Developing a Predictive Model of Quality of Experience for Internet Video

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

Improving users' quality of experience (QoE) is crucial for sustaining the advertisement and subscription based revenue models that enable the growth of Internet video. Despite the rich literature on video and QoE measurement, our understanding of Internet video QoE is limited because of the shift from traditional methods of measuring video quality (e.g., Peak Signal-to-Noise Ratio) and user experience (e.g., opinion scores). These have been replaced by new quality metrics (e.g., rate of buffering, bitrate) and new engagement-centric measures of user experience (e.g., viewing time and number of visits). The goal of this paper is to develop a predictive model of Internet video QoE. To this end, we identify two key requirements for the QoE model: (1) it has to be tied in to observable user engagement and (2) it should be actionable to guide practical system design decisions. Achieving this goal is challenging because the quality metrics are interdependent, they have complex and counter-intuitive relationships to engagement measures, and there are many external factors that confound the relationship between quality and engagement (e.g., type of video, user connectivity). To address these challenges, we present a data-driven approach to model the metric interdependencies and their complex relationships to engagement, and propose a systematic framework to identify and account for the confounding factors. We show that a delivery infrastructure that uses our proposed model to choose CDN and bitrates can achieve more than 20% improvement in overall user engagement compared to strawman approaches.

Categories and Subject Descriptors

C.4 [**Performance and Systems**]: measurement techniques, performance attributes; C.2.4 [**Computer-Communication Networks**]: Distributed Systems—Client/server

Keywords

Video quality, Measurement, Performance, Human Factors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGCOMM'13, August 12–16, 2013, Hong Kong, China.
Copyright 2013 ACM 978-1-4503-2056-6/13/08 ...$15.00.

1. INTRODUCTION

The growth of Internet video has been driven by the confluence of low content delivery costs and the success of subscription- and advertisement-based revenue models [2]. Given this context, there is agreement among leading industry and academic initiatives that *improving users' quality of experience* (QoE) is crucial to sustain these revenue models, especially as user expectations of video quality are steadily rising [3, 20, 26].

Despite this broad consensus, our understanding of Internet video QoE is limited. This may surprise some, especially since QoE has a very rich history in the multimedia community [5, 6, 11]. The reason is that Internet video introduces new effects with respect to both *quality* and *experience*. First, traditional quality indices (e.g., Peak Signal-to-Noise Ratio (PSNR) [7]) are now replaced by metrics that capture delivery-related effects such as rate of buffering, bitrate delivered, bitrate switching, and join time [3, 15, 20, 28, 36]. Second, traditional methods of quantifying experience through user opinion scores are replaced by new *measurable engagement measures* such as viewing time and number of visits that more directly impact content providers' business objectives [3, 36].

The goal of this paper is to develop a *predictive model* of user QoE in viewing Internet video. To this end, we identify two key requirements that any such model should satisfy. First, we want an **engagement-centric** model that accurately predicts user engagement in the wild (e.g., measured in terms of video play time, number of user visits). Second, the model should be **actionable** and useful to guide the design of video delivery mechanisms; e.g., adaptive video player designers can use this model to tradeoff bitrate, join time, and buffering [12, 13, 21] and content providers can use it to evaluate cost-performance tradeoffs of different CDNs and bitrates [1, 28].

Meeting these requirements, however, is challenging because of three key factors (Section 2):

- **Complex relationship between quality and engagement:** Prior measurement studies have shown complex and counter-intuitive effects in the relationship between quality metrics and engagement. For instance, one might assume that increasing bitrate should increase engagement. However, the relationship between bitrate and engagement is strangely non-monotonic [20].
- **Dependencies between quality metrics:** The metrics have subtle interdependencies and have implicit tradeoffs. For example, bitrate switching can reduce buffering. Similarly, aggressively choosing a high bitrate can increase join time and also cause more buffering.
- **Confounding factors:** There are several potential confounding factors that impact the relationship between quality and engagement: the nature of the content (e.g., live vs. Video on Demand

	Engagement-centric	Actionable
PSNR-like (e.g., [17])	✗	✓
Opinion Scores(e.g., [6])	✓	✗
Network-level (e.g., bandwidth, latency [35])	✗	✓
Single metric (e.g., bitrate, buffering)	✗	✓
Naive learning	✗	✗
Our approach	✓	✓

Table 1: A summary of prior models for video QoE and how they fall short of our requirements

(VOD), popularity), temporal effects (e.g., prime time vs. off-peak), and user-specific attributes (e.g., connectivity, device, user interest) [26].

As Table 1 shows, past approaches fail on one or more of these requirements. For instance, user opinion scores may be reflective of actual engagement, but these metrics may not be actionable because these do not directly relate to system design decisions. On the other hand, network- and encoding-related metrics are actionable but do not directly reflect the actual user engagement. Similarly, one may choose a single quality metric like buffering or bitrate, but this ignores the complex metric interdependencies and relationships of other metrics to engagement. Finally, none of the past approaches take into account the wide range of confounding factors that impact user engagement in the wild.

In order to tackle these challenges, we present a data-driven approach to develop a robust model to predict user engagement. We leverage large-scale measurements of user engagement and video session quality to run machine learning algorithms to automatically capture the complex relationships and dependencies [23]. A direct application of machine learning, however, may result in models that are not intuitive or actionable, especially because of the confounding factors. To this end, we develop a systematic framework to identify and account for these confounding factors.

Our main observations are:

- Compared to machine learning algorithms like naive Bayes and simple regression, a decision tree is sufficiently expressive enough to capture the complex relationships and interdependencies and provides close to 45% accuracy in predicting engagement. Furthermore, decision trees provide an intuitive understanding into these relationships and dependencies.
- Type of video (live vs. VOD) , device (PC vs. mobile devices vs. TV) and connectivity (cable/DSL vs. wireless) are the three most important confounding factors that affect engagement. In fact, the QoE model is considerably different across different types of videos.
- Refining the decision tree model that we developed by incorporating these confounding factors can further improve the accuracy to as much as 70%.
- Using a QoE-aware delivery infrastructure that uses our proposed model to choose CDN and bitrates can lead to more than 20% improvement in overall user engagement compared to other approaches for optimizing video delivery.

Contributions and Roadmap: To summarize, our key contributions are

- Systematically highlighting challenges in obtaining a robust video QoE model (Section 2);

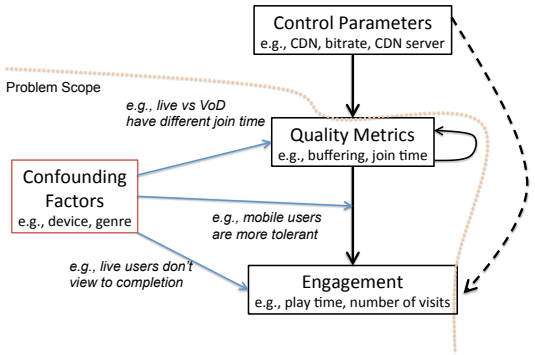


Figure 1: Our goal is to develop a predictive model of engagement that accounts for the complex relationship between quality and engagement, the interdependencies between quality metrics, and the confounding factors that impact different aspects of this learning. We do not attempt to model the impact of the control parameters on quality or engagement in this paper.

- A roadmap for developing Internet video QoE that leverages machine learning (Section 3);
- A methodology for identifying and addressing the confounding factors that affect engagement (Section 4 and Section 5); and
- A practical demonstration of the utility of our QoE models to improve engagement (Section 6)

We discuss outstanding issues in Section 7 and related work in Section 8 before concluding in Section 9.

2. MOTIVATION AND CHALLENGES

In this section, we provide a brief background of the problem space and highlight the key challenges in developing a unified QoE model using data-driven techniques.

2.1 Problem scope

Multiple measurement studies have shown that video quality impacts user engagement [20, 26]. Given that engagement directly affects advertisement- and subscription-based revenue streams, there is broad consensus across the different players in the Internet video ecosystem (content providers, video player designers, third-party optimizers, CDNs) on the need to optimize video quality according to these metrics. In this paper, we focus on the fraction of video that the user viewed before quitting as the measure of engagement and the following industry-standard quality metrics:

- *Average bitrate*: Video players typically switch between different bitrate streams during a single video session. Average bitrate, measured in kilobits per second, is the time average of the bitrates played during a session weighted by the time duration each bitrate was played.
- *Join time*: This represents the time it takes for the video to start playing after the user initiates a request to play the video and is measured in seconds.
- *Buffering ratio*: It is computed as the ratio of the time the video player spends buffering to the sum of the buffering and play time and is measured in percentage.
- *Rate of buffering*: It captures the frequency at which buffering events occur during the session and is computed as the ratio of the number of buffering events to the duration of the session.

In this paper, we are not reporting the impact of rate of bitrate switching due to two reasons. First, we were unable to collect this data for the large time frames that we are working on. Second, the providers use specific bitrate adaptation algorithms and we want to

avoid reaching conclusions based on them. That said, the framework and the techniques that we propose are applicable if we want to later include more quality metrics to the above list.

Our goal in this paper is to develop a principled *predictive model* of the user engagement as a function of these quality metrics. Such a model can then be used to inform the system design decisions that different entities in the video ecosystem make in order to tune the video according to the quality metrics to maximize engagement. For example, a video control plane may choose the CDN and bitrate for a video session based on a global optimization scheme [28], CDNs choose the servers and the specific bitrates [3], video player designers can adaptively switch bitrates [13] and so on.

An alternative direction one may consider is building a predictive model that relates the control parameters used (e.g., CDN, CDN server, adaptive bitrate selection logic) to the user engagement. While this is certainly a viable approach that a specific entity in the video delivery chain (e.g., CDN, player, "meta-CDN")[1] can employ, we believe that we can decouple this into two problems: (a) learning the relationship between the quality metrics and engagement (quality → engagement) and (b) learning the relationship between the control parameters and engagement (knobs→ quality). We do so because these two relationships evolve at fundamentally different timescales and depend on diverse factors (e.g., user behavior vs. network/server artifacts). First, the control parameters available to different entities in the ecosystem may be very different; e.g., the control plane [28] operates at a coarse granularity of choosing the CDN whereas the CDN can choose a specific server. Second, the control knobs for each entity may themselves change over time; e.g., new layered codecs or more fine-grained bitrates. One can view this as a natural layering argument—decoupling the two problems allows control logics to evolve independently and helps us reuse a reference QoE model across different contexts (e.g., control plane, CDN, video player).

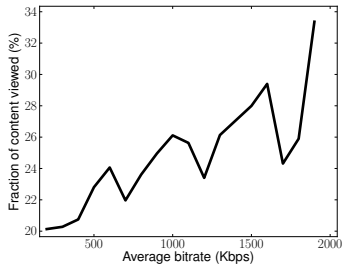
While modeling the knobs→quality problem is itself an interesting research challenge, this is outside the scope of this paper; the focus of this paper is on the problem of modeling the quality → engagement relationship.[2] As we discuss next, there are three key challenges in addressing this problem.

2.2 Dataset

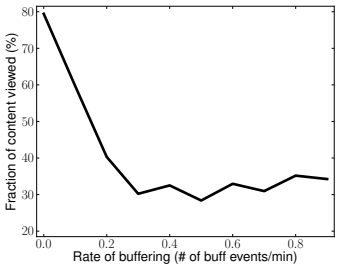
The data used in this paper was collected by `conviva.com` in real time using a client-side instrumentation library. This library gets loaded when users watch video on `conviva.com`'s affiliate content providers' websites. We collect all the quality metrics described earlier as well as play time for each individual session. In addition we also collect a range of user-specific (e.g., location, device, connectivity), content (e.g., live vs. VOD, popularity), and temporal attributes (e.g., hour of the day).

The dataset that is used for various analysis in this paper is based on 40 million video viewing sessions collected over 3 months spanning two popular video content providers (based in the US). The first provider serves mostly VOD content that are between 35 minutes and 60 minutes long. The second provider serves sports events that are broadcast while the event is happening. Our study is limited to clients in the United States.[3]

[1]Some services offer CDN switching services across multiple physical CDNs to offer better cost-performance tradeoffs (e.g., [9]).
[2]To evaluate the potential improvement due to our approach, however, we need to model this relationship as well. We use a simple quality prediction model in Section 6.
[3]These are distinct providers and there is no content overlap; i.e., none of the VOD videos is a replay of a live event.



(a) Non-monotonic relationship between average bitrate and engagement



(b) Threshold effect between rate of buffering and engagement

Figure 2: Complex relationship between quality metrics and engagement

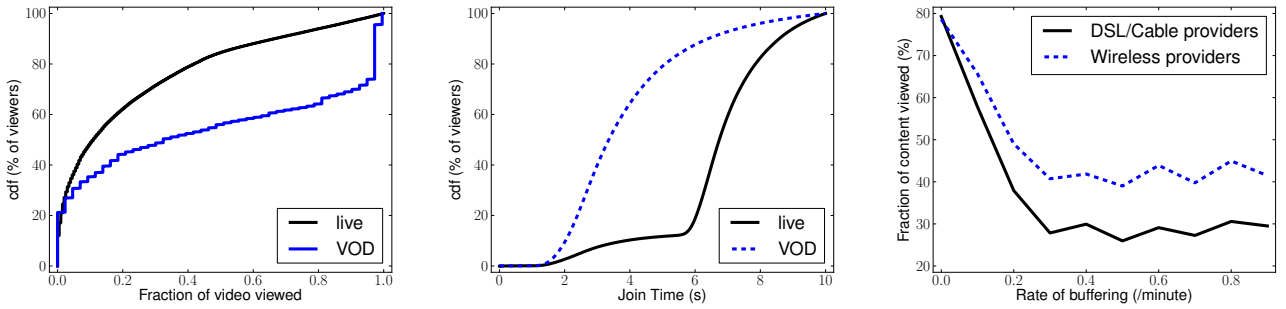
2.3 Challenges in developing video QoE

We use our dataset to highlight the main challenges in developing an engagement-centric of model for video QoE. Although past measurement studies (e.g., [10, 15, 20, 22, 32, 33, 40, 41]) have also presented some of these observations in a different context, our contribution lies in synthesizing these observations in the context of Internet video QoE.

Complex relationships: The relationships between different individual quality metrics and user engagement are very complex. These were shown by Dobrian et al., and we reconfirm some of their observations [20]. For example, one might assume that higher bitrate should result in higher user engagement. Surprisingly, there is a non-monotonic relationship between them as shown in Figure 2a. The reason is that videos are served at specific bitrates and hence 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, which led to a drop in engagement. Similarly, engagement linearly decreases with increasing rate of buffering up to a certain threshold (0.3 buffering events/minute). Beyond this, users get annoyed and they quit early as shown in Figure 2b.

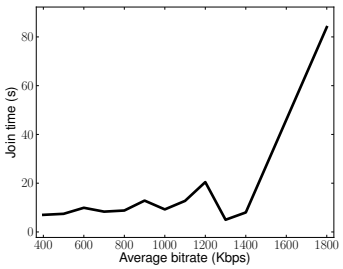
Interaction between metrics: Naturally, the various quality metrics are interdependent on each other. For example, streaming video at a higher bitrate would lead to better quality. However, as shown in Figure 3a, it would take longer for the video player buffer to sufficiently fill up in order to start playback leading to higher join times. Similarly, streaming video at higher bitrates leads to higher rates of buffering as shown in Figure 3b.

Confounding factors: In addition to the quality metrics, several external factors also directly or indirectly affect user engagement [26]. For instance, user-attributes like user interest, content attributes like genre and temporal attributes like age of the content have effects on user engagement. A confounding factor could affect engagement and quality metrics in the following three ways

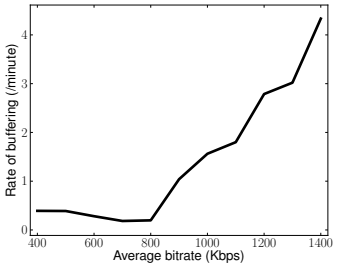


(a) User viewing pattern for live and VOD *(b)* Join time distribution for live and VOD *(c)* Tolerence to rate of buffering on different connectivity

Figure 4: Various confounding factors directly or indirectly affect engagement



(a) Higher bitrates cause higher join times



(b) Higher bitrates cause higher rates of buffering

Figure 3: The quality metrics are interdependent on each other

(Figure 1). First, some factors may affect user viewing behavior itself and result in different observed engagements. For instance, Figure 4a shows that live and VOD video sessions have significantly different viewing patterns. While a significant fraction of the users view VOD videos to completion, live sessions are more short-lived. Second, the confounding factor can impact the quality metric. As Figure 4b shows, the join time distribution for live and VOD sessions are considerably different. Finally, and perhaps most importantly, the confounding factor can affect the relationship between the quality metrics and engagement. For example, we see in Figure 4c that users on wireless connectivity are more tolerant to rate of buffering compared to users on DSL/cable connectivity.

3. APPROACH OVERVIEW

At a high-level, our goal is to express *user engagement* as a function of the quality metrics. That is, 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 metrics such as buffering ratio, average bitrate etc. Ideally, we want this function f to be *accurate*, *intuitive*, and *actionable* in order to be adopted by content providers, video player designers, CDNs, and third-party optimization services to evaluate different provisioning and resource management tradeoffs (e.g., choosing different CDNs and bitrates).

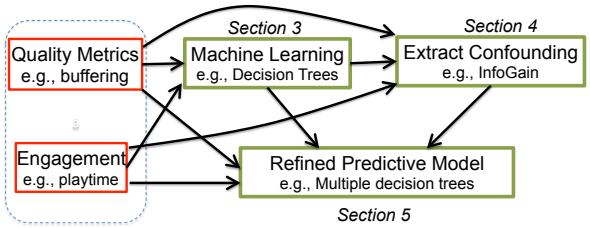


Figure 5: High level overview of our approach. We begin by using standard machine learning approaches to build a basic predictive model and also to extract the key confounding factors. Having identified the confounding factors, we refine our predictive model to improve the accuracy.

As we saw in the motivating measurements in the previous section, developing such a model is challenging because of the complex relationships between the quality metrics and engagement, interdependencies between different quality metrics, and the presence of various confounding factors that affect the relationship between the quality metrics and engagement. In this section, we begin by presenting a high-level methodology for systematically tackling these challenges. While the specific quality and engagement metrics of interest may change over time and the output of the prediction model may evolve as the video delivery infrastructure evolves, we believe that the data-driven and machine learning based roadmap and techniques we envision will continue to apply.

3.1 Roadmap

Figure 5 provides a high-level overview showing three main components in our approach. A key enabler for the viability of our approach is that several content providers, CDNs and third-party optimization services today collect data regarding individual video sessions using client-side instrumentation on many popular video sites. This enables a *data-driven machine learning approach* to tackle the above challenges.

Tackling complex relationships and interdependencies: We need to be careful in using machine learning as a black-box on two accounts. First, the learning algorithms must be expressive enough to tackle our challenges. For instance, naive approaches that assume that the quality metrics are independent variables or simple regression techniques that implicitly assume that the relationships be-

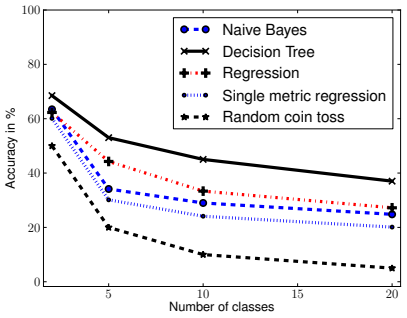


Figure 6: Decision tree is more expressible than naive Bayes and regression based schemes

tween quality and engagement are linear are unlikely to work. Second, we do not want an overly complex machine learning algorithm that becomes unintuitive or unusable for practitioners. Fortunately, as we discuss in Section 3.2 we find that decision trees, which are generally perceived as usable intuitive models [24, 27, 37, 39][4] are also the most accurate. In some sense, we are exploiting the observation that given large datasets, simple non-parametric machine learning algorithms (e.g., decision trees) actually work [23].

Identifying the important confounding factors: Even though past studies have shown that external factors such as users' device and connectivity affect engagement [26], there is no systematic method to identify these factors. In Section 4, we propose a taxonomy for classifying potentially confounding factors. As we saw in the previous section, the confounding factors can affect our understanding in all three respects: affecting quality, affecting engagement, and also affecting how quality impacts engagement. Thus, we need a systematic methodology to identify the factors that have an impact on all three dimensions.

Refinement to account for confounding factors: As we will show in Section 3.2, decision trees are expressive enough to capture the relationships between quality metrics and engagement. It may not, however, be expressive enough to capture the impact of all the confounding factors on engagement. In Section 5, we evaluate different ways by which we can incorporate these confounding factors to form a unified model.

3.2 Machine learning building blocks

Decision trees as predictive models: We cast the problem of modeling the relationship between the different quality metrics as a discrete classification problem. We begin by categorizing engagement into different classes based on the fraction of video that the user viewed before quitting. 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 before quitting. We can select the granularity at which the model predicts engagement by appropriately setting the number of classes (e.g., 5 classes means 20% bins vs. 20 classes means 5% bins). We use similar domain-specific discrete classes to bin the different quality metrics. For join time, we use bins of 1 second interval; for buffering ratio we use 1% bins; for rate of buffering we use 0.1/minute bins; and for average bitrate we use 100 kbps-sized bins.

Figure 6 compares the performance of three different machine learning algorithms: binary decision trees, naive Bayes, and classification based on linear regression. The results are based on 10-fold cross-validation—the data is divided into 10 equally sized subsets and the model is trained 10 times, leaving out one of the subsets each time from training and tested on the omitted subset [31]. Naturally, the prediction accuracy decreases when the model has to predict at a higher granularity. We observe that decision trees perform better than naive Bayes and linear regression. This is because naive Bayes algorithm assumes that the quality metrics are independent of each other and hence it does not attempt to capture interdependencies between them. Similarly, linear regression is not expressible enough to capture the complex relationships between quality metrics and engagement. Also, as shown in Figure 6, performing linear regression based on just a single "best" metric (average bitrate) yields even lower accuracy since we are ignoring the complex metric interdependencies and the relationships between other metrics and engagement.

Information gain analysis: *Information gain* is a standard approach for uncovering hidden relationships between variables. More importantly, it does so without making any assumption about the nature of these relationships (e.g., monotone, linear effects); it merely identifies that there is some potential relationship. Information gain is a standard technique used in machine learning for feature extraction—i.e., identifying the key features that we need to use in a prediction task. Thus, it is a natural starting point for systematically identifying confounding factors.

The information gain is based on the idea of entropy of a random variable Y which is defined as $H(Y) = \sum_i P[Y = y_i] log \frac{1}{P[Y=y_i]}$ where $P[Y = y_i]$ is the probability that $Y = y_i$. It represents the number of bits that would have to be transmitted to identify Y from n equally likely possibilities. The lesser the entropy the more uniform the distribution is. The conditional entropy of Y given another random variable X is $H(Y|X) = \sum_j P[X = x_j]H(Y|X = x_j)$. It represents the number of bits that would be required to be transmitted to identify Y given that both the sender and the receiver know the corresponding value of X. Information gain is defined as $H(Y) - H(Y|X)$ and it is the number of bits saved on average when we transmit Y and both sender and receiver know X. The relative information gain can then be defined as $\frac{H(Y)-H(Y|X)}{H(Y)}$.

In the next section, we use the information gain analysis to reason if a confounding factor impacts either engagement or quality.

Compacted decision trees: Decision trees help in categorizing and generalizing the data given in the dataset and provide a visual model representing various if-then rules. One main drawback while dealing with multi-dimensional large datasets is that these techniques produce too many rules making it difficult to understand and use the discovered rules with just manual inspection or other analysis techniques [27]. In order to get a high-level intuitive understanding of the impact of different quality metrics on engagement, we compact the decision tree. First, we group the quality metrics into more coarse-grained bins. For instance, we classify average bitrate into three classes—very low, low, and high. The other quality metrics (buffering ratio, buffering rate, and join time) and engagement are classified as either high or low. We then run the decision tree algorithm and compact the resulting structure to form general rules using the technique described in [27]. The high-level idea is to prune the nodes whose majority classes are significant; e.g., if more than 75% of the data points that follow a particular rule belong to a particular engagement class then we prune the tree at that level. The tree formed using this technique may not be highly accurate. Note that the goal of compacting the decision tree is only to get a high-level understanding of what quality metrics affect engagement the most and form simple rules of how they impact engagement. Our predictive model uses the original (i.e.,

[4]Decision trees can be directly mapped into event processing rules that system designers are typically familiar with [39].

uncompressed) decision tree; we do not sacrifice any expressive power. In the next section, we use this technique to test if a confounding factor impacts the relationship between quality metrics and engagement—particularly to check if it changes the relative importance of the quality metrics.

3.3 Limitations

We acknowledge three potential limitations in our study that could apply more broadly to video QoE measurement.

- **Fraction of video viewed as a metric for engagement:** While fraction of video viewed before quitting may translate into revenue associated from actual advertisement impressions, it does not capture various psychological factors that affect engagement (e.g., user may not be interested in the video and might be playing the video in the background). We use fraction of video viewed as a measure of engagement since it can be easily and objectively measured and it provides a concrete starting point. The high-level framework that we propose can be applied to other notions of engagement.
- **Coverage over confounding factors:** There might be several confounding factors that affect engagement that are not captured in our dataset (e.g., user interest in the content). Our model provides the baseline in terms of accuracy—uncovering other confounding factors and incorporating them into the model will lead to better models and higher prediction accuracy.
- **Early quitters:** A large fraction of users quit the session after watching the video for a short duration. These users might be either "sampling" the video or quitting the session because of quality issues. They can be treated in three ways: (1) Remove them completely from the analysis, (2) Separate them into two groups based on their quality metrics (high quality population and low quality population) and learn a separate model for each group (3) Profile users based on their viewing history and predict whether they will quit early or not based on their interest in the video content. We use (1) in this paper as it provides a clearer understanding of how quality impacts engagement. That said, approaches (2) and (3) are likely to be useful and complementary in a system-design context; e.g., to guide resource-driven tradeoffs on which users to prioritize.

4. IDENTIFYING CONFOUNDING FACTORS

In this section, we propose a framework for identifying confounding factors. To this end, we begin with a taxonomy of potentially confounding factors. Then, we use the machine learning building blocks described in the previous section to identify the factors that have a non-trivial impact on engagement.

4.1 Approach

We identify three categories of potential confounding factors from our dataset:

- **Content attributes:** This includes the *type of video* (i.e., live vs. VOD) and the *overall popularity* (i.e., number of views).
- **User attributes:** This includes the user's *location* (region within continental US), *device* (e.g., smartphones, tablets, PC, TV), and *connectivity* (e.g., DSL, cable, mobile or wireless).
- **Temporal attributes:** Unlike live content that is viewed during the event, VOD objects in the dataset are available to be accessed at any point in time since its release. This opens up various temporal attributes that can possibly affect engagement including the *time of day* and *day of week* of the session and the *time since release* for the object (e.g., day of release vs. not).

Confounding Factor	Engagement	Join Time	Buff. Ratio	Rate of buff.	Avg. bitrate
Type of video (live or VOD)	8.8	15.2	0.7	0.3	6.9
Overall popularity (live)	0.1	0.0	0.0	0.2	0.4
Overall popularity (VOD)	0.1	0.2	0.4	0.1	0.2
Time since release (VOD)	0.1	0.1	0.1	0.0	0.2
Time of day (VOD)	0.2	0.6	2.2	0.5	0.4
Day of week (VOD)	0.1	0.2	1.1	0.2	0.1
Device (live)	1.3	1.3	1.1	1.2	2.7
Device (VOD)	0.5	11.8	1.5	1.5	10.3
Region (live)	0.6	0.7	1.3	0.5	0.4
Region (VOD)	0.1	0.3	1.2	0.2	0.2
Connectivity (live)	0.7	1.1	1.4	1.1	1.5
Connectivity (VOD)	0.1	0.4	1.1	1.4	1.3

Table 2: Relative information gain (%) between different potential confounding factors and the engagement and quality metrics. We mark any factor with more than 5% information gain as a potential confounding factor

We acknowledge that this list is only representative as we are only accounting for factors that can be measured directly and objectively. For example, the user's interest in the particular content is also a potential confounding factors that we cannot directly measure. Our goal here is to develop a systematic methodology to identify and account for these factors. Given more fine-grained instrumentation to measure other types of factors (e.g., use of gaze tracking in HCI), we can use our framework to evaluate these other factors as well.

In Section 2, we saw that confounding factors can act in three possible ways:

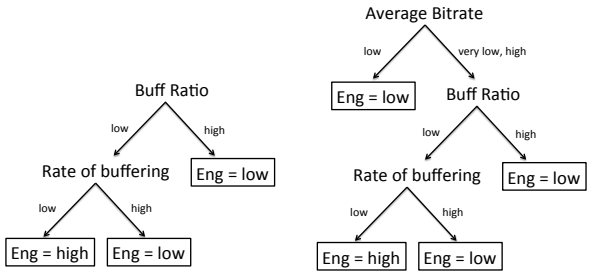
1. They can affect the observed engagement (e.g., Figure 4a)
2. They can affect the observed quality metric and thus indirectly impact engagement (e.g., Figure 4b);
3. They can impact the nature and magnitude of quality $\rightarrow$ engagement relationship (e.g., Figure 4c).

For (1) and (2) we use *information gain analysis* to identify if there is a hidden relationship between the potential confounding factor w.r.t engagement or the quality metrics. For (3), we identify two sub-effects: the impact of the confounding factor on the quality $\rightarrow$ engagement relationship can be *qualitative* (i.e., the relative importance of the different quality metrics may change) or it can be *quantitative* (i.e., the tolerance to one or more of the quality metrics might be different). For the qualitative effect, we use the compacted decision tree separately for each class (e.g., TV vs. mobile vs. PC) using the method described in Section 3.2 and compare their structure. Finally, for the quantitative sub-effect in (3), we simply check if there is any significant difference in tolerance.

4.2 Analysis results

Next, for each user, content, and temporal attribute we run the different identification techniques to check if it needs to be flagged as a potential confounding factor. Table 2 presents the information gain between each factor with respect to engagement (fraction of video viewed before quitting) and the four quality metrics.

Type of video: Classified as live or VOD session, as Table 2 shows, the type of video has high information gain with respect



(a) VOD *(b)* Live

Figure 7: Compacted decision tree for live and VOD are considerably different in structure

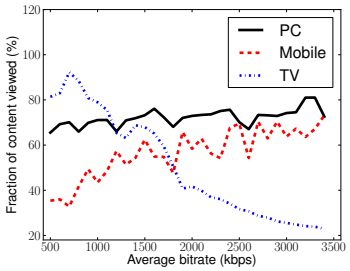


Figure 8: Anomalous trend : Higher bitrate led to lower engagement in the case of TV in the VOD dataset

to engagement confirming our earlier observation that the viewing behavior for live and VOD are different (Section 2.3). Again, since join time distributions for live and VOD sessions are also different (Section 2.3), it is not surprising that we observe high information gain in join time. Similarly, the set of bitrates used by the live provider and the VOD provider are quite different leading to high information gain for average bitrate as well.

We learn the compacted decision tree for VOD and live sessions separately as shown in Figure 7 and see that the trees are structurally different. While buffering ratio and rate of buffering have the highest impact for VOD, average bitrate has the highest impact for live events. Somewhat surprisingly, some live users tolerate very low bitrates. We believe this is related to an observation from prior work which showed that users viewing live sporting events may run the video in background and the player automatically switches to lower quality to reduce CPU consumption [20].

Since the differences between live and VOD sessions are considerably large, for the remaining attributes, we perform the analysis separately for live and VOD data.

Device: We classify the devices as PC (desktops and laptops) or mobile devices (smartphones and tablets) or TV (e.g., via Xbox). In the VOD dataset, 66% of the traffic were initiated from PC and around 33% were from TV. Mobile users formed a small fraction. However, in the live dataset, 80% of the traffic were initiated from PCs and almost 20% of the traffic from mobile users—users on TV formed a small fraction.

For a subset of the VOD data, we observed that the compacted decision tree for the TV users was different from that of mobile and PC users. While PC and mobile users showed a tree structure similar to Figure 7a, we observed Figure 9 in the case of TV. Intrigued by this difference, we visualized the impact of bitrate on engagement. Curiously, we find in Figure 8 that increased bitrate led to lower engagement in the case of TV. This is especially surprising as one would expect that users would prefer higher bitrates

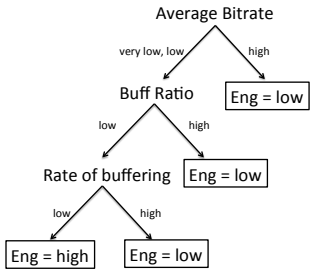
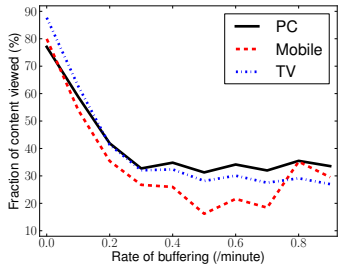
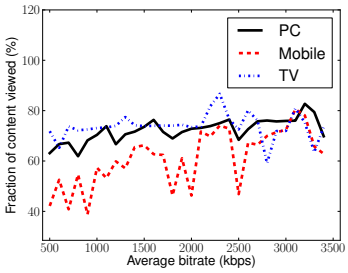


Figure 9: Compacted decision tree for TV for the VOD data that showed the anomalous trend



(a) Rate of buffering



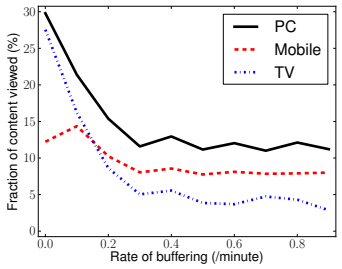
(b) Average bitrate

Figure 10: VOD users on different devices have different levels of tolerance for rate of buffering and average bitrate

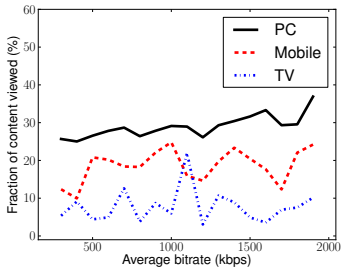
on larger screens. Investigating this further, we saw complaints on the specific content provider's forum regarding contrast issues at higher bitrate when viewing the video on TV. This issue was later corrected by the provider and the newer data did not show this anomaly. As shown in Table 2, we observe a high information gain in terms of join time and average bitrate for VOD data.

Even though the compacted tree was similar in structure for TV, PC and mobile users (not shown), Figure 10 and 11 show substantial differences in tolerance levels for average bitrate and rate of buffering. This is consistent with a recent measurement study that shows that mobile users are more tolerant toward low quality [26]. The one difference with live data, however, is that device does not lead to high information gain for engagement or any of the quality metrics (Table 2). Because of the differences in tolerance, we consider device as an important confounding factor.

Connectivity: Based on the origin ISP, we classify the video session as originating from a DSL/cable provider or from a wireless provider. In the VOD dataset, 95% of the sessions were initiated from DSL/cable providers. In the live dataset, 90% were from DSL/cable providers. We see that sessions with wireless connection had slightly lower bitrates and higher buffering values compared to the sessions in cable and DSL connection. This accounts for the slight information gain that we observe in Table 2.



(a) Rate of buffering



(b) Average bitrate

Figure 11: Live users on different devices have different levels of tolerance for rate of buffering and average bitrate

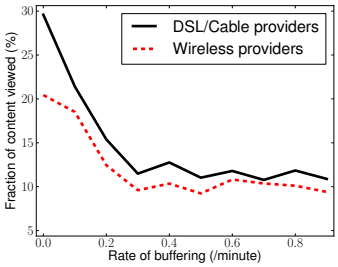


Figure 12: For live content, users on DSL/cable connection and users on wireless connection showed difference in tolerance for rate of buffering

The compacted decision trees had the same structure for cable vs. wireless providers for both live and VOD data. But we observed difference in tolerance to rate of buffering for both live and VOD content. As we observed earlier in Section 2.3, users were more tolerant to buffering rate when they were on a wireless provider for VOD content. For live content, as shown in Figure 12, we observed difference in tolerance for rate of buffering. Due to these differences, we consider connectivity as a confounding factor.

Time of day: Based on the time of the day, the sessions were classified as during night time (midnight-9am), day time (9am-6pm) or peak hours(6pm-midnight). We observed that 15% of the traffic were during night time, 30% during day time and 55% during peak hours. We also observed that users experienced slightly more buffering during peak hours when compared to late nights and day time. The compacted decision trees were similar for peak hours vs. day vs. night. Users were, however, slightly more tolerant to rate of buffering during peak hours as shown in Figure 13. Since we want to take a conservative stance while shortlisting confounding factors, we consider time of day to be a confounding factor.

Other factors: We did not observe high information gain or significant differences in the compacted decision tree or the tolerance to quality for other factors such as region, popularity, day of week,

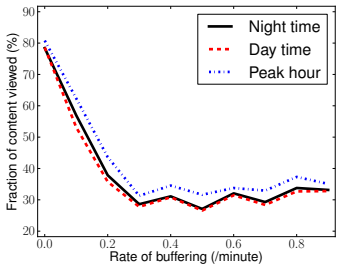


Figure 13: For VOD, users tolerance for rate of buffering is slightly higher during peak hours

Confounding Factor	Engmnt	Quality	Q→E Qual	Q→E Quant
Type of video - live or VOD	✓	✓	✓	✓
Overall popularity (live)	✗	✗	✗	✗
Overall popularity (VOD)	✗	✗	✗	✗
Time since release (VOD)	✗	✗	✗	✗
Time of day (VOD)	✗	✗	✗	✓
Day of week (VOD)	✗	✗	✗	✗
Device (live)	✗	✗	✗	✓
Device (VOD)	✗	✓	✓ ✗	✓
Region (live)	✗	✗	✗	✗
Region (VOD)	✗	✗	✗	✗
Connectivity (live)	✗	✗	✗	✓
Connectivity (VOD)	✗	✗	✗	✓

Table 3: Summary of the confounding factors. Check mark indicates if a factor impacts quality or engagement or the quality→engagement relationship. The highlighted rows show the key confounding factors that we identify and use for refining our predictive model

and time since video release (not shown). Thus, we do not consider these as confounding factors.

4.3 Summary of main observations

Table 3 summarizes our findings from the analysis of various potential confounding factors. Our main findings are:

- The main confounding factors are type of video, device, and connectivity.
- The four techniques that we proposed for detecting confounding factors are *complementary* and expose different facets of the confounding factors.
- Our model also reconfirmed prior observations on player-specific optimizations for background video sessions. It was also able to reveal interesting anomalies due to specific player bugs.

5. ADDRESSING CONFOUNDING FACTORS

Next, we describe how we refine the basic decision tree model we saw in Section 3.2 to take into account the key confounding factors from the previous section. We begin by describing two candidate approaches for model refinement and the tradeoffs involved. We study the impact of both candidate approaches and choose a heuristic "splitting" based approach.

5.1 Candidate approaches

There are two candidate approaches to incorporate the confounding factors into the predictive model:

- **Add as new feature:** The simplest approach is to add the key confounding factors as additional features in the input to the machine learning algorithm and relearn the prediction model.

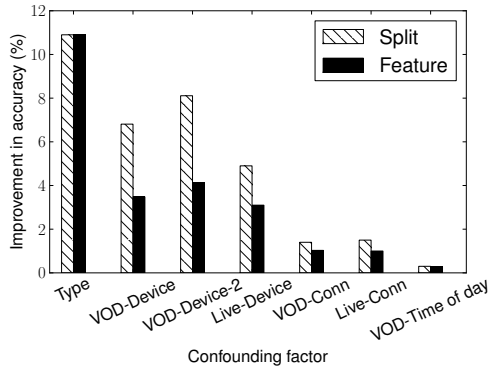


Figure 14: Comparing feature vs split approach for the different confounding factors

- **Split Data:** Another possibility is to split the data based on the confounding factors (e.g., live on mobile device) and learn separate models for each split. Our predictive model would then be the logical union of multiple decision trees—one for each combination of the values of various confounding factors.

Both approaches have pros and cons. The feature-addition approach has the appeal of being simple and requiring minimal modifications to the machine learning framework. (This assumes that the learning algorithm is robust enough to capture the effects caused by the confounding factors). Furthermore, it will learn a single unified model over all the data. The augmented model we learn, however, might be less intuitive and less amenable to compact representations. Specifically, in the context of the decision tree, mixing quality metrics with confounding factors may result in different levels of the tree branching out on different types of variables. This makes it harder to visually reason about the implications for system design. For instance, consider the scenario where we want to know the impact of quality on engagement for mobile users in order to design a new mobile-specific bitrate adaptation algorithm for live sports content. This is a natural and pertinent question that a practitioner would face. To answer this question, we would in effect have to create a new "projection" of the tree that loses the original structure of the decision tree. Moreover, we would have to create this projection for every such system design question.

In contrast, the split data approach will explicitly generate these intuitive projections for different combinations of the confounding factors by construction. It also avoids any doubts we may have about the expressiveness of the machine learning algorithm. The challenge with the split approach is the "curse of dimensionality". As we have more factors to split, the available data per split becomes progressively sparser. Consequently, the model learned may not have sufficient data samples to create a robust enough model.[5] Fortunately, we have two reasons to be hopeful in our setting. First, we have already pruned the set of possibly confounding external factors to the key confounding factors. Second, as Internet video grows, we will have larger datasets to run these algorithms and that will alleviate concerns with limited data for multiple splits.

Following in the data-driven spirit of our approach, we analyze the improvements in prediction accuracy that each approach gives before choosing one of these techniques.

5.2 Results

[5]Note that this dimensionality problem is not unique to the split data approach. A decision tree (or any learning algorithm for that matter) faces the same problem as we go deeper into the tree as well. The split approach just elevates the dimensionality problem to the first stage of the learning itself.

Model	Accuracy (in %)
Simple decision tree	45.02
Without early-quitters	51.20
Multiple decision tree	68.74

Table 4: Summary of the model refinements and resultant accuracy when number of classes for engagement is 10

For this study, we set the number of classes for engagement to 10. We observe similar results for other number of classes as well. Figure 14 compares the increase in accuracy using the feature and the split approach for the three key confounding factors.

As shown in Figure 14, splitting based on type of video vs. adding it as a feature (*Type*) results in the same increase in accuracy. In the case of the split approach, we observe that both splits (live and VOD) do not have the same accuracy—live is more predictable than VOD. However, splitting based on the device type gives better improvement compared to adding device as a feature for both VOD and live (*VOD-Device*, *VOD-Device-2* and *Live-Device*). But, we observed that the accuracy across the splits were not the same. For the VOD dataset, splits corresponding to TV and PC had higher accuracy compared to the split corresponding to smartphones. This is because, as we saw in Section 4, only a small fraction of users viewed VOD content on mobile phones in our dataset. *VOD-Device-2* corresponds to the data in which we observed an anomalous trend in viewing behavior on TV. Here, we observed that the split corresponding to TV had very high accuracy leading to better gains from splitting. For the live dataset, we however observed that the TV split had lower gains compared to mobile and smartphones. This is again because of the inadequate amount of data—the fraction of users watching live content on TV in our dataset was negligible.

Splitting works better than feature addition for both live (*Live-Conn*) and VOD (*VOD-Conn*) in the case of connectivity and for time of day in the case of VOD (*VOD-Time of day*). Time of day did not lead to a huge improvement in improvement in accuracy and hence we ignore it. The other external factors that we considered in Section 4 led to negligible increase in accuracy when addressed using both these approaches.

Why does split perform better? Across the various confounding factors, we see that the split data approach is better (or equivalent) to the feature addition approach. The reason for this is related to the decision tree algorithm. Decision trees use information gain for identifying the best attribute to branch on. Information gain based schemes, however, are biased towards attributes that have multiple levels [19]. While we bin all the quality metrics at an extremely fine level, the confounding factors have only few categories (e.g., TV or PC/laptop or smartphone/tablet in the case of devices). This biases the decision tree towards always selecting the quality metrics to be more important. In the case of type of video, the information gain in engagement is very high since user viewing behavior is very different (i.e, it satisfies criteria number (1) that we have for identifying confounding factors). So it gets chosen at the top level and hence splitting and adding as a feature led to same gain.

5.3 Proposed predictive model

As mentioned in Section 3.3, we observed many users who "sample" the video and quit early if it is not of interest [41]. Taking into account this *domain-specific observation*, we ignore these early-quitter sessions from our dataset and relearn the model leading to $\approx 6\%$ increase in accuracy.

Further incorporating the three key confounding factors (type of device, device and connectivity), we propose a unified QoE model

based on splitting the dataset for various confounding factors and learning multiple decision trees—one for each split. Accounting for all the confounding factors further leads to around 18% improvement. Table 4 summarizes the overall accuracies when number of classes for engagement is set to 10. This implies that about 70% of the predictions are within the same 10% bucket as the actual user viewing duration.

6. IMPLICATIONS FOR SYSTEM DESIGN

In this section, we demonstrate the practical utility of the QoE model using trace-driven simulations. We simulate a video control plane setting similar to previous work and use our QoE model to guide the choice of CDN and bitrate [28]. We compare the potential improvement in engagement using our QoE model against other strawman solutions.

6.1 Overview of a video control plane

The QoE model that we developed can be used by various principals in the Internet video ecosystem to guide system design decisions. For instance, video player designers can use the model to guide the design of efficient bitrate adaptation algorithms. Similarly, CDNs can optimize overall engagement by assigning bitrates for each individual client using our QoE model.

Prior work makes the case for a coordinated control plane for Internet video based on their observation that a purely client- or server- driven bitrate and CDN adaptation scheme for video sessions might be suboptimal [28]. This (hypothetical) control plane design uses a global view of the network and CDN performance to choose the CDN and bitrates for each session based on a global optimization framework. The goal of the optimization is to pick the right control parameters (in this case, CDN and bitrate) in order to maximize the overall engagement. As shown in Figure 15, this control plane takes as input control parameters (CDN, bitrate) and other attributes (device, region, ISP etc.) as input and predicts the expected engagement.

Our QoE model can be used to guide the design of such a video control plane. Note, however, that the QoE model from Section 5 takes various quality metrics and attributes that are confounding as input and predicts the expected engagement. Thus, as discussed in Section 2, we also need to develop a quality model which takes CDN, bitrate, and client attributes as input and predicts the quality metrics (buffering ratio, rate of buffering and join time) in order to fully realize this control plane design. Figure 15 shows the various components and their inputs and outputs. Our key contribution here is in demonstrating the use of the QoE model within this control plane framework and showing that a QoE-aware delivery infrastructure could further improve the overall engagement.

6.2 Quality model

We use a simplified version of the quality prediction model proposed from prior work [28]. It computes the mean performance (buffering ratio, rate of buffering and join time) for each combination of attributes (e.g., type of video, ISP, region, device) and control parameters (e.g., bitrate and CDN) using empirical estimation. For example, we estimate the performance of all Comcast clients in the east coast of the United States that streamed live content over an Xbox from Akamai at 2500 Kbps by computing the empirical mean for each of the quality metrics.

When queried with specific attributes (CDN and bitrate) the models returns the estimated performance. One challenge, however, is that adding more attributes to model often leads to data sparsity. In this case, we use a hierarchical estimation heuristic—i.e, if we do not have sufficient data to compute the mean performance value for a specific attribute, CDN and bitrate combination, we use a coarser-grained granularity of attribute elements [28]. For example, if we do not have enough data regarding the performance of Xbox over Akamai over Comcast connection from the east coast at 2500 Kbps, we return the mean performance that we observed over all the devices over Akamai at 2500 Kbps over Comcast connection from the east coast. We follow the following hierarchy for this estimation: {Type of video, ISP, region, device} $<$ {Type of video, ISP, region} $<$ {Type of video, ISP}.

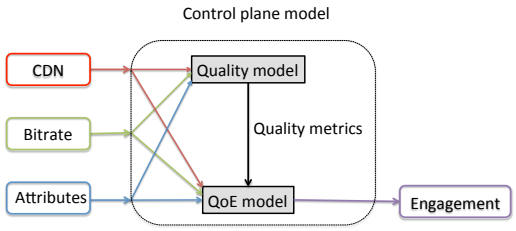


Figure 15: We use a simple quality model along with our QoE model to simulate a control plane. The inputs and outputs to the various components are shown above.

6.3 Strategies

We compare the following strategies to pick the control parameters (CDN and bitrate):

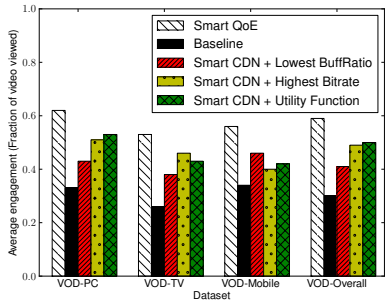
1. Smart QoE approach: For our smart QoE approach, we use a *predicted quality model* and a *predicted QoE model* based on historical data. For choosing the best control parameters for a particular session, we employ the following brute force approach. We estimate the expected engagement for all possible combinations of CDNs and bitrates by querying the *predicted quality model* and the *predicted QoE model* with the appropriate attributes (ISP, device etc.). This approach assigns the CDN, bitrate combination that gives the best predicted engagement.

2. Smart CDN approaches: We find the best CDN for a given combination of attributes (region, ISP and device) using the *predicted quality model* by comparing the mean performance of each CDN in terms of buffering ratio across all bitrates and assign clients to this CDN. We implement three variants for picking the bitrate:
2(a) Smart CDN, highest bitrate: The client always chooses to stream at the highest bitrate that is available.
2(b) Smart CDN, lowest buffering ratio: The client is assigned the bitrate that is expected to cause the lowest buffering ratio based on the *predicted quality model*
2(c) Smart CDN, control plane utility function: The client is assigned the bitrate that would maximize the utility function $(-3.7 \times BuffRatio + \frac{Bitrate}{20})$ [28].

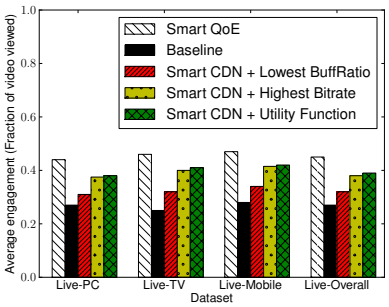
3. Baseline: We implemented a naive approach where the client picks a CDN and bitrate randomly.

6.4 Evaluation

To quantitatively evaluate the benefit of using the QoE model, we perform a trace based simulation. We use week-long trace to simulate client attributes and arrival times. In each epoch (one hour time slots), a number of clients with varying attributes (type of video, ISP, device) arrive. For each client session, we assign the CDN and bitrate based on the various strategies mentioned earlier. For simplicity, we assume the CDNs are sufficiently provisioned and do not degrade their performance throughout our simulation. To evaluate the performance of these strategies, we develop *actual engagement models* and an *actual quality models* based on the empirical data from the current measurement epoch and compare the engagement predicted by these models for each session. (Note that



(a) VOD



(b) Live

Figure 16: Comparing the predicted average engagement for the different strategies

the models that we use for prediction are based on historical data). Since the arrival patterns and the client attributes are the same for all the strategies, they have the same denominator in each epoch.

Figure 16 compares the performance of the different strategies for live and VOD datasets broken down by performance on each device type. As expected, the baseline scheme has the worst performance. The smart QoE approach can potentially improve user engagement by up to 2× compared to the baseline scheme. We observed that the smart CDN and lowest buffering ratio scheme picks the lowest bitrates and hence the expected engagements are lower compared to the other smart schemes (except in the case of VOD on mobile phones where it outperforms the other smart CDN approaches). The smart CDN with utility function approach and smart CDN highest bitrate approaches have very comparable performances. This is because the utility function favors the highest bitrate in most cases. Our smart QoE approach picks intermediate bitrates and dynamically shifts between picking the lower and the higher bitrates based on the various attributes and the predicted quality. Thus, it can potentially improve user engagement by more than 20% compared to the other strategies.

7. DISCUSSION

Other engagement measures: Content providers are also interested in other measures of engagement involving different time scales of user involvement such as customer return probability to the particular service. The quality metrics might impact these other engagement measures differently [20]; e.g., join time may affect the return probability of the customer even though it does not have a huge impact on the user engagement during a particular session. We may have to weigh in these different notions of engagement to compute an aggregate engagement index. Having said that, we believe the high-level data-driven approach we propose can be applied to other notions of engagement as well.

Evolution of the QoE model: As the network and the user expectations for quality change with time, the QoE model also needs to evolve to capture these effects. For instance, the specific bitrates at which providers serve content might change with time. Similarly, with time, users might have higher expectations with respect to quality from the video delivery system. In this context, we envision a live refinement system that constantly observes and analyzes the user viewing habits and continuously adapts the QoE model based on these observations.

QoE model for other Internet services: The methodology that we proposed can be generalized to be used for developing QoE models for other Internet services as well. The specific metrics, confounding factors and inferences might be different, but the general methodology of developing a data-driven predictive QoE model using machine learning techniques can be applied to new domains like VoIP, online gaming etc.

8. RELATED WORK

Engagement in Internet video: Past measurement studies have shown that video quality impacts user engagement [20, 26]. However, they provide a simple quantitative understanding of the impact of individual quality metrics (e.g., buffering) on engagement. We shed further light and provide a unified understanding of how all the quality metrics when put together impact engagement by developing a QoE model. Similarly, previous studies have also shown that a few external factors (e.g., connectivity) affect user engagement [26]. Recent work suggests the use of Quasi Experimental Design (QED) to eliminate any possible bias that can be caused by confounding factors and establish causal relationships [26]. However, these factors have to be provided a priori and there does not exist any techniques to identify if an external factor is potentially confounding or not. We extend our previous work [15] by developing techniques to identify external factors that are confounding and incorporate these factors to form a unified QoE model.

User studies: Prior work by the multimedia community try to assess video quality by performing subjective user studies and validating objective video quality models against the user study scores [11, 18,25,30,32]. User studies are typically done at a small-scale with a few hundred users and the perceptual scores given by users under a controlled setting may not translate into measures of user engagement in the wild. The data-driven approach that we proposed is scalable and it produces an engagement-centric model.

Control plane: Liu et al., make a case for a co-ordinated video control plane that uses measurement-driven parameters to improve video quality by adapting CDN and bitrate of clients using a global optimization scheme [28]. As we showed, our QoE model can be used under a similar framework to further improve the benefits.

Adaptive video players design: Commercial video players today perform client-side bitrate adaptation based on current bandwidth conditions [4]. Studies that have analyzed these players have found that there is significant scope for improving their adaptation schemes [13]. Video player designers typically use ad hoc mechanisms to trade-off between various network parameters [12, 13]. Our video QoE model can be potentially used to develop engagement-centric video player adaptation algorithms.

Diagnosis: Past work has looked at techniques to proactively diagnose quality issues in video delivery in order to minimize its impact on users [29, 38]. In Section 4.2, we show that our model can also

detect anomalous behavior among users watching VOD content on TV, and potentially other quality issues as well.

QoE metrics in other media: There have been attempts to study the impact of network factors on user engagement and user satisfaction in the context of other media technologies. For example, in [16], the authors study the impact of bitrate, jitter and delay on call duration in Skype and propose a unified user satisfaction metric as a combination of these factors. Our approach derives a unified QoE model in the context of Internet video and it is very timely given that Internet video has gone mainstream in the past few years.

Other video measurement studies: Several measurement studies of Internet video have focused on content popularity, user behavior and viewing patterns [22, 41]. The observations made in these works have implications on understanding measurement-driven insights and performing domain-specific refinements to improve the QoE model. For instance, Yu et al., also observed that users sample videos and quit the session early [41]. Similarly, we observed that some users tolerate low bitrate while watching live content. Previous work also observed this phenomena which is a result of the player running in the background [20].

9. CONCLUSIONS

An imminent challenge that the Internet video ecosystem—content providers, content delivery networks, analytics services, video player designers, and users—face is the lack of a reference methodology 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., [34]), such a measure will become a fundamental requirement to promote further innovation by allowing us to objectively compare different competing designs [8, 14].

Internet video presents both a challenge and an opportunity for QoE measurement. On one hand, the nature of the delivery infrastructure introduces new complex relationships between quality and engagement and between quality metrics themselves. To further make matters worse, there are many confounding factors introduced by different aspects of this ecosystem that directly or indirectly impact engagement (e.g., genre, popularity, device). At the same time, however, we have an unprecedented opportunity to obtain a systematic understanding of QoE because of the ability to collect large client- and server-side measurements of actual user behavior *in the wild*.

This paper is a significant first step in seizing this opportunity and addressing the above challenges. We developed a data-driven machine learning approach to capture the complex interactions as well as confounding effects. We also demonstrated significant practical benefits that content providers can obtain by using an improved QoE prediction model over current ad hoc approaches.

Acknowledgements

We thank our shepherd Jon Crowcroft and the anonymous reviewers for their feedback that helped improve the paper. We also thank the Conviva staff for answering several questions about the dataset and the data collection infrastructure. This work is partially supported by the National Science Foundation under grants CNS-1050170, CNS-1017545, CNS-0905134 and CNS- 0746531

10. REFERENCES

[1] Buyer's Guide: Content Delivery Networks. http://goo.gl/B6gMK.
[2] Cisco study. http://goo.gl/tMRwM.
[3] Driving Engagement for Online Video. http://goo.gl/pO5Cj.
[4] Microsoft Smooth Streaming. http://goo.gl/6JOXh.
[5] P.800 : Methods for subjective determination of transmission quality. http://www.itu.int/rec/T-REC-P.800-199608-I/en.
[6] P.910 : Subjective video quality assessment methods for multimedia applications. http://goo.gl/QjFhZ.
[7] Peak signal to noise ratio. http://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio.
[8] SPEC philosophy. http://www.spec.org/spec/#philosophy.
[9] Turbobytes: How it works. http://www.turbobytes.com/products/optimizer/.
[10] Video quality metrics. http://goo.gl/Ga9Xz.
[11] Vqeg. http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx.
[12] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen. What Happens when HTTP Adaptive Streaming Players Compete for Bandwidth? In *Proc. NOSSDAV*, 2012.
[13] S. Akhshabi, A. Begen, and C. Dovrolis. An Experimental Evaluation of Rate Adaptation Algorithms in Adaptive Streaming over HTTP. In *MMSys*, 2011.
[14] R. H. Allen and R. D. Sriram. The Role of Standards in Innovation. *Elsevier: Technology Forecasting and Social Change*, 2000.
[15] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. A quest for an internet video quality-of-experience metric. In *Hotnets*, 2012.
[16] K. Chen, C. Huang, P. Huang, and C. Lei. Quantifying skype user satisfaction. In *SIGCOMM*, 2006.
[17] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. In *IEEE Transactions on Broadcasting*, 2011.
[18] N. Cranley, P. Perry, and L. Murphy. User perception of adapting video quality. *International Journal of Human-Computer Studies*, 2006.
[19] H. Deng, G. Runger, and E. Tuv. Bias of importance measures for multi-valued attributes and solutions. In *ICANN*, 2011.
[20] 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.
[21] J. Esteban, S. Benno, A. Beck, Y. Guo, V. Hilt, and I. Rimac. Interactions Between HTTP Adaptive Streaming and TCP. In *Proc. NOSSDAV*, 2012.
[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] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. In *IEEE Intelligent Systems*, 2009.
[24] 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.
[25] A. Khan, L. Sun, and E. Ifeachor. Qoe prediction model and its applications in video quality adaptation over umts networks. In *IEEE Transactions on Multimedia*, 2012.
[26] S. S. Krishnan and R. K. Sitaraman. Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs. In *IMC*, 2012.
[27] B. Liu, M. Hu, and W. Hsu. Intuitive Representation of Decision Trees Using General Rules and Exceptions. In *Proc. AAAI*, 2000.
[28] 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 *SIGCOMM*, 2012.
[29] A. Mahimkar, Z. Ge, A. Shaikh, J. Wang, J. Yates, Y. Zhang, and Q. Zhao. Towards Automated Performance Diagnosis in a Large IPTV Network. In *Proc. SIGCOMM*, 2009.
[30] V. Menkvoski, A. Oredope, A. Liotta, and A. C. Sanchez. Optimized online learning for qoe prediction. In *BNAIC*, 2009.
[31] T. Mitchell. *Machine Learning*. McGraw-Hill.
[32] 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.
[33] L. Plissonneau and E. Biersack. A Longitudinal View of HTTP Video Streaming Performance. In *MMSys*, 2012.
[34] I. Sodagar. The MPEG-DASH Standard for Multimedia Streaming Over the Internet. *IEEE Multimedia*, 2011.
[35] H. H. Song, Z. Ge, A. Mahimkar, J. Wang, J. Yates, Y. Zhang, A. Basso, and M. Chen. Q-score: Proactive Service Quality Assessment in a Large IPTV System. In *Proc. IMC*, 2011.
[36] M. Watson. Http adaptive streaming in practice. In *MMSys - Keynote*, 2011.
[37] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2000.
[38] C. Wu, B. Li, and S. Zhao. Diagnosing Network-wide P2P Live Streaming Inefficiencies. In *Proc. INFOCOM*, 2009.
[39] W. Xu, L. Huang, A. Fox, D. Patterson, and M. Jordan. Detecting large-scale system problems by mining console logs. In *Proc. SOSP*, 2009.
[40] H. Yin et al. Inside the Bird's Nest: Measurements of Large-Scale Live VoD from the 2008 Olympics.
[41] H. Yu, D. Z. B. Y. Zhao, and W. Zheng. Understanding User Behavior in Large-Scale Video-on-Demand Systems. In *Proc. Eurosys*, 2006.