A reactive crowdsourcing-based QoE monitoring platform

Abstract

Measuring the Quality of Experience (QoE) in a real-world environment is challenging. Even though a number of platforms have been deployed to gauge network path performance from the edge of the Internet, one cannot easily infer QoE from that data because of the subjective nature of QoE. On the other hand, crowdsourcing-based QoE assessment, namely QoE crowdtesting, is increasingly popular in conducting subjective assessments for various services, including video streaming, VoIP, and IPTV. Workers on crowdsouring platforms can access and participate in assessment tasks remotely through the Internet. The experimenter can also select a pool of potential workers according to their geolocation or their historical accuracy. Existing QoE crowdtesting mainly evaluates emulated scenarios, instead of studying the impact of Internet events. This is because the launching of QoE crowdtesting is usually not based on network measurement results. Although we can measure network path quality from the workers, it will be difficult to correlate the assessment results with Internet events because of differences in the assessment time and network path being measured.

In this project, we propose a framework which integrates network measurement infrastructures and crowdsourcing platforms to measure the QoE on network paths. Our approach is to use existing network measurement infrastructures to detect network events, such as link congestion. Based on information of the events, the framework initiates QoE crowdtesting to recruit workers who are potentially affected, who then provide feedback on their perceived QoE. The main advantage of this reactive approach is to improve the effectiveness of launching QoE crowdtesting tasks. The deliverables are expected to include (1) a platform which can automatically launch QoE crowdtesting in response to network events, (2) a mechanism for creating suitable QoE crowdtesting and recruiting an appropriate set of workers from the crowd, and (3) a set of data obtained from the platform and the models derived from them.

1 Research Approach

Measuring the QoE of users is known to be a hard problem for network/service providers. Traditional active and passive network measurement can only obtain objective metrics from which it can be hard to infer the QoE for users. QoE crowdtesting is becoming popular, because it can obtain feedback from human subjects with a lower cost and within a shorter period of time than traditional laboratory-based assessments. Although the workers access the tests through the Internet and are able to conduct some simple network measurement tests, QoE crowdtesting is often evaluated on emulated scenarios (e.g., [5, 9]). This is because there is no existing method to determine *when* and *what* to measure with QoE crowdtesting. In this project, we propose a reactive framework to launch QoE crowdtesting in a timely manner to evaluate the impact of network events such as link congestion on user QoE.

The main novelty of this project is on the methodology of integrating network measurement infrastructures and public crowdsourcing platforms, so that the results collected from the QoE crowdtesting can better correlate with network measurement data. Existing research mainly focuses either on measuring network performance or measuring user QoE. Large-scale network measurement platforms seldom carry out subjective assessments of QoE, whereas QoE crowdtesting cannot continuously monitor network path performance. Some mobile projects (e.g., Mobilyzer [16]) attempt to build private crowds to continuously measure the network and obtain feedback from users about QoE. However, the measurements are executed through mobile applications, and the incentive for users to use those applications is low. The network measurements also have limited scope and frequency due to resource constraints (e.g., battery life) on the mobile device and the cost of cellular data for the end user.

The proposed project facilitates the strength of the two parties. CAIDA has already developed a large-scale measurement infrastructure, Archipelago (Ark), to measure the Internet. Using the Ark infrastrcuture, CAIDA has developed a system to continuously collect Internet topology and network quality data. The data collected from the system can be used to detect and localize different types of network events including route changes and link congestion. Currently, the congestion measurement system uses RTT as the signal used to detect the occurrence of link congestion. We envision that in addition to latency-based detection of congestion, we can launch supplementary measurements to collect more data on the network path performance, such as loss rate or available bandwidth. Unlike passive monitoring in the network operation center (NOC), the end-to-end network performance and routes are expected to be comparable to that of experienced by users. Furthermore, the framework can measure scheduled events, such as changing configurations and migrating links and quantify the impact of these incidents and events on user QoE.

In this project, the crowd is one of the key factors that affects the effectiveness of the framework. For example, we can evaluate the QoE of more links if we are able to enlist workers in a diverse set of ISPs. We can also increase the robustness of the QoE assessment by increasing the number of suitable workers participating in the assessment. We will take advantage of the mature crowdsourcing platforms in the US, such as Amazon Mechanical Turk, Microworkers, and CrowdFlower. According to a survey from CrowdFlower [6], the top two countries where the workers live are the US and India. Some previous studies also show that US workers have a better accuracy than Indian workers [17].

2 The Reactive Framework

This framework consists of three major steps as shown in Figure 1.

(1) The first step is to collect network event information from the measurement infrastructure and prepare the QoE crowdtesting. The congestion measurement infrastructure provides the ASes at the ends of an interdomain link, the relevant IPs, and possibly geolocation information. We propose to explore the web services that may possibly be hosted *behind* that link using hints in reverse DNS names. We are particularly interested in video streaming services, which account for a large fraction of peak-hour Internet consumption by the end user. For example, if the congestion measurement system finds evidence of link congestion between an an ISP and Netflix, then the platform can automatically launch a QoE crowdtesting campaign on the crowdsourcing platforms to invite workers who use that ISP, to assess

the QoE of watching a video streamed from Netflix.

(2) The crowdsourcing platforms only support coarse-grained filters for selecting workers, such as the country, the language proficiency, and the past accuracy of workers. Some workers who wish to enroll in our task will not be suitable for assessing the targeted network events. To save the cost and time, we will devise a qualification test, which is a short task that the workers will performn which will give us access to their network information, such as the IP and city-level geolocation. This step will allow us to choose only qualified workers, i.e, workers that are most likely to be in a position to evaluate the QoE effects of the network events we consider.

(3a) The qualified workers then conduct QoE crowdtesting which we prepared in Step (1), thus enabling us to evaluate the effect of the network events on QoE.

(3b) We will also log the data for unqualified workers, as they can be useful for testing future events. After collecting sufficient data, we can "pre-approve"/"blacklist" workers who are known to be suitable/unsuitable.

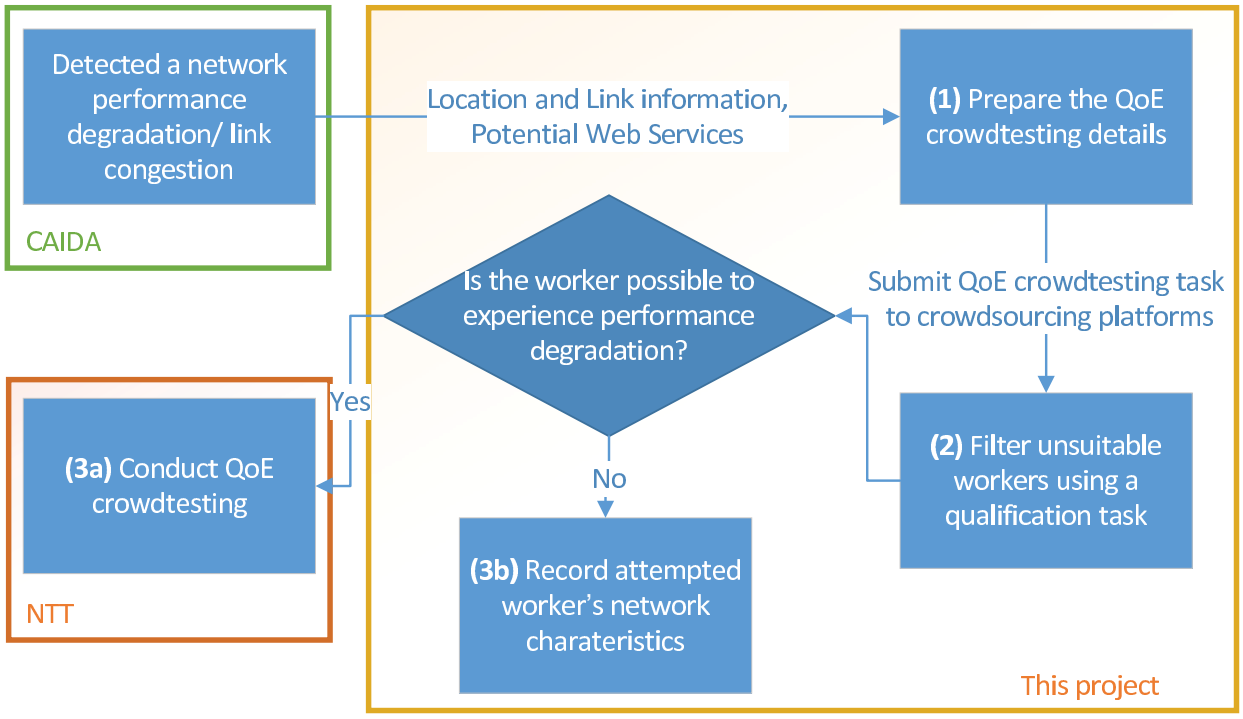

Figure 1: The overall framework.

2.1 An Example

We illustrate the framework with the following example. Figure 2 shows a time-series of the round-trip time (RTT) from a probe in Hong Kong to a local newspaper website using Akamai's CDN service [1]. From the traceroute data, we find that the RTT inflation is probably caused by a congestion event between the university and Wharf T&T. Therefore, we expect that the framework

[1]The data is extracted from the HARNET measurement platform in the Hong Kong Polytechnic University.

will launch QoE crowdtesting by enlisting workers from that university to stream video clips from Akamai in order to assess the impact on QoE.

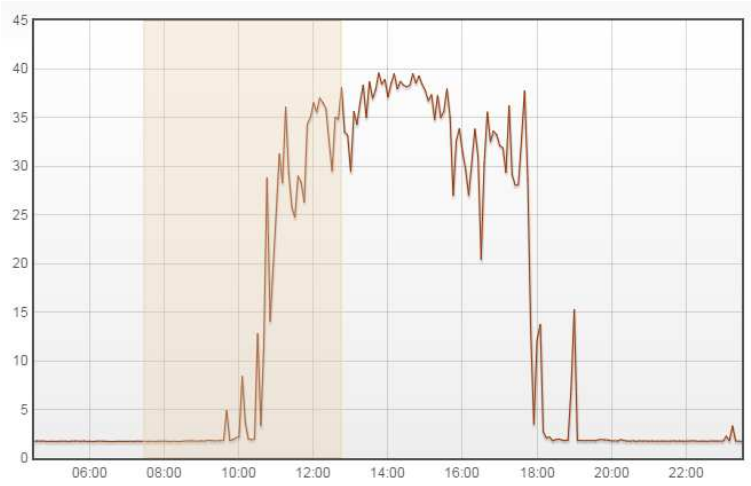

Figure 2: A time-series of RTT from a university in Hong Kong to `www.atnext.com`.

3 Research Challenges

There are a number of technical challenges that we will address in this project:

3.1 Choice of Crowd

It is important to choose the correct crowd (a pool of people) to conduct QoE crowdtesting, because the size and diversity of the crowd can affect the opportunity of measuring from workers who are influenced by a network event. One possible way is to build our own crowd and require the participants to install software (e.g., Dasu [18] and Mobilyzer [16]). However, this approach is hard to scale and the participants tend to have low diversity. Furthermore, there is no real incentive for the participants to conduct QoE assessments, which results in sparse measurements. In this project, we propose a framework which utilizes the large and diverse pool of workers available from existing crowdsourcing platforms. For example, Amazon Mechanical Turk claims to have more than half a million registered workers from 190 countries. Around 70% of MTurk workers are from the US [3]. Furthermore, a recent study [20] collects worker's information from seven MTurk accounts registered and submitted tasks at different geographical locations. They find that each MTurk account can, on average reach a population of 7,300 Amazon MTurk workers. Apart from MTurk, we can also utilize other platforms such as Microworkers [2] and CrowdFlower [1] to increase the accessible workers.

Even though the worker base in existing public crowdsourcing platforms is large, there is still a possibility that the framework cannot find sufficient workers, or cannot find workers in the "right" place to assess the observed network events. In this case, we will investigate the possibility of inviting participants from the network where the Ark probes are located. In several cases, our Ark probes are in residential locations, hosted by volunteers who may be willing to perform QoE

measurements on request. Testing from the Ark vantage points will ensure that the QoE tests will traverse the same paths as the network measurement tests.

3.2 Worker Selection

Selecting the workers who are affected by a particular network event (Step (2)) can be a challenging problem. This is because the browser has limited capability in performing network measurement. For example, `traceroute` cannot be run on the browser. Therefore, we can rely on control plane information (e.g., BGP) and data from other measurement infrastructures (e.g., iPlane and RIPE Altas) to predict the network path. Furthermore, the worker database in our framework can help speed up the worker selection process. The completion time of the QoE crowdtesting task is also critical, because the network degradation may be short-lived. To alleviate this problem, we can actively invite suitable workers by sending email or personal messages via the platform. Furthermore, we propose to dynamically adjust the wages for the crowdsourced workers, which can attract a sufficient number of workers in a timely manner, i.e., before the network degradation event ends.

3.3 Network Event Selection

A number of network events on the Internet can degrade the network performance and hence impact the QoE of end-users. The measurement infrastructure we have built can detect the presence of network congestion in a timely manner. However, the platform still has to determine which event may cause impact on the QoE. In this project, we can focus on inter-domain congestion events, because congested links can result in significant delay inflation and packet loss. Existing studies [12] showed that these two network impairments can cause QoE degradation in video streaming services.

To predict the level of QoE degradation, we propose to apply existing models (e.g., [12] and IRate [13]) to estimate the video streaming performance from network metrics. For example, using the decision tree from IRate, we can predict the best start-up video bitrate by using the median RTT and server-to-client packet loss rate. When the server to client packet loss rate and the RTT are higher than 1.75% and 101.7ms, respectively, the network cannot support a minimum video bitrate of 300Kbps, which is also the minimum bitrate for the live video encoder employed by YouTube [21]. Our subjective experiment in [13] also shows that the MOS of a smoothly played 300-Kbps video clip is around 3, which is barely acceptable. Therefore, we believe that the QoE will be unsatisfactory when the network cannot support the minimum bitrate. We can use the estimation from the model to determine which network events can cause significant QoE degradation. Furthermore, we will use the collected results to further refine the model.

Another criteria of selecting network events is the scale of potential influence to Internet users. If the event only affects a small number of users, it can be hard to find suitable workers in the QoE crowdtesting platform. To estimate the scale of the network event, we can use the AS rank [4] of two ASes on either side of the congested link. The Higher the rank of the two ASes, the larger is the potential impact in terms of the number of users possibly affected. We will also investigate using the estimated number of subscribers of access networks (using publicly available information on their webpages) to gauge the impact of network events.

In the first phase of the project, we will focus on more "predictable" congestion patterns, such as diurnal patterns that appear during peak hours every day. During the course of our measurements, we have found evidence that such a repeating pattern is commonly observed on links that are persistently congested. The predictable nature of such events will allow us to find the appropriate workers and test the framework. In the next phase of the project we will extend the system to cover network degradations due to non-periodic events such as cable cuts, misconfigurations, or popular live events such as sporting events or webcasts.

3.4 QoE Crowdtesting Task Preparation and Design

The design of the QoE crowdtesting task is important to reliably measure QoE. For example, the description of the QoE crowdtesting task displayed on the dashboard can help attract suitable workers. We will consider publicly available information, such as BGP, IP-geolocation, and ISP-AS mapping, to convert technical details into layman's terms. For example, consider an inter-domain congestion event between an ISP and a video streaming provider. The IP-geolocation mapping of the links may narrow down to a physical area, say California. Therefore, we can specify in the task that we are interested in users in California connecting with the ISP. Therefore, workers can understand whether s/he suits our task. Besides, other factors including wages and the length of the task can also affect the incentive of participating the task. The collaboration with NTT can further facilitate the design of the QoE assessment.

The task design must consider the limitations of conducting network measurements in the browser environment on the client side. In the browser environment, we can conduct simple delay measurements or speedtest-like available bandwidth measurement to the web server and the DNS server by using Adobe Flash or javascript. However, it is hard to conduct sophisticated network path measurement to obtain packet loss rate, more accurate available bandwidth measurement, and routing information due to the userspace limitation. To validate whether the worker traversed the congested path, we can instruct the workers to run traceroute and report their results as one of the screening criteria.

The QoE is application specific. In this project, we mainly focus on video streaming QoE for its popularity and high sensitivity to changes in available bandwidth. In the QoE crowdtesting task, we implement a customized video player to record the video playback information, including playback status (playing/rebuffering), video buffer status, video bitrate, and the download speed of video segments. We can use this information to derive a number of quality metrics to correlate with the QoE. An example of such metrics is the application performance metrics (APM) we proposed in [12] for capturing the smoothness of the playback. Apart from rebuffering events, a number of studies showed that the change of video bitrate/quality can significantly affect the QoE. Table 1 shows a summary of different metrics and their impact on the QoE as investigated in previous work (references are shown in the leftmost column). The second to the fifth columns from the left are metrics related to the selection or adaptation of video bitrate. The rightmost three columns are metrics related to the temporal structure of the video streaming, which are applicable to both DASH and HTTP streaming. The information we collected in the experiment can compute all the metrics listed in the table.

Another challenge is to find a suitable video streaming service to measure in the task. Employing existing video services can measure the QoE which is comparable to that experienced by real users, but dissecting the video streaming systems can be hard. We can utilize research infras-

Table 1: Summary of different quality metrics for video streaming investigated in existing literatures.

Work	Bitrate		Bitrate switching		Initial delay	Rebuffering events	Rebuffering duration
	Initial	Highest	Frequency	Amplitude			
[15]			↓				
[9]						↓	↓
[8]					↓	↓	
[11]	↑		↓			↓	
[19]			–	↓		↓	
[10]		↑	–	↓			
[7]			–	–		–	
[12]					↓	↓	↓
[14]				↓			
[13]	↑		↓				

Note: ↓, –, ↑ represents higher frequency or longer duration of that event is found to be improving, having no effect, or degrading the QoE, respectively.

tructures (e.g., Planetlab) or cloud services (e.g., Amazon EC2) to host simple video streaming services to conduct the test. We are pursuing several collaborations with video service providers which will open up the possibility of server-side network measurement [14, 13] which can reveal more network path metrics.

References

[1] Crowdflower. http://www.crowdflower.com/.

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

[3] mturk tracker. http://demographics.mturk-tracker.com.

[4] CAIDA. AS Ranking. http://as-rank.caida.org/.

[5] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. Quadrant of euphoria: a crowdsourcing platform for QoE assessment. *IEEE Netw.*, 24(2):28–35, 2010.

[6] CrowdFlower. Crowd demographics. https://success.crowdflower.com/hc/en-us/articles/202703345-Crowd-Demographics.

[7] S. Egger, B. Gardlo, M. Seufert, and R. Schatz. The impact of adaptation strategies on perceived quality of HTTP adaptive streaming. In *Proc. ACM VideoNext*, 2014.

[8] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen. Initial delay vs. interruptions: Between the devil and the deep blue sea. In *Proc. QoMEX*, 2012.

[9] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz. Quantification of YouTube QoE via crowdsourcing. In *Proc. IEEE ISM*, 2011.

[10] T. Hoßfeld, M. Seufert, C. Sieber, and T. Zinner. Assessing effect sizes of influence factors towards a QoE model for HTTP adaptive streaming. In *Proc. QoMEX*, 2014.

[11] Y. Liu, Y. Shen, Y. Mao, J. Liu, Q. Lin, and D. Yang. A study on quality of experience for adaptive streaming service. In *Proc. IEEE IIMC*, 2013.

[12] R. Mok, E. Chan, and R. Chang. Measuring the quality of experience of HTTP video streaming. In *Proc. IFIP/IEEE IM (pre-conf)*, 2011.

[13] R. Mok, W. Li, and R. Chang. IRate: Initial video bitrate selection system for HTTP streaming. *IEEE JSAC*, in press, 2016.

[14] R. Mok, X. Luo, E. Chan, and R. Chang. QDASH: A QoE-aware DASH system. In *Proc. ACM MMSys*, 2012.

[15] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen. Spatial flicker effect in video scaling. In *Proc. IEEE QoMEX*, 2011.

[16] A. Nikravesh, H. Yao, S. Xu, D. Choffnes, and Z. M. Mao. Mobilyzer: An open platform for controllable mobile network. In *Proc. ACM MobiSys*, 2015.

[17] J. Redi, T. Hoßfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel. Crowdsourcing-based multimedia subjective evaluations: A case study on image recognizability and aesthetic appeal. In *Proc. ACM CrowdMM*, 2013.

[18] M. Sánchez, J. Otto, Z. Bischof, D. Choffnes, F. Bustamante, B. Krishnamurthy, and W. Willinger. Dasu: Pushing experiments to the Internet's edge. In *Proc. USENIX NSDI*, 2013.

[19] N. Staelens, J. D. Meulenaere, M. Claeys, G. V. Wallendael, W. V. den Broeck, J. D. Cock, R. V. de Walle, P. Demeester, and F. D. Turck. Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices. *IEEE Trans. Broadcast.*, In press, 2014.

[20] N. Stewart, C. Ungemach, A. J. L. X. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, and J. Chandler. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, 10(5), 2015.

[21] YouTube. Live encoder settings, bitrates, and resolutions. https://support.google.com/youtube/answer/2853702?hl=en.