks and how to manage their customers' traffic to capture as many new customers as possible. In this scenario, the concept of Quality of Experience (QoE) has the potential to become one of the main guiding paradigms for managing quality in cellular networks. Closely linked to the subjective perception of the end-user, QoE enables a broader, more holistic understanding of the factors that influence the performance of systems, complementing traditional technology-centric concepts such as Quality of Service (QoS).

In this paper we study the QoE of popular apps in smartphones (YouTube, Facebook and Gmaps) from two different yet complementary perspectives: subjective tests performed in a controlled lab, and passive end-device measurements with QoE user feedback in operational networks, through a field trial. Our study considers the impact of the most relevant QoS-based characteristics of the access network: the downlink bandwidth. Besides providing a solid ground-truth (based on the experience of real users) regarding the QoE-requirements of popular apps such as YouTube and Facebook (e.g., a downlink bandwidth of 4 Mbps/1 Mbps respectively is high enough to reach near optimal results in terms of overall quality and acceptability), our results suggest that lab study results are highly applicable in the live setting, as the mappings obtained between network QoS and user QoE are highly similar in both scenarios. This a major contribution, as it permits to gain high insight about QoE in mobile devices, even by running experiments in the lab.

The standard approach to assess the performance of networks and services from a QoE end-user perspective is to conduct controlled lab experiments [16]–[18]. The key benefits of such an approach rely on the full control the experimenter has on the overall evaluation process. Indeed, content and context are fully known and controlled, and users are directly briefed and observed on the spot, providing as such tangible and solid results. However, lab experiments miss out many important QoE influence factors such as usage context, content preferences by individual users, or device usability among others, potentially introducing differences w.r.t. evaluations conducted in the field [21]. Field trial experiments place the end-user and the evaluated components (i.e. network, apps, etc.) as closest as possible to their daily usage scenarios and running environments, providing more representative evaluations. This augmented degree of realism w.r.t. lab experiments yields in principle more reliable results in terms of end-user experience, to the cost of higher complexity in acquiring and processing the results (e.g., traffic monitoring, QoE feedback, app-level measurements, etc.). Indeed, we developed different tools to conduct the field trial, including a passive monitoring tool to measure the traffic of the field trial participants at their end devices, a QoE-feedback app to gather user experience data (e.g., quality ratings), and a YouTube passive monitoring

The research leading to these results has received partial funding from the European Union under the FP7 Grant Agreement n. 318627, "mPlane", and has been done within the project ACE 3.0 at the Telecommunications Research Center Vienna (FTW).

978-3-901882-77-7

tool to measure initial playback delays, playback stallings, and video quality switches (induced by the adaptive video streaming protocols used by YouTube).

The remainder of the paper is organized as follows: Sec. II presents an overview of the related work on QoE, focusing on the specific case of mobile devices. Sec. III describes the subjective tests' setup and presents the obtained results, including the impact of the downlink bandwidth on the overall experience and acceptability of the end-user. Sec. IV describes the tools and the approach followed in the field trial, and discusses the obtained results, particularly in terms of similarity to those obtained in the lab. Sec. V discusses several implications, limitations and topics related to the passive monitoring of QoE in end-user devices, including privacy, network neutrality, and incentives among others. Finally, Sec. VI concludes this work.

II. Related Work

The study of the QoE requirements for cloud-based applications as the ones we target in this paper has a long list of fresh and recent references. A good survey of the QoE-based performance of cellular networks when accessing different cloud services is presented in [7]. The specific case of QoE in YouTube deserves particular attention, due to the overwhelming popularity and omnipresence of the service. Studies have both considered the "standard" HTTP video streaming flavour of YouTube, as well as the more recent Dynamic Adaptive Streaming (DASH) version. Previous papers [9], [10] have shown that stalling (i.e., stops of the video playback) and initial delays on the video playback are the most relevant Key Performance Indicators (KPIs) for QoE in standard HTTP video streaming. In the case of adaptive streaming, a new KPI becomes relevant in terms of QoE: quality switches. In particular, authors in [12] have shown that quality switches have an important impact on QoE, as they increase or decrease the video quality during the playback. A comprehensive survey of the QoE of adaptive streaming can be found in [13].

There has been a recent surge in the development of tools and software libraries for measuring network performance on mobile devices: some examples are Mobiperf [24], Mobilyzer [23], and the Android version of Netalyzr [22]. When it comes to our specific analysis of QoE in cellular networks and mobile devices, most references are very new, showing that there is still an important gap to fill. In [14], authors study the QoE of YouTube in mobile devices through a field trial, exclusively considering the non-adaptive version of the YouTube player. Authors in [15] recently introduced Prometheus, an approach to estimate QoE of mobile apps, using both passive in-network measurements and in-device measurements, applying machine learning techniques to obtain mappings between QoS and QoE. In [6], authors introduce QoE Doctor, a tool to measure and analyze mobile app QoE, based on active measurements at the network and the application layers. Additional papers in a similar direction tackle the problem of modeling QoE for Web [4] in cellular networks, and video [5].

The main limitation of these approaches is the lack of real user experience ground truth in their analyses. Most of the papers study QoE-related metrics such as page-load times, interface latency, or video stallings but without any reference to real user experience, reflected for example in terms of Mean Opinion Scores. In addition, many of the proposed approaches rely on active measurements only (e.g., [6]), which is less attractive when thinking on large scale user traffic monitoring and analysis. Our approach considers both real users QoE feedback and passive monitoring at end devices, improving and extending the state of the art.

This paper elaborates on our previous study recently presented in [3], particularly extending the analysis by performing measurements at the end devices and conducting a field trial in operational cellular networks.

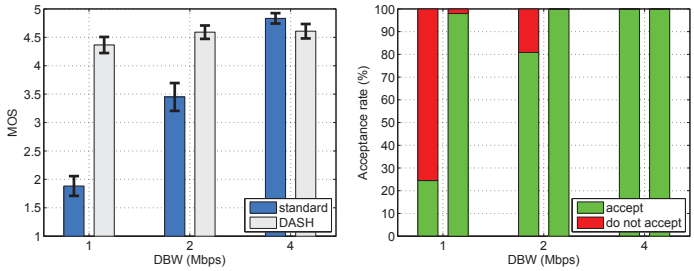
III. Mobile QoE in the Lab

Let us begin by reporting the results of the conducted subjective lab tests. The subjective study consists of 52 participants interacting with the aforementioned services while experiencing different downlink bandwidth profiles in the background data connection. Android smartphone devices are used in the study (Samsung Galaxy S4, OS Android 4.4 KitKat). Devices are connected to the Internet through separate WiFi access networks. The downlink traffic between the different evaluated services and the devices is routed through a modified version of the very well known NetEm network emulator so as to control the different access network profiles under evaluation.

Different constant bandwidth profiles are instantiated at the network emulators, changing downlink bandwidth logarithmically, from 0.5 Mbps to 16 Mbps. These profiles are selected from operational experience, particularly following typical operational values reported in [7] for different access network technologies (LTE, 3G/2G, etc.). Note that while we do not emulate the particular characteristics of a cellular access network (which would mainly impact the RTT profiles), results obtained in the field (c.f. Sec. IV) suggest that our lab results are accurate in real cellular access networks.

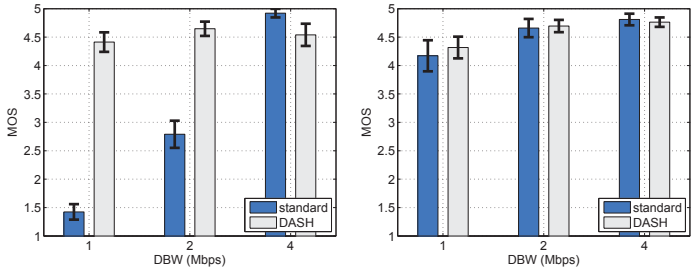
Participants were instructed to perform independent tasks for each of the three considered applications. For YouTube, they were requested to watch two-minutes HD YouTube videos, considering both the usage of the standard (i.e., non-DASH) and the DASH versions of the YouTube player. Videos correspond to 4K ultra-HD videos (i.e., 2160p), which are down-scaled to HD resolution (i.e., 720p) due to the device's display capabilities (i.e., screen size and resolution). The average video bit rate (vbr) of the corresponding HD videos is in all cases around 1.6 Mbps. In the case of Facebook, participants were instructed to access the application with a specific user account, browse the timeline of this user, and browse through specific photo albums created for this user. Finally, Gmaps tasks consisted of exploring different city maps using the Gmaps application, in satellite view, which consumes more bandwidth.

Tests were performed in a dedicated lab for subjective studies, compliant with the QoE subjective studies standards [16]–[18]. Regarding participants' demographics, 29 participants were female and 23 male, the average age was 32 years old, with 40 participants being less than 30 years old. Around half of the participants were students and almost 43% were employees, and 70% of the participants have completed university or baccalaureate studies.



(a) Overall quality (MOS). (b) Acceptance rate.

Figure 1. Overall quality and acceptability in YouTube standard (i.e., non-DASH) and DASH. DASH is capable of handling lower DBW connections with high QoE, trading image quality by lower download throughput.



(a) Stalling. (b) Video Image Quality.

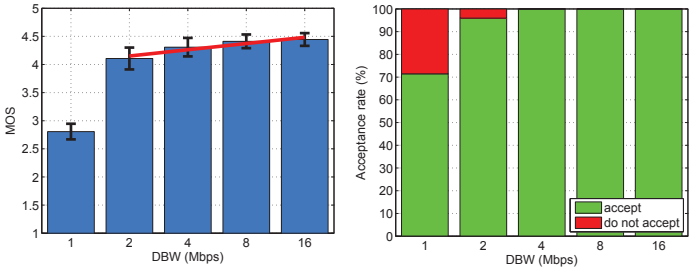
Figure 2. QoE for YouTube Mobile, considering playback stallings and video image quality. Video image quality is perceived as almost excellent for the lowest DBW condition, even if video resolution is lower.

Regarding QoE feedback, participants were instructed to rate their *overall experience* according to a continuous ACR Mean Opinion Score (MOS) scale [16], ranging from "bad" (i.e., MOS = 1) to "excellent" (i.e., MOS = 5). MOS ratings were issued by participants through a custom questionnaire application running on separate laptops, which pops up immediately after a condition has been tested. Participants also provided feedback on the *acceptability* of the application, stating whether they would continue using the application under the corresponding conditions or not. For the specific case of YouTube, two additional questions were asked to participants: (i) *stalling annoyance* (did you perceive stalling as disturbing?); (ii) *video image quality* (rate the image quality of the video). The reader shall note that the maximum MOS ratings declared by the participants are never 5 but somewhere between 4.2 and 4.6. This is a well known phenomenon in QoE studies called *rating scale saturation*, where users hardly employ the limit values of the scale for their ratings [7].

A. QoE in YouTube Mobile

The Downlink BandWidth (DBW) takes values 1 Mbps, 2 Mbps, and 4 Mbps in YouTube tests. Figure 1 reports the overall quality and acceptability results obtained for the YouTube tests. Recall that in the YouTube scenario, we compare the standard, non-adaptive version of the YouTube player (videos are selected to play in HD quality) against the DASH-capable one. In the DASH case, videos are also requested in HD quality, but the server adapts the subsequent video quality resolutions to the bandwidth estimated by the player.

Figure 1(a) compares the overall QoE experienced by the participants using both player versions. It is quite impressive to appreciate how the DASH approach results in a nearly optimal



(a) Overall quality (MOS). (b) Acceptance rate.

Figure 3. QoE in Gmaps. Overall quality and acceptability for different DBW. A DBW of 2 Mbps is high enough to achieve good QoE and almost full acceptability.

QoE for all the tested conditions (from 1 Mbps to 4 Mbps), whereas the fixed HD quality approach results in poor QoE for downlink bandwidth below 4 Mbps. As expected for the standard player, heavy stalling occurs for the 1 Mbps condition, taking into account that the average vbr is 1.6 Mbps. Indeed, as we have shown in [19], the DBW should be in the order of 30% higher than the average video bitrate to avoid stalling when non-adaptive streaming is used. This dimensioning rule also explains the results obtained for the 2 Mbps condition, as some stalling still occurs. No stalling seems to occur for the DASH version. The main difference is that DASH changes the video quality without incurring in playback stalling, whereas the fixed quality configuration definitely results in video stalling.

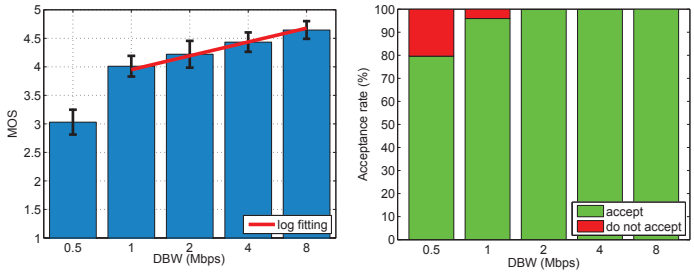
Figure 1(b) reports the results in terms of acceptability of the participants. This is one of the key features that an operator has to consider, because low acceptance rate may sooner or later turn into churn. As observed, acceptance rate is as low as 23% for the standard streaming at 1 Mbps, whereas it's close to 99% in the case of DASH.

To complement the picture for YouTube QoE in mobile devices, Figure 2 depicts the results obtained in terms of (a) annoyance caused by stalling (stop of the video playback), and (b) video image quality. In Figure 2(a), a MOS = 5 means not disturbing at all, whereas a MOS = 1 means unbearable (very annoying). Stalling has a very strong impact on the user's level of annoyance, confirming what has been already seen in previous studies for desktop and laptop like devices.

The most interesting result is presented in Figure 2(b), which reports the perceived image quality of the video. According to previous studies [12], quality switches induced by DASH have an important impact on QoE. However, in the case of smartphones, where displays are smaller than laptops or desktop devices, quality switches do not seem to have an important impact on the perception of the user. While these results are directly linked to the specific quality-switching patterns induced by the tested DBW conditions, they represent a main contribution to assess QoE for YouTube in smartphones when using DASH. As a summary, using DASH highly reduces the chances of playback stalling, at no apparent perceived image quality cost.

B. QoE in Gmaps and Facebook Mobile

Gmaps is tested with a fully logarithmic scale: 1 Mbps, 2 Mbps, 4 Mbps, 8 Mbps, and 16 Mbps. Figure 3 reports the overall quality and acceptability results obtained for the Gmaps tests. Figure 3(a) shows that a DBW of 4 Mbps results



(a) Overall quality (MOS). (b) Acceptance rate.

Figure 4. QoE in Facebook. Overall quality and acceptability for different DBW. A DBW of 1 Mbps is high enough to achieve good QoE and almost full acceptability.

in near optimal QoE (MOS $\approx$ 4.5), and from this value on, QoE saturation already occurs. This means that no major QoE improvements are then obtained for additional bandwidth provisioning. A DBW of 2 Mbps provides good quality results and almost full acceptance, but a DBW of 1 Mbps rapidly brings Gmaps into bad user experience.

Similarly, Facebook is tested with DBW = 0.5 Mbps, 1 Mbps, 2 Mbps, 4 Mbps and 8 Mbps. Figure 4 reports the results obtained in the Facebook tests for different DBW configurations, considering both (a) the overall quality and (b) the acceptance rate. A DBW of 500 kbps is not high enough to reach full user satisfaction in Facebook mobile for Android devices, as participants declared a fair quality with an acceptance rate of about 80%. Still, a DBW of 1 Mbps results in good overall quality, with almost full acceptance of the participants. Excellent QoE results are attained for 8 Mbps, which shows that even if a 2 Mbps DBW allocation is high enough to reach full acceptance (cf. Figure 4), the overall experience of the user can still marginally improve.

In both cases, the relation between QoE and DBW is clearly logarithmic when not considering the most restrictive DBW configuration in both apps (1 Mbps and 0.5 Mbps respectively). Next we show that such logarithmic mappings are also observed in the field trial.

IV. From the Lab to the Field

In this section we overview the details of the conducted field trial and analyze the obtained results, particularly comparing them with the observations and conclusions drawn from the subjective lab study. The main question we try to answer is to which extent, subjective lab studies conducted under WiFi networks are applicable to operational cellular networks.

A. Field Trial Overview

The field trial consisted of 30 participants using their own smartphones and cellular ISPs to access the same apps tested in the lab as part of their normal daily Internet activity. Participants were requested to perform the same kind of tasks to those performed by the lab study participants, to improve comparison of results. QoE feedback was provided for each session through a customized QoE crowd-sourcing app (details next), for a total span of 2 weeks. In this paper we only focus on the overall experience declared by participants, but the QoE feedback provided actually includes the same questions as those evaluated in the lab study. In addition, all the traffic flows

Table I. Metrics recorded for each data flow, using the Android-based passive monitoring tool. All metrics are extracted from the Android developers' API.

Metric ID	Metric Name	Units	Example
1	device id (IMEI)	–	352668049725157
2	flow start time	s	1430825689
3	flow direction (up/down)	–	downlink
4	flow duration	s	10,24
5	flow size	KB	4041,00
6	avg. flow throughput	kbps	3157,03
7	app (Android API package)	–	com.android.browser
8	signal strength	dBm	-71
9	operator (MCC.MNC)	–	295.4
10	cell id	–	16815
11	cell location (lat-lon)	deg (^{o})	{40,198-12,347}
12	RAT	–	LTE

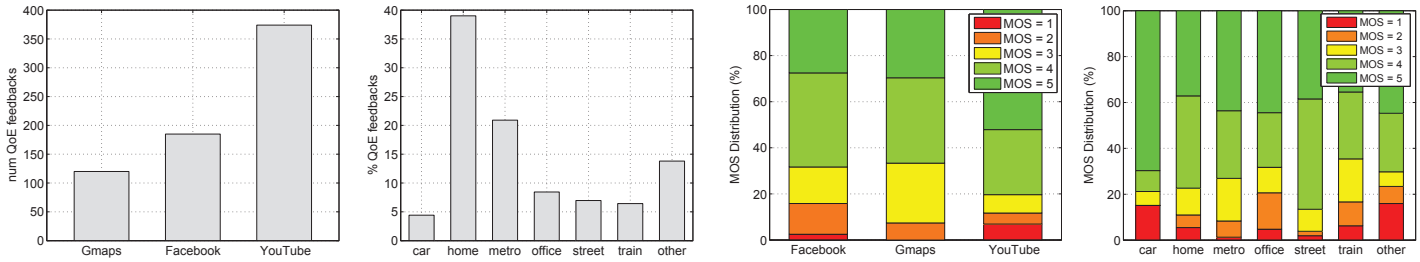
generated by the participants were passively monitored with an Android-based monitoring application developed for this field trial (details next). Besides QoE feedback, participants indicated their location at the moment of performing the corresponding task (e.g., at home, in the underground - metro, walking, etc.). Field trial participants were compensated with vouchers for their participation, which proved to be sufficient for achieving correct involvement in the study.

Figure 5 depicts the distribution of ratings issued by participants in terms of (a) number of ratings per app, (b) per location, and (c-d) MOS values distributions for both apps and locations. In total, almost 700 ratings were issued by the participants during the span of the field trial. As a-priori expected, the biggest share of ratings were done for YouTube, which is currently the most popular app in the Internet. The preferred location was home, which is coherent with the results that we have obtained in previous similar field trials [21]. Interestingly, the second most preferred location to access the requested apps was the underground, evidencing that mobile traffic and smartphone usage in such mobility scenario is highly frequent, at least within the users' community represented by the field trial participants.

Figure 5(c) and Figure 5(d) report the MOS scores distributions. Surprisingly, the MOS distributions are rather similar, both when considering the tested apps (cf. Figure 5(c)) and the selected locations (cf. Figure 5(d)). This suggests that network performance was rather stable during the span of the study, and uniform for both fixed mobility profiles (e.g., home) and highly dynamic mobility profiles (i.e., metro). Indeed, tests were performed in the city of Vienna, where all ISPs have very good network coverage, even in the underground, justifying as such the observed results.

B. End-device Monitoring Tools

To monitor the traffic of the field-trial participants and to log their QoE feedbacks, we developed two specific Android-based applications. The traffic monitoring tool consists of a simple Android-based passive monitoring tool which captures several metrics for all the traffic flows generated by the device. We decided to develop our own tool and not to use those available in the literature (e.g., [22]–[24]), as these either rely on active measurements only or are too specific for their original purpose.



(a) Ratings per App. (b) Ratings per location. (c) MOS dist. per App. (d) MOS dist. per location.

Figure 5. Distribution of QoE feedbacks in the field. The biggest share of ratings were done for YouTube. The preferred location was home, followed by the underground, evidencing the usability scenarios mostly preferred by mobile users. MOS distributions are rather similar wrt tested apps and selected locations, suggesting that network performance was rather stable during the span of the study.

Table II. POPULAR APP NAMES, ACCORDING TO THE ANDROID API NAMING SCHEME.

App	Android API-based Name
YouTube	`system.android.media` `com.google.android.youtube`
Web Browsing (Chrome)	`com.android.chrome`
Web Browsing (Firefox)	`org.mozilla.firefox`
Web Browsing (Android)	`com.android.browser`
WhatsApp	`com.whatsapp`
Gmaps	`com.google.android.apps.maps`
Instagram	`com.instagram.android`
Facebook	`com.facebook.katana` `com.facebook.orca`
Dropbox	`com.dropbox.android`

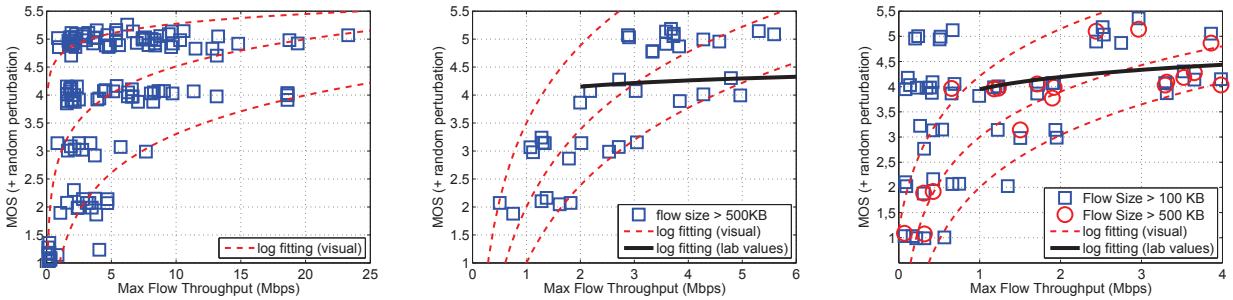
Table I reports the different metrics passively monitored for each traffic flow by our tool. Flows in this context correspond to the standard 5-tuple flow definition, and are associated to the specific app generating them, using the Android developers' APIs. The first metric is a simple device identifier known as IMEI (International Mobile Station Equipment Identity), which is a unique number identifying a 3GPP device. Metrics with ID from 2 to 6 correspond to traffic flow measurements, including the flow start time, the flow direction (uplink or downlink), the flow duration, the size of the flow, and most importantly, the average flow transfer throughput, which is simply computed as the ratio between the flow size and the flow duration. Metric ID 7 indicates the app which generated the corresponding flow, using as naming scheme the Android API notation. For example, YouTube video flows are associated to the app name `system.android.media` (`com.google.android.youtube` is associated to the rest of the YouTube player content, such as thumbnails of videos), Google maps flows are associated to the app name `com.google.android.apps.maps`, Google Chrome web browsing flows are associated to name `com.android.chrome` and so on. Table II provides a list of Android API apps' names for popular mobile apps. Metric ID 8 provides the strength of the signal at the smartphone when the corresponding traffic flow starts. Metrics with ID from 9 to 11 correspond to the operator providing the Internet access and the cell to which the smartphone is attached to at the time of the flow start, particularly including the geographical location of the cell (i.e., longitude and latitude). Finally, metric ID 12 indicate the Radio Access Technology (RAT) used by the smartphone (e.g., LTE, 3G, 2G, EDGE, etc.) when the flow starts.

All these metrics are logged locally at the smartphone, and are periodically sent to a centralized server for post-processing and analysis.

QoE feedbacks are provided by the participants through a web-based app, which is manually run by the user immediately after completing a specific task, such as watching a short YouTube video, exploring a city map using Gmaps, or using Facebook to browse photo albums. This app keeps a local database to store QoE feedbacks even when the device has lost connectivity. For the sake of the analysis presented in this paper, a QoE feedback entry consists of the following 4 fields: {`timestamp; app; location; MOS`}. Given that the QoE feedback tool and the traffic monitoring tool use both the same time reference (i.e., from the local smartphone), a MOS score given by the participant to certain application would always have a timestamp bigger than the timestamps indicating the start of the flows associated to the rated app.

In order to correlate the traffic measurements and the MOS scores provided by the field trial participants, we group flows into sessions. A session corresponds to a group of flows generated by the same app which are continuous in time, based on a pre-defined maximum inter-flows timeout. Evidently, the inter-flows time for a specific session is partially determined by the type of application being accessed by the user, as well as by its usage behavior; for example, the inter-flows time for a web browsing session is generally larger than the inter-flows time for a google maps session. To become independent of such issues, we follow a simple and pragmatic approach to identify relevant sessions. By relevant we refer to sessions which have an associated QoE feedback/MOS rating. The procedure is as follows: given a MOS rating at time t_{MOS} for app app_{MOS}, we define a session as all the flows associated to app app_{MOS} and started within the time window $[t_{\text{MOS}} - Th_{session}; t_{\text{MOS}}]$. The threshold $Th_{session}$ defines the maximum session duration, and it is set to 4 minutes, which is the average time requested to participants to take to perform a specific task.

The final step is to define a proper session-based KPI which could be used to correlate sessions and MOS scores. Recall that the results presented for the lab study considered the downlink bandwidth as the independent network feature being tested in terms of QoE. Hence, we would define a KPI that tries to capture this downlink bandwidth for the rated session. The best approximation one could get for the downlink bandwidth when using passive throughput measurements is the Maximum Flow Throughput (MFT) achieved within the session. The through-



(a) YouTube QoE vs. MFT. (b) Gmaps QoE vs. MFT. (c) Facebook QoE vs. MFT.

Figure 6. QoE for YouTube, Gmaps and Facebook in the field. Squares and circles correspond to individual sessions reported/rated by participants. Red/black lines correspond to log fitting curves. Filtering out small flows improves the correlations between flow throughput measurements and QoE, specially by avoiding protocol impact on the achieved downlink speed.

put of a flow is limited by multiple components, including the application itself, the server providing the flows, the TCP congestion and flow control, and the available bandwidth of the connection. Throughput limitations by the application itself or by the server are less relevant to us, because they are not linked to performance of the cellular network. The impact of the TCP protocol, and specially the slow start phase, can be limited by filtering out small flows from the analysis (we shall come back to this issue later on). Therefore, when targeting the performance of the cellular connection, the MFT achieved for a specific session would be the closest indication to the downlink bandwidth.

In the next section, we analyze the results obtained by correlating the MOS scores and the corresponding session MFT values for the three tested apps.

C. QoE in the Field

Figure 6 depicts the results obtained from the field trial measurements, reporting the MOS scores as a function of the MFT per session for (a) YouTube, (b) Gmaps, and (c) Facebook. To improve visualization of results, MOS scores are plotted with a very small random perturbation (basically to avoid overlapping as much as possible).

Figure 6(a) presents the results obtained in the case of YouTube. Squares correspond to individual sessions rated by participants. Red lines correspond to log fitting curves, with the only purpose of showing such a logarithmic relation between MOS and MFT, in a purely visual basis. High MFT values result in good QoE; indeed, MOS $>$ 4 for almost all sessions with MFT $>$ 5 Mbps, which is highly similar to the results observed in the lab study (cf. Figure 1), where QoE is optimal for a DBW $>$ 4 Mbps. In addition, most of the sessions having very poor QoE (i.e., MOS = 1) have a very low MFT. However, as expected, the picture becomes very fuzzy in the most relevant MFT gap, between 1 Mbps and 4 Mbps, having MOS scores between 2 and 5, i.e., from sessions with poor QoE to excellent QoE. This is coherent with the fact that the QoE of YouTube is strictly linked to the stallings observed in the video playback, and this can happen for both high video bitrate and low video bitrate videos. In addition, as we have shown in Figure 1, using fixed video image quality or adaptive quality completely changes the obtained results, this adding more noise to the overall mapping. As a consequence, even if we can estimate good and bad QoE video sessions for very high and very low MFT values, we need application-layer measurements (i.e., stallings, video bitrate, etc.) to estimate the QoE of YouTube, specially for 1 Mbps $<$ MFT $<$ 4 Mbps. Even if we do not report results in such a direction in this paper, we have developed a tool to provide such application-layer measurements for YouTube in mobile devices [25], which we expect will highly increase the QoE estimations.

Figure 6(b) presents the results obtained in the case of Gmaps. In the case of Gmaps, sessions are composed of both big and small flows, linked to the different components of the app. As we said before, to improve the correlation to network performance, we filter out small flows from the computation of the MFT values. In particular, squares in Figure 6(b) correspond to individual sessions rated by participants, with flows smaller than 500 KB kept aside for the computation of the corresponding MFT. The threshold of 500 KB comes directly from the practice, as we noticed that this represents a good tradeoff between accuracy and coverage of the complete set of Gmaps flows. As before, red curves show the visual log fitting of the MOS vs MFT curve, but in this case, we also add the log fitting curve obtained from the lab study results (cf. Figure 3(a)). Besides some small number of outliers which received MOS scores of 3 (i.e., fair quality), results clearly show that good QoE can be expected for a MFT $>$ 2 Mbps, exactly as suggested by the lab study results in Figure 3(a). In addition, also similarly to the lab indications, QoE rapidly degrades for MFT $\leq$ 1 Mbps. Therefore, we can say that for the case of Gmaps, the mappings between MOS and MFT observed in the field trial are pretty much aligned to the MOS vs DBW curves obtained in the lab study, suggesting that conclusions drawn from such studies have a direct and accurate applicability in the practice.

Figure 6(c) presents the results obtained in the case of Facebook. Facebook flows are rather smaller than in the case of Gmaps, therefore we also consider a similar filtering approach, but considering a less restrictive threshold. In Figure 6(c), squares correspond to sessions with flows smaller than 100 KB filtered out of the computation of the MFT values, whereas circles consider a threshold of 500 KB. As in the case of Gmaps, we include both the visual log fitting curves and the log curve obtained from the lab study results. Mappings follow the lab study results when considering flows $>$ 500

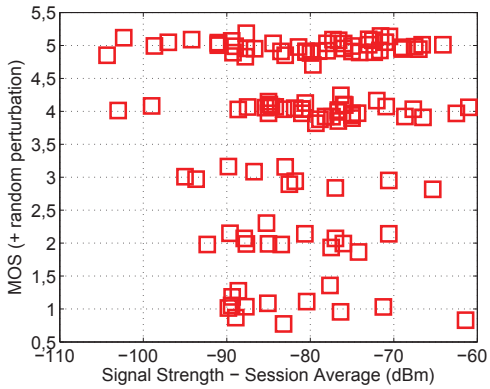


Figure 7. MOS vs signal strength in YouTube. The signal strength metric corresponds to the average single strength when considering all the flows of a single session. There is no apparent correlation between the MOS declared by participants and the measured average signal strength.

KB, resulting in good QoE for MFT $\geq$ 1 Mbps. A MFT $\leq$ 0.5 Mbps results in poor QoE (i.e., MOS = 1 or 2), similar to the observations in the lab, cf. 4. Thus, similar to what we observed in the Gmaps app, mappings between MOS and MFT in the field trial are aligned to the MOS vs DBW curves obtained in the lab study.

As a summary, the MFT observed in a session seems to be a good QoE indicator in the field, specially when considering apps generating big traffic flows. Apps such as Gmaps and Facebook can be reliably monitored in the field using passive flow measurements as the ones conducted by our tool, but considering only big flow instances (flow size $>$ 500 KB). The case of YouTube is a challenging one: high and low MFT values relate well to good and bad QoE, but mappings are very poor for more commonly observed throughputs. Thus, it's necessary to additionally perform measurements at the application layer (e.g., stallings, page-load-times, etc.) to capture QoE indications.

V. Discussion and Perspectives

In this section we provide some additional discussion on the obtained results, and then move on to address several implications, limitations and topics related to the passive monitoring of QoE in end-user devices.

Firstly, considering both the lab and the field results, we can claim that conclusions drawn from both approaches are highly similar and coherent between them, suggesting that subjective lab studies results are applicable to operational cellular networks. In our particular scenario, the usage of WiFi technology in the lab study setup did not have an appreciable impact on the quality of the results when considering real cellular networks.

More in general, obtained results suggest that a downlink bandwidth of 4 Mbps is high enough to reach near optimal results in terms of overall quality and acceptability for YouTube when accessed in smartphones. This threshold drops to 2 Mbps and 1 Mbps for Gmaps and Facebook apps respectively. As a consequence, cellular network operators should target such downlink bandwidth thresholds as their short term goal for dimensioning their access networks. Given these relatively low requirements, resources could be re-allocated or scheduled to manage the network more easily and with a more efficient cost-benefit trade-off, avoiding over-provisioning while keeping high QoE. The implications for the end-user are straightforward: you do not need a super high speed cellular contract with your operator if your target is on the studied applications. So in particular, an expensive LTE contract is not necessary to have a near optimal experience today.

Our results show that dynamic applications such as YouTube DASH are much better suited to smartphone scenarios, providing the same level of experience as the non-adaptive version of the YouTube application in terms of image quality, but with much lower QoS-based requirements in terms of downlink bandwidth. This is a major finding, as DASH has been shown to degrade the video image quality and the associated user experience when considering standard, laptop or PC devices. The main difference with smartphones is their inherent small size displays, which to some extent filter out the impact of quality switches. A direct implication of this finding is that cellular network operators willing to monitor the QoE of its YouTube customers must know which type of technology is used by the YouTube app in the smartphone to understand its QoE. Even more, as also reflected by the results obtained in the field, the only reliable way to monitor QoE in the case of YouTube is to measure application layer features such as stallings and quality levels. Our work in this direction has recently provided very promising results [25].

A particular question that arises in this study is whether other KPIs related to the end-device measurements could also be used to estimate the QoE of a session. The signal strength is a-priori a relevant metric related to the health of the connection, thus it could in principle a good KPI to our purpose. However, we could not find any relevant correlation between the strength of the signal and the MOS scores provided by the participants. As an example, Figure 7 reports the results obtained for the case of YouTube. The signal strength metric corresponds to the average single strength among all the flows of a single session. There is no apparent correlation between MOS scores and the measured average signal strength.

The last part of this section is devoted to present and discuss different implications and topics related to the usage of passive monitoring and QoE-feedback tools at the end-device as the ones we have used in this study.

A. QoE Crowdsourcing Approach

In the conducted field trial, participants rated the quality of their sessions through our tools as part of their participation to the study. However, a quite novel and interesting perspective for QoE-based network performance analysis at the large scale is to employ similar QoE-feedback tools to obtain the feedback of those customers who are willing to do so. Services such as Skype are already taking advantage of its large population of users for doing such an outsourcing of its QoE-based performance monitoring, resulting in a very rich and powerful input to enhance its service and improve the engagement of the users. In a nutshell, every time a user completes a Skype call, the application automatically presents a short questionnaire asking for the experienced quality. We envision a similar approach for the benefit of cellular ISP, where its customers could potentially receive an automatic pop-up like questionnaire after completion of randomly selected sessions.

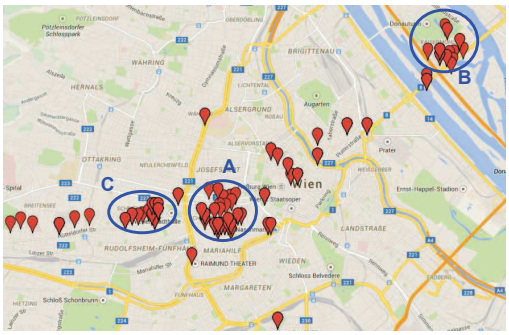


Figure 8. End-device location monitoring and privacy issues. End-user activity and private location can be guessed by simply measuring the location of the cell where smartphone are attached to. In this example scenario, participants' home is located at region A, working office is located at region B, and high activity occurs at region C, linked to daily train traveling.

B. Incentives

Previous discussion brings to the light a highly relevant topic linked to the large scale usage of end-device monitoring system: the incentives a customer receives to install such tools on his phone. End-device measurement tools only become relevant to an operator when these are used at the large-scale, so as to provide meaningful and representative information. Free tools available at the Google Play store such as Onavo[1] and RadioOpt[2] are smartly designed such that the customer is attracted to install and maintain the app running on its phone, based on side applications provided by the tools, such as widgets measuring the data consumption, or proxies offering data compression to reduce the usage of the contracted data volume. Google is for sure the leader in terms of incentives, as all of its apps are highly valuable to the end user (gmail, gmaps, gdocs, etc.), and as a side effect, the company has a full visibility of its worldwide overlay.

C. Privacy Issues

Conducting measurements at end devices can have a detrimental and undesirable effect on the privacy of the monitored customers, as metrics available through the Android API are good enough to sniff on the customers habits. Unfortunately, most of the apps we install today in our smartphones have access to lots of information related to our private life. As an example, Figure 8 shows a simple map in which all the session QoE ratings provided by one of the participants of the field trial are geo-located using metric ID 11 (cf. Table I). Three regions concentrate the majority of the ratings of this participant, and these correspond to (A) his home, (B) his working office and (C) his daily train traveling activity. So even if the participant does not provide for example his home address, this can be easily retrieved from such simple measurements.

D. Network Neutrality

The last topic we address is the case of network neutrality and the identification of traffic differentiation through end-device measurements. End-device throughput measurements can be used to identify potential traffic differentiation policies done by an ISP, based on types of traffic. This is highly relevant, as many cellular operators are today tempted to mistreat some classes of traffic to discourage its usage or for other internal interests such as traffic engineering. As an example of identification of such a potential traffic differentiation, Figure 9 depicts the distribution of the downlink average flow throughput (metric ID 6, cf. Table I) for two participants of the field trial having a contract with two different ISPs. ISP 1 seems to treat differently the traffic corresponding to YouTube videos, as the flow throughput in the download is abruptly shaped down to 4 Mbps (see the slope in the CDF) whereas no shaping is observed for other traffic apps such as Gmaps. While we are not sure about the root causes of such a differentiation, a similar approach could be applied to understand and to assess the application of such policies by cellular operators.

[1]http://www.onavo.com/
[2]https://www.radioopt.com/

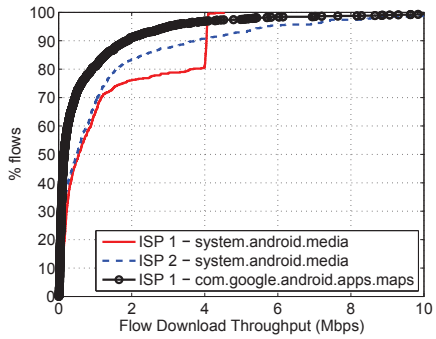


Figure 9. Network neutrality and identification of traffic differentiation. End-device throughput measurements can be used to identify potential traffic differentiation policies done by an ISP, based on types of traffic.

VI. Concluding Remarks

Smartphones are becoming the Internet-access devices by default, and we claim that network operators must understand how to manage and dimension their networks to correctly provision popular services accessed in smartphones, avoiding wasting additional unnecessary resources while keeping end users happy, and most importantly, reducing the chances of churning due to quality dissatisfaction. We have presented an overview on the QoE of different services and apps with different network-level QoS requirements for the specific case of smartphone devices, including both lab study results as well as measurements in the field. We have shown that the results obtained in the lab are highly applicable in the live scenario, as mappings track the QoE provided by users in real networks. Our results are highly relevant to future 5G design and LTE evolution in better understanding the mapping between network performance and customer experience. In addition, they provide hints and many insights about how and to which extent, end device measurements and QoE-based monitoring at end devices can be applied in the practice, complementing lab studies. We are aware that our results only tackle one side of the problem: the experience of the customers, from a very simple perspective: the downlink bandwidth. We agree with other researchers in that a more holistic perspective incorporating QoE, energy-consumption, data (re)transmission, and radio resource impact (among others) should be considered. This paper provides some initial components of such a holistic analysis. Finally, we are currently working on a deeper analysis regarding the impact of user location and mobility on field results. We also plan to better study the correlation between lab and field results.

References

[1] Cisco Visual Networking Index, "Global Mobile Data Traffic Forecast Update 2014-2019 White Paper", 2014.

[2] eMarketer Newsletter, "2 Billion Consumers Worldwide to Get Smart(phones) by 2016", 2014.

[3] P. Casas et al., "Exploring QoE in Cellular Networks: How Much Bandwidth do you Need for Popular Smartphone Apps?", in *ACM All Things Cellular Workshop*, 2015.

[4] A. Balachandran et al., "Modeling Web Quality of Experience on Cellular Networks", in *ACM MOBICOM*, 2014.

[5] M. Shafiq et al., "Understanding the Impact of Network Dynamics on Mobile Video User Engagement", in *ACM SIGMETRICS*, 2014.

[6] Q. Chen et al., "QoE Doctor: Diagnosing Mobile App QoE with Automated UI Control and Cross-layer Analysis", in *ACM IMC*, 2014.

[7] P. Casas et al., "Quality of Experience in Cloud Services: Survey and Measurements", in *Computer Networks*, vol. 68, 2014.

[8] P. Casas at al., "A First Look at Quality of Experience in Personal Cloud Storage Services", in *IEEE ICC Workshops*, 2013.

[9] T. Hoßfeld et al., "Quantification of YouTube QoE via Crowdsourcing", in *MQoE*, 2011.

[10] R. K. P. Mok et al., "Inferring the QoE of HTTP Video Streaming from User-Viewing Activities", in *W-MUST*, 2011.

[11] T. Hoßfeld et al., "'Initial Delay vs. Interruptions: Between the Devil and the Deep Blue Sea", in *QoMEX*, 2012.

[12] B. Lewcio et al., "Video Quality in Next Generation Mobile Networks – Perception of Time-varying Transmission," in *CQR*, 2011.

[13] M. Seufert et al., "A Survey on Quality of Experience of HTTP Adaptive Streaming", in *IEEE Communications Surveys & Tutorials*, 2014.

[14] G. Gómez et al., "YouTube QoE Evaluation Tool for Android Wireless Terminals," in *Journal on Wireless Communications and Networking*, vol. 164, 2014.

[15] V. Aggarwal et al., "Prometheus: Toward Quality-of-Experience Estimation for Mobile Apps from Passive Network Measurements," in *ACM HotMobile*, 2014.

[16] Int. Telecommunication Union, "Methods for Subjective Determination of Transmission Quality", *ITU-T Rec. P.800*, 1996.

[17] ——, "Subjective Video Quality Assessment Methods for Multimedia Applications", *ITU-T Rec. P.910*, 2008.

[18] ——, "Subjective Testing Methodology for Web Browsing", *ITU-T Rec. P.1501*, 2013.

[19] P. Casas et al., "YOUQMON: A System for On-line Monitoring of YouTube QoE in Operational 3G Networks," in *ACM SIGMETRICS PER*, vol. 41, 2013.

[20] A. Molavi Kakhk et al., "Identifying Traffic Differentiation on Cellular Data Networks", in *ACM SIGCOMM*, 2014.

[21] R. Schatz et al., "Vienna Surfing: Assessing Mobile Broadband Quality in the Field", in *ACM SIGCOMM W-MUST*, 2011.

[22] C. Kreibich et al., "Netalyzr: Illuminating the Edge Network", in *ACM IMC*, 2010.

[23] H. Yao et al., "Mobilyzer: An Open Platform for Controllable Mobile Network Measurements", in *ACM MobiSys*, 2015.

[24] Mobiperf, Measuring Network Performance on Mobile Platforms, https://sites.google.com/site/mobiperfdev/home.

[25] F. Wamser et al., "YoMoApp: a Tool for Analyzing QoE of YouTube HTTP Adaptive Streaming in Mobile Networks", in *European Conference on Networks and Communications*, 2015.