

Scoring high: analysis and prediction of viewer behavior and engagement in the context of 2018 FIFA WC live streaming

Nikolas Wehner, Michael Seufert, Sebastian Egger-Lampl, Bruno Gardlo, Pedro Casas, Raimund Schatz

Angaben zur Veröffentlichung / Publication details:

Wehner, Nikolas, Michael Seufert, Sebastian Egger-Lampl, Bruno Gardlo, Pedro Casas, and Raimund Schatz. 2020. "Scoring high: analysis and prediction of viewer behavior and engagement in the context of 2018 FIFA WC live streaming." In Proceedings of the 28th ACM International Conference on Multimedia, October 12-16, 2020, Seattle, WA, USA, edited by Chang Wen Chen, Rita Cucchiara, and Xian-Sheng Hua, 807-15. New York, NY: ACM.
<https://doi.org/10.1145/3394171.3414016>.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>



Scoring High: Analysis and Prediction of Viewer Behavior and Engagement in the Context of 2018 FIFA WC Live Streaming

Nikolas Wehner*
nikolas.wehner@uni-wuerzburg.de

Michael Seufert*
michael.seufert@uni-wuerzburg.de

Sebastian Egger-Lampl
sebastian.egger-lampl@ait.ac.at

Bruno Gardlo
bruno.gardlo@ait.ac.at

Pedro Casas
pedro.casas@ait.ac.at
AIT Austrian Institute of Technology
Vienna, Austria

Raimund Schatz
raimund.schatz@ait.ac.at

ABSTRACT

Large-scale events pose severe challenges to live video streaming service providers, who need to cope with high, peaking viewer numbers and the resulting fluctuating resource demands, keeping high levels of Quality of Experience (QoE) to avoid end-user frustration and churn. In this paper, we analyze a unique dataset consisting of more than a million 2018 FIFA World Cup mobile live streaming sessions, collected at a large national public broadcaster. Different from previous work, we analyze QoE and user engagement as well as their interaction, in dependency to specific soccer match events, which have the potential to trigger flash crowds during a match. Flash crowds are a particular challenge to video service providers, since they cause sudden load peaks and consequently, the likelihood of quality problems. We further exploit the data to model viewer engagement over the course of a soccer match, and show that client counts follow very similar patterns of change across all matches. We believe that the analysis as well as the resulting models are valuable sources of insight for service providers, equipping them with tools for customer-centric resource and capacity management.

CCS CONCEPTS

• **Networks** → **Network performance analysis**; *Network measurement*; • **Computing methodologies** → Classification and regression trees.

KEYWORDS

Quality of Experience (QoE); User Engagement; Live Video Streaming; Network Performance

ACM Reference Format:

Nikolas Wehner, Michael Seufert, Sebastian Egger-Lampl, Bruno Gardlo, Pedro Casas, and Raimund Schatz. 2020. Scoring High: Analysis and Prediction of Viewer Behavior and Engagement in the Context of 2018 FIFA

*Now with: University of Würzburg, Germany. This work was performed while Nikolas Wehner and Michael Seufert were with AIT Austrian Institute of Technology.

This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in:

MM '20, October 12–16, 2020, Seattle, WA, USA

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7988-5

<https://doi.org/10.1145/3394171.3414016>

WC Live Streaming. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, October 12–16, 2020, Seattle, WA, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3394171.3414016>

1 INTRODUCTION

Live and on-demand video streaming services are exhibiting remarkable growth in terms of popularity and adoption rates. This development is illustrated best by a four-fold global video traffic increase between 2015 and 2020 as forecasted by Cisco, with video constituting 82 % of consumer Internet traffic by 2020 [4]. Understanding user engagement as well as the causes and the impact of quality degradation in video streaming services is of prime relevance to content and service providers, who have to run and maintain a complex content delivery infrastructure, providing high levels of Quality of Experience (QoE) to the audience. To this end, large-scale collection and analysis of quality and viewer engagement data have become key ingredients for video streaming quality monitoring and related scientific research [5]. Also, proactive scaling of infrastructure capacity based on system load predictions has become a cornerstone of QoE management in this domain [3, 30].

In this context, streaming of *live* events in real-time poses exceptional requirements for QoE management, not only because of the need to minimize playout latency, but also because of high peak loads and levels of strain caused by unexpected surges in user activity and viewer fluctuations (flash crowds) [33]. In particular, streaming of sport events such as soccer matches represents a major challenge, due to the high emotional involvement of the audience, the importance of key events such as goals and penalties, and the resulting risk of disappointing and frustrating viewers due to quality problems occurring in critical moments of peak attention. While QoE and user engagement in the context of video streaming have been fairly well studied in lab and field settings (e.g., [5, 16]), research work investigating and modeling QoE and user engagement in the context of streaming large-scale sport events is still rare (cf. [10, 33]), particularly when it comes to analyzing and modeling the impact of match events and flash crowd phenomena.

In this paper, we investigate viewer behavior and engagement in the context of live online video streaming of soccer matches. The specific context is the 2018 FIFA Soccer World Cup, which is one of the most popular broadcasted sport events worldwide. We collect and analyze a unique, large-scale dataset with more than 1.3 million live video streaming sessions, consisting of multiple quality metrics monitored at the mobile devices of viewers watching FIFA

WC football matches. Measurements are collected for users of a live IPTV streaming service provided by a national public broadcaster.

The first goal of the study is to obtain a deeper understanding of the interplay between match events, quality, and user behavior, and quantify key relationships therein. The second goal is to reliably forecast levels of viewer engagement, to enable better capacity management and consequently, better QoE of live video streaming of soccer events, specially in the context of flash crowds. To address these goals, we structure the study around the following research questions:

RQ1: What is the impact of key match events (e.g. goals) on user engagement and QoE?

RQ2: To which extent do QoE and user engagement correlate with each other in the context of soccer live streaming?

RQ3: How accurately can user engagement over the course of a soccer match be predicted using machine-learning models?

Answering these questions is far from trivial, specially due to the complexities associated to the identification of causal relationships, and the usual pitfalls when mixing up correlation and causality. Indeed, how to properly discover and analyze causal relationships in data captured in the wild are open research questions. We therefore followed standard recommendations and paid special attention to the different steps involved in the statistical analysis of the data, to avoid common pitfalls and biased conclusions.

The remainder of the paper is structured as follows: after a discussion on background and related work, we describe the technical setup and methodology used on the measurement campaign (Section 3). A description and characterization of the collected data is presented in Section 4. Section 5 reports the results of the analysis, targeting the first two research questions. In Section 6 we introduce a model to predict user engagement over time, and evaluate its applicability to load management. Finally, we discuss conclusions and provide an outlook on future work.

2 BACKGROUND AND RELATED WORK

In this section, we discuss video QoE in general, followed by a survey of related work for analyzing video QoE and user engagement in the field and in particular for large events.

QoE for HTTP adaptive video streaming (ABR/HAS) is a well-investigated research topic [26]. The main QoE influence factors include stalling or re-buffering events [8, 13, 24, 34], initial playback delay [7, 12], and quality adaptation [14, 18, 22, 23, 31, 32]. Stalling has certainly the strongest negative impact on QoE, and while viewers can better cope with quality changes and playback delays, they are key metrics to understand overall video QoE.

The impact of QoE on user engagement and the prediction of user engagement are also widely investigated research topics [5, 9, 17, 28, 29]. These studies show that especially the visual quality and stalling strongly impact the abandonment rate. Regarding mobile streaming, mobile users tend to have lower user engagement than non-mobile users. Further, mobile users with cellular access usually abort their streams even faster than users with WiFi [19].

QoE for live and video on demand (VoD) streaming is also a widely investigated research topic. In [20], authors conduct a behavior study for live and VoD IPTV. Their findings include that the

average viewing time for VoD is significantly longer than that for live services, and that video quality does not play such a big role for live streaming compared to VoD. In [5], authors analyze VoD and live streaming with respect to user engagement on user and video level. They show that the stalling ratio is the most important metric at the view-level, and that initial delays are critical for user engagement. The impact of video stream quality on viewer behavior is studied in [16]. The study concludes that users tend to quit the startup of the video playback if the startup delay exceeds 2 seconds.

However, only a few papers so far analyze live streaming for large-scale events, in particular for sports events. The studies reported in [10], [2], [1], [6], and [33] come closest to our work, as they also address specific aspects of our paper, but in a slightly different manner. In [10], authors analyzed video streaming data of a major South American content provider for the 2014 FIFA WC. They correlate the usual QoE quality indicators with the session duration, and show that in the context of world cup matches, especially bitrate degradation and frequent stalling events negatively impact the session duration. The authors of [1, 2] use a quasi-experimental framework to investigate the causal impact of QoE on user engagement for live streams of the 87th Academy Awards. They also show that the stalling rate and the average bitrate impact the user engagement the strongest. Finally, the study reported in [6] also focuses on the analysis and impact of video streaming for large sport events – the Super Bowl 2013, but here for the specific case of cellular network performance. Their findings include that the uplink configuration plays a key role in the performance of the video streaming.

Flash crowds for large-scale events are investigated in [33], where the authors analyze live and VoD streams of the 2008 Beijing Olympics. They reveal that the observed flash crowds, in this context a specific competition, focus mainly on Chinese athletes winning a gold medal. This indicates that the location in which the content is consumed plays an important role when considering flash crowds for worldwide sports events, due to the attachments of the audience to national teams or athletes. Other research investigating the relationship between QoE and flash crowds for video streaming can be found in [11], [25], [15], [21]. However, the impact of flash crowds on video QoE is a sparsely researched topic in general, and in particular in the context of sports events.

To the best of our knowledge, no research on the interaction between QoE and user engagement in dependency to content-driven flash crowd events has been done so far. Based on the 2018 FIFA WC matches, we treat specific soccer match events, e.g., goals, as potential flash crowd triggering events, and investigate the impact of these events on user engagement and QoE. Additionally, we are the first to model user engagement for large-scale soccer events over the course of a match, which offers a valuable tool and source of insights to service providers in terms of resource provisioning.

3 METHODOLOGY

To investigate the relationships between match events, quality, and user engagement during live soccer event streaming, we performed two types of data collection activities: streaming playback monitoring, and extraction of relevant match events.

3.1 QoE and player actions monitoring

The data collection was done through a custom video streaming analytics platform (<https://www.ait.ac.at/loesungen/experience-tools/qoestream/>) used by the public broadcaster for large scale quality monitoring. At the client-side, the different metrics and player events were extracted from the video player provided by the broadcaster, gathered on session and clip-view level. For practical reasons, we focused in particular on the iOS version of the player, which uses the native AVPlayer API to report video metrics. The collected data was properly anonymized, centralized in an analytics platform, and properly curated for further analysis.

3.2 Match event extraction

To obtain a complete picture of the live streaming experience and user engagement during the soccer matches, as well as to assess the nature and importance of a specific moment of a match, it is necessary to know which kind of match event (goal, kick-off, etc.) happened at which point in time. We extracted the relevant match information and the match events from the live blogs of the official FIFA website. We used Selenium to automatically open the live blogs in Chrome, then scroll the blogs to the bottom of the site to guarantee that all events are loaded, and finally iterate through all the blog entries to extract the match events and the corresponding timestamps. The types of match events were characterized by the FIFA with a numerical identifier. Along with the identifier, a description of the event was provided which specified the event, e.g., which player of which team had scored. The monitored match events included the goals, bookings, the start and end of halftime, penalties, and others. We excluded Video Assistant Referee (VAR) events since the blog entries did not contain timestamps for these events. We performed the extraction for all matches and ended up with around 1200 match events distributed over all 64 matches. However, in this work, we focus mainly on the halftime and goal events, and hence consider only 316 events for the analysis.

4 DATASET CHARACTERISTICS

The 2018 FIFA WC took place in Russia from June 14 to July 15, with a total of 64 matches played. Of these, only 56 matches were broadcasted by the national service provider due to the fact that the last matches in the group stage were played in parallel, and only one match could be broadcasted at once. Two of the 56 monitored matches did not contain a sufficient number of views and were thus removed from the dataset. Further, we removed sessions with more than 15 stalling events or with an initial delay higher than 20 seconds. In both cases, this corresponds approximately to the 99% percentile of the data. Sessions with an unusual length were also filtered out. After the filtering, 54 matches with 1,325,479 live sessions remained for the study.

Figure 1a depicts the distributions of the most relevant QoE Key Quality Indicators (KQIs), including initial playback delay, the video bitrate, the number of bitrate changes, the number of stalling events, and the average stalling length, for all the sessions. The x-axis represents the unit values and the corresponding unit for each metric can be found in the legend, e.g., the unit for the initial delay is seconds, while the unit for bitrate changes is the number only. The y-axis denotes the CDF of the observed metric. For 70%

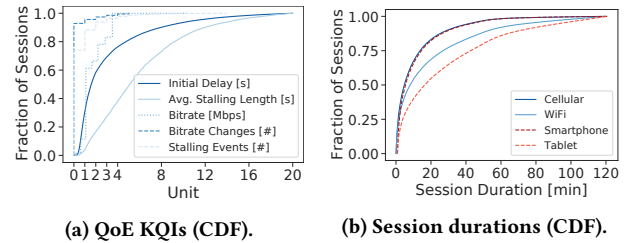


Figure 1: Distributions of QoE KQIs (left) and session durations by platform (right).

of the sessions, the initial delay is below 2.5 seconds, and below 10 seconds for 90% of the sessions. For the bitrate distribution, six steps can be seen, corresponding to the different quality levels of the streams. The majority of sessions show a bitrate of around 1.09 Mbps, which is also the median. The most prevalent bitrates besides 1.09 Mbps are around 0.64 Mbps and 2.15 Mbps. Only a small share shows very low or very high bitrates. The distributions for the number of bitrate changes and the number of stalling events show that the majority of the sessions did not experience bitrate changes or stalling events. Nevertheless, there are sessions with one, two, and three stalling events, with a decreasing occurrence, respectively. The average stalling length is highly distributed, from slightly above 0 to 10 seconds, while only a small share of the streams shows stalling events with an average length higher than 10 seconds. The mean average stalling length is around 6 seconds, while the median is around 5 seconds. All in all, our results show that the broadcaster provided a stable streaming experience with short waiting times, few interruptions, and a low bitrate fluctuation.

The distributions of the session duration, grouped by the used end device and the network access type, are depicted in Figure 1b. The blue lines represent the network access types, WiFi and cellular access, which includes 2G, 3G, and 4G. The red dashed lines represent the used end device type, smartphone or tablet. The distributions for cellular and smartphone overlap, as cellular access had been used mostly by smartphones. In general, sessions played out on a smartphone show a much shorter session duration (mean of 10 minutes) compared to sessions played out on tablets (mean of 26 minutes). When comparing WiFi and cellular access, sessions in WiFi are longer on average (mean of 19 minutes) than sessions watched with cellular access (mean of 10 minutes). These differences suggest that tablet users are more prone to watch matches more attentively, while smartphone users tend to just peek and check for specific events. Further, users at home usually watch a match in a more comfortable situation, using WiFi and devices with larger screens, e.g. tablets. More than 90% of our tablet users also used WiFi. On the other hand, people on the road usually have to use the inconvenient way of cellular access and small handheld devices. These usage contexts are also mirrored in the data.

The relative number of session starts, i.e., arrivals, and session ends, i.e., departures, during the course of a world cup match averaged over all matches are shown in Figure 2. The y-axis presents the arrivals and departures, respectively, in an interval of one minute, whereby the number is normalized for each match by the maximum number of arrivals and departures observed for the corresponding match, respectively. The arrival and departure behaviors are mostly

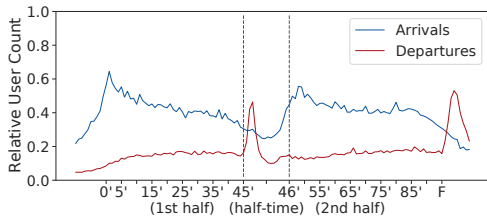


Figure 2: Arrivals and departures during a match averaged over all matches. The x-axis denotes the match time.

as expected. Shortly before the start of a match, the arrivals increase strongly. After the match start, new arrivals decrease at a steady pace, until a second burst occurs with the start of the second half-time. At the end of the match, arrivals decrease strongly. In contrast, departures stay at a similarly low level during the match, except for two strong bursts, happening at the end of the first half-time and the end of the match. Interestingly, most users stop the stream at the end of the first half-time and start a new stream as soon as the second half starts.

5 INTERPLAY BETWEEN EVENTS, USER ENGAGEMENT, AND QOE

We now focus on questions RQ1 and RQ2, and thus, investigate the interplay between match events, user engagement, and QoE.

In general, the relationship between events, user engagement, and QoE can be described with a uni-directional chain. Match events can solely impact user engagement, which in turn can influence the QoE. The opposite direction is clearly not valid, since user engagement can not impact match events. However, QoE can for itself also impact user engagement, as demonstrated in the literature [27] and in the common practice.

For a more precise definition of the time before and after an event – let us say a goal for example, a multi-window concept is applied to the data. Figure 3 shows a sketch of this concept, where A represents the time window before the specific event, and B represents the time window after the event took place – indicated by the red bold line. Window C serves for the analysis of the interaction between user engagement and QoE KQIs, which is performed later in Section 5.3. The length of the windows is described by different interval sizes, which we use to define the analysis boundaries. Indeed, we consider multiple intervals for the analysis, since we can only guess how much time has to pass until the impact on the audience is visible. The used intervals are 30, 60, 120, and 180 seconds. The upper boundary of 180 seconds has been chosen because we could observe that goal events affected the user engagement up until 165 seconds after the goal occurred. This is further described in Section 6.1. For each window, we compute the mean values of the target metrics, which correspond either to the user engagement metrics, or to the QoE KQIs. All the investigated metrics are listed in Table 1. We take four specific metrics reflecting user engagement, related to the start and end of a video streaming session. These include: the session start ratio, the session end ratio, and the average elapsed times between started/ended sessions, respectively.

After the calculation of the metrics for all relevant events, we perform a paired *t-test* to check for any differences between the user engagement metrics on the A and B windows. Note that the

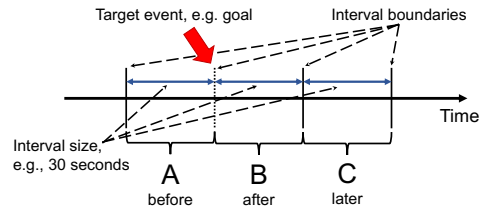


Figure 3: Multi-window concept for analyzing QoE and engagement before (A) and after (B, C) a match event.

assumptions of normality and homogeneity hold. To compensate for the multiple comparisons, we apply the Bonferroni-Holm correction on the obtained *p-values* of the *t-tests*.

To validate that any differences are not by chance, we use randomly sampled timestamps from periods without events to check whether the match events were actually responsible for influencing the user engagement metrics. For a valid comparison with the match events, our sample size equals the number of considered match events (316). For the validation, we perform independent *t-tests* which compare the differences between the A and B windows for the relevant events, and the differences between the A and B windows for the sampled baseline of non-event periods. Again, we adjust the *p-values* with the Bonferroni-Holm correction. Similar to [16] we assume causality here, because we rule out other possible influences, as we work only in the context of the events. In addition, we average the considered metrics over multiple matches considering the same events for all matches, which further helps in filtering out and averaging erratic correlations not linked to causality.

Our analysis is limited to the most frequently occurring events, including *goal*, *start half*, and *end half*. Other events of interest include penalties and yellow/red cards; however, we chose to omit them for this analysis due to the relatively small sample size. Note also that we excluded five matches that resulted in extra times and penalty shootouts from the analysis, since they present significant outliers in terms of user engagement at the end of a match as compared to the rest of the matches.

5.1 Impact of match events on engagement

To study the impact of events on user engagement, we compute the change of user engagement metrics (cf. Table 1) before and after an event occurred (cf. Fig. 3, windows A and B).

Figure 4 shows the *t-test* results for the impact of the goal events on user engagement. Figure 4a considers solely goal events with a scoring difference of one or zero goals, i.e., exciting matches, while Figure 4b considers solely goal events with a scoring difference of two or more goals, i.e., mostly already decided matches. The goal-scoring difference can serve as an approximate indicator for the match excitement, which very likely influences the viewing behavior of the audience. To avoid complexity, we do not take the current match time into account, which could for sure play an important role in terms of excitement. The x-axis denotes the different user engagement metrics, the y-axis denotes the used window interval size. Heatmap tiles are colored according to the obtained *p-value* of the corrected *t-tests*, whereby tiles with a darker blue state a lower *p-value* and tiles with a brighter blue state a higher *p-value*. Tiles with *p-values* below the significance level of

Table 1: Metrics along with unit, abbreviation, and description used for the impact analysis.

Group	Metric	Unit	Abbreviation	Description
User Engagement	Session Start Ratio	%	SR	Ratio between started sessions and active sessions
	Session Start Inter-arrival Time	ms	SI	Average elapsed time between started sessions
	Session End Ratio	%	ER	Ratio between ended sessions and active sessions
	Session End Inter-arrival Time	ms	EI	Average elapsed time between ended sessions
QoE KQIs	Initial Delay	s	ID	The average initial delay
	Bitrate	kbps	BR	The average played out bitrate
	Number of Bitrate Changes	#	BC	The average number of bitrate changes
	Number of Stalling Events	#	SE	The average number of stalling events
	Average Stalling Length	s	ASL	The average stalling length

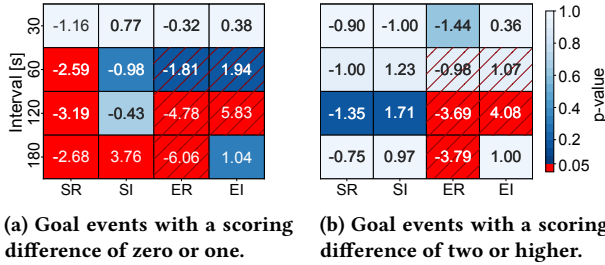


Figure 4: Impact of different goal event types on user engagement metrics (columns). Each tile contains the respective t-test statistic. Red tiles mark significant A/B differences, hatch patterns indicate significant deviation of the respective metric from the event-free baseline.

5% are colored in red. Numeric values within tiles represent the t-test statistic. When an independent t-test results in a corrected p-value below 5%, then the corresponding tile is overlaid with a hatch pattern. This means that the metric values for the time periods containing the event are significantly different compared to the baseline periods without match events. Hence, for red tiles, the hatch overlay suggests that the found differences are actually caused by the event, since the presence/non-presence of an event is the only difference here. In contrast, hatched blue/white tiles represent metric/window-duration combinations for which a significant average metric difference before vs. after the event was *not* detected, but still, the respective metric averages deviate significantly from the event-free baseline. Since these cases are typically located next to hatched red tiles, we assume that in fact the event has a relevant impact, but the window size is too small or too large to detect temporal change.

Our results suggest that exciting matches (characterized by teams' goal counts being to each other) cause stronger audience interest and thus an increase in the arrivals, which is observed from the increased session start ratio and the decreased session start inter-arrival times. In contrast, this session starting behavior is not visible for matches with a higher scoring difference, for which the winner is already more certain. While this result is expected, it confirms that the match situation is crucial for the streaming behavior of the crowd, and that goals in an exciting match can cause a growth of the audience in only a few minutes. When considering the session ending behavior, both scenarios exhibit a similar user engagement and validation pattern. Besides the analysis of goal events, the kickoff

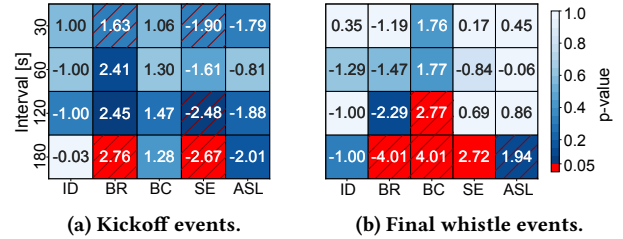


Figure 5: Impact of different event types on QoE KQIs (columns). Each tile contains the respective t-test statistic. Red tiles mark significant A/B differences, hatch patterns indicate significant deviation of the respective metric (ID, etc.) from the baseline.

event and the final whistle event proved to be the events with the strongest impact on the user engagement, while the first halftime end and second halftime start showed slightly weaker effects. This result is also not surprising, especially when considering the arrival and departure patterns depicted in Figure 2.

5.2 Impact of match events on QoE

We follow the same analysis to study the impact of an event on the QoE KQIs, using the specific KQIs listed in Table 1.

The analysis revealed that especially the kickoff event, i.e., the start of the match, and the final whistle, i.e., the end of the match, strongly impact the QoE KQIs for the intervals of 120 and 180 seconds. This comes as expected, since a high number of users start or leave the streaming at the match begin or match end, using/releasing available network and system resources. For example, we observed that after the kickoff event, the bitrate decreases for more than 35 kbps on average, and that the number of stalling events increases approximately 2-3% when relating it to the mean stalling number of 0.54. This applies for all intervals, but the results are only significant and validated for the interval of 180 seconds (cf. Figure 5a). In contrast, in the context of the final whistle, an increase of the bitrate and a decrease of the number of bitrate changes, the number of stalling events, and the average stalling length is observed (cf. Figure 5b). Only the bitrate and the number of bitrate changes revealed to be significant and validated. In general, this indicates that an interval size of 120 to 180 seconds is required to see any effects caused by an event.

We can also observe that the actual halftime events do not nearly impact the QoE KQIs as strongly as the kickoff and the final whistle event. Further, goal events showed no significant impact on the QoE KQIs. This applies also when additionally considering the goal-scoring differences of the matches, i.e., the excitement of the matches. Finally, the initial delay showed no significant increase or decrease for any of the events.

As a **conclusion regarding RQ1**, we can say that the impact on the user engagement varies in dependency of the match event type, but we could observe an impact for all the considered events. Regarding QoE, results showed that the QoE is influenced only by those match events that start, interrupt, or end a match, and that goals play a negligible role. Furthermore, our results show that a match event’s impact seems to be visible within one to three minutes after its occurrence.

5.3 Interplay between engagement and QoE

In this section, we address RQ2. To analyze the interaction between user engagement and QoE, we quantify how changes of the QoE KQIs or changes of the user engagement metrics at an earlier time impact the QoE KQIs or user engagement metrics in a subsequent interval. In particular, we compute trends of source metrics between the A and B windows, and assess the subsequent, possibly causally connected trends of a target metric in the B and C windows (cf. Figure 3). For all scenarios, we consider both an increase and a decrease of the source metric from window A to B, and investigate the resulting trend of the target metric from window B to C. We do so by computing correlations between the differences of the windows A-B and the differences of the windows B-C with the Spearman’s rank-order correlation coefficient (SROCC). As before, multiple window sizes and the same metrics are considered. Additionally, we account for the impact of match events by randomly sampling from appropriate timestamps, such that the timestamp between windows A and B is either a certain match event (dependent) or a random timestamp (independent of match events).

If a match event happened between windows A and B, results show low to moderate correlations between ± 0.3 to ± 0.5 . The findings include that reduced arrivals after a goal are correlated with a reduced number of stalling events (0.41) and shorter stalling lengths (0.36). This can happen, e.g., if the crowd loses its interest in the match due to a deciding goal, and thus, the load decreases, which causes the subsequent QoE improvement. Another QoE improvement in terms of bitrate is correlated with halftime end events (0.35) where again the load decreases due to many users dropping out of the stream. In contrast, QoE degradations are correlated with halftime start events, i.e., when the load suddenly peaks due to many users tuning in, which results in longer initial delays (0.52).

When analyzing the impact of the QoE KQIs on the user engagement in the context of the match events, no remarkable findings were identified. This is likely caused by the fact that users who start a stream after a match event are not aware of the QoE, so it is not reasonable to investigate correlations with arrivals. For departures after match events, we neither observed any remarkable impacts, which suggests that QoE KQIs do not additionally motivate or preclude that users drop out of the stream after a match event.

Next, we consider the interactions independent of events, i.e., using randomly sampled timestamps between windows A and B.

Regarding the impact of user engagement on QoE, we observe that a higher load caused by many arriving users is correlated to longer waiting times with respect to both initial delay (0.40) and stalling events (0.51), and thus, QoE degradations. On the other hand, fewer arrivals correlate with QoE improvements, which are represented by higher bitrates (0.45) and a lower number of bitrate changes (0.47). The analysis of the impact of QoE KQIs on user engagement shows that a lower bitrate is correlated with more users dropping out (0.43). As expected, the dropout rate is decreased when less stalling events occur (0.41) and subsequently, the average stalling length also decreases (0.49). These findings are in line with related work and again emphasize the impact of stalling and video quality on the dropout rate. Note that all these results were obtained again with the 120 and 180 second intervals.

As a **conclusion regarding RQ2**, we can state that the observed correlations for the event-independent analysis are mainly in line with earlier findings, even though in a weaker form. In contrast, for the event-dependent analysis, we could not observe any impact of the QoE on user engagement, but only that user engagement can affect the QoE. Thus, we observed different interaction patterns for both analyses, which shows that match events can impact both, user engagement and QoE.

6 USER ENGAGEMENT PREDICTION

In this section, we target RQ3 by developing a quantitative model that predicts the user count (and thus: system load) over the course of a match. A model of the current user count over the course of a match provides two benefits to streaming service providers. First, it allows to detect significant deviations from typical user behavior patterns, useful for locating failures and bottlenecks limiting the number of concurrent streams. In foresight, it allows predicting future load, which is helpful for planning and adjusting the capacity of the streaming system. In addition, for the research in this paper, it allows investigating the correlations between match events and user access behavior. For example, it could be analyzed whether after a goal has been scored, one also has to expect significant deviations from normal access behavior patterns, e.g. in terms of an increased number of clients accessing the live stream. We therefore developed two models for this purpose. The *general model* aims at characterizing the typical user count over the course of a soccer match; the *prediction model* aims at predicting the future user counts, based on the general model and historical user count data available at a given point in time during a match.

6.1 General model

To model the user count over time, we inspected all monitored matches manually. Here, a generic pattern was detected, which is depicted in Figure 6a. The figure shows the progress of the relative user counts over the match time, averaged over all matches, as a dashed red line. It can be observed that especially the end of the match, but also the end of the first half, are the phases in which most users concurrently access the stream. At the beginning of each half, the user count increases quickly, but the increase slows down towards the end of the half. In the halftime, a fast and large dropout of users occurs, which saturates quickly.

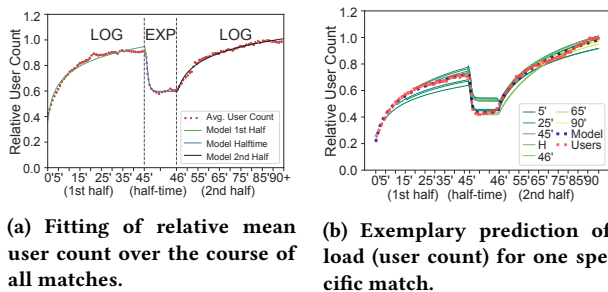


Figure 6: Fundamental engagement/load model and example for user count predictions. The x-axis denotes the match time, the y-axis denotes the user count, relative to the maximum observed user count for a match.

This generic pattern confirms that a match can be naturally divided into three phases, namely, the *first half*, the *half-time*, and the *second half*. Note that the opening and final match and the five matches with extra time had to be excluded for this analysis, due to highly different loads. The *first half* can be typically characterized by a logarithmic increase (green solid line), and the user count u at time t is modeled with three parameters: $u(t) = a_1 \cdot \log(b_1 + t) + c_1$. The *half-time* shows an exponential decrease (blue solid line), which is also modeled with three parameters: $u(t) = a_2 \cdot \exp(b_2 \cdot t) + c_2$. Finally, the *second half* follows the same logarithmic trend as the first half (black solid line), and thus, the user count is described with the same kind of model: $u(t) = a_3 \cdot \log(b_3 + t) + c_3$. For the rest of the monitored matches of the world cup, the user count could be fitted very well with this three-phases model, reaching always a coefficient of determination R^2 above 0.921, and a relative mean absolute error below 3.24%.

To investigate if a goal causes a significant deviation from normal access behavior patterns, we further investigated the deviation of the actual user count and the modeled user count. We found that the average deviation is positive, which means that the user count after a goal is larger than explained by the general model. Moreover, this average deviation shows first an increase from 0% to the maximum of 2% at 93s after the goal, and then a decrease until around 165s after the goal, where it stabilizes at a slightly positive level of around 0.05%. This deviation can be well fitted with a parabolic function in the range of 0 to 180s after the goal. This also validates our use of the selected window sizes in the previous analyses. If the parabolic function is added to the general model after each goal, the fitting of the load improves. However, for the matches with goals, the mean improvement of R^2 is 0.0004, and the mean reduction of the relative mean absolute error is 0.02%. This shows that applying the goal correction only gives a very marginal improvement. Thus, in the remainder of this work, only the general three-phases model is used, which already shows highly accurate fitting performance.

6.2 Prediction model

In general, streaming providers are interested in estimating the future development of user count (and consequently, system load) while airing a given match. Thus, based on the presented general model, we develop a model that relies on machine learning to predict the future, final parameters of the model from the currently

observed fitting parameters. For this purpose, the historical data of observed user counts of a match are fitted in steps of 5 minutes, and the parameters of the model are extracted. Note that only the past and current model phases are considered. This means that, e.g., to fit the historical data of a match up to the 20th minute, only the first phase of the model is used, while to fit the historical data of a match up the 75th minute, all three phases are used.

Extracted features include the current match time, the final fitting parameters of all already completely observed past model phases, the current fitting parameters of the current model phase, the goodness of the current fit in terms of R^2 , and the difference between the current parameters and the average parameters for that phase (over all matches). Moreover, due to match-specific events (e.g., injury time), the lengths of the three phases slightly varied over the different matches. Thus, in addition, the best split time after each already completely observed phase and the initial user counts in each observed phase are stored as features. This means that the number of features increases after each phase, because more information is available, which can be considered final. For example, at minute 20 of a match, only the initial user counts at the beginning of the match and the current fitting parameters of the first model phase are available. In contrast, at minute 75, final information about the first two model phases are available, in addition to the initial user counts and the current fitting parameters of the third phase. The corresponding labels, i.e., the prediction targets, are the final parameters of the current model phase, which would be obtained after fitting the whole historical data of that phase.

The prediction model is trained on the extracted features for all monitored matches, and uses three random forest (RF) regressors - one for each phase - to predict the future, final model parameters in each phase, in an iterative way. This takes into account the three different numbers of features as discussed above. If the RF-predicted behavior deviates too much from the actually fitted behavior (difference in R^2 larger than 0.1), the current fitting parameters are used instead of the RF-predicted parameters, to mitigate the propagation of poor predictions. The future behavior of the current phase is then obtained and used as a starting point to predict the model parameters of the next phases. Average parameter values are used for the yet unknown features in these models. This allows to chain the trained random forest models, and eventually obtain a prediction for the entire course of the match. To smooth the potential discontinuities between the different phases of the model, a Savitzky-Golay filter with a window length of 91 seconds and a polynomial order of 1 is applied as a final post-processing step. The training was implemented in scikit-learn, using 5-fold cross-validation on the training set to optimize the hyperparameters of each random forest regressor; these include: the number of decision trees in the forest, the maximum depth of the trees, and the split criterion. To generate training and test sets, an "outer" 5-fold cross-validation is performed, such that the set of all matches was divided into five parts. Four parts of the matches were used for training the models, and the performance was tested on the fifth part. The parts were rotated five times, such that each part once served as the test set, which means that for each match, a prediction could be obtained from a model that was not trained on that match.

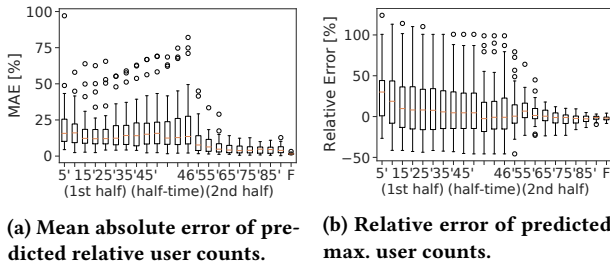


Figure 7: Results of user count prediction model validation. The x-axis denotes the match time at (and thus available data on) which the model was executed.

6.3 Evaluation of the prediction model

Figure 6b shows the exemplary prediction of the match Serbia vs Switzerland after every 5 minutes. The prediction of the future user counts are visualized as green lines with decreasing intensity of the color, i.e., the first prediction at 5' has the most intense dark green color, while the last prediction at 90' has the least intense light green color. The red line shows the actual user counts over the course of the match, and the blue line is the final fitted model, which is based on all observed data. The very first prediction after 5' is already quite good, and overestimates the maximum user count by less than 10%. Also, all other predictions approximate the real user count very well for this match. The maximum relative error of around 20% is observed for the prediction after 46'. Next, we evaluate the performance of the prediction model over all matches.

Figure 7a shows the performance in terms of the mean absolute error for predicted user counts over the future course of the match, relative to the maximum user counts of each match. The last entry (F) of the x-axis shows the performance of the final model, which was fitted on all historical data, and serves as performance baseline. The y-axis shows the box plots of the error distributions over all matches. The plot shows that even the early predictions already perform very well, and all show a median error below 20%. While outliers of up to 100% are observed for the very first prediction, outliers of at most 80% prediction errors can be observed for subsequent predictions, e.g., at the first prediction on the second half. Prediction errors become smaller as the match runs and more data is available, with maximum errors below 15% from the 65th minute onward, and median errors below 5%.

Figure 7b presents results on the error of the predicted maximal user counts, which is especially relevant for system capacity planning. As expected, the variance of the error distribution over the different matches becomes smaller for increasing time. The median is close to 0, which shows that the prediction is very good. In addition, the trained prediction model rather tends to overestimate the maximum user counts, which is more advantageous than underestimation for provisioning purposes. For some matches, the early predictions overestimate the maximum count by up to 112%; however, these extreme errors become smaller and also less likely throughout the match. Latest from minute 55, the predictions yield a very high accuracy for all matches, with only marginal errors.

To summarize, and as a **conclusion regarding RQ3**, we developed a model that can accurately predict the future user counts, which represents highly valuable information for load and capacity

management of streaming providers. Already early predictions provided by the model, e.g., after the 5th minute of the match, can be used to forecast the user counts over the whole course of the match with very high accuracy. We also showed that the goal events had only a minor impact on the user counts, but that the model's accuracy could still be further improved when considering this impact, albeit only marginally.

7 CONCLUSION

In this paper, we investigated the potential of specific match events to act as a trigger for flash crowds, based on more than a million live streams of the 2018 FIFA WC broadcasted by a large national public IPTV provider. For this purpose, we analyzed the impact of match events on user engagement and QoE, as well as their interplay in the context of match events. We also compared measurements obtained in the context of events with those obtained from event-free match periods, to quantify the influence of the presence of events on user engagement and QoE metrics.

Our results show that match events have significant impact on user engagement (and in turn, QoE KQIs) and that the intensity of the impact strongly depends on the type of match event. In addition, we found that match events typically impact user engagement and QoE within one to three minutes after their occurrence. Finally, we developed a novel three-phases model, which describes user engagement over the course of a soccer match with very high accuracy. On top of this model, we developed a machine learning based approach to predict the future number of concurrent users, which enables streaming providers to forecast likely shape and amplitude of system load over the course of a match already since its very beginning.

Our analysis would have certainly benefited from including measurement data obtained from additional client platforms beyond iOS (Android, Web), as this would have enabled the development of even more accurate models, as well as the comparative investigation of quality and behavior patterns across client platforms. However, such multi-platform data was not available to the authors in 2018 as it is today. Regarding future work, we thus envisage repeating the presented analysis and validate the proposed prediction model on a larger scale, when the next soccer championship takes place. We also aim to apply our analysis and modeling approach to data from live streaming of other types of large sports events which are prone to generate flash crowds (e.g., olympic games), to verify whether our method is limited to soccer streaming or not. Finally, a large scale analytics dataset collected from multiple platforms will also enable deeper analysis of cause-and-effect relationships on behalf of methods like quasi-experimental designs (cf. [16]), allowing the isolation of single influencing factors in field data.

All in all, we believe that the conducted analysis as well as the resulting models are valuable sources of insight to video streaming service providers, empowering their network operation and management capabilities with tools enabling customer-centric resource and capacity management.

ACKNOWLEDGMENTS

This work was performed together with the Austrian Broadcasting Corporation (ORF) and NOUS Wissensmanagement GmbH.

REFERENCES

- [1] Adnan Ahmed, Zubair Shafiq, Harkeerat Bedi, and Amir Khakpour. 2017. Suffering from buffering? Detecting QoE impairments in live video streams. In *2017 IEEE 25th International Conference on Network Protocols (ICNP)*. IEEE, 1–10.
- [2] Adnan Ahmed, Zubair Shafiq, and Amir Khakpour. 2016. QoE analysis of a large-scale live video streaming event. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. 395–396.
- [3] Alcardo Alex Barakabitze, Nabajeet Barman, Arslan Ahmad, Saman Zadtootaghaj, Lingfen Sun, Maria G. Martini, and Luigi Atzori. 2019. QoE Management of Multimedia Streaming Services in Future Networks: A Tutorial and Survey. *IEEE Communications Surveys & Tutorials* (2019), 1–1. <https://doi.org/10.1109/COMST.2019.2958784>
- [4] Cisco. 2018. Cisco Visual Networking Index: Forecast and Trends, 2017–2022. (2018). <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf>
- [5] Florian Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. 2011. Understanding the Impact of Video Quality on User Engagement. In *Proceedings of the ACM SIGCOMM*. Toronto, Canada.
- [6] Jeffrey Erman and K.K. Ramakrishnan. 2013. Understanding the Super-Sized Traffic of the Super Bowl. In *Proceedings of the 2013 Conference on Internet Measurement Conference (IMC '13)*. Association for Computing Machinery, New York, NY, USA, 353–360. <https://doi.org/10.1145/2504730.2504770>
- [7] Marie-Neige Garcia, Dominika Dytko, and Alexander Raake. 2014. Quality Impact Due to Initial Loading, Stalling, and Video Bitrate in Progressive Download Video Services. In *Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, Singapore.
- [8] Deepti Ghadiyaram, Janice Pan, and Alan C Bovik. 2015. A Time-varying Subjective Quality Model for Mobile Streaming Videos with Stalling Events. In *Applications of Digital Image Processing XXXVIII*.
- [9] Thiago Guarnieri, Jussara Almeida, and Alex Vieira. 2019. An Adaptation Aware Model to Predict Engagement on HTTP Adaptive Live Streaming. In *2019 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–6.
- [10] Thiago Guarnieri, Idilio Drago, Alex B Vieira, Italo Cunha, and Jussara Almeida. 2017. Characterizing QoE in large-scale live streaming. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 1–7.
- [11] Jian He, Yonggang Wen, Jianwei Huang, and Di Wu. 2013. On the cost-QoE tradeoff for cloud-based video streaming under Amazon EC2's pricing models. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 4 (2013), 669–680.
- [12] Tobias Hoßfeld, Sebastian Egger, Raimund Schatz, Markus Fiedler, Kathrin Masuch, and Charlott Lorentzen. 2012. Initial Delay vs. Interruptions: Between the Devil and the Deep Blue Sea. In *Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX)*. Yarra Valley, Australia.
- [13] Tobias Hoßfeld, Raimund Schatz, Michael Seufert, Matthias Hirth, Thomas Zinner, and Phuoc Tran-Gia. 2011. Quantification of YouTube QoE via Crowdsourcing. In *Proceedings of the International Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE)*. Dana Point, CA, USA.
- [14] Tobias Hoßfeld, Michael Seufert, Christian Sieber, and Thomas Zinner. 2014. Assessing Effect Sizes of Influence Factors Towards a QoE Model for HTTP Adaptive Streaming. In *Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX)*. Singapore.
- [15] Tobias Hoßfeld, Lea Skopin-Kapov, Yoram Haddad, Peter Pocta, Vasilius A Siris, Andrej Zgank, and Hugh Melvin. 2015. Can context monitoring improve QoE? A case study of video flash crowds in the internet of services. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*. IEEE, 1274–1277.
- [16] S Shunmuga Krishnan and Ramesh K Sitaraman. 2013. Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-experimental Designs. *IEEE/ACM Transactions on Networking* 21, 6 (2013), 2001–2014.
- [17] Pierre Lebreton, Kimiko Kawashima, Kazuhisa Yamagishi, and Jun Okamoto. 2018. Study on viewing time with regards to quality factors in adaptive bitrate video streaming. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.
- [18] Blazej Lewcio, Benjamin Belmudez, Amir Mehmood, Marcel Wältermann, and Sebastian Möller. 2011. Video Quality in Next Generation Mobile Networks – Perception of Time-varying Transmission. In *Proceedings of the IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR)*. Naples, FL, USA.
- [19] Zhenyu Li, Gaogang Xie, Mohamed Ali Kaafar, and Kave Salamatian. 2015. User behavior characterization of a large-scale mobile live streaming system. In *Proceedings of the 24th International Conference on World Wide Web*. 307–313.
- [20] Ning Liu, Huajie Cui, S-H Gary Chan, Zhipeng Chen, and Yirong Zhuang. 2014. Dissecting user behaviors for a simultaneous live and VoD IPTV system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 10, 3 (2014), 1–16.
- [21] Eliseu César Miguel, Ítalo Cunha, Cristiano M Silva, Fernando Carvalho, and Sérgio VA Campos. 2017. Resource-constrained P2P streaming overlay construction for efficient joining under flash crowds. In *2017 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 639–644.
- [22] Pengpeng Ni, Ragnhild Eg, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen. 2011. Flicker Effects in Adaptive Video Streaming to Handheld Devices. In *Proceedings of the 19th ACM International Conference on Multimedia (MM)*. Scottsdale, AZ, USA.
- [23] Ozgur Oyman and Sarabjot Singh. 2012. Quality of Experience for HTTP Adaptive Streaming Services. *IEEE Communications Magazine* 50, 4 (2012), 20–27.
- [24] Yining Qi and Mingyuan Dai. 2006. The Effect of Frame Freezing and Frame Skipping on Video Quality. In *Proceedings of the 2nd International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IHH-MSP)*. Pasadena, CA, USA.
- [25] Julius Rückert, Björn Richerzhagen, Eduardo Lidanski, Ralf Steinmetz, and David Hausheer. 2015. Topt: Supporting flash crowd events in hybrid overlay-based live streaming. In *2015 IFIP Networking Conference (IFIP Networking)*. IEEE, 1–9.
- [26] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2015. A Survey on Quality of Experience of HTTP Adaptive Streaming. *IEEE Communications Surveys & Tutorials* 17, 1 (2015).
- [27] Michael Seufert, Sarah Wassermann, and Pedro Casas. 2019. Considering User Behavior in the Quality of Experience Cycle: Towards Proactive QoE-aware Traffic Management. *IEEE Communications Letters* PP (01 2019), 1–1. <https://doi.org/10.1109/LCOMM.2019.2914038>
- [28] Michael Seufert, Nikolas Wehner, Florian Wamser, Pedro Casas, Alessandro D'Alconzo, and Phuoc Tran-Gia. 2017. Unsupervised QoE Field Study for Mobile YouTube Video Streaming with YoMoApp. In *Proceedings of the 9th International Conference on Quality of Multimedia Experience (QoMEX)*. Erfurt, Germany.
- [29] Muhammad Zubair Shafiq, Jeffrey Erman, Lusheng Ji, Alex X. Liu, Jeffrey Pang, and Jia Wang. 2014. Understanding the Impact of Network Dynamics on Mobile Video User Engagement. In *Proceedings of the ACM SIGMETRICS*. Austin, TX, USA.
- [30] Maria Torres Vega, Cristian Perra, Filip De Turck, and Antonio Liotta. 2018. A Review of Predictive Quality of Experience Management in Video Streaming Services. *IEEE Transactions on Broadcasting* 64, 2 (Jun 2018), 432–445. <https://doi.org/10.1109/TBC.2018.2822869>
- [31] Huyen TT Tran, Thang Vu, Nam Pham Ngoc, and Truong Cong Thang. 2016. A Novel Quality Model for HTTP Adaptive Streaming. In *Proceedings of the 6th IEEE International Conference on Communications and Electronics (ICCE)*. Ha Long, Vietnam.
- [32] Fei Wang, Zesong Fei, Jing Wang, Yifan Liu, and Zhikun Wu. 2017. HAS QoE Prediction Based on Dynamic Video Features with Data Mining in LTE Network. *Science China Information Sciences* 60, 4 (2017), 042404.
- [33] Hao Yin, Xuening Liu, Feng Qiu, Ning Xia, Chuang Lin, Hui Zhang, Vyas Sekar, and Geyong Min. 2009. Inside the bird's nest: measurements of large-scale live VoD from the 2008 olympics. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. 442–455.
- [34] Kai Zeng, Hojatollah Yeganeh, and Zhou Wang. 2016. Quality-of-experience of Streaming Video: Interactions between Presentation Quality and Playback Stalling. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. Phoenix, AZ, USA.