QUINCE: A unified crowdsourcing-based QoE measurement platform

Ricky K. P. Mok
CAIDA/UC San Diego
cskpmok@caida.org

Ginga Kawaguti
NTT, Japan
ginga.kawaguti.nr@hco.ntt.co.jp

kc claffy
CAIDA/UC San Diego
kc@caida.org

ABSTRACT

Assessing QoE *in situ* is a challenging task. Researchers have employed crowdsourcing-based approaches to achieve scale and diversity of subjects, but not without confounding factors. Apart from subjective bias of users, environmental factors introduce variance to QoE measurements. We propose QUINCE, a QoE measurement platform, which uses a gamified approach to enable longitudinal study with repeated and varying measurements in a single platform. We leveraged existing Internet measurement data and infrastructures to integrate three different types of network and QoE measurements to yield a more comprehensive view of subjects. Our preliminarily results show that QUINCE achieves a high level of engagement from subjects and collects data that is useful for correlating network performance and YouTube video streaming QoE.

CCS CONCEPTS

• **Networks** → **Network measurement**; • **Human-centered computing** → *Web-based interaction.*

KEYWORDS

QoE, network measurement, crowdsourcing

ACM Reference Format:
Ricky K. P. Mok, Ginga Kawaguti, and kc claffy . 2019. QUINCE: A unified crowdsourcing-based QoE measurement platform. In *SIGCOMM '19: ACM SIGCOMM 2019 Conference (SIGCOMM Posters and Demos '19), August 19–23, 2019, Beijing, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3342280.3342307

1 INTRODUCTION

Researchers use crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) [1], to acquire large and diverse pools of human subjects. Often this approach includes a web-based platform to serve experiments to remote subjects via their web browser, and collect measurement results. In addition to network measurement [5], such web-based platforms have been used to assess the quality of experience (QoE) of video streaming [4, 8] and web performance [3, 6]. But typically the design of these platforms is such that each subject can perform only one type of measurement within a short experiment session, which is usually less than 30 minutes. The overhead of this approach is high, because the subject must spend a large portion of the experiment time reading instructions and procedures. Each subject can only contribute a few useful data points on one type of experiment, whereas environmental factors, such as time-of-day, network speed, or user equipment, could increase variance of the results.

We propose a novel measurement framework – QUINCE (Quality of Internet Consumer Experience) – to unify network measurement and QoE assessments. We employ gamification techniques to enrich each subject's experience and improve their engagement level. These enhancements enable us to carry out longitudinal studies on the same set of subjects. Therefore, we can reduce variance introduced by differences in environmental factors between subjects.

By incorporating data from existing Internet measurement data and infrastructure, we have implemented and deployed three types of measurements in QUINCE: video QoE assessments; network performance measurements; and Internet topology measurements. Parameters for these measurements dynamically adapt to recent Internet congestion events. We implemented QUINCE and used it to conduct two IRB-approved 1-week studies using MTurk. We found that a significant number of (70) subjects participated in our experiment during the two weeks. We analyzed the resulting measurement data to study the correlation between user-reported QoE of YouTube and measured available bandwidth of the subject's access link.

2 QUINCE

Figure 1 shows the three main components of the QUINCE architecture.[1] The green components on the left represent the datasets we use to support the choice of measurement targets and visualization. The middle part of the figure is one screen shot of the user interface of QUINCE, which subjects interact with to execute measurement tasks, denoted as blue boxes on the right. We incorporate existing Internet measurement datasets to generate three kinds of measurement tasks:

Video streaming QoE assessment. We implemented a customized video player to simulate artifacts, such as rebuffering and video quality change, and to record video streaming performance in the background. We ask the subject to rate the QoE after the end of video playback using a 5-point Likert scale (1:Bad–5:Excellent). In addition to inserting impairments in the video player, we leveraged the YouTube player API [7] to embed YouTube video streams into QUINCE, and assess the user QoE of these streams in the wild.

Network throughput measurement. We coordinated the use of two existing speed test measurement infrastructures (M-Lab NDT, and fast.com) to conduct throughput measurements from

SIGCOMM Posters and Demos '19, August 19–23, 2019, Beijing, China

ACM ISBN 978-1-4503-6886-5/19/08…$15.00
https://doi.org/10.1145/3342280.3342307

[1] The prototype of QUINCE is available online at https://crowdtrace.caida.org.

subjects. This coordination allowed us to gain insights into correlation between available access link bandwidth and perceived performance.

Traceroute measurement. Topology data is important to diagnose network performance degradation. However, traceroute cannot be directly executed in the browser. We carefully designed the user interface and interactive tutorial to instruct subjects to execute the traceroute software module pre-installed in their operating systems and upload the output the QUINCE system. QUINCE selects traceroute destinations according to recent congestion signals provided by the MANIC project, described at ACM SIGCOMM 2018 [2].

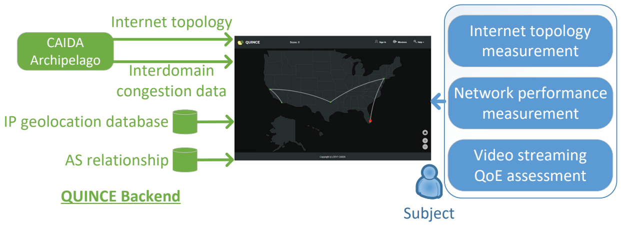
Internet topology
CAIDA Archipelago
Interdomain congestion data
IP geolocation database
AS relationship
QUINCE Backend
Internet topology measurement
Network performance measurement
Video streaming QoE assessment
Subject

Figure 1: The overall architecture of the QUINCE (Quality of INternet Consumer Experience) measurement platform. The green components are the datasets QUINCE uses. The blue portions denote the measurement tasks that subjects complete. The middle of the figure shows a snapshot of the main user interface of QUINCE, plotting the geolocated traceroute measurement.

2.1 Experiment design

Our experiment employs a two-phase approach. We first require each subject to conduct all the measurement tasks once, which is also the minimum requirement for receiving any reward. The subjects can then optionally enroll in the second phase, which is a longitudinal study that covers the rest of experiment period. They can choose to perform any task and receive extra credit according to the number of tasks submitted. We designed a simple scoring system to quantify the work a user has completed. To mitigate a few heavily contributing subjects from dominating the dataset, we assign a cool-down timer, from 2 minutes to 1 hour, for each task. The subject cannot re-perform the same task before the timer expires.

3 RESULTS

We conducted two one-week IRB-approved preliminarily studies on MTurk in October and November, 2018. After subjects completed all the measurement tasks once, they received a credit of USD $2. In these two studies, we recruited 70 distinct subjects. Figure 2 showed the cumulative distribution function (CDF) of the aggregated scores of all QUINCE subjects in the two MTurk studies. The minimum score requirement for receiving the credit is 500 points. We rewarded users USD$0.1 for every additional 100 points. Therefore, the incremental reward was only one-fourth of the initial reward for conducting the same number of measurement tasks. We found that our experiment design successfully retained a significant number of users who continued to perform measurements. Specifically, 25% of users performed at least 50 tasks and earned at least 5,000 points, 10 times the minimum requirement, while 8 subjects earned more than 20,000 points. The points earned by all subjects were more than 460,000, which cost less than USD$700 for the two studies. To acquire the same number of measurements, a one-off approach (where users cannot repeat measurements) would require more than 900 subjects and could cost more than double (USD$900 × 2 = USD$1,800).

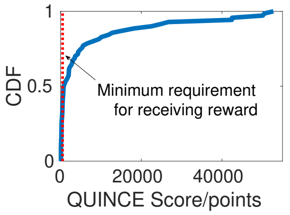

Figure 2: CDF of scores earned by subjects; 25% of subjects earned 10 times more points than the minimum requirement (500 points).

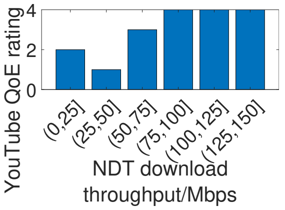

Figure 3: The lowest YouTube QoE rating we observed in each range of NDT download throughput.

We analyzed the data to study the relationship between the Internet access speed and the QoE of YouTube. We compared the lowest downlink throughput of each subject measured by the M-Lab NDT on QUINCE with their lowest YouTube QoE rating. Figure 3 shows the lowest YouTube QoE rating in each throughput category. We found that even for subjects with downlink throughput as high as 50Mbps, the YouTube performance could be unsatisfactory (rating less than 3). This could be because a small capacity access link can be easily congested by other users sharing it.

In the future, we will conduct large-scale measurement studies to measure the impact of interdomain congestion on video streaming QoE. By leveraging the longitudinal experiment approach, we will also measure the change in network throughput and video streaming QoE at different times of day from the same subject. Furthermore, we will expand the coverage of the network throughput measurement by including more speed test infrastructures, and extend QUINCE’s QoE measurement to other popular video streaming services, such as Vimeo and DailyMotion.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for valuable comments. This research is supported by NTT Communications under the project *A reactive crowdsourcing-based QoE monitoring* (20193641).

REFERENCES

[1] Amazon. 2018. Mechanical Turk. https://www.mturk.com.
[2] Amogh Dhamdhere, David Clark, Alexander Gamero-Garrido, Matthew Luckie, Ricky Mok, Gautam Akiwate, Kabir Gogia, Vaibhav Bajpai, Alex Snoeren, and kc claffy. 2018. Inferring Persistent Interdomain Congestion. In *Proc. ACM SIGCOMM.*
[3] Qingzhu Gao, Prasenjit Dey, and Parvez Ahammad. 2017. Perceived Performance of Top Retail Webpages In the Wild. *ACM SIGCOMM Computer Communication Review* 47, 5 (Oct 2017), 42–47.

[4] Bruno Gardlo, Sebastian Egger, Michael Seufert, and Raimund Schatz. 2014. Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing. In *Proc. IEEE ICC*.

[5] Gokay Huz, Steven Bauer, kc claffy, and Robert Beverly. 2015. Experience in using MTurk for Network Measurement. In *Proc. ACM C2B(1)D*.

[6] Matteo Varvello, Jeremy Blackburn, David Naylor, and Konstantina Papagiannaki. 2016. EYEORG: A Platform For Crowdsourcing Web Quality Of Experience Measurements. In *Proc. ACM CoNEXT*.

[7] YouTube. 2018. YouTube Player API Reference for iframe Embeds. https://developers.google.com/youtube/iframe_api_reference.

[8] Lingyan Zhang, Shangguang Wang, Fangchun Yang, and Rong N. Chang. 2017. QoECenter: A Visual Platform for QoE Evaluation of Streaming Video Services. In *Proc. IEEE ICWS*.