Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-Experimental Designs

S. Shunmuga Krishnan and Ramesh K. Sitaraman, *Member, IEEE*

***Abstract*—The distribution of videos over the Internet is drastically transforming how media is consumed and monetized. Content providers, such as media outlets and video subscription services, would like to ensure that their videos do not fail, start up quickly, and play without interruptions. In return for their investment in video stream quality, content providers expect less viewer abandonment, more viewer engagement, and a greater fraction of repeat viewers, resulting in greater revenues. The key question for a content provider or a content delivery network (CDN) is whether and to what extent changes in video quality can *cause* changes in viewer behavior. Our work is the first to establish a causal relationship between video quality and viewer behavior, taking a step beyond purely correlational studies. To establish causality, we use Quasi-Experimental Designs, a novel technique adapted from the medical and social sciences. We study the impact of video stream quality on viewer behavior in a scientific data-driven manner by using extensive traces from Akamai's streaming network that include 23 million views from 6.7 million unique viewers. We show that viewers start to abandon a video if it takes more than 2 s to start up, with each incremental delay of 1 s resulting in a 5.8% increase in the abandonment rate. Furthermore, we show that a moderate amount of interruptions can decrease the average play time of a viewer by a significant amount. A viewer who experiences a rebuffer delay equal to 1% of the video duration plays 5% less of the video in comparison to a similar viewer who experienced no rebuffering. Finally, we show that a viewer who experienced failure is 2.32% less likely to revisit the same site within a week than a similar viewer who did not experience a failure.**

***Index Terms*—Causal inference, Internet content delivery, multimedia, quasi-experimental design, streaming video, user behavior, video quality.**

I. Introduction

THE INTERNET is radically transforming all aspects of human society by enabling a wide range of applications for business, commerce, entertainment, news, and social networking. Perhaps no industry has been transformed more radically than the media and entertainment segment of the economy. As media such as television and movies migrate to the Internet, there are twin challenges that *content providers* face whose ranks include major media companies (e.g., NBC, CBS), news outlets (e.g., CNN), sports organizations (e.g., NFL, MLB), and video subscription services (e.g., Netflix, Hulu).

The first major challenge for content providers is providing a *high-quality streaming experience* for their viewers, where videos are available without failure, start up quickly, and stream without interruptions [1]. A major technological innovation of the past decade that allows content providers to deliver higher-quality video streams to a global audience of viewers is the content delivery network (CDN for short) [2], [3]. CDNs are large distributed systems that consist of hundreds of thousands of servers placed in thousands of ISPs close to end-users. CDNs employ several techniques for transporting [4], [5] media content from the content provider's origin to servers at the "edges" of the Internet where they are cached and served with higher quality to the end-user. (See [3] for a more detailed description of a typical CDN architecture.)

The second major challenge of a content provider is to actually *monetize* their video content through ad-based or subscription-based models. Content providers track key metrics of viewer behavior that lead to better monetization. Primary among them relate to viewer *abandonment, engagement, and repeat viewership*. Content providers know that reducing the abandonment rate, increasing the play time of each video watched, and enhancing the rate at which viewers return to their site increase opportunities for advertising and upselling, leading to greater revenues. The key question is whether and by how much increased stream quality can *cause* changes in viewer behavior that are conducive to improved monetization. Relatively little is known from a scientific standpoint about the all-important *causal* link between video stream quality and viewer behavior for online media. *Exploring the causal impact of quality on behavior and developing tools for such an exploration are the primary foci of our work.*

While understanding the link between stream quality and viewer behavior is of paramount importance to the content provider, it also has profound implications for how a CDN must be architected. An architect is often faced with tradeoffs on *which* quality metrics need to be optimized by the CDN. A scientific study of which quality metrics have the most impact on viewer behavior can guide these choices. As an example of viewer behavior impacting CDN architecture, we performed

Manuscript received January 25, 2013; revised May 17, 2013 and September 09, 2013; accepted September 09, 2013; approved by IEEE/ACM Transactions on Networking Editor M. Reisslein. Date of publication September 30, 2013; date of current version December 13, 2013. A preliminary version of this work appeared in the ACM Internet Measurement Conference (IMC), Boston, MA, USA, November 14–16, 2012 and was selected as a fast-track submission to the journal.

S. S. Krishnan is with Akamai Technologies, Bangalore 560 037, India (e-mail: sarumuga@akamai.com).

R. K. Sitaraman is with the Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA (e-mail: ramesh@cs.umass.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2013.2281542

small-scale controlled experiments on viewer behavior a decade ago that established the relative importance of the video to start up quickly and play without interruptions. These behavioral studies motivated an architectural feature called prebursting [5] that was deployed on Akamai's live streaming network that enabled the CDN to deliver streams to a media player at higher than the encoded rate for short periods of time to fill the media player's buffer with more data more quickly, resulting in the stream starting up faster and playing with fewer interruptions. It is notable that the folklore on the importance of startup time and rebuffering were confirmed in two recent important large-scale scientific studies [6], [7]. Our current work sheds further light on the important nexus between stream quality and viewer behavior and, importantly, provides the first evidence of a causal impact of quality on behavior.

A. *Measuring Quality and Viewer Behavior*

The advent of customizable media players supporting major formats such as Adobe Flash, Microsoft Silverlight, and Apple HTTP streaming has revolutionized our ability to perform truly large-scale studies of stream quality and viewer behavior as we do in this paper, in a way not possible even a few years ago. It has become possible to instrument media players with an analytics plugin that accurately measures and reports both quality and behavioral metrics from every viewer on a truly planetary scale.

B. *From Correlation to Causality*

The ability to measure stream quality and viewer behavior on a global scale allows us to correlate the two in a statistically significant way. For each video watched by a viewer, we are able to measure its quality including whether the stream was available, how long the stream took to start up, and how much rebuffering occurred causing interruptions. We are also able to measure the viewer's behavior including whether he/she abandoned the video and how long he/she watched the video.

As a first step, we begin by simply correlating important quality metrics experienced by the viewers to the behavior that they exhibit. For instance, we discover a strong correlation between an increase in the delay for the video to start up and an increase in rate at which viewers abandon the video. Several of our results are the first quantitative demonstration that certain key streaming quality metrics are correlated with key behavioral metrics of the viewer. However, the deeper question is not just whether quality and behavior are *correlated*, but whether quality can *causally* impact viewer behavior. While correlation is an important first step, correlation does not necessarily imply causality. The holy grail of a content provider or a CDN architect is to discover causal relationships rather than just correlational ones since they would like to know with *some certainty* that the significant effort expended in improving stream quality will in fact result in favorable viewer behavior.

In fact, a purely correlational relationship could even lead one astray, if there is no convincing evidence of causality, leading to poor business decisions. For instance, both video quality (say, video bit rates) and viewer behavior (say, play time) have been steadily improving over the past decade and are hence correlated in a statistical sense. However, that fact alone is not sufficient to conclude that higher bit rates cause viewers to watch longer, unless one can account for other potential "confounding" factors such as the available video content itself becoming more captivating over time.

While inferring causality is generally difficult, a key tool widely used in the social and medical sciences to infer causality from observational data is a quasi-experimental design (QED) [8]. Intuitively, a QED is constructed to infer if a particular "treatment" (i.e., cause) results in a particular "outcome" (i.e., effect) by pairing each person in the observational data who has had treatment with a random untreated person who is "significantly identical" to the treated person in all other respects. Thus, the pairing eliminates the effect of the hidden confounding variables by ensuring that both members of a pair have sufficiently identical values for those variables. Thus, evaluating the differential outcomes between treated and untreated pairs can either strengthen or weaken a causal conclusion that the treatment causally impacts the outcome. While it is impossible to completely eliminate all hidden factors, our causal analysis using QEDs should be viewed as strengthening our correlational observations between treatments and outcomes by eliminating the common threats to a causal conclusion.

C. *Our Contributions*

Our study is one of the largest of its kind of video stream quality and viewer behavior that collects and analyzes a data set consisting of more than 23 million video playbacks from 6.7 million unique viewers who watched an aggregate of 216 million minutes of 102 000 videos over 10 days.

To our knowledge, our work is the first to provide evidence that video stream quality *causally* impacts viewer behavior, a conclusion that is important to both content providers and CDNs. Furthermore, our adaptation of QEDs is a unique contribution and is of independent interest. QEDs have been used extensively in medical research and the social sciences in the past decades. We expect that our adaptation of QEDs for measurement research in networked systems could be key in a variety of other domains that have so far been limited to correlational studies.

Our work is also the first to quantitatively explore viewer abandonment rates and repeat viewership in relation to stream quality, last-mile connectivity, and video duration. In addition, we study viewer engagement (e.g., play time) in relation to stream quality (e.g., rebuffering) that has also been recently studied in [6] in a correlational setting, but we take a step beyond correlational analysis to establish a causal relationship between quality and engagement using QEDs. Our work makes the following specific contributions on the impact stream quality on viewer behavior.

- We show that an increase in the startup delay beyond 2 s causes viewers to abandon the video. Using regression, we show that an additional increase of the startup delay by 1 s increases the abandonment rate by 5.8%.
- Viewers are less tolerant to startup delay for a short video such as a news clip than a long video such as an hour-long TV episode. In a quasi-experiment, the likelihood of a viewer of a short video abandoning earlier than a similar

viewer of a long video exceeded the likelihood that the opposite happens by 11.5%.

- Viewers watching video on a better connected computer or device have less patience for startup delay and abandon sooner. In particular, viewers on mobile devices have the most patience and abandon the least, while those on fiber-based broadband abandon the soonest. In a quasi-experiment, the likelihood that a viewer on fiber abandoned earlier than a similar viewer on a mobile device exceeded the likelihood that the opposite happens by 38.25%.
- Viewers who experienced an increase in the normalized rebuffer delay, i.e., they experienced more interruptions in the video, played the video for lesser time. In a quasi-experiment, a viewer who experienced a rebuffer delay that equals or exceeds 1% of the video duration played 5.02% less of the video in comparison to a similar viewer who experienced no rebuffering.
- A viewer who experienced a failed visit is less likely to return to the content provider's site to view more videos within a specified time period than a similar viewer who did not experience the failure. In a quasi-experiment, the likelihood that a viewer who experienced failure returns to the content provider's site within a week is less than the likelihood of a similar viewer who did not experience failures by 2.32%.

We show that the above results are statistically significant using the sign test. Furthermore, these results show a significant level of causal impact of stream quality on viewer behavior. In this regard, it is important to recall that small changes in viewer behavior can lead to large changes in monetization since the impact of a few percentage points over tens of millions of viewers can accrue to large impact over a period of time. Finally, our work on deriving a causal relationship by systematically accounting for the confounding variables must not be viewed as a definitive proof of causality, as indeed there can be no definitive proof of causality. Rather, our work *significantly increases the confidence in a causal conclusion* by eliminating the effect of major confounding factors that could threaten such a conclusion.

II. Background

We describe the process of a user watching a stream, defining terms along the way that we will use in this paper.

Viewer: A viewer is a user who watches one or more streams using a specific media player installed on the user's device. A viewer is uniquely identified and distinguished from other viewers by using a globally unique identifier (GUID) value that is set as a cookie when the media player is accessed. To identify the viewer uniquely, the GUID value is generated to be distinct from other prior values in use.

Views: A view represents an attempt by a viewer to watch a specific video stream. A typical view would start with the viewer initiating the video playback, for instance, by clicking the play button of the media player[1] (see Fig. 1). During a view, the media player begins in the startup state where it connects to the server and downloads a certain specified amount of data, before transitioning to the play state. In the play state, the player uses the data from its buffer and renders the video on the viewer's screen. Meanwhile, the player continues to download data from the server and stores it in the buffer. Poor network conditions between the server and the player could lead to a situation where the buffer is drained faster than it is being filled. This could lead to a condition where the buffer is empty, causing the player to enter the rebuffer state where the viewer experiences an interruption or "freeze" in the video play back. While in the rebuffer state, the player continues to fill its buffer from the server. When the buffer has a specified amount of data, the player enters the play state and the video starts to play again (see Fig. 1). A view can end in three ways: A *successful view* ends normally when the video completes; a *failed view* ends with a failure or error due to a problem with the server, network, or content; and, finally, an *abandoned view* ends with the viewer *voluntarily* abandoning the stream either before the video starts up or after watching some portion of it. Note that a viewer may abandon the view by closing the browser, stopping the stream, or clicking on a different stream. There are other secondary player-initiated events or viewer-initiated events that are part of the viewing process. For instance, a viewer could initiate actions such as pausing, fast-forwarding, or rewinding the video stream. Furthermore, the player may switch the bit rate of the encoded media in response to network conditions, such as reduce the bandwidth if there is packet loss. We do not explicitly analyze behaviors associated with these secondary events in this paper, though these could be part of future work.

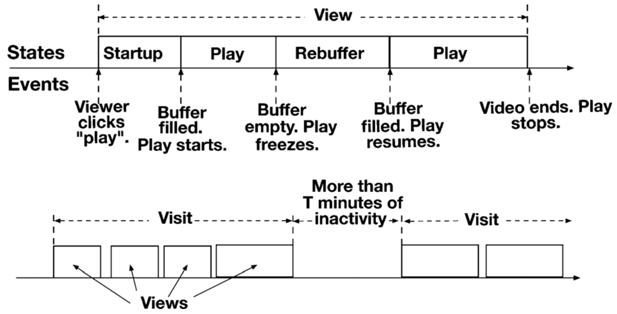

View
States
Startup
Play
Rebuffer
Play
Events
Viewer clicks "play".
Buffer filled. Play starts.
Buffer empty. Play freezes.
Buffer filled. Play resumes.
Video ends. Play stops.
Visit
More than T minutes of inactivity
Visit
Views

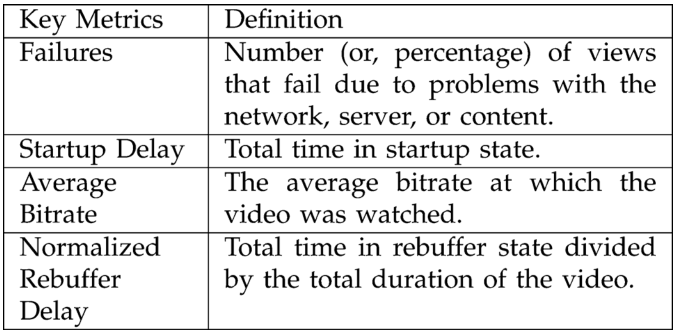
Fig. 1. Views and visits.

Visits: A visit is intended to capture a single session of a viewer visiting a content provider's site to view videos. A visit is a maximal set of contiguous views from a viewer at a specific content provider site such that each visit is separated from the next visit by at least T minutes of inactivity, where we choose $T = 30$ min[2] (see Fig. 1).

Stream Quality Metrics: At the level of a view, the key metrics that measure the quality perceived by the viewer are shown in Fig. 2. Failures address the question of whether the stream was available or if viewing of the video initiated by the viewer failed due to a problem with the network or the server or the content itself (such as a broken link). A failed view can be frustrating to the viewer as he/she is unable to watch the video. A

[1] For some content providers, a "pre-roll" advertisement is shown before the actual content video is requested by the media player. In that case, our view starts at the point where the actual video is requested.

[2] Our definition is similar to the standard notion of a visit (also called a session) in Web analytics, where each visit is a set of page views separated by a period of idleness of at least 30 min (say) from the next visit.



Key Metrics	Definition
Failures	Number (or, percentage) of views that fail due to problems with the network, server, or content.
Startup Delay	Total time in startup state.
Average Bitrate	The average bitrate at which the video was watched.
Normalized Rebuffer Delay	Total time in rebuffer state divided by the total duration of the video.

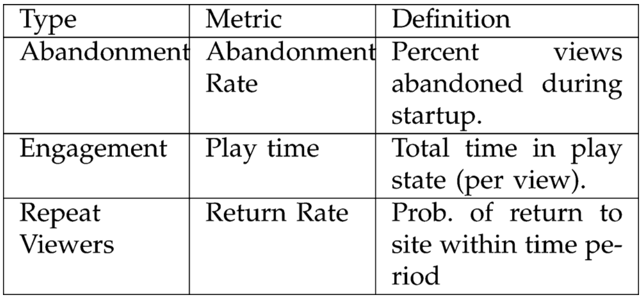
Fig. 2. View-level stream quality metrics.

second key metric is startup delay, which is the amount of time the viewer waits for the video to start up. Once the video starts playing, the average bit rate at which the video was rendered on the viewer's screen is a measure of the richness of the presented content. This metric is somewhat complex since it is a function of how the video was encoded, the network connectivity between the server and the client, and the heuristics for bit-rate switching employed by the player. Finally, a fourth type of metric quantifies the extent to which the viewer experienced rebuffering. Rebuffering is also frustrating for the viewer because the video stops playing and "freezes." We can quantify the rebuffering by computing the rebuffer delay, which is total time spent in a rebuffer state, and normalizing it by dividing it by the duration of the video.

Many of the above view-level metrics can be easily extended to visit-level or viewer-level metrics. One key visit-level metric that we examine in this paper is a *failed visit*, which is a visit that ends with a failed view. A failed visit could have had successful views prior to the failed view(s). However, a failed visit is important because the viewer tries to play a video one or more times but is unable to do so and leaves the site right after the failure, presumably with a level of frustration.

In our paper, we use many of the key metrics in Fig. 2 in our evaluation of the impact of quality on viewer behavior, though these are by no means the only metrics of stream quality. It should be noted that many of the above metrics were incorporated into measurement tools within Akamai and have been in use for more than a decade [1], [9]. The lack of client-side measurements in the early years led to measurements based on automated "agents" deployed around the Internet that simulated synthetic viewers [9], [10] that were then supplemented with server-side logs. In recent years, there is a broad consensus among content providers, CDNs, and analytics providers that these metrics or variations of these metrics matter.

Metrics for Viewer Behavior: Our metrics are focused on the key aspects of viewer behavior that are often tracked closely by content providers that we place in three categories (see Fig. 3). The first category is abandonment where a viewer voluntarily decides to stop watching the video.*Here, we are primarily concerned with abandonment where the viewer abandons the video even before it starts playing.* A viewer can also abandon a stream after watching a portion of the video that results in a smaller play time, which we account for in the next category of metrics. The second category is viewer engagement that can be measured by play time that is simply the amount of video that the viewer watches. The final category speaks to the behavior of viewers over longer periods of time. A key metric is the return rate of viewers measured as the probability that a viewer returns to the content provider's site over period of time, say, returning within a day or returning within a week.



Type	Metric	Definition
Abandonment	Abandonment Rate	Percent views abandoned during startup.
Engagement	Play time	Total time in play state (per view).
Repeat Viewers	Return Rate	Prob. of return to site within time period

Fig. 3. Key metrics for viewer behavior.

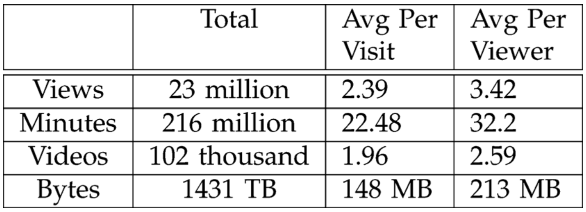
III. Data Sets

The data sets that we use for our analysis are collected from a large cross section of actual users around the world who play videos using media players that incorporate the widely deployed Akamai's client-side media analytics plug in.[1] When content providers build their media player, they can choose to incorporate the plugin that provides an accurate means for measuring a variety of stream quality and viewer behavioral metrics. When the viewer uses the media player to play a video, the plugin is loaded at the client side, and it "listens" and records a variety of events that can then be used to stitch together an accurate picture of the playback. For instance, player transitions between the startup, rebuffering, seek, pause, and play states are recorded so that one may compute the relevant metrics. Properties of the playback, such as the current bit rate, bit-rate switching, and state of the player's data buffer are also recorded. Furthermore, viewer-initiated action that leads to abandonment such as closing the browser or browser tab, clicking on a different link, etc., can also be accurately captured. Once the metrics are captured by the plugin, the information is "beaconed" to an analytics back end that can process huge volumes of data. From every media player at the beginning and end of every view, the relevant measurements are sent to the analytics back end. Furthermore, incremental updates are sent at a configurable periodicity even as the video is playing.

A. Data Characteristics

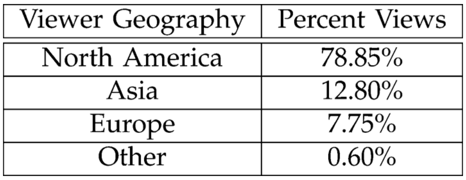
While the Akamai platform serves a significant amount of the world's enterprise streaming content accounting for several million *concurrent* views during the day, we choose a smaller but representative slice of the data from 12 content providers so as to include videos from all the major genres of news, entertainment, sports, television shows, and movies. We consider only on-demand videos in this study, leaving live videos for future work. We tracked the viewers and views for the chosen content

[3] While all our data is from media players that are instrumented with Akamai's client-side plugin, the actual delivery of the streams could have used *any* platform and not necessarily just Akamai's CDN.



	Total	Avg Per Visit	Avg Per Viewer
Views	23 million	2.39	3.42
Minutes	216 million	22.48	32.2
Videos	102 thousand	1.96	2.59
Bytes	1431 TB	148 MB	213 MB

Fig. 4. Summary of views, minutes watched, distinct videos, and bytes downloaded for our data set.



Viewer Geography	Percent Views
North America	78.85%
Asia	12.80%
Europe	7.75%
Other	0.60%

Fig. 5. The geography of viewers in our trace at the continent-level.

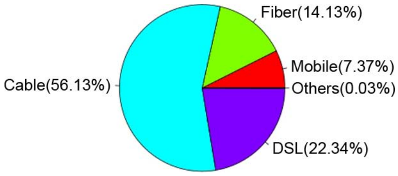


Fig. 6. Connection type as percent of views.

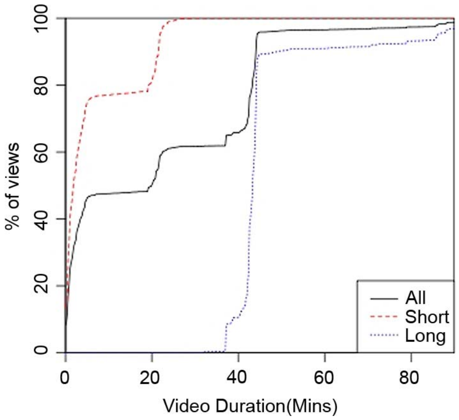


Fig. 7. A CDF of the total video duration. The median duration is 19.92 minutes over all videos, 1.8 minutes for short, and 43.2 minutes for long videos.

providers for a period of 10 days (see Fig. 4). Our data set is extensive and captures 23 million views from 6.7 million unique viewers, where each viewer on average made 3.42 visits over the period and viewed a total of 32.2 min of video. In each visit, there were on average 2.39 views but only 1.96 unique videos viewed, indicating that sometimes the viewer saw the same video twice. The geography of the viewer was mostly concentrated in North America, Europe, and Asia, with small contributions from other continents (see Fig. 5). More than half the views used cable, though fiber, mobile, and DSL were significant. The fiber category consisted mostly of AT&T Uverse and Verizon FiOS that contributed in roughly equal proportion. The other connection types such as dialup were negligible (see Fig. 6). Video duration is the total length (in minutes) of the video (see Fig. 7). We divide the videos into short that have a duration of less than 30 min and long that have a duration of more than 30 min. Examples of short video include news clips, highlight reels for sports, and short television episodes. The median duration was 1.8 min, though the mean duration was longer at 5.95 min. In contrast, long video consists of long television episodes and movies. The median duration for long videos was 43.2 min, and the mean was 47.8 min.

IV. Analysis Techniques

A key goal is to establish a *causal* link between a stream quality metric X and viewer behavior metric Y. The first key step is to establish a *correlational* link between X and Y using the statistical tools for correlation and regression. Next, in accordance with the maxim that “correlation does not imply causation,” we do a more careful analysis to establish causation. We adapt the innovative tool of QED used extensively in the social and medical sciences to problem domains such as ours.

A. Correlational Analysis

To study the impact of a stream quality metric X (say, startup delay) with a viewer behavioral metric Y (say abandonment rate), we start by visually plotting metric X versus metric Y in the observed data. The visual representations are a good initial step to estimating whether or not a correlation exists. As a next step, we also quantify the correlation between X and Y. There are many different ways to calculate the correlation. Primary among them are Pearson’s correlation and Kendall’s correlation that is a type of rank correlation. As observed in [6], Kendall’s correlation is more suitable for a situation such as ours since it does not assume any particular distributional relationship between the two variables. Pearson’s correlation is more appropriate when the correlated variables are approximately linearly related, unlike the relationships that we explore in our work. Kendall’s correlation measures the whether the two variables X and Y are statistically dependent (i.e., correlated) without assuming any specific functional form of their relationship. Kendall’s correlation coefficient τ takes values in the interval $[-1, 1]$ where $\tau = 1$, meaning that X and Y are perfectly concordant, i.e., larger values of X are always associated with larger values for Y; $\tau = -1$, meaning that X and Y are perfectly discordant, i.e., larger values of X are always associated with smaller values of Y; and τ near 0 implying that X and Y are independent.

B. Causal Analysis

A correlational analysis of stream quality metric X (say, startup delay) and a viewer behavior metric Y (say, abandonment rate) could show that X and Y are associated with each other. A primary threat to a causal conclusion that an *independent variable* X causes the *dependent variable* Y is the existence of *confounding variables* that can impact both X and Y. To take a recent example from the medical literature, a study published in *Nature* [11] made the causal conclusion that children who sleep with the light on are more likely to develop myopia later in life. However, as it turns out, myopic parents tend to leave the light on more often, as well as pass their genetic predisposition to myopia to their children. Accounting for the confounding variable of parent's myopia, the causal results were subsequently invalidated or substantially weakened.

More relevant to our own work, let us consider a potential threat to a causal conclusion that a stream quality metric X (say, startup delay) results in a viewer behavior Y (say, abandonment). As a hypothetical example, suppose that mobile users tend to have less patience for videos to start up as they tend to be busy and are "on the go," resulting in greater abandonment. Further assume that mobile users tend to have larger startup delays due to poor wireless connectivity. In this situation, a correlation between startup delay and abandonment may not imply causality unless we can account for the confounding variable of how the viewer is connected to the Internet. In our causal analyses, we systematically identify and account for all or a subset of the following three types of confounding variables as relevant.

1) *Content*: The video[4] being watched could itself influence both quality and viewer behavior. For instance, some videos are more captivating than others, leading viewers to watch more of it. Or, some videos may have higher perceived value than others, leading viewers to tolerate more startup delay. The manner in which the video is encoded and the player heuristic used by the media player could also impact stream quality. For instance, the player heuristics that could differ from one content provider to another specifies how much of the video needs to be buffered before the stream can start up or resume play after rebuffering.
2) *Connection type*: The manner in which a viewer connects to the Internet, both the device used and typical connectivity characteristics, can influence both stream quality and viewer behavior. We use the connection type of the viewer as a confounding variable, where the connection type can take discrete values such as mobile, DSL, cable, and fiber (such as AT&T's Uverse and Verizon's FiOS).
3) *Geography*: Geography of viewer captures several social, economic, religious, and cultural aspects that can influence viewer behavior. For instance, it has been observed by social scientists that the level of patience that consumers exhibit toward a delay in receiving a product varies based on geography of the consumer [12]. Such a phenomena might well be of significance in the extent to which the viewer's behavior is altered by stream quality. In our work, we analyze viewer's geography at the granularity of a country.

1) QED Method: A primary technique for showing that an independent variable X (called the treatment variable) has a causal impact on a dependent variable Y (called the outcome variable) is to design a *controlled experiment*. To design a true experiment in our context, one would have to randomly assign viewers to differing levels of stream quality (i.e., values of X) and observe the resultant viewer behaviors (values of Y). The random assignment in such an experiment removes any systematic bias due to the confounding variables that are threats to a causal conclusion. However, the level of control needed to perform such an experiment at scale for our problem is either prohibitively hard, expensive, or even impossible. In fact, there are legal, ethical, and other issues with intentionally degrading the stream quality of a set of viewers to do a controlled experiment. However, there are other domains where a controlled experiment can be and is performed, e.g., $A|B$ testing of Web page layouts [13].

Given the inability to perform true experiments, we adapt a technique called QED to discover causal relationships from observational data that already exist. QEDs were developed by social and medical scientists, as a similar inability to perform controlled experiments is very common in those domains [8]. In particular, we use a specific type of QED called the matched design [14] where a treated individual (in our case, a view or viewer) is randomly matched with an untreated individual, where both individuals have identical values for the confounding variables. Consequently, any difference in the outcome for this pair can be attributed to the treatment. Our population typically consists of views or viewers and the treatment variable is typically binary. For instance, in Section VII, viewers who experienced "bad" stream quality in the form of a failed visit are deemed to be treated, and viewers who had normal experience are untreated. We form comparison sets by randomly matching each treated viewer with an untreated viewer such that both viewers are as identical as possible on the confounding variables. Needless to say, the more identical the viewers are in each pair, the more effective the matching is in neutralizing the confounding variables. Note that matching ensures that the distributions of the confounding variables in the treated and untreated set of viewers are identical, much as if viewers were randomly assigned to treated and untreated sets in a controlled experiment. Now, by studying the behavioral outcomes of matched pairs, one can deduce whether or not the treatment variable X has a causal effect on variable Y, with the influence of the confounding variables neutralized. Note that treatment variable need not always be stream quality. Depending on the causal conclusion, we could choose the treatment variable to content length or connection type, if we would like to study their impact on viewer behavior.

a) Statistical Significance: It Is important to evaluate whether the results are *statistically significant* or if they could have occurred by random chance. As is customary in hypothesis testing [15], we state a null hypothesis H_o that contradicts the assertion that we want establish. That is, H_o contradicts the assertion that x impacts y and states that the treatment variable x has no impact on the outcome variable y. We then compute the "p-value" defined to be the probability that the null hypothesis H_o is consistent with the observed results. A "low" p-value lets us reject the null hypothesis, bolstering our conclusions from the QED analysis as being statistically significant. However, a "high" p-value would not allow us to reject the null hypothesis. That is, the QED results could have happened through random chance with a "sufficiently" high probability that we cannot reject H_o. In this case, we conclude that the results from the QED analysis are not statistically significant.

The definition of what constitutes a "low" p-value for a result to be considered statistically significant is somewhat arbitrary. The p-value is compared to a chosen significance level α, and it is customary in the medical sciences to conclude that a treatment

[4] Note that our notion of video content is URL-based and thus also incorporates the content provider. If the same movie is available from two content providers, they would constitute two different pieces of content for our analysis.

is effective if the p-value is at most $\alpha = 0.05$. We choose a more stringent $\alpha = 0.01$ as our significance level—a level achievable in our field given the large amount of experimental subjects (tens of thousands treated–untreated pairs), but less achievable in medicine with human subjects (usually in the order of hundreds of treated–untreated pairs). However, our results are unambiguously significant and not very sensitive to the choice of significance level. All our results turned out to be highly significant with p-values of 4×10^{-5} or smaller, except for one conclusion with a larger p-value that we deemed statistically insignificant.

The primary technique that we employ for evaluating statistical significance is the sign test that is a nonparametric test that makes no distributional assumptions and is particularly well suited for evaluating matched pairs in a QED setting [16]. We sketch the intuition of the technique here, while deferring the specifics to the technical sections. For each matched pair (u, v), where u received treatment and v did not receive treatment, we define the differential outcome denoted by *outcome* (u, v) as the numerical difference in the outcome of u and the outcome of v. If H_o holds, then the outcomes of the treated and untreated individuals are identically distributed since the treatment is assumed to have no impact on the outcome. Thus, the differential outcome is equally likely to be a positive number as a negative number. Thus, for n independently selected matched pairs, the number of positive values of the differential outcome (call it X) follows the binomial distribution with n trials and probability $1/2$. In a measured sample consisting of a total of n nonzero values of the differential outcome, suppose that x have positive values. Given that H_o holds, the probability (i.e., p-value) of such an occurrence is at most $\mathrm{Prob}\left(|X - n/2| \geq |x - n/2|\right)$, which is sum of both tails of the binomial distribution. Evaluating the above tail probability provides us the required bound on the p-value.

b) Statistical Power: It is important to design experiments that have "sufficient" statistical power to detect a "significant" effect, assuming such an effect does exist [17]. By convention, sufficient power is often defined as power that is at least 80%. Let an alternate hypothesis H_1 assert that a variable X has an effect of a certain magnitude on a variable Y. The magnitude of the effect posited by H_1 is called the *effect size*. Statistical power is the probability that the null hypothesis H_o is correctly rejected, given that the alternative hypothesis H_1 holds. Three key factors that influence statistical power are: 1) the sample size, which is simply the number of matched pairs in our quasi-experiment; 1) the statistical significance level α , which in our case equals 0.01; and 3) the effect size. The statistical power of the experiment increases with both the number of matched pairs and the effect size. Our aim is to design experiments with sufficient power of at least 80% for the effect sizes of real-world significance. In this regard, it is important to recall that small changes in viewer behavior can lead to large changes in monetization, hence even an effect size of a few percentage points is of significance.

Our primary use of power analysis is to ensure that our sample sizes are sufficiently large to detect the effect sizes of interest. As an example, in the viewer abandonment experiments in Section V, each matched pair (u, v) has an outcome of either $+1$ or -1. The effect size is the difference between the percentage of matched pairs with positive outcomes and pairs with negative outcomes. In Fig. 8, we plot the number of matched pairs required for detecting a given effect size with power of at least 80%. Note that detecting a smaller effect of 2% requires a larger sample size of about 29 000 pairs, but detecting a "larger" effect of 5% requires just 4600 pairs. As we show later, all our experiments are designed to have sufficient power to detect the effect sizes of interest.

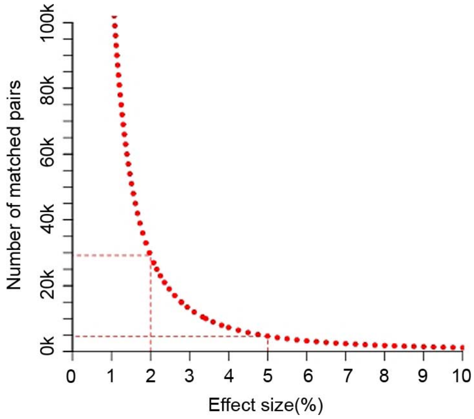


Fig. 8. Minimum number of matched pairs required in a viewer abandonment experiment to detect a given effect size with statistical power at least 80%.

c) Some Caveats: It is important to understand the limitations of our QED tools, or for that matter *any* experimental technique of inference. Care should be taken in designing the quasi-experiment to ensure that the major confounding variables are explicitly or implicitly captured in the analysis. If there exists confounding variables that are not easily measurable (example, the age and gender[5] of the viewer) and/or are not identified and controlled, these unaccounted dimensions could pose a risk to a causal conclusion, if indeed they turn out to be significant. *Our work on deriving a causal relationship by systematically accounting for the confounding variables must not be viewed as a definitive proof of causality, as indeed there can be no definitive proof of causality, but rather, our work increases the confidence in a causal conclusion by accounting for potential major sources of confounding.* This is of course a general caveat that holds for all domains across the sciences that attempt to infer causality from observational data.

In the context of these caveats, it is instructive to compare our approach to a more controlled study that one could undertake for understanding the causal rules of viewer behavior. The advantage of a controlled study is that more parameters of the viewer are knowable such as age, gender, and education level. However, a controlled study has the disadvantage of not studying the viewers "in the wild," and the population studied is necessarily much smaller (in the hundreds) and much less representative of the wide range of users, geographies, devices, video formats, genre, and content that exist in the real world. We believe that

[5] Even though we cannot measure age and gender, certain types of influences of these unmeasured parameters are already accounted for by matching the content and connection type of the viewers, e.g., older people are more likely to watch certain shows (say, *60 Minutes*) or are less likely to watch on certain devices (say, mobile).

both controlled studies and studies in the wild such as ours have a place in understanding the causal rules of viewer behavior.

V. Viewer Abandonment

We address the question of how long a viewer will wait for the stream to start up, a question of great importance that has not been studied systematically to our knowledge. However, the analogous problem of how long a user will wait for Web content to download has received much attention. In 2006, Jupiter Research published a study based on interviewing 1058 online shoppers and postulated what is known in the industry as the "4-second rule" that states that an average online shopper is likely to abandon a Web site if a Web page does not download in 4 s [18]. However, a recent study [19] implied that the users have become impatient over time and that even a 400-ms delay can make users search less. Our motivation is to derive analogous rules for streaming where startup delay for video is roughly analogous to download time for Web pages.

Assertion 5.1: An increase in startup delay causes more abandonment of viewers.

To investigate if our assertion holds, we classify each view into 200-ms buckets based on their startup delay. We then compute for each bucket the percentage of views assigned to that bucket that were abandoned. The percent of abandoned views and startup delay are positively correlated with a Kendall correlation of 0.72.

Suppose now that we build a media delivery service that provides a startup delay of exactly x seconds for every view. What percent of views delivered by this system will be abandoned before the stream starts up? To *estimate* this metric, we define a function called $\mathrm{AbandonmentRate}(x)$ that equals

$$100 \times \mathrm{Impatien}t(x)/(\mathrm{Impatient}(x) + \mathrm{Patient}(x))$$

where $\mathrm{Impatient}(x)$ is all views that were abandoned after experiencing less than x seconds of startup delay and $\mathrm{Patient}(x)$ are views where the viewer waited at least x time without abandoning. That is, $\mathrm{Impatient}(x)$ (resp., $\mathrm{Patient}(x)$) corresponds to views where the viewer did not (resp., did) demonstrate the patience to hold on for x seconds without abandoning. Note that a view in $\mathrm{Patient}(x)$ could still have been abandoned at some time greater than x. Also, note that a view where the video started to play before x seconds does not provide any information on whether the viewer would have waited until x seconds or not, and so is considered neither patient or impatient. Fig. 9 shows the abandonment rate estimated from our data, which is near zero for the first 2 s, but starts to rise rapidly as the startup delay increases. Fitting a simple regression to the initial part of the curve shows that abandonment rate increases by 5.8% for each 1-s increase in startup delay.

Assertion 5.2: Viewers are less tolerant of startup delay for short videos in comparison to longer videos.

Researchers who study the psychology of queuing [20] have shown that people have more patience for waiting in longer queues if the perceived value of the service that they are waiting for is greater. Duration of the service often influences its perceived value with longer durations often perceived as having greater value. People often tolerate the 30-min delay for the checkin process for a 4-h plane ride but would find the same wait excessive for a 10-min bus ride. On the same principle, is it true that viewers would be more patient for the video to start up if they expect to be watching the video for a longer period of time?

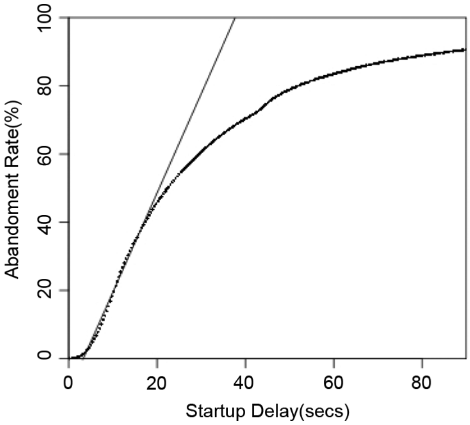


Fig. 9. Viewers start to abandon the video if the startup delay exceeds about 2 s. Beyond that point, a 1-s increase in delay results in roughly a 5.8% increase in abandonment rate.

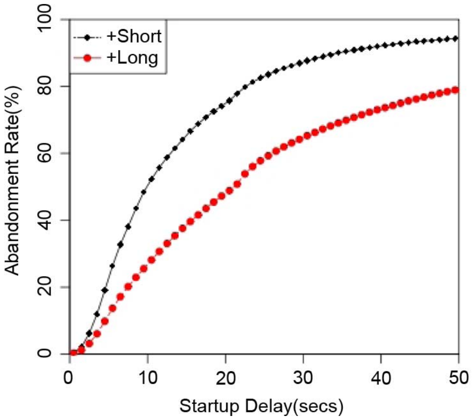


Fig. 10. Viewers abandon at a higher rate for short videos than for long videos.

To investigate our assertion, we first classify the views based on whether the content is short with duration smaller than 30 min (e.g., news clip) or long with duration longer than 30 min (e.g., movies). The Kendall correlations between the two variables, percent of abandoned videos and startup delay, were 0.68 and 0.90 for short and long videos, respectively, indicating a strong correlation for each category. Furthermore, Fig. 10 shows abandonment rate for each type of content as a function of the startup delay. One can see that viewers typically abandon at a larger rate for short videos than for long videos.

Assertion 5.3: Viewers watching videos on a better connected computer or device have less patience for startup delay and so abandon sooner.

The above assertion is plausible because there is some evidence that users who expect faster service are more likely to be disappointed when that service is slow. In fact, this is often touted as a reason for why users are becoming less and less able to tolerate Web pages that download slowly. To study whether or not this is true in a scientific manner, we segment our views

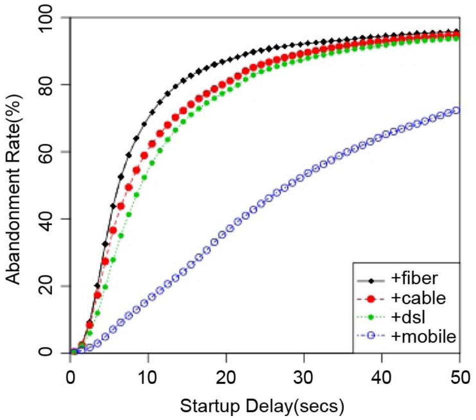


Fig. 11. Viewers who are better connected abandon sooner.

into four categories based on their connection type that indicates how the corresponding viewer is connected to the Internet. The categories in roughly the increasing order of connectivity are mobile, DSL, cable modem, and fiber (such as Verizon FIOS or AT&T Uverse). In all four categories, we see a strong correlation between the two variables, percent of abandoned views, and startup delay. The Kendall correlations for mobile, DSL, cable, and fiber are 0.68, 0.74, 0.71, and 0.75, respectively. Furthermore, in Fig. 11, we show the abandonment rate for each connection type. We can see that viewers abandon significantly less on mobile in comparison to the other categories, for a given startup delay. Some difference in abandonment is discernible between the other categories in the rough order of cable, DSL, and fiber, though they are much smaller.

A. QED for Assertion 5.2

First, we devise a QED to study the impact of content length on abandonment (Assertion 5.2). Therefore, we make the content length (long or short) the treatment variable, and the outcome measures patience of the viewer to startup delay. The viewer's patience to startup delay can be influenced by both the viewer's geography and connection type, which we use as the confounding variables. Specifically, we form matched pairs (u, v) such that view u is a short video that was abandoned, view v is a long video that was abandoned, u and v are watched by viewers from the same geography, and the viewers have the same connection type. We also make the viewers and the content they are watching more similar by ensuring that both u and v are watching videos from the same content provider. The matching algorithm is described as follows.

1) *Match step*: Let the treated set T be all abandoned views for short content, and let untreated set C be all the abandoned views for long content. For each $u \in T$, we pick uniformly and randomly a $v \in C$ such that u and v belong to viewers in the same geography, have the same connection type, and are watching content from the same content provider. The matched set of pairs $M \subseteq T \times C$ have the same attributes for the confounding variables and differ only on the treatment.
2) *Score step*: For each pair $(u, v) \in M$, we compute an $outcome(u, v)$ to be $+1$ if u was abandoned with a smaller startup delay than v. If u was abandoned with a larger startup delay than v, then $outcome(u, v) = -1$. Moreover, $outcome(u, v) = 0$, if the startup delays when u and v were abandoned are equal. Now

$$Net\ Outcome = \left(\frac{\sum_{(u,v)\in M} outcome(u, v)}{|M|} \right) \times 100.$$

Note that a positive value for net outcome provides positive (supporting) evidence for Assertion 5.2, while a negative value provides negative evidence for the assertion. The results of the matching algorithm produced a net outcome of 11.5%. The net outcome shows that the matched pairs that support Assertion 5.2 exceed those that negate the assertion by 11.5%. The positive net outcome provides evidence of causality that was not provided by the prior correlational analysis alone by eliminating the threats posed by the identified confounding variables.

To derive statistical significance of the above QED result, we formulate a null hypothesis H_o that states that the treatment (long versus short video) has no impact on abandonment. If H_o holds, the $outcome(u, v)$ is equally likely to be positive $(+1)$ as negative (-1). We now use the sign test that we described in Section IV-B to derive a bound on the p-value. Since we matched $n = 78\,840$ pairs, if H_o holds, the expected number pairs with a positive outcome is $n/2 = 78\,840/2 = 39\,420$. Our observational data however had $x = 43\,954$ pairs with positive scores, i.e., $x - n/2 = 4534$ pairs in excess of the mean. We bound the p-value by showing that it is extremely unlikely to have had 4534 positive pairs in excess of the mean by computing the two-sided tail of the binomial distribution with n trials and probability 1/2

$$\text{p-value} \leq \text{Prob}\left(|X - \frac{n}{2}| \geq |x - \frac{n}{2}\right) \leq 3.3 \times 10^{-229}. \quad (1)$$

The above bound for the p-value is much smaller than the required significance level of 0.01 and leads us to reject the null hypothesis H_o. Thus, we conclude that our QED analysis is statistically significant. Finally, note that our experiment has a sample size of 78 K pairs and can detect effect size as small as 1.2% with statistical power of at least 80% (cf. Fig. 8). Furthermore, for the observed effect size of 11.5%, power is nearly 100%.

B. QED for Assertion 5.3

To investigate a causal conclusion for Assertion 5.3, we set up a QED where the treatment is the connection type of the user and the outcome measures the relative tolerance of the viewer to startup delay. For each pair of network types A and B, we run a matching algorithm where the treated set T is the set of all abandoned views with connection type A and untreated set is all abandoned views with connection type B. The matching algorithm used is identical to the one described earlier except that the match criterion in step 1 is changed to match for identical content and identical geography. That is, for every matched pair (u, v), view u has network type A and view v has network type B, but both are views for the same video and belong to viewers in the same geography.

The results of the matching algorithm are shown in Fig. 12. For instance, our results show that the likelihood that a mobile

Untreated Treated	dsl	cable	fiber
mobile	33.81	35.40	38.25
dsl	-	-0.75	2.67
cable	-	-	3.65

Fig. 12. Net QED outcomes support the causal impact of connection type on viewer patience for startup delay, though the impact is more pronounced between mobile and the rest. The p-value for all entrees are very small ($< 10^{-17}$), except dsl-versus-cable (0.06) and dsl-versus-fiber (4.6×10^{-5}).

viewer exhibited more patience than a fiber viewer is greater than the likelihood that opposite holds by a margin of 38.25%. Much as in Section V-A, we use the sign test to compute the p-value for each QED outcome in the table. The number of matched pairs was large for each QED: Comparison of cable to the other three connection types ranged from 56 639 to 62 614 pairs each, and the other comparisons ranged from 15 239 to 24 482 pairs. All QED outcomes in Fig. 12 turned out to be statistically significant with exceedingly small p-values, except the dsl-versus-cable comparison that was inconclusive. Specifically, our results show that a that a mobile viewer exhibits more patience than other (non-mobile) viewers, and the result holds with exceedingly small p-values ($< 10^{-17}$). Our results also provide strong evidence for DSL and cable users being more patient than fiber users, though the p-value for the dsl-versus-fiber was somewhat larger (4.6×10^{-5}) but still statistically significant. The dsl-versus-cable comparison was however inconclusive and not statistically significant as the p-value of the score was 0.06, which is larger than our required significance level of 0.01. Finally, note that our sample size varied from 15 K pairs (mobile-versus-fiber) to 62 K pairs (dsl-versus-cable). As a result, the smallest effect size detectable by our experiments with a power of at least 80% varied from 2.8% to 1.4%, respectively (cf. Fig. 8). For the observed values of effect size, with the exception of dsl-versus-cable, the power values ranged from 93% to nearly 100%.

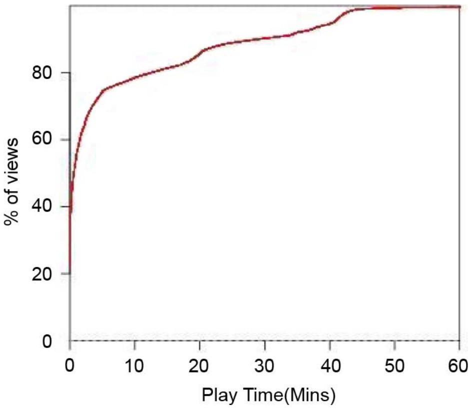


Fig. 13. Significant fraction of the views have small duration.

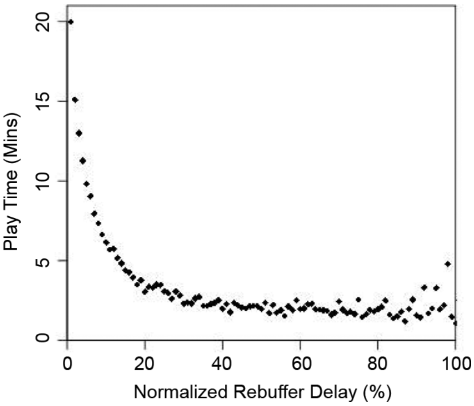


Fig. 14. Correlation of normalized rebuffer delay with play time.

VI. Viewer Engagement

We study the extent to which a viewer is engaged with the video content of the content provider. A simple metric that measures engagement is play time. Here, we study play time on a per-view basis, though one could study play time aggregated over all views of a visit (called *visit play time*) or play time aggregated over all visits of a viewer (called *viewer play time*). Fig. 13 shows the cumulative distribution function (CDF) of play time over our entire data set. A noticeable fact is that a significant number of views have very small play time with the median play time only 35.4 s. This is likely caused by "video surfing," where a viewer quickly views a sequence of videos to see what might of interest to him/her before settling in on the videos that he/she wants to watch. The fact that a viewer watched on average of 22.48 min per visit (cf. Fig. 4) is consistent with this observation. Play time is clearly impacted by both the interest level of the viewer in the video and the stream quality. Viewer interest could itself be a function of complex factors. For instance, Italian viewers might be more interested in soccer World Cup videos than American viewers, even more so if the video is of a game where Italy is playing. In understanding the impact of stream quality on viewer engagement, the challenge is to neutralize the bias from confounding variables *not* related to stream quality such as viewer interest, geography, and connection type. Since more rebuffer delay is expected of videos with a longer duration, we use normalized rebuffer delay[6] that equals $100 \times (\text{rebuffer delay}/\text{video duration})$.

Assertion 6.1: An increase in (normalized) rebuffer delay can cause a decrease in play time.

To evaluate the above assertion, we first classify views by bucketing their normalized rebuffer delay into 1% buckets. Then, we compute and plot the average play time for all views within each bucket (see Fig. 14). The decreasing trend visualizes the negative correlation that exists between normalized rebuffer delay and play time. The Kendall correlation between the two metrics is -0.421, quantifying the negative correlation.

A. QED Analysis

To examine the causality of Assertion 6.1, we devise a QED where the treatment set T consists of all views that suffered normalized rebuffer delay more than a certain threshold $\gamma\%$. Given a value of γ as input, the treated views are matched with untreated views that did not experience rebuffering as follows.

[6] Note that normalized rebuffer delay can go beyond 100% if we rebuffer for longer than the total duration of the video.

Normalized Rebuffer Delay γ (percent)	Net Outcome (percent)	P-Value
1	5.02	$< 10^{-143}$
2	5.54	$< 10^{-123}$
3	5.7	$< 10^{-87}$
4	6.66	$< 10^{-86}$
5	6.27	$< 10^{-57}$
6	7.38	$< 10^{-47}$
7	7.48	$< 10^{-36}$

Fig. 15. Viewer who experienced more rebuffer delay on average watched less video than an identical viewer who had no rebuffer.

1) *Match step*: We form a set of matched pairs M as follows. Let T be the set of all views who have a normalized rebuffer delay of at least $\gamma\%$. For each view u in T, suppose that u reaches the normalized rebuffer delay threshold $\gamma\%$ when viewing the tth second of the video, i.e., view u receives treatment after watching the first t seconds of the video, though more of the video could have been played after that point. We pick a view v uniformly and randomly from the set of all possible views such that the following applies.
    a) The viewer of v has the same geography, connection type, and is watching the same video as the viewer of u.
    b) View v has played at least t seconds of the video without rebuffering until that point.
2) *Score step*: For each pair $(u, v) \in M$, we compute

$$outcome(u, v) = \frac{\text{play time of } v - \text{play time of } u}{\text{video duration}}.$$

$$Net\ Outcome = \left(\frac{\sum_{(u,v)\in M} outcome(u, v)}{|M|} \right) \times 100. \tag{2}$$

Note that the closer we can make the matched views u and v in variables other than the treatment, the more accurate our QED results. Though as a practical matter, adding too many matching parameters can highly reduce the availability of matches, eventually impacting the statistical significance of the results. It is worth noting step 1(b), where we ensure that v watches the video to at least the same point as when u first received treatment. Thus, at the time both u and v play the tth second of the video, they have viewed the same content, and the only difference between them is one had rebuffering and the other did not. The net outcome of the matching algorithm can be viewed as the difference in the play time of u and v expressed as a percent of the video duration. Fig. 15 shows that on average a view that experienced normalized rebuffer delay of 1% or more played 5.02% of less of the video. There is a general upward trend in the net outcome when the treatment gets harsher with increasing values of γ. Much as in Section V-A, we use the sign test to compute the p-values for each QED outcome. All p-values were extremely small as shown in Fig. 15, making the results statistically significant. Finally, minimum effect sizes detectable with sufficient power varied from 1.25% for the experiment with the largest sample size ($\gamma = 1\%$) to 3.6% for the experiment with the smallest one ($\gamma = 7\%$). Furthermore, for the observed effect sizes in Fig. 15, power is nearly 100%.

VII. Repeat Viewership

We study the viewers who, after watching videos on a content provider's site, return after some period of time to watch more. Repeat viewers are valued highly valued by media content providers as these viewers are more engaged and more loyal to the content provider's site. Even a small decrease (or increase) in the return rate of viewers can have a large impact on the business metrics of the content provider. Clearly, a number of factors, including how captivating the video content is to the viewer, influence whether or not a viewer returns. However, we show that stream quality can also influence whether or not a viewer returns.

The most drastic form of quality degradation is failure when a viewer is unable to play a video successfully. Failures can be caused by a number of issues, including problems with the content (broken links, missing video files, etc.), the client software (media player bugs, etc.), or the infrastructure (network failure, server overload, etc.). More frustrating than a failed view is a failed visit where a viewer tries to play videos from the content providers site but fails and leaves the site immediately after the failure, presumably with some level of frustration. (Note that the definition of a failed visit does not preclude successful views earlier in that visit before the last view(s) that failed.) We focus on the impact of a failed visit experienced by a viewer on his/her likelihood of returning to the content provider's site.

Assertion 7.1: A viewer who experienced a failed visit is less likely to return to the content provider's site to view more videos within a specified time period than a similar viewer who did not experience a failed visit.

To examine if the above assertion holds, we classify each of our views as either failed or normal (i.e., not failed). For each failed visit (resp., normal visit), we compute the *return time*, which is defined to be the next time the viewer returns to the content provider's site. (Return time could be infinite if they do not return to the site within our trace window.) Fig. 16 shows the CDF of the return time for both failed visits and normal visits. It can be seen that there is significant reduction in the probability of return following a failed visit as opposed to a normal one. For instance, the probability of returning within 1 day after a failed visit is 8.0% versus 11% after a normal one. Likewise, the probability of returning within 1 week after a failed visit is 25% versus 27% after a normal one.

A. QED Analysis

We perform a QED analysis to strengthen Assertion 7.1 by considering viewers[7] with a failed visit to be the treated set T. For each $u \in T$, we find a matching viewer v that is similar to u in all the confounding variables. As before, we ensure that viewers u and v are from the same geography, have the same connection type, and are viewing content from the same content provider. However, there is a subtle characteristic that needs to be matched. Specifically, we need also ensure that the

[7] Note that, in this matching, we are matching viewers and not views as we are evaluating the repeat viewership of a viewer over time.

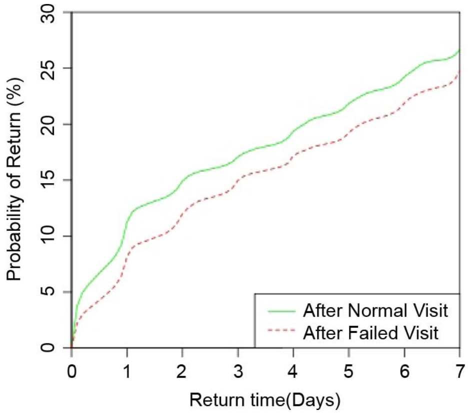


Fig. 16. Probability of the return after a failed visit and after a normal visit. The probability of returning within a specified return time is distinctly smaller after a failed visit than after a normal one.

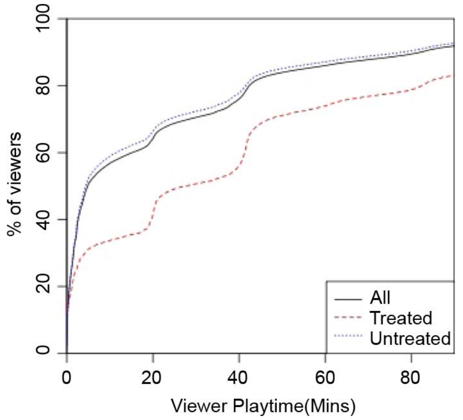


Fig. 17. CDF of the viewer play time for all, treated, and untreated viewers.

propensity of u to watch videos *prior* to when u received treatment is equivalent to the corresponding propensity of v. This ensures that any differential behavior *after* the treatment can be attributed to the treatment itself.

To reinforce the last point, a viewer who watches more video at a site is more likely to have had a failed view. Therefore, the treated set T of viewers has a bias toward containing more frequent visitors to site who also watch more video. Fig. 17 shows the CDF of the aggregate play time of a viewer across all visits. It can be seen that the treated set T has viewers who have watched for more time in aggregate. To neutralize this effect, we match on the number of prior visits and aggregate play time in step 1(c) below and make them near identical, so that we are comparing two viewers who have exhibited similar propensity to visit the site *prior* to treatment. The use of a similarity metric of this kind for matching is common in QED analysis and is similar in spirit to propensity score matching of [21].

The matching algorithm follows.

1) *Match step*: We produce a matched set of pairs M as follows. Let T be the set of all viewers who have had a failed visit. For each $u \in T$, we pick the first failed visit of viewer u. We then pair u with a viewer v picked uniformly and randomly from the set of all possible viewers such that the following applies.

Return Time δ (in days)	Outcome (percent)	P-Value
1	2.38	$< 10^{-57}$
2	2.51	¡10^{-51}
3	2.42	$< 10^{-44}$
4	2.35	$< 10^{-37}$
5	2.15	$< 10^{-22}$
6	1.90	$< 10^{-11}$
7	2.32	$< 10^{-6}$

Fig. 18. Viewer who experienced a failed visit is less likely to return within a time period than a viewer who experienced a normal visit.

   a) Viewer v has the same geography, same connection type as u, and is watching the content from the same content provider as u.
   b) Viewer v had a normal visit at about the same time (within ±3 h) as the first failed visit of viewer u. We call the failed visit of u and the corresponding normal visit of v that occurred at a similar time as matched visits.
   c) Viewer u and v have the same number of visits and about the same total viewing time (±10 min) prior to their matched visits.

2) *Score step*: For each pair $(u, v) \in M$ and each return time δ, we assign $outcome(u, v, \delta)$ to -1 if u returns within the return time and v does not, $+1$ if v returns within the return time and u does not, and 0 otherwise

$$Net\ Outcome(\delta) = \left(\frac{\sum_{(u,v)\in M} outcome(u, v, \delta)}{|M|}\right) \times 100.$$

Fig. 18 shows the outcome of the matching algorithm for various values of the return time (δ). The positive values of the outcome provide strong evidence of the causality of Assertion 7.1 since it shows that viewers who experienced a normal visit returned more than their identical pair with a failed visit. To take a numerical example, for $\delta = 1$ day, 458 621 pairs were created. The pairs where the normal viewer returned but its identical failed pair did not exceed the pairs where the opposite happened. The amount of pairs in excess was 10 999 pairs, which is 2.38% of the total pairs. Using the sign test, we show that the p-value is extremely small (2.2×10^{-58}), providing strong evidence of statistical significance for the outcome. Note that as δ increases, the outcome score remained in a similar range. However, one would expect that for very large δ values, the effect of the failed event should wear off, but we did not analyze traces that were long enough to evaluate if such a phenomenon occurs. All p-values remain significantly smaller than our threshold of significance of 0.01, allowing us to conclude that the results are statistically significant. Finally, the sample size of our experiments were large, ranging from 45 K to 458 K pairs. The smallest effect sizes detectable with sufficient power varied from 0.5% for the experiment with the largest sample size ($\delta = 1$ day) to 1.6% for the experiment with smallest sample size ($\delta = 7$ days). Furthermore, for the observed effect sizes in Fig. 18, power is nearly 100%.

VIII. Related Work

The quality metrics considered here have more than a dozen years of history within industry where early measurement systems used synthetic "measurement agents" deployed around the world to measure metrics such as failures, startup delay, rebuffering, and bit rate, for example, Akamai's Stream Analyzer measurement system [1], [9]. There have been early studies at Akamai on streaming quality metrics using these tools [10]. However, truly large-scale studies were made possible only with the recent advent of client-side measurement technology that could measure and report detailed quality and behavioral data from actual viewers. To our knowledge, the first important large-scale study and closest in spirit to our work is the study of viewer engagement published in 2011 [6] that shows several correlational relationships between quality (such as rebuffering), content type (such as live, short/long VoD), and viewer engagement (such as play time). A recent sequel to the above work [7] studies the use of quality metrics to enhance video delivery. A key differentiation of our work from prior work is our focus on establishing *causal* relationships, going a step beyond just correlation. While our viewer engagement analysis was also correlationally established in [6], our work takes the next step in ascertaining the causal impact of rebuffering on play time. Besides our results on viewer engagement, we also establish key assertions pertaining to viewer abandonment and repeat viewership that are the first quantitative results of its kind. However, it must be noted that [6] studies a larger set of quality metrics, including join time, average bit rate, and rendering quality, and a larger class of videos including live streaming, albeit without establishing causality.

The work on quasi-experimental design in the social and medical sciences has a long and distinguished history stretching several decades that is well documented in [8], though its application to data mining is more recent. In [14], the authors use QEDs to answer questions about user behavior in social media such as Stack Overflow and Yahoo Answers.

There are a number of other studies on perceived quality, though they tend to be small-scale studies or do not link the quality to user behavior [22], [23]. There has also been prior work for other types of systems—for instance, the relationship between page download times and user satisfaction [24] for the Web and quantifying user satisfaction for Skype [25]. There has also been work on correlating QoS with QoE (quality of experience) for multimedia systems using human subjects [26]. These of course have a very different focus from our work and do not show causal impact. There has been significant amount of work in workload characterization of streaming media, P2P, and Web workloads [27], [28]. Even though we do characterize the workload to a degree, our focus is quality and viewer behavior.

IX. Conclusion

Our work is the first to demonstrate a *causal* nexus between stream quality and viewer behavior. The results presented in our work are important because they are the first quantitative demonstration that key quality metrics causally impact viewer behavioral metrics that are key to both content providers and CDN operators. As all forms of media migrate to the Internet, both video monetization and the design of CDNs will increasingly demand a true causal understanding of this nexus. Establishing a causal relationship by systematically eliminating the confounding variables is immensely important, as mere correlational studies have the potential costly risk of making incorrect conclusions.

Our work breaks new ground in understanding viewer abandonment and repeat viewership. Furthermore, it sheds more light on the known correlational impact of quality on viewer engagement by establishing its causal impact. Our work on startup delay shows that more delay causes more abandonment; for instance, a 1-s increase in delay increases the abandonment rate by 5.8%. We also showed the strong impact of rebuffering on the video play time. For instance, we showed that a viewer experiencing a rebuffer delay that equals or exceeds 1% of the video duration played 5.02% less of the video in comparison to a similar viewer who experienced no rebuffering. Finally, we examined the impact of failed visits and showed that a viewer who experienced failures is less likely to return to the content provider's site in comparison to a similar viewer who did not experience failures. In particular, we showed that a failed visit decreased the likelihood of a viewer returning within a week by 2.32%. While reviewing these results, it is important to remember that small changes in viewer behavior can lead to large changes in monetization since the impact of a few percentage points over tens of millions of viewers can accrue to large impact over a period of time.

As more and more data become available, we expect that our QED tools will play an increasing larger role in establishing key causal relationships that are key drivers of both the content provider's monetization framework and the CDN's next-generation delivery architecture. The increasing scale of the measured data greatly enhances the statistical significance of the derived conclusions and the efficacy of our tools. Furthermore, we expect that our work provides an important tool for establishing causal relationships in other areas of measurement research in networked systems that have so far been limited to correlational studies.

Acknowledgment

The authors thank E. Bommaiah, H. Kammanahalli, and D. Jensen for insightful discussions. Furthermore, they thank Prof. M., Prof. M. Reisslein, and the anonymous referees for their detailed comments that resulted in significant improvements to the paper. Any opinions expressed in this work are solely those of the authors and not necessarily those of Akamai Technologies.

References

[1] R. Sitaraman and R. Barton, "Method and apparatus for measuring stream availability, quality and performance," US Patent 7,010,598, Feb. 2003.

[2] J. Dilley, B. M. Maggs, J. Parikh, H. Prokop, R. K. Sitaraman, and W. E. Weihl, "Globally distributed content delivery," *IEEE Internet Comput.*, vol. 6, no. 5, pp. 50–58, Sep.–Oct. 2002.

[3] E. Nygren, R. Sitaraman, and J. Sun, "The Akamai network: A platform for high-performance Internet applications," *Oper. Syst. Rev.*, vol. 44, no. 3, pp. 2–19, 2010.

[4] K. Andreev, B. Maggs, A. Meyerson, and R. Sitaraman, "Designing overlay multicast networks for streaming," in *Proc. 15th Annu. ACM Symp. Parallel Algor. Archit.*, 2003, pp. 149–158.

[5] L. Kontothanassis, R. Sitaraman, J. Wein, D. Hong, R. Kleinberg, B. Mancuso, D. Shaw, and D. Stodolsky, "A transport layer for live streaming in a content delivery network," *Proc. IEEE*, vol. 92, no. 9, pp. 1408–1419, Sep. 2004.

[6] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2011, pp. 362–373.

[7] 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. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2012, pp. 359–370.

[8] W. Shadish, T. Cook, and D. Campbell*, Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA, USA: Houghton Mifflin, 2002.

[9] Akamai, Cambridge, MA, USA, "Akamai stream analyzer service description," 2009 [Online]. Available: http://www.akamai.com/dl/feature_sheets/Stream_Analyzer_Service_Description.pdf

[10] Akamai, Cambridge, MA, USA, "Akamai streaming: When performance matters," 2004 [Online]. Available: http://www.akamai.com/dl/whitepapers/Akamai_Streaming_Performance_Whitepaper.pdf

[11] G. E. Quinn, C. H. Shin, M. G. Maguire, and R. A. Stone, "Myopia and ambient lighting at night," *Nature*, vol. 399, no. 6732, pp. 113–113, 1999.

[12] H. Chen, S. Ng, and A. Rao, "Cultural differences in consumer impatience," *J. Market. Res.*, vol. XLII, pp. 291–301, 2005.

[13] R. Kohavi, R. Longbotham, D. Sommerfield, and R. Henne, "Controlled experiments on the Web: survey and practical guide," *Data Mining Knowl. Discovery*, vol. 18, no. 1, pp. 140–181, 2009.

[14] H. Oktay, B. Taylor, and D. Jensen, "Causal discovery in social media using quasi-experimental designs," in *Proc. 1st Workshop Social Media Anal.*, 2010, pp. 1–9.

[15] E. Lehmann and J. Romano*, Testing Statistical Hypotheses*. New York, NY, USA: Springer-Verlag, 2005.

[16] D. Wolfe and M. Hollander*, Nonparametric Statistical Methods*. New York, NY, USA: Wiley, 1973.

[17] J. Cohen*, Statistical Power Analysis for the Behavioral Sciences*. Evanston, IL, USA: Routledge, 1988.

[18] Akamai, Cambridge, MA, USA, "Retail Web site performance," 2006 [Online]. Available: http://www.akamai.com/html/about/press/releases/2006/press_110606.html

[19] S. Lohr, "For impatient Web users, an eye blink is just too long to wait," *New York Times*, Feb. 2012.

[20] R. Larson, "Perspectives on queues: Social justice and the psychology of queueing," *Oper. Res.*, vol. 35, no. 6, pp. 895–905, 1987.

[21] P. Rosenbaum and D. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *Amer. Statist.*, vol. 39, no. 1, pp. 33–38, 1985.

[22] S. Gulliver and G. Ghinea, "Defining user perception of distributed multimedia quality," *Trans. Multimedia Comput., Commun., Appl.*, vol. 2, no. 4, pp. 241–257, 2006.

[23] M. Claypool and J. Tanner, "The effects of jitter on the peceptual quality of video," in *Proc. 7th ACM Int. Conf. Multimedia (Part 2)*, 1999, pp. 115–118.

[24] N. Bhatti, A. Bouch, and A. Kuchinsky, "Integrating user-perceived quality into Web server design," *Comput. Netw.*, vol. 33, no. 1, pp. 1–16, 2000.

[25] K. Chen, C. Huang, P. Huang, and C. Lei, "Quantifying Skype user satisfaction," *Comput. Commun. Rev.*, vol. 36, no. 4, pp. 399–410, 2006.

[26] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang, "Quality of experience in distributed interactive multimedia environments: toward a theoretical framework," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 481–490.

[27] K. Sripanidkulchai, B. Maggs, and H. Zhang, "An analysis of live streaming workloads on the Internet," in *Proc. 4th ACM SIGCOMM Conf. Internet Meas.*, 2004, pp. 41–54.

[28] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 1–14.

S. Shunmuga Krishnan received the B.E. degree in computer science from the PSG College of Technology, Coimbatore, India, in 2006.

He is currently a Senior Software Engineer with Akamai Technologies, Bangalore, India, in the Media Analytics Division. His research interests include machine learning, computer vision, and data analytics.

Ramesh K. Sitaraman (M'10) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, India, and the Ph.D. degree in computer science from Princeton University, Princeton, NJ, USA, in 1993.

He is currently with the School of Computer Science, University of Massachusetts, Amherst, MA, USA. As a principal architect, he helped create the Akamai network and is an Akamai Fellow. He is best known for his pioneering role in helping build the first large content delivery networks (CDNs) that currently deliver much of the world's Web content, streaming videos, and online applications. His research spans all aspects of Internet-scale distributed systems, including algorithms, architectures, performance, energy efficiency, user behavior, and economics.

Prof. Sitaraman is a recipient of an NSF CAREER Award and a Lilly Fellowship.