Shedding Light on the Structure of Internet Video Quality Problems in the Wild

Junchen Jiang [*] , Vyas Sekar [†] , Ion Stoica [††] [◦], Hui Zhang [*] [◦]
[*] Carnegie Mellon, [†] Stony Brook University, [††] UC Berkeley, [◦] Conviva

ABSTRACT

The key role that video quality plays in impacting user engagement, and consequently providers' revenues, has motivated recent efforts in improving the quality of Internet video. This includes work on adaptive bitrate selection, multi-CDN optimization, and global control plane architectures. Before we embark on deploying these designs, we need to first understand the nature of video of quality problems to see if this complexity is necessary, and if simpler approaches can yield comparable benefits.

To this end, this paper is a first attempt to shed some light on the structure of video quality problems. Using measurements from 300 million video sessions over a two-week period, we identify recurrent problems using a hierarchical clustering approach over the space of client/session attributes (e.g., CDN, AS, connectivity). Our key findings are that: (1) a small number (2%) of critical clusters account for 83% of join failure problems (44–84% for other metrics); (2) many problem events (50%) persist for at least 2 hours; (3) a majority of these problems (e.g., 60% of join failures, 30–60% for other metrics) are related to content provider, CDN, or client ISP issues. Building on these insights, we evaluate the potential improvement by focusing on addressing these recurrent problems and find that fixing just 1% of these clusters can reduce the number of problematic sessions by 55% for join failures (15%–40% for other metrics).

Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed systems—*Distributed applications*; C.4 [**Performance of Systems**]: [measurement techniques]

General Terms

Performance, Measurement

1. INTRODUCTION

With the rapid growth of Internet video, content providers and delivery systems are constantly striving to deliver a higher quality of viewing experience to users. Delivering high quality content is especially critical given the subscription- and advertisement-driven revenue models that have driven this growth in recent years. Several independent studies have confirmed what industry practitioners have known implicitly for several years—quality impacts user engagement and determines the likelihood of users viewing the content and returning to the content providers [13, 1, 19].

Motivated by these observations, there is a growing body of recent work that aims to improve the video delivery quality. This includes work on developing better bitrate adaptation and transport-layer algorithms (e.g., [8, 16]); better CDN and server selection strategies (e.g., [4, 5]); the use of multi-CDN optimizations (e.g., [25, 20]); and even cases made for global control plane solutions that take a centralized approach to optimize video quality [21].

While these efforts are valuable and demonstrate potential improvements, a key missing piece is an understanding of the structure of Internet video quality problems. More specifically, given that some of these aforementioned efforts require significant deployment effort, it is important to analyze if the complexity envisioned by them (e.g., global coordination or TCP fixes) is really necessary or if most of these improvements can be achieved with simpler alternatives. For instance, if there are a common set of recurrent problems and root causes (e.g., specific ISPs or CDNs or content providers) that account for most of the observed problems, then we can improve the aggregate quality by focusing our efforts on these few "bad apples" rather than a wholesale change.

This paper is a first step to understand the structure of the observed quality problems in the wild and thus bridge this gap in our understanding of Internet video quality. To this end, we use a dataset of observed video quality collected over 300 million sessions across a diverse set of 379 content providers (i.e., video hosting sites) with viewers distributed across 213 countries. This provides us a unique opportunity to obtain a *panoramic view* across a wide spectrum of content providers, content delivery networks (e.g., traditional CDNs vs. data center CDNs), user viewing platforms (e.g., mobile vs. desktop vs. TV set-top boxes), and content genres (e.g., live vs. VoD). This is especially relevant as this means that our observations are not tightly coupled to a single content provider or content delivery system, which has been a perceived drawback of prior large-scale measurement studies (e.g., [22]).

Using this dataset, we identify video viewing sessions suffering quality issues (*problem sessions*) with respect to four key video quality metrics: buffering, bitrate, join time (delay in loading the video), and join failures (video failed to load), and we study these metrics independently. We group together problem sessions that share one or more client or session attributes (e.g., ISP, CDN, content provider, connection type, user platform) into *problem clusters*. Intuitively, these clients potentially share some common underlying phenomenon that led to these sessions manifesting as problem ses-

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.
CoNEXT'13, December 9-12, 2013, Santa Barbara, California, USA.
Copyright 2013 ACM 978-1-4503-2101-3/13/12 ...$15.00.
http://dx.doi.org/10.1145/2535372.2535394 ...$15.00.

sions. Starting from the problem clusters, we also identify *critical clusters* that represent likely root causes of the vast majority of the problem clusters. For example, a specific poorly performing ISP may manifest as several distinct problem clusters each with a different CDN, but the underlying problem can be logically attributed to the ISP.

Our key observations are:

- The problem clusters show a natural skewed distribution in terms of (a) *prevalence:* more than 20% of the problem clusters appear for more than 25% of the time; and (b) *persistence:* more than 50% of the problem clusters last for more than 2 hours.
- A small number of critical clusters (50× lower than the number of problem clusters) can account for up to 74% of the observed problem sessions.
- While the types of attributes (e.g., provider, ISP) defining the dominant critical clusters are common across different quality metrics, the specific content providers or ISPs that exhibit problems are different across the quality metrics.
- Analyzing the most prevalent critical clusters, we see interesting patterns with a few less-provisioned content providers that do not offer multiple bitrates, ISPs in non-US regions, and users viewing video from mobile wireless (i.e., 3G or 4G) connections.

Motivated by these observations, we analyze the potential improvement if we focus our efforts on addressing problems associated with the few critical clusters and find that:

- Across all quality metrics, a *proactive* approach, which focuses on "fixing" the poor performance of the top 1% of the critical clusters, can alleviate up to 58% of all problem sessions.
- Even a *reactive* approach, which detects persistent critical clusters after they occur and addresses them, can reduce the number of problem session by up to 51%.

These results have important (and positive) implications for improving the current state of Internet video quality. In some sense, we can have considerable benefits in globally improving the state of video quality by focusing on these problems which are amenable to simple (and well known) solutions rather than embarking on wholesale deployment efforts. For instance, the problems associated with non-US users may be alleviated by contracting with local CDN operators. Similarly, simple solutions such as offering a more fine-grained selection of bitrates can alleviate issues w.r.t. specific providers and mobile users. We do acknowledge that our analysis does not provide a prescriptive alleviation strategy or perform a cost-benefit analysis; the goal of this paper is to quantify the number of quality problems that are potentially amenable to proactive or reactive strategies.

In the rest of the paper, we begin by describing our dataset in Section 2 and methodology in Section 3. We analyze the spatial and temporal properties of the problem clusters and critical clusters in Section 4 and the potential for improvement in Section 5. We discuss directions for extending our analysis in Section 6 and related work in Section 7, before concluding in Section 8.

2. DATASET AND MOTIVATION

In this section, we begin by describing our dataset. We also present preliminary statistics to motivate the types of structural insights we would like to obtain regarding the nature of video quality problems. As an ongoing effort, we are working on anonymizing the relevant data for releasing of the dataset used in this work.

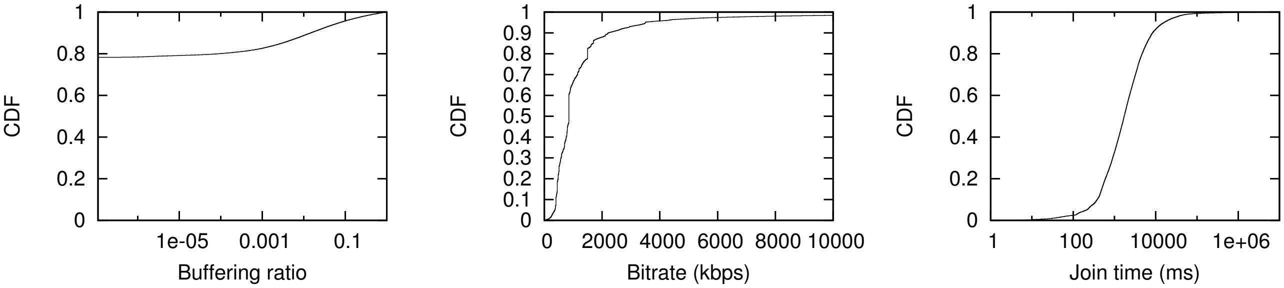
Dataset: Our dataset is based on client-side measurements of video quality from over 300 million sessions over a duration of two weeks. The unique feature of our dataset is that it is collected over 379 distinct content providers spanning diverse genres, both live and video-on-demand content, different content delivery platforms, different types of bitrate adaptation algorithms, and device/browser platforms. Though US viewers dominates the dataset (∼55%), there are a fair number of European (∼12%) and Chinese (∼8%) users in the dataset. This is especially relevant as it provides us with a panoramic view of state of Internet video delivery today.

The basic unit in our dataset is a *video session*. A session represents a user viewing a video on one of our affiliates' sites for some duration of time. Each session is associated with a set of *seven attributes*:

1. *ASN:* The Autonomous System Number (ASN) that the client IP belongs to. Note that a single ISP (e.g., Comcast) may own different ASNs both for management and business reasons. We focus on the ASN as it is more fine-grained than the ISP granularity. We observe in aggregate 15K unique ASNs spanning multiple countries.
2. *CDN:* In total, we observe 19 unique CDNs spanning popular CDN providers as well as several in-house and ISP-run CDNs. (Some providers use proprietary CDN switching logic; in this case we pick the segment of the session with the CDN used for the longest duration.)
3. *Content provider (Site):* This is the specific affiliate content provider from which the client requested some content. We have 379 content providers that span different genres of content. We use the terms site and content provide interchangeably.
4. *VoD or Live:* Video content falls in one of two categories: video-on-demand (VoD) or Live. We use a binary indicator to see if the particular content was a Live event or a VoD video.
5. *Player type:* We see diverse players such as Flash, Silverlight, and HTML5.
6. *Browser:* We see diverse client browsers including Chrome, Firefox, MSIE, and Safari.
7. *Connection type:* Finally, we have the type of access network connection such as mobile/fixed wireless, DSL, fiber-to-home. These annotations come from third party services [2].

We focus on *four key quality metrics* that are common across different content providers and have been shown to be critical for measuring quality as well as user engagement [13]:

1. *Buffering ratio:* Given a video session of duration T seconds, if the player spent B seconds in buffering (i.e., waiting for the player buffer to replenish midstream, the buffering ratio is defined as $\frac{B}{T}$. Prior work has shown that buffering ratio is a key metric that impacts user engagement [13].
2. *Join time:* This is the time taken for the video to start playing from the time the user clicks on the "play" button on the player. While join time may not directly impact the amount of a specific video viewed, it does have long term effects as it reduces the likelihood of repeated visits [13, 19].
3. *Average bitrate:* Many video players today support adaptive bitrate selection and midstream bitrate switching to adapt to changing bandwidth availability. The average bitrate of a session is simply the time-weighted average of the bitrates used in a given session. (Bitrate refers to the video playback rate, rather than throughput or download rate.)
4. *Join failures:* Some sessions may not even start playing the video; either the content is not available on the CDN server or the CDN is under overload or other unknown reasons. We mark



(a) Buffering ratio (b) Bitrate (c) Join time

Figure 1: ***CDF of observed quality metrics – buffering ratio, bitrate, and the join time. We see that a non-trivial number of sessions suffer quality problems. For instance, more than 5% of sessions have a buffering ratio larger than 10%.***

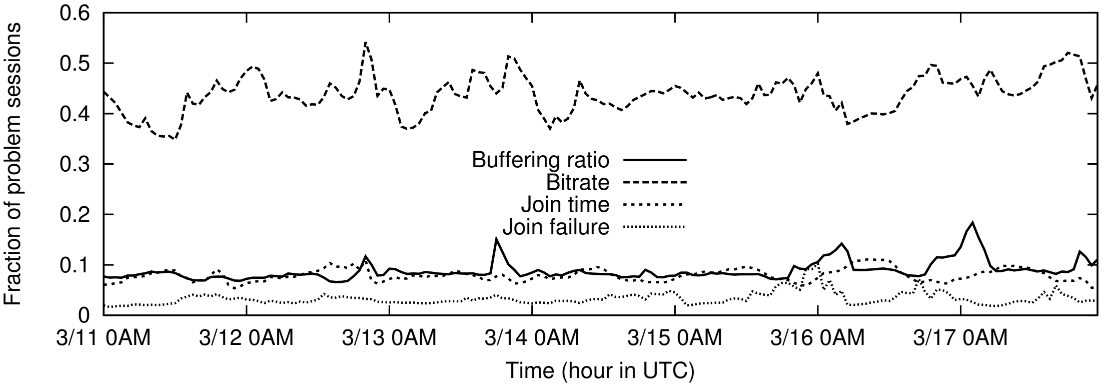


Figure 2: ***Timeseries of the fraction of problem sessions for the four quality metrics. The fraction of problem sessions is consistently high over time and the different quality metrics seem slightly uncorrelated.***

as a session as a join failure if no content was played during this session.[1]

Figure 1 shows the distribution of the first three quality metrics over the 1-week dataset. (Join failures are binary events; it is not meaningful to look at a distribution.) The result reconfirms prior observations that there are a non-trivial number of sessions with less-than-ideal quality [13, 21]. The key difference here is that these past efforts only considered a small set of 3–4 content providers. In contrast, we are considering the aggregate data from over 300 content providers. For instance, more than 5% of all sessions have a join time greater than 10 seconds; i.e., users had to wait for 10 seconds before the video even started playing! Similarly, more than 5% of sessions had a buffering ratio that was greater than 10%. This is particularly bad as past studies show that even a 1% increase in buffering ratio can lead to 3-4 minutes of lost viewership [13]. Finally, we also see that more than 80% of sessions observe an average bitrate less than 2 Mbps; i.e., less than the lower end of today's "HD" content.

Identifying problem sessions: Our focus is on understanding *quality problems* as they appear in the wild. To this end, we identify problem sessions w.r.t. each of the quality metrics. Note that a given session may appear as a problem session on a subset of metrics; i.e., it might have a low join time but may have a high buffering ratio or vice versa. We consider the metrics separately to avoid implicitly assuming that the metrics or failures are correlated.

- For join failures, we use a binary indicator if the session failed or not. For the remaining metrics, we choose specific thresholds based on domain-specific knowledge and observations in prior work. Our specific thresholds and rationale are follows.
- For buffering ratio, we identify a problem session if the value is greater than 5%; this is based on the observation that beyond this value there is a sharp decrease in amount of video viewed [13].
- For bitrate, we mark a problem session if the average bitrate is less than 700kbps; this value roughly corresponds to the recommended "360p" setting on video providers. We use a fixed threshold of bitrate in this work for simplicity, but we do acknowledge that bitrate settings are content-dependent (e.g., some contents do not provide high resolution streams).
- Third, we mark all sessions with a join time greater than 10 seconds; this represents a conservative upper bound on the tolerance of users [3, 19].

We do acknowledge that there is no ideal choice of threshold and it is likely that these thresholds will evolve as user expectations and network conditions improve. As such, the choice of thresholds is illustrative of the structure of video quality problems that occur today. The methodology and qualitative observations we present are not tied to the specific thresholds. We have confirmed that the results are qualitatively similar for other choices of these thresholds as well.

Aggregate statistics: Figure 2 shows a timeseries of the *fraction of problem sessions* or *problem ratio* over the week-long trace. Each

[1] Join failures are reported by the client-side measurement module that sends a "heartbeat" on the player status.

point here is the fraction of problem sessions on each metric for each hour-long epoch. First, the result shows that the fraction of quality problems is relatively consistent over time and the quality problems from the previous CDFs are not concentrated in time and are quite evenly spread out. For instance, for buffering ratio the average problem ratio is 0.097 per hour and the standard deviation is less than 10^{-3}. Second, we see that the different quality metrics do exhibit slightly different patterns of activity; we do not see a significant temporal correlation between metrics and we do see a small number of uncorrelated peaks.

Motivating questions: The above results reconfirm that quality problems exist and that the fraction of problem sessions is consistently high. These results raise several natural questions about the *structure* of these quality problems w.r.t. the different session attributes:

- Are the problems uniformly spread through the space of attribute combinations or are there specific combinations that have a higher concentration?
- While the fraction of problem sessions is relatively consistent, a natural question is if the "events" (e.g., outages with specific sites or CDNs) underlying these problems are also consistent?
- Is each problem a transient or one-off event for a specific ISP, CDN, or provider (or combination of these) or are these problems persistent over long periods?
- While the different quality metrics appear to be slightly temporally uncorrelated, a natural question is if these metrics are also structurally uncorrelated; e.g., do the same set of ISPs or CDNs contribute to the problematic sessions across all metrics?

These questions have key implications for video delivery design and optimization. For instance, recurrent problem events may guide us to *proactively* identify a few "bad apples" to improve the aggregate quality. Similarly, the duration of problem events has key implications for systems that attempt to *reactively* identify and alleviate quality problems [21]—do we have enough time to observe and react or are the problems very transient? Finally, the structural correlation across metrics has key implications in terms of the effort required to address the quality problems; i.e., will fixing one provider alleviate problems w.r.t. all metrics or do we need a multi-pronged approach catering to each metric independently?

3. METHODOLOGY

The previous section raises several natural questions regarding the *structure* of the problem events both across the space of client-side attributes as well as in time. In this section, we describe our basic data analysis building blocks.

3.1 Identifying Problem Clusters

We begin by dividing our dataset into discrete one hour epochs.[2] As a first step to analyze the structure, we *cluster*[3] together sessions that share one or more client/session attributes within the same epoch. For instance, the cluster "ASN=$ASN1$" describes all sessions where the user belongs to $ASN1$ and the cluster "ASN=$ASN1$, CDN=$CDN1$", describes all sessions where the user belongs to $ASN1$ and the session was assigned to a server from $CDN1$.

In order for our observations to be reliable, we want to focus on clusters that are deemed to be statistically significant sources of problem sessions. To this end, we define the problem ratio of a cluster as the ratio $\frac{\text{\# of problem sessions}}{\text{\# of sessions}}$. Then, we cull out the clusters whose problem ratio is significantly higher than the global average problem ratio. We also remove all clusters that have a small number of sessions in aggregate; i.e., problems observed within a small cluster may not be statistically significant. Combining these two steps, we define a problem cluster as a cluster that has a problem ratio $\geq 1.5\times$ the global problem ratio,[4] and the number of sessions in this cluster is ≥ 1000. In the rest of this paper, we start from the problem clusters as our basis and subsequently refine the analysis.

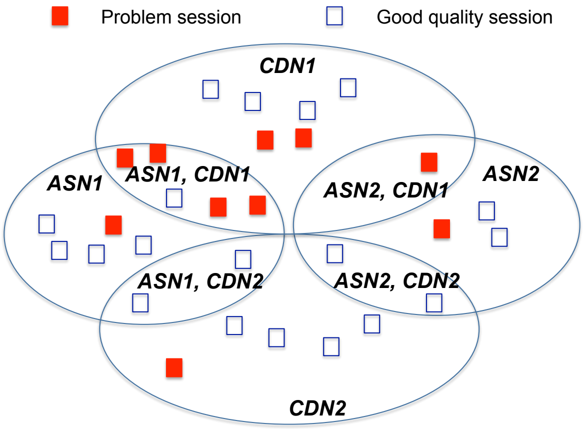


Figure 3: Simple example to illustrate the notion of problem clusters (with respect to one quality metric).

Figure 3 shows a simple example to illustrate the definition of problem clusters. Here, we have several sessions (the filled rectangles represent problem sessions and the non-filled rectangles are "good" sessions) spanning 2 ASNs and 2 CDNs. In this example, we have eight clusters in total, 2 each for ASN, CDN, and 2 each for ASN-CDN combinations. However, not all clusters are interesting for our analysis. For instance, the cluster "$CDN2$" has only one problem session out of 9 sessions, so we may remove this as it does not have significantly high problem ratio. Cluster "$ASN1$, $CDN2$", "$ASN2$, $CDN1$" and "$ASN2$, $CDN1$" each has a very small number of sessions, so we may also remove them as they are not significant enough.

Note that we are not simply identifying clusters that have high volume—we are identifying clusters of sufficient volume that have a problem ratio that is significantly higher than the global average. In fact, we found close to zero correlation between the size of a problem cluster and its problem ratio (not shown).

3.2 Identifying Critical Clusters

While the grouping of problem sessions into problem clusters provides some insights into the structure of problems, there is still one key missing aspect. Specifically, we may have different granularities of problem clusters that may be intrinsically related to the same underlying root cause. Thus, our next step is to refine these problem clusters to identify such potential causal structures across the problem sessions.

The set of all clusters can be viewed as a *hierarchical* structure across the space of client/session attributes with natural parent-child relationships. We can visualize these parent-child relationships as a DAG as shown in Figure 4. A cluster $C1$ is a parent

[2] One hour is the finest granularity of the dataset and thus we cannot analyze effects at smaller timescales.

[3] The term "cluster" represents a group of sessions that share common attributes, and it is indeed different from traditional clustering algorithms where a cluster can be a group of any data points.

[4] This value roughly represents two standard deviations away from the mean of the per-cluster problem ratio distribution.

of cluster $C2$, if the set of attributes defining the cluster $C1$ is a strict subset of that of $C2$. For instance, the cluster "$ASN1$" is a parent of the more specific clusters "$ASN1$, $CDN1$" and " $ASN1$, $CDN2$". Note that a single cluster may have multiple parents; e.g., "$ASN1$" and "$CDN1$" are parents of the cluster "$ASN1$, $CDN1$".

Our goal is to identify a small number of *critical clusters* that can potentially explain the occurrences of different problem clusters. In fact, it will be shown later that there is a relative small number of critical clusters that explain the occurrences of most problem clusters (see Table 1). In our example in Figure 4, intuitively we should pick the "$CDN1$" cluster rather than pick "$ASN1$, $CDN1$" and "$ASN2$, $CDN2$" clusters separately. Given that we do not have ground truth, critical clusters can serve as starting points for further investigation.

An intuitive criterion for identifying a critical cluster is analogous to the notion of the minimum description length (or Occam's razor) from the machine learning literature. Conceptually, we should pick the most compact description to explain an observation. Building on the above intuition, we can identify a critical cluster as consisting of the minimal set of attributes that when combined together can lead to significantly high problem ratio in its cluster (e.g., a problem cluster) and removing even one attribute from this set will reduce the problem ratio. To this end, we identify critical clusters using a *phase transition* algorithm as follows. For each session, we construct all logical paths in the DAG from the root to the leaf. Then, for each of these paths, we identify the point closest to the root along this path such that every cluster that is a descendant is a problem cluster and once removing it every cluster that is an ancestor is not a problem cluster.

We use Figure 5 to explain the intuition. In this figure, "$CDN1$, $ASN1$" is the critical cluster—every cluster that is a child of this combination is a problem cluster and if we remove the sessions in this combination, the parents "$CDN1$" and "$ASN1$" cease to be problem clusters. That is, this combination of attributes represents a key "transition point" in this hierarchy between problem clusters and non-problem clusters.

There are two subtle concerns with this algorithm. First, we may not be able to clearly identify such phase transition points if the data is quite noisy; i.e., we may not be able to attribute problem clusters to a specific critical cluster. Fortunately, as we will in the next section the coverage over problem sessions that appear in problem clusters is quite high. Second, there might be corner cases where we may find two potential phase transitions. This can happen if some of the attributes are themselves correlated; e.g., if a specific Site only uses a single CDN or most of its clients appear from a single ISP. In such low probability events, we equally divide the attribution across both potential critical clusters.

4. ANALYSIS OF PROBLEM AND CRITICAL CLUSTERS

In this section, we analyze the properties of the problem clusters and critical clusters. For brevity, we present results from the first week of the dataset in this section noting that the results are consistent across both weeks.

At a high level, we find that: (1) There are a non-trivial number of problem clusters that are *prevalent* (i.e., recurrent problems) and *persistent* (i.e., long lasting); (2) The majority of these problem clusters are covered by a small number of critical clusters—a few potential causes that can explain most of these observations; (3) Most of critical clusters correspond to either the Site, the ASN, or the CDN; and (4) While the critical attributes for different quality metrics are similar, the specific values of ASN, Site, or CDN that appear as critical clusters varies quite significantly.

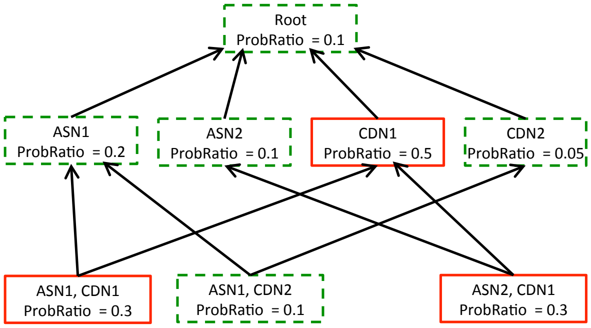


Figure 4: Representing the relationship between clusters using a DAG. The dashed-green boxes show clusters without a high problem ratio and the solid-red boxes identify the problem clusters.

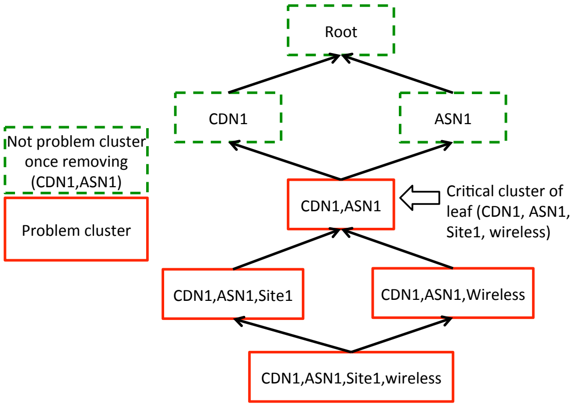

Root
CDN1
ASN1
Not problem cluster once removing (CDN1,ASN1)
Problem cluster
CDN1,ASN1
Critical cluster of leaf (CDN1, ASN1, Site1, wireless)
CDN1,ASN1,Site1
CDN1,ASN1,Wireless
CDN1,ASN1,Site1,wireless

Figure 5: An example to illustrate the phase transition idea for identifying a critical cluster. Intuitively, removing any one attribute from this critical cluster will cease to be a problem cluster and adding any attribute to it will continue to be a problem cluster.

4.1 Prevalence and Persistence

Recall that a problem cluster is a group of problem sessions that occur in the same one-hour epoch that has a problem ratio $\geq 1.5\times$ global problem ratio and that has at least $\geq$ 1000 sessions. That is, we are focusing on statistically significant problem events. Here, we begin by analyzing the temporal *prevalence* and *persistence* of the problem clusters.

We define the *prevalence* of a problem cluster as the fraction of the total number of epochs in which this cluster appears as a problem cluster. Consider the example in Figure 6 with a total of 6 epochs, the prevalence of the cluster "$ASN1$, $CDN1$" is $\frac{4}{6} = 0.67$ and similarly the prevalence of the cluster "$CDN2$" is $\frac{5}{6} = 0.83$. Figure 7 shows the distribution of the prevalence of the problem clusters for the different quality metrics. We see a consistent pattern across all quality metrics that around 10% of the clusters have a prevalence greater than 8% across all metrics. In other words, many of these problem clusters are repeated observations that are recurrent problem events.

We define the *persistence* of a problem cluster in terms of the length of the consecutive occurrences of this cluster as a problem

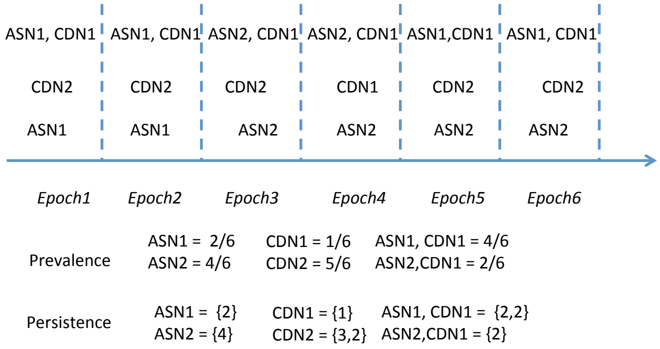


Figure 6: ***Illustrating the notion of prevalence and persistence. For persistence analysis, we group together "events" that appear continuously.***

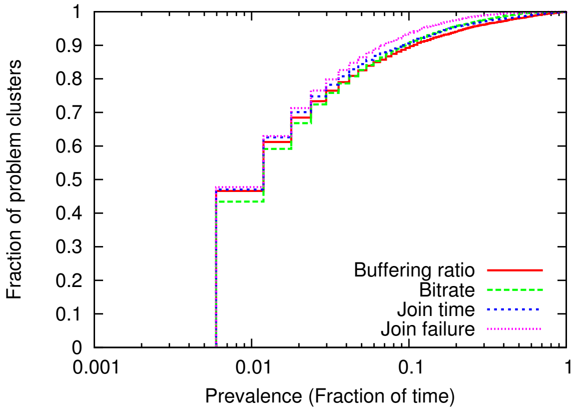


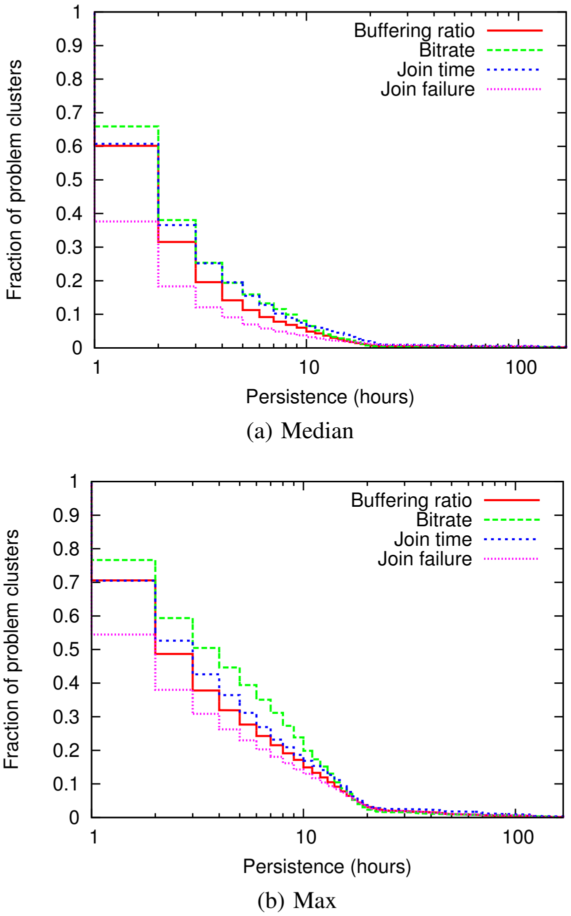
Figure 7: ***Distribution of the prevalence of problem clusters across quality metrics. We find a natural skewed distribution with a few clusters having high prevalence.***

cluster. To this end, we coalesce consecutive occurrences of the cluster into a single logical event that lasts for multiple hours. For each problem cluster, we consider the distribution of the length of these "streaks" and report the *median* and the *maximum* value. For the timeseries in Figure 6, the "$ASN1$,$CDN1$" cluster has a median and maximum persistence of 2 while the "$ASN2$" cluster has a maximum persistence of 4. (In this simple series, the median and max coincide, but more generally they will not.)

Figure 8(a) shows the distribution of the median persistence and Figure 8(b) shows the distribution of the longest persistent event across the problem clusters for the 4 quality metrics. For three of the metrics, more than 60% of the problem clusters have a median duration that last more than 2 hours. Furthermore, we see that more than 1% of clusters have a peak duration that last more than a day!

As we will see later, these observations have key implications for addressing video quality problems. The prevalence analysis suggests that we may be able to alleviate video quality problems by proactively diagnosing the pathological clusters and taking some remedial measures. The persistence analysis suggests that we may also be able to reactively diagnose and alleviate problems events even after they occur because many of these events last several hours.

4.2 Critical cluster analysis



(a) Median

(b) Max

Figure 8: ***Inverse CDF of the median and max persistence of problem clusters. Many problem clusters last multiple hours and that a non-trivial number of problem clusters last for tens of hours.***

The previous results showed that there are a non-trivial number of persistent/prevalent problem clusters that last for several hours . As we discussed earlier, multiple problem clusters may be implicitly related by a single root cause as we saw in Figure 4. To this end, we focus next on the critical clusters using the algorithm described in Section 3. Recall that every critical clusters is also a problem cluster; i.e., it has a sufficiently high problem ratio and it has a significant number of sessions. The motivation to focus on a few critical cluster rather than all problem clusters is the observation (as shown shortly) that a small fraction of problem clusters cover most of the problem sessions.

Figure 9 shows the number of problem clusters relative to the number of critical clusters in the case of the Join Time metric. We see that number of critical clusters is almost 50× lower than the number of problem clusters suggesting that there are indeed a small number of events that might have "caused" most problems. One natural question is whether the critical clusters *cover* most of the problem sessions. Table 1 summarizes the mean coverage and reduction of the critical clusters for the four quality metrics and in all cases, we see that the number of critical clusters is only 2-3% of the number of problem clusters (i.e., 50× fewer), but they manage to cover 44–84% of the problem sessions. As a point of reference, we also show the coverage of the problem clusters; i.e., not all ses-

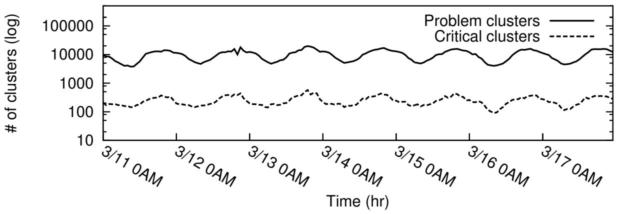


Figure 9: ***The number of critical clusters is significantly smaller than the number of problem clusters. The timeseries shown here is for the join time; we see similar results for the other quality metrics too.***

Metric	Mean problem clusters	Mean critical clusters	Mean problem cluster coverage	Mean critical cluster coverage
BufRatio	10433	286 (2%)	0.8	0.66 (82%)
JoinTime	9953	247 (2%)	0.86	0.83 (96%)
JoinFailure	9620	302 (3%)	0.87	0.84 (96%)
Bitrate	9437	287 (3%)	0.57	0.44 (77%)

Table 1: ***Reduction via focusing only on critical clusters and the effective coverage of the critical clusters.***

sions are part of a problem cluster as they may be part of small clusters or clusters with very small problem ratio. We see that the critical clusters cover almost all problem sessions that are part of some problem cluster; i.e., many of the coverage gaps are really due to problem sessions that belong to a statistically insignificant cluster (i.e., either with too few sessions or with too few problem sessions).

We also repeated the prevalence and persistence analysis for the critical clusters and find similar patters of skewed distribution and a few pathological cases that span several hours. We do not show these for brevity.

4.3 Types of Critical Clusters

Next, we analyze the structure of the critical clusters for the different quality metrics. First, we analyze the types of client/session attribute combinations that appear frequently in the critical clusters. Then, we analyze if the different metrics are correlated in the critical clusters. Finally, we highlight some interesting observations and some hypothesis to explain the most prevalent critical clusters.

Types of clusters: Figure 10 shows a breakdown of the types of critical clusters for different quality metrics. We aggregate critical clusters into the different attribute dimension(s) they represent. For instance, if we see a critical cluster for $CDN1$ and $CDN2$, we count these toward the CDN contribution in the pie chart. Some problem sessions may remain unaccounted for in this breakdown for two reasons: (a) they are not part of a significant enough problem cluster or (b) our algorithm did not assign a critical cluster for a problem cluster. This mirrors the coverage observation we saw earlier in Figure 9 and Table 1. In almost all cases, most of the unaccounted for sessions fall outside any problem cluster; i.e., this is not due to the critical cluster detection algorithm. The result shows that the most dominant category of critical clusters actually corresponds to a *content provider* (labeled as "Site"). We also see that CDN, ASN, and ConnectionType are also prominent types of critical clusters across all quality metrics. This suggests that most quality issues are potentially caused by server-side (Site or CDN) or client-side (ASN, ConnectionType) problems rather than a combination (which indicates a bad path between client and server) or other attributes.

BufRatio vs. Bitrate	BufRatio vs. JoinTime	BufRatio vs. JoinFailure	Bitrate vs. JoinTime	Bitrate vs. JoinFailure	JoinTime vs. JoinFailure
0.07	0.23	0.13	0.08	0.01	0.09

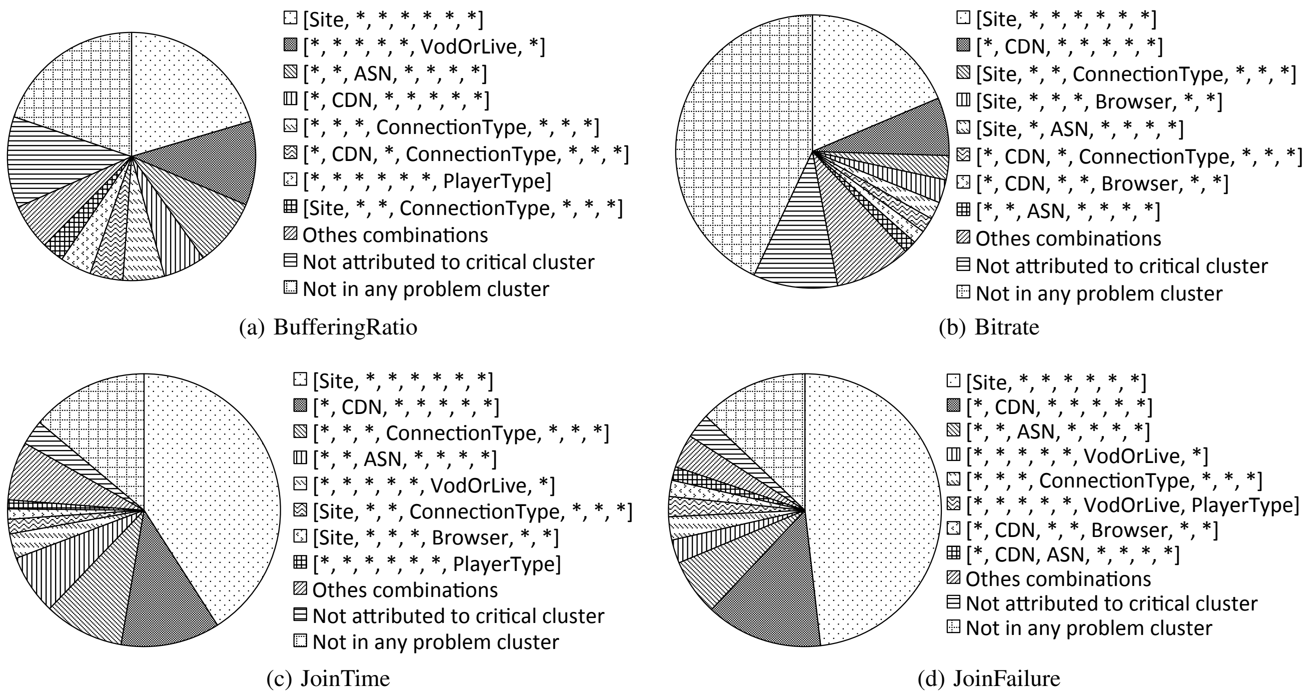
Table 2: ***Average Jaccard similarity index between the top 100 critical clusters for the different metrics. We see that most metrics are relatively uncorrelated, possibly because the critical attributes are very different.***

Overlap across metrics: We saw in the previous graph that the types of critical clusters that contribute the most problem sessions are very similar across different quality metrics. Note, however, that this does not necessarily mean that the actual set of critical clusters are identical. In other words, a different set of CDNs or Sites may be responsible for problems across buffering ratio and join time. To analyze this, we compute the *Jaccard similarity* index between the top-100 in terms of the total number of problem sessions covered critical clusters for the different metrics. (The Jaccard similarity measure for two sets A and B is $\frac{|A \cap B|}{|A \cup B|}$.) We find that the overlap between the different metrics is only around 23% in the best case (buffering ratio and join time) and in the worst case is only around 1% (between bitrate and join failure). We manually analyzed the specific clusters and we found that the actual set of Site, CDN, and ASN critical clusters are indeed very different.

Understanding most prevalent critical clusters: In order to illustrate the causes for the problem, we consider the critical clusters with a prevalence higher than 60% for the different quality metrics. For clarity of presentation, we only consider the critical clusters whose attributes fall in one of the following categories: ASN, CDN, Site, and ConnectionType as our previous breakdown shows these as the most dominant attributes. We present this analysis with two disclaimers. First, due to the sensitive nature of this data, we do not present the names of the actual providers, but focus on their characteristics. Second, this involves a fair amount of manual analysis and domain knowledge. As such, we intend this result to be illustrative (and somewhat speculative) rather than attempt to be conclusive. This said, we still believe that the high-level insights are still useful to inform future video delivery architectures.

Table 3 presents some of the anecdotal examples we observed. The empty cells simply indicate that there were no critical clusters in this category with a prevalence higher than 60%. We see a few interesting patterns here. In terms of buffering ratio, we see that the top ASNs are typically in Asia, and the content providers that had issues typically only had a single bitrate of content. The CDNs with buffering/join time problems are also typically "in-house" CDNs run by the Site itself; i.e., not a third-party CDN like Akamai or Limelight. We also see that wireless connections and wireless ISPs appear in the buffering and bitrate cells respectively, which is somewhat expected.

One interesting artifact we uncovered in the case of join time was that these were mostly ASNs in China accessing content from Chinese CDNs but there were third-party player modules loaded from US providers that led to higher join times. Another curious observation is that all the Sites with significant join failures tended to use the same global CDN. However, the CDN in aggregate does not have a significant presence in terms of failures, except in the



(a) BufferingRatio

(b) Bitrate

(c) JoinTime

(d) JoinFailure

Figure 10: ***Analyzing the structure of the critical clusters: The result show a breakdown of the total number of sessions attributed to a specific type of critical cluster. Note that there may be multiple values of these attributes; i.e., there can be many Sites and many CDNs contributing to the Site and CDN sector.***

	ASN	CDN	Site	ConnType
BufRatio	Asian ISPs	In-house, single bitrate	Single bitrate	Mobile wireless
JoinTime	Chinese ISPs accessing CDNs in China, but player loads modules from US CDN	In-house CDNs of UGC providers	High bitrates	
JoinFailure		Same set as buffering ratio	Same single global CDN, maybe low priority providers	
Bitrate	Wireless provider		UGC Sites	

Table 3: ***Analysis of the most prevalent critical clusters. A empty cell implies that we found no interesting cluster in this combination.***

case of these Sites.[5] We speculate that these, presumably low-end, providers may have lower priority service and could have potentially benefited from using multiple CDNs.

4.4 Summary of main observations

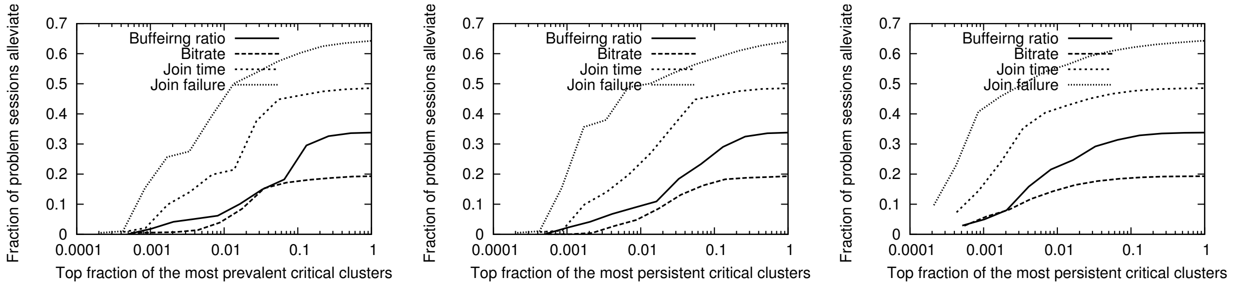
Our key observations from the analysis of problem clusters and critical clusters are:

- There is a distinct skewed distribution in the prevalence; around 8-12% of the problem clusters appear more than 10% of the time.
- There is also a skewed distribution in the persistence; more than 20% of problem clusters have a median duration greater than 2 hours and 1% of clusters have a peak duration lasting more than 1 day.
- We find that a small number of critical clusters (2-3% of the number of problem clusters) can account for 44-84% of all problem sessions.
- While the set of attribute combinations in the critical clusters that cover the most number of problem sessions is very similar across the quality metrics (i.e., Site, CDN, ASN), the actual values of these attributes is very different (with a max overlap of 23%).
- We see a few expected patterns such as Asian and wireless ISPs appearing as most prevalent critical clusters. We see some unexpected patterns that can be easily alleviated (e.g., the player modules loaded remotely for Chinese users) and Sites that could benefit from standard strategies such as using more fine-grained bitrates or using multiple CDNs.

5. WHAT-IF IMPROVEMENT ANALYSIS

In the previous section, we observed that a small number of critical clusters can potentially explain most common video quality problems and that many of these problems are both persistent and prevalent. These observations seem to suggest that focusing on these few critical clusters can yield quite significant improvements. In this section, we run a series of *what-if* analyses to estimate the

[5] These Sites used a single CDN; recall that our critical cluster algorithm will prefer more compact descriptions and thus attributes these problems to the Site rather than the single Site-CDN combination.



(a) Prevalence (b) Persistence (c) Coverage

Figure 11: ***Overall improvement by fixing the top-k critical clusters ranked in order prevalence, persistence, and coverage of attributed volume for the four quality metrics. Across the metrics we see a manifestation of the Pareto rule where 1% of the top-k clusters can alleviate up to 60% of the potential problems.***

potential improvement in video quality if we focus our efforts on some of these key clusters. Our goal in this section is not to evaluate a specific deployment strategy for alleviation; e.g., through multiple CDNs or better initial bitrate selection or better bitrate adaptation algorithms. Rather, we want to evaluate the potential for improvement and quantify the number of sessions amenable to simple strategies rather than an observed/actual improvement. We cannot conclusively say that a) the specific sessions we consider are actually fixable or b) fixing them incurs zero/low cost or that c) we have a prescriptive strategy for the specific "principals" such as ISP or Site or CDN.

Methodology: At a high-level, we consider two key dimensions that together define the what-if analyses:

1. What *types* of critical clusters?

   For instance, should we focus on prevalent vs. persistent clusters or clusters that cover the most sessions? Should we focus only a few key attributes such as Site or ASN or should we consider a broader approach? Finally, how *many* of these critical clusters should we select; e.g., what is the marginal utility of the top-10% clusters vs. the top-1% clusters?

2. Do we use a *proactive* or *reactive* approach?

   A proactive approach would involve offline analysis to identify the clusters and address the problems before new problems can occur. A reactive approach waits for problem incidents to occur and then considers temporarily alleviating the problems (e.g., [21]).

Once we identify the *candidate* critical clusters along these axes, we evaluate the impact of logically "fixing" problem sessions. First, we identify the epochs in which this candidate cluster was flagged as a critical cluster. Then, we evaluate a scenario in which the problem ratio for the clusters attributed to this critical cluster can be reduced to global observed average problem ratio across all clusters. Intuitively, this conversion to the global average is simulating the fact that there are likely to be some background effects due to which some baseline number problem sessions are unavoidable. Thus, the best we can do is roughly to reduce the problem ratio of the given critical cluster to the global average.

5.1 Type of clusters to select

We begin by analyzing the types of critical clusters we want to select and consider three natural approaches by selecting the top-k critical clusters in terms of: (1) prevalence, (2) persistence, and (3) coverage in terms of number of problem sessions attributed to it.

In Figure 11, we see a consistent pattern across all three approaches. First, we see a manifestation of the Pareto rule, with the top few clusters account for a significant fraction of all problem sessions; e.g., in the case of Join failure and coverage combination (Figure 11(c)), the top 1% can account for almost 60% of all problem sessions. Second, we see that the relative benefit for different metrics are slightly different, with join failure and join time showing significantly higher potential improvement rates compared to buffering ratio and bitrate. In these results, the maximum possible values do not reach 1 on the y-axis; this is because even choosing all critical clusters will not cover all problem sessions as we already saw Table 1.

We also see that choosing the top-k critical clusters in terms of problem session coverage yields significantly more improvement than the persistence or prevalence based selection. In some sense, this is not surprising as the persistence and prevalence rankings are volume agnostic; they simply look at the number of epochs in which the critical clusters appear. That said, the critical clusters with the highest coverage may also have a large volume of traffic and thus incur a higher cost (e.g., potential upgrades or disruption to users). We revisit this aspect in the next section.

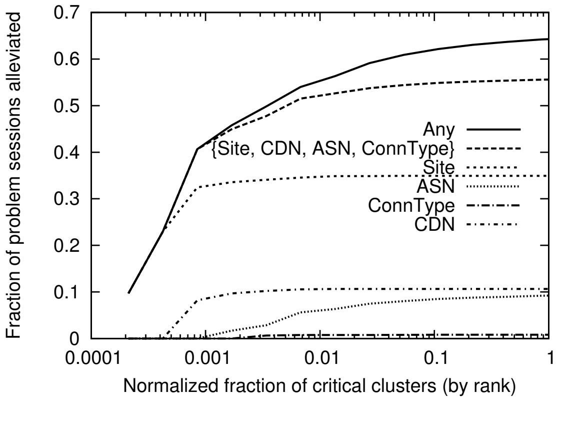


Figure 12: ***Comparison between heuristic selection of critical clusters using on specific attributes vs. broader approaches that consider more attribute combinations.***

One natural question, building on the breakdown in Figure 10, could be that focusing on only one of the attributes can itself yield most of the benefit. To analyze this, we compare the effect of picking the top-k critical clusters but focusing only on specific attributes such as CDN, ASN, or Site in Figure 12. (The x-axis here is normalized by the total number of critical clusters; thus, attributes that only have a few distinct values such as CDN do not appear throughout the x-axis.) For clarity, we only consider the join failure and the top-k clusters in terms of the coverage criterion. We see that focusing on any one attribute alone cannot provide significant improvements compared to considering all types of critical clusters (marked as "any" in the graph). We do, however, see that considering the union of the top-4 attributes does provide comparable improvement.

5.2 Proactive history-based approaches

Metric	Intra-week		Inter-week	
	New	Potential	New	Potential
BufRatio	0.35 (71%)	0.49	0.19 (61%)	0.31
Bitrate	0.13 (68%)	0.19	0.09 (64%)	0.14
JoinTime	0.47 (84%)	0.56	0.42 (85%)	0.49
JoinFailure	0.68 (85%)	0.8	0.54 (86%)	0.63

Table 4: ***Trace-driven simulation of a proactive alleviation strategy. Here, we identify the top 1% critical clusters observed in offline data, simulate the effect of reducing the problem ratio for these clusters ("New") in the future epochs, and compare it with the effect of reducing the problem ratio for 1% critical clusters of the future epochs ("Potential"). Percentage in the bracket shows how close to the potential improvement the proactive alleviation strategy is.***

The previous result considers an "oracle" setting where we alleviate the critical clusters after-the-fact. Next, we consider a *proactive* alleviation strategy where we use offline analysis based on historical data to identify the key critical clusters. Then, we consider the potential improvement in the future epochs assuming that these few critical clusters have improved performance. As before, we assume that the problem ratio for the problem clusters attributed to these chosen critical clusters are reduced to the global average for the epoch.

We consider two settings here. In the first *intra-week* setting, we identify the top 1% critical clusters (by coverage) using the first 4 days of the first week trace, and test the improvement on the remaining three days. In the second *inter-week* setting, we use the first week to identify the top 1% critical clusters and analyze the improvement in the second week. Table 4 shows the mean problem ratio for the intra-week and inter-week simulation. As a point of reference, we also show the upper bound if we did the "oracle" simulation from the earlier experiment where we identify the top 1% critical clusters in each epoch and fix them; i.e., using after-the-fact analysis rather than history. We see that even using historical measurement data can yield close-to-optimal (70–85%) benefits across the four metrics.

5.3 Reactive approaches

Next, we consider a setting where we detect and *react* to problem incidents *after* they occur. Conceptually, this approach focuses on the persistent events by detecting them after the first hour in which it appears as a critical cluster and then using some remedial action(s) to reduce the problem ratio for the remaining duration of the event to the global average problem ratio. Again, our goal is not to focus on the specific type of remedial action such as reducing bitrates or switching CDNs [21]. Rather, we want to highlight the potential for improvement in the spirit of the previous analysis.

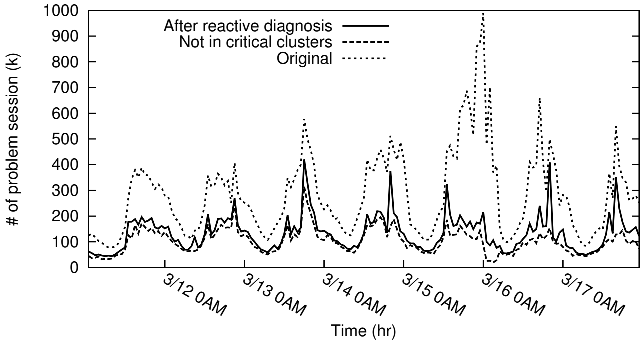


Figure 13: ***Improvement with a reactive approach for the join failure. We see that even reactive approaches can yield significant improvement. "Not in critical clusters" refers to those sessions cannot be associated with any critical cluster and thus cannot be alleviated by fixing critical clusters. They are more likely to be "random" problems.***

	New	Potential
BufRatio	0.43 (95%)	0.45
Bitrate	0.12 (70%)	0.17
JoinTime	0.48 (78%)	0.61
JoinFailure	0.51 (81%)	0.63

Table 5: ***The average improvement with a reactive approach for different quality metrics.***

Figure 13 shows a trace-driven simulation of this reactive approach in the case of the join failure metric. The result shows that even a reactive strategy that takes 1 hour to detect a problem and then attempts to remedy it, can still reduce the overall problem ratio by 50%. We summarize the mean improvements for the remaining metric in Table 5, and find consistent trends across all quality metrics.

5.4 Summary of main observations

There are three main takeaways from our what-if analysis:

- We see that even just selecting the top 1% of critical clusters (in terms of coverage) can yield a potential improvement of 15-55% across the quality metrics.
- Proactively alleviating the dominant critical clusters observed in history can also yield close-to-optimal improvement (60–85% of the upperbound) in the future.
- Even a simple reactive strategy of waiting for critical clusters to emerge after 1 hour and taking remedial actions to address these critical clusters can reduce up to 51% of problem sessions.

6. DISCUSSION

While our work provides several insights into the structure of Internet video quality problems, we acknowledge that this is only a first step. Here, we identify some potential limitations and suggest some directions to extend our analysis.

Cost of remedial measures: Our improvement analysis does not capture the costs that might be incurred to logically "fix" a particular critical cluster; e.g., does it need infrastructure upgrades or contracting new CDN service or using multiple CDNs or multiple bitrates. It will be interesting to also consider a natural cost-benefit analysis that considers the complexity of upgrading or taking remedial actions for each critical cluster. A more comprehensive solution will involve an automated system that identifies the bottleneck as well as provides remedial actions.

Hidden attributes: There might be hidden attributes associated with each viewing session on two fronts: (1) the attribute may be measurable but implicit (e.g., we found many ASNs in non-US regions, so it is natural to consider geography as an additional attribute) and (2) the attribute may not be immediately measurable (e.g., does a particular ASN use rate limiting or excessive buffers in switches). That said, the analysis techniques we use is quite general and can be repeated when more client attributes are available.

More diagnostic capabilities: In our current framework, we largely rely on domain knowledge and manual diagnosis to *explain* the phenomena we observe (as seen in Section 4.3). Furthermore, we currently consider a static setting where the set of attributes and problem sessions are given as input. One natural extension to this framework if we can trigger more fine-grained measurements (possibly from third-party sources) when we observe a specific critical cluster. For instance, if we observe a specific CDN having quality issues, then we may request server load statistics from that CDN. Similarly, if a specific combination of ASN-CDN has quality issues, then we may need deployment maps for that CDN near that ASN to diagnose problems. We leave this for future work.

7. RELATED WORK

We briefly highlight some of the key related work to place our work in context.

Impact of quality on users: Dobrian et al. showed that the buffering percentage is the most critical metric [13]. More recent studies demonstrate a more causal relationship between quality and engagement [19] and also identify specific externalities (e.g., mobile device) that impact the relationship between quality and engagement [9]. These efforts focus on the impact of quality on engagement and do not analyze why and where these quality problems occur.

Improving video quality: This includes the work on identifying problems existing in client-adaptation algorithms (e.g., [8]), interactions with TCP control loops (e.g., [15, 14]), and techniques to improve bitrate adaptation (e.g., [7, 17]). Other efforts have demonstrated inefficiencies in CDN and server selection strategies [27, 4]. Recent work has suggested cross-CDN optimizations to improve video quality [21, 20]. Each of these efforts focuses on particular aspect of the video delivery ecosystem. Our work takes a broader view of the entire ecosystem. Our findings can inform the design of these optimizations and further suggests simpler strategies that can alleviate a significant number of potential problems.

Other video measurements: There is a large literature in understanding content popularity and access patterns (e.g., [11][24]), flash crowds during highly popular events (e.g., [28]), and their implications for CDN and caching designs. Our focus is not on the popularity or caching implications but on understanding the structure of quality problems.

Video streaming bottlenecks: Mahimkar et al. characterize performance problems in a large scale IPTV network [22] and Wu et al. have analyzed performance problems in a P2P live streaming system [10]. While our work follows in this spirit, our work differs in two aspects. First, our vantage point gives us a unique opportunity to study multiple content providers rather than focus on inefficiencies in one specific provider. Second, we focus on web-based Internet video which is the dominant fraction of video traffic rather than IPTV or P2P deployments.

Network performance variability: Past measurement studies have shown that the network performance can be quite variable (e.g., [18, 6, 23]). Recent work on crowdsourcing focuses on network performance bottlenecks [26, 12]. Our specific focus in this work is not on identifying network bottlenecks. That said, a natural direction of future work is to integrate such bottleneck detection techniques for deeper diagnosis.

Clustering algorithms: The problem of detecting critical clusters is conceptually similar to detecting hierarchical heavy hitters (HHH) [29, 22]. The goal with HHH is to detect all clusters that contribute a significantly high fraction to the total volume even after removing all of its descendants already marked as a HHH cluster. While we also seek to identify common patterns in multidimensional data, there is a key difference. The critical cluster generation step is not a simple volume counting application; we want to attribute problems to one specific parent cluster. Thus, HHH approaches are not directly applicable.

8. CONCLUSIONS

The growth of Internet video and the role that video quality plays in user engagement (and thus revenues) has sparked a renewed interest in redesigning various aspects of the content delivery ecosystem ranging from video players, CDNs, multi-CDN optimizations, and global control planes. While these efforts are valuable, what is critically lacking today is a broad spectrum understanding of the nature of video quality problems as they occur in the wild. In some sense, the trajectory of these efforts, including our own prior work in this space, has been somewhat backwards—as a community we have proposed solutions without necessarily understanding if they are necessary or why they provide improvements.

This paper was an initial attempt to bridge this gap. We find, perhaps surprisingly, that a small number of potential problem causes can account for a large number of problem sessions. Furthermore, these problem causes are amenable to simple solutions, either via using offline traces to identify the sources of these problems or by reacting only to long-lasting outages. We believe that these observations bodes well for the Internet video ecosystem going forward as many of the aforementioned efforts to improve video quality could be simplified to achieve the same benefits.

Acknowledgments

We thank our shepherd Fabián E. Bustamante and the anonymous reviewers for their feedback. This research was supported in part by the National Science Foundation under awards CNS-1040757, CNS-1040800, and CNS-1040801.

9. REFERENCES

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

[2] Quova. http://developer.quova.com/.

[3] 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.

[4] V. K. Adhikari, Y. Chen, S. Jain, and Z.-L. Zhang. Where Do You 'Tube'? Uncovering YouTube Server Selection Strategy. In *Proc. IEEE ICCCN*, 2011.

[5] V. K. Adhikari, Y. Guo, F. Hao, V. Hilt, , and Z.-L. Zhang. A Tale of Three CDNs: An Active Measurement Study of Hulu and Its CDNs. In *Proc. IEEE Global Internet Symposium*, 2012.

[6] A. Akella, S. Seshan, and A. Shaikh. An empirical evaluation of wide-area internet bottlenecks. In *Proc. Internet Measurement Comference*, 2003.

[7] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen. What Happens when HTTP Adaptive Streaming Players Compete for Bandwidth? In *Proc. NOSSDAV*, 2012.

[8] S. Akhshabi, A. C. Begen, and C. Dovrolis. An Experimental Evaluation of Rate Adaptation Algorithms in Adaptive Streaming over HTTP. In *Proc. MMSys*, 2011.

[9] 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.

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

[11] 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.

[12] D. R. Choffnes, F. E. Bustamante, and Z. Ge. Crowdsourcing service-level network event monitoring. In *ACM SIGCOMM Computer Communication Review*, volume 40, pages 387–398. ACM, 2010.

[13] 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.

[14] J. Esteban, S. Benno, A. Beck, Y. Guo, V. Hilt, and I. Rimac. Interactions Between HTTP Adaptive Streaming and TCP. In *Proc. NOSSDAV*, 2012.

[15] M. Ghobadi, Y. Cheng, A. Jain, and M. Mathis. Trickle: Rate Limiting YouTube Video Streaming. In *Proc. USENIX ATC*, 2012.

[16] T.-Y. Huang, N. Handigol, B. Heller, N. Mckeown, and R. Johari. Confused, timid, and unstable: picking a video streaming rate is hard. In *Proc. IMC*, 2012.

[17] J. Jiang, V. Sekar, and H. Zhang. Improving Fairness, Efficiency, and Stability in HTTP-Based Adaptive Streaming with Festive . In *ACM CoNext*, 2012.

[18] C. Kreibich, B. Nechaev, V. Paxson, and N. Weaver. Netalyzr: Illuminating The Edge Network. In *Proc. IMC*, 2010.

[19] S. S. Krishnan and R. K. Sitaraman. Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs. In *IMC*, 2012.

[20] H. Liu, Y. Wang, Y. R. Yang, A. Tian, and H. Wang. Optimizing Cost and Performance for Content Multihoming. In *in Proc. SIGCOMM*, 2012.

[21] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang. A Case for a Coordinated Internet Video Control Plane. In *Proc. SIGCOMM*, 2012.

[22] 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.

[23] J. S. Otto, M. A. Sánchez, D. R. Choffnes, F. E. Bustamante, and G. Siganos. On blind mice and the elephant. In *Proc. of ACM SIGCOMM*, 2011.

[24] L. Plissonneau and E. Biersack. A longitudinal view of http video streaming performance. In *Proc. MMSys*, 2012.

[25] D. Rayburn. Telcos and carriers forming new federated cdn group called ocx (operator carrier exchange). June 2011. StreamingMediaBlog.com.

[26] M. A. Sánchez, J. S. Otto, Z. S. Bischof, D. R. Choffnes, F. E. Bustamante, B. Krishnamurthy, and W. Willinger. Dasu: Pushing experiments to the internetâĂŹs edge. In *Proc. of USENIX NSDI*, 2013.

[27] R. Torres, A. Finamore, J. R. Kim, M. Mellia, M. M. Munafo, and S. Rao. Dissecting Video Server Selection Strategies in the YouTube CDN. In *ICDCS*, 2011.

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

[29] Y. Zhang, S. Singh, S. Sen, N. Duffield, and C. Lund. Online identification of hierarchical heavy hitters: algorithms, evaluation, and applications . In *Proc. IMC*, 2004.