Benchmarking Video Service Quality: Quantifying the Viewer Impact of Loss-Related Impairments

Vidhyalakshmi Karthikeyan, Brahim Allan, Detlef D. Nauck, and Miguel Rio

***Abstract*—We present the first empirical study of the impact of loss-related errors on TV viewing engagement across disparate platforms, delivery technologies and performance measures. Our dataset comprises anonymised video viewing sessions and data about quality of delivery from a content service provider with a nationwide customer base. We study buffering events on streaming apps, mild and severe packet loss errors on multicast-delivered IPTV to a Set-Top-Box (STB) and signal strength errors on Digital Terrestrial TV. Since these metrics cannot be directly compared to each other, we use engagement as our proxy measure. We first characterise the relationship between each impairment and viewing engagement, investigating confounding factors such as type of content, asset length and connection type. We conclude that the loss of engagement due to poor quality delivery is incurred immediately for on-demand content and in the long-term for live content. We rank impairments across platforms by their impact on engagement.**

***Index Terms*—Streaming, IPTV, DTT, Video quality, Engagement**

I. Introduction

Video viewing has rapidly evolved in recent years and is expected to dominate Internet traffic by 2021 [1]. Major content providers already support converged usage, enabling customers to seamlessly watch the same content item across multiple platforms, for example, on one or more set-top-boxes (STBs) and a multitude of apps on mobile devices. Customers can watch live content or catch-up on selected items for a limited time period after broadcast. Platform diversity is common – a service provider achieves end-to-end delivery by managing integration and interoperation of components from multiple vendors. Delivery technologies, error recovery mechanisms and, most importantly, the data captured about session performance varies vastly by platform. Irrespective of end-to-end heterogeneity, video quality delivered across multiple platforms to a converged customer should to be equivalent from the customer's perspective, suggesting the need for cross-platform benchmarking of end-to-end video quality. However, customer expectations and context play a key role in benchmarking – the same viewer has a different tolerance to (re)buffering experienced on an application (app) on a portable device whilst travelling on a train than on a STB connected to a family room television during prime time [2]. A viewer may also pay different levels of attention in both scenarios.

Content service providers therefore look for ways to compare disparate measures in order to prioritise improvements in customer experience across platforms. Which service degradation has more customer impact – five seconds of buffering time on an app over unicast, ten uncorrected packet loss errors on a STB over multicast or one thresholded signal strength error over Digital Terrestrial Television (DTT)?

Service providers quantify customer impact in many ways. Monetary cost can be clearly attributed to complaint calls, engineer appointments and service churn. However, they are not representative of the impact on the entire viewer base as not all customers with poor service get in touch or are surveyed and the delay between impacted experience and contact can be arbitrary. Viewing engagement, on the other hand, is unreserved customer feedback. We note that viewing disengagement can result from at least three aspects of content consumption: video quality, disinterest in the content and changes in lifestyle. Whilst subjective and objective testing models explore user-perceived video quality and promise a more causal relationship, such models are highly parameter-dependent. Service providers find themselves grappling with ever-changing, bespoke implementations and error messages that may not have been previously modelled. Therefore, there is value in studying observable behaviour of their entire nationwide customer base and the relationship to levers that service providers can control to deliver a better experience. These avenues of research are complementary. We acknowledge that quality metrics can have interdependent and counter-intuitive relationships to each other and user behaviour [3]. We explore these topics in this paper and develop ways of representing engagement and performance such that the relationship becomes clearer. We propose that viewing engagement is a common proxy measure that can be used at scale through which we can compare different metrics across different platforms that are otherwise uncomparable.

The purpose of this study is to quantify the impact that loss-related impairments have on viewing engagement across different video delivery platforms using anonymised data from a nationwide video service provider in the UK. Our findings and methodology apply to a wide range of video platforms beyond the exact implementation of this service provider. We characterise the relationship between engagement and each of the impairment types on STB and streaming apps on portable devices. We then develop a method to benchmark loss-related errors across all delivery methods to rank the scale of impact on viewers.

The rest of the paper is structured as follows. We present related work in Section II, identifying our original contribution

V. Karthikeyan, B. Allan and D. Nauck are with BT, Applied Research, Adastral Park, Ipswich, IP5 3RE, United Kingdom
M. Rio is with University College London, Department of Electrical Engineering, Torrington Place London, WC1E 7JE, United Kingdom

to the field. We describe our measurement setup and dataset in Section III and evaluate key metrics that capture different facets of viewing engagement in Section IV. We present baseline viewing and error performance in the population in Section V. Section VI characterises the impact of loss-related errors on session abandonment, highlighting confounding factors. Section VII presents our analysis methods and findings on the long-term impact of loss-related errors on STB and app viewing engagement. Finally, we describe our engagement-based method to benchmark video quality across platforms, delivery technologies and metrics measured in Section VIII. We rank different loss-related impairments that occur in a complete video delivery ecosystem by their impact on user engagement.

II. Related Work

Traditional indices of video quality scoring use subjective and objective testing. Modelling Quality of Experience (QoE) from Quality of Service (QoS) metrics is characterised by [4]–[9]. Studies [10]–[13] focus on IPTV. Studies [14]–[16] focus on QoE modelling in HTTP streaming environments and we also refer readers to [17] for a survey of this area. ITU-T frameworks on subjective and objective testing for media delivery also exist [18]–[21]. Note that perceptual quality depends on the nature of the service and its implementation. Resolution, compression ratios, video/audio formats and recovery strategies such as buffering, error correction [22] and re-transmission impact perceived quality loss. Where interactivity such as pause/seek actions are included in the service, response times also contributes to system QoE. Increasingly, however, such video quality scores are being replaced by relating delivery quality to measurable engagement metrics to better align with business objectives and also take into account individual delivery implementations. We review the most relevant literature on the impact of quality of service delivery on user engagement within the scope of our work.

An objective User Satisfaction Index that statistically correlates QoS metrics (bitrate, delay, jitter and round trip time) to Skype call quality is proposed in [23]. The index is shown to correlate well to call duration and speech quantity during the call. Note, however, that the requirements for two-way conversations differ from one-way viewing of video. In the mobile video delivery space, [24] characterise and model the relationship of 31 different mobile network parameters on session abandonment using a decision tree approach. Using Yahoo! toolbar browsing data, [25] use clustering methods on user behaviour to categorise websites visited in terms of metrics that represent popularity, activity and loyalty. An objective measurement on the impact of end-to-end application QoE such as join time, buffering ratio and frequency, average bitrate and rendering quality on per-video and per-viewer engagement is presented in [26]. It concludes that buffering ratio, a consequence of packet loss, has the highest impact on engagement. The magnitude of impact depends upon content length and type (live/VoD). Similarly, a decision tree-based method to determine application QoS metric interdependencies and their complex relationship to viewing engagement is developed in [3]. The authors conclude that type of video (live/on-demand), viewing device (mobile/PC/TV), connectivity (wired/wireless) and time of day (peak/off-peak) are inter-related confounding factors that impact viewing engagement. Understandably, customer expectations vary by context and affect overall satisfaction. Other authors model video buffering, QoE and user engagement as three cornerstones of content delivery. The authors of [27] fit a mixture of two exponentials to model video playtime and buffering ratio in a HTTP streaming application and examine the correlation between QoE and video engagement. Using data from a live tennis event broadcast to a Francophone audience over the Internet using HTTP, [28] describes and correlates the impact of a spectrum of video quality metrics on video playtime. Focused on VoD content delivered from an Akamai Content Delivery Network (CDN), the investigation in [29] concludes that fewer content chunks are viewed on lossy sessions overall with early loss being particularly detrimental to engagement. Using CDN-delivered VoD sessions, [30] propose a method to demarcate correlation and causation. The authors use a customer matching algorithm to design quasi-experiments from data that identifies causal links between start-up delay and session abandonment by connection strength, rebuffering ratio and viewing duration, and video failure to start and low likelihood of viewer to return to platform.

There is a rich area of literature on modelling and predicting Mean Opinion Score (MOS) from video quality metrics. Such modelling requires a wide variety of metrics in order to produce a robust model. Our dataset focuses on one aspect of quality per session, i.e. loss-related impairments, unique to each delivery platform. We also find that existing literature tends to focus on a single delivery mechanism in each study. The strength of our dataset lies in its variety across streaming, multicast and terrestrial broadcast technologies for a nationwide base of real users 'in the wild'. Studies also tend to focus on immediate session abandonment following impairments whereas we cover long-term effects as well. We present point comparisons to literature where possible as well as adding novel insights to user behaviour within each platform. We then make further unique contributions that, to our knowledge, are the first of their kind as stated below.

We present the first study that benchmarks the quality of TV delivered across:

- Disparate platforms with different customer context and expectation, namely app streaming and STB viewing
- Disparate delivery technologies, namely unicast over wireless and mobile networks, multicast and DTT over fixed networks
- Uncomparable quality metrics measured per platform, namely buffering events, multicast packet loss and thresholded low DTT signal strength errors

Our empirical study is based on a common customer base. We have developed a method using various facets of viewing engagement to make this cross-comparison. Our aim is to rank different loss-related impairments by their impact on video viewing.

III. Log data composition

Viewers may subscribe to either or both of the app and STB platforms, although some app content is only available to viewers who also have a STB. App content is delivered using HTTP adaptive streaming over fixed and mobile networks, typically to a portable device. On the STB, a mixed subset of High Definition (HD), Standard Definition (SD) and Ultra-High Definition (UHD) channels are delivered using non-adaptive multicast streams over Real-time Transport Protocol (RTP) with unicast recovery mechanisms that fetch missing packets, ideally before playout occurs. Other channels are delivered using DTT by radio waves. Customers can purchase the STB product only if they meet a minimum access line quality criterion and multicast traffic is prioritised through the network. Both live and video-on-demand (VoD) can be watched on both platforms. This is henceforth referred to as the content type. End user access speeds on the STB depend upon access technology and line performance. Possible technologies are ADSL, ADSL2+, FTTC and FTTP. The bandwidths for app content consumed over mobile technologies depends on the mobile service provider. The service provider in this study uses one vendor solution for collection of telemetry data from the STB and a different vendor to collect application-level performance data from the app platform.

A summary of loss-related metrics collected on each platform is shown in Table I. The app dataset used for this analysis includes an anonymised device identifier, session duration, metadata including connection type, and two aspects of session quality, namely duration of buffer underrun and number of interruptions. Session duration and buffering duration do not include startup delay. Note that buffer underrun relates to video stalling duration but may not be identical. This depends on the individual player implementation as some players may wait to fill the buffer to a specific level before playout starts again. We do not know the policy implemented on all the players but we know the duration in milliseconds for which the buffer was empty. The STB log dataset includes a (different) anonymised device identifier, session duration and a thresholded event that marks a loss-related impairment detected over a preceding time window (in the order of seconds). We can distinguish between DTT and multicast delivery. Multicast loss errors distinguish mild and severe packet loss and the DTT error is triggered due to low signal strength, based on vendor-implemented thresholds for packet loss and poor signal respectively. A session in VoD content demarcates a user starting and exiting the viewing of a single VoD asset, including any start-up delay, buffering and trick play operations. A live session in all platforms starts when the user expresses the intention to view the channel through the player or the remote control and ends when the user exits that channel. A viewer leaving and returning to the same asset or channel is recorded as a new session.

All loss-related impairments manifest to viewers as video and/or audio glitchiness and/or stalling of varying severity. Additionally, severe packet loss and DTT signal strength error messages are also displayed on the TV screen, persisting until resolved or dismissed by the customer. The transport protocols and the impact on viewers is summarised per error type in Table I.

TABLE I
Summary of loss-related impairment metrics on app and STB platforms

Type of TV viewing	Loss-related errors
TV Apps (Live and VoD, Unicast using HTTP streaming)	**Buffering events** • Presents as glitchiness and/or video stall due to buffer underrun • When resumed, stream skips content in live TV or plays from point of stall in VoD
IPTV (Live only, Multicast using RTP)	**Multicast ribbon error/notification** • Presents as glitchiness and/or video stall due to mild packet loss • When resumed, stream skips content **Multicast dialog error** • Presents as severe glitchiness and stalling due to severe packet loss • Also presents an on-screen dialog that can be dismissed by viewer
Freeview (Live only, DTT)	**DTT signal strength error** • Presents as severe glitcheness and stalling due to low signal strength • Also presents an on-screen dialog that can be dismissed by viewer

Note that performance metrics are not captured for any third-party on-demand content viewing on the STB, which is the most popular type of VoD consumption on that platform. Therefore, all analysis of on-demand viewing is restricted to the app platform. Due to the commercially sensitive nature of our data, we are unable to report specific absolute numbers and instead report orders of magnitude or relative values. Our data covers nationwide usage over a six week period starting August 2017, obtained with customer consent, anonymised, stored and analysed in a secure big data Hadoop cluster.

We apply filters to our dataset to only retain valid sessions with play start and end events, and reasonable viewing durations using respective distributions as explained in Section V. Our insights are derived from analysing over half a billion sessions from over a million customers across both STB and app platforms, and delivery technologies. DTT contributes the largest proportion of viewing, followed by viewing on multicast and finally app-based unicast content.

IV. Facets of engagement

Viewing engagement may be quantified in many ways, including:

- Absolute duration of video play (playtime) or session volume
- Asset completion ratio
- Time to return to platform

Fig. 1 shows the cumulative distribution of session playtime by buffering performance for live and VoD content on the app. Playtime does not include rebuffering or start up time. It shows that errored sessions are typically longer than error-free sessions. Independent of content type, the longer the session,

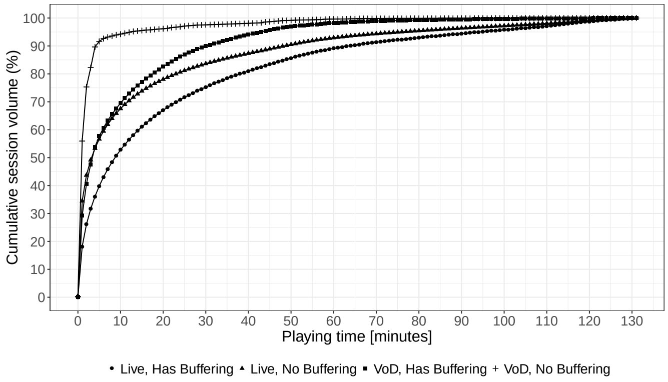
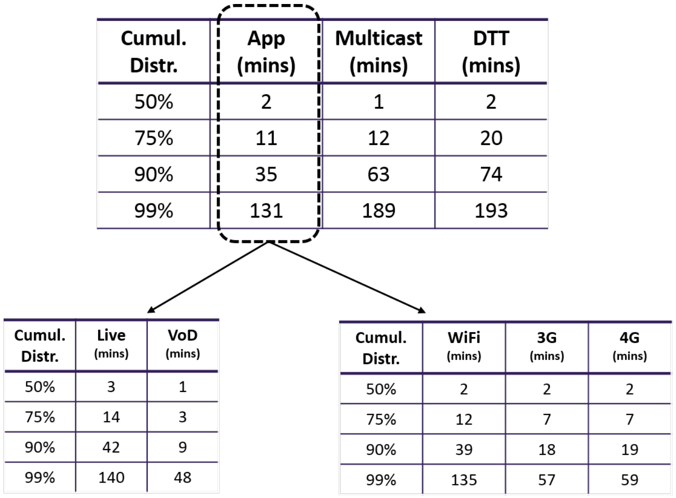


Fig. 1. CDF of errored and errorless session playtime by type of content

TABLE II
CUMULATIVE DISTRIBUTION VALUES OF VIEWING DURATION ON STREAMING TV APPS AND STB



Cumul. Distr.	App (mins)	Multicast (mins)	DTT (mins)
50%	2	1	2
75%	11	12	20
90%	35	63	74
99%	131	189	193

Cumul. Distr.	Live (mins)	VoD (mins)
50%	3	1
75%	14	3
90%	42	9
99%	140	48

Cumul. Distr.	WiFi (mins)	3G (mins)	4G (mins)
50%	2	2	2
75%	12	7	7
90%	39	18	19
99%	135	57	59

the more chance there is of a buffering event happening. Similarly, the higher the session volume, the greater the likelihood of packet loss materialising during the viewing experience. The same holds for STB viewing (figure not shown). Even if the probability of loss at any point in the session is uniform, the probability of the error materialising during the session increases linearly with longer viewing time. Longer sessions are more likely to experience errors. Therefore, absolute playtime and session volume cannot be directly compared between generic errored and error-free populations. However, we can frame performance measures in relation to the content length to assess their severity and this can be correlated to absolute playtime as in Section VI. Deviations from an expected growth in playtime is also comparable across platforms for increasing volume of errors, as proposed in Section VII.

An alternative facet is asset completion ratio, defined per session as the percentage ratio of asset playtime to asset length. For example, a viewer who watches thirty seconds of a one minute asset has watched 50% of the item. In order to relate errors to item abandonment, identifying customer intent is key – a VoD viewer clearly intends to watch the chosen item but a live TV viewer may watch a single session spanning multiple assets. Given that our dataset contains performance measures per session, we cannot attribute the errors to an individual live content item within the session and its completion ratio. Completion ratio is also typically lower for longer assets and higher for popular items. Our dataset also does not mark the position of advertisements in live content or start/end credits of the asset, if any.

In this paper, we derive a growth rate metric from absolute playtime for our long-term engagement analysis in Section VII. We primarily use asset completion ratios to characterise the immediate consequence of errors on session abandonment in Section VI.

V. BASELINE VIEWING DISTRIBUTIONS

Any viewing benchmarking based on loss-related errors must be preceded by an understanding of baseline viewing and error distributions. This section draws a cross-platform comparison in typical viewing behaviour. Table II shows selected percentile viewing durations by the three delivery platforms investigated in this paper (top table). Since the app platform has both variety in content type and access technology which informs customer behaviour, we also show the breakdown of viewing durations on the app platform by content type (bottom left) and by access technology (bottom right). Other access technologies are used in the population but we only study the three most popular to ensure sufficient data volume. The findings are discussed below.

App viewing tends to be shorter than STB viewing but viewing duration varies by connection type and content type. App viewing is shortest on mobile connections and VoD sessions are typically shorter than live content sessions. STB session durations show most spread on DTT. Viewing live content is most prevalent on the app and WiFi is the most common connection type – 65% of all app viewing is of live streams over WiFi and 11% is of live streams over 4G. VoD streaming over WiFi accounts for 18% of all app viewing in session volume.

Sessions on the three delivery mechanisms (app unicast, STB multicast and STB DTT) experience different profiles of loss-related errors. Over the six week period, the proportion of customers that experience one or more multicast packet loss errors or one or more DTT signal strength errors are approximately, and coincidentally, equal. An app viewer is 1.8 times more likely than a STB viewer to experience any loss-related errors. At a session level, STB multicast and app unicast sessions are 4.3 and 34.3 times more likely than DTT sessions respectively to record one or more loss-related impairments. Note, however, that preset error thresholds affect their likelihood and impact, and therefore cannot be directly compared. A DTT signal strength error may be substantially less likely to happen than a single app buffering event but could be more intrusive.

Buffering ratio is the percentage proportion of the total session duration that was spent buffering. It is a key performance indicator (KPI) in industry since it takes into account viewing time and buffering time. The authors of [28] investigate a live sport event. They report that 65% of their sessions have a buffering ratio of 2% or less and we find the same in our live app dataset as well. On the other hand, [26] report that 7% of sessions viewing long VoD content experience buffering of over 10%. We find that 22% of VoD sessions experience the same buffering across all content durations. Note that [26] analyse data across numerous content providers over all access technologies, whilst our app VoD viewing population connects primarily using WiFi or mobile technologies. We build upon these results to show a breakdown by connection type in Section VI-E.

For the rest of this analysis, we only consider sessions with durations less than the 99th percentile of their content type as very long sessions may well not have a viewer. Note that although we do not explicitly separate out and analyse channel surfing behaviour, a substantial portion of all viewing is channel surfing and is therefore included in our dataset.

VI. Impact of buffering on app session abandonment

In this section, we characterise the relationship between loss-related impairments and abandonment within the same session using the app dataset due to availability of logs for both VoD and live content types. We compare buffering impact on video playtime to findings in literature. We examine the effect of the following confounding factors on the relationship between aspects of buffering and session abandonment:

- Asset length
- Content type
- Connection type

We then study the importance of the two dimensions of buffer underrun, duration and number of interrupts, on session abandonment.

A. Impact of buffering ratio on video play time

Authors of [27] propose that a mixture of two exponential decay functions is a good fit to model the relationship between buffering ratio and asset playtime in long VoD content. They use data about long duration VoD content from [26] to do so and achieve a Root Mean Squared Error (RMSE) of 0.659. We have fitted the same type of decay function to our app VoD dataset, resulting in the following fitting function with asset playtime T and buffering ratio R:

$$T = 14.96219 \cdot e^{-0.47372 \cdot R} + 3.44452 \cdot e^{-0.05076 \cdot R} \quad (1)$$

Our dataset contains both long and short assets delivered primarily over WiFi and mobile technologies. The VoD assets in our dataset are typically short sport-related content including promo videos due to the nature of the product offered by the service provider on the app platform. We do not have actual asset lengths in our dataset and we estimate this from the percentage completion and absolute video play time, noting that any fastforward/rewind actions will impact this variable. We find that 50% and 70% of the VoD assets in our dataset are up to 2 minutes and 4 minutes long respectively. Some assets last many hours. Moreover, the data also includes channel surfers and users who abandon due to disinterest in content.

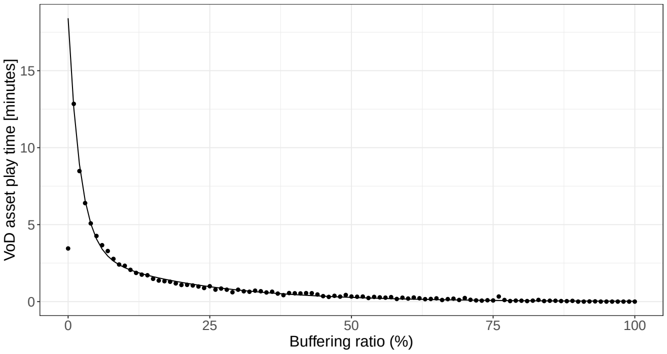


Fig. 2. Relationship between buffering ratio and play time showing a fitted mixture of two exponentials

This is exhibited in a noticeably low average playtime for buffering-free sessions ($R = 0$). We exclude this data point from our fitted model but it is included in Fig. 2. This plot shows the distribution between buffering ratio and VoD asset play time overlaid on the fitted curve of the mixture of two exponentials.

The RMSE of our model is 0.103 and the Pearson correlation coefficient of the fitted model to the raw data is 0.998, showing a better fit than that in [27] (RMSE: 0.659, Pearson correlation: 0.996). Using a mixture of two exponential decay curves does indeed show a good fit for this relationship, although the parameters of the function vary significantly between the two datasets due to the reasons described above. Therefore, service providers must understand the attributes of their assets on the platform and perform empirical studies to determine the best parameters for further modelling.

B. Role of content type on buffered session abandonment

Whilst all valid VoD sessions that meet the duration filter are taken into account, live sessions must be prepared differently to be comparable to VoD viewing. Raw live sessions are matched to a programme guide and split by programme item. A single session that spans three items is split into three viewing chunks and a completion ratio computed per asset based on programme duration. We then only retain sessions that start close to the start of the programme. The spread of start times around programmes varies by programme length – the shorter the programme, the closer to its start the viewers typically arrive. Therefore, for this analysis, we retain any session that starts within 10% of the programme duration on either side of the programme start time. For example, any session that starts within three minutes on either side of the start of a thirty-minute programme is included. Since it is conceivable that a viewer arrives just at the end of a previous programme and leaves just after the start of the subsequent programme, sessions that span more than three programmes are filtered out. This leaves us with a subset of live viewing sessions where we have a better estimate of the viewer's intent to watch the specific programme as a similar comparison to intentional VoD viewing. We also ensure that the impact of errors in the session on the completion ratio of

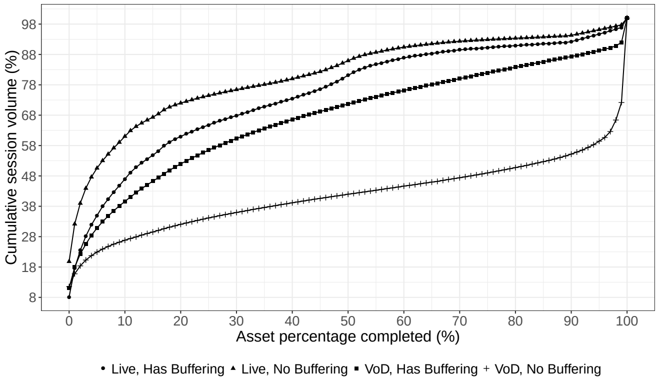


Fig. 3. Cumulative distribution of asset completion ratio by content type and buffering performance

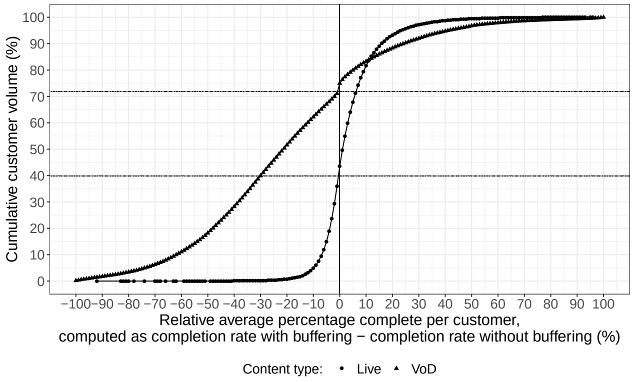


Fig. 4. Relative average asset completion ratio per customer viewing live and VoD content with and without buffering

the intended asset is clear. We find that 91% of all live sessions view one programme only and therefore our filter still retains a substantial portion of live sessions that were successfully matched to the programme schedule.

Fig. 3 shows the cumulative distribution of video asset completion ratios by content type and buffering performance. The four combinations of sessions with buffering and no buffering, live and VoD content are represented using unique symbols.

Both VoD and live content show baseline abandonment behaviour in the absence of any buffering. This may be explained by viewers losing interest in the item for reasons unrelated to quality. Also, since VoD assets tend to be shorter on the app than live assets, completion ratios tend to be higher on VoD than on live programmes. Compared to the error-free baseline VoD completion ratios, Fig. 3 shows that VoD sessions with buffering have higher asset abandonment than those without buffering. This shows that customers are likely to abandon VoD viewing in the presence of buffering. Live content, on the other hand, shows the opposite trend – sessions with buffering tend to complete more of the intended asset than sessions without buffering. We cannot interpret this to mean that buffering encourages completion but rather than customers persevere in spite of buffering to view the asset.

Whilst the results in Fig. 3 characterises a large volume of sessions on the platform, individual viewer tolerance to buffering is a significant confounding variable. We construct a quasi-experiment from our data to control for this. We identify all app viewer devices who consume live and VoD content and experience errored and error-free sessions in both content types over a three month summer period. This selects 5.8% of all viewers in that time. Fig. 4 shows the distribution of relative average completion ratio per customer per content type, computed as the difference between the average completion rate on sessions with and without buffering. A negative value corresponds to a viewer who had a lower average completion ratio when the session experienced buffering. For example, assuming a viewer watched on average 100% of all VoD asset where there was no buffering and 25% of all VoD assets when there was buffering, their average relative completion ratio would be $25 - 100 = -75\%$. Every customer contributes to one value on the x-axis. We expect to see a distribution of relative completion rates across our common viewer base, ranging from customers who abandon all their sessions with buffering and fully complete when without $((0 - 100)\%)$ to vice versa $((100 - 0)\%)$.

Completion ratios per customer also show the same trend across the viewing base. Fig. 4 shows that 71.8% of VoD viewers have lower average completion rates on sessions with buffering than on sessions without buffering. However, only 39.9% of viewers abandon earlier when encountering buffering on live viewing. This confirms that VoD viewers are more likely to abandon when encountering buffering and live viewers are more likely to persist despite buffering.

The contrast between live and VoD completion ratios in the presence of buffering can be explained by two competing forces that inform viewer engagement in every platform interaction. The first is the viewer's intent to watch the item. The second is the annoyance felt by poor delivery quality. In every viewing experience, one effect outweighs the other. Our result shows that the annoyance factor has the stronger influence when content is available on-demand, and in live viewing, intent to view outweighs the disruption caused by buffering. Time-sensitive content like sport or the news are typically watched live and even though some content is available on-demand after broadcast, one hypothesis for this user behaviour is perceived scarcity, where a user perceives increased value when the specific content item is viewed live than after the event. An analysis of drivers for perseverance with live content and session abandonment in VoD content is a valuable next step.

C. Cost of buffering interruptions

We have shown that streaming VoD viewers abandon sessions in the presence of buffering. In this subsection, we aim to quantify the disengagement cost of every additional interruption in a session.

Noting that longer sessions typically have more interruptions, discussed in Section V, we still find that asset completion ratios fall with increasing number of interruptions per session. However, the cost of disengagement is non-linear. The first interruption is most expensive – we report a 22% drop

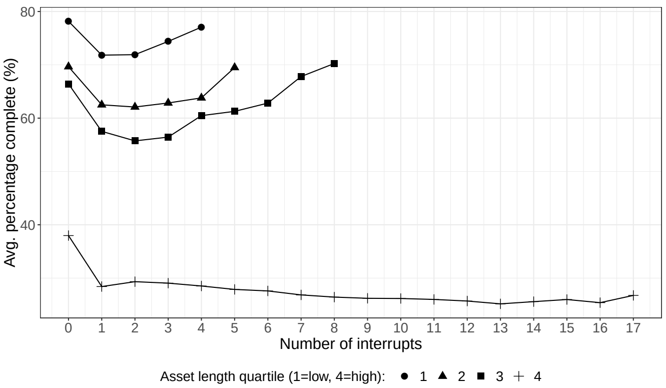


Fig. 5. Average asset completion over volume of interrupts per asset length quartile

in overall average asset completion rates from an error-free baseline of 62%. The next three interrupts cost an additional 10% loss in completion. The next five interrupts cost the final 4% loss in completion before a plateau is reached at 26–27%. This highlights the importance of error-free delivery but also shows some tolerance to interruption. We conclude that, on average, every additional interruption costs session engagement.

The rest of this section addresses the impact of other key variables on this overall relationship.

D. Role of asset length on buffered session abandonment

In this subsection, we present our findings on the impact that VoD asset length on the app platform has on the relationship between number of interruptions and asset completion ratios. We estimate the asset length from the completion ratio and absolute video play time per session as this is not available in our original dataset. Our estimates show that 70% of our content is up to 4 minutes long with much longer durations also present. We then bucket the estimated lengths into quartiles with the following approximate boundaries: 1=up to 1 minute, 2=up to 2 minutes, 3=up to 5 minutes, 4=over 5 minutes. Fig. 5 shows the plot of average completion ratio to the number of interruptions in the session for the four asset length quartiles.

We expect to see that short assets have higher completion ratios and higher number of interrupts taking place on longer assets, both confirmed by the figure. Whilst we observe an immediate drop in completion ratios with the first interrupt across all asset length quartiles, we then observe a recovery in asset completion, which appears counterintuitive. We hypothesise that viewers, having already persevered through the first interruption and knowing that the item is not longer than 5 minutes, persevere with the asset despite interruptions beyond a certain point. It would be interesting to see how abandonment behaviour develops within the session following each additional interruption but our data lacks this granularity.

We can quantify the average impact of interruptions per usage quartile q using data from this figure. We compute the normalised relative average completion rate $\bar{d}_q$ for each asset length quartile q over all interruptions I_q in that quartile. We define $d_{q,i}$ and $\bar{d}_q$ as follows with $\bar{p}_{q,i}$ being the average playtime for all sessions in quartile q with number of interruptions i:

$$d_{q,i} = \frac{\bar{p}_{q,i} - \bar{p}_{q,0}}{\bar{p}_{q,0}} \cdot 100\% \quad (2)$$

$$\bar{d}_q = \frac{1}{I_q} \sum_{i=1}^{I_q} d_{q,i} \quad (3)$$

We obtain the following values: $\bar{d}_1 = -4.48\%$, $\bar{d}_2 = -6.59\%$, $\bar{d}_3 = -6.54\%$, $\bar{d}_4 = -27.6\%$. We conclude from this result that not only do interruptions impact asset completion rates, the effect is greater with increasing asset length.

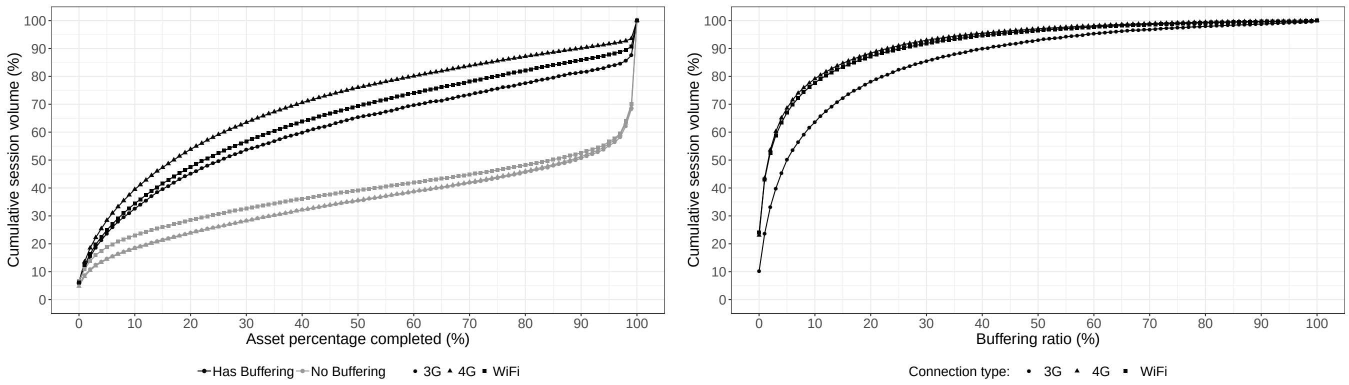
E. Role of connection type on VoD buffered session abandonment

We now focus on streaming VoD asset completion ratios in the presence of buffering by connection type. Fig. 6a shows the cumulative distribution of VoD asset completion ratios by connection type with and without buffering. We show the three most popular connection methods – WiFi, 4G and 3G. Every connection type has two lines on the plot. The lighter line shows the distribution of completion rates across all error-free sessions for that connection type. The black line shows the same for errored sessions. The error-free distributions for 3G and 4G overlap in the figure. Since WiFi is the most prevalent, it forms a good reference for VoD content completion ratios on cellular mobile connections. We find that overall completion ratios on error-free sessions are higher on mobile than on WiFi. This is explained by the nature of viewed assets – we find that viewers choose shorter assets on mobile than on WiFi, therefore reaching higher completion ratios on the former.

Fig. 6b shows the cumulative distribution of VoD session buffering ratios by connection type. This plot shows that sessions on WiFi and 4G have almost identical distributions of buffering ratios, and 3G sessions perform worst overall. Note that the viewing population is not necessarily common across the different connection types.

The two figures show that although viewers on 4G reach higher completion rates than those on WiFi when there is no buffering, they are the least tolerant population overall when encountering buffering. In contrast, 3G customers suffer the most buffering but are most persistent. This offers an interesting insight to service providers in determining how to prioritise between 4G and 3G performance improvements. Whilst the 3G network is unsurprisingly more error-prone, viewers are more tolerant. Those on 4G, however, are quicker to disengage than those on WiFi, despite identical buffering ratios seen in both networks.

Fig. 7 shows the cost of every additional interrupt per connection type, in line with Figs. 5 and 8 for the other confounding variables we study in this section. Fig. 7 shows the average completion ratio per number of interruptions for VoD sessions by connection type. It confirms the findings of Fig. 6b that error-free completion rates vary slightly by connection type and also that viewers on 4G show the highest



(a) Cumulative VoD asset completion ratio by buffering performance and connection type

(b) Cumulative VoD buffering ratios by connection type

Fig. 6. Impact of connection type on VoD session abandonment with buffering

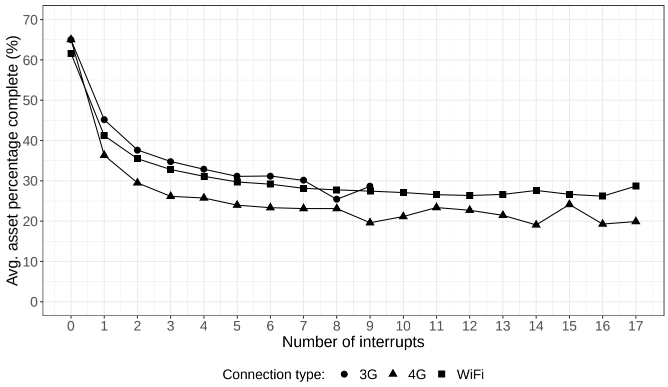


Fig. 7. Average VoD asset completion ratios by number of interrupts per connection type

disengagement for every added interruption whilst 3G viewers remain more tolerant.

We compute the normalised relative average completion rate $\bar{d}_c$ for each connection type c defined in analogy to Eq. 2 and 3.

We obtain the following values: $\bar{d}_{\text{3G}} = -44.3\%$, $\bar{d}_{\text{WiFi}} = -49.5\%$, $\bar{d}_{\text{4G}} = -60.1\%$. We conclude from this result sessions on 4G have the highest overall abandonment in the presence of interruptions followed by sessions on WiFi and finally 3G. Note that $\bar{d}_q$ is not comparable to $\bar{d}_c$ because of the different ways of aggregating the data. The values in $\bar{d}_c$ are brought down by the lower completion ratios observed when more than eight number of interruptions occur. This is typically experienced on long assets which have a lower completion ratio overall (quartile 4 of Fig. 5). Therefore, the effect of asset length influences $\bar{d}_c$. Future multi-dimensional analyses of the confounding factors could help isolate the individual effects.

F. Buffering duration vs. number of interruptions

Buffering interruptions within a session have two dimensions: duration and volume. We now compare the two facets in terms of cost of engagement on app VoD sessions through Fig. 8.

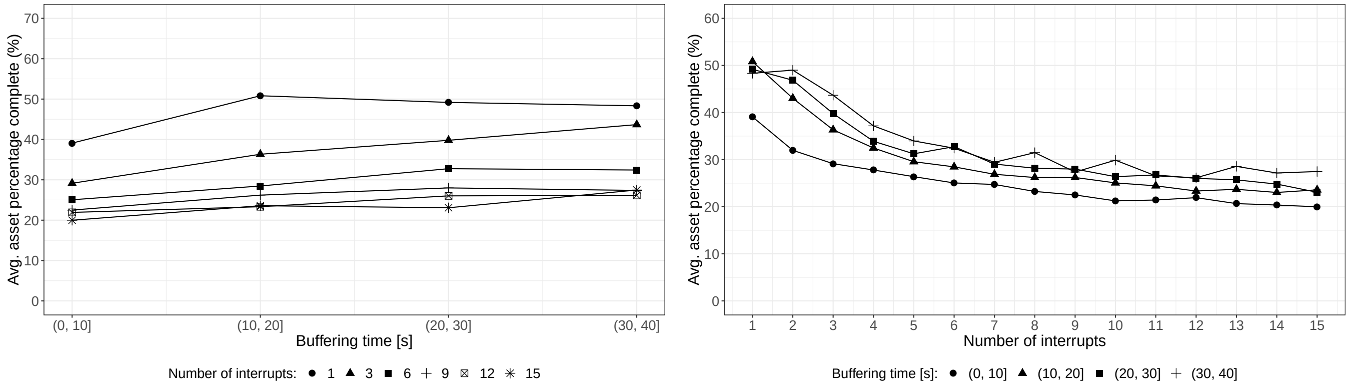
Fig. 8a shows the average completion ratio of increasing buffering duration with an overlay of the number of interruptions over which the buffering duration was spread. Fig. 8b is the corollary to Fig. 8a and shows the average completion rate by number of interrupts and each line shows an interval of buffering duration. A minimum session volume per data point is enforced to ensure credibility of the average value. We only show selected number of interruptions in Fig 8a and we bucket buffering durations in 10 second intervals for clarity. Cut-offs on both buffering duration and number of interrupts have been informed by respective population distributions to retain a large but robust proportion of our dataset.

Fig. 8 shows that VoD viewers persist to reach higher asset completion rates despite longer buffering durations (Fig. 8a). However, the more interruptions for the same buffering duration, the lower the completion ratio, i.e. higher abandonment (Fig. 8b). We recognise again that total asset length impacts tolerance to absolute buffering duration and number of interruptions. But given a fixed buffering duration, we aim to study the impact of number of interruptions, which can be done independent of asset length.

We compute the normalised relative average completion rate $\bar{d}_b$ for each buffering duration interval b defined in analogy to Eq. 2 and 3.

We obtain the following values: $\bar{d}_{(0,10]} = -36.0\%$, $\bar{d}_{(10,20]} = -41.9\%$, $\bar{d}_{(20,30]} = -36.1\%$, $\bar{d}_{(30,40]} = -31.2\%$.

We conclude that the number of interruptions is more detrimental to asset completion than buffering duration itself, within a boundary of tolerance. Given a fixed buffering duration, every additional interruption results in more abandonment. However, for a fixed number of interruptions, an increase in buffering duration typically does not show the same trend. Therefore, viewing engagement with the app VoD content may be increased by delivering items such that the number of interruptions is minimised whilst remaining within an envelope of buffering duration. We note that the same result has been observed in [2] through subjective studies using assets of fixed duration.



(a) Average VoD asset completion ratios by buffering duration per number of interruptions

(b) Average VoD asset completion ratios by number of interruptions for fixed buckets of buffering duration

Fig. 8. Comparison of cost of buffering duration vs. cost of number of interruptions on asset completion

VII. Impact of errors on long-term viewing engagement on apps and STB

Long-term loyalty to a platform is important for service providers for customer retention and show a different facet of viewing engagement. Completion ratios of individual assets can be very variable for each viewer and are also affected by interest in the specific content item and lifestyle. We find that live viewing shows higher asset completion in the presence of buffering, which may be explained by perceived content scarcity and the importance of time-sensitive viewing. This does not mean that there is no engagement benefit to error-free delivery in live content.

In this section, we investigate the impact of the presence and intensity of loss-related errors on long-term viewing engagement across both streaming app and STB platforms. Note that the potential types of impairments are greater in live viewing since there is more diversity in delivery technology and platform. Nonetheless, we present our findings for both live and VoD viewing.

A. Method to evaluate impact of error persistence on long-term engagement

Average errored session playtime is higher than average error-free session playtime (Fig. 1) since the likelihood of errors materialising is cumulative with increasing session duration. Therefore, absolute playtime cannot be directly compared to determine any loss of viewing due to errors. We developed a relative method based on rate of growth of viewing with worsening error performance that enables this comparison to be made.

A heavy user is differently impacted by errors than a light user, as shown later. Therefore, we first group customers into usage level categories $U = \{1, \ldots, 5\}$, defined by grouping the total number of sessions N viewed in the six week data period into intervals. Subsequently, we calculate the percentage of sessions that saw any loss-related errors e in the time period. For example, of all viewers who watched two sessions $(N = 2)$ on the app, most will have experienced buffering on none of those sessions, some on one session and others on both sessions. This gives us subpopulations of viewers at discrete percentage errored-session values of $e = 0\%$, $e = 50\%$ and $e = 100\%$ respectively. As the total session volume viewed increases, the potential percentage errored-session values becomes continuous. We are more interested in the spread of buffering across sessions than in the intensity of buffering within each session. We compute the average playtime $\bar{p}_{u,e}, u \in U$ for each subpopulation (u, e) with percentage of errored-sessions e. It is to be expected that the higher the value of u, the higher the average playtime $\bar{p}_{u,e}$, especially for the error-free subpopulations ($e = 0$). Therefore, we compute the percentage difference in average playtime from the error-free subpopulation for each u as shown in Eq. 4.

$$d_{u,e} = \frac{\bar{p}_{u,e} - \bar{p}_{u,0}}{\bar{p}_{u,0}} \cdot 100\% \tag{4}$$

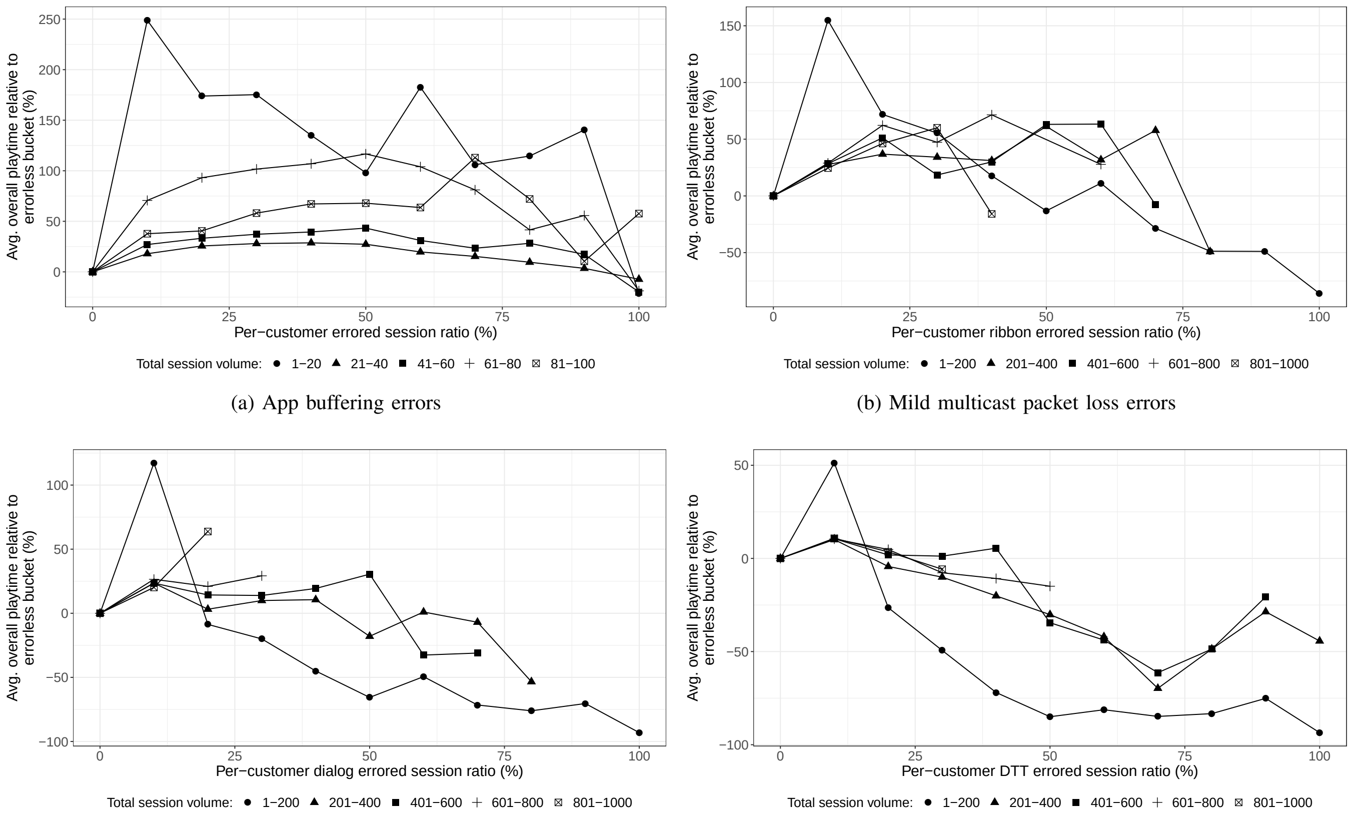
This key final step enables cross-comparisons across and within each usage level.

If error intensity has no impact on long-term viewing engagement, we expect to see a linear plot for all usage levels.

B. Loss-related errors and long-term live TV engagement by platform

This subsection presents our findings for live viewing. Fig. 9 plots the $d_{u,e}$ values for each of our four loss-related errors across multicast, DTT and app streaming platforms to quantify the impact of loss-related errors on long-term average playtime. We determine the cutoff for the maximum N shown on the figure using distributions of session volume by platform and error type. We find that our session volume cutoff for N includes a minimum of 95.5% of all sessions on the respective platform. We bucket N into intervals of 20 sessions to create the set of usage categories U in Fig. 9a and 200 sessions in Figs. 9b to 9d. Customers in the fifth highest session volume interval ($u = 5$) are the heaviest users of the platform. The x-axis shows values of e and has been bucketed into 10% intervals for clarity.

We observe non-linear growth rate in playtime with increasing error intensity. The concave curves in Fig. 9 capture the



(a) App buffering errors

(b) Mild multicast packet loss errors

(c) Severe multicast packet loss errors

(d) DTT signal strength errors

Fig. 9. Long-term average playtime by usage volume relative to error-free baseline per platform and type of impairment for live viewing

level of disengagement that groups of customers at similar usage levels show as they experience more errored sessions. The higher the number of live sessions that experience loss-related impairments at a given usage level, the lower the rate of growth of live viewing relative to the error-free subpopulation. The steeper the curvature, the more the disengagement. We conclude that in the long term, customers view less than we expect as they experience more errored sessions.

We now discuss observations on the impact of usage category u on viewer disengagement. The lightest users show a high peak in the first errored-session bucket $e = 10\%$ followed by a rapid decline in relative average playtime with increasing percentage errored sessions. Heavy users show a shallower rise to start followed by a moderate decline in comparison. This trend is most visible for viewers who experience DTT signal strength errors, followed by multicast dialog errors, ribbon errors and app buffering events.

We find that the fall in engagement is more pronounced on the STB platform compared to the app platform. Light users on the STB where $e > 10\%$ disengage aggressively irrespective of delivery technology and type of loss impairment. This leads to the convex shape of the curve, especially notable in the low usage category of viewers who experience DTT signal strength errors (Fig. 9d).

This low-usage DTT disengagement might be explained by content commoditisation and disinterest in the specific content item. Typically popular content delivered on DTT is easily accessible for free on other platforms in the UK including a popular streaming app. Low engagers who can afford to abandon DTT content on this service provider's STB could watch the same content elsewhere. In contrast, popularly-watched content delivered over multicast and the apps are more exclusive to the service provider and typically paid-for by subscription. Viewers who choose to stream content on the app may do so in the absence of a TV or when mobile and are more invested in the platform, potentially due to a lack of options in viewing live content. Therefore frustration due to buffering may not result in disengagement or may indeed be seen as normal for an Internet streaming experience [2].

We conclude this section with the finding that live content viewers continue watching a programme despite buffering (Section VI). However, they watch less than we expect in the long term with increasing errors. This holds across both app and STB platforms on unicast, multicast and DTT delivery technologies.

C. App buffering and long-term VoD engagement

Fig. 10 shows the relationship between average playtime relative to the error-free baseline for each usage level u for VoD content on the TV streaming apps.

In contrast to live content, VoD viewers on the app do not disengage with increasing buffering levels as represented by the linear relationship between rate of growth of average playtime and percentage errored sessions in Fig. 10. Therefore,

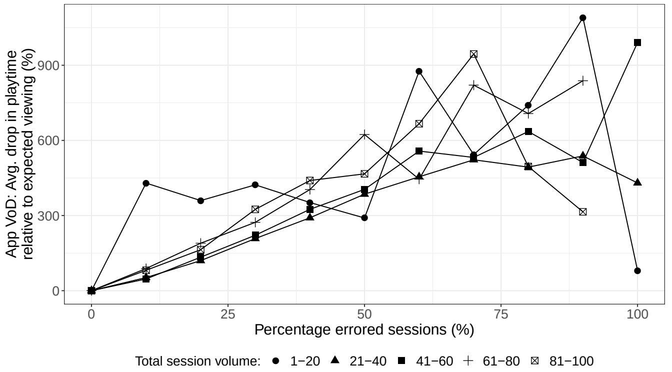


Fig. 10. Long-term average playtime by usage volume relative to error-free baseline for VoD viewing on the app

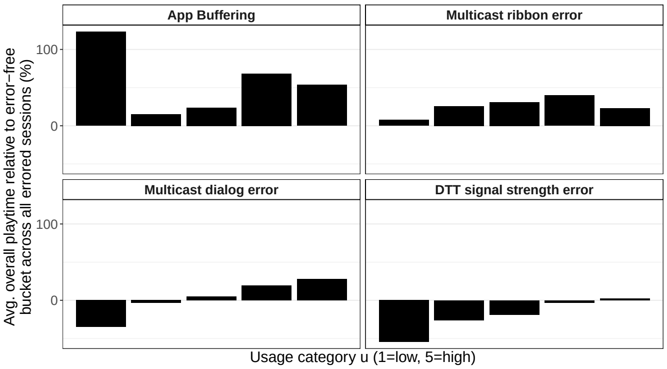


Fig. 11. Cross-platform benchmarking of the impact of loss-related impairments on long-term live playtime by usage category as given in Fig. 9

we conclude that the engagement cost of buffering in app VoD content is incurred immediately on the same session (Section VI) but not in the long term.

VIII. Cross-platform benchmarking of loss-related errors

This section is focused on quantifying the level of disengagement in a way that enables cross-comparison across platforms and types of impairments. We have investigated four types of loss-related impairments – buffering events on unicast delivery to TV streaming apps on portable devices, mild and severe multicast stream packet loss on a STB, and signal strength errors in DTT delivery to the STB. Our aim is to rank the various types of viewing impairments by their size of impact on viewing engagement.

Since live viewing forms the largest proportion of content consumption at present [31] and can be viewed across both platforms with this specific service provider, we focus on relative benchmarking of long-term disengagement of live content.

A. Method to quantify error-driven loss of growth in long-term engagement

In Section VII, we defined a relative average playtime metric $d_{u,e}$ to characterise the diminishing rate of increase in playtime with increasing percentage errored sessions e. In order to compare our four types of loss impairments across usage categories and platforms, we summarise $d_{u,e}$ over all error levels e to create $\bar{d_u}$, the mean of the average playtime relative to the error-free baseline for usage category u. Given that we bucket our percentage errored session values e into intervals of 10% from 0% to 100%, we expect 11 discrete steps to compute this average. However not each usage category u displays high values of e with sufficient representation in the population. For example, in order for a viewer who watches 1000 DTT sessions in our data period to have $e = 80\%$, 800 sessions must have experienced a severe signal strength error. We therefore find that our more severe error types have $\max(e) < 100\%$ for high usage levels. We define E_u to be the total number of discrete steps of e observed for usage category u. Our cross-platform comparison metric $\bar{d_u}$ is defined as follows:

$$\bar{d_u} = \frac{1}{E_u} \sum_{e=0}^{E_u} d_{u,e} \tag{5}$$

Every combination of loss impairment and usage category has one $\bar{d_u}$ value, resulting in $4 \cdot 5 = 20$ values for cross-platform comparison.

B. Relative benchmarking of error-driven long-term live TV disengagement

Fig. 11 shows all $\bar{d_u}$ values in our live viewing dataset across the four loss impairment types. The lower the value, the greater the disengagement in the presence of the loss impairment.

We find that users of the app platform show the most persistence in viewing across all usage categories, followed by multicast viewers who experience ribbon errors. All $\bar{d_u}$ are positive, which means that average playtime for errored-sessions is greater than average playtime of error-free sessions.

We also observe that STB viewers in higher usage categories appear more persistent to loss impairments on their respective platform. This trend is especially pronounced for viewers experiencing multicast dialog errors and DTT signal strength errors, both of which are displayed on the screen to the customer and are clearly intrusive to the viewing experience.

Comparing the height of the bars in Fig. 11 across usage categories, we conclude that DTT signal strength errors on the STB have the most impact on long term loss of playtime, followed by severe packet loss (dialog) and mild packet loss (ribbon) errors on multicast delivered to the STB. Buffering events on the app have the least impact on playtime for the same number of errored sessions per usage category.

Content and context influence viewer engagement. Three hypotheses explain why app users are more persistent:

- App users have more intent to watch a specific live item despite the inconvenience of the less immersive experience.
- The thresholds set for each impairment error is different and a single buffering event has less impact on viewing quality on a small screen than a signal strength error during evening entertainment on a large-screen TV.

- App users take a 'lean forwards' approach to TV viewing, whereas STB viewers are traditionally opt for a 'lean backwards' experience. Whilst hard to quantify, differences in state of mind of the typical user impacts their willingness to take action and abandon a viewing session.

It is noteworthy that DTT signal strength errors have greater impact on disengagement than severe packet loss (dialog errors) on multicast, although the perceptual impact to a viewer is very similar on the television. This effect is attributable to the difference in perceived value of content on each platform. Whilst popular multicast content is exclusive to the service provider's TV product, some DTT-delivered channels are available on other platforms such as catch-up players over broadband. Noting also that the DTT platform has a wider spread in distribution of session durations observed, it is plausible that channel surfers disengage sooner than viewers invested in the content. Recognising that usage category in our analysis is a function of session volume, repeating the analysis by grouping sessions in each usage category by their viewing duration will show the impact of errors on each platform on channel surfers and invested viewers.

We have concluded this section with a ranking and discussion of the four types of loss-related impairments that are fundamentally uncomparable to each other. In a world where more errors typically occurs with more viewing, it is challenging for service providers to establish the cost of those errors on viewing engagement at scale across their entire customer base. It is furthermore challenging to compare that cost across different bespoke delivery technologies and customer contexts. Whilst the ranking of the four impairments in this study is a novel finding in itself, we have also derived a proxy measure based on long-term viewing engagement to enable us to make this cross-comparison. This method may also be applied in other scenarios with different error types and service delivery technologies.

IX. Conclusion

We present the first study that benchmarks the quality of TV delivered across disparate platforms, delivery technologies and uncomparable quality metrics to a nationwide audience. We have shown that loss-related impairments have a tangible impact on viewer engagement, independent of the viewing platform. However, the nature of disengagement is influenced by a number of attributes. Viewers on 4G and 3G are less and more tolerant to buffering respectively than viewers on WiFi. The shorter the asset, the more tolerance viewers show to buffering interruptions on app VoD content. Live viewers continue watching a programme despite buffering. However, they watch less than we expect in the long term with increasing errors. This holds true across both app and STB platforms. In contrast, we conclude that the engagement cost of buffering in app streamed VoD content is incurred immediately, not in the long term. We find that viewers have some tolerance to buffering duration but little tolerance to increasing number of interrupts. An alternative delivery method that minimises the number of interrupts may increase user engagement with the app platform.

Potential next steps include a more detailed multi-dimensional drivers analysis to identify pockets of customers who show behaviour that differs from the global trends. We intend to model the cost to serve consequences of disengagement to highlight the impact of video quality on the business model of a service provider. We have found that subjective testing corroborates a number of our population-wide results but more subjective tests are underway to further explore the impact of visual impairments arising from measured errors on viewer engagement across disparate platforms and modes of content consumption.

References

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016–2021," Sep. 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[2] B. Allan, M. Nilsson, and I. Kegel, "A Subjective Comparison of Broadcast and Unicast Transmission Impairments," in *SMPTE 2018*, Oct. 2018, pp. 1–20.

[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," in *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, ser. SIGCOMM '13. New York, NY, USA: ACM, 2013, pp. 339–350. [Online]. Available: http://doi.acm.org/10.1145/2486001.2486025

[4] R. Mok, E. Chan, and R. Chang, "Measuring the quality of experience of HTTP video streaming," in *2011 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, May 2011, pp. 485–492.

[5] A. Khan, L. Sun, and E. Ifeachor, "Content Clustering Based Video Quality Prediction Model for MPEG4 Video Streaming over Wireless Networks," in *IEEE International Conference on Communications, 2009. ICC '09*, Jun. 2009, pp. 1–5.

[6] A. Murshed, A. Khalifeh, and M. Al-Taee, "Quality of experience analysis of real-time video streaming over lossy networks," in *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, Dec. 2013, pp. 1–6.

[7] M. Venkataraman, M. Chatterjee, and S. Chattopadhyay, "Evaluating Quality of Experience for Streaming Video in Real Time," in *IEEE Global Telecommunications Conference, 2009. GLOBECOM 2009*, Nov. 2009, pp. 1–6.

[8] J. Ahmed, A. Johnsson, R. Yanggratoke, J. Ardelius, C. Flinta, and R. Stadler, "Predicting SLA Violations in Real Time using Online Machine Learning," *arXiv:1509.01386 [cs, stat]*, Sep. 2015, arXiv: 1509.01386. [Online]. Available: http://arxiv.org/abs/1509.01386

[9] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, 2010.

[10] M. Nuhbegović, A. Čolaković, and A. Hasković, "Validating IPTV service quality under realistic triple play network conditions," in *2014 X International Symposium on Telecommunications (BIHTEL)*, Oct. 2014, pp. 1–6.

[11] H. J. Kim and S. G. Choi, "A study on a QoS/QoE correlation model for QoE evaluation on IPTV service," in *2010 The 12th International Conference on Advanced Communication Technology (ICACT)*, vol. 2, Feb. 2010, pp. 1377–1382.

[12] K.-J. Kim, W.-S. Shin, D.-K. Min, H.-J. Kim, J.-S. Yoo, H.-M. Lim, S.-H. Lee, and Y.-K. Jeong, "Analysis of key features in IPTV service quality model," in *2008 IEEE International Conference on Industrial Engineering and Engineering Management*, Dec. 2008, pp. 595–598.

[13] H. Meng, R. Huang, X. Wei, Y. Qian, and Q. Liu, "QoE prediction model for IPTV based on machine learning," in *2016 8th International Conference on Wireless Communications Signal Processing (WCSP)*, Oct. 2016, pp. 1–5.

[14] T. Hossfeld, C. Moldovan, and C. Schwartz, "To each according to his needs: Dimensioning video buffer for specific user profiles and behavior," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, May 2015, pp. 1249–1254.

[15] T. Hossfeld, M. Seufert, C. Sieber, T. Zinner, and P. Tran-Gia, "Identifying QoE Optimal Adaptation of HTTP Adaptive Streaming Based on Subjective Studies," *Comput. Netw.*, vol. 81, no. C, pp. 320–332, Apr. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2015.02.015

[16] H. T. T. Tran, N. P. Ngoc, T. Hossfeld, and T. C. Thang, "A Cumulative Quality Model for HTTP Adaptive Streaming," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2018, pp. 1–6.

[17] 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 Communications Surveys Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[18] ITU-T Publications, "Methodology for the subjective assessment of the quality of television pictures," Tech. Rep. BT.500, 2012. [Online]. Available: https://www.itu.int/rec/R-REC-BT.500

[19] ——, "Parametric non-intrusive assessment of audiovisual media streaming quality - Lower resolution application area," Tech. Rep. P.1201.1, 2012. [Online]. Available: https://www.itu.int/rec/T-REC-P.1201.1/en

[20] ——, "Objective perceptual multimedia video quality measurement in the presence of a full reference," Tech. Rep. J.247, 2008. [Online]. Available: https://www.itu.int/rec/T-REC-J.247/en

[21] W. Robitza, M. Garcia, and A. Raake, "A modular HTTP adaptive streaming QoE model — Candidate for ITU-T P.1203 ("P.NATS")," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2017, pp. 1–6.

[22] A. C. Begen, "Error Control for IPTV over xDSL Networks," in *2008 5th IEEE Consumer Communications and Networking Conference*, Jan. 2008, pp. 632–637.

[23] K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei, "Quantifying Skype User Satisfaction," in *Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '06. New York, NY, USA: ACM, 2006, pp. 399–410. [Online]. Available: http://doi.acm.org/10.1145/1159913.1159959

[24] 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 *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '14. New York, NY, USA: ACM, 2014, pp. 367–379, event-place: Austin, Texas, USA. [Online]. Available: http://doi.acm.org/10.1145/2591971.2591975

[25] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret, "Models of User Engagement," in *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*, ser. UMAP'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 164–175, event-place: Montreal, Canada. [Online]. Available: https://doi.org/10.1007/978-3-642-31454-4_14

[26] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the Impact of Video Quality on User Engagement," in *Proceedings of the ACM SIGCOMM 2011 Conference*, ser. SIGCOMM '11. New York, NY, USA: ACM, 2011, pp. 362–373. [Online]. Available: http://doi.acm.org/10.1145/2018436.2018478

[27] C. Moldovan and F. Metzger, "Bridging the Gap between QoE and User Engagement in HTTP Video Streaming," in *2016 28th International Teletraffic Congress (ITC 28)*, vol. 01, Sep. 2016, pp. 103–111.

[28] M. T. Diallo, F. Fieau, and J. Hennequin, "Impacts of video Quality of Experience on User Engagement in a live event," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–7.

[29] 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, Dec. 2013.

[30] M. Ghasemi, P. Kanuparthy, A. Mansy, T. Benson, and J. Rexford, "Performance Characterization of a Commercial Video Streaming Service," in *Proceedings of the 2016 Internet Measurement Conference*, ser. IMC '16. New York, NY, USA: ACM, 2016, pp. 499–511. [Online]. Available: http://doi.acm.org/10.1145/2987443.2987481

[31] "The Communications Market Report: United Kingdom," Tech. Rep., 2015. [Online]. Available: http://stakeholders.ofcom.org.uk/market-data-research/market-data/communications-market-reports/cmr15/uk/

Vidhyalakshmi Karthikeyan completed her PhD in 2019, MSc in Telecommunications in 2009 and BEng in Electrical and Electronic Engineering in 2008, all from University College London, UK. Following a decade spent in Applied Research at BT specialising in data science & machine learning applications for better TV service assurance and network analytics, she currently leads the Product Analytics team at BT across all of BT's consumer products and brands. She is an inventor on 24 patents and applications. Her interests lie in bringing the entire data ecosystem from governance, data collection, reporting, data science and machine learning models to the heart of organisational decision-making from product design to in-life support, enabling organisations to deliver superior customer experience and operational efficiency.

Brahim Allan received BEng and MSc degrees in Telecommunication engineering from Kings College London (University of London), United Kingdom in 2007 and 2009 respectively. He joined BT's research department in 2011 as a researcher working in the future content and application services team. His research interests includes ultrahigh-definition imaging, high dynamic range imaging, video analysis and their evaluation. His most recent interest is on Big Data, specialising in TV data analytics. He holds 2 patents and has published 7 papers. He is a member of BSI, British Standards Institution and represent BT in MPEG (Moving Picture Experts Group) standard meetings.

Detlef Nauck is the Head of AI and Data Science Research for BT's Applied Research Division located at Adastral Park, Ipswich, UK. Detlef leads a programme spanning the work of 30 international researchers who develop capabilities underpinning modern AI systems. A key part of the work is to establish best practices in Data Science and Machine Learning for conducting data analytics professionally and responsibly leading to new ways of analysing data for achieving better insights. Detlef is a computer scientist by training and holds a PhD and a Postdoctoral Degree (Habilitation) in Machine Learning and Data Analytics. He is a Visiting Professor at Bournemouth University and a Private Docent at the Otto-von-Guericke University of Magdeburg, Germany. He has published 3 books, over 120 papers, and holds over 20 patents.

Miguel Rio received the Ph.D. degree from the University of Kent, Canterbury, U.K., and the M.Sc. and M.Eng. degrees in informatics from the University of Minho, Braga, Portugal. He is a Professor in Computer Networks, Department of Electronic and Electrical Engineering, University College London, London, U.K. He has authored extensively in top ranked conferences and journals and has been a principal investigator in numerous research projects. His research interests include network measurement, network routing, new network control architectures and the application of machine learning techniques to computer networking problems.