A Quest for an Internet Video Quality-of-Experience Metric

Athula Balachandran   Vyas Sekar‡   Aditya Akella†

Srinivasan Seshan   Ion Stoica*   Hui Zhang

Carnegie Mellon University † *University of Wisconsin Madison*
‡ *Stony Brook University* * *University of California Berkeley*

ABSTRACT

An imminent challenge that content providers, CDNs, third-party analytics and optimization services, and video player designers in the Internet video ecosystem face is the lack of a single "gold standard" to evaluate different competing solutions. Existing techniques that describe the quality of the encoded signal or controlled studies to measure opinion scores do not translate directly into user experience at scale. Recent work shows that measurable performance metrics such as buffering, startup time, bitrate, and number of bitrate switches impact user experience. However, converting these observations into a quantitative quality-of-experience metric turns out to be challenging since these metrics are interrelated in complex and sometimes counter-intuitive ways, and their relationship to user experience can be unpredictable. To further complicate things, many confounding factors are introduced by the nature of the content itself (e.g., user interest, genre). We believe that the issue of interdependency can be addressed by casting this as a machine learning problem to build a suitable predictive model from empirical observations. We also show that setting up the problem based on domain-specific and measurement-driven insights can minimize the impact of the various confounding factors to improve the prediction performance.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: measurement techniques, performance attributes

General Terms

Human Factors, Measurement, Performance

1. INTRODUCTION

With the decreasing cost of content delivery and the growing success of subscription and ad-based business models (e.g., [2]), video traffic over the Internet is predicted to increase in the years to come, possibly even surpassing television based viewership in the future [3]. An imminent challenge that all players in the Internet video ecosystem—content providers, content delivery networks, analytics services, video player designers, and users—face is the lack of a standardized approach to measure the *Quality-of-Experience* (QoE) that different solutions provide. With the "coming of age" of this technology and the establishment of industry standard groups (e.g., [13]), such a measure will become a fundamental requirement to promote further innovation by allowing us to objectively compare different competing designs [11, 17].

The notion of QoE appears to many forms of media and has a rich history in the multimedia community (e.g., [9, 10, 14, 15]). However, Internet video introduces new effects in terms of measuring both quality and experience:

- *Measuring quality:* Internet video is delivered using HTTP-based commodity technology over a largely unreliable network via existing CDN infrastructures. Consequently, the traditional encoding-related measures of quality like Peak Signal-to-Noise Ratio are replaced by a suite of quality metrics that capture several effects introduced by the delivery mechanism—buffering, bitrate delivered, frame rendering rate, bitrate switching, and startup delay [6, 33].
- *Measuring experience:* In the context of advertisement- and subscription-supported services, the perceptual opinion of a user in a controlled study does not necessarily translate into objective measures of engagement that impact providers' business objectives. Typical measures of engagement used today to approximate these business objectives are in-the-wild measurements of user behavior; e.g., fraction of a particular video played and number of visits to the provider [6, 33].

To obtain a robust QoE measure, we ideally need a *unified* and *quantitative* understanding of how low-level quality metrics impact measures of experience. By unified, we want

to see how the *set* of quality metrics taken together impact quality, as opposed to each metric in isolation. This is especially relevant since there are natural tradeoffs between the metrics; e.g., lower bitrate can ensure lower buffering but reduces the user experience. Similarly, by quantitative, we want to go beyond a simple correlational understanding of "metric M impacts engagement", to a stronger statement of the form "changing metric M from x to x' changes engagement from y to y'".

Unfortunately, the state of the art in our understanding of video QoE is limited to a simple qualitative understanding of how individual metrics impact engagement [19]. This leads to severe shortcomings for every component of the video ecosystem. For example, adaptive video players today resort to ad hoc tradeoffs between bitrate, startup delay, and buffering [16, 20, 32]. Similarly, frameworks for multi-CDN optimization use primitive QoE metrics that only capture buffering effects without accounting for the impact of bitrate or bitrate switching [28, 29]. Finally, content providers do not have systematic ways to evaluate the cost-performance tradeoffs that different CDNs or multi-CDN optimizations offer [1].

We observe that there are three key factors that make it challenging to obtain a unified and quantitative understanding of Internet video QoE:

- **Complex relationships:** The relationships between the quality metrics and the effective user experience can be quite complex and even counter-intuitive. For example, while one would naturally expect a higher video bitrate leading to better user experience, we observe a non-monotonic relationship between the two.
- **Metric dependencies:** The metrics themselves have subtle interdependencies and have implicit tradeoffs. For example, although switching bitrates to adapt to the bandwidth conditions can reduce buffering, we observe that high rates of switching can annoy users.
- **Impact of content:** There are many confounding factors introduced by the nature of the content itself. For example, different genres of content such as live and video-on-demand (VOD) show very different viewing patterns. Similarly, users' interest in content also affects their tolerance non-trivially.

Our goal in this paper is to identify a feasible roadmap toward developing a robust, unified and quantitative QoE metric that can address these challenges. We have two intuitive reasons to be hopeful. The challenges raised by complex relationships and subtle interdependencies can be addressed by casting QoE inference as a *machine learning* problem of building an appropriate model that can predict the user engagement (e.g., play time) as a function of the various quality metrics. The second issue of content-induced effects can be addressed using domain-specific and measurement-driven insights to carefully set up the learning tasks.

Our preliminary results give us reason to be optimistic. For example, a decision tree based classifier can provide
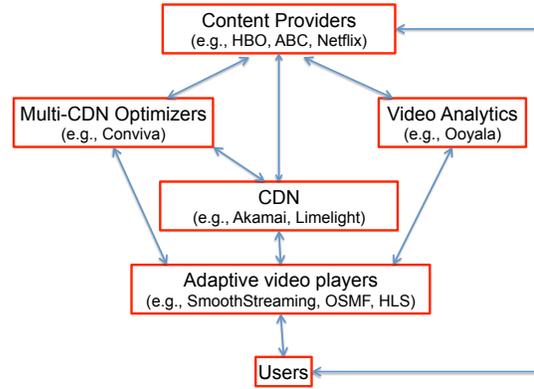

Figure 1: Overview of the Internet video ecosystem; a robust QoE metric is critical for every component in this ecosystem.

close to 50% accuracy in predicting the engagement. Carefully setting up the inputs and features for the learning process could lead to as high as 25% gain in accuracy of the prediction model.

The rest of this paper is organized as follows. Section 2 describes how a standardized QoE metric would impact the different players in the video ecosystem. Section 3 describes the main challenges in developing a QoE metric. Section 4 makes the case for a predictive model for developing a QoE metric. In Section 5, we present some preliminary results before discussing various challenges in Section 6. We conclude in Section 7.

2. USE CASES FOR VIDEO QOE

We begin with a brief overview of the Internet video ecosystem today and argue why there is an immediate need for a standardized QoE metric and how this impacts the different players in the video ecosystem (Figure 1).

- *Content providers* like HBO, ABC, and Netflix would like to maximize their revenues from subscription and ad-based business models while trying to minimize their distribution costs. To this end, content providers have business arrangements with CDNs (e.g., Akamai, Limelight) and also with third-party analytics (e.g., Ooyala [8]) and optimization services (e.g., Conviva [5]). A robust QoE metric enables content providers to objectively evaluate the cost-performance tradeoffs offered by the CDNs and the value that such third-party services offer.
- *Content Distribution Networks* need to allocate their distribution resources (e.g., server and bandwidth capacity) across user population. They need standard metrics to demonstrate superior cost-performance tradeoffs. CDNs also need such metrics to guide the design of their delivery infrastructures to minimize their delivery costs while maximizing their performance [24].
- Recent studies have argued the case for *cross-CDN optimization* [28, 29] and there are already commercial services (e.g., Conviva [5]) that provide these capabilities. These services need standard measures to demonstrate

quantifiable value to the content providers. An open challenge that such optimization frameworks face is the choice of a suitable quality metric that needs to be optimized [29]. Similarly, third-party *video analytics* services need concrete ways to translate their insights with respect to user demographics and user behaviors into quantitative engagement effects.

- *Video player* designers have to make conscious tradeoffs in their bitrate adaptation algorithms. For example, moving to a higher bitrate may offer better engagement but increases the risk of buffering that is known to annoy users. Similarly, user studies suggest that users cannot tolerate too frequent bitrate switches as it impacts their perceptual experience [18]. The lack of a systematic understanding of video QoE forces player designers to use ad hoc adaptation strategies without a clear optimization goal [16, 20, 32].

- Ultimately, the success of this ecosystem depends on the *users*' experience. Increasingly, the same content (TV shows, movies) is available from multiple providers (e.g., Apple iTunes, Amazon, Hulu Plus, Google). Beyond issues of content availability, users would prefer services that give them a better cost-experience tradeoff. Another issue relates to the recent introduction of ISP bandwidth quotas [4]; an understanding of video QoE enables users and delivery providers to better customize the experience under such constraints.

There appears to be rough consensus among the leading industry players on two accounts. First, there is implicit agreement on the set of quality metrics and the measures of engagement [6, 19, 33]. Second, there is also a growing realization in this community of the need for a data-driven approach using measurements in-the-wild as opposed to traditional methods of using controlled user studies. The key challenge that remains is providing a unified and quantitative understanding of the relationship between the quality metrics and the engagement measures. As we will show in the next section, this turns out to be non-trivial.

3. CHALLENGES IN MEASURING QOE

In this section, we use real world measurements of client viewing sessions from two large content providers, one serving TV episodes and the other providing live sports events, to highlight challenges in QoE measurement. We use industry-standard video quality metrics for our study [6]. For concreteness we focus on play time as the measure of user engagement in this section. A subset of the observations we present have also appeared in other measurement studies (e.g., [12, 19, 21–23, 26, 31]), albeit in other contexts. Our specific contribution lies in highlighting the challenges these raise for developing a unified QoE metric.

3.1 Complex relationships

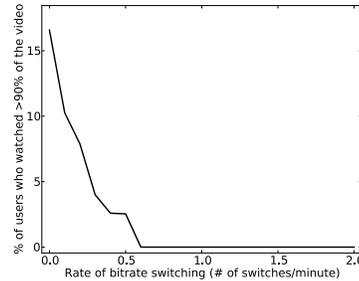The relationship between different quality metrics and user

Figure 2: Switching annoys users after a certain threshold

engagement are complex. These were extensively studied in [19] and we reconfirm some of these observations.

Counter-intuitive effects: Although, intuitively higher rendering quality (frames per second) should lead to higher user engagement, Dobrian et al. noticed several instances where lower rendering quality led to long play times, especially in the case of live videos [19]. This is because of an optimization by the video player to reduce CPU consumption by reducing the frame rate when the video is played in the background. Users may run live videos in the background while focusing on other work, but would be compelled to close the player if the CPU usage was high.

Non-monotone effects: Although higher average bitrate should result in higher user engagement, prior work has observed that there is a non-monotonic relationship between the two [19]. This is because CDNs serve content at specific bitrates and the values of average bitrates in between these standard bitrates correspond to clients that had to switch bitrates during the session. These clients likely experienced higher buffering and hence the lower engagement.

Threshold effects: We also observed that the rate of switching has a threshold effect on user engagement. Rates up to 0.5 switches/minute do not have any effect on the user engagement. However, at higher rates, users seem to quit early as shown in Figure 2. Our observation corroborates the user studies made in [18].

3.2 Interaction between metrics

Naturally, the various quality metrics are not independent of each other. The interaction and implicit tradeoffs between these metrics also needs careful examination. Next, we highlight some of these interesting tradeoffs:

Switching vs buffering: An efficient bitrate switching algorithm should pro-actively switch bitrates to avoid buffering events within a session and also let the user experience the best possible video quality based on the current network conditions at the client. However, as shown in Figure 2, high rates of switching annoys users leading them to quit early.

Join time vs bitrate: Although higher bitrate would imply higher quality, it would also imply higher join time since it would take longer time for the player video buffer to fill up sufficiently to start rendering video frames.
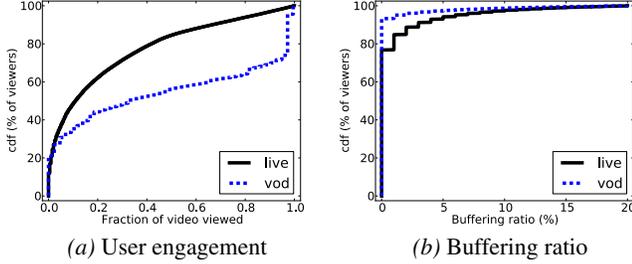
(a) User engagement   *(b)* Buffering ratio

Figure 3: User viewing pattern of live and VOD videos

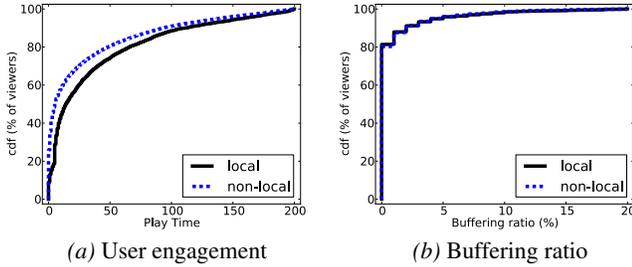*(a)* User engagement   *(b)* Buffering ratio

Figure 4: User interest induced by regional games

3.3 Externalities

There are many confounding external factors that affect user engagement and are not captured by the quality metrics. For instance, user-attributes like bandwidth at the client and its variability, and content attributes like genre, popularity and age of the content have effects on user engagement.

Genre of the content: We observed that live and VOD video sessions experience similar quality—e.g., Figure 3b shows the cdf of buffering ratio (fraction of session time spent in buffering) for live and VOD sessions. However, as shown in Figure 3a, the distribution of the fraction of video viewed by users over several popular video objects is considerably different. This shows that the viewing pattern of live and VOD sessions are very different.

User interest: As observed in Figure 3a, several users tend to sample videos resulting in many of them quitting early based on interest despite the lack of any quality issues during the session. Another example of how user interest affects user engagement is in the case of regional sporting events. Figure 4a shows the distribution of play time of users watching a regional sporting event segregated into whether the viewer was from the region or not. On average, local users watch 10 minutes more of the video than non-local users despite the fact that the quality of the video was the same (e.g., Figure 4b shows the cdf of the buffering ratio in both cases).

4. TOWARD A PREDICTIVE MODEL FOR QOE INFERENCE

At a high-level, our goal is to express engagement as a function of the quality metrics, given other factors such as
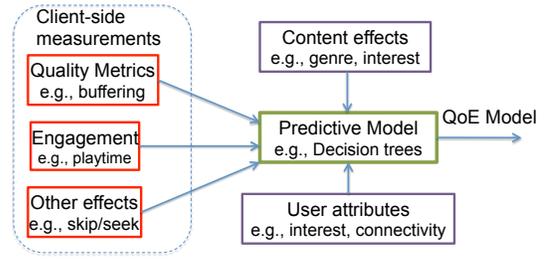

Figure 5: High-level overview of our proposed approach for deriving a QoE metric via a predictive model

content and user attributes. We want to capture a relationship $Engagement = f(\{QualityMetric_i\})$, where $Engagement$ can be the video playtime, number of visits to a website, and each $QualityMetric_i$ represents observed indices such as buffering ratio, average bitrate. Assuming that this function f is accurate, we can simply use this as our QoE metric to compare two competing solutions. For e.g., a content provider can use observed values of the quality indices for different CDNs or multi-CDN optimizers, and compute the QoE they offer.

The foremost question here is whether we can build an *accurate* model. As we saw in the previous section, there are three high-level challenges in going from low-level quality indices (e.g., buffering, bitrate, bitrate switching, startup latency) to a quantitative and unified QoE metric—there are complex relationships between these metrics and the engagement; there are hidden dependencies among the metrics themselves (e.g, bitrate vs. startup latency); and there confounding effects of user interest, content genre, and user connectivity. We believe that these issues are fundamental and that any video QoE metric must address these challenges.

We see two intuitive reasons to be hopeful about the feasibility of building such a predictive model:

- First, the issue of dependencies and hidden relationships can be handled by choosing a suitable *machine learning* approach that is expressive enough to capture complex dependencies. Beyond choosing a robust learning approach, the key to success here is the availability of sufficiently large datasets. Fortunately, many content providers, CDNs, and third-party optimizers and analytic services already have the requisite instrumentation in client- and server-side infrastructure to collect such measurements [6, 19].

- Second, the issue of confounding effects can be tackled by carefully incorporating these domain-specific insights into the learning process. This can be achieved in one of two ways: (1) carefully set up the input data for the machine learning task and (2) explicitly identify such confounding features and let the learning algorithm automatically account for these factors.

Figure 5 presents a high-level overview of our vision for building a unified and quantitative QoE metric. We use measurements of quality and engagement metrics and optionally
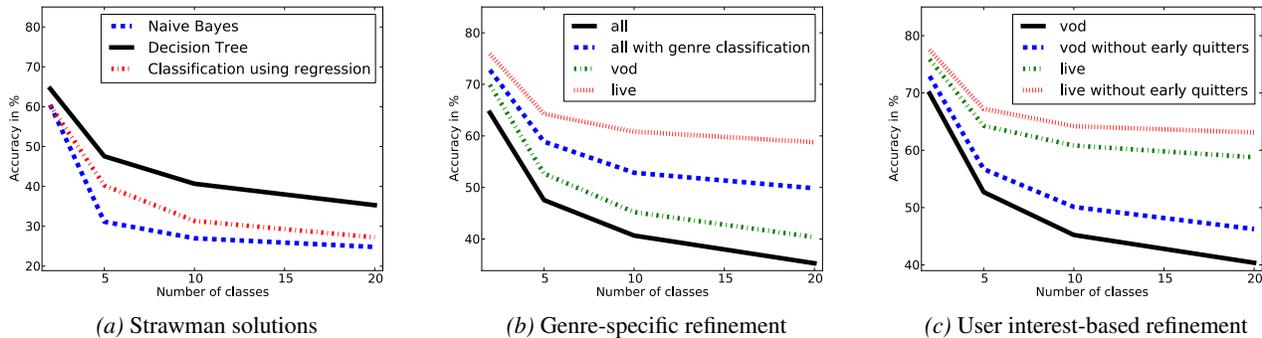
4

(a) Strawman solutions   *(b) Genre-specific refinement*   *(c) User interest-based refinement*

Figure 6: Evaluating different models that predict user engagement from quality metrics

take into account user actions during video sessions to build a basic prediction model. We further refine the prediction using domain-specific knowledge of content-induced and user-induced effects.

5. PRELIMINARY RESULTS

In this section, we show preliminary results to confirm the intuition outlined in the previous section. Our results are based on 10-fold cross-validation [30] on around 10 million video sessions spanning a month of viewership.

Strawman solutions: As mentioned in Section 4, we set up a learning problem where the objective is to predict the user engagement given the various quality metrics. We use play time to measure user engagement. We categorize the play time into classes based on the fraction of the video that the user viewed. For example, when the number of classes is set to 5, the model tries to predict if the user viewed 0-20% or 20-40% or 40-60% or 60-80% or 80-100% of the video. By varying the number of classes, we change the granularity at which the model predicts user engagement. The quality metrics are also appropriately classified into fine-grained classes.

The choice of the learning algorithm is crucial as it should be adept at learning the various complex relationships and interdependencies between the different quality metrics. We employed learning algorithms such as naive Bayes, simple regression and classical binary decision tree from the machine learning literature [30] to learn a predictive model.

As shown in Figure 6a, decision trees perform better than naive Bayes and regression-based techniques. Naive Bayes is a probablistic classifier based on Bayes theorem [30] and has strong independence assumption. However, the quality metrics are intrinsically related to each other as shown in Section 3. Similarly, simple regression based techniques cannot capture the complex relationships (eg. non-monotonic effects) between the quality metrics and user engagement. As expected, all the learning algorithms have lower accuracy with increasing number of classes since it would require the model to predict at a finer granularity.

Domain-specific refinements: While decision trees are expressive enough to capture the various complex relationships, the confounding effects caused by various externalities are not captured in the previous model. However measurement-driven insights can be used to perform certain domain-specific refinements such as:

(1) *Genre-specific refinement*: Based on the observation that the viewing pattern of live and VOD video are different (Figure 3a), we segment the data into two parts (live sessions and VOD sessions) and run the decision tree algorithm separately on them. As observed in Figure 3, this leads to about 20% increase in accuracy.

(2) *User interest-based refinement*: Based on the observation that users tend to "sample" the video, we ignore these early quitters (users with play time < 5 minutes) from our dataset and relearn the model. This leads to about 5% further improvement in the accuracy of the model (Figure 6c).

These results lead us to believe that the roadmap we envisioned in Section 4—build a basic predictive model based on machine learning techniques and then employ domain-specific insights to carefully set up the inputs—is a promising starting point that can ultimately lead to the development of a robust Internet video QoE metric.

6. DISCUSSION

While our preliminary results from the previous section are promising, there are several open challenges that need to be addressed.

Other engagement metrics: As a concrete starting point, we used the play time as the measure of engagement. Content providers care about other measures of engagement as well; e.g., actual ad impressions, user loyalty in terms of likelihood of return, total number of videos viewed. We believe that the high-level framework we envisioned can be applied to these metrics as well. An interesting observation from past work is that different quality metrics may impact engagement metrics differently [19]. For example, startup delay may not affect a specific viewing session, but may hurt the likelihood of the customer returning to the same provider.

Thus, we may need to weigh in the different engagement measures when computing an aggregate QoE metric.

Coverage over externalities: External effects raise two obvious questions:

(1) Have we identified for all possible externalities? For example, a user's ISP or the device from which she is viewing video may also have a substantial impact on the relationship between quality and engagement.

(2) For a specific effect, have we adequately accounted for its impact?

Both problems are challenging and we believe that this motivates the need for more measurement studies in this space to further refine our understanding. That said, we have reason to believe that machine learning approaches can help us on both accounts. First, in terms of identifying externalities, one could explicitly enumerate a large set of user and content attributes and use *feature selection* techniques to identify the external (i.e., non quality) factors that have the greatest impact on engagement. Second, in the case of effects induced by user behavior, we can learn *user profiles* of interest or past viewing patterns (e.g., is this an early quitter) to guide the modeling process.

Intuitive models: Inferring the underlying cause for lower user engagement given the quality metrics is a tough problem because of various confounding factors like user interest and the user's tolerance to the video quality. However, in order to achieve widespread adoption, any unified metric must be intuitive so that system designers and practitioners can actually make sense of the tradeoffs that the QoE model predicts. This can be tackled by identifying and including the various confouding factors into the model. Another concern with applying black-box machine learning models is that the outputs may be too complex; e.g., models like PCA are particularly notorious for producing unintuitive outputs. Our choice of decision trees here as a starting point turns out to be beneficial. There are known techniques to convert the outputs to more intuitive explanations (e.g., [27]) and to generate more closed-form equations that capture quantitative tradeoffs (e.g., [25]).

Validation: The last question is how can we ultimately validate that the QoE metric is useful to the different participants in the video ecosystem from Section 2. One option is to use A/B testing where a subset of users are given the optimized "treatment" using our new QoE metric to drive decisions on how to allocate server resources [29]. We can subsequently examine if the optimized users have higher engagement relative to the control group that received no optimization.

7. CONCLUSIONS

Many industries suffer an unfortunate fate where the lack of understanding of what metrics really matter leads to deceptive marketing—vendors will inevitably quote individual metrics that make them look good, without attempting to explain how this fits in the grander scheme of things, and how it actually impacts the end user experience (e.g., clock

frequency in computer architecture or megapixels in digital cameras [7]). With the proliferation of quality factors, Internet video stands to suffer a similar fate. Our overarching goal is to avoid this and spur further innovation by developing a robust, unified, and quantitative QoE metric. As we showed, there are several challenges in this regard and it is likely that there are others that we have overlooked in our initial exploration. Our preliminary results give us reason to be hopeful that we can systematically address these challenges in ongoing and future work.

8. REFERENCES

[1] Buyer's Guide: Content Delivery Networks. http://goo.gl/B6gMK.
[2] CBS Scores 37 Million In Online Ad Sales For March Madness. http://goo.gl/C3bSV.
[3] Cisco forecast. http://goo.gl/7rilD.
[4] Comcast to cap data transfers at 250 gb in oct. http://goo.gl/X5A11.
[5] Conviva. http://www.conviva.com.
[6] Driving Engagement for Online Video. http://goo.gl/pO5Cj.
[7] Figure of merit. http://goo.gl/pNdwn.
[8] Ooyala. http://www.ooyala.com/.
[9] P.800 : Methods for subjective determination of transmission quality. http://www.itu.int/rec/T-REC-P.800-199608-I/en.
[10] P.910 : Subjective video quality assessment methods for multimedia applications. http://goo.gl/QjFhZ.
[11] SPEC philosophy. http://www.spec.org/spec/#philosophy.
[12] Video quality metrics. http://goo.gl/Ga9Xz.
[13] I. Sodagar. The MPEG-DASH Standard for Multimedia Streaming Over the Internet. *IEEE Multimedia*, 2011.
[14] K. Chen, C. Huang, P. Huang, C. Lei. Quantifying Skype User Satisfaction. In *Proc. SIGCOMM*, 2006.
[15] A. Bouch, A. Kuchinsky, and N. Bhatti. Quality is in the Eye of the Beholder: Meeting Users' Requirements for Internet Quality of Service. In *Proc. CHI*, 2000.
[16] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen. What Happens when HTTP Adaptive Streaming Players Compete for Bandwidth? In *Proc. NOSSDAV*, 2012.
[17] R. H. Allen and R. D. Sriram. The Role of Standards in Innovation. *Elsevier: Technology Forecasting and Social Change*, 2000.
[18] N. Cranley, P. Perry, and L. Murphy. User perception of adapting video quality. *International Journal of Human-Computer Studies*, 2006.
[19] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. A. Joseph, A. Ganjam, J. Zhan, and H. Zhang. Understanding the impact of video quality on user engagement. In *Proc. SIGCOMM*, 2011.
[20] J. Esteban, S. Benno, A. Beck, Y. Guo, V. Hilt, and I. Rimac. Interactions Between HTTP Adaptive Streaming and TCP. In *Proc. NOSSDAV*, 2012.
[21] H. Y. et al. Inside the Bird's Nest: Measurements of Large-Scale Live VoD from the 2008 Olympics. In *Proc. IMC*, 2009.
[22] A. Finamore, M. Mellia, M. Munafo, R. Torres, and S. G. Rao. Youtube everywhere: Impact of device and infrastructure synergies on user experience. In *Proc. IMC*, 2011.
[23] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding User Behavior in Large-Scale Video-on-Demand Systems. In *Proc. Eurosys*, 2006.
[24] L. Kontothanassis, R. Sitaraman, J. Wein, D. Hong, R. Kleinberg, B. Mancuso, D. Shaw, and D. Stodolsky. A Transport Layer for Live Streaming in a Content Delivery Network. *Proc. IEEE*, 92, 2004.
[25] L. Huang, J. Jia, B. Yu, B.-G. Chun, P. Maniatis, and M. Naik. Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression. In *Proc. NIPS*, 2010.
[26] E. B. L. Plissonneau. A Longitudinal View of HTTP Video Streaming Performance. In *MMSys*, 2012.
[27] B. Liu, M. Hu, and W. Hsu. Intuitive Representation of Decision Trees Using General Rules and Exceptions. In *Proc. AAAI*, 2000.
[28] H. Liu, Y. Wang, Y. R. Yang, A. Tian, and H. Wang. Optimizing Cost and Performance for Content Multihoming. In *Proc. SIGCOMM*, 2012.
[29] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang. A Case for a Coordinated Internet Video Control Plane. In *Proc. SIGCOMM*, 2012.
[30] T. Mitchell. *Machine Learning*. McGraw-Hill.
[31] R. K. P. Mok, E. W. W. Chan, X. Luo, and R. K. C. Chang. Inferring the QoE of HTTP Video Streaming from User-Viewing Activities . In *SIGCOMM W-MUST*, 2011.
[32] S. Akhshabi, A. Begen, C. Dovrolis. An Experimental Evaluation of Rate Adaptation Algorithms in Adaptive Streaming over HTTP. In *Proc. MMSys*, 2011.
[33] M. Watson. Http adaptive streaming in practice. http://goo.gl/yXjjt.