Are You Still Watching? Streaming Video Quality and Engagement Assessment in the Crowd

Werner Robitza*†, Alexander M. Dethof†, Steve Göring*, Alexander Raake*, André Beyer‡, and Tim Polzehl‡
*Audiovisual Technology Group, TU Ilmenau, Germany – Email: {firstname.lastname}@tu-ilmenau.de
†AVEQ GmbH, Vienna, Austria – Email: {firstname.lastname}@aveq.info
‡Crowdee GmbH, Berlin, Germany – Email: {firstname}@crowdee.de

***Abstract*—As video streaming accounts for the majority of Internet traffic, monitoring its quality is of importance to both Over the Top (OTT) providers as well as Internet Service Providers (ISPs). While OTTs have access to their own analytics data with detailed information, ISPs often have to rely on automated network probes for estimating streaming quality, and likewise, academic researchers have no information on actual customer behavior. In this paper, we present first results from a large-scale crowdsourcing study in which three major video streaming OTTs were compared across five major national ISPs in Germany. We not only look at streaming performance in terms of loading times and stalling, but also customer behavior (e.g., user engagement) and Quality of Experience based on the ITU-T P.1203 QoE model. We used a browser extension to evaluate the streaming quality and to passively collect anonymous OTT usage information based on explicit user consent. Our data comprises over 400,000 video playbacks from more than 2,000 users, collected throughout the entire year of 2019. The results show differences in how customers use the video services, how the content is watched, how the network influences video streaming QoE, and how user engagement varies by service. Hence, the crowdsourcing paradigm is a viable approach for third parties to obtain streaming QoE insights from OTTs.**

***Index Terms*—video quality, video streaming, Quality of Experience, video streaming, adaptive streaming, user behavior, user engagement, YouTube, Netflix, Amazon Prime Video, crowdsourcing**

I. Introduction

Video streaming is one of the main drivers of current Internet traffic. As the delivered streaming quality continues to improve—also considering technological developements like 4K/UHD, HDR and high framerate content—customer demands increase simultaneously. Studies indicate that users with faster connections and better services have higher expectations and may be disappointed faster in case of service problems [1]. Considering this, video streaming over-the-top (OTT) providers generally optimize their services to increase user engagement. In these contexts, engagement is often a proxy for Quality of Experience (QoE)—or vice versa. Bad QoE results from streaming interruptions or low visual quality. Consequently, all providers along the service chain are incentivized to monitor and manage network quality to keep customer experience high.

While OTTs have access to all parameters relevant for a valid and in-depth estimation of QoE at the customer side, Internet Service Providers (ISPs) and academia usually do not. ISPs can generally only estimate OTT streaming quality based on simple bandwidth-related quality models. They can also set up automated monitoring probes that regularly measure Key Performance Indicators (KPIs) informing about Quality of Service (QoS). However, these methods can only provide individual samples and do not necessarily reflect what is happening at the customer side. Similarly, academic researchers can often only resort to setting up laboratory measurements for estimating OTT QoE.

Hence, crowdsourcing is a viable alternative to automated probing systems, with the benefit of providing large-scale longitudinal measures of real customer experience. In this paper, we present the results from a customer-centric, real-life crowdsourcing study in which we monitored and evaluated the QoE of YouTube, Netflix and Amazon Prime Video streaming on the desktop in Germany throughout the entire year 2019. After presenting related work in Section II, we describe our methodology in Section III. Our results are presented in Section IV and discussed in Section V. They show that crowdsourcing can provide ISPs or regulators with results that are representative of what customers really experience, similarly to QoE analytics data that otherwise only OTTs have. Our paper is concluded in Section VI.

II. Related Work

A number of large-scale studies focusing on user engagement in video streaming have been published, e.g. [1]–[3], making use of third-party analytics platforms to correlate video properties and network performance with engagement metrics such as viewing time. The authors could show strong relationships between stalling and video abort rates [1], or overall video quality and video view duration [2]. The cited studies rely on proprietary datasets of millions of video views that can only be created by video OTT or CDN providers, but are not available to academia (or the public). Also, the datasets do not cover the most popular streaming platforms like YouTube—or the services are kept secret.

From an academic perspective, crowdsourcing studies on OTT streaming quality have been performed with dedicated testing tools such as *YouSlow* [4] (a browser extension) or *YoMoApp* [5] (a mobile app). The benefit of browser extensions is the fact that real video sessions can be measured, similar to proprietary analytics data. Due to platform constraints, in the mobile crowdsourcing case, however, users need to actively watch videos in a research app instead of, e.g., the official YouTube app, which consequently leads to smaller datasets and therefore lower statistical inference power.

In the laboratory, dedicated tests to infer user engagement from video streaming issues have been reported in [6]–[8]. A subjective test methodology for assessing the impact of initial loading delay on QoE and user behavior was published in ITU-T Rec. P.917. These methods can help to understand the underlying reasons for certain user actions, but are prone to experimental biases if subjects are in a laboratory situation and aware that they are monitored.

In general, there is also no consensus on the definition of the term *user engagement*: from a macro perspective, it can be interpreted as the collective customer behavior over time (e.g., churn rates, usage times); from a more fine-grained perspective, it could be the individual actions of a user when watching a single video.

To summarize, previous approaches have covered video streaming KPIs and user engagement, but to the extent of our literature survey, so far, no large-scale crowdsourcing campaign comparing *multiple* popular services has been published. The use of a large dataset in our case makes it possible to relate user engagement and video quality to different underlying factors and characterize user habits and streaming performance for several major ISPs in Germany, both from a macro and fine-grained perspective.

III. Methodology

A. Measurement Software

We created a web browser extension called *YTCrowdMon* (for Google Chrome and Mozilla Firefox) that allows end users to measure their video streaming performance on YouTube, Amazon Prime, and Netflix—and in turn, it allows us to measure user behavior and video QoE.

The software inspects both technical events and user events. These events occur at different layers: The *network layer* is accessible through web browser APIs, where data about HTTP requests can be recorded, including URL parameters, size and timing. The *player layer* and its events (e.g., stalling) can be inspected via JavaScript. For YouTube, a JavaScript API is available. For Netflix and Amazon Prime, extracted the same QoE-relevant events and KPIs from the services' proprietary APIs. We validated the accuracy of this method based on comparison with screen recordings. Finally, we gathered user events, that is, interaction with tabs (e.g., closing, navigating away), or with the player (e.g., pausing or seeking).

B. User Acquisition / Crowdsourcing

The *YTCrowdMon* extension was distributed to users across Germany via the crowdsourcing platform *Crowdee*. Users were offered a small amount of money to install the extension and perform active measurements with it. After receiving a description of the task and accepting the privacy policies (which clearly explained to the users which data was collected for which purposes), the extension could be installed. In terms of personal data, IP addresses and geolocations were gathered, with informed consent. No names, email addresses, or socio-economic information were collected. Users could keep the add-on installed for as long as they wished and temporarily deactivate it at any time.

To maintain a steady number of measurements, we repeatedly incentivized users throughout the year to perform speed tests and video quality measurements with the extension (one every week of participation). For this purpose, a dedicated speed test and active video quality test mode was developed. Speed tests were performed on speedtest.net. The results from these active tests are not part of this publication; however, we used the users' speed test results to classify their bandwidth later.

C. Data Processing

The data from the extension was collected centrally and processed in daily batches. Data processing steps included the following items:

1) ISP determination: Based on the IP addresses, we resolved the ISP and Autonomous System ID using the `ipapi.co` service. Mappings were developed to group aliases of the same ISP together. We associated each user with his/her home ISP based on where the majority of video views came from (e.g., if one user used ISP *X* in 75% of all video sessions, this was considered his home ISP). For privacy reasons, IP addresses were anonymized immediately after resolving the ISP data.

2) Bandwidth estimation: By inspecting the maximum achieved speed test result for each user/ISP combination, we classified users into bandwidth groups for later analysis. These bandwidth groups are $[0, 16]$ Mbit/s, $(16, 50]$ Mbit/s, and $(50, 500]$ Mbit/s, which are typical for Internet products in Germany.

3) Statistics calculation: Using the raw events gathered from the layers described in Section III-A, and additional video/audio quality-related metadata about the watched videos, statistics were calculated. These statistics include the video watch duration, initial loading delay and stalling (count, duration), video quality level or resolution changes, and the Mean Opinion Score (MOS) according to ITU-T P.1203.

4) Filtering: Data were filtered to prune invalid measurements and unreliable users, and thus create a more homogenous picture for later analyses, as usually done for crowdsourcing studies. We filtered out video sessions from ISPs other than the user's home ISP, measurements from outsideGermany, and video sessions with extremely long loading times of over one minute. About 25% of playbacks were filtered out according to the above criteria.

D. QoE Calculation

In order to estimate the QoE of each video streaming session, we used the HAS QoE model ITU-T Rec. P.1203 [9], based on the implementation in [10]. The model was chosen for it having been extensively trained and validated on 29 subjective databases. For video, P.1203.1 Mode 0 was used, which requires information about bitrates, resolution, framerate, and codec. For H.265 and VP9, which the P.1203 standard does not support, we used an extended open-source model.[1] Audio quality was estimated based on codec and bitrate. The overall P.1203 model works by first calculating video and audio quality per second, then integrating these data over time, taking into account initial loading delay and stalling events, up to a maximum of 5 min. P.1203 outputs an overall audiovisual quality on the MOS scale (*O.46*) as well as other diagnostic data (e.g., *O.23*, the stalling quality). For every video playback's first 5 min, we calculated the P.1203 score if at least 10 s of video were played.

The benefits of this model are three-fold: first, it can quantify the effects of stalling and quality variations over time (temporal pooling), which is not possible with pure pixel-based quality metrics/models like PSNR, SSIM or VMAF. Hence, we get a complete picture of an entire session's QoE. Second, in its Mode 0, P.1203.1 only requires metadata and consequently can be calculated easily, whereas pixel-based models are computationally much more complex and cannot realistically be run in a crowdsourced scenario on users' machines. Further, pixel-based models need access to the decoded video, which is not practically feasible in web browsers.

IV. Results

A. Overall Statistics

Overall, we collected 447,489 video playbacks in the year 2019, from the five major ISPs studied in this work. The data, which comprises more than 33,000 hours of streamed video, stems from 2,002 unique installations of the *YTCrowdMon* extension, 35% of which using Firefox, the remainder Chrome. Most playbacks (93.6%) stem from YouTube, which is mostly due to the service's popularity and availability, and the overall shorter length of videos. 3.61% and 2.51% of playbacks are from Amazon and Netflix, respectively. Most playbacks happen in the speed group $(16, 50]$ Mbit/s, acccording to the larger number of users in that group.

In terms of network performance, the distribution of achieved download speeds are seen in Figure 1. Here we can easily notice the heterogeneous market situation by distinguishing the different access technologies: ISPs *A*, *C*, and *E* primarily are DSL-based, with peaks at roughly 50 Mbit/s corresponding to VDSL50 technology. ISPs *B* and *D* on the other hand sell both DSL and cable-based access, with ISP *B* being able to serve the highest throughputs in the field, and ISP *D* having a few high-speed users, but most of the clients at speeds $\leq$ 8 Mbit/s. We will later investigate which impact these speeds have on video streaming performance/QoE.

B. User Behavior and Engagement

In the following, we will look at the behavior and engagement of our users, starting with a high-level perspective of the user pool and its charateristics. Later, we will interpret it on a per-session basis.

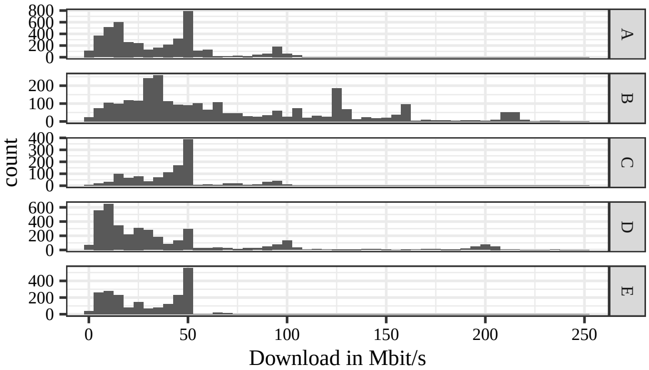

Fig. 1. Speed test result distribution per ISP.

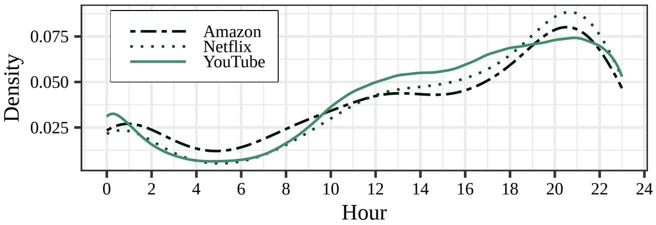

Fig. 2. Hours of video playback.

The majority of 64% of our users are exclusive users of YouTube. Only about 8% have used all three video services at least once, which includes the paid plans for Netflix and Amazon. 10% and 14%, respectively, used Amazon and YouTube, or Netflix and YouTube. This can likely be explained by the "premium" cost of Amazon and Netflix compared to free YouTube access, and the fact that Netflix/Amazon may be watched on Smart TVs or tablets more often than in the browser. Users are avid streamers though: an average user watches two complete YouTube videos per day.

When are people typically watching video? Figure 2 shows the hours of the day at which video playbacks were started. Notably, YouTube is watched slightly more often during the day, but less so during traditional "prime time" hours (20:00), where Netflix and Amazon peak. Usage does not vary across the weekdays; each day contributes about 14% of playbacks, except for Amazon, which shows a peak on Thursdays. This is most likely linked to their Thursday "Prime Deals", where series and movies are offered at reduced prices for members.

To get an idea of what people are watching, we first show the distribution of available video durations in Figure 3. Note the similarity between Amazon and Netflix, with peaks at 25 and 45 (mostly series), and 100 minutes (movies). YouTube's

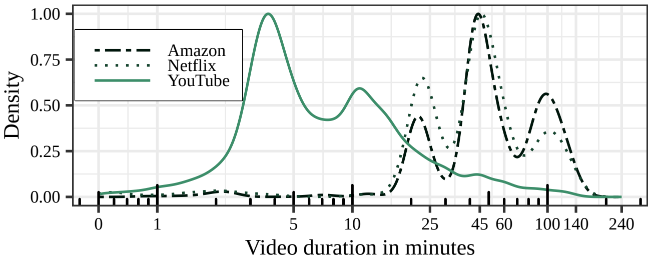

Fig. 3. Video durations per service. x-axis is logarithmic.

[1] https://github.com/Telecommunication-Telemedia-Assessment/itu-p1203-codecextension

TABLE I
CUTOFF POINTS FOR DURATION AND COMPLETION RATIO.

	Duration (min)			**Completion Ratio**	
Service	S/M	MS/ML	ML/L	Beg./Mid.	Mid/End
YouTube	3.53	5.76	12.8	0.16	0.94
Amazon	21.85	43.96	89.8	0.31	0.99
Netflix	24.17	43.38	58.87	0.19	0.82

mostly user-generated content typically only lasts slightly below 5 minutes, or around 10 minutes.

In the following, we define user engagement as the video completion ratio (*CR*), i.e., the ratio between total time spent watching the video and the overall video duration. The value of *CR* is generally between 0 and 1. Figure 4 shows the *CR* distribution for all services. For easier analysis, we classified the *CR* per service into three categories with the same number of observations (i.e., imagine the area below the curve split into three equal-sized parts): *beginning* (B), *middle* (M), and *end* (E), depending on where users exited the video. Accordingly, for example, if a YouTube video was exited before 16% of its duration was reached, it was determined as "quit in the *beginning*". Likewise, we classified video durations into *short* (S), *medium-short* (MS), *medium-long* (ML), and *long* (L). The cutoff points between those classifications are shown in Table I.

Users finish video playbacks when they close the respective tab or the browser, navigate to another website, or watch another video. In Figure 5, we give an overview of the frequency of these actions, dependent on when users quit in the video, and how long the video is. Going from left to right (completion ratio classification), it can be observed that when the user is still at the beginning, the likelihood of closing the tab (green dashed line) or choosing another video (black dashed line) is high. For short videos, users choose another video even more often. When in the middle of the video, however, users are much more likely to just close the tab instead of choosing another video, particularly for Netflix. Once users have reached the end of a video, it is most likely that they will continue with another video—this is similar for all three services. Going from top to bottom (video duration), we notice that the likelihood of continuing with another playback generally decreases with the duration of the video, but only when users have actually watched the video until the end. When users quit at the beginning, the video duration itself has only a small impact (except for short videos).

To get a first indication as to *why* users actually abort a video, we selected all YouTube videos and ran a 10-fold cross-validated Random Forest regression with *CR* as target, and a set of video metadata, KPIs, and P.1203 diagnostic scores as features. A feature importance analysis showed that the P.1203 O.23 score ("stalling quality") and duration were the most relevant features for predicting *CR*, with higher relevance than video popularity measures such as the number of views, or the ratio between likes and dislikes. A more in-depth prediction of *CR* will be part of future work, including time-series–based modeling.

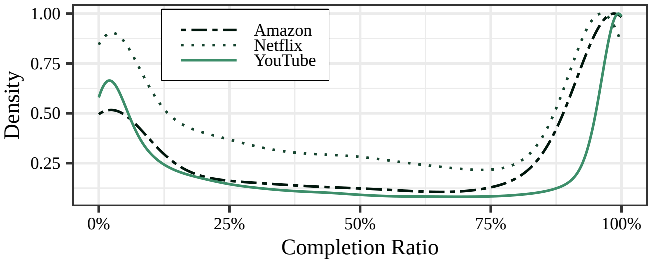

Fig. 4. Completion ratio for different services.

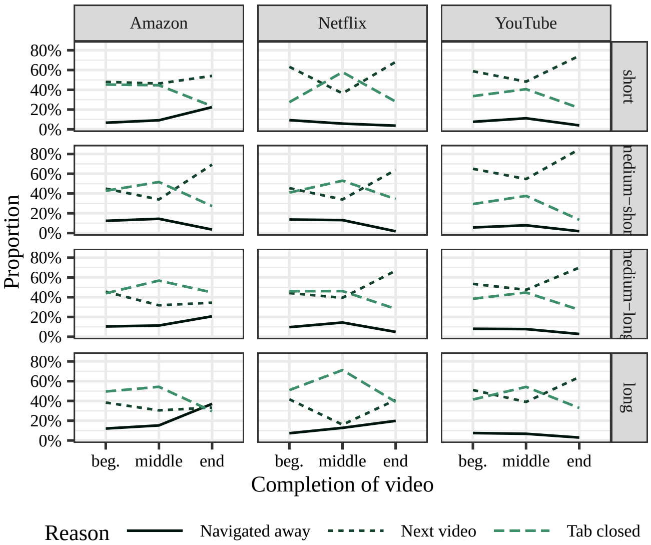

Fig. 5. Reasons for stopping playbacks.

C. Video Streaming QoE

Video streaming QoE is influenced by many factors (cf. [11]), including the audiovisual quality itself (and its variation over time) as well as discontinuities in terms of initial loading delay (ILD), the number of stalling events, or the total stalling ratio (i.e., how much of the playback time was spent buffering). Figure 6 shows the distribution of initial loading delay over all services and bandwidth groups. It can be observed that the overall median ILD is lower for YouTube in comparison to Amazon and Netflix. This difference is generally more pronounced for lower bandwidths. Also, the higher the bandwidth, the lower the median ILD. For YouTube, the median drops from 1.03 s (0–16M) to 0.67 s (16–50M) and 0.55 s (50–500M), whereas for Amazon and Netflix, the median ILDs for 0–16M are 2.58 s and 2.83 s, respectively. They drop to 1.75 s and 1.66 s at 16–50M and 1.25 s at 50–500M. As expected, while the median is low, ILD varies greatly and may reach much higher values.

Stalling events happen rarely. Most sessions have no stalling at all: for YouTube, 77% of sessions are free of stalling; for Amazon and Netflix it is 93% and 96%, respectively. When there is stalling, it is typically very short. 20% of YouTube sessions have one stalling event with a median time of 0.35 s. We believe this to be the case due to mid-roll advertisements and subsequent short stalling while loading the original video.

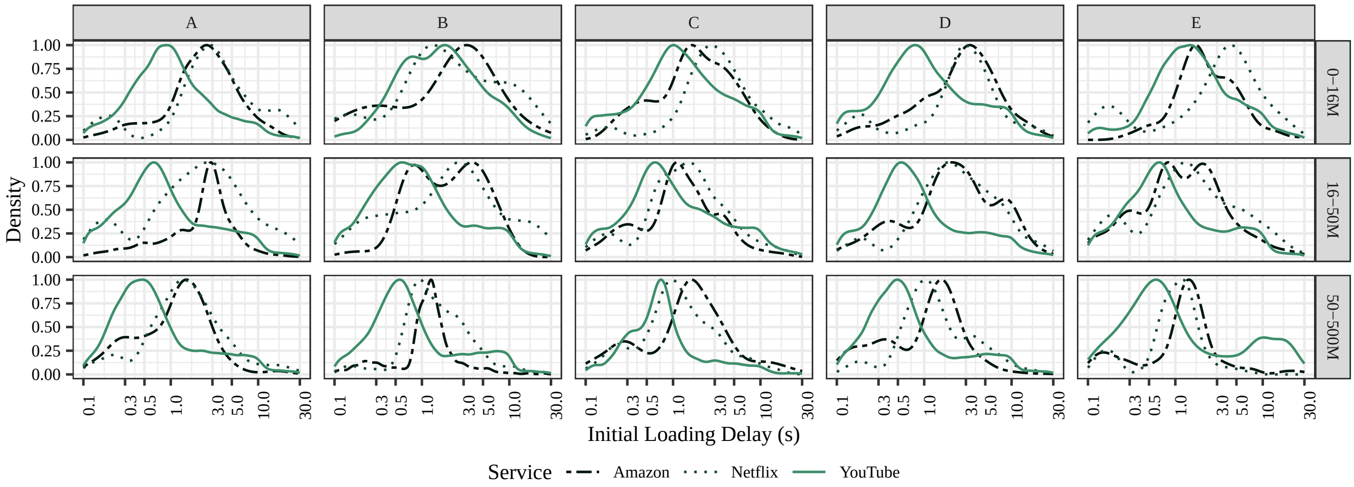

Fig. 6. Initial loading delay for different services and speeds. x-axis is scaled logarithmically.

This has to be verified in a more in-depth data analysis. Amazon videos that have one or more stallings have a median total stalling duration of 0.26 s (mean: 0.53 s). Netflix videos with at least one stalling have a median total stalling duration of 1.53 s (mean: 3.75 s). This hints at a more conservative player buffering strategy (i.e., a larger buffer that takes longer to fill), which we will analyze in future work.

The overall MOS for a session was calculated by integrating the collected KPIs using the ITU-T P.1203 QoE model. We show the MOS distribution per ISP and service in Figure 7. Here, we can observe that MOS scores for YouTube are generally higher than for Amazon and Netflix. However, given the content-agnostic calculation of P.1203.1 Mode 0, a higher MOS can also be the consequence of higher video bitrates without necessarily higher video quality. Hence, a direct inter-service comparison is not valid without further video content or encoding analysis. The small "bump" at MOS ≈ 4.25 in YouTube scores can be explained by the presence of stalling: if there is at least one stalling event, a video can never reach the maximum MOS. Since stalling is more prevalent for YouTube, the curve is shaped differently.

The quality offered by the ISPs is not the same: ISP C, for example, has lower QoE for Netflix compared to Amazon, while for ISP D, it is the opposite. The generally lower quality can be explained by the low-bandwidth access speeds, which is particularly noticeable in the QoE for ISP D (which also has the lowest speed tests in the field). The difference in performance for Amazon and Netflix may be due to different performance in Content Delivery Network peering, but a more detailed investigation of the origin of these differences is required and will be part of future work.

V. Discussion

The presented results show that the chosen crowdsourcing paradigm can be employed to analyze QoE and user engagement in a real life setting. It can therefore complement subjective lab tests, which can provide deeper insights, but result in fewer data points, and—most importantly—cannot capture real customer behavior. The use of passive QoE monitoring with real customers has its benefits over laboratory-based methods and testbeds: first, a much broader range of locations can be covered, and a significantly higher number of measurements can be performed. QoE analysis in this case requires the use of a well-trained and validated streaming QoE model, for which P.1203 may be chosen, since its application scope matches the scope of our campaign. Second, for academia or ISPs, this method allows gathering data that otherwise only OTT providers would have, namely about users' *real* streaming behavior. This holds for global data like hours of use, but also per-session data like video completion and abort behavior, which in turn are proxies for user experience. Such behavior could not be realistically tested in a lab. Particularly in the case of behavior analysis, it is known that users may act differently in a lab test, when they know that they are being watched [7]. Hence, we see a high value in the passive data.

Although in principle more cost-effective, large scale crowd studies can cause considerable costs with higher volumes. Regular incentivization for the users to perform active measurements is required in order to keep the number of passive measurements at a meaningful level, particularly for paid services. If there are no other extrinsic incentives, the biggest intrinsic motivation to use a monitoring tool may only be to participate in research, or to "do a favor" to the researchers conducting the study.

While it would be preferable to use the same approach for mobile networks, it cannot be reasonably applied here, as the external collection of passive video watching data is not possible with native mobile apps. Hence, one could only collect active video data, limiting the amount and representativeness of data.

Further, the number of sessions with stalling—and consequently noticeable quality issues—was very low. The absence of strong network issues can be attributed to the desktop-based fixed-net evaluation. This means that higher numbers of sam-

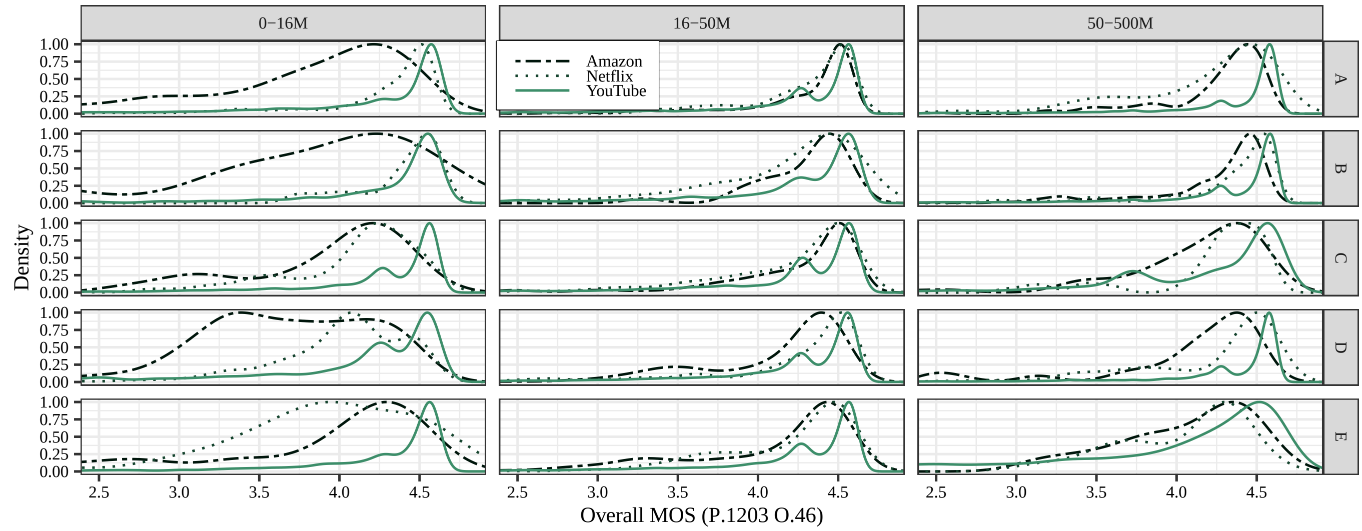

Fig. 7. P.1203 MOS per ISP and speed group.

ples are required for in-depth statistical analysis. Otherwise, a root-cause analysis—or an attempt at explaining reasons for certain user behavior—may be invalid. However, it also means that any (laboratory or crowd) tests that show significantly more stalling than what a user would get at home may yield biased ratings or reactions.

Finally, while the P.1203 model provides a good view on a per-session quality, it is limited to a few minutes of playback. As we have observed, this does not suffice for hour-long videos. Hence, further research and extensive subjective testing is needed to validate temporal pooling methods for existing methods, or extrapolate QoE from shorter measurements.

VI. Conclusio and Future Work

This paper presents a first analysis of a large-scale crowdsourcing dataset of passively measured video streaming KPIs and QoE, collected over an entire year. Covering five different national ISPs and three major streaming services, the novelty of the research consists in comparing these aspects with respect to general streaming behavior, user engagement, and overall QoE.

Our results show that on PCs, Amazon and Netflix have much lower overall usage and viewership in contrast to YouTube. This also has implications on possible crowd campaigns. However, the average user still watches at least one video per day. The Amazon and Netflix catalog offers series and movies, and thus much larger video durations. Still, users typically finish these videos once they have tuned in. We could show that the way users abort their sessions changes depending on the service, the duration of the video, and how far they are in completing the clip. Finally, a KPI/QoE analysis showed differences in initial loading times across services and user bandwidths, as well as notable QoE differences between the observed ISPs.

Future work will focus on individual session engagement and why users abort videos. Is it because they are bored, or did something happen that annoyed them? Also, we will investigate how current QoE models can be extended for the case of predicting longer session QoE (i.e., videos longer than 5 min), and how they can be used to predict user engagement.

References

[1] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 2001–2014, 2013.

[2] F. Dobrian, A. Awan, D. Joseph, and A. Ganjam, "Understanding the impact of video quality on user engagement," *Communications of the ACM*, vol. 56, no. 3, pp. 91–99, 2013.

[3] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a Predictive Model of Quality of Experience for Internet Video Categories and Subject Descriptors," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 339–350, 2013.

[4] H. Nam, K.-H. Kim, and H. Schulzrinne, "QoE Matters More Than QoS: Why People Stop Watching Cat Videos," in *IEEE International Conference on Computer Communications*, 2016.

[5] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-gia, and R. Schatz, "YoMoApp: a Tool for Analyzing QoE of YouTube HTTP Adaptive Streaming in Mobile Networks," Tech. Rep., 2014.

[6] R. K. P. Mok, E. W. W. Chan, X. Luo, and R. K. C. Chang, "Inferring the QoE of HTTP video streaming from user-viewing activities," in *Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack - W-MUST '11*, 2011, p. 31.

[7] W. Robitza and A. Raake, "(Re-)Actions Speak Louder Than Words? A Novel Test Method for Tracking User Behavior in Web Video Services," in *Eighth International Workshop on Quality of Multimedia Experience (QoMEX)*, Lisbon, 2016.

[8] P. Lebreton and K. Yamagishi, "Study on user quitting rate for adaptive bitrate video streaming," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, Sep 2019.

[9] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Goring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. Erfurt: IEEE, may 2017, pp. 1–6.

[10] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software," in *9th ACM Multimedia Systems Conference*, Amsterdam, 2018.

[11] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hossfeld, and P. Tran-gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communication Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2014.