Next to You: Monitoring Quality of Experience in Cellular Networks From the End-Devices

Pedro Casas, Michael Seufert, Florian Wamser, Bruno Gardlo, Andreas Sackl,
and Raimund Schatz, *Member, IEEE*

***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 laboratory tests with end-device passive measurements and QoE crowd-sourced feedback obtained in operational cellular networks. The study addresses the impact of both access bandwidth and latency on the QoE of five different services and mobile apps: YouTube, Facebook, Web browsing through Chrome, Google Maps, and WhatsApp. We evaluate the influence of both constant and dynamically changing network access conditions, tackling in particular the case of fluctuating downlink bandwidth, which is typical in cellular networks. As a main contribution, we show that the results obtained in the laboratory 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 laboratory tests, and operational networks.**

***Index Terms*—QoE, smartphones, end-device measurements, field trial, subjective lab tests, mobile apps, crowdsourcing, YoMoApp.**

Manuscript received December 9, 2015; accepted February 20, 2016. Date of publication March 2, 2016; date of current version June 8, 2016. The research leading to these results has received funding from the European Union under the FP7 Grant Agreement no. 318627, "mPlane" project, as well as from the EU ICT Project INPUT (H2020-2014-ICT-644672). The work has been partially performed within the framework of the projects ACE 3.0 and N-0 at the Telecommunications Research Center Vienna (FTW), and has been partially funded by the Austrian Government and the City of Vienna through the program COMET. The work has also been partly funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO4770/1-2, TR257/31-2 (OekoNet) and TR257/41-1. The associate editor coordinating the review of this paper and approving it for publication was A. Pras.

P. Casas, B. Gardlo, A. Sackl, and R. Schatz are with the FTW-The Telecommunications Research Center, Vienna 1220, Austria and also with AIT - Austrian Institute of Technology, Vienna 1220, Austria (e-mail: pedro.casas@ait.ac.at; bruno.gardlo@ait.ac.at; andreas.sackl@ait.ac.at; raimund.schatz@ait.ac.at).

M. Seufert and F. Wamser are with the University of Würzburg, Institute of Computer Science, Würzburg 97074, Germany (e-mail: seufert@informatik.uni-wuerzburg.de; florian.wamser@informatik.uni-wuerzburg.de).

Digital Object Identifier 10.1109/TNSM.2016.2537645

I. Introduction

SMARTPHONES are becoming the most typical mobile device to access Internet today. Recent projections [1] 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 [2], 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).

The standard approach to assess the performance of networks and services from a QoE end-user perspective is to conduct controlled lab experiments [3]–[5]. 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 [6]. 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.).

In this paper we study the QoE of popular apps in smartphones (YouTube, Facebook, Gmaps, Web Browsing and WhatsApp) 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 firstly the impact of the most relevant QoS-based characteristics of the access network: the downlink bandwidth. In addition, we take two relevant network-related metrics into the study, evaluating the network access latency, and the *stability* of the cellular network. Given the natural mobility context in which users operate smartphones in cellular networks, we evaluate both constant and dynamically changing network bandwidth conditions, tackling in particular the case of fluctuating downlink bandwidth. This is highly important and a major contribution, as the bandwidth of a cellular connection naturally fluctuates due to interference, handover, etc. We have 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 tool to measure initial playback delays, playback stallings, and video quality switches (induced by the adaptive video streaming protocols used by YouTube).

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. In addition, our study shows the benefits of monitoring the QoE directly from the end-users' devices, as it becomes also possible to include contextual information (e.g., location, mobility, etc.) into the QoE analysis.

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. Sec. IV describes the tools developed to measure QoE-related metrics directly at the end-devices. Sec. V describes the approach followed in the field trial and discusses the obtained results, particularly in terms of similarity to those obtained in the lab. Sec. VI discusses the obtained results and our main findings. Sec. VII overviews several implications, limitations and topics related to the passive monitoring of QoE at end-user devices, including privacy, network neutrality, and incentives among others. Finally, Sec. VIII concludes this work.

This work is an extended and more complete version of a recently published paper [7], and it elaborates on our recent studies on QoE for cellular networks [8], [9]. In particular, we extend the subjective lab studies by adding new services as well as evaluating the impact of other QoS-related metrics such as latency at the access and network stability in terms of bandwidth fluctuations. The paper additionally extends the field trial results by adding an analysis on the impact of user location/mobility on QoE for some of the evaluated services. We also include details on the development of a novel end-device application which passively monitors Key Performance Indicators (KPIs) of YouTube such as stallings and quality switches, and present application results in both the lab studies and the field trial. Last but not least, we extend the discussion and interpretation of results to provide more useful conclusions to the reader.

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 [10]. 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 [11], [12] 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 [13] 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 [14].

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 [15], Mobilyzer [16], and the Android version of Netalyzr [17]. 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 [18], 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 [19] 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 [20], 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 [21] in cellular networks, and video [22].

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. Other limitation of some of the proposed approaches is that they rely on active measurements only (e.g., [20]), 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.

Finally, the problem of analyzing the impact of network bandwidth fluctuations on QoE has received little attention in the past, but we are giving strong steps in this direction, to make researchers and practitioners aware of the relevance of

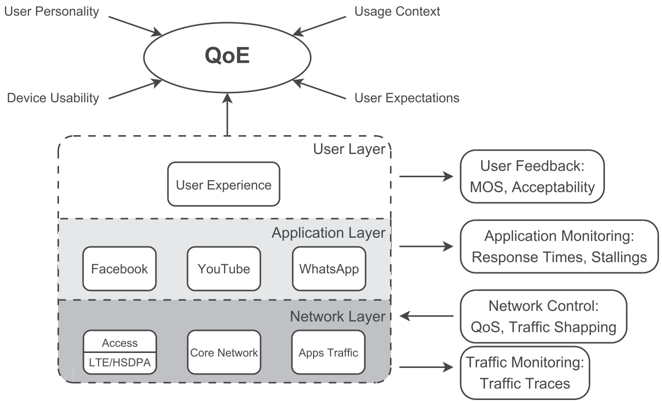


Fig. 1. Layered QoE evaluation methodology for networking services.

this issue. In particular, we have presented in [23] a study on the impact of network bandwidth fluctuations and network outages on the QoE of web-based services, using subjective lab studies and measurements in fixed-line networks.

III. Mobile QoE in the Lab

Let us begin by reporting the results of the conducted subjective lab tests. Lab tests are realized through the layered evaluation methodology depicted in Fig. 1. The experience of a user with any application is conditioned by multiple features, including dimensions such as technical characteristics of the application, user personality and expectations, user demographics, device usability, and usage context among others. Particularly when evaluating networking-based applications, the influence of the network itself as well as its interplay with the particular application have to be linked to the user's opinions, additionally identifying those perceivable performance parameters that are most relevant to the user experience. This mapping is realized by analyzing and correlating the three layers depicted in Fig. 1: the *network layer* accounts for the influence of the network QoS parameters (e.g., network bandwidth, RTT, etc.); the *application layer* considers both the technical characteristics (e.g., screen resolution, video bit-rate, web-page complexity) and the perceivable performance parameters of the application (e.g., page-load times, response time, video stalling, etc.); finally, the *user layer* spans the user subjective opinions on the evaluated application (e.g., MOS values, acceptability, etc.). The experimental evaluations reported in this section are designed in such a way that all the three aforementioned layers could be properly measured. In particular, there is a strong emphasis on monitoring part of these layers directly at the end-user device, enriching as such the contextual-information gathering and the visibility on the QoE monitoring problem, as we come as close as possible to the user and applications running on his device.

The subjective study consists of 52 participants interacting with the aforementioned services while experiencing different downlink bandwidth and access delay profiles in the background data connection. Fig. 2 depicts a high-level diagram of the experimental testbed employed in the subjective tests. 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 [24] so as to control the different access network profiles under evaluation. Next we report firstly the results obtained in terms of constant downlink bandwidth (section III-A), and then focus on the impact of access latency (section III-B) and dynamically changing downlink bandwidth, additionally including outages (section III-C).

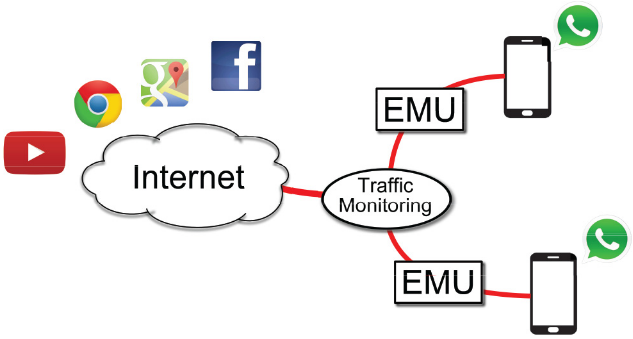


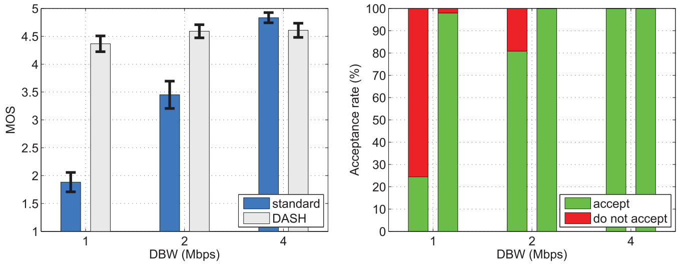
Fig. 2. Experimental setup used in the study. Devices are connected to the Internet through independent, controlled WiFi connections.

A. QoE for Constant Downlink Bandwidth

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 [10] for different access network technologies (LTE, 3G/2G, etc.). Note that while we do not emulate the particular characteristics of a cellular access network, results obtained in the field (c.f. Sec. V) suggest that our lab results are accurate in real cellular access networks.

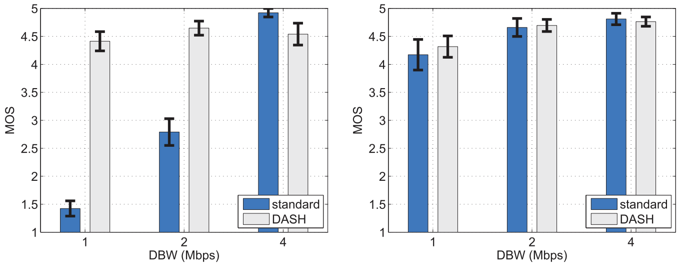
Participants were instructed to perform independent tasks for each of the three considered applications. For YouTube, they were requested to watch two-minutes HD YouTube videos, considering both the usage of the standard (i.e., non-DASH) and the DASH versions of the YouTube player. Videos correspond 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 [3]–[5]. 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,



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

Fig. 3. 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.

Fig. 4. 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.

and 70% of the participants have completed university or baccalaureate studies.

Regarding QoE feedback, participants were instructed to rate their *overall experience* according to a continuous ACR Mean Opinion Score (MOS) scale [3], 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, three 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); (iii) *initial playback delay annoyance* (did you perceive the initial loading time of the video as disturbing?). 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 [10].

1) QoE in YouTube Mobile: The Downlink BandWidth (DBW) takes values 1 Mbps, 2 Mbps, and 4 Mbps in YouTube tests. Fig. 3 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.

Fig. 3(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 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 [25], 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.

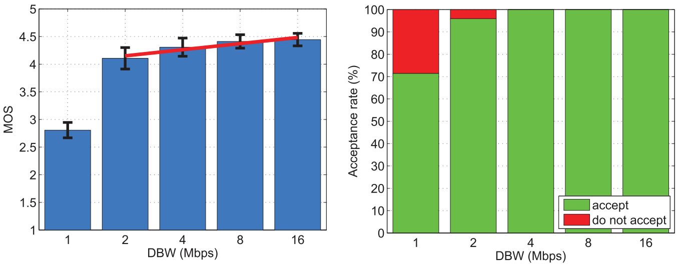
Fig. 3(b) reports the results in terms of acceptability of the participants. This is one of the key features that an operator has to consider, because low acceptance rate may sooner or later turn into churn. As observed, acceptance rate is as low as 23% for the standard streaming at 1 Mbps, whereas it’s close to 99% in the case of DASH.

To complement the picture for YouTube QoE in mobile devices, Fig. 4 depicts the results obtained in terms of (a) annoyance caused by stalling (stop of the video playback), and (b) video image quality. In Fig. 4(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 Fig. 4(b), which reports the perceived image quality of the video. According to previous studies [13], 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.

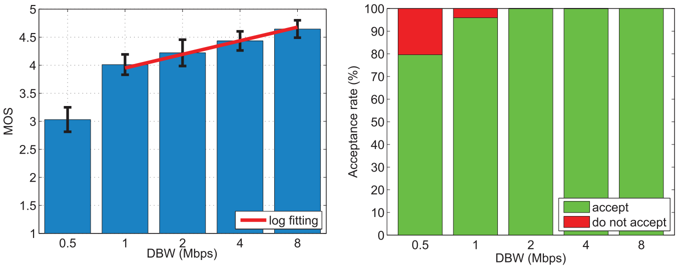
2) QoE in Gmaps and Facebook Mobile: Gmaps is tested with a fully logarithmic scale: 1 Mbps, 2 Mbps, 4 Mbps, 8 Mbps, and 16 Mbps. Fig. 5 reports the overall quality and acceptability results obtained for the Gmaps tests. Fig. 5(a) shows that a DBW of 4 Mbps results in near optimal QoE (MOS $\approx$ 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. Fig. 6 reports the results obtained



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

Fig. 5. 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.



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

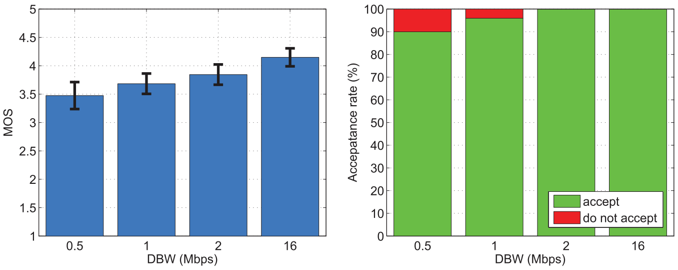
Fig. 6. QoE in Facebook. Overall quality and acceptability for different DBW. A DBW of 1 Mbps is high enough to achieve good QoE and almost full acceptability.

in 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. Fig. 6), 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.

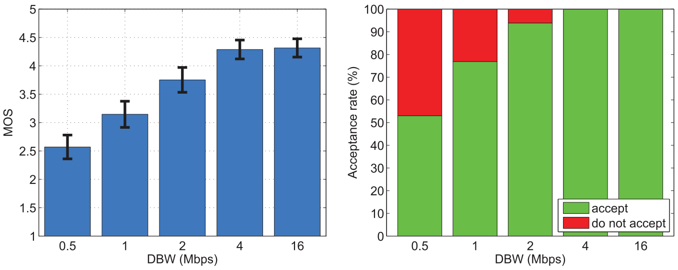
3) QoE in Mobile Web Browsing and WhatsApp: Web browsing is tested with DBW = 0.5 Mbps, 1 Mbps, 2 Mbps, and 16 Mbps. Fig. 7 reports the overall quality and acceptability results obtained for the News website browsing tests. Note first how the quality increases in a logarithmic fashion with increasing values of the DBW. Good experience (MOS $\approx$ 4) is obtained for a DBW of 2 Mbps, and only slight QoE differences are obtained when increasing the bandwidth to up to 16 Mbps, going to MOS $\approx$ 4.15. Going in the DBW decreasing direction, the slowest tested condition still results in fair quality (MOS $\approx$ 3.5) and high acceptance rate, close to 90%.

For WhatsApp, we add an additional test at DBW = 4 Mbps, given the file sizes used and the occurrence of saturation. Fig. 8 shows the QoE results for different DBW values. Users tolerate



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

Fig. 7. QoE in Web browsing (news website). Overall quality and acceptability for different downlink bandwidth configurations.



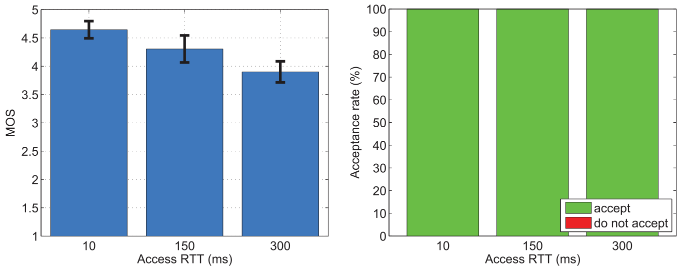
(a) Overall quality (MOS). (b) Acceptance rate.

Fig. 8. QoE in WhatsApp. Overall quality and acceptability for different downlink bandwidth configurations.

WhatsApp downloads with a good overall experience and high acceptability as long as the DBW is above 2 Mbps, but user experience heavily degrades for slower connections, resulting in very bad quality for a DBW of 500 kbps. In this case, a DBW threshold of 2 Mbps permits to approximately discriminate between good and bad experience. Given the file size used in the tests (5 MB), there is a clear saturation effect after 4 Mbps, as QoE does not increase for higher DBW values. Finally, even if the obtained results are partially biased by both the specific file size used in the tests and the participants task briefing, obtained results are similar to those we obtained in [26] for the specific case of Dropbox file sharing, suggesting that the main take aways are potentially more generic than expected when considering file downloads, either in mobile devices or in fixed ones.

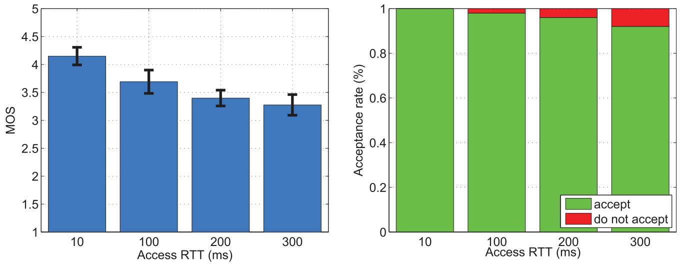
B. QoE for Access RTT

Constant access RTT profiles are tested for two out of the five studied services: Web browsing and Facebook. Our decision to only focus on these two services is based on the findings of previous work [27] stating that network delay is one of the most impacting network features on such type of interactive services. In addition, we were bounded to the maximum number of tests that could be run with participants without causing a degradation on the quality of the results due to fatigue. In both cases, access RTT is increased from an optimal condition (RTT = 10 ms) to a very slow access network scenario, considering a maximum access RTT of 300 ms. RTT profiles are selected from operational experience. In particular, RTT in operational LTE and HSPA networks is close to 50 ms [28], whereas 500 ms are common values observed on EDGE scenarios.



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

Fig. 9. QoE in Facebook. Overall quality and acceptability for different access RTT configurations.



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

Fig. 10. QoE in Web browsing (news website). Overall quality and acceptability for different access RTT configurations.

1) QoE in Facebook Mobile: Fig. 9 shows that the QoE degrades when the access RTT increases far beyond 150 ms, but the impact is not as significant as one might expect a-priori for a browsing-like application, and acceptance seems not to be impacted at all. Indeed, overall quality remains almost optimal for an access RTT of 150 ms, suggesting that the new evaluations of ultra-low latency cellular networks are not really necessary for applications such as Facebook Mobile. The main reason for such a result is that the degree of interactivity of the Facebook application is not as high as for other applications such as video-conferencing or gaming, suggesting that all in all, the operator should focus the dimensioning on the downlink bandwidth rather than the access RTT for this type of application. Still, Facebook is not the most network resources demanding application, so the dimensioning should probable not be done based on its specific latency requirements, as we see next.

2) QoE in Mobile Web Browsing: As reported in Fig. 10, the impact of access RTT is more marked in the case of Web browsing. QoE rapidly degrades when the access RTT increases above optimal values, and a MOS close to 3.6 (fair quality) is obtained for an access RTT = 100 ms. Still, acceptance rate is only slightly affected by the increasing RTT, suggesting that even if users can rapidly notice a non-responsive access network when browsing standard web pages in a smartphone, they still agree on using the application. A very interesting observation is that bigger access RTTs do not necessarily result in a highly increased QoE degradation, which is probably linked to the local caching and rendering techniques used by web browsers in mobile devices.

C. QoE under Bandwidth Fluctuations

As we have recently shown in [23] and as we see next, the experience of a user for certain applications is very sensitive to bandwidth fluctuations. Throughput fluctuations due to bandwidth variation are very common in cellular networks, but unfortunately, its QoE-effect is not captured in today's network measurements, as only average throughput values are typically considered. To better understand the QoE of mobile services under bandwidth fluctuations, we tested two types of bandwidth fluctuation patterns: periodic increase/decrease of downlink bandwidth, and downlink bandwidth outages, where bandwidth suddenly drops to zero, mimicking a disconnection scenario.

In particular, we tested the following downlink bandwidth profiles in YouTube, Web browsing, and Gmaps: periodical increase from 1 Mbps to 3 Mbps in YouTube (we refer to this profile as "1/3"), 3 times per minute for 5 second periods (Average Downlink Bandwidth, ADW = 1.5 Mbps); periodical drops from 4 Mbps to 0 Mbps (we refer to this profile as "4/0"), twice per minute for 10 second periods in YouTube DASH only (ADW = 2.7 Mbps), and twice per minute for 15 second periods in Web browsing (ADW = 2 Mbps); finally, a 7/1 profile (ADW = 4 Mbps), a 16/0 profile (ADW = 8 Mbps) and a 4/0 profile (ADW = 2 Mbps) in Gmaps, shifting bandwidth twice per minute for 15 second periods.

1) QoE in YouTube Mobile: Fig. 11 presents the results obtained in terms of (a) overall quality, (b) acceptance, and annoyance caused by (c) initial delays and (d) stalling. The 4/0 profile is only tested with the DASH flavor of YouTube, as the non-adaptive version provides too low quality results in the case of 10 second outages. The short-duration bandwidth increases do not have any significant impact on the QoE of both YouTube versions. Indeed, such a spiky bandwidth increase does not compensate for the low average downlink bandwidth, which causes the expected stalling impact for the non-adaptive application. The DASH version keeps offering optimal results, but interestingly enough, the acceptance rate slightly drops as compared to the 1 Mbps condition (cf. Fig. 3), which is probably caused by the additional quality changes triggered by fluctuations. When it comes to bandwidth outages (10 seconds-long), the reader can appreciate that even YouTube DASH can suffer from important QoE degradations when throughput drops to zero for short periods. The YouTube DASH version is not predictive, and quality switches respond to current bandwidth estimations. Given the image quality results reported in Fig. 4(b), a good way to avoid QoE degradations in the case of outages would be to preemptively caching as many low-quality video chunks as possible when the bandwidth is above certain predefined threshold.

2) QoE in Gmaps Mobile and Web Browsing: Fig. 12 reports the overall quality and acceptance results obtained for the Web browsing and Gmaps tests. Figs. 12(a) and 12(b) show the impact of a 4/0 profile on Web browsing QoE. The interesting part comes when comparing the constant 2 Mbps bandwidth condition (cf. Fig. 7) with the outage bandwidth profile. While both conditions correspond to an average downlink bandwidth of 2 Mbps, the fluctuation profile 4/0 results in a much degraded

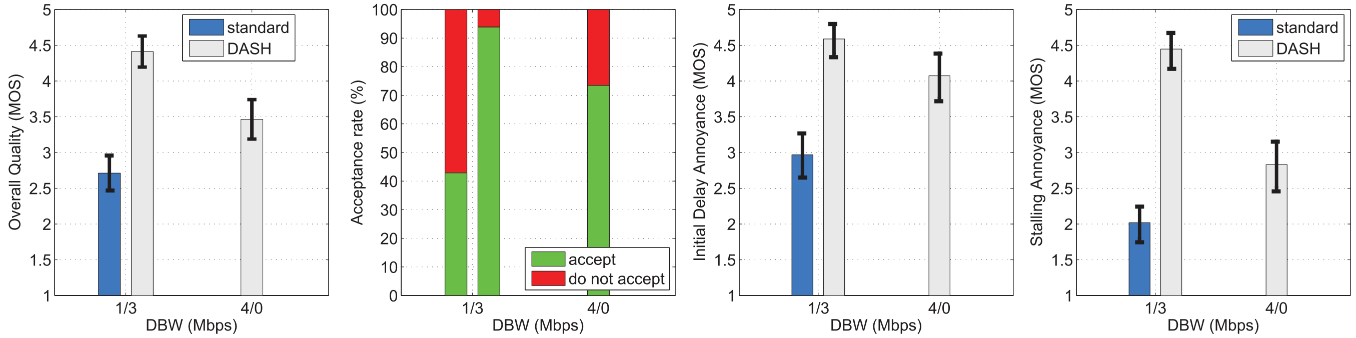


Fig. 11. QoE in YouTube under bandwidth fluctuations. Overall quality, acceptability, and annoyance caused by initial delays and stalling for 1/3 (average downlink bandwidth = 1.5 Mbps) and 4/0 (average downlink bandwidth = 2.7 Mbps) downlink bandwidth profiles.

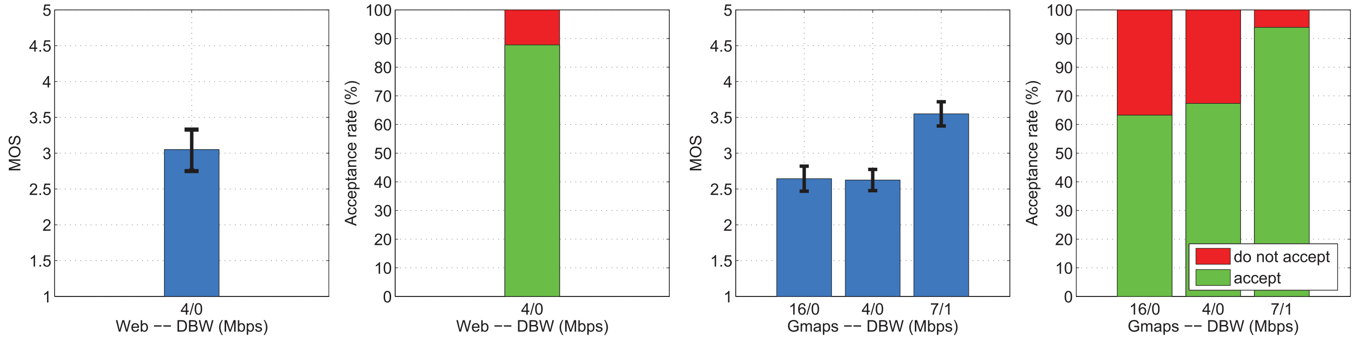


Fig. 12. QoE in Gmaps and Web browsing under bandwidth fluctuations. Overall quality and acceptability for different downlink bandwidth fluctuation profiles. Bandwidth outages have a very marked impact on the QoE of these services, and higher downlink bandwidth values do not compensate for such drops.

experience, with a MOS score dropping to 3, and an acceptance rate dropping to 88%.

Figs. 12(c) and 12(d) report the Gmaps results. Note first how both the 16/0 and the 4/0 outage profiles cause a very strong QoE degradation, with quality dropping to MOS $\approx$ 2.6 and acceptance rate to about 65%. When comparing to the constant bandwidth scenario, Gmaps QoE is actually near optimality and full acceptance for a downlink bandwidth higher than 2 Mbps (cf. Fig. 5), evidencing the important impact of the outages. Interestingly, results for both outage profiles are almost identical, even if the peak bandwidth values are very different, i.e., 16 Mbps and 4 Mbps respectively. This suggests that higher peak downlink bandwidth values do not compensate for the impact of outages.

The impact of outages on Gmaps is much stronger than in the case of the Web browsing, which is directly tied to the degree of interactivity of the application, which is much higher in Gmaps. Finally, the impact of the 7/1 profile is much less important as compared to the outages, but quality degradation is also very noticeable. An important take away from this evaluation is that the average downlink bandwidth is not as informative as one might expect when considering QoE in mobile devices, as results can greatly change, depending on the specific bandwidth profile.

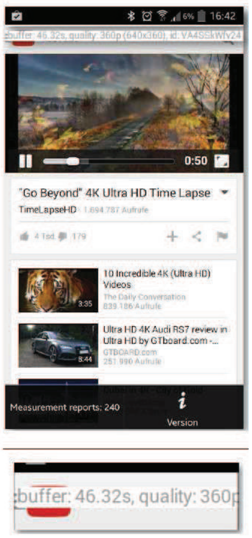
IV. End-Device Monitoring Tools

To monitor the traffic of the field-trial participants and to log their QoE feedbacks, we developed three specific Android-based applications. The first one is YoMoApp [9], an application which passively monitors QoE-relevant KPIs of YouTube adaptive video streaming on end-user smartphones. The second tool consists of a passive, flow-level traffic monitor, capable of sniffing all the incoming and outgoing traffic, additionally labeling the corresponding flows according to the application generating the traffic. The final tool consists of a web-based app which permits users to provide feedback on their experienced quality. We describe these tools next.

A. The YoMoApp Tool

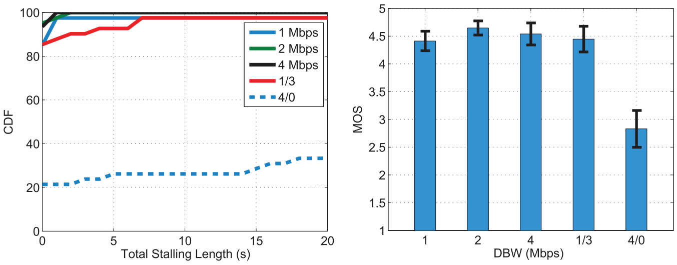
The goal of the tool is to monitor application layer KPIs of YouTube that have a high correlation with the actual QoE of mobile app users. As we said before, the main influence parameters of the YouTube QoE are stallings and video quality. To obtain these parameters, we monitor the buffer filling levels and the resolution of the YouTube videos.

YoMoApp works as follows. The original YouTube app is fully replicated in functionality and design, see Fig. 13. To this end, existing libraries from YouTube are used that are available for YouTube developers. An Android web view browser element is embedded for the YouTube video playback, such that HTML5 video playback is possible, including adaptive streaming according to the MPEG DASH approach of YouTube. Additional functions are added, which ultimately perform the monitoring of the application parameters in the newly created app. The monitoring is done at runtime via JavaScript, which queries the embedded HTML5 $\langle video \rangle$ object. In Fig. 13, the utilized parameters are listed. Note that the obtained parameters can be displayed in YoMoApp for validation, but are


16:42
6%
buffer: 46.32s, quality: 360p (640x360), id: VA4SSkWfv24
0:50
"Go Beyond" 4K Ultra HD Time Lapse
TimeLapseHD 1.694.787 Aufrufe
4 Tsd. 179
10 Incredible 4K (Ultra HD) Videos
The Daily Conversation
839.186 Aufrufe
3:35
Ultra HD 4K Audi RS7 review in Ultra HD by GTboard.com -...
GTBOARD.com
251.990 Aufrufe
8:44
Measurement reports: 240
Version
buffer: 46.32s, quality: 360p

Name	Description
buffered	List of time ranges of the media content that have been buffered
height/ width	Height and width of the video's display area in CSS pixels
played	Object indicating all the ranges of the video that have been played
currentTime	Current video playtime
youtubeId	Object indicating YouTube identifier of the video content
totalVideoFrames	Total number of frames that would have been displayed if no frames are dropped
droppedVideoFrames	Total number of frames dropped predecode or dropped because the frame missed its display deadline
corruptedVideoFrames	A Total number of corrupted frames that have been detected
Timestamp	Timestamp of the data query
Name	A pre-defined device name
Session timestamp	A session timestamp to identify the YouTube session.

Fig. 13. Screenshot of the app and selected parameters from the HTML5 ⟨*video*⟩ object, Media Source Extensions, and device, which can be tracked by the app.



(a) Total stalling time.

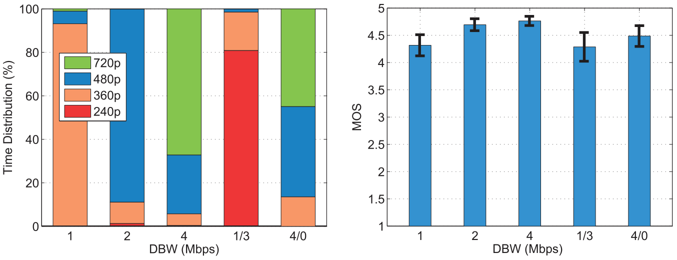
(b) Impacts of stallings on QoE.

Fig. 14. Monitoring of stallings and their impact on QoE with YoMoApp.

usually hidden. We are preparing a publicly available version of YoMoApp, which shall be soon available to download from the Android Google Play apps store.

To show the applicability of YoMoApp in the practice, we employed YoMoApp in the previously presented subjective lab study, tracking the performance of YouTube in the DASH version. Fig. 14(a) shows the distribution of the total stalling time for each of the bandwidth-related tested conditions. Almost no stalling occurs for the constant bandwidth conditions. Stalling occurs in about 14% of the variable DBW = 1/3 conditions, ranging up to a total stalling time of 34 s. The outage scenario (DBW = 4/0) is the one more impacted by stalling, as more than 75% of the tests result in video stalling. The average total stalling time in this condition is 25 s, with a maximum of up to 41 s. Fig. 14(b) shows the corresponding MOS values in terms of stalling annoyance (same results presented in previous Sec., but condensed in one single Fig. for better interpretation), which are very in-line with the stalling distribution as tracked by YoMoApp.

Fig. 15(a) shows the percentage of time on each quality level per condition, i.e., the percentage of time which each video resolution was played out during the streaming. The three constant bandwidth conditions at 1 Mbps, 2 Mbps and 4 Mbps result in a straightforward mapping to video resolution, resulting in a major share of 360p, 480p and 720p resolution respectively. The outage condition has similar quality shares to the 4 Mbps one, which is not surprising considering that it is a 4/0 Mbps on/off pattern. The variable 1/3 condition contains a large percentage of the lowest resolution, which indicates that the YouTube adaptation is very conservative when the network conditions fluctuate considerably. Finally, Fig. 15(b) shows the resulting image quality MOS values as rated by participants, confirming once again that resolution adaptation does not have a relevant impact on the subjectively perceived image quality in smartphones, given the small screen size.



(a) Distribution of quality levels.

(b) Video image QoE.

Fig. 15. Monitoring of video quality switches and the resulting image QoE with YoMoApp.

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

B. Passive Traffic Monitoring and QoE Feedback

The passive 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., [15]–[17]), as these either rely on active measurements only or are too specific for their original purpose.

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

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`

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.

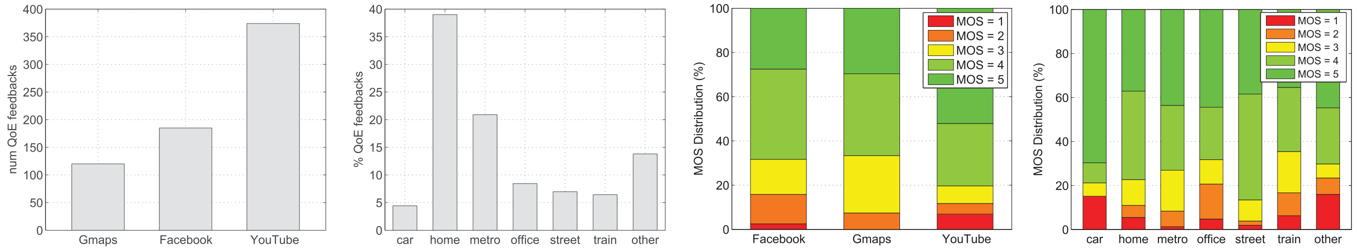
The final step is to define a proper session-based KPI which could be used to correlate sessions and MOS scores. Recall that the results presented for the lab study considered the downlink bandwidth as the independent network feature being tested in terms of QoE. Hence, we would define a KPI that tries 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 throughput 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 analysis of the field-trial measurements, we analyze the results obtained by correlating the MOS scores and the corresponding session MFT values for three of the tested apps (YouTube, Facebook and Gmaps).

V. 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. For the sake of brevity, we focus on only three out of the five applications tested in the lab, as drawn conclusions remain unchanged. Also, given the complexity of the problem, the study considers only the impact of the downlink bandwidth for the field trial scenario. We plan to extend the analysis to the monitoring of bandwidth fluctuations and access latency in the future.

A. Field Trial Overview

The field trial consisted of 30 participants using their own smartphones and cellular ISPs to access the same apps



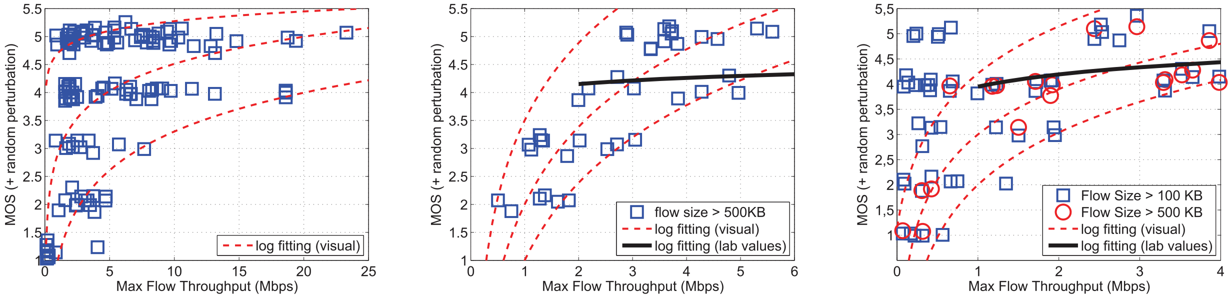
(a) Ratings per App.

(b) Ratings per location.

(c) MOS dist. per App.

(d) MOS dist. per location.

Fig. 16. 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.



(a) YouTube QoE vs. MFT.

(b) Gmaps QoE vs. MFT.

(c) Facebook QoE vs. MFT.

Fig. 17. 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.

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 generated by the participants were passively monitored with the tools described in previous section, including the monitoring of YouTube performance at the application layer. 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.

Fig. 16 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 for YouTube, Facebook and Gmaps. 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 [6]. 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.

Fig. 16(c) and Fig. 16(d) report the MOS scores distributions. Surprisingly, the MOS distributions are rather similar, both when considering the tested apps (cf. Fig. 16(c)) and the selected locations (cf. Fig. 16(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. QoE in the Field

Fig. 17 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).

Fig. 17(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

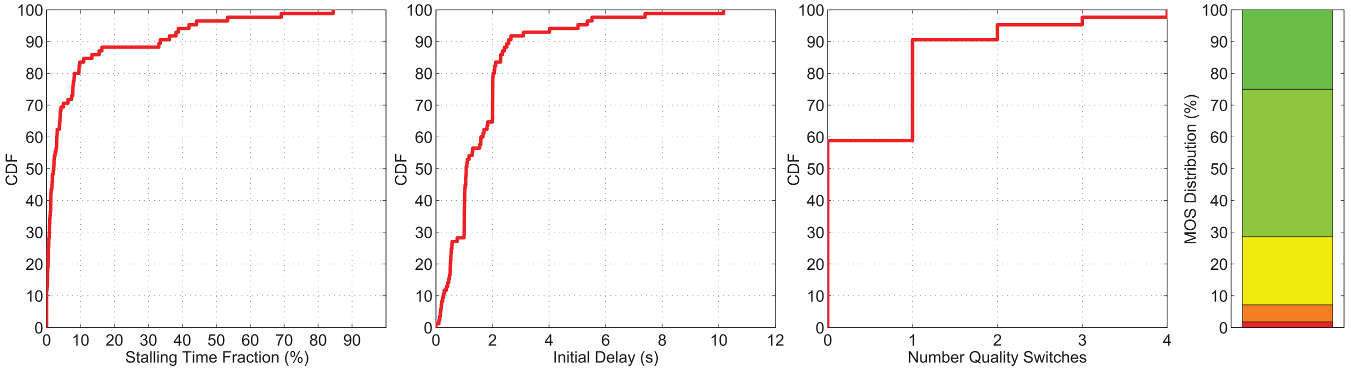


Fig. 18. Using YoMoApp to monitor YouTube in smartphones. Results correspond to the monitoring of one single participant. Quality is good (i.e., MOS $\geq$ 4 4) for about 70% of the video sessions, with an initial playback delay below 2 seconds, a total stalling fraction below 4%, and with almost no quality switches for these sessions.

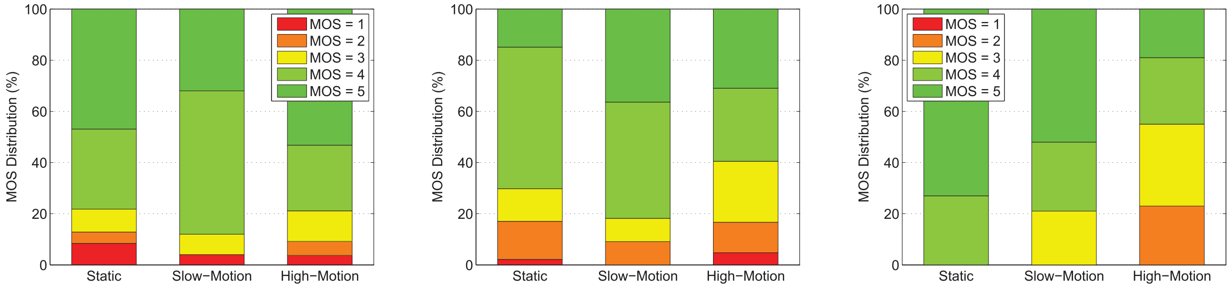
observed in the lab study (cf. Fig. 3), where QoE is optimal for a DBW > 4 Mbps. In addition, most of the sessions having very poor QoE (i.e., MOS = 1) have a very low MFT. However, as expected, the picture becomes very fuzzy in the most relevant MFT gap, between 1 Mbps and 4 Mbps, having MOS scores 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 Fig. 3, 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.

Fig. 17(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 Fig. 17(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. Fig. 5(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 Fig. 5(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.

Fig. 17(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 Fig. 17(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 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. 6. 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, using YoMoApp.

Fig. 18 shows the results obtained by using YoMoApp for one selected participant. A complete analysis of the field test results with YoMoApp is still ongoing, but nevertheless, we present example results to better motivate the usefulness of YoMoApp. The Fig. shows the distribution of the total stalling time as a fraction of the video length, the distribution of the initial delay, the distribution of the tracked quality switches, and the distribution of overall quality MOS values. Quality is good (i.e., MOS $\geq$ 4) for about 70% of the video sessions, with a total stalling time fraction below 4%, an initial playback delay below 2 seconds, and with almost no quality switches for these sessions. For the remaining 30% of the sessions rate as average or worse, there is a marked increase in the total stalling time fraction and initial delay, with about 10% of the sessions with



(a) YouTube QoE vs. Mobility.

(b) Facebook QoE vs. Mobility.

(c) Gmaps QoE vs. Mobility.

Fig. 19. Potential impact of mobility on overall QoE. Mobility patterns are constructed based on location as declared by participants when performing the evaluation tasks. *Static* refers to locations "home" and "office", *slow-motion* refers to location "street", whereas *high-motion* refers to locations "car", "metro" and "train".

4 or more seconds of initial delay and a stalling time fraction above 30%.

C. Impact of Context on QoE - the Case of Mobility

To conclude with the analysis of the field trial results, we present an evaluation on the impact of context on QoE, considering the specific case of mobility. The overall results presented in Fig. 16(d) do not reveal a major impact of the location (and potentially the associated degree of mobility) on the reported QoE for the considered applications. However, by taking a closer look into the results of each application, and by doing some raw hypothesis on the relation between location and degree of mobility, we can obtain some interesting results. Our hypothesis is as follow: we assume that participants at in-door locations are in a static situation when conducting the tests, walking while conducting the tests at the street, and moving while taking the tests at a train, underground or even car. We verified the accuracy of such an hypothesis by directly asking to some of the participants of the field trial, but we are not 100% sure that it applies to all performed tests, and additional filtering based on passively tracked location would be required to get better results. In any case, the initial results provided next are in line with our expectations and provide a first look into the problem.

Fig. 19 depicts the distribution of overall QoE values reported by participants for the three analyzed applications, grouped by degree of mobility. We consider three different mobility patterns: *static* refers to locations "home" and "office", *slow-motion* refers to location "street", whereas *high-motion* refers to locations "car", "metro" and "train". As before (cf. Fig. 16(d)), there is no apparent impact of mobility on the QoE results for (a) YouTube and (b) Facebook; this is most probably linked to the potentially good networking QoS offered by cellular networks in Vienna, but also to the degree of interactivity of these two applications. For example, recalling the impact of bandwidth fluctuations on YouTube QoE reported in Fig. 11, even short network outages might remain unnoticed when watching YouTube videos, thanks to the pre-buffering done by the app. However, results are much more interesting when considering a highly interactive application such as Gmaps in Fig. 19(c). In this case, there is a clear difference on the QoE distributions when considering static and a high-motion mobility patterns, with a much worse quality when moving faster. Indeed, note that the ratio of fair and bad QoE values goes from 0% for a static context to more than 50% when moving on a car, train or underground. These results are further confirmed by the results obtained in the lab when considering bandwidth fluctuations (cf. Fig. 12), which show how sensitive might be Gmaps when network conditions do not remain stable. A further and deeper analysis on the impacts of contextual information, particularly including mobility, are part of our ongoing work.

VI. Findings and Discussion

In this section we provide some additional discussion on the obtained results. 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

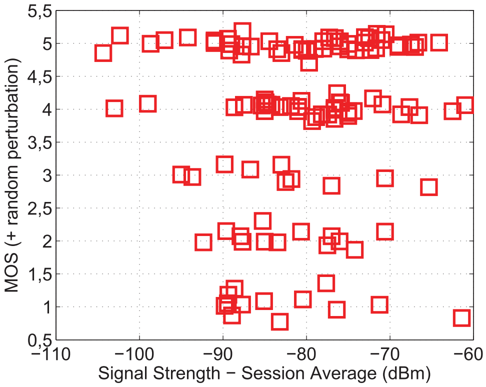


Fig. 20. MOS vs signal strength in YouTube. The signal strength metric corresponds to the average single strength when considering all the flows of a single session. There is no apparent correlation between the MOS declared by participants and the measured average signal strength.

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. We believe that the YoMoApp tool will play a key role in the short-term future to address this issue.

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, Fig. 20 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.

Let us now focus on some additional relevant aspects worth to comment on. In particular, we further elaborate on four specific topics of the study: (i) access network latency; (ii) downlink bandwidth fluctuations; (iii) downlink bandwidth outages and (iv) contextual information tracking.

A. *Do we Need Super Responsive Networks Today?*

Finding: even if we have only tested the impact of the access latency on Facebook and Web browsing, we have seen that the access RTT should be kept below 100 ms to achieve good user-perceived quality and high acceptability.

Implications: this means that super low latency access networks such as LTE are not needed today for the tested mobile applications. Still, we expect that more interactive applications such as Gmaps would require lower access RTTs, and thus believe that highly responsive networks would soon become highly relevant in terms of QoE-provisioning for mobile devices.

B. *Fast and Responsive, or Stable?*

Finding: we have shown that downlink bandwidth fluctuations can have an important impact on the experience of the end user, particularly when using mobile devices.

Implications: this finding has two major implications for the cellular network operator: (i) firstly, it evidences that faster and more responsive cellular networks should not be the only guidelines to follow when designing and dimensioning their networks, but that stability in terms of bandwidth, an even if we did not evaluate it, also in terms of latency, should be a major concern; (ii) secondly, when it comes to monitor and measure throughput in todays' cellular networks, operators should realize that traditional KPIs (Key Performance Indicators) based on average throughput are not as informative as have been assumed so far, and should evolve their monitoring systems to capture such fluctuations.

C. *Keep Connected*

Finding: we have found that even short-duration bandwidth outages (i.e., drops to 0 Mbps for some milliseconds) have a major negative impact on the experience of the end user.

Implications: these results suggest that besides targeting more stable cellular networks, a major effort would have to be carried in the near future in terms of multi network technology convergence. Indeed, the usage of multiple types of access technologies either in parallel or to perform fast handovers would become an integral part of the future 5G network, and our results suggest that such transitions should be done without impacting the connectivity of the device, not even for a few seconds.

D. *Context Matters*

Finding: last but not least, even if only preliminary, we have found that mobility plays a key role in the quality experienced by users, at least in the tested cellular networks, and for highly interactive applications such as Gmaps.

Implications: it is generally agreed among the QoE research community that context is critical when assessing the quality of an applications from the eyes of the end user, and our preliminary findings suggest to cellular ISPs that they have to consider means to catch as much contextual information as possible to take better conclusions, and therefore more informed decisions, about the QoE of their customers.

VII. Implications and Perspectives

The last part of the paper 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. In particular, we address four main topics: crowdsourcing for QoE analysis, incentives to achieve large participation of end-users, privacy issues related to measurements at end-devices, and additional perspectives from end-device measurements.

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.

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, Fig. 21 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.

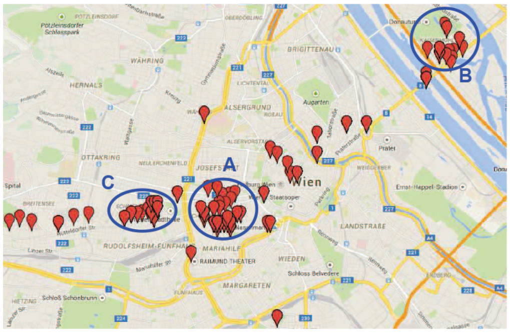


Fig. 21. End-device location monitoring and privacy issues. End-user activity and private location can be guessed by simply measuring the location of the cell where smartphone are attached to. In this example scenario, participants' home is located at region A, working office is located at region B, and high activity occurs at region C, linked to daily train traveling.

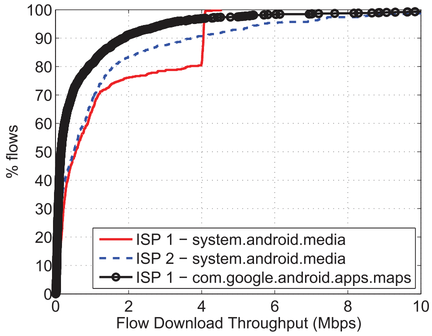


Fig. 22. Network neutrality and identification of traffic differentiation. End-device throughput measurements can be used to identify potential traffic differentiation policies done by an ISP, based on types of traffic.

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, Fig. 22 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

[1] http://www.onavo.com/

[2] https://www.radioopt.com/

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.

VIII. 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 believe that QoE has the potential to become the next guiding paradigm for managing quality provisioning and applications' design in cellular networks and mobile devices, and conducted an study shedding light in this direction.

We have presented an overview on the QoE of different services and applications 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. By considering both constant and dynamically changing network QoS conditions in our study, we have shown that downlink bandwidth fluctuations play a key role in determining the QoE of the evaluated services, specially for those highly interactive. We have also shown that dynamic applications such as YouTube DASH are much better suited to smartphone scenarios, providing the same level of experience to the non-adaptive version of the YouTube application, but with much lower QoS-based requirements in terms of downlink bandwidth. We additionally claim that the involvement of end users in the assessment process of the QoE in mobile devices is essential to obtain reliable QoE ground truths.

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. 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] eMarketer Newsletter. (2014). *2 Billion Consumers Worldwide to Get Smart (Phones) by 2016* [Online]. Available: http://www.emarketer.com/Article/2-Billion-Consumers-Worldwide-Smartphones-by-2016/1011694.

[2] Cisco. (2015). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020*, White Paper [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[3] Int. Telecommunication Union, "ITU-T Rec. P.800: Methods for subjective determination of transmission quality," 1996.

[4] Int. Telecommunication Union, "ITU-T Rec. P.910: Subjective video quality assessment methods for multimedia applications," 2008.

[5] Int. Telecommunication Union, "ITU-T Rec. P.1501: Subjective testing methodology for web browsing," 2013.

[6] R. Schatz and S. Egger, "Vienna surfing: Assessing mobile broadband quality in the field," in *Proc. 1st ACM SIGCOMM Workshop Meas. Up Stack (W-MUST'11)*, 2011, pp. 19–24.

[7] P. Casas, B. Gardlo, M. Seufert, F. Wamser, and R. Schatz, "Taming QoE in cellular networks: From subjective lab studies to measurements in the field," in *Proc. 11th Int. Conf. Netw. Service Manage. (CNSM)*, Nov. 2015, pp. 237–245.

[8] P. Casas, R. Schatz, F. Wamser, M. Seufert, and R. Irmer, "Exploring QoE in cellular networks: How much bandwidth do you need for popular smartphone apps?" in *Proc. 5th Workshop All Things Cell. Oper. Appl. Challenges (AllThingsCellular'15)*, 2015, pp. 13–18.

[9] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, "Understanding YouTube QoE in cellular networks with YoMoApp: A QoE monitoring tool for YouTube mobile," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw. (MobiCom'15)*, 2015, pp. 263–265.

[10] P. Casas and R. Schatz, "Quality of experience in cloud services: Survey and measurements," *Comput. Netw.*, vol. 68, pp. 149–165, 2014.

[11] T. Hofeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2011, pp. 494–499.

[12] R. K. Mok, E. W. Chan, X. Luo, and R. K. Chang, "Inferring the QoE of HTTP video streaming from user-viewing activities," in *Proc. 1st ACM SIGCOMM Workshop Meas. Up Stack (W-MUST'11)*, 2011, pp. 31–36.

[13] B. Lewcio, B. Belmudez, A. Mehmood, M. Wältermann, and S. Möller, "Video quality in next generation mobile networks: Perception of time-varying transmission," in *Proc. IEEE Int. Workshop Tech. Committee Commun. Qual. Reliab. (CQR)*, May 2011, pp. 1–6.

[14] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hobfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, Mar. 2015.

[15] Mobiperf. (2013). *Mobiperf, Measuring Network Performance on Mobile Platforms* [Online]. Available: http://mobiperf.com.

[16] A. Nikravesh, H. Yao, S. Xu, D. Choffnes, and Z. M. Mao, "Mobilyzer: An open platform for controllable mobile network measurements," in *Proc. 13th Annu. Int. Conf. Mobile Syst. Appl. Serv. (MobiSys'15)*, 2015, pp. 389–404.

[17] C. Kreibich, N. Weaver, B. Nechaev, and V. Paxson, "Netalyzr: Illuminating the edge network," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas. (IMC'10)*, 2010, pp. 246–259.

[18] G. Gómez, L. Hortigüela, Q. Pérez, J. Lorca, R. García, and M. Aguayo-Torres, "YouTube QoE evaluation tool for android wireless terminals," *EURASIP J. Wireless Commun. Netw.*, vol. 164, pp. 1–14, 2014.

[19] V. Aggarwal, E. Halepovic, J. Pang, S. Venkataraman, and H. Yan, "Prometheus: Toward quality-of-experience estimation for mobile apps from passive network measurements," in *Proc. 15th Workshop Mobile Comput. Syst. Appl. (HotMobile'14)*, 2014, pp. 18:1–18:6.

[20] Q. A. Chen *et al.*, "QoE doctor: Diagnosing mobile app QoE with automated UI control and cross-layer analysis," in *Proc. Conf. Internet Meas. Conf. (IMC'14)*, 2014, pp. 151–164.

[21] A. Balachandran *et al.*, "Modeling web quality-of-experience on cellular networks," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom'14)*, 2014, pp. 213–224.

[22] M. Z. Shafiq, J. Erman, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Understanding the impact of network dynamics on mobile video user engagement," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst. (SIGMETRICS'14)*, 2014, pp. 367–379.

[23] A. Sackl, P. Casas, R. Schatz, L. Janowski, and R. Irmer, "Quantifying the impact of network bandwidth fluctuations and outages on web QoE," in *Proc. 7th Int. Workshop Qual. Multimedia Exp. (QoMEX)*, May 2015, pp. 1–6.

[24] S. Hemminger, "Network emulation with NetEm," in *Proc. Linux Conf. Aust. (LCA'05)*, 2005, pp. 1–9.

[25] P. Casas, M. Seufert, and R. Schatz, "YOUQMON: A system for on-line monitoring of YouTube QoE in operational 3G networks," *SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 2, pp. 44–46, Aug. 2013.

[26] P. Casas, H. R. Fischer, S. Suette, and R. Schatz, "A first look at quality of experience in personal cloud storage services," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Jun. 2013, pp. 733–737.

[27] P. Casas, M. Seufert, S. Egger, and R. Schatz, “Quality of experience in remote virtual desktop services,” in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM’13)*, May 2013, pp. 1352–1357.

[28] M. Laner and P. Svoboda, P. Romirer-Maierhofer, N. Nikaein, F. Ricciato, and M. Rupp, “A comparison between one-way delays in operating HSPA and LTE networks,” in *Proc. 10th Int. Symp. Model. Optim. Mobile Ad Hoc Wireless Netw. (WiOpt)*, May 2012, pp. 286–292.

Pedro Casas received the electrical engineering degree from the University of the Republic, Montevideo, Uruguay, in 2005, and the Ph.D. degree in computer science from Télécom Bretagne, Plouzané, France, in 2010. He is a Scientist with the Austrian Institute of Technology (AIT), Vienna, Austria. He held Research and Teaching Assistant positions with the University of the Republic, between 2003 and 2012, and was at the French Research Laboratory LAAS-CNRS, Toulouse, France, as a Postdoctoral Research Fellow between 2010 and 2011. Between 2011 and 2015, he was a Senior Researcher with the Telecommunications Research Center Vienna (FTW), Vienna, Austria. His research interests include the monitoring and analysis of network traffic, network security and anomaly detection, QoE modeling and automatic assessment, as well as machine-learning and data mining-based approaches for Networking. He has authored more than 80 networking research papers (50 as main author) in major international conferences and journals. He was the recipient of the seven best paper awards in the last 6 years.

Michael Seufert received degrees in computer science, mathematics, and education from the University of Würzburg, Würzburg, Germany, the diploma degree in computer science (additionally passed the state examinations for teaching mathematics and computer science in secondary schools) in 2011. He is currently pursuing the Ph.D. degree at University of Würzburg. From 2012 to 2013, he was with FTW Telecommunication Research Center, Vienna, Austria, working in the area of user-centered interaction and communication economics. He is currently a Researcher with the Chair of Communication Networks, University of Würzburg. His research interests include the QoE of Internet applications, social networks, performance modeling and analysis, and traffic management solutions.

Florian Wamser received the diploma degree in computer science, in 2009, and the Ph.D. degree, in 2015. He studied at the University of Würzburg and at the Helsinki University of Technology, Espoo, Finland. He is a Research Associate with the Chair of Communication Networks, University of Würzburg, Würzburg, Germany. He leads the group on cloud networks and internet applications at the Chair of Prof. Dr.-Ing. Phuoc Tran-Gia. During his thesis studies, he worked on the topics characterization and modeling of application Internet traffic in broadband wireless access networks. The title of his dissertation is “Performance Assessment of Resource Management Strategies for Cellular and Wireless Mesh Networks.” His research interests include analytical and simulative performance evaluation and optimization of cloud networks and related fields.

Bruno Gardlo received the M.Sc. and Ph.D. degrees in telecommunications from the Faculty of Electrical Engineering, University of Zilina, Zilina, Slovakia, in 2009 and 2012, respectively. He is currently an Expert Advisor with the Austrian Institute of Technology (AIT), Seibersdorf, Austria. Before he joined AIT, he was working as a Researcher for Forschungszentrum Telekommunikation Wien, with specialization on QoE and crowdsourcing. His research interests include perceptual video and audiovisual quality evaluation, video and audiovisual metric design or user interface design and its effect on user experience, developing web applications for improving crowdsourcing efficiency, and reliability.

Andreas Sackl received the media and computer science degree from the Technical University of Vienna and Mass Media and Communication Science, University of Vienna, Vienna, Austria. He is currently pursuing the Ph.D. degree in computer science at the Technical University of Berlin, Berlin, Germany. He is a Scientist at Austrian Institute of Technology (AIT), Seibersdorf, Austria. Before working at AIT, he was a Researcher with FTW (Telecommunications Research Center Vienna) in the field of quality of experience and usability. He is the author of numerous workshop and conference papers and acts as Reviewer (QoMEX, TVX, ACM SIGCHI) and TPC member (PQS, QoE-FI).

Raimund Schatz (M’08) received the M.Sc. degree in telematics, TU-Graz, Graz, Austria, the Ph.D. degree in informatics, TU-Vienna, Vienna, Austria, the M.B.A. and M.Sc. degrees in international and finance management both from Open University, Milton Keynes, U.K. He has recently joined the Austrian Institute of Technology (AIT), business unit, Technology Experience. Until end of 2015, he was a Key Researcher with the Telecommunications Research Center Vienna (FTW), Vienna, Austria and the Manager of the User-Centered Interaction, Services, and Systems Quality Department. From 2009 to 2015, he lead FTW’s research projects on quality-of-experience assessment and monitoring of broadband services in wireless and wireline networks, conducted together with a number of industry partners in the telecom industry. Furthermore, he is or has been actively involved in various QoE-related EU projects and networking activities, including Optiband (FP7), CELTIC QuEEN and COST Action IC1003 (Qualinet) and IC1304 (ACROSS) as well as the organization of various QoE-related workshops and events (e.g., Special Session Chair for QoMEX 2013; Organizer of PQS 2013, General Chair of QoENAM 2014, QoE-FI 2015 and QoE-FI 2016). Being an active member of ACM, he is the author of more than 100 publications in the areas of quality of experience, HCI, and pervasive computing.