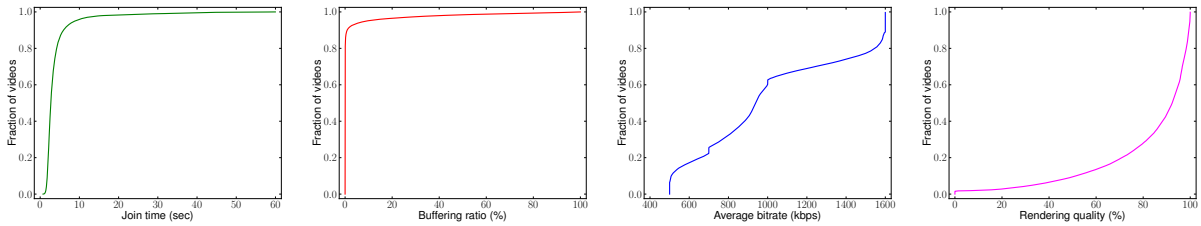
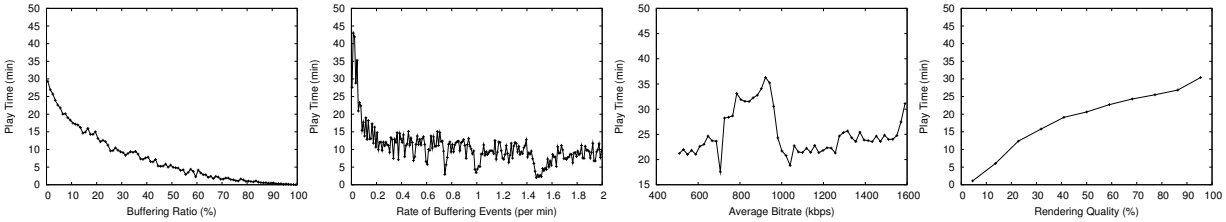


(a) Join time (b) Buffering ratio (c) Average bitrate (d) Rendering quality

Figure 2: CDFs for four quality metrics for dataset $LvodA$.



(a) Buffering ratio (b) Rate of buffering events (c) Average bitrate (d) Rendering quality

Figure 3: Qualitative relationships between four quality metrics and the play time for a video from $LvodA$.

trailers, short interviews, and short skits. The two short VoD datasets are labeled as $SvodA$ and $SvodB$.

- **Live:** Sports events and news feeds are typically delivered as live video streams. There are two key differences between the VoD-type content and live streams. First, the client buffers in this case are sized such that the viewer does not lag more than a few seconds behind the video source. Second, all viewers are roughly synchronized in time. The two live datasets are labeled $LiveA$ and $LiveB$. As a special case study, dataset $LiveH$ corresponds to the three of the final World Cup games with almost a million viewers per game on average (1.2 million viewers for the last game from this dataset).

3. ANALYSIS TECHNIQUES

In this section, we begin with real-world measurements to motivate the types of questions we want to answer and explain our analysis methodology toward addressing these questions.

3.1 Overview

To put our work in perspective, Figure 2 shows the cumulative distribution functions (CDF) of four quality metrics for dataset $LvodA$. As expected, most viewing sessions experience very good quality, i.e., have very low $BufRatio$, low $JoinTime$, and relatively high $RendQual$. However, the number of views that suffer from quality issues is not trivial. In particular, 7% of views experience $BufRatio$ larger than 10%, 5% of views have $JoinTime$ larger than 10s, and 37% of views have $RendQual$ lower than 90%. Finally, only a relatively small fraction of views receive the highest bit rate. Given that a non-negligible number of views experience quality issues, it is critical for content providers to understand if improving the quality of these sessions could have potentially increased the user engagement.

To understand how the quality could potentially impact the engagement, we consider one video object each from $LiveA$ and $LvodA$. For this video, we bin the different sessions based on the value of the quality metrics and calculate the average play time for each bin. Figures 3 and 4 show how the four quality metrics interact with the play time. Looking at the trends visually confirms that *quality matters*. At the same time, these initial visualizations spark several questions:

- How do we *identify* which metrics matter the most?
- Are these quality metrics *independent* or are they manifestations of the same underlying phenomenon? In other words, is the observed relationship between the engagement and the quality metric M really due to M or due to a hidden relationship between M and another more critical metric M'?
- How do we *quantify* how important a quality metric is?
- Can we explain the seemingly counter-intuitive behaviors? For example, $RendQual$ is actually negatively correlated for the $LiveA$ video (Figure 4(d)), while the $AvgBitrate$ shows an unexpected non-monotone trend for $LvodA$ (Figure 3(c)).

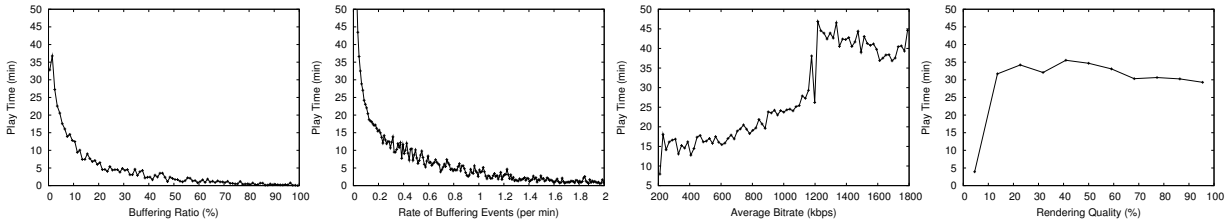
To address the first two questions, we use the well-known concepts of correlation and information gain from the data mining literature that we describe next. To measure the quantitative impact, we also use linear regression based models for the most important metric(s). Finally, we use domain-specific insights and experiments in controlled settings to explain the anomalous observations.

3.2 Correlation

The natural approach to quantify the interaction between a pair of variables is the correlation. Here, we are interested in quantifying the magnitude and direction of the relationship between the engagement metric and the quality metrics.

To avoid making assumptions about the nature of the relationships between the variables, we choose the Kendall correlation, instead of the Pearson correlation. The Kendall correlation is a *rank correlation* that does not make any assumption about the underlying distributions, noise, or the nature of the relationships. (Pearson correlation assumes that the noise in the data is Gaussian and that the relationship is roughly linear.)

Given the raw data–a vector of (x,y) values where each x is the measured quality metric and y the engagement metric (play time or number of views)–we *bin* it based on the value of the quality metric. We choose bin sizes that are appropriate for each quality metric of interest: for $JoinTime$, we use 0.5 second intervals, for $BufRatio$ and $RendQual$ we use 1% bins, for $RateBuf$ we use 0.01/min



(a) Buffering ratio (b) Rate of buffering events (c) Average bitrate (d) Rendering quality

Figure 4: Qualitative relationships between four quality metrics and the play time for a video from $LiveA$.

sized bins, and for $AvgBitrate$ we use 20 kbps-sized bins. For each bin, we compute the empirical *mean* of the engagement metric across the sessions/viewers that fall in the bin.

We compute the Kendall correlation between the mean-per-bin vector and the values of the bin indices. We use this "binned" correlation metric for two reasons. First, we observed that the correlation coefficient[2] was biased by a large mass of users that had high quality but very low play time, possibly because of low user interest. Our primary goal, in this paper, is not to study user interest in the specific content. Rather, we want to understand if and how the quality impacts user engagement. To this end, we look at the average value for each bin and compute the correlation on the binned data. The second reason is scale. Computing the rank correlation is computationally expensive at the scale of analysis we target. The binned correlation retains the qualitative properties that we want to highlight with lower compute cost.

3.3 Information Gain

Correlations are useful for quantifying the interaction between variables when the relationship is roughly *monotone* (either increasing or decreasing). As Figure 3(c) shows, this may not always be the case. Further, we want to move beyond the single metric analysis. First, we want to understand if a pair (or a set) of quality metrics are complementary or if they capture the same effects. As an example, consider $RendQual$ in Figure 3; $RendQual$ could reflect either a network issue or a client-side CPU issue. Because $BufRatio$ is also correlated with $PlayTime$, we suspect that $RendQual$ is mirroring the same effect. Identifying and uncovering these hidden relationships, however, is tedious. Second, content providers may want to know the top k metrics that they should to optimize to improve user engagement. Correlation-based analysis cannot answer such questions.

To address the above challenges, we augment the correlation analysis using the notion of *information gain* [32], which is based on the concept of entropy. The entropy of random variable Y is $H(Y) = \sum_i P[Y = y_i] \log \frac{1}{P[Y=y_i]}$, where $P[Y = y_i]$ is the probability that $Y = y_i$. The conditional entropy of Y given another random variable X is defined as $H(Y|X) = \sum_j P[X = X_j]H(Y|X = x_j)$ and the information gain is then $H(Y) - H(Y|X)$, and the relative information gain is $\frac{H(Y)-H(Y|X)}{H(Y)}$. Intuitively, this metric quantifies how our knowledge of X reduces the uncertainty in Y.

Specifically, we want to quantify what a quality metric informs us about the engagement; e.g., what does knowing the $AvgBitrate$ or $BufRatio$ tell us about the play time distribution? As with the correlation, we bin the data into discrete bins with the same bin specifications. For the play time, we choose different bin sizes depending on the duration of the content. From this binned data, we compute $H(Y|X_1, \ldots, X_N)$, where Y is the discretized play time and $X_1, \ldots, X_N$ are quality metrics. From this estimate, we calculate the relative information gain.

Note that these two classes of analysis techniques are *complementary*. Correlation provides a first-order summary of monotone relationships between engagement and quality. The information gain can corroborate the correlation or augment it when the relationship is not monotone. Further, it provides a more in-depth understanding of the interaction between the quality metrics by extending to the multivariate case.

3.4 Regression

Rank correlation and information gain are largely qualitative analyses. It is also useful to understand the *quantitative* impact of a quality metric on user engagement. Specifically, we want to answer questions of the form: *What is the expected improvement in the engagement if we optimize a specific quality metric by a given amount?*

For quantitative analysis, we rely on *regression*. However, as the visualizations show, the relationships between the quality metrics and the engagement are not always obvious and several of the metrics have intrinsic dependencies. Thus, directly applying regression techniques with complex non-linear parameters could lead to models that lack a physically meaningful interpretation. While our ultimate goal is to extract the relative quantitative impact of the different metrics, doing so rigorously is outside the scope of this paper.

As a simpler alternative, we use linear regression based curve fitting to quantify the impact of specific ranges of the most critical quality metric. However, we do so only after visually confirming that the relationship is approximately linear over the range of interest. This allows us to employ simple linear data fitting models that are also easy to interpret.

4. VIEW LEVEL ENGAGEMENT

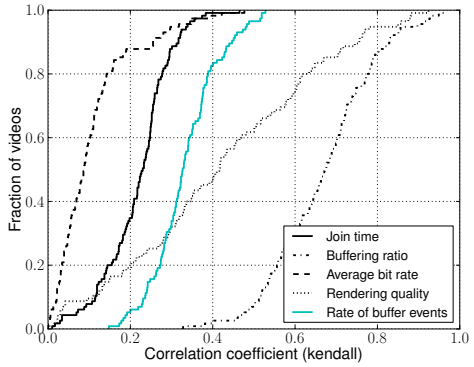
The engagement metric of interest at the view level is $PlayTime$. We begin with long VoD content, then proceed to live and short VoD content. In each case, we start with the basic correlation based analysis and augment it with information gain based analysis. Note that we compute the binned correlation and information gain coefficients on a per-video-object basis. Then we look at the distribution of the coefficients across all video objects. Having identified the most critical metric(s), we quantify the impact of improving this quality using a linear regression model over a specific range of the quality metric.

In summary, we find that $BufRatio$ consistently has the highest impact on user engagement among all quality metrics. For example, for a 90 minutes live event, an increase of $BufRatio$ by 1% can decrease $PlayTime$ by over 3 minutes. Interestingly, the relative impact of the other metrics depend on the content type. For live video, $RateBuf$ is slightly more negatively correlated with $PlayTime$ as compared to long VoD; because the player buffer

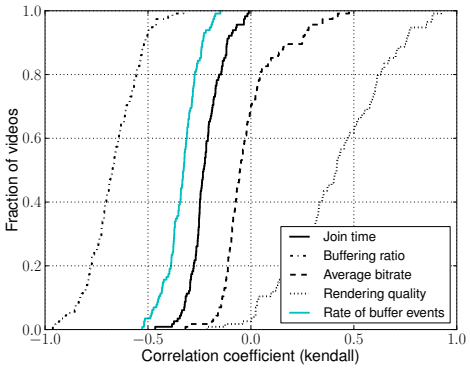
[2]This happens with Pearson and Spearman correlation metrics also.

is small there is little time to recover when the bandwidth fluctuates. Our analysis also shows that higher bitrates are more likely to improve user engagement for live content. In contrast to live and long VoD videos, for short videos $RendQual$ exhibits correlation similar to $BufRatio$. We also find that various metrics are not independent. Finally, we explain some of the anomalous observations from Section 3 in more depth.

4.1 Long VoD Content



(a) Absolute values



(b) Actual values (signed)

Figure 5: Distribution of the Kendall rank correlation coefficient between the quality metrics and play time for $LvodA$.

Figure 5 shows the distribution of the correlation coefficients for the quality metrics for dataset $LvodA$. We include both *absolute value* and *signed* values to measure the magnitude and the nature (i.e., increasing or decreasing) of the correlation. We summarize the median values for both datasets in Table 2. The results are consistent across both datasets for the common quality metrics $BufRatio$, $JoinTime$, and $RendQual$. Recall that the two datasets correspond to two different content providers; these results confirm that our observations are not unique to dataset $LvodA$.

The result shows that $BufRatio$ has the strongest correlation with $PlayTime$. Intuitively, we expect a higher $BufRatio$ to decrease $PlayTime$ (i.e., a negative correlation) and a higher $RendQual$ to increase $PlayTime$ (i.e., a positive correlation). Figure 5(b) confirms this intuition regarding the nature of these relationships. We notice that $JoinTime$ has little impact on the play duration. Surprisingly, $AvgBitrate$ has very low correlation as well.

Next, we proceed to check if the univariate information gain analysis corroborates or complements the correlation results in Figure 6. Interestingly, the relative order between $RateBuf$ and $BufRatio$

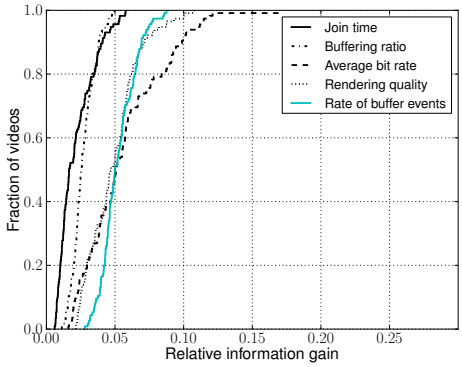


Figure 6: Distribution of the univariate gain between the quality metrics and play time, for dataset $LvodA$.

Quality metric	Correlation coefficient	
	$LvodB$	$LvodA$
$JoinTime$	-0.17	-0.23
$BufRatio$	-0.61	-0.67
$RendQual$	0.38	0.41

Table 2: Median values of the Kendall rank correlation coefficients for $LvodA$ and $LvodB$. We do not show $AvgBitrate$ and $RateBuf$ for $LvodB$ because the player did not switch bitrates or gather buffering event data. For the remaining metrics the results are consistent with dataset $LvodA$.

is reversed compared to Figure 5. The reason (see Figure 7) is that most of the probability mass is in the first bin (0-1% $BufRatio$) and the entropy here is the same as the overall distribution. Consequently, the information gain for $BufRatio$ is low; $RateBuf$ does not suffer this problem (not shown) and has higher information gain. We also see that $AvgBitrate$ has high information gain even though its correlation was very low. We revisit this observation in Section 4.1.1.

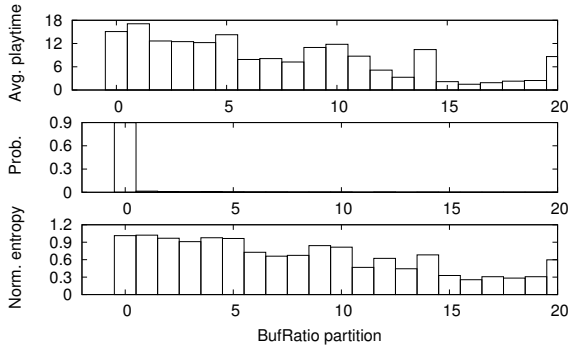


Figure 7: Visualizing why buffering ratio does not result in a high information gain even though it is correlated.

So far we have looked at each quality metric in isolation. A natural question is: Does combining two metrics provide more insights? For example, $BufRatio$ and $RendQual$ may be correlated with each other. In this case knowing that both correlate with $PlayTime$ does not add new information. To evaluate this, we show the distribution of the bivariate relative information gain in Figure 8. For clarity, rather than showing all pairwise combinations, for each metric we include the bivariate combination with the highest relative information gain. For all metrics, the combination with the $AvgBitrate$ provides the highest bivariate information

gain. Also, even though $BufRatio$, $RateBuf$, and $RendQual$ had strong correlations in Figure 5(a), their combinations do not add much new information because they are inherently correlated.

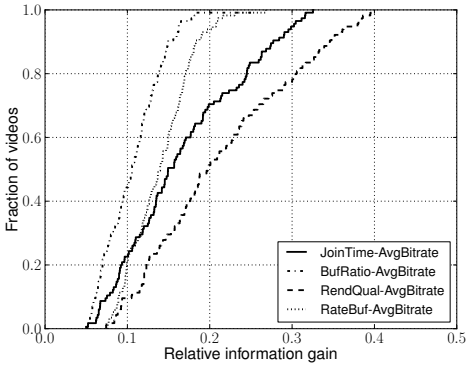


Figure 8: Distribution of the best bivariate relative information gains for $LvodA$

4.1.1 Strange behavior in $AvgBitrate$

Between Figures 5 and 6, we notice that $AvgBitrate$ is the metric with the weakest correlation but the second highest information gain. This observation is related to Figure 3 from Section 3. The relationship between $PlayTime$ and $AvgBitrate$ is not monotone; it shows a peak between the 800-1000 Kbps, is low on either side of this region, and increases slightly at the highest rate. Because of this non-monotone relationship, the correlation is low. However, knowing the value of $AvgBitrate$ allows us predict the $PlayTime$; there is a non-trivial information gain.

Now this explains why the information gain is high and the correlation is low, but does not tell us why the $PlayTime$ is low for the 1000-1600 Kbps band. The reason is that the values of bitrates in this range correspond to clients having to switch bitrates because of buffering induced by poor network conditions. Thus, the $PlayTime$ is low here mostly as a consequence of buffering, which we already observed to be the most critical factor. This also points out the need for robust bitrate selection and adaptation algorithms.

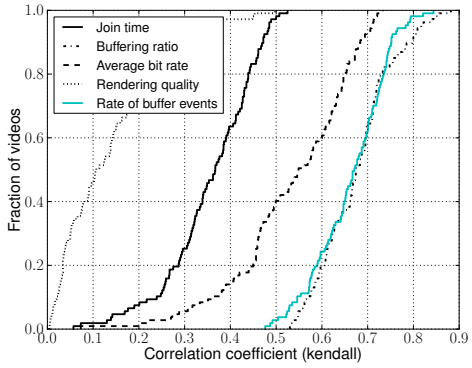
4.2 Live Content

Figure 9 shows the distribution of the correlation coefficients for dataset $LiveA$. The median values for the two datasets are summarized in Table 3. We notice one key difference with respect to the $LvodA$ results: $AvgBitrate$ is more strongly correlated for live content. Similar to dataset $LvodA$, $BufRatio$ is strongly correlated, while $JoinTime$ is weakly correlated.

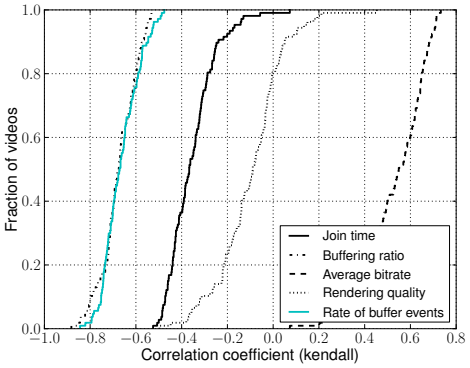
Quality metric	Correlation coefficient	
	$LiveB$	$LiveA$
$JoinTime$	-0.49	-0.36
$BufRatio$	-0.81	-0.67
$RendQual$	-0.16	-0.09

Table 3: Median values of the Kendall rank correlation coefficients for $LiveA$ and $LiveB$. We do not show $AvgBitrate$ and $RateBuf$ because they do not apply to $LiveB$. For the remaining metrics the results are consistent with dataset $LiveA$.

For both long VoD and live content, $BufRatio$ is a critical metric. Interestingly, for live, we see that $RateBuf$ has a much stronger negative correlation with $PlayTime$. This suggests that the Live users are *more sensitive* to each buffering event compared to the



(a) Absolute values



(b) Actual values (signed)

Figure 9: Distribution of the Kendall rank correlation coefficient between the quality metrics and play time for $LiveA$.

Long VoD audience. Investigating this further, we find that the average buffering duration is much smaller for long VoD (3 seconds), compared to live (7s), i.e., each buffering event in the case of live content is more disruptive. Because the buffer sizes in long VoD are larger, the system fares better in face of fluctuations in link bandwidth. Furthermore, the system can be more proactive in predicting buffering and hence preventing it by switching to another server, or switching bitrates. Consequently, there are fewer and shorter buffering events for long VoD. For live, on the other hand, the buffer is shorter, to ensure that the stream is current. As a result, the system is less able to proactively predict throughput fluctuations, which increases both the number and the duration of buffering events. Figure 10 further confirms that $AvgBitrate$ is a critical metric and that $JoinTime$ is less critical for Live content. The bivariate results (not shown for brevity) mimic the same effects from Figure 8, where the combination with $AvgBitrate$ provides the best information gains.

4.2.1 Why is $RendQual$ negatively correlated?

We noticed an anomalous behavior for $PlayTime$ vs. $RendQual$ for live content in Figure 4(d). The previous results from both $LiveA$ and $LiveB$ datasets further confirm that this is not an anomaly specific to the video shown earlier, but a more pervasive phenomenon in live content.

To illustrate why this negative correlation arises, we focus on the relationship between the $RendQual$ and $PlayTime$ for a particular live video in Figure 11. We see a surprisingly large fraction of viewers with low rendering quality and high play time. Further, the

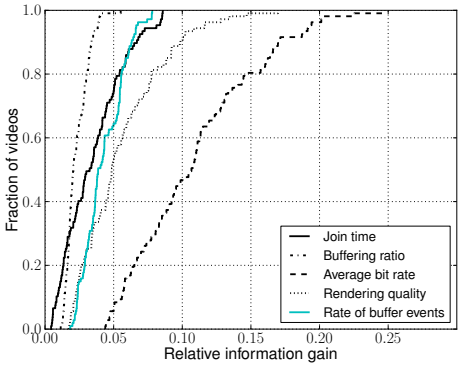


Figure 10: Distribution of the univariate gain between the quality metrics and play time for $LiveA$.

$BufRatio$ values for these users is also very low. In other words, these users have no network issues, but see a drop in $RendQual$, but continue to watch the video for a long duration despite this poor frame rate.

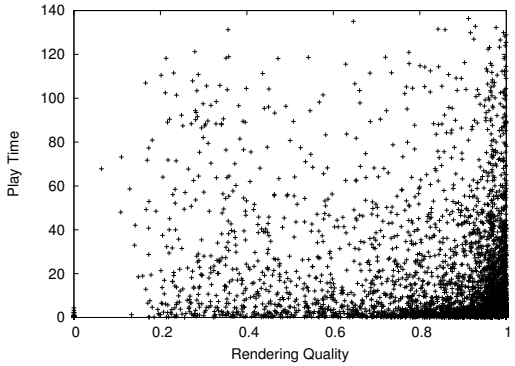


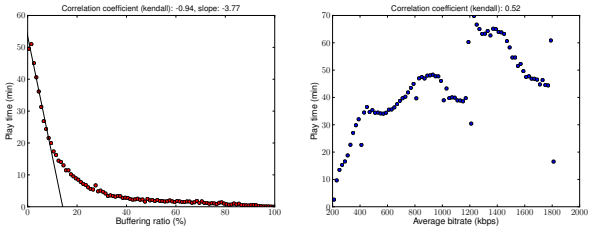
Figure 11: Scatter plot between the play time and rendering quality. Notice that there are a lot of points where the rendering quality is very low but the play time is very high.

We speculate that this counter-intuitive negative correlation between $RendQual$ and $PlayTime$ arises out of a combination of two effects. The first effect has to do with user behavior. Unlike long VoD viewers (e.g., TV episodes), live video viewers are also likely to run the video player in background (e.g., listening to the sports commentary). In such situations the browser is either minimized or the player is in a hidden browser tab. The second effect is an optimization by the player to reduce the CPU consumption when the video is being played in the background. In these cases, the player decreases the frame rendering rate to reduce CPU use. We replicated the above scenarios–minimizing the browser or playing a video in a background window–in a controlled setup and found that the player indeed drops the $RendQual$ to 20% (e.g., rendering 6-7 out of 30 frames per second). Curiously, the $PlayTime$ peak in Figure 4(d) also occurs at a 20% $RendQual$. These controlled experiments confirm our hypothesis that the anomalous relationship is in fact due to these player optimizations for users playing the video in the background.

4.2.2 Case study with high impact events

A particular concern for live content providers is whether the observations from typical events can be applied to *high impact* events [22]. To address this concern, we consider the $LiveH$ dataset.

Because the data collected during the corresponding period of time does not provide the $RendQual$ and $RateBuf$, we only focus on $BufRatio$ and $AvgBitrate$, which we observed as the most critical metrics for live content in the previous discussion. Figures 12(a) and 12(b) show that the trends and correlation coefficients for $LiveH1$ match closely with the results for datasets $LiveA$ and $LiveB$. We also confirmed that the values for $LiveH2$ and $LiveH3$ are almost identical to $LiveH1$; we do not show these for brevity. These results, though preliminary, suggest that our observations apply to such singular events as well.



(a) $BufRatio$ (b) $AvgBitrate$

Figure 12: Impact of two quality metrics for $LiveH1$, one of the three final games from the 2010 FIFA World Cup. A linear data fit is shown over the 0-10% subrange of $BufRatio$. The results for $LiveH2$ and $LiveH3$ are almost identical and not shown for brevity.

With respect to the average bitrate, the play time peaks around a bitrate of 1.2 Mbps. Beyond that value, however, the engagement decreases. The reason for this behavior is similar to the previous observation in Section 4.1.1. Most end-users (e.g., DSL, cable broadband users) cannot sustain such a high bandwidth stream. As a consequence, the player encounters buffering and also switches to a lower bitrate midstream. As we already saw, buffering adversely impacts the user experience.

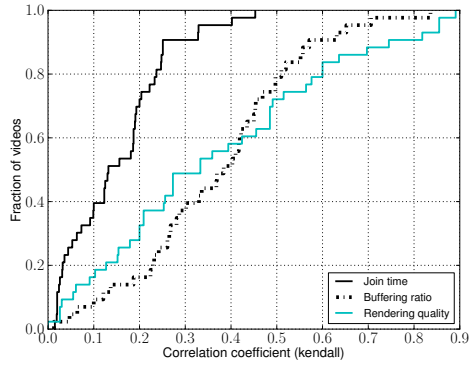
Quality metric	Correlation coefficient	
	$SvodB$	$SvodA$
$JoinTime$	0.06	0.12
$BufRatio$	-0.53	-0.38
$RendQual$	0.34	0.33

Table 4: Median values of the Kendall rank correlation coefficients for $SvodA$ and $SvodB$. We do not show $AvgBitrate$ and $RateBuf$ because the player did not switch bitrates and did not gather buffering event data. The results are consistent with $SvodA$.

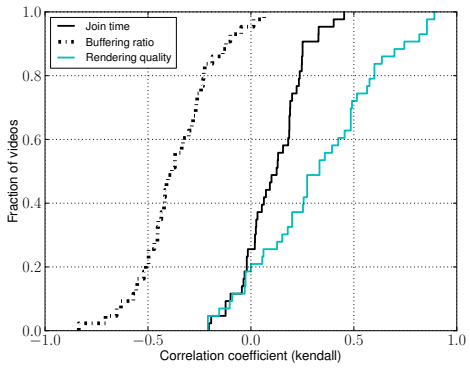
4.3 Short VoD Content

Finally, we consider the short VoD category. For both datasets $SvodA$ and $SvodB$ the player uses a discrete set of 2-3 bitrates (without switching) and was not instrumented to gather buffering event data. Thus, we do not show the $AvgBitrate$ (it is meaningless to compute the correlation on 2 points) and $RateBuf$. Figure 13 shows the distribution of the correlation coefficients for $SvodA$ and Table 4 summarizes the median values for both datasets.

We notice similarities between long and short VoD: $BufRatio$ and $RendQual$ are the most critical metrics that impact $PlayTime$. Further, $BufRatio$ and $RendQual$ are themselves strongly correlated (not shown). As before, $JoinTime$ is weakly correlated. For brevity, we do not show the univariate/bivariate information gain results for short VoD because they mirror the results from the correlation analysis.



(a) Absolute values



(b) Actual values (signed)

Figure 13: Distribution of the Kendall rank correlation coefficient between the quality metrics and play time for $SvodA$. We do not show $AvgBitrate$ and $RateBuf$ because the player did not switch bitrates and did not gather buffering event data.

4.4 Quantitative Impact

As we discussed earlier and as our measurements so far highlight, the interaction between the $PlayTime$ and the quality metrics can be quite complex. Thus, we avoid blindly applying quantitative regression models on our dataset. Instead, we only apply regression when we can visually confirm that this has a meaningful real-world interpretation and when the relationship is roughly linear. Thus, we restrict this analysis to the most critical metric, $BufRatio$. Further, we only apply regression to the 0-10% range of $BufRatio$, where we confirmed a simple linear relationship.

We notice that the distribution of the linear-fit slopes are very similar within the same content type in Figure 14. The median magnitudes of the slopes are one for long VoD, two for live, and close to zero for short VoD. That is, $BufRatio$ has the strongest quantitative impact on live, then on long VoD, then on short VoD.

Figure 12(a) also includes linear data fits on the 0-10% subrange for $BufRatio$ for the $LiveH$ data. These show that, within the selected subrange, a 1% increase in $BufRatio$ can reduce the average play time by more than *three minutes* (assuming a game duration of 90 minutes). Conversely, providers can increase the average user engagement by more than three minutes by investing resources to reduce $BufRatio$ by 1%.

4.5 Summary of view-level analysis

The key observations from the view-level analysis are:

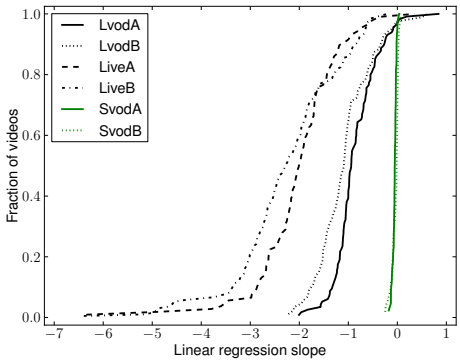


Figure 14: CDF of the linear-fit slopes between $PlayTime$ and the 0-10% subrange of $BufRatio$.

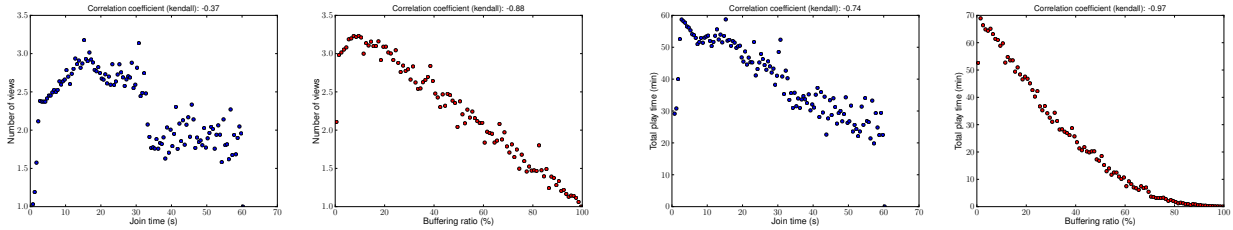
- For long and short VoD content, $BufRatio$ is the most important quality metric.
- For live content, $AvgBitrate$ in addition to $BufRatio$ is a key quality metric. Additionally, the requirement of small buffer for live videos exacerbates buffering events.
- A 1% increase in $BufRatio$ can decrease 1 to 3 minutes of viewing time.
- $JoinTime$ has significantly lower impact on view-level engagement than the other metrics.
- Finally, our analysis for the negative correlation of $RendQual$ in live video highlights the need to put the statistics in the context of actual user and system behavior.

5. VIEWER LEVEL ENGAGEMENT

Content providers also want to understand if good video quality improves customer retention or if it encourages users to try more videos. To address these questions we analyze the user engagement at the viewer level in this section. For brevity, we highlight the key results and do not duplicate the full analysis as in the previous section.

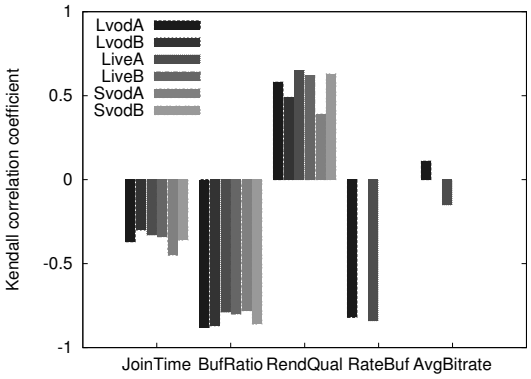
For this analysis, we look at the *number of views* per viewer and the *total play time* aggregated over all videos watched by the viewer in a one week interval. Recall that at the view level we filtered the data to only look at videos with at least 1000 views. At the viewer level, however, we look at the aggregate number of views and play time per viewer across all objects irrespective of that video's popularity. For each viewer we correlate the average of each quality metric with the two engagement metrics.

Figure 15 visually confirms that the quality metrics also impact the number of views. One curious observation is that the number of views increases in the range 1–15 seconds before starting to decrease. We also see a similar effect for $BufRatio$, where the first few bins have fewer total views. This effect does not, however, occur for the total play time. We speculate that this is an effect of user interest. Many users have very good quality but little interest in the content; they "sample" the content and leave without returning. Users who are actually interested in the content are more tolerant of longer join times (and buffering). However, the tolerance drops beyond a certain point (around 15 seconds for $JoinTime$). Figure 16 summarizes the values of the correlation coefficients for the six datasets. The values are qualitatively consistent across the different datasets and also similar to the trends we observed at the view level. One significant difference is that while $JoinTime$ is uninteresting at the view level, it has a more pronounced impact

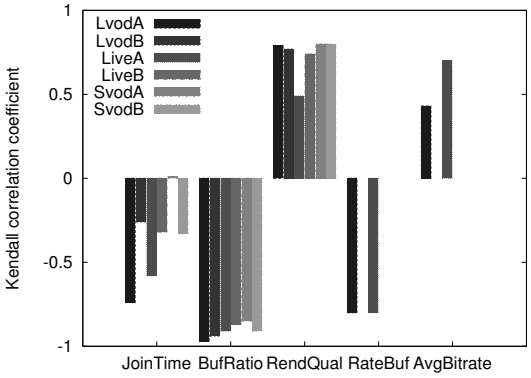


(a) Join time, # views (b) Buffering ratio, # views (c) Join time, Play time (d) Buffering ratio, Play time

Figure 15: Visualizing the impact of $JoinTime$ and $BufRatio$ on the number of views and play time for $LvodA$



(a) Number of views

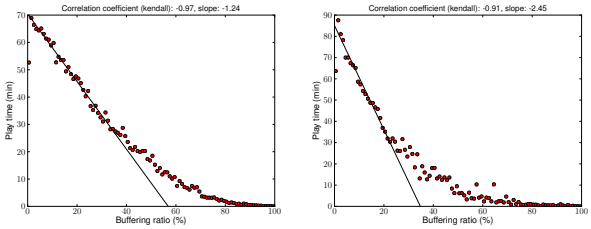


(b) Total play time

Figure 16: Viewer-level correlations w.r.t the number of views and play time. $AvgBitrate$ and $RateBuf$ values do not apply for $LvodB$,$LiveB$,$SvodA$, and $SvodB$.

on the total play time at the viewer level. This has interesting system design implications. For example, consider a scenario where a provider decides to increase the buffer size to alleviate the buffering issues. However, increasing buffer size can increase join time. The above result shows that doing so without evaluating the impact at the viewer level may be counterproductive, as increasing the buffer size may reduce the likelihood of a viewer visiting the site again.

As with the view-level analysis, we complement the qualitative correlations with quantitative results. Figure 17 shows linear data fitting for the total play time as a function of the buffering ratio for $LvodA$ and $LiveA$. This shows that reducing $BufRatio$ by 1% translates to an effective increase the total play time by 1.2 minutes for long VoD content and by 2.4 minutes for live content on average per user.



(a) Long VoD (b) Live

Figure 17: Linear data fitting between the buffering ratio and the total play time, for datasets $LvodA$ and $LiveA$.

Summary of viewer-level analysis:

- Both the number of views and the total play time are impacted by the quality metrics.
- The quality metrics that impact the view level engagement consistently impact the viewer level engagement. We confirm that these results are consistent across different datasets.
- The correlation between the engagement metrics and the quality metrics becomes visually and quantitatively even more striking at the viewer level.
- Additionally, the join time, which seemed less relevant at the view level, has non-trivial impact at the viewer level.

6. DISCUSSION

The findings presented in this paper are the result of an iterative process that included more false starts and misleading interpretations than we care to admit. We present two of the main lessons we learned in this process. Then, we discuss an important direction of future research for Internet video measurement.

6.1 The need for complementary analysis

> *All of you are right. The reason every one of you is telling it differently is because each one of you touched a different part of the elephant. So, actually the elephant has all the features you mentioned. [10]*

For the Long VoD case, we observed that the correlation coefficient for the average bitrate was weak, but the univariate information gain was high. The process of trying to explain this discrepancy led us to visualize the behaviors similar to Figure 3(c). In this case, the correlation was weak because the relationship was non-monotone. The information gain, however, was high because the intermediate bins near the natural modes had significantly lower engagement and consequently low entropy in the play time distribution.

This observation guided us to a different phenomenon, sessions that were forced to switch rates because of poor network quality.

If we had restricted ourselves to a purely correlation-based analysis, we may have missed this effect and incorrectly inferred that $AvgBitrate$ was not important. This highlights the value of using multiple *views* from complementary analysis techniques in dealing with large datasets.

6.2 The importance of context

Lies, damned lies, and statistics

Our second lesson is that while statistical data mining techniques are excellent tools, they need to be used with caution and with a judicious appreciation of the context in which they are applied. That is, we need to take the results of these analysis together with the context of the *human* and *operating* factors. For example, naively acting on the observation that the rendering quality is negatively correlated for live content can lead to an incorrect understanding of its impact on engagement. As we saw, this negative correlation is the outcome of both user behavior and player optimizations. Users who intend to watch a live event for a long time may run these in background windows; the player cognizant of this background window effect tries to reduce CPU consumption by reducing the rendering quality. This highlights the importance of backing the statistical analysis with a more in-depth domain knowledge and controlled experiments in replicating the observations.

6.3 Toward a video quality index

Our ultimate vision is to use measurement-driven insights to develop an *empirical Internet video quality index*, analogous to the notion of mean opinion scores and subjective quality indices [8, 9]. Given that there are multiple quality metrics, it is difficult for video providers and consumers to objectively compare different video services. At the same time, the lack of a concrete metric makes it difficult for delivery infrastructures and researchers to focus their efforts. If we can derive such a quality index, content providers and consumers can use it to choose delivery services while researchers and delivery providers can use it to guide their efforts for developing better algorithms for video delivery and adaptation.

However, as our measurements and lessons show the interactions between the quality metrics and engagement can be complex, interdependent, and counterintuitive even for a somewhat simplified view with just three content types and five quality metrics. Furthermore, there are other dimensions that we have not explored rigorously in this paper. For example, we considered three broad genres of content: Live, Long VoD, and Short VoD. It would also be interesting to analyze the impact of quality for other aspects of content segmentation. For example, is popular content more/less likely to be impacted by quality or is the impact likely to differ depending on the types of events/videos (e.g., news vs. sports vs. sitcoms)? Our preliminary results in these directions show that the magnitude of quality impact is marginally higher for popular videos but largely independent of content genres. Similarly, there are other fine-grained quality measures which we have not explored. For example, anecdotal evidence suggests that temporal effects can play a significant role; buffering during the early stages or a sequence of buffering events are more likely to lead to user frustration. Working toward such a unified quality index is an active direction of future research.

7. RELATED WORK

Content popularity: There is an extensive literature on modeling content popularity and its subsequent implications for caching (e.g., [15, 18, 22, 24, 26]). Most of these focus on the heavy-tailed nature of the access popularity distribution and its system-level impact. Our work on analyzing the interplay between quality and engagement is orthogonal to this extensive literature. One interesting question (that we address briefly) is to analyze if the impact of quality is different across different popularity segments. For example, providers may want to know if niche video content is more or less likely to be impacted by poor quality.

User behavior: Yu *et al.* present a measurement study of a VoD system deployed by China Telecom [24] focusing on modeling user arrival patterns and session lengths. They also observe that many users actually have small session times, possibly because many users just "sample" a video and leave if the video is of no interest. Removing the potential bias from this phenomenon was one of the motivations for our binned correlation analysis in Section 3. Other studies of user behaviors also have significant implications for VoD system design. For example, there are measurement studies of channel switching dynamics in IPTV systems (e.g., [19]), and understanding seek-pause-forward behaviors in streaming systems (e.g., [16]). As we mentioned in our browser minimization example for live video, understanding the impact of such behavior is critical for putting the measurement-driven insights in context.

P2P VoD: In parallel to the reduction of content delivery costs, there have also been improvements in building robust P2P VoD systems that can provide performance comparable to a server-side infrastructure at a fraction of the deployment cost (e.g., [14, 25, 26, 34]). Because these systems operate in more dynamic environments (e.g., peer churn, low upload bandwidth), it is critical for them to optimize judiciously and improve the quality metrics that really matter. While our measurements are based on a server-hosted infrastructure for video delivery, the insights in understanding the most critical quality metrics can also be used to guide the design of P2P VoD systems.

Measurements of deployed video delivery systems: The networking community has benefited immensely from measurement studies of *deployed* VoD and streaming systems using both "black-box" inference (e.g., [12, 25, 33, 35]) and "white-box" measurements (e.g., [22, 27, 30, 37]). Our work follows in this rich tradition of providing insights from real deployments to improve our understanding of Internet video delivery. At the same time, we believe that we have taken a significant step forward in qualitatively and quantitatively measuring the impact of the video quality on user engagement.

User perceived quality: There is prior work in the multimedia literature on metrics that can capture user perceived quality (e.g., [23, 38]) and how specific metrics affect the user experience (e.g., [20]). Our work differs on several key fronts. The first is simply an issue of timing and scale. Internet video has only recently attained widespread adoption and revisiting user engagement is ever more relevant now than before. Prior work depend on small-scale experiments with a few users, while our study is based on real-world measurements with millions of viewers. Second, these fall short of linking the perceived quality to the actual user engagement. Finally, a key difference is with respect to methodology; user studies and opinions are no doubt useful, but difficult to objectively evaluate. Our work is an empirical study of engagement in the wild.

Engagement in other media: The goal of understanding user engagement appears in other content delivery mechanisms as well. The impact of page load times on user satisfaction is well known (e.g., [13, 21, 28]). Several commercial providers measure the impact of page load times on user satisfaction (e.g., [6]). Chen *et al.* study the impact of quality metrics such as bitrate, jitter, and delay

on call duration in Skype [11] and propose a composite metric to quantify the combination of these factors. Given that Internet video has become mainstream only recently, our study provides similar insights for the impact of video quality on engagement.

Diagnosis: In this paper, we focused on measuring the quality metrics and how they impact user engagement. A natural follow up question is whether there are mechanisms to pro-actively diagnose quality issues to minimize the impact on users (e.g., [17, 31]). We leave this as a direction for future work.

8. CONCLUSIONS

As the costs of video content creation and dissemination continue to decrease, there is an abundance of video content on the Internet. Given this setting, it becomes critical for content providers to understand if and how video quality is likely to impact user engagement. Our study is a first step towards addressing this goal.

We present a systematic analysis of the interplay between three dimensions of the problem space: quality metrics, content types, and quantitative measures of engagement. We study industry-standard quality metrics for Live, Long VoD, and Short VoD content to analyze engagement at per view and viewer-level.

Our key takeaways are that at the view-level, buffering ratio is the most important metric across all content genres and the bitrate is especially critical for Live (sports) content. Additionally, we find that the join time becomes critical in terms of the viewer-level engagement and thus likely to impact customer retention.

These results have key implications both from commercial and technical perspectives. In a commercial context, they inform the policy decisions for content providers to invest their resources to maximize user engagement. At the same time, from a technical perspective, they also guide the design of the technical solutions (e.g., tradeoffs in the choice of a suitable buffer size) and motivate the need for new solutions (e.g., better pro-active bitrate selection, rate switching, and buffering techniques).

In the course of our analysis, we also learned two cautionary lessons that more broadly apply to measurement studies of this nature: the importance of using multiple complementary analysis techniques when dealing with large datasets and the importance of backing these statistical techniques with system-level and user context. We believe our study is a significant step toward an ultimate vision of developing a unified quality index for Internet video.

Acknowledgments

We thank our shepherd Ratul Mahajan and the anonymous reviewers for their feedback that helped improve this paper. We also thank other members of the Conviva staff for supporting the data collection infrastructure and for patiently answering our questions regarding the player instrumentation and datasets.

9. REFERENCES

[1] Alexa Top Sites. http://www.alexa.com/topsites/countries/US.

[2] Cisco forecast. http://blogs.cisco.com/sp/comments/cisco_visual_networking_index_forecast_annual_update/.

[3] Driving Engagement for Online Video. http://events.digitallyspeaking.com/akamai/mddec10/post.html?hash=ZDlBSGhsMXBidnJ3RXNWSW5mSE1HZz09.

[4] Hadoop. http://hadoop.apache.org/.

[5] Hive. http://hive.apache.org/.

[6] Keynote systems. http://www.keynote.com.

[7] Mail service costs Netflix 20 times more than streaming. http://www.techspot.com/news/42036-mail-service-costs-netflix-20-times-more-than-streaming.html.

[8] Mean opinion score for voice quality. http://www.itu.int/rec/T-REC-P.800-199608-I/en.

[9] Subjective video quality assessment. http://www.itu.int/rec/T-REC-P.910-200804-I/en.

[10] The tale of three blind men and an elephant. http://en.wikipedia.org/wiki/Blind_men_and_an_elephant.

[11] K. Chen, C. Huang, P. Huang, C. Lei. Quantifying Skype User Satisfaction. In *Proc. SIGCOMM*, 2006.

[12] Phillipa Gill, Martin Arlitt, Zongpeng Li, Anirban Mahanti. YouTube Traffic Characterization: A View From the Edge. In *Proc. IMC*, 2007.

[13] 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.

[14] B. Cheng, L. Stein, H. Jin, and Z. Zheng. Towards Cinematic Internet Video-On-Demand. In *Proc. Eurosys*, 2008.

[15] B. Cheng, X. Liu, Z. Zhang, and H. Jin. A measurement study of a peer-to-peer video-on-demand system. In *Proc. IPTPS*, 2007.

[16] C. Costa, I. Cunha, A. Borges, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto. Analyzing Client Interactivity in Streaming Media. In *Proc. WWW*, 2004.

[17] C. Wu, B. Li, and S. Zhao. Diagnosing Network-wide P2P Live Streaming Inefficiencies. In *Proc. INFOCOM*, 2009.

[18] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. IMC*, 2007.

[19] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain. Watching Television Over an IP Network. In *Proc. IMC*, 2008.

[20] M. Claypool and J. Tanner. The effects of jitter on the peceptual quality of video. In *Proc. ACM Multimedia*, 1999.

[21] D. Galletta, R. Henry, S. McCoy, and P. Polak. Web Site Delays: How Tolerant are Users? *Journal of the Association for Information Systems*, (1), 2004.

[22] H. Y. et al. Inside the Bird's Nest: Measurements of Large-Scale Live VoD from the 2008 Olympics. In *Proc. IMC*, 2009.

[23] S. R. Gulliver and G. Ghinea. Defining user perception of distributed multimedia quality. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(4), Nov. 2006.

[24] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding User Behavior in Large-Scale Video-on-Demand Systems. In *Proc. Eurosys*, 2006.

[25] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross. A measurement study of a large-scale P2P IPTV system. *IEEE Transactions on Multimedia*, 2007.

[26] Y. Huang, D.-M. C. Tom Z. J. Fu, J. C. S. Lui, and C. Huang. Challenges, Design and Analysis of a Large-scale P2P-VoD System. In *Proc. SIGCOMM*, 2008.

[27] W. W. Hyunseok Chang, Sugih Jamin. Live Streaming Performance of the Zattoo Network. In *Proc. IMC*, 2009.

[28] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson, and B. Shneiderman. Determining Causes and Severity of End-User Frustration. In *International Journal of Human-Computer Interaction*, 2004.

[29] K. Cho, K. Fukuda, H. Esaki. The Impact and Implications of the Growth in Residential User-to-User Traffic. In *Proc. SIGCOMM*, 2006.

[30] K. Sripanidkulchai, B. Maggs, and H. Zhang. An Analysis of Live Streaming Workloads on the Internet. In *Proc. IMC*, 2004.

[31] 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.

[32] T. Mitchell. *Machine Learning*. McGraw-Hill.

[33] S. Ali, A. Mathur, and H. Zhang. Measurement of Commercial Peer-to-Peer Live Video Streaming. In *Proc. Workshop on Recent Advances in Peer-to-Peer Streaming*, 2006.

[34] S. Guha, S. Annapureddy, C. Gkantsidis, D. Gunawardena, and P. Rodriguez. Is High-Quality VoD Feasible using P2P Swarming? In *Proc. WWW*, 2007.

[35] S. Saroiu, K. P. Gummadi, R. J. Dunn, S. D. Gribble, and H. M. Levy. An Analysis of Internet Content Delivery Systems. In *Proc. OSDI*, 2002.

[36] H. A. Simon. Designing Organizations for an Information-Rich World. Martin Greenberger, Computers, Communication, and the Public Interest, The Johns Hopkins Press.

[37] K. Sripanidkulchai, A. Ganjam, B. Maggs, and H. Zhang. The Feasibility of Supporting Large-Scale Live Streaming Applications with Dynamic Application End-Points. In *Proc. SIGCOMM*, 2004.

[38] K.-C. Yang, C. C. Guest, K. El-Maleh, and P. K. Das. Perceptual Temporal Quality Metric for Compressed Video. *IEEE Transactions on Multimedia*, Nov. 2007.