An Educated Guess on QoE in Operational Networks through Large-Scale Measurements

An Educated Guess on QoE in Operational Networks through Large-Scale Measurements

Pedro Casas (1), Bruno Gardlo (1), Raimund Schatz (1), Marco Mellia (2)
(1) AIT Austrian Institute of Technology, (2) Politecnico di Torino
(1) name.surname@ait.ac.at, (2) marco.mellia@polito.it

ABSTRACT

Downlink throughput is the most widely used and accepted performance feature within the networking community, specially in the operational field. Current network monitoring and reporting systems as well as network quality benchmarking campaigns use the Average Downlink Throughput (ADT) as the main Key Performance Indicators (KPIs) reflecting the health of the network. In this paper we address the problem of network performance monitoring and assessment in operational networks from a user-centric, Quality of Experience (QoE) perspective. While we have shown in the past that accurate QoE estimation requires measurements and KPIs collected at multiple levels of the communications stack – including network, transport, application and end-user layers, we take a practical approach and provide an educated guess on QoE using only a standard ADT-based KPI as input. We do so to maximize the utilization of throughput measurements currently collected with common network traffic monitoring systems. Armed with QoE models mapping downlink bandwidth to user experience – derived from subjective QoE lab tests, we estimate the QoE undergone by customers of both cellular and fixed-line networks, using large-scale passive traffic measurements. In particular, we study the performance of three highly popular end-customer services, including YouTube video streaming, Facebook social networking, and WhatsApp multimedia sharing. Surprisingly, our results suggest that up to 33% of the observed traffic flows might result in sub-optimal – or even poor, end-customer experience in both cellular and fixed-line networks, for the monitored services.

Keywords

QoE; Subjective Lab Tests; Network Measurements; Cellular Networks; Performance.

1. INTRODUCTION

Quality of Experience (QoE) is becoming one of the leading concepts for network management and performance evaluation in operational networks. The intensifying competition among network operators is forcing Internet Service Providers (ISPs) to integrate QoE into the core of their network management systems, from network monitoring and reporting to traffic engineering. This need is even more relevant for cellular network operators, who need to offer high quality levels to reduce the risks of customers churning for quality dissatisfaction in a complex and bandwidth-restrictive context. Mobile users consume a wide variety of data services such as video streaming, social networks, web-browsing, VoIP and video calls, file-sharing, etc., all of them imposing different performance requirements to the network in terms of user experience. For example, bandwidth-intensive applications such as YouTube require high speed connections, whereas interactive applications such as Skype video-calling are additionally sensitive to network latency.

Traditional Key Performance Indicators (KPIs) reflecting network performance include network throughput, latency, packet loss, etc. In particular, downlink throughput is the most widely used and accepted metric in the operational field for network performance monitoring and reporting. Also in the case of cellular ISPs quality benchmarking, network performance drive tests report downlink throughput as the most relevant KPI revealing the health and performance of a cellular network. And even more, both end-customer Internet service offers as well as government regulatory-bodies targets are strongly – or even solely, based on downlink throughput.

When it comes to the experience of the end-customer, it is well recognized within the research community that application-layer metrics such as page load times in web-browsing, the number of re-buffering events in video distribution, the waiting times in file sharing, etc. define the key features to understand the network performance from a QoE perspective. However, monitoring such features in a large-scale basis is highly challenging and arduous in current networks. With the term *large-scale* we refer to the passive observation of the traffic generated by *all* the customers of an ISP – or at least a large fraction of them. Passive monitoring at the large-scale is typically performed in-network at the access and/or at the core of the ISP, and current massive adoption of end-to-end encryption (e.g., HTTPS adoption by YouTube, Facebook, etc.) makes it very difficult – or even not possible, to passively monitor application-layer metrics on those vantage points. For this reason, the tendency nowadays is to additionally monitor customers traffic directly at their end devices [1], to directly capture application-layer metrics as well as other relevant contextual information. End-device based monitoring is performed in a crowdsourcing basis, relying on the willingness of the end-user to install and run monitoring applications on their own devices. As such, end-device monitoring is less scalable (i.e., captures a small share of users) and less reliable (e.g., the vantage point is not under the control of the ISP) than in-network monitoring.

In this paper we take a practical approach to the problem of large-scale QoE monitoring in operational networks:

we estimate the QoE of popular end-customer services in both cellular and fixed-line networks, using as input the most readily available KPI, the Average flow Downlink Throughput (ADT). By doing so, we expect that our results would improve the visibility of operators on the QoE of their customers, without doing any modifications to their current standard monitoring systems. To achieve the goal, we rely on models mapping ADT to QoE for different services and different types of networks and devices, obtained from multiple subjective QoE lab tests we have performed in the past. QoE is service-dependent, thus it is not possible to build models mapping ADT to QoE for each of the potential services consumed by the customer. However, it is well known that a small number of services are responsible for the largest share of the traffic and users in any network – the mice and elephants phenomenon also applies to Internet services [2], thus limiting the study to the most popular services already gives the operator a pretty good estimation of the QoE undergone by its customers. We therefore study the three most popular services in western countries for both cellular and fixed-line networks: YouTube, Facebook, and WhatsApp.

We apply the derived mappings to large-scale flow measurements collected at both cellular and fixed-line EU ISPs in 2013 and 2014. The complete dataset consists of a full week of flow measurements from each network, aggregating thousands of customers and resulting in tens of millions of flows. Even if the QoE results provided by our study are indicative – we do not have the ground truth in terms of QoE for all the monitored customers, our estimations suggest that: (i) up to 30%/33% of the monitored flows might result in sub-optimal QoE (i.e., MOS scores below 3) for YouTube in cellular/fixed-lined networks respectively, (ii) this fraction increases to almost 40% in the case of both Facebook and media sharing through WhatsApp in the monitored cellular network, and (iii) bad quality events are likely to occur for about 15% and 18% of the monitored YouTube flows (cellular and fixed-line respectively), 20% of the Facebook flows and 25% of the WhatsApp flows. As we explain next, the proposed study is conservative and estimations are based on worst-case scenarios; still, our results evidence that poor quality events are far from negligible in both fixed-line and cellular networks for the studied services, pointing to the strong need of better network monitoring and traffic analysis KPIs reflecting the experience of customers in operational networks.

The main contribution of our study is to scale subjective QoE studies out of the lab to enhance the visibility of ISPs on the performance of their operational networks. We use practical, readily available and well understood downlink throughput-based KPIs to bridge large-scale network measurements to end-customer experience. We provide an assorted set of QoE mapping models and KPIs for different services, different specific contents (e.g., SD and HD video) and different end-devices (e.g., smartphone, PC/laptop). We collect and analyze large-scale measurements in two different operational networks. Finally, we discuss limitations of our approach in the concluding remarks.

The remainder of the paper is organized as follows: Sec. 2 presents an overview of the related work on QoE for the studies services and QoE in operational networks. Sec. 3 presents and discusses different models and KPIs mapping ADT to QoE, derived from past QoE subjective tests and recent updates. Sec. 4 describes the collected network measurements and presents the results obtained by combining the QoE models with the network measurements. Finally, Sec. 5 concludes this work, pointing to some limitations of our study and future work.

2. RELATED WORK

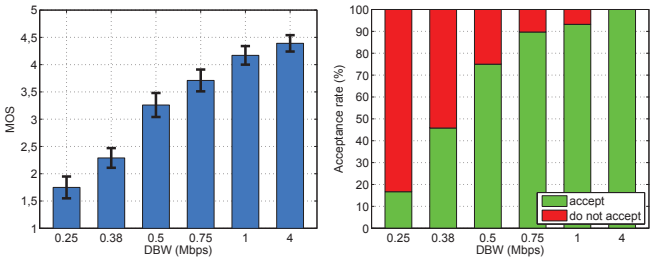
The study of the QoE requirements for services as the ones we target in this paper has a long list of fresh and recent references. A good survey on the QoE-relevant performance of cellular networks when accessing many different web and cloud services is presented in [12]. Among them, YouTube deserves particular attention, due to its overwhelming popularity. Previous papers [3, 13–15] have shown that stalling (i.e., stops of the video playback) and initial delays on the video playback are the most relevant KPIs for QoE in standard, non-adaptive HTTP video streaming. In the case of adaptive streaming (DASH), a new KPI becomes relevant in terms of QoE: quality switches. Authors in [4] have shown that quality switches may have an important impact on QoE, as they increase or decrease the video quality during the playback. However, in [11] we recently found that QoE for YouTube in modern smartphones is actually slightly impaired by resolution switches, as the size of the screens is rather small and users are much used to watching YouTube in such devices. A comprehensive survey of the QoE of adaptive streaming can be found in [16].

The study of QoE in Facebook has received less attention in the past [3], but some newer studies are available, specially for the case of Facebook's QoE in smartphones [1, 7]. WhatsApp is a new service and its study has been so far quite limited. In [18] we have recently addressed the characterization of its traffic, including a QoE outlook.

When it comes to assessing the performance of operational networks, there is a growing number of papers pushing QoE concepts and methodologies within the analysis. Video streaming services are by far the mostly analyzed [6, 8–10, 20]. In [8], authors study the problem of network buffers dimensioning for optimal QoE in UDP video streaming. In [20] we introduced the first on-line, large-scale monitoring system for assessing the QoE of YouTube in cellular networks using passive, in-network measurements only. Different papers [6, 9, 10] study the problem of QoE and user engagement prediction for HTTP video streaming in both fixed-line and cellular networks. Particularly in cellular networks, recent papers tackle the problem of modeling QoE for Web browsing [5] and QoE for mobile apps [19] using passive in-network measurements, radio measurements and in-device measurements, applying machine learning techniques to obtain mappings between QoS and QoE.

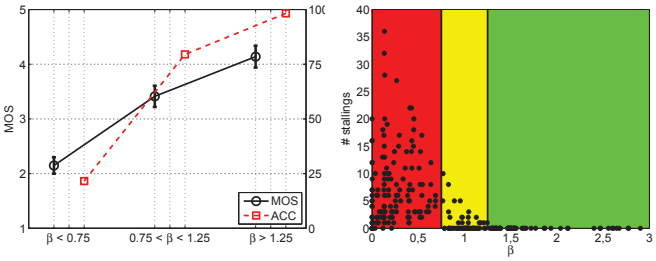
There has also been a recent surge in the development of tools for measuring QoE and network performance on mobile devices: some examples are Mobiperf[1], Mobilyzer [23], the Android version of Netalyzr [21], and our recent YoMoApp tool for YouTube QoE in smartphones [22]. In a similar direction, authors in [7] introduced QoE Doctor, a tool to measure and analyze mobile app QoE, based on active measurements at the network, application, and user-interface levels.

[1]Measuring Network Performance on Mobile Platforms, http://mobiperf.com



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

Figure 1: QoE in YouTube 360p – PC/laptops.



(a) MOS and acceptance vs. β. (b) # stallings vs. β.

Figure 2: β = ADT/VBR as a KPI reflecting user experience. Users have a much better experience when $\beta > 1.25$ – no stalling events. We refer to this + 25% over-provisioning as the β rule.

3. QOE MODELS

We have conducted a series of subjective QoE tests for the studied services in previous work [1, 3, 12], both in controlled lab settings as well as through field trials. In this section we revise some of the obtained results, analyze their main QoE characteristics and provide some updates. In particular, we focus on those studies analyzing the dependence of QoE on downlink throughput, to further apply the obtained results to the large-scale network measurements. QoE is evaluated along two dimensions: overall quality and acceptability. The overall quality is rated according to a standard Mean Opinion Score (MOS) scale [12], where 1 means *bad* and 5 means *excellent*. Acceptability is a binary indicator, stating whether the user would be willing to continue using the service under the corresponding conditions or not. We split the QoE results for two different classes of end-devices: (i) PC/laptops (YouTube only) and (ii) smartphones (YouTube, Facebook and WhatsApp). In Sec. 4 we apply the QoE mappings coming from (i) to measurements from the fixed-line network, whereas we use those mappings coming from (ii) on the cellular network measurements.

3.1 QoE in YouTube

In the case of YouTube QoE for PC/laptops, we resort to the results in [3,12], which were obtained through subjective field trial testing. Field trial testing places the end-user as close as possible to his daily usage context (location, own device, preferred content, etc), providing highly representative results. In these specific tests, 33 participants watched their preferred YouTube videos in their own laptops at their premises for a time span of about two weeks, and rated the undergone experience. Downlink traffic was passively modified through traffic shaping, done at the core of the network – participants were provided with specific Internet access connections for the study. Tests were performed in 2012, using the default video resolution set by the YouTube player, which corresponded to 360p resolution by the time of the testing.

Fig. 1 reports the (a) overall quality and (b) acceptance rate as a function of the downlink bandwidth (DBW) configured in the downlink traffic shaping. A DBW of about 750 kbps is sufficient to achieve a 90% share of positive acceptance with good QoE, whereas QoE degrades rapidly for a DBW below 0.5 Mbps. QoE saturation starts at 1 Mbps, as the QoE gain is marginal even when quadrupling the DBW.

In [3, 20] we introduced and evaluated an intuitive and very practical traffic flow-based monitoring KPI reflecting the QoE of non-adaptive HTTP video streaming, the ratio $\beta = ADT/VBR$, where ADT corresponds to the average flow downlink throughput and VBR to the video bitrate. Fig. 2(a) depicts the relationship between QoE and β for the aforementioned field trial testing. Users have a much better experience when $\beta > 1.25$, which corresponds to videos without stalling. Using measurements and results from [20] – in a nutshell, we measure the number of stallings in YouTube videos streamed to a DBW-controlled host, Fig. 2(b) actually shows that video stalling does not occur when the DBW is about 25% higher than the VBR. On the contrary, the number of stallings tends to be very high when $\beta < 0.75$.

Both the DBW-QoE thresholds coming from Fig. 1 as well as the β KPI are applied to the fixed-line network traffic traces, collected back in 2013. In Sec. 4 we verify that the monitored videos correspond mostly to YouTube 360p contents. In addition, the adoption of HTTPS by YouTube in 2013 was still limited, thus we could directly observe the exact VBR values – needed to compute β, by DPI techniques.

3.2 QoE in YouTube Mobile

In [1] we performed a series of subjective QoE lab tests in modern, 5" smartphones, for YouTube, Facebook and WhatsApp – among other services. A total of 52 participants accessed these services in smartphones connected to a fully controlled access network, where different DBW values were set. In YouTube we tested both DASH and non-DASH contents, the latter considering the highest available resolution for standard 5" smartphones, i.e., HD 720p contents. For current study, and considering that the cellular network measurements were collected in 2014, we extended the obtained results through additional subjective QoE tests to also cover SD contents in smartphones, including 360p and 480p resolutions. To get a clear idea of the typical VBR values of different YouTube contents and different encodings – including DASH and non-DASH streaming, Tab. 1 summarizes the average VBR values for the most popular YouTube videos – w.r.t. number of views, as declared at the YouTube video gallery website. For different video resolutions, the table reports the targeted device type (according to screen size), and the average VBR values available at YouTube for different codecs, including DASH ACV/H.264 and DASH VP9, as well as non-DASH codes. The last column of the table reports the β-based (i.e., 25% over-provisioning), ideally minimum ADT requirements to avoid video stalling, taking a conservative approach in which contents are assumed to be non-DASH – indeed, note that non-DASH codecs result in the highest VBR values.

Table 1: Average video bitrates for YouTube popular contents – different codecs/streaming strategies.

Quality	Device Type	DASH AVC	DASH VP9	non-DASH	β–approach (+25%)
240p	Smartphone < 4.5"	250 kbps	130 kbps	275 kbps	**340 kbps**
360p	Smartphone < 4.5"	380 kbps	250 kbps	570 kbps	**710 kbps**
480p	Smartphone, Tablet	700 kbps	800 kbps	850 kbps	**1060 kbps**
720p	Smartphone 5", Tablet, Laptop/PC	1400 kbps	1000 kbps	2000 kbps	**2500 kbps**

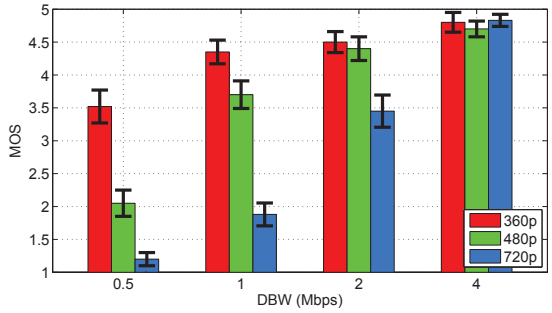
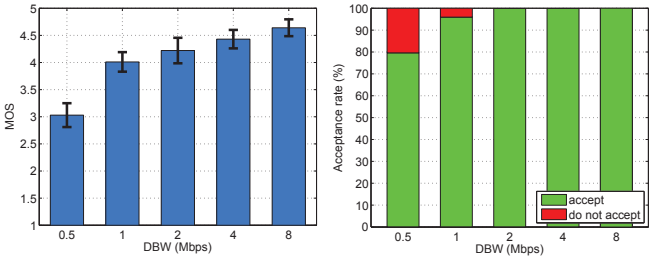


Figure 3: QoE in YouTube mobile – smartphones, for different video resolutions.

Fig. 3 reports the overall quality MOS results for YouTube mobile considering three different video resolutions and non-DASH coding: 360p, 480p and 720p. Good QoE is attained for a DBW of 0.5, 1 and 2 Mbps for the three video resolutions respectively, and quality saturation is clearly observed for 360p videos after 1 Mbps. As expected from the average VBR values reported in Tab. 1, optimal QoE is obtained for DBW above 1, 2 and 4 Mbps respectively. The DBW threshold of 0.5 Mbps is not low enough to identify bad quality in 360p videos, but both 480p and 720p contents are heavily impaired at this DBW value. We stress again the fact that these QoE thresholds are conservative, as we are considering the highest VBR values – non-DASH content (cf. Tab. 1).

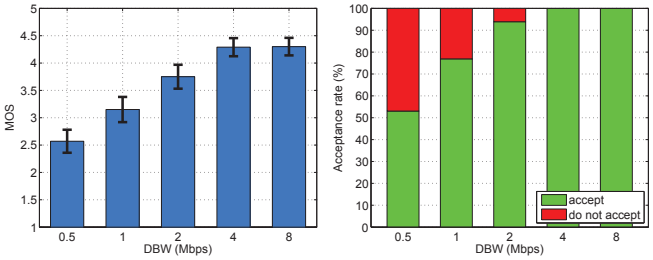
3.3 QoE in Facebook Mobile

Testing Facebook from a QoE perspective is challenging, as the application consists of multiple sub-applications and contents, which in most cases generate very different traffic patterns. For this reason, and based on our original experiences [3], we evaluate specific Facebook sub-applications which are either used by most users, or that cause a higher load on the network. Thus, participants were instructed to access the application with a specific user account, browse the timeline of this user – composed of pictures and assorted multimedia contents, and browse through specific photo albums created for this user. Such an approach tries to capture an average usage of Facebook besides simple message posting. Fig. 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 0.5 Mbps 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, and QoE saturation is already observed for higher DBW values. Full acceptability is attained for a 2 Mbps DBW allocation.



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

Figure 4: QoE in Facebook Mobile.

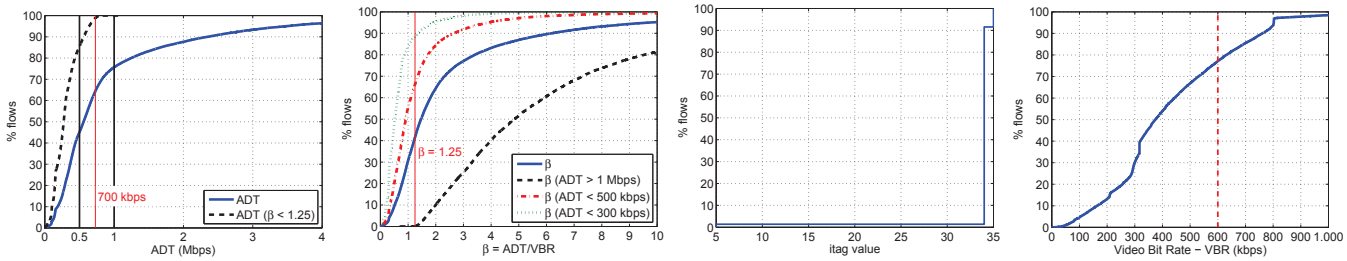


(a) Overall quality (MOS). (b) Acceptance rate.

Figure 5: QoE in WhatsApp.

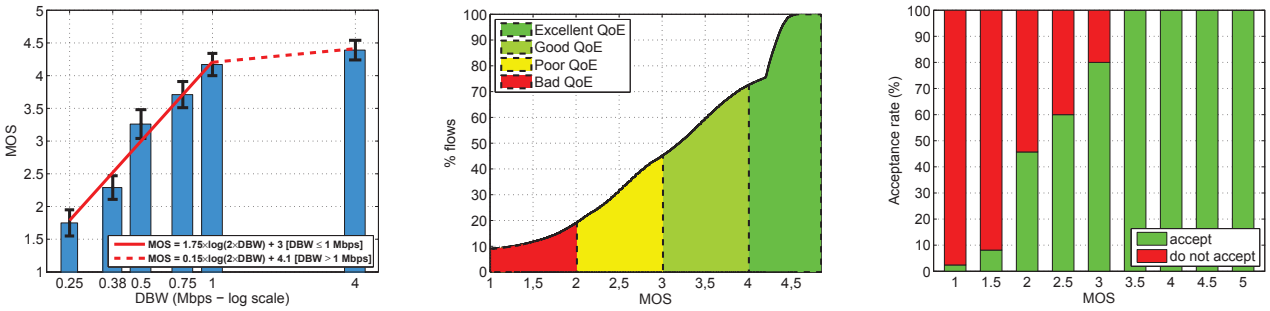
3.4 QoE in WhatsApp

We evaluate the most bandwidth-demanding type of traffic for WhatsApp, which corresponds to multi-media data sharing – chat generates negligible traffic, and the WhatsApp calling service was still unavailable in 2014, when the large-scale cellular traffic measurements were collected. Participants worked in couples and exchanged specific video files of fixed size (i.e., 5 MB), and the participant downloading the video file was the one providing a QoE evaluation, based on the experienced waiting time. Fig. 5 shows the QoE results for different DBW values. Users tolerate WhatsApp downloads with a good overall experience and high acceptability as long as the DBW is above 2 Mbps, but experience heavily degrades for slower connections, resulting in bad quality for a DBW of 0.5 Mbps. A DBW of 1 Mbps defines the QoE limit to fair quality. Given the file size used in the tests, there is a clear saturation effect after 4 Mbps, as QoE does not increase for higher DBW values. Note however that these results are partially biased by both the specific file size used in the tests and the participants task briefing – an end customer might tolerate longer waiting times if the multi-media download does not represent a hot content or an interactive exchange. Still, these results and thresholds are highly similar to those we have obtained in [12] for the specific case of cloud file sharing, suggesting a correct trend.



(a) ADT distribution. (b) β distribution. (c) YouTube video types – itag. (d) VBR values.

Figure 6: YouTube QoE in a fixed-line network, including both ADT and β = ADT/VBR as KPIs.



(a) DBW-QoE mapping function. (b) Large-scale QoE. (c) ACC vs MOS.

Figure 7: Direct application of DBW-QoE mapping functions to the fixed-line network measurements.

4. QOE IN OPERATIONAL NETWORKS

In this section we use the previously presented QoE results to asses the QoE-relevant performance of operational fixed-line and cellular networks, relying on large-scale flow throughput measurements collected in 2013 and 2014. Next we describe the collected measurements and perform a QoE-based evaluation of both networks.

4.1 Data Description

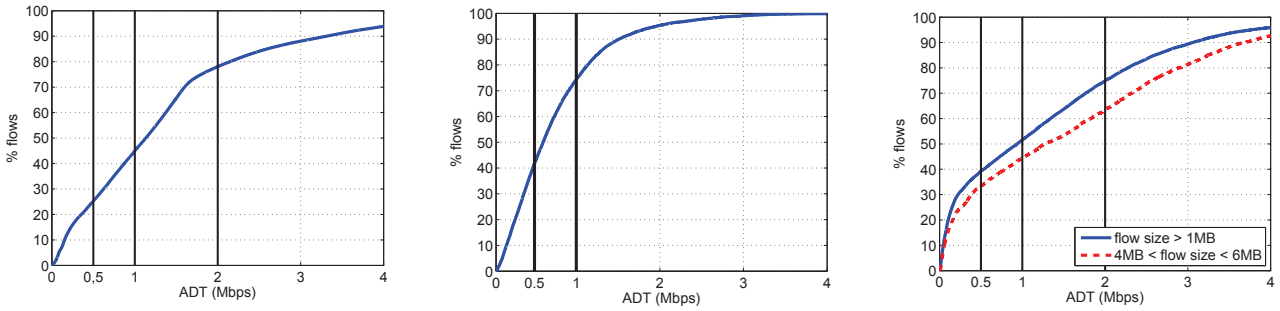
The evaluation of the fixed-line network is performed on top of YouTube flows collected by mid-2013 at a link of a European fixed-line ISP aggregating 20,000 residential customers who access the Internet through ADSL connections. The dataset spans a full week and consists of several millions of YouTube video flows. For each YouTube flow, the dataset includes the achieved ADT along with additional meta-data describing the video content – in particular, the average VBR and the specific video format, through its itag code. The itag is an undocumented code used internally by YouTube to identify video formats.

The analysis of the cellular network is performed on top of flow measurements collected at the core of a European national-wide cellular ISP during one week in early 2014. Flows are collected directly at the well-known Gn interface. The complete dataset consists of several tens of millions of YouTube, Facebook and WhatsApp flows. Only the ADT value is reported for each flow in this network. In both cases, user related data are fully anonymized.

4.2 QoE in a Fixed-line Network

Fig. 6(a) depicts the distribution of the YouTube flow ADT values. Given that the fixed-line dataset includes also the VBR value for each video flow, Fig. 6(b) additionally reports the distribution of the β = ADT/VBR metric. Before analyzing both distributions, Fig. 6(c) and Fig. 6(d) characterize the specific YouTube video contents watched in this network by 2013. Fig. 6(c) confirms that more than 90% of the video flows have itag = 34, which corresponds to 360p, FLV videos; in addition, Fig. 6(d) shows that almost 80% of the collected flows have an average VBR below 600 kbps. We can therefore apply directly the QoE results presented in Sec. 3.1, which were obtained specifically for 360p YouTube videos.

According to Fig. 6(a), about 55% of the flows achieve an ADT above 0.5 Mbps, resulting in good QoE. However, as much as 33% of the flows show an ADT below 400 kbps, potentially resulting in poor QoE. Bad quality is most surely occurring for about 18% of the flows, which achieve an ADT below 250 kbps. Finally, only 35% of the flows achieve an ADT above 700 kbps, which would result in optimal quality according to the 25%+ β over-provisioning rule. Note that 700 kbps corresponds exactly to the DBW settings recommended by large video providers for 360p videos [17]. The picture completes with the ADT values obtained by flows with $\beta < 1.25$ (dotted curve in Fig. 6(a)), which in all cases is strictly below 700 kbps, further confirming the validity of the β over-provisioning rule. When further analyzing the QoE from the β metric perspective, Fig. 6(b) shows that 60% of the flows have a $\beta > 1.25$, resulting in optimal quality settings. The difference with the predicted 35% of flows with ADT > 700 kbps from Fig. 6(a) comes from the variability in the VBR values. Indeed, β is a better QoE indicator than ADT, as it considers the particular VBR value of each flow. Still, about 32% of the flows have $\beta < 1$, resulting in potentially poor quality, and about 18% have $\beta < 0.75$, resulting in bad quality. These results are the same as those



(a) ADT in YouTube. (b) ADT in Facebook. (c) ADT in WhatsApp.

Figure 8: Average flow Downlink Throughput in a cellular network for YouTube, Facebook and WhatsApp.

predicted from Fig. 6(a) using the ADT, suggesting that poor and bad quality flows can be properly analyzed from an ADT perspective.

A final exercise we perform is that of directly translating the ADT values to quality MOS scores, by extracting a simple model from the QoE subjective results. Fig. 7(a) presents a basic curve-fitting approach to map the results presented in Fig. 1 to MOS scores. As usual in QoE modeling [12, 13, 15], we employ log-fitting curves to map DBW to MOS: the resulting model takes the form $\text{MOS} = a \times \log(b \times \text{DBW}) + c$, with $\{a, b, c\} = \{1.75, 2, 3\}$ for DBW ≤ 1 Mbps, and $\{a, b, c\} = \{0.15, 2, 4.1\}$ for DBW > 1 Mbps. Fig. 7(b) depicts the distribution of the mapped ADT to MOS results. Results are much more graphically appealing from a QoE-analysis perspective, as they directly show the predicted QoE values in terms of MOS scores. Fig. 7(c) complements this QoE picture, showing the specific acceptance ratios for each of the QoE levels, using the field trial results from Sec. 3.1. As before, we can see that about 55% of the flows have a MOS score > 3, resulting in good QoE and high acceptance rate – below 80% in the worst case. About 33% of the flows have a MOS < 2.5 (i.e., poor quality), and about 18% of the flows have MOS < 2, resulting in bad quality.

4.3 QoE in a Cellular Network

Fig. 8 reports the ADT values observed in the monitored cellular network for (a) YouTube mobile, (b) Facebook mobile and (c) WhatsApp. Note that ADT values are computed only for flows bigger than 1 MB in Facebook and WhatsApp, to obtain more reliable results – computing ADT for small flows is highly error-prone. Let is begin by the QoE of YouTube. Contrary to previous fixed-line analysis, in the YouTube mobile dataset we do not have access to the VBR values, thus we can not resort to a β-based analysis. In addition, the evaluation from an ADT perspective becomes more challenging in this case, as we do not know the specific video resolutions of the monitored flows. However, given that measurements were collected in early 2014, we expect that the largest share of videos watched in this network would correspond to 360p and 480p resolutions, using smartphones – HD content support for YouTube mobile became massively available in late 2014. Under such an assumption, and considering the QoE results of Fig. 3, Fig. 8(a) shows that 55% of the flows have an ADT > 1 Mbps, resulting in good quality for both 360p and 480p resolutions. About 30% of the flows have an ADT below 700 kbps, which would potentially result in sub-optimal quality, specially for 480p videos. Finally, about 15% of the flows have an ADT < 250 kbps, which would most probably result in bad QoE, even for 360p videos watched in small-screen smartphones, according to the expected VBR values in Tab. 1.

In the case of Facebook, Fig. 8(b) shows that about 40% of the flows achieve an ADT < 0.5 Mbps, resulting in sub-optimal QoE, whereas 25% of the flows achieve an ADT $>$ 1 Mbps, which corresponds to potentially excellent quality and full acceptability. The DBW-QoE mappings provided in Fig. 4 do not offer high visibility in the bad QoE region, which is located for some DBW value around 250 kbps – based on simple log-extrapolation. Still, we can estimate that about 20% of the flows result in bad QoE, with an ADT < 250 kbps.

Finally, when it comes to WhatsApp, Fig. 8(c) shows the distribution of ADT values for flows bigger than 1 MB, as well as for flows with size between 4 MB and 6 MB. Recall that the QoE results in Fig. 5 correspond to 5 MB files, therefore this discrimination. About 60% of the flows in the size range $[4, 6]$ MB achieve an ADT > 1 Mbps, resulting in good quality and high acceptability. Download waiting times become slightly annoying for about 10% of the flows – 0.5 Mbps $\leq$ ADT ≤ 1 Mbps, whereas bad QoE potentially occurs for about 20% to 25% of the flows, which achieve an ADT < 250 kbps.

5. CONCLUDING REMARKS

QoE is becoming increasingly relevant for ISPs, and there is a growing number of research studies focusing on the analysis of operational networks from a QoE perspective. In this paper we have proposed a simple yet powerful approach to shed light on the QoE undergone by customers of both fixed-line and cellular networks, using standard and readily available throughput measurements collected in operational networks. Quite surprisingly, our results confirm that sub-optimal and bad QoE occurrences are far from negligible in both networks for highly popular end-customer services, with about 30% of QoE-impaired traffic flows. This is highly relevant for ISPs, which might not have a clear overview on their performance when it comes to the experience of their customers.

The presented assessment methodology is technically sound and relies on real QoE subjective measurements, which provide a solid ground basis for interpretation of end user ex-

perience. Still, as we claimed throughput the paper, there are multiple limitations on our study, coming both from the QoE modeling perspective as well as from the large-scale in-network measurements. Firstly, the QoE results used as input depend on the specific characteristics of the analyzed contents, which are not easy to get from in-network measurements, as explained in Sec. 1. Whereas we do a per-content discrimination for the YouTube analysis at the fixed-line network, our predictions are potentially less accurate for the cellular network measurements, where contents are harder to discriminate from the available data. Still, recall that we have considered worst-case QoE predictions in Sec. 4.3, assuming non-DASH contents in YouTube, and higher than average volume flows for Facebook and WhatsApp. Secondly, the QoE mappings presented in Sec. 3 consider the relationship between MOS scores and the DBW values set at the traffic shapers, and not the particularly measured flow ADT values. Hence, predictions based on such mappings offer an upper bound to QoE, as in general, ADT values would be lower than the DBW ones. In any case, given that we deal with services and/or specific tasks generating high volume flows, we expect that QoE underestimations would be limited. Finally, MOS predictions done by modeling the QoE in Fig. 7 correspond to average QoE values, without considering the confidence intervals observed in the lab tests. This also applies to the evaluations done for the other services, where we have applied QoE and ADT thresholds based on the reported average MOS scores.

To conclude, we stress once more that accurate QoE estimation requires measurements and KPIs collected at multiple levels of the communications stack, including network, transport, application and end-user layers. Still, an educated guess on QoE can be done based on simple throughput measurements, as we have shown in this paper.

6. REFERENCES

[1] P. Casas et al., "Next to You: Monitoring Quality of Experience in Cellular Networks from the End-devices", in *IEEE Trans. Netw. Serv. Manag.*, to appear, 2016.

[2] P. Casas et al., "IP Mining: Extracting Knowledge from the Dynamics of the Internet Addressing Space", in *ITC*, 2013.

[3] P. Casas et al., "YouTube & Facebook Quality of Experience in Mobile Broadband Networks", in *IEEE Globecom Workshops*, 2012.

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

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

[6] A. Balachandran et al., "Developing a Predictive Model of Quality of Experience for Internet Video", in *ACM SIGCOMM*, 2013.

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

[8] O. Hohlfeld et al., "A QoE Perspective on Sizing Network Buffers", in *ACM IMC*, 2014.

[9] F. Dobrian et al., "Understanding the Impact of Video Quality on User Engagement", in *ACM SIGCOMM*, 2011.

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

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

[12] P. Casas et al., "Quality of Experience in Cloud Services: Survey and Measurements", in *Computer Networks*, vol. 68, pp. 149-165, 2014.

[13] T. Hoßfeld et al., "Quantification of YouTube QoE via Crowdsourcing", in *IEEE ISM*, 2011.

[14] R. K. P. Mok et al., "Inferring the QoE of HTTP Video Streaming from User-Viewing Activities", in *ACM SIGCOMM Workshops*, 2011.

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

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

[17] J. Jiang et al., "Shedding Light on the Structure of Internet Video Quality Problems in the Wild", in *ACM CoNEXT*, 2013.

[18] P. Fiadino et al., "Vivisecting WhatsApp in Cellular Networks: Servers, Flows, and Quality of Experience", in *TMA*, 2015.

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

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

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

[22] F. Wamser et al., "Understanding YouTube QoE in Cellular Networks with YoMoApp – a QoE Monitoring Tool for YouTube Mobile", in *ACM MOBICOM*, 2015.

[23] A. Nikravesh et al., "Mobilyzer: An Open Platform for Controllable Mobile Network Measurements", in *ACM MOBISYS*, 2015.