**Summary: CNS Core: Small: A Unified Approach to Internet Performance Measurement**

**Overview**

The use of online applications, such as video streaming and video conferencing, significantly surged during the COVID-19 pandemic to facilitate virtual meetings and classroom instruction. Service providers and regulators face increasing pressure to understand the quality of experience (QoE) of these applications. Poor QoE can greatly degrade the usability of applications and thus impair the effectiveness of communication and learning. But scientific measurement of QoE faces a daunting set of challenges.

First, unlike quality of service (QoS) measurements (e.g., latency, loss, throughput) which are relatively straightforward to write software tools to measure, QoE metrics reflect inherently subjective factors of human-computer interaction, including a user's past experience and expectations. Second, studying correlations between QoS and QoE measurements – although critical to our understanding of the impact of performance degradations on users – is challenged by today's traffic engineering practices, which often cause measurement traffic and video traffic to traverse different paths. To overcome this obstacle, researchers have begun to use public crowdsourcing platforms to recruit a diverse set of human subjects willing to report their QoE with online applications. But thus far none of these studies have performed simultaneous network-level measurements to/from those subjects. Another gap particularly relevant to today's world: current QoE crowdtesting approaches cannot measure QoE of cloud-based video conferencing, because users carry out assessments independently. Systematic scientific study of the QoE of modern video conferencing applications requires inducing conversation and interaction among users, as well as capturing network path performance metrics between the cloud and multiple users.

We propose to develop and apply a fundamentally new approach to Internet performance measurement research that will narrow these critical gaps. First, we will extend our unified assessment platform (QUINCE- Quality of Internet Consumer Experience) to crowdsource novel experiments for measuring the QoE of video streaming and video conferencing. We will leverage gamification techniques to incentivize subject participation and improve overall efficiency of experimental effort. Our second task will combine analyses of QUINCE data and external network topology and performance measurements to diagnose QoE degradations induced by congestion events on interconnections and last-mile access links.

**Intellectual Merit**

The intellectual merit of this work lies in our novel methodologies to integrate research areas–QoE crowdtesting and network performance measurement–in order to identify network bottlenecks and their impact on the real-world QoE of popular applications. Our research will yield innovative techniques to obtain high quality QoE assessments for video streaming and conferencing applications. We will develop methods to mitigate issues caused by fluctuations in subject availability. We will strategically select measurement targets based on performance of paths observed from existing Internet measurement platforms. This coupling will reveal critical performance information at the time of user QoE assessments.

**Broader Impacts**

We will broadly disseminate our methods and results to interested communities via publications, conferences, workshops, and our web site. Our analysis of QoE degradation will enable us to realistically understand the performance of essential tools for remote learning and telecommuting in the United States. The data that we will collect can facilitate training and improve the accuracy of machine learning and artificial intelligence models to infer QoE from network QoS metrics. We will mentor a diverse set of undergraduate/high school students, who will learn about real-world Internet measurement through participating in platform development.

**Keywords:** quality of experience; crowdsourcing; network measurement; CloudAccess

## Contents

## Project Description

## 1 Introduction and Motivation

Amid the unprecedented COVID-19 pandemic, people have practiced social distancing to slow the spread of the virus. Telecommuting, distance learning, and virtual gathering have become a new normal in the United States. By all accounts, Internet traffic surged during the pandemic [18, 16, 10, 41, 61, 52, 26]. Video conferencing traffic exploded by 210% - 700% [16, 41, 61], which contributed to a 30+% growth in upstream traffic [16, 52]. ISPs observed a 20-30% increase in on-demand and live video streaming traffic [16, 52, 75]. The traffic surge contributed to the overflow of interconnections with content providers and degraded video quality of experience (QoE) [11, 25]. We are now seeing light at the end of the tunnel for this pandemic, but significant usage of video streaming and conferencing tools will continue as companies permit their employees to work from home for a prolonged period. Higher education also accelerated the adoption of various modes of digital delivery of classes [29], such as fully virtual and hybrid learning models. We are entering a phase where we need a more rigorous quantitative understanding of the user-perceived performance of these tools, because they will now compete more aggressively with in-person alternatives.

Researchers have leveraged public crowdsourcing platforms to recruit human subjects to perform QoE assessments [34]. However, on the Internet, many different platforms provide on-demand and live video streaming using various technologies, CDNs, and cloud infrastructures. Different implementations, including the location and bandwidth of video caches, the use of multiple CDNs, and the quality adaptation algorithm in video players, make measuring and comparing *in-situ* user QoE of multiple video streaming providers difficult.

Measuring the QoE of *video conferencing* is even more challenging because such measurement requires multiple parties to interact with each other at the same time. Existing laboratory-based studies (e.g., [66, 65, 68, 67, 72]) are small-scale and conducted in controlled settings, where performance and environment factors can differ from user-perceived experience at home.

The goal of this research is to develop novel, reliable, and scalable QoE crowdtesting methodologies to diagnose QoE degradation issues in video streaming and video conferencing applications. We formulate two tasks to accomplish our goal. Our first task is to design and implement new experiments to extend the capability of our gamified web platform – Quality of Internet Consumer Experience (QUINCE) [56] – to crowdsource network measurements and QoE assessments from users of major video streaming platforms. We will also create an interactive environment with QUINCE connecting multiple subjects to perform QoE assessments on cloud-based video conferencing applications. We will leverage our longitudinal experiment design to attract subjects and incentivize sustained participation [57].

In the second task we will analyze the correlation between network performance metrics and reported QoE. In addition to network measurements embedded in QUINCE, we will leverage external measurement platforms capable of inferring evidence of interdomain congestion and reactively triggering throughput measurements to speed test servers from the cloud. The combined data will support our analysis of the dynamics of video streaming/conferencing QoE when congestion occurs in interconnections and access networks.

This project is directly responsive to the NSF CNS core program's research goals of producing practical abstractions, techniques, tools, artifacts, or datasets that address/enhance both general and functional requirements of network systems.

## 2 Background

**Crowdsourcing-based QoE assessment.** QoE crowdtesting is becoming popular because of the growing size and diversity of the subject pool, and its cost-effectiveness. Experimenters can access more than 100K prospective human subjects [21] (also called workers) via public web-based crowdsourcing platforms [7, 4]. Subjects participate remotely, completing experiments in exchange for monetary rewards. In 2012, researchers used such techniques to conduct QoE crowdtesting on videos and picture quality [40, 39, 62]. In 2014, Kraft and Zölzer used HTML5 and JavaScript to crowdsource subjective listening tests. In 2016, researchers developed the Eyeorg platform [70] to record and replay web page loading processes to evaluate the QoE of Web browsing. In 2020, Yann *et al.* [77] re-transmitted TV signals as Internet live streaming to perform randomized experiments that evaluated the performance of video quality adaptation algorithms. However, these measurements neither asked users for their subjective scores nor collected network topology measurements for performance diagnosis. In 2018 PI Mok developed a unified platform QUINCE [56] to perform network measurement and QoE assessments, and found that user's last-mile throughput correlated with reported YouTube QoE. *We propose to use this platform as a foundation for introducing new types of measurements and analysis that transform the community's capability to scientifically study QoE on the Internet.*

Researchers have also proposed efforts to improve the reliability and robustness of subjective QoE crowdtesting results, e.g., filtering out unreliable scores [63, 30]. We will incorporate these as well as our own techniques [54] to filter out low quality measurements in our experiments.

**QoE measurement of video conferencing.** Small-scale (8-25 participants) laboratory-based studies have correlated video conferencing QoE with factors such as coding quality and communication delay [44] as well as packet buffer sizes [76]. A Dutch research group has been especially active, developing a testbed with a custom video client to support small (fewer than 30 subjects) QoE studies [66], such as investigating the effect of network delay and jitter [65], video encoding bitrates and packet loss rate [68, 67], and user age and prior experience [67] on reported QoE of multi-party video conferencing. They found that participants who led the conversation were more sensitive to network delay [65], while young or engaged participants were more forgiving to impairments in video calls [67]. Video quality degradation induced by packet loss was noticeable but acceptable to users in high video bitrate settings [68, 67]. Their customized client used a peer-to-peer approach to transport, not representative of today's cloud-based video conferencing applications running over WebRTC protocol in browsers. Two recent laboratory-based studies ($N < 30$) investigated the impact of packet loss [72] and video quality [73] on the QoE of WebRTC-based conferencing on mobile phones, and showed that slow recovery from packet loss in congestion control algorithms led to sub-optimal video encoding bitrate.

**Gamified QoE assessment.** Gamification, defined as the use of game design elements, e.g., scores/points, leaderboards, and badges, in non-game contexts [19] has demonstrated utility in crowdsourcing experiment tasks [33, 58, 59], including improving data quality [42, 31] and user activity [32, 74]. The use of monetary reward in a gamified image-tagging experiment [27] established its ability to incentivize subjects to culminate in more work done to increase efficiency. PI Mok applied four gamification elements (stories, scores, levels, and badges) in the QUINCE platform to incentivize subjects to perform longitudinal experiments that improved overall efficiency [57].

## 3 Existing infrastructure components to support our research agenda

In this research, we will enhance the measurement capability and coverage of QUINCE (§3.1) by incorporating and advancing techniques from inter-related performance measurement projects. We will leverage three network measurement platforms/tools (§3.2 - §3.4) to gather network and video streaming performance data to support our analysis.

### 3.1 QUINCE - Quality of INternet Consumer Experience

We undertook a two-year effort to build a prototype of QUINCE [56], our gamified web platform for conducting crowdsourcing-based experiments that capture real-world network performance and video streaming QoE. Figure 1 shows the overall architecture of QUINCE, including modules we propose to develop. We partially integrated CAIDA's MANIC platform with IP geolocation databases to strategically select measurement destinations for end users and to visualize observed Internet topology. We implemented four related types of measurements into a unified interface.

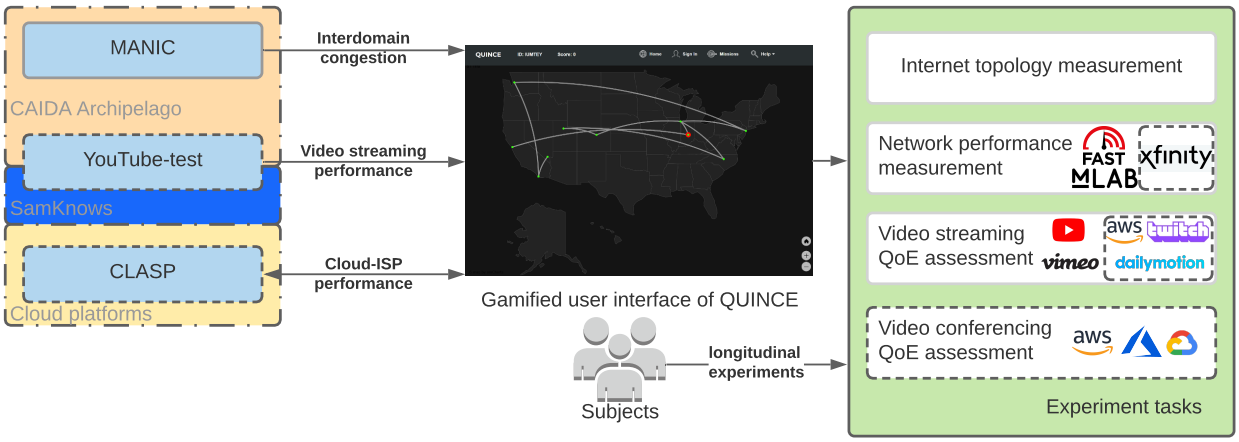



*Figure 1: Overall architecture of the QUINCE measurement platform. Boxes with solid line are existing components of QUINCE. The dot-dashed boxes show the underlying infrastructure that the platforms/tools rely on. Dotted boxes denote measurement tasks proposed for this project. The middle of the figure shows a snapshot of the main user interface of QUINCE.*

*Video streaming QoE assessment.* We streamed a short (60-90s) video clip from our own web server using HTTP Adaptive Streaming (HAS) or from large-scale video service providers (YouTube and Vimeo) with a customized JavaScript-based video player. Upon completion of the video playback, we asked subjects to rate their QoE using an Absolute Category Rating (ACR) method (1:Bad–5:Excellent) [38]. We also simulated different streaming performance conditions by inserting different impairments, such as re-buffering and switching video quality, with the video player.

*File download.* We asked subjects to download dedicated web pages, so we could extract hostnames of CDN caches from the source code of these pages for use as target destinations in subsequent traceroute measurements.

*Network performance measurement.* We used web-based speed tests to measure network throughput between a subject's computer and speed test servers across the Internet. We incorporated two tests into QUINCE: M-Lab Network Diagnostic Tool (NDT) [47] (downlink and uplink throughput), and a customized version of fast.com [24] (downlink throughput).

*Network topology measurement.* We instructed subjects to execute their system's built-in `traceroute` command to measure paths from their computer to IP destinations. Our platform determines the IP destinations based on evidence of congestion captured by CAIDA's MANIC platform [20] (§3.2) and the hostnames we extracted in the file download task.

In addition to experiment tasks, we recently designed a novel experiment framework to improve the efficiency of QoE crowdtesting [57]. QoE crowdtestings are typically much shorter than laboratory experiments (less than 30 minutes [34, 23] *vs.* more than one hour [64]), and a significant portion of experiment time is spent training subjects to operate the assessment interface. Because each assessment can take up to 3 minutes depending on the length of the test videos, subjects can only assess a few stimuli in each experiment. Thus, the efficiency of the crowdsourcing campaign can be low. The few samples collected from each subject make it hard to examine intra-rater reliability and mitigate variances introduced by different assessment environments between subjects.

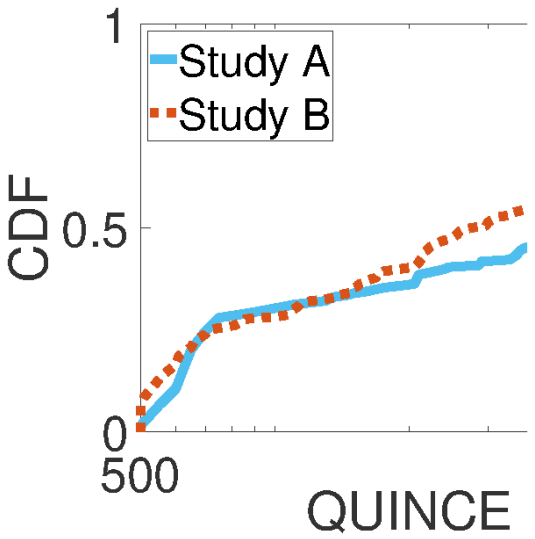



*Figure 2: CDFs of points that subjects earned in QUINCE. Subjects continued to conduct experiments after meeting the minimum requirement. Dotted grey line indicates the score that QUINCE notifies subjects about reaching the maximum reward.*

QUINCE's experiment framework differs from traditional QoE crowdtesting in that we introduce an *extended study* after the subjects complete the initial one. That is, subjects can revisit the platform to perform more tasks over a period of time. This design has three major advantages. First, the extended study increases the overall efficiency of the experiment campaign, because we do not need to repeat instructions again to returning participants. Second, by collecting more ratings from the same subjects, we reduce variances induced by environmental factors across subjects, which leads to more reliable results. Third, the experiment campaign looks more attractive to subjects, as it can lead to a larger reward than other one-off crowdsourcing tasks.

**Preliminary results.** We performed two IRB-approved studies in July 2019 (Study A) and December 2019 (Study B) on Amazon Mechanical Turk. We recruited more than 500 subjects from over 150 cities and 60 ISPs in the United States. More than 70% of subjects enrolled in the extended study. Half of them achieved at least 5.5 times more than the minimum required points (Figure 2). Task completion time for all four types of experiments dropped by 18.8%-46.1% from the subject's first attempt over the course of the experiment, reducing the cost per QoE rating up to 67%. Most important, despite lowering the monetary cost, our framework did not jeopardize the reliability and achieved similar inter-rater reliability to one-off QoE crowdtesting.

### 3.2 MANIC - Measurement and ANalysis of Internet Congestion

CAIDA's MANIC platform [20] uses bdrmap [46] to identify all interdomain links (links interconnecting two networks) visible from vantage points (VPs) in access ISPs. The VPs run Time-Series Latency Probing (TSLP) [45] to continuously measure the round-trip latency (RTT) to the near and far side of the router interfaces of the interdomain links. MANIC analyzes the patterns in RTTs to infer congestion on the interdomain links. MANIC is currently performing measurements from

51 active CAIDA's Ark VPs in 39 ISPs. We will use MANIC's data APIs [12] to obtain interdomain link information, including the IP addresses of the router interfaces and the inferred level of congestion, as one of the data sources for analyzing the QoE and network topology data that we will collect with QUINCE.

### 3.3 CLASP - CLoud-based Applications Speed Platform

Apart from measuring from the edge, PI Mok developed CLASP [55] to monitor the network performance between the cloud and access ISPs during the pandemic. The results revealed evidence of congestion both downlink/uplink directions, potentially affecting video conferencing/streaming performance in some ISPs. CLASP launches virtual machines (VMs) in different cloud regions of Amazon AWS, Google Cloud, and Microsoft Azure via the cloud platform APIs as VPs. The VMs execute speed tests to a selected set of servers in multiple speed test infrastructures (Ookla [60], Comcast [15], and M-Lab [3]) to measure download/upload throughput measurements, and run bdrmap [46] to discover interdomain links between each cloud region and ISPs. We will couple QUINCE with CLASP to deploy targeted measurements based on information from QUINCE to enable a more comprehensive and accurate measurement capabilities.

### 3.4 YouTube-test

YouTube-test [6], developed by our collaborator, Dr. Vaibhav Bajpai and his team, measures the performance of YouTube video streaming by mimicking a normal video watching session that streams a short (60-90 seconds) YouTube video. It then reports the hostname and IP address of the video cache and performance metrics including the throughput, start-up delay, number of rebuffering events, and video bitrate. The latest version of the test streams YouTube videos over QUIC and is running on over 100 SamKnows VPs.

## 4 Research agenda

This research divides into two tasks, tackling six fundamental challenges in QoE crowdtesting and data acquisition and analytics for diagnosing real-life QoE of Internet users. Our first task (§4.1) will investigate three problems with traditional QoE crowdtesting that cause unrealistic measurement parameters, constrain the type of applications, and restrict the scale of experiments. We will implement the solutions in QUINCE and evaluate them by deploying experiments on public crowdsourcing platforms.

1. *Impractical measurement parameters.* Simulated impairments cannot capture realistic dynamics of modern streaming services, including proprietary bitrate adaptation algorithms and video encoding schemes to mitigate network quality degradation in close to real-time.
2. *Unable to conduct video conferencing QoE assessments.* Existing QoE crowdtesting platforms only deliver content to subjects. Inter-subject interaction is necessary for conducting multi-party video conferencing QoE assessments.
3. *Low overall efficiency.* One-off experiment campaigns are inefficient, as subjects spend a significant portion of time on training rather than the actual assessment task [57]. It is also hard to obtain repeated measurements and longitudinal data from the same set of subjects.

The enhanced platform will enable us to conduct scalable and reliable network and QoE measurements from end users. Our second task (§4.2) is to analyze these data to diagnose QoE degradation in the wild. We will focus on how congestion events on interdomain links and the quality of last-mile links impact the QoE, because these two sections of the network paths are prone to

performance degradation. The characteristics of QUINCE data will bring us three challenges (I - III) to perform robust data analysis.

*I. Limited amount of topology measurement data.* QUINCE relies on subjects to conduct traceroutes manually, but we cannot expect them to perform the comprehensive scans of the Internet needed to identify interdomain links [46, 48].

*II. Low measurement frequency.* Revealing congestion events on interdomain links requires high frequency measurements around-the-clock to capture inflated round-trip latency or decreased throughput. However, the volume and time granularity of QUINCE measurement data largely depends on the subjects' revisit frequency (e.g., a few times per day) and availability (e.g., after work hours). Without strategic sampling, the data will be too sparse to capture trends and changes in network performance and allow us to isolate confounding factors.

*III. Variability in throughput measurement.* The deployment, location, and selection of speed test servers can impact on the accuracy of web-based throughput measurements for evaluating user's last-mile available bandwidth.

### 4.1 Task 1: Assessing the in-situ QoE from the crowd

In the first task, we will propose three transformative changes to overcome limitations in QoE crowdtesting (1 - 3) to improve efficiency, scalability, and capability.

1. We propose three new types of video streaming QoE assessments to reveal real-life video streaming QoE perceived by end-users (§4.1.1).
2. We will design and implement reliable experiment protocols and a cloud-based measurement infrastructure to crowdsource video conferencing experiments (§4.1.2).
3. We will investigate gamification techniques to improve subject engagement and incentivize participation in multi-party experiments (§4.1.3).

#### 4.1.1 Measuring in-situ video streaming QoE

**Embedding videos streamed from large-scale video content providers.** Measuring the QoE of large-scale video streaming platforms is vital to understanding the performance perceived by most Internet users. However, it is infeasible to capture fine-grain streaming performance data or monitor subjects on third-party web sites because of the same-origin policy [50]. Instead, we will use video player APIs provided by video streaming platforms to embed videos into QUINCE. The embedded player downloads videos in the same fashion as the native player on the platform's website. These APIs also provide basic information from which to infer streaming performance: video quality level, length of buffered video, and current playtime. We will select Creative Common license videos from four major video platforms (YouTube [78], Vimeo [71], Dailymotion [17], and Twitch [69]) and embed them into QUINCE. In addition to video-on-demand streaming, our player will support live streaming (in all four platforms) and 360° videos (only available in YouTube and Vimeo). Because we cannot preview the content of live streaming, we will prepare our own videos and broadcast them as live streams on these platforms.

The APIs we will use do not provide network layer information about streaming sessions, such as video cache assignment and network protocols (i.e., IPv4 *vs.* IPv6), which is essential to analyze performance problems. We will study the mechanism used to initialize streaming in each platform to identify the cache that serves the video content. It is challenging to locate the web page or RESTful API that contains the cache information, because the source code of video streaming websites is often encoded and the implementation varies. We plan to use Chrome's DevTools [14] to download the source code and capture HTTP transactions during video streaming. We will

extract the hostnames of servers that sent video data and search the source code for the same hostnames (Figure 3). We will use the *File Download* task in QUINCE to request that users download those URLs upon completion of a video QoE test. We will parse the uploaded file on-the-fly and instruct subjects to conduct traceroutes toward the caches to capture forward network paths and their IPv6 connectivity to the video caches.

```
n4v7sn76%26mm%3D31%252C26%26itag%3D22%26mt%3D1535242901%26mime%3Dvideo%252Fmp4%26e
%252Cpl%252Cratebypass%252Crequiressl%252Csource%252Cexpire%26initcwndbps%3D597506
0,itag=43\u0026url=https%3A%2F%2Fr1---sn-a5mekned.googlevideo.com%2Fvideoplayback%
rs03IUMnTq9WCJ4oSfikqDuech%26mn%3Dsn-a5mekned%252Csn-
n4v7sn76%26mm%3D31%252C26%26itag%3D43%26expire%3D1535264658%26mime%3Dvideo%252Fwe
```

*Figure 3: A code segment of a YouTube video page, which embeds the YouTube CDN cache hostname assigned to our test session. We will apply this technique to extract cache assignment information for other video streaming platforms.*

**Measuring video streaming services on CDNs and cloud platforms.** CDNs (e.g., Cloudflare and CloudFront) and cloud platforms (e.g., Amazon AWS, Microsoft Azure, and Google Cloud Platform) are popular options to deliver video content to end-users, particularly for small-scale websites. We will leverage cloud computing resources through CloudBank to host video clips on different regions of the three cloud platforms. Apart from streaming from the nearest region to the location of subjects, we will assign subjects to stream from other cloud regions. This approach will allow us to conduct measurements across various cloud providers and regions.

**Simulating realistic streaming performance impairments.** We will use the QUINCE video player to inject impairments commonly found in HTTP Adaptive streaming, including re-buffering and changing video bitrate. Our parameterization of impairment severity will leverage historical streaming performance measurement data captured by the YouTube-test [6].

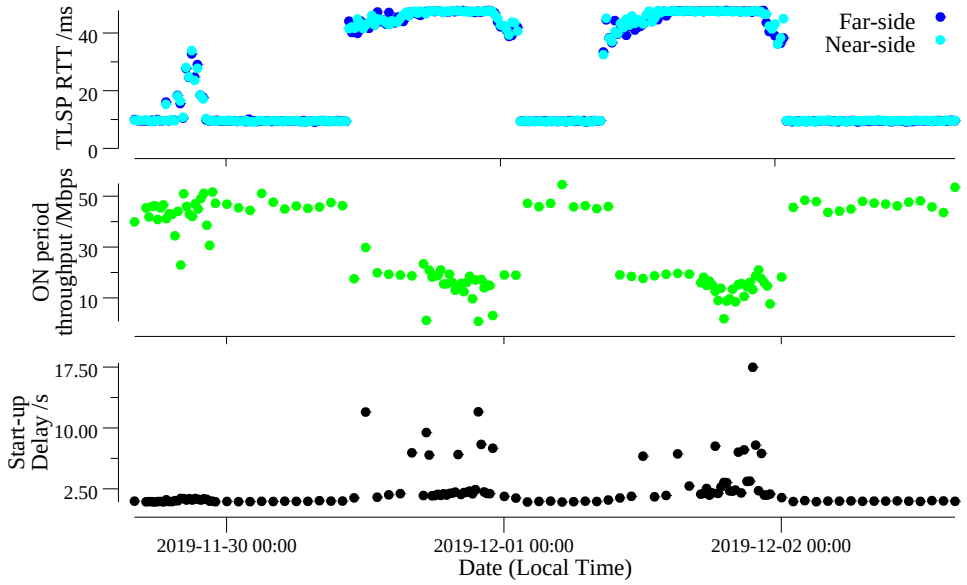



*Figure 4: TSLP and YouTube-test measurements from an Ark node in Las Vegas. Both near and far side round-trip delay were elevated, indicating the access link was probably the bottleneck. We did not observe re-buffering events during this time, but start-up delay showed significant degradation (elevated black dots).*

Figure 4 shows three days of measurement data obtained by this tool running on an Ark VP. The top figures plot the round-trip delay measured by TSLP [45] to the near and far-side of the interdomain link traversed toward the YouTube video cache. The middle and bottom figures show the average download throughput of video chunks (ON period throughput) and the start-up delay of the test video captured by the YouTube-test, respectively. In the evening of November 30 and December 1, 2019, network performance was severely degraded, as indicated by the inflated RTTs and reduced throughput. Over the same period, the start-up delay increased up to 17.5 seconds. We will use such historical measurement data to inform our instantiation of impairment parameters for reconstructing video playback in QUINCE and allow subjects to evaluate the QoE degradation caused by this simulated network congestion. We will collaborate with Dr. Vaibhav Bajpai to obtain the measurement data he collected from more than 100 SamKnows vantage points (LoC) (§3.4). We will further expand the coverage by running the YouTube-test on more Ark vantage points and measuring other popular video streaming providers.

#### 4.1.2 Measuring the QoE of video conferencing

Crowdsourcing subjective assessment of video conferencing is much more challenging than video streaming, because it requires multiple subjects from different geographic locations to interact in a meaningful context that can resemble a real-life video conferencing experience. Previous standardization efforts (e.g., ITU-T Recommendation P.920 [36] and P.1301 [37]) only applied to a controlled environment. In this task, we will adapt and gamify protocols used in controlled experiments into crowdsourcing scenarios and will integrate the assessment capability into QUINCE.

We will design new tasks for subjects to assess impairments that can affect the QoE of video conferencing. Our assessments will examine how speech delay and interactivity using screen sharing impact the QoE. We will display random numbers or names on the subjects screen after he/she has started a call. We will instruct the subject to say the displayed number or name to the other subject(s). At the same time, we will ask the other subject(s) to type the number or name as soon as they hear it. The subjects will alternate roles as the speaker or the listener. Our task will allow subjects to repeat the experiment as many times as possible within a fixed time period.

We will develop a collaboration task to evaluate interactivity by adapting the building block task described in [36], which requires subjects to rebuild an object based on instructions provided by remote participants. The original in-person design required experimenters to provide a block puzzle to one of the subjects, who then shows the rebuilding process to another subject through a webcam. However, this design is impractical in crowdsourcing scenarios due to lack of physical contact with subjects, and the use of a webcam at home leads to privacy concerns.

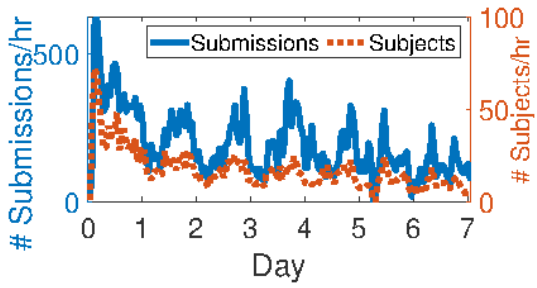
Our design will leverage the screen sharing function to facilitate information exchange between subjects, which is representative of virtual presentation or remote learning scenarios. We will display a complete picture of an object to only one of the subjects and show virtual loose blocks on others interfaces. We will ask the subjects to share the screen in the video call. The subject who has the complete picture will be responsible for guiding the others to rebuild the object by observing their screen sharing. After completing the tasks, we will instruct subjects to use the scale described in [36, 37], and a 5-point Likert scale, to rate their perceived quality of experience for video/audio quality and the interactive conversation, respectively. Subjects' reliability and performance (task completion) could be degraded due to loss of attention, instead of poor QoE. Apart from gamifying the protocol to boost the subject's attention span, we will display pop-up messages at random times, requiring subjects to click and dismiss them, as an attention check.

We will deploy a cloud-based video conferencing testbed [8] using open-source media servers, such as Jitsi [1], Kurento[2], and OpenVidu[5], to support our experiments. We will deploy at least

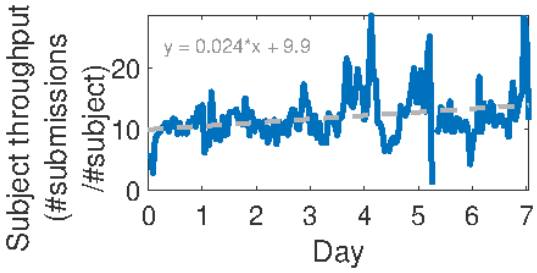
one video conferencing server in different cloud regions in the U.S. to handle video calls between subjects. Similar to commodity platforms, these media servers employ the WebRTC protocol to transfer video/audio data directly to/from browsers. We will record the experiment calls in the cloud and the performance data (e.g., video bitrate) collected from the subjects' browsers to capture a comprehensive view of application performance. The video conferencing servers will also reactively perform traceroutes toward connected clients for subsequent use in path diagnosis.

#### 4.1.3 Improving subject engagement via gamification

Our current QUINCE platform applies several gamification techniques to motivate users to contribute and improve the richness and quality of measurement data captured. QUINCE uses a map-based design (middle of Figure 1), and presents a mission to help diagnose Internet paths. QUINCE uses a scoring system to quantify work that users accomplish, and awards badges to recognize achievements.




*(a) The number of submissions and number of active subjects showed a strong diurnal pattern, correlated with Internet usage.*




*(b) Total number of submissions divided by number of unique subjects per hour. Each subject submitted on average 10 measurements per hour. The linear fit (dashed line) indicates an increasing trend in this throughput.*

*Figure 5: Hourly performance of subjects in our MTurk study.*

We studied the behavior of subjects in our preliminary study (Study B). We published the study to MTurk at 1:35am, and within an hour, subjects had performed almost 600 measurement tasks. We received over 200 submissions per hour from at least 30 subjects throughout the first day (Figure 5a. We observed a diurnal pattern, which peaked in the evening (7pm-11pm). In peak hours, more than 20 subjects submitted over 200 measurements per hour. In off-peak hours, about 10 subjects still submitted at least 50 measurements per hour.

We further analyzed the subject throughput ($= \frac{\text{\# of submission}}{\text{\# of subjects}}$) to understand their task performance throughout the campaign (Figure 5b). After the first two hours when subjects went through the tutorial, performance (subject task completion rate) increased over time.

This motivating result showed that our experiment framework can effectively support large-scale longitudinal QoE assessments. Furthermore, we collect measurements around-the-clock with more samples during peak hours, providing us opportunities to diagnose QoE degradation due to Internet congestion.

We will introduce new gamification features to further boost subject's intrinsic motivation. We will use rankings and leaderboards to inspire competition among users. Instead of using an "all-time" leaderboard that may discourage new users, we will present a "last-day" leaderboard to show scores earned the day before [35]. By introducing inter-subject competitions, we will investigate the effectiveness of new features by comparing the number of subjects participating in the extended study with our previous results.

We will study the use of game design patterns [19] to moderate the behavior of subjects. We will focus on not only on attracting subjects to revisit our platform, but also performing tasks at a requested time. This functionality is essential to video conferencing QoE assessments, which require two or more *simultaneously* available subjects. We plan to use two methods to enhance the game pattern. First, we will use HTML5's service worker API [51] to emit (optional) sound and visual notifications about new measurement tasks, as we expect some subjects may keep the QUINCE browser tab running in the background. Second, we will organize *virtual events* in QUINCE to invite subjects to perform tasks at a scheduled time. These two elements will improve participation in the collaborative video conferencing experiments. We will also initiate additional measurements during peak hours, when congestion events are most likely.

#### 4.1.4 Expected outcomes

This task will deliver reliable crowdsourcing-based methodologies for measuring the QoE of large-scale video streaming and video conferencing services. We will also investigate the effect of various gamification techniques on subject's performance and behavior.

Results of our development efforts will enable us to collect network performance/topology data, application performance data, and QoE assessments on a single unified platform. We will use the extended QUINCE platform to carry out at least three experiment campaigns to collect over 100K subjective scores from more than 1,000 subjects in the United States. The data will lead us to the second task on understanding network-induced QoE degradation.

**Limitations of approach.** QUINCE supports measurements on desktops and laptops but not mobile devices. Although the mobile user base of crowdsourcing platforms is expanding it is still a minority. Some of QUINCE's tools (e.g., the video player) can run in a mobile environment with a browser, but we believe scientific measurements of QoE on mobile devices will require implementing a native application. First, it is very complicated (sometimes impossible) for mobile users to execute traceroute measurements in smartphones without installing additional applications. Furthermore, native applications can minimize overhead in user space, improving accuracy [43]. We leave mobile QoE measurement to future work.

### 4.2 Task 2: Diagnosing QoE degradation with network measurement

Our second task will involve analyzing performance and topology data on both interdomain links between ISPs and content providers/cloud platforms and intra-ISP paths that connected the end-users. We will integrate three measurement platforms (§3) to tackle the three challenges outlined in §4. We will develop two methodologies to obtain supplementary data to map the interconnections traversed by subjects, infer performance of these interconnections, and determine whether subjects experience congestion events when they perform QUINCE experiments. We will also improve the coverage of speed test measurements in QUINCE for reliable estimates of last-mile performance.

I. We will integrate data and inference from CAIDA's MANIC platform [20] (§3.2) to identify the interconnects that QUINCE subjects and YouTube-test (§3.4) use to access the content providers and the cloud.
II. We will reactively target measurements using CLASP (§3.3) to measure throughput between cloud platforms and the subject's access ISPs throughout QUINCE experiment campaigns.
III. We will leverage multiple speed test platforms to provide comprehensive network coverage and accurate estimation of the subject's last mile capacity. Therefore, we will be able to

determine whether the last-mile link is a bottleneck to the video streaming performance.

The coverage of network paths by each platform is different, depending on the location and network of the VPs and the measurement methodologies. Figure 6 shows an overview of network paths and end-points covered by the four measurement platforms. Measurements in MANIC, CLASP, and YouTube-test may share common interdomain links with QUINCE measurements, providing opportunities for joint analysis.

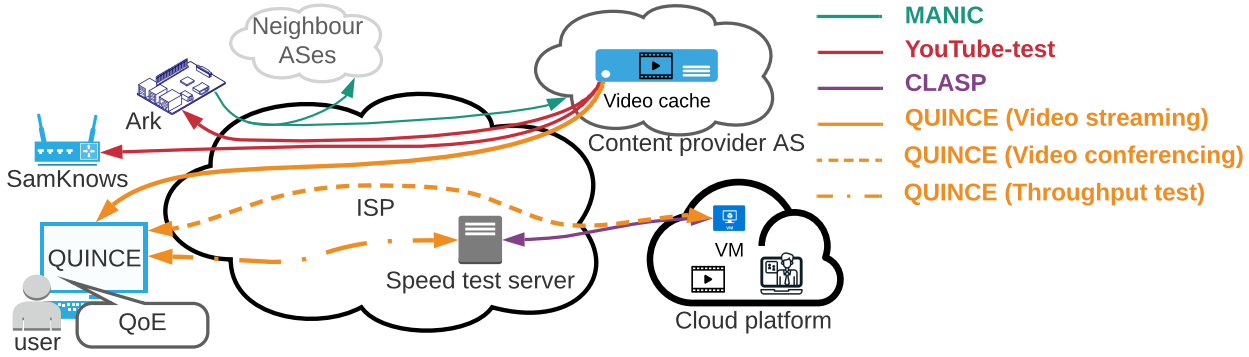



*Figure 6: Coverage of network paths by four measurement infrastructures. The YouTube-test from Ark and SamKnows VPs measures the end-to-end YouTube streaming performance (red paths). The MANIC platform monitors congestion events on interdomain links to neighboring ASes of the access ISPs hosting Ark VPs (green paths). The CLASP platform [55] measures the network throughput between the cloud regions and the core network of ISPs (purple path). The QUINCE platform initiates measurements from end-users to video streaming services for video streaming QoE and performance (solid orange path), speed test servers for network throughput (dotted dash orange path), and cloud platforms for video streaming and teleconferencing for application QoE.*

#### 4.2.1 Correlating QoE degradation with interdomain congestion events

**Annotating topology measurement data with MANIC's interdomain link inference.** Our approach to investigating the correlation between interdomain congestion events and QoE is to seek overlapping interdomain links observed by both QUINCE subjects and CAIDA's MANIC platform. To this end, we will ask subjects to perform traceroute measurements toward the IP addresses of video caches or video conferencing servers right before or after the assessments. We will use the interdomain links discovered by the MANIC platform to identify interdomain links in traceroutes submitted by QUINCE subjects. Each interdomain link is represented by a pair of IP addresses indicating the near and far side interfaces of the routers. We will compare consecutive IP hops in traceroutes with the IP pairs observed by MANIC's Ark VPs in the same ISP [53]. We will use MANIC data APIs to extract the level of congestion for identified links, and correlate it with reported QoE measurements. We will apply the same method to traceroute data collected from YouTube-test.

Figure 7 shows the reported QoE of Vimeo hosted by Akamai from QUINCE subjects in four US major ISPs (AT&T, Comcast, Spectrum, and Centurylink) that we collected in our preliminary studies. The *x*-axis shows 9 interdomain links, denoted by *ASN-Link ID*, that we identified using

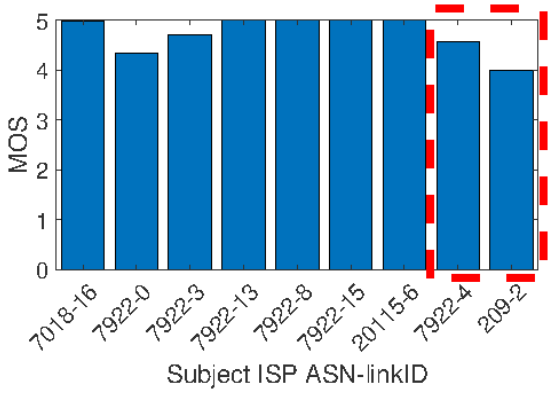



*Figure 7: Vimeo MOSes for subjects who traversed*

the traceroute data submitted by subjects. MANIC also monitored these links during the experiment period. Two links (7922-4 and 209-2), connected between Comcast (AS7922)/CenturyLink (AS209) and Akamai, showed evidence of congestion. The MOS rated by subjects traversing the congested links were sightly lower than those using uncongested links. As we increase the scale of measurement, we expect to improve the coverage of interdomain links measured by QUINCE. We will apply robust statistical analysis, such as analysis of variance (ANOVA) and student $t$-tests, and compare differences in QoE under different network scenarios. We will also build machine learning models (e.g., [9]) to help operators infer the QoE from network performance metrics.

**Triggering reactive measurements from the cloud.** To extend our visibility beyond Ark vantage points, we will use CLASP to reactively launch throughput measurements from virtual machines (VMs) in cloud platforms to speed test servers in QUINCE users' ISPs [55]. Therefore, we will be able to continuously capture network performance perceived by cloud-based video streaming/conferencing traffic to a subject's ISP.

When a subject first accesses the QUINCE platform, we will use prefix-to-AS mapping [13] and IP geolocation databases (e.g., MaxMind GeoIP2 [49], NetAcuity [22]) to resolve his/her IP address to the AS number and physical location, respectively. We will then select the nearest available server in Ookla, M-Lab NDT, or Comcast speed test platforms in the same ISPs as the subjects. We will conduct traceroute measurements from VMs to both subjects and speed test servers to examine whether the speed test servers share the same interconnections with the network paths toward the subjects. We will launch hourly throughput measurements to the selected servers from cloud regions and platforms that will serve video clips or host video calls for subjects. Longitudinal measurement data will capture trends and variations in network performance that likely impact videoconferencing performance, e.g., Figure 8. We will compare the assessments we conduct in congested/non-congested periods to understand the relationship between network congestion and video conferencing QoE. On the other hand, upload throughput (from cloud to ISPs) will be useful for us to study cloud-based video streaming QoE.

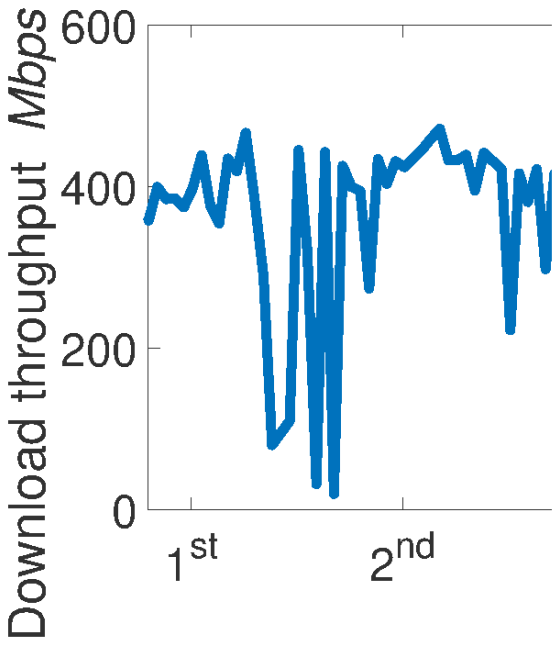



*Figure 8: Hourly download throughput from Cox (Las Vegas) speed test server to GCP West 1 region. Throughput drops reveal evidence of congestion in the direction toward the cloud, which may impair cloud-based videoconferencing QoE.*

#### 4.2.2 Analyzing the impact of the last-mile bandwidth on the QoE

Another factor influencing application performance and QoE is the quality of the last-mile networks. Increased usage of remote learning or telecommuting application during the COVID-19 pandemic increased the occurrence of last-mile congestion in the U.S. [28].

We will use web-based speed tests embedded in QUINCE to measure download and upload throughput. Although speed test servers may not be in the same network as the video caches or video conferencing servers, the tests are effective to saturate and measure the available bandwidth of the last-mile link.

Placement of speed test servers and measurement methodology can influence test results. Using servers far from users may imply a latency so high that it prevents a user from saturating the access link. We will capture the latency between subjects and test servers from the browser to ensure that they are sufficiently close. We will also embed multiple tests (M-Lab NDT, fast.com, and Comcast speed test) in QUINCE to cross-validate measurement results (§3.1).

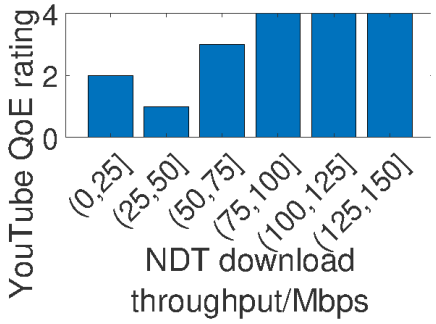



*Figure 9: The lowest YouTube QoE rating we observed in each range of NDT download throughput [56].*

We recently studied the relationship between YouTube QoE and download throughput measured by M-Lab NDT [56] (Figure 9) Even for subjects with downlink throughput as high as 50Mbps, YouTube performance could be unsatisfactory (rating $< 3$). These results illustrate the non-obvious relationship between observable QoS metrics and user-reported QoE metrics, the target of this research task.

### 4.3 Expected research outcome

We will develop and evaluate the two methodologies to overcome limitations of crowdsourcing-based measurements in identifying and inferring congestion events on interdomain links across CDNs and cloud platforms. We will also conduct robust measurements to estimate the bandwidth of the subject's last-mile link to control for this confounding factor in our congestion analysis. With subjective ratings on various network conditions, we will be able to scientifically understand how interdomain link congestion and the quality of access links impact user QoE in the wild.

## 5 Evaluation and validation plan

**System evaluation.** We will use four metrics to benchmark the performance of QUINCE.

*Response rate.* The response rate quantifies how many users correctly perform various measurement tasks. We will also consider how many video conferencing tests the subjects perform to assess the effectiveness of QUINCE in incentivizing crowdsourcing subjects to perform interactive experiments.

*Subject revisit intention and throughput.* The return of trained subjects is the key to efficiency of our experiments. We will quantify subjects' throughput using the number of measurements contributed in each visit. We will also examine the effectiveness of gamification techniques

on the subjects' revisit pattern by evaluating the participation rate and responsiveness to browser notifications.

*Dwell time.* To avoid excessive repetition of experiments, we set a cool-down time for each measurement task, during which subjects will not find any available task. The dwell time measures the time the subjects remain on our platform waiting for new tasks, an indicator of engagement.

*Cost.* The cost is the average financial cost per subject and per measurement. We expect the cost of our longitudinal experiment design will lower this cost by at least 30% from one-off experimental approaches.

**Evaluating the reliability of QoE measurements.** Due to the subjective nature of QoE measurements, *ground truth* is not so meaningful a concept as with other network measurements. We will use two approaches to assess the reliability of subjective assessments performed in QUINCE. First, we will evaluate inter-rater reliability between subjects, to quantify differences in rating between subjects under the same conditions. High inter-rater reliability provides more confidence to the results because low-quality subjects tend to provide random ratings.

Our second approach will cross-validate QoE assessments in QUINCE with traditional laboratory experiments. QUINCE collects the video playback using information from video player and records video conferencing sessions in the experiments. We will reproduce part of the playback episodes and recordings that have large disagreement between subjects, and re-evaluate them in laboratory experiments. We will recruit students on campus to perform pilot experiments under close supervision. The ratings will serve as the *gold standard data*. We will compare the statistical significance of the difference between the MOSes obtained from the lab and crowdsourcing environment. We will also use the user behavior that we captured in the pilot experiments to train machine learning models to identify low-quality crowdsourcing workers (e.g., [54]).

**Evaluating topology coverage and performance measurements.** We will analyze the diversity of ISPs and geographic coverage of QUINCE subjects. Furthermore, we will examine the number of overlapping interdomain links observed by both QUINCE and other measurement platforms. We will study the effectiveness of the two methods in task 2 for supplementing interdomain congestion data using data from MANIC and CLASP.

We will use a semi-controlled testbed, where we can emulate network performance (e.g., link capacity, packet loss rate, latency) between a test host and the Internet, to reproduce throughput tests, video streaming and video conferencing performance. We will compare measured performance from QUINCE subjects against testbed subjects. We will also cross-validate measurement data captured on the server side, including the TCP_INFO statistics recorded by M-Lab NDT [47].

## References


[1] Jitsi. https://jitsi.org.

[2] Kurento. https://www.kurento.org.

[3] Measurement lab. https://www.measurementlab.net.

[4] Microworkers. https://microworkers.com/.

[5] Openvidu. https://openvidu.io.

[6] S. Ahsan, V. Bajpai, J. Ott, and J. Schönwälder. Measuring youtube from dual-stacked hosts. In *Proc. PAM*, 2015.

[7] Amazon. Mechanical turk. https://www.mturk.com.

[8] E. André, N. L. Breton, A. Lemesle, L. Roux, and A. Gouaillard. Comparative study of WebRTC open source SFUs for video conferencing. In *Proc. IPTComm*, 2018.

[9] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of quality of experience fro Internet video. In *Proc. ACM SIGCOMM*, 2013.

[10] A. Bergman and J. Iyengar. How COVID-19 is affecting internet performance, 2020. https://www.fastly.com/blog/how-covid-19-is-affecting-internet-performance.

[11] T. Böttger, G. Ibrahim, and B. Vallis. How the Internet reacted to Covid-19 - A perspective from Facebook's edge network. In *Proc. ACM IMC*, 2020.

[12] CAIDA. MANIC (Measurement and ANalysis of Internet Congestion) API. https://api.manic.caida.org/v1/.

[13] CAIDA. Routeviews prefix to AS mappings dataset (pfx2as) for IPv4 and IPv6. https://www.caida.org/data/routing/routeviews-prefix2as.xml.

[14] Chome Developers. Chrome DevTools. https://developer.chrome.com/docs/devtools/.

[15] Comcast. Xfinity xfi speed test. https://speedtest.xfinity.com.

[16] Comcast. COVID-19 network update, 2020. https://corporate.comcast.com/covid-19/network/may-20-2020.

[17] dailymotion. Video player documentation. https://developer.dailymotion.com/player.

[18] DE-CIX. Highest jump ever: DE-CIX Frankfurt reaches 9.1 Tbps, 2020. https://www.de-cix.net/en/news-events/news/de-cix-frankfurt-reaches-9-1-tbps.

[19] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From game design elements to gamefulness: defining "gamification". In *Proc. MindTrek*, 2011.

[20] A. Dhamdhere, D. Clark, A. Gamero-Garrido, M. Luckie, R. Mok, G. Akiwate, K. Gogia, V. Bajpai, A. Snoeren, and kc claffy. Inferring persistent interdomain congestion. In *Proc. ACM SIGCOMM*, 2018.

[21] D. Difallah, E. Filatova, and P. Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proc. of ACM WSDM*, 2018.

[22] digitalelement. Netacuity. https://www.digitalelement.com/solutions/.

[23] S. Egger-Lampl, J. Redi, T. Hoßfeld, M. Hirth, S. Möller, B. Naderi, C. Keimel, , and D. Saupe. Crowdsourcing quality of experience experiments. In *Proc. Crowdsourcing and Human-Centred Experiments (Dagstuhl Seminar 15481)*, 2017.

[24] Fast.com. Internet speed test. https://fast.com.

[25] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia. Campus traffic and e-learning during COVID-19 pandemic. 176:107290, 2020.

[26] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber, J. Tapiador, N. Vallina-Rodriguez, O. Hohlfeld, and G. Smaragdakis. The lockdown effect: Implications of the COVID-19 pandemic on Internet traffic. In *Proc. ACM IMC*, 2020.

[27] O. Feyisetan, E. Simperl, and M. V. K. abd Nigel Shadbolt. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proc. WWW*, 2015.

[28] R. Fontugne, A. Shah, and K. Cho. Persistent last-mile congestion: Not so uncommon. In *Proc. ACM IMC*, 2020.

[29] S. Gallagher and J. Palmer. The pandemic pushed universities online. the change was long overdue, 2020. `https://hbr.org/2020/09/the-pandemic-pushed-universities-online-the-change-was-long-overdue`.

[30] B. Gardlo, S. Egger, M. Seufert, and R. Schatz. Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing. In *Proc. IEEE ICC*, 2014.

[31] J. Goncalves, S. Hosioa, J. Rogstadius, E. Karapanos, and V. Kostakos. Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. *Computer Networks*, 90(2015):34–48, 2015.

[32] J. Hamari. Do badges increase user activity? a field experiment on the effects of gamification. *Computers in Human Behavior*, 71:469–478, 2017.

[33] J. Hamari, J. Koivisto, and H. Sarsa. Does gamification work? - A literature review of empirical studies on gamification. In *Proc. HICSS*, 2014.

[34] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *IEEE Trans. Multimedia*, 16(2):541–558, 2014.

[35] P. G. Ipeirotis and E. Gabrilovich. Quizz: targeted crowdsourcing with a billion (potential) users. In *Proc. WWW*, 2014.

[36] ITU-T. Interactive test methods for audiovisual communications, 2000.

[37] ITU-T. Subjective quality evaluation of audio and audiovisual multiparty telemeetings, 2017.

[38] ITU-T Recommendation P.913. *Audiovisual quality in multimedia services*, Jan 2013.

[39] C. Keimel, J. Habigt, and C. Horch. Video quality evaluation in the cloud. In *Proc. IEEE PV*, 2012.

[40] C. Keimel, J. Habigt, C. Horch, and K. Diepold. QualityCrowd - A framework for crowd-based quality evaluation. In *Proc. IEEE Picture Coding Symposium*, 2012.

[41] C. Labovitz. Network traffic insights in the time of COVID-19: March 23-29 update, 2020. `https://www.nokia.com/blog/network-traffic-insights-time-covid-19-march-23-29-update/`.

[42] T. Y. Lee, C. Dugan, W. Geyer, T. Ratchford, J. Rasmussen, N. S. Shami, and S. Lupushor. Experiments on motivational feedback for crowdsourced workers. In *Proc. ICWSM*, 2013.

[43] W. Li, D. Wu, R. Chang, and R. Mok. Toward accurate network delay measurement on android phones. *IEEE Transactions on Mobile Computing*, 17(3):717–732, 2018.

[44] Y. Lu, Y. Zhao, F. Kuipers, and P. V. Mieghem. Measurement study of multi-party video conferencing. In *Proc. IFIP Networking*, 2010.

[45] M. Luckie, A. Dhamdhere, D. Clark, B. Huffaker, and kc claffy. Challenges in inferring internet interdomain congestion. In *Proc. ACM IMC*, 2014.

[46] M. Luckie, A. Dhamdhere, B. Huffaker, D. Clark, and k. claffy. bdrmap: Inference of Borders Between IP Networks. In *Proc. ACM IMC*, Nov 2016.

[47] M-Lab. NDT (Network Diagnostic Tool). `https://www.measurementlab.net/tests/ndt/`.

[48] A. Marder, M. Luckie, A. Dhamdhere, B. Huffaker, kc claffy, and J. M. Smith. Pushing the boundaries with bdrmapIT. In *Proc. ACM IMC*, 2018.

[49] MaxMind. GeoIP databases and services. https://www.maxmind.com/en/geoip2-services-and-databases.
[50] MDN Web Docs. Same-origin policy. https://developer.mozilla.org/en-US/docs/Web/Security/Same-origin_policy.
[51] MDN web docs. Serviceworkerregistration.shownotification(). https://developer.mozilla.org/en-US/docs/Web/API/ServiceWorkerRegistration/showNotification.
[52] S. Mitchko-Beale. How Charter is meeting higher demand for reliable Internet during COVID-19 crisis, 2020. https://corporate.charter.com/newsroom/chief-technology-officer-how-charter-is-meeting-higher-demand-for-reliable-internet-during-covid-19-crisis.
[53] R. Mok, V. Bajpai, A. Dhamdhere, and kc Claffy. Revealing the load balancing behavior of YouTube traffic on interdomain links. In *Proc. PAM*, 2018.
[54] R. Mok, R. Chang, and W. Li. Detecting low-quality workers in qoe crowdtesting: A worker behavior based approach. *IEEE Trans. on Multimedia*, 19(3):530–543, 2017.
[55] R. Mok and k. claffy. Measuring the impact of COVID-19 on cloud network performance. In *COVID-19 Network Impacts Workshop*, 2020.
[56] R. Mok, G. Kawaguti, and K. Claffy. QUINCE: A unified crowdsourcing-based QoE measurement platform. In *Proc. ACM SIGCOMM poster session*, 2019.
[57] R. Mok, G. Kawaguti, and J. Okamoto. Improving the efficiency of qoe crowdtesting. In *Proc. ACM QoEVMA*, 2020.
[58] B. Morschheuser, J. Hamari, and J. Koivisto. Gamification in crowdsourcing: A review. In *Proc. HICSS*, 2016.
[59] L. E. Nacke and S. Deterding. The maturing of gamification research. *Computers in Human Behavior*, 2017.
[60] Ookla. Speedtest. http://www.speedtest.net.
[61] A. Periyannan. BlueJeans Statement: How we are helping customers during the coronavirus outbreak, 2020. https://www.bluejeans.com/blog/bluejeans-statement-how-we-are-helping-customers-during-coronavirus-outbreak.
[62] B. Rainer, M. Waltl, and C. Timmerer. A Web based subjective evaluation platform. In *Proc. QoMEX*, 2013.
[63] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer. CrowdMOS: An approach for crowdsourcing mean opinion score studies. In *Proc. IEEE ICASSP*, 2011.
[64] R. Schatz, S. Egger, and K. Masuch. The impact of test duration on user fatigue and reliability of subjective quality ratings. *Journal of the AES*, 60(1/2):63–73.
[65] M. Schmitt, S. Gunkel, P. Cesar, and D. Bulterman. The influence of interactivity patterns on the quality of experience in multi-party video-mediated conversations under symmetric delay conditions. In *Proc. ACM SAM*, 2014.
[66] M. Schmitt, S. Gunkel, P. Cesar, and P. Hughes. A QoE testbed for socially-aware video-mediated group communication. In *Proc. SAM*, 2013.
[67] M. Schmitt, J. Redi, D. Bulterman, and P. S. Cesar. Towards individual QoE for multiparty videoconferencing. *IEEE Transactions on Multimedia*, 20(7):1781–1795, 2018.
[68] M. Schmitt, J. Redi, P. Cesar, and D. Bulterman. 1mbps is enough: Video quality and individual idiosyncrasies in multiparty HD video-conferencing. In *Proc. QoMEX*, 2016.
[69] Twitch developers. Embedding video and clips. https://dev.twitch.tv/docs/embed/video-and-clips.
[70] M. Varvello, J. Blackburn, D. Naylor, and K. Papagiannaki. EYEORG: A platform for crowdsourcing web quality of experience measurements. In *Proc. ACM CoNEXT*, 2016.

[71] Vimeo. Vimeo player API. https://github.com/vimeo/player.js.

[72] D. Vucic and L. Skorin-Kapov. The impact of packet loss and google congestion control on QoE for WebRTC-based mobile multiparty audiovisual telemeetings. In *Proc. MultiMedia Modeling*, 2019.

[73] D. Vučić and L. Skorin-kapov. QoE assessment of mobile multiparty audiovisual telemeetings. *IEEE Access*, 8:107669–107684, 2020.

[74] X. Wang, D. H.-L. Goh, E.-P. Lim, A. W. L. Vu, and A. Y. K. Chua. Examining the effectiveness of gamification in human computation. *International Journal of Human-Computer Interaction*, 33(10):813–821, 2017.

[75] H. Waterman and K. Schulz. Verizon delivers network reliability during COVID-19 while accelerating 5G deployments, 2020. https://www.verizon.com/about/news/how-americans-are-spending-their-time-temporary-new-normal.

[76] J. Xu and B. W. Wah. Exploiting just-noticeable difference of delays for improving quality of experience in video conferencing. In *Proc. ACM MMSys*, 2013.

[77] F. Y. Yan, H. Ayers, C. Zhu, S. Fouladi, J. Hong, K. Zhang, P. Levis, and K. Winstein. Learning in situ: a randomized experiment in video streaming. In *Proc. USENIX NSDI*, 2020.

[78] YouTube. Youtube player API reference for iframe embeds. https://developers.google.com/youtube/iframe_api_reference.