Towards Internet Scale Quality-of-Experience Measurement with Twitter

Dennis Kergl, Robert Roedler, Gabi Rodosek

Towards Internet Scale Quality of Experience Measurement with Twitter

Dennis Kergl, Robert Roedler, and Gabi Dreo Rodosek

Universität der Bundeswehr München
Department of Computer Science, 85577 Neubiberg, Germany
{dennis.kergl,robert.roedler,gabi.dreo}@unibw.de
http://www.unibw.de

Abstract. At present, Quality of Experience (QoE) measurements are accomplished by interrogating users for the perceived quality of a service they just have used. Influenced by many factors and often limited by domain or geographical region, this technique has several drawbacks when a general state of QoE for the internet as a whole is prospected. To achieve such a general metric, we leverage user complaints that we observe in real-time in social media. Such approaches have been successfully applied for the monitoring of specific and single services. We aim to extend existing methods in order to create an overall metric, define an internet wide QoE baseline, monitor changes and hence, provide a context for assessing smaller scale findings against a ground truth. The contribution of this work is to demonstrate the feasibility of using social media analysis for generating a meaningful value for quantifying the actual QoE of the internet.

1 Introduction

Management and operation of communication and networking services rely on holistic knowledge of interrelationships between technical values and perceived service quality. Especially perceived quality of internet services is fundamental for both developing web applications and planning network infrastructure [1]. The shift from technology-oriented to user-centric development, operation and measurement correlates with the trends of network architectures that evolve from host-oriented to information-centric models, and infrastructure networks from static to Software Defined Network (SDN) technologies [22].

On a technical level, there exist plenty of measured parameters that can describe characteristics of network links, protocols, connected systems and applications and form the well-defined Quality of Service (QoS) concept [16]. Although the International Telecommunication Union (ITU) definition of QoS includes the ability to satisfy stated and implied needs of the user of the service, the QoS concept does not provide insights to users' satisfaction on consumed services. To close this gap and to provide specific and measurable objectives to application and infrastructure developers, the concept of QoE was introduced.

There are many different application domains of the QoE concept resulting in slightly different understandings. Basically, existing work can be classified in either concentrating on specific network technologies, e.g., mobile networks like 3G [20], 4G [28], 5G [22], on specific media (e.g., video [32], voice [12]), on specific services like Internet Protocol television (IPTV) [31], Mobile Social Networks (MSN) [6] and YouTube [33], or on the type of service deployment like cloud services [2] or peer-to-peer networks [11]. Also combinations of the aforementioned categories are actively researched, e.g., in [5].

QoE can be thought of as QoS plus a human factor. This simple definition might be misleading, as modeling of the human factor is an unresolved issue, that includes different fields of psychology like cognitive psychology, memory psychology, and psychophysics [30]. That is why QoE measurements often include conducting real-world experiments with test persons, asking them in various ways for their opinion on used services. Using this black box test setting, the inscrutability of the human mind is bypassed and the aimed value is achieved. The disadvantage is the requirement of strictly controlled testing procedures, high personnel demand and lack of scalability.

In order to address these shortcomings, this paper is about the question of whether it is possible to turn complaints of globally distributed social media users into valuable signals for inferring perceived levels of web service quality. With a positive outcome, continuous QoE measurement at large scale would become feasible and advanced questions might be identified. With Twitter, people can publish messages (tweets) using various devices and follow other users to subscribe to their tweets. The public Twitter Application Programming Interfaces (APIs) offer programmatic access to public tweets in a well-documented JavaScript Object Notation (JSON) format. Twitter is the most widespread service of its kind and, due to its openness and popularity, current subject of research in several disciplines covering a broad range of examined topics [34]. In this work, we investigate the feasibility of using tweets as an indicator for QoE drops. Twitter users are not a representative share of all internet users. More than that, we expect the population to be biased in various ways. This is a restriction to our approach, which let us detect only complaints about problems without providing an unbiased base line. Nevertheless, we expect the outcome to be actionable in a way that subsequent work can build upon it and support the presented use cases.

The remainder of this paper is organized as follows. Section 2 gives an overview of QoE metrics for web services, how measurements are performed, and which shortcomings exist in current methods. Also, use cases are presented and requirements to the solution are derived. In Section 3 we investigate existing approaches on leveraging social media content for detecting disruptions of web services. A description of our experimental setup, message processing and signal extraction methods to receive correlations between web services and user experience are presented in Section 4. In Section 5, we give insights to the generated data and evaluate our findings with respect to the research question that is raised. The results are concluded in Section 6 and future work is outlined.

2 Problem Statement

Assessing accurate QoE metrics is a challenging problem. Schatz et. al. give comprehensive insights in [29]. They define two different kinds of testing techniques for QoE: The first kind is made of *subjective* tests, typically conducted in a controlled laboratory setting but also as field tests or using crowdsourcing methods. All of these methods aim to gather answers from humans to predefined questions and include the downside of being costly, time-consuming and require careful planning. The second kind of tests is *objective* measurement that include measuring physiological aspects of test persons or technical parameters of the utilized systems and infrastructure. These assessment methods need to be mapped to a resulting user experience score, requiring a proper model. Whilst also being affected by the drawbacks of the subjective methods, the big advantage of objective methods is the possibility of automation and therefore, some degree of scalability.

2.1 Modeling and formalization of QoE

QoE is defined as a metric for the relationship of a person that interacts as user with an application [25, 21]. While QoS focuses on the relationship between systems, the authors recognize that a change in QoS only affects QoE if a person's expectation is affected. Analogous to this approach, also the concept of Quality of Business (QoBiz) is introduced, the value of which only is affected by changes in QoE if a company's revenue is impaired. The key finding of these publications is that values of different quality aspects can be seen independent, even though they build upon each other, so that only weak coupling between these metrics can be assumed.

2.2 Web QoE metrics and assessment

In contrast to QoS that is well defined and standardized [16], and even adapted to specific technologies like mobile networks [17], QoE is much harder to quantify. A common factor of most approaches, is the assignment of an average value for perceived quality on a scale from 1 to 5 (representing *bad*, *poor*, *fair*, *good*, and *excellent*), what is known as the Mean Opinion Score (MOS) [14].

Streijl et. al. give a comprehensive summary of methods, applications, limitations and alternatives of the MOS in [30]. They describe the influence of psychological aspects, test design, testing methods and even the choice of scales to the result of MOS measurement. Stating the costly and time-consuming nature and limited scope of subjective quality tests, they also review objective models that exist in various types (e.g., arithmetic models, statistical models, parametric network planning models). While these models can be considered correct, as long as the calculated MOS lies within the confidence interval of the subjective MOS, the authors conclude that slight broadening of distortions results in higher complexity and disagreement between perceived qualities.

2.3 Challenges of objective methods for network related QoE

The ITU outlines a framework for estimating end-to-end-performance in IP networks in [15] and recommends to focus on technical metrics like bandwidth, delay and packet loss rate in order to gain insights to perceived web quality. While concluding that perceived quality can be derived with a correlation that is high enough for most use cases (> 0.9), more detailed methods for addressing factual challenges and considering a higher number of variables have been published during the last years. Most of these approaches incorporate in some way the complexity of human emotion, that are not considered in the framework of the ITU. Some of the human mind's complex relationships have been researched in context of QoE: Egger et. al. show in [8] the direct applicability of the Weber-Fechner Law (see [10] for a brief historical outline) to the relationship of waiting time and download experience. They proof this finding empirically for simple waiting tasks and furthermore, they also investigate the applicability of logarithmic relations between bandwidth and mean opinion score for more complex tasks like web browsing. Instead of a logarithmic relationship, rather an exponential relationship was discovered, as has also been shown before by Fiedler et. al. in [9]. The explanation for this outcome lies in the complex, non-linear models of network-level page load times, which were investigated in detail by Belshe [4]. Also, a memory effect has to be considered as psychological influence factor as described in [13]. With [7] Egger et. al. provide a condensed summary of many of the intertwined aspects. From these insights into technical and psychological background of perceived web quality, we can derive that a purely technical approach to measure web QoE is a hard problem.

2.4 Use Cases

To demonstrate the tangibility of the problem statement, we look at the following exemplary stakeholders that can benefit from internet wide QoE measurements.

Network Providers need to optimize investments on new infrastructure in a way that costs are minimized while turnover is maximized, aiming at ultimately maximizing profit. QoE is a valid metric for customer satisfaction, which in turn we imply is positively correlated with turnover. Due to this correlation, optimizing for QoE is more target-oriented than optimizing for technical QoS parameters. The knowledge of a base level of customer satisfaction and the ability to detect changes is key either to assess the effect of investments already carried out and to identify weak points in network infrastructure that are most in need for further investments.

Service Providers that offer their business to worldwide customers, often rely on both own and third-party infrastructure to deliver contents. Ensuring continuous availability and convenient response times, as two key service level metrics, is business-critical to them. Their challenge in monitoring customer experience is manifold: Services are frequently added and changed so that automatic or synthetic monitoring of technical key performance indicators often lags behind and covers only a small fraction of all service functions. Also a service provider would like to be aware of a shift of customers' experience during the lifetime of a service. Challenging is that the underlying infrastructure is very heterogeneous in most times, not only because of third-party services but also because of implementing novel cloud technologies as demanded by service expansion. In case of a problem, they also want to identify whether the problem affects only their own service or services of other providers as well to communicate accordingly to their customers.

Security Actors may observe disruptions of network segments or services of central importance for the reliability of internet infrastructure as a result of large-scale attacks. In such scenarios, it is crucial to gain as much information as possible as quickly as possible. This is to make up the information advantage of the attackers and become able to successfully deploy counter measures in a timely manner. To know whether, which, where and to what extend web services are affected, can support this process effectively.

2.5 Requirements

To conclude the former stated shortcomings and limitations of current approaches, we derive the following requirements on real-time QoE measurement at internet scale, matching the demands of the presented use cases.

1. Identify an overall baseline for web service QoE.
2. Recognize changes in customer experience with web services, especially drops.
3. Monitor for QoE problems independently of underlying network technology.
4. Monitor new services immediately after deployment and adapt to changes.
5. Provide continuous insights to changes and affected service.
6. Provide measurements near real-time.

3 Related Work

In order to examine to which extend the identified requirements are met by existing approaches and also to eventually identify the open points that have to be considered, we give the necessary overview of the most significant work in the relevant fields.

3.1 Measuring QoE

There are several approaches to derive MOS for specific applications from measurable network parameters and traffic monitoring, most of which include elaborate field trials interviewing test persons. In [5] Casas et. al. present YOUQMON, an approach to calculate the MOS for YouTube videos in 3G networks by passively monitoring network packets within the network core. To evaluate the model, they conducted a field trial with 16 different videos to compare the calculated MOSs with the ones perceived by test persons.

Mok et. al. investigate how network path qualities (i.e., bandwidth, round-trip time (RTT) and loss rate) affects QoE of Hypertext Transfer Protocol (HTTP) video streaming [23]. They measure the MOS in a sophisticated experimental setup under strictly controlled test conditions. Furthermore, they present first results for a correlation between video category (i.e., sports, news, comedy, music video) and the perceived quality, while keeping technical attributes like stall times and re-buffering frequency fixed. The dependency between MOS and video category is a good example of the human factor in MOS measurement and shows the non-linear connection between technical values and perceived quality. Both publications show the need for access to core network components to automatically measure a QoE score. While fulfilling some of the requirements, these approaches cannot adapt to new services and are strongly dependent on the underlying network topology.

The same authors investigate in a recent work the quality of crowd-sourced approaches to QoE measurement [24]. Though being relatively cost-effective, for long running settings, costs are still a disadvantage. Advantages over one-time experiments are, e.g., the ability of conducting an ongoing assessment of certain services, and due to using humans as sensors, adaptability to changes in the assessed services. A disadvantage is still the management effort for planning, supporting and evaluating the questionnaires. Also the quality can be an issue, as the authors investigate in the paper.

3.2 Using Twitter to detect outages

Principally, not only Twitter is suited as a data source for detecting opinions about web services. Other social media platforms offer also a wealth of user generated content. The decisive criterion for choosing Twitter is easy accessibility of data. This is meant in a technical manner, as Twitter offers a well-documented API with free and open access for many use cases. Apart from that, using Twitter is motivated in the text-focused format of the data that can be exploited with well-established techniques.

Motoyama et. al. were the first to leverage the unique characteristics of Twitter messages for detecting outages of internet services [26]. They identified terms that qualify tweets to report about service outages by investigating tweets that occurred in temporal correlation with major service outage reports in the media. To further refine their filter, they developed a heuristic that leveraged customs of Twitter users, like using hashtags that include the word *fail*. To clean up the derived signals, they made use of exponential smoothing and gave insights into their chosen parameters to achieve optimal results. Their outcome is to be able to identify outages of online services by observing between 4 and approximately 200 reports about a specific service outage. The suggested solution was later implemented by Augustine and Cushing [3]. They used the approach to monitor outages and network problems of the NETFLIX content delivery network. They were able to evaluate the accuracy of their system because a list of outages of the monitored web service was available to them and showed the practical applicability and value of leveraging tweets for their use case.

Qiu et. al. evaluated the relationship between tweets and customer care tickets that both address mobile network experience issues [27]. They found that tweets, relating to the same problem, preceded customer care tickets by approximately 10 minutes. Furthermore, tweets reported a wider range of problems while also addressing a slightly different set of problems. Qiu et. al. mapped the problems reported via Twitter to incidents they knew from the ticket system. In addition to the already known incidents, they were able to identify short-term problems that have not been reported via the ticket system. Summing up their findings, we emphasize that these correspond with our motivation to exploit tweets for measuring QoE in real-time: Timely detection of drops in experience, high sensitivity for a broad range of problems and open availability of continuous monitoring data.

3.3 Open points

We conclude the review of related work with summarizing how the formulated requirements are met in Table 1. A global baseline for an internet wide QoE score is not provided by any of the mentioned publications. Furthermore, to the best of our knowledge, there is no such approach in existing scientific literature. While the approaches that use network parameter measurements to obtain an MOS are able to map the results to a continuous scale between 1 and 5, the approaches that leverage social media messages to detect outages are only able to make a binary decision between service *available* and *not available*. Also to the best of our knowledge, there is no approach so far that would investigate other service disruptions like increased latency. Network measurement based approaches have an obvious dependency on the underlying technology. In contrast, approaches that use humans as sensor are free of this dependency. Also, human based test methods are able to adapt to new services and service changes. In the case of crowd-sourced test methods, questionnaires and manuals have to be adapted. Provided that services and technology conditions are stable, all methods can be used for continuous monitoring. Though crowd-sourcing methods have limited real-time response times, as the setup and management overhead can be significant.

Table 1. Assessment how the requirements are met by existing work.

	Casas	Mok'11	Mok'16	Motoyama	Qiu	Augustine
Requirement (see Sec. 2.5)	[5]	[23]	[24]	[26]	[27]	[3]
1: Global baseline	○	○	○	○	○	○
2: Detect score changes	●	●	●	◐	◐	◐
3: Independent of technology	○	○	●	●	●	●
4: Adapt new services	○	○	◐	●	●	●
5: Continuous monitoring	●	●	●	●	●	●
6: Results in real-time	●	●	◐	●	●	●

Requirement ○=not met, ◐=partially met, ●=fully met

4 Internet Scale QoE Measurement

We address the identified open points in the following way. In contrast to existing work, our approach aims to isolate a signal that is suitable for inferring an internet wide QoE score, rather than concentrating on a specific service. Furthermore, we add distinction between a total loss of availability and response time of a service, which can be used to derive a graduated score of disruption, rather than a binary decision. According to existing approaches that use social media content and therefore humans as sensors, we also use tweets to meet the remaining requirements.

4.1 Experimental setup

In this section, we briefly describe the source of the analyzed data, the ETL process (i.e., extract, transform, load), and used methods of feature isolation, data smoothing and signal extraction. All steps were performed using two mid-class notebooks and one office workstation, equipped with 4–12 CPU cores at 2.7 GHz–3.16 GHz and 8 GB–32 GB RAM. For loading and analyzing the data, these systems formed a small cluster running Elasticsearch on Apache Lucene as main database supported by a powerful indexing and search environment, and Kibana for gaining insights into the data. Once the data has been loaded, typical requests involving a keyword filter took approximately 60–80 seconds.

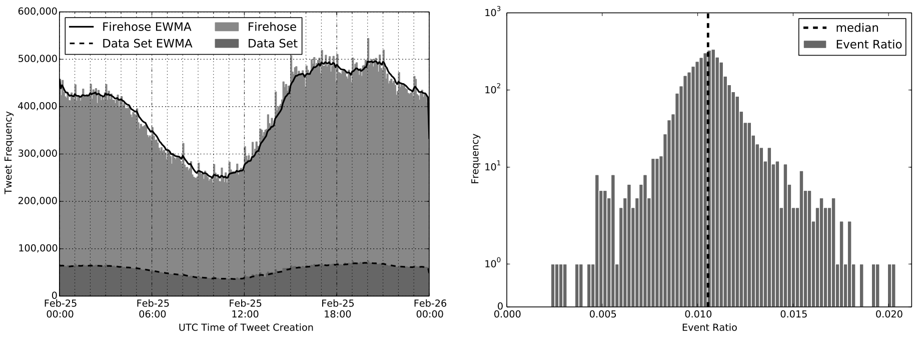
4.2 Data Source

In order to gather a reasonable data set for our analysis, we used Twitter's public API. Combining several API methods, we captured and requested tweets that have been created in the time interval from 13-Feb-2017 12:00:00.000 UTC to 02-Mar-2017 08:59:59.999 UTC. The obtained data set is not complete in the sense that a complete data set would include every single tweet that has been published during the considered period. First reason is, for being able to analyze textual content efficiently, we dropped all tweets with a language attribute that differs from *en* (i.e., English). This restriction is not as strict as it may seem. First, English is the most used language on Twitter, and second, internet related issues are often reported in English, even if it is not the native language of the reporting user. A possible reason might be that error messages are mostly in English and are simply cited by reporting users. We have observed this behavior very often in our data set. In addition, there are several more reasons for the data set not being complete: Besides public tweets there are direct tweets between users. Direct tweets are private and not accessible by anyone else but the sending and the receiving user. Another category of tweets that we missed consists of such tweets that have already been deleted at the time of our request. Ultimately our data set lacks of tweets that we simply did not cover with our query parameters.

Taking into account that a productive implementation of our findings, if appropriate, would most certainly also use the public Twitter API, working on an equivalent data set with common shortfalls appears to be justified. Beyond this, an eventual implementation should be able to work on a much smaller data set than we used to examine the feasibility of the concept.

Figure 1(a) shows a typical Twitter-Day for English content of our data set. We are able to show the amount of total tweets in the firehose (i.e., the stream of all public tweets), taking advantage of the nature of Twitter's sample stream that we described in [19] and also captured for this analysis. Also the proportion of our data set can be derived from the figure.

Furthermore, we have the requirement to our data of being as random as possible. This is due to a limitation in Twitter's Streaming API: If requested with a set of keywords (that seems appropriate for our problem at the first glance), Twitter will deliver all tweets that contain these keywords, applying a logical *OR* between the requested keywords and there is no possibility to use a *NOT* operator for this request. However, according to the documentation of the Streaming API, Twitter caps the delivery rate of this stream to 1% of the current firehose rate. Hence, if the hit rate of our filter exceeds this limit, we would not be able to derive an accurate result value due to capped measurement values. Whether this limitation is a real problem or practical systems could rely on Twitter's Streaming API, can be derived from Figure 1(b): The histogram shows the distribution of the resulting percentage of the firehose for all 5 minute intervals of our data set when we use the keywords that are presented in the next section. According to this analysis, we would hit the limit of one percent in more than a half of the time.



(a) Tweet Frequency for 5-minute-bins for a typical English subset of 24 hours.

(b) Event ratio using *OR*-linked keywords.

Fig. 1. Properties of the derived data set

Data Extraction The extraction challenge is typically twofold: First, we have to identify relevant terms, that appear in tweets of our interest. Second, tweets containing the identified terms have to be collected from the data set to form generating components for the desired signal.

Inevitably, at first we have to define what characterizes a *tweet of interest*. As Twitter is to be leveraged as a sensor for drops in web QoE, tweets that represent a suitable signal combine the following **Properties**:

1. Related to a specific web service (no need to mention which in particular).
2. Describe present reduction of availability or speed of (parts of) a web service.
3. May be formulated as a question, may use humor, sarcasm, or irony.
4. Do not notify about intentional down time.

Hereinafter the occurrence of a tweet of interest will be referenced as *event*. As a starting point for suitable terms to identify events, we analyzed and used phrases presented in [26]. The first filter approach used the following **Conditions** for tweet content (not case sensitive):

1. Must contain: *website* OR *site* OR *server*
2. Must contain: *down* OR *unreachable* OR *error*

While Condition 1 identifies tweets about web services, Condition 2 restricts the result set to terms that most likely describe the problem component of Property 2. Using only these two conditions, a one-time training set of 400 tweets showed a selectivity of 0.77.

After this manual review we optimized the filter choice for selectivity. As a result, these additional **Conditions** were added to the filter:

3. Must NOT contain: (*going to be* OR *will be* OR *was* OR *is not*) *down*
4. Must NOT contain: (*close** OR *shut** OR *take** OR *took* OR *torn*) *down*
5. Must NOT contain: (*clean* OR *count** OR *dress* OR *low* OR *right* OR *scroll* OR *settle* OR *sit* OR *top* OR *written*) *down*
6. Must not be a retweet of an original tweet.

Condition 3 was added to ensure the temporal relation of Property 2. To address Property 4, Condition 4 was introduced. The conditions were further restricted by introducing Condition 5, covering the most common semantic ambiguities of the term *down*, and by Condition 6 as retweets in the training set in most instances did not fulfill the listed properties. Finally, this set of conditions was evaluated against a test set of 840 tweets that is distinct to the training set. For the refined condition set, a selectivity of 0.88 was identified, yielding 12% false positives. We consider this rate as being sufficient for conducting a proof of concept, considering the simplicity of our approach.

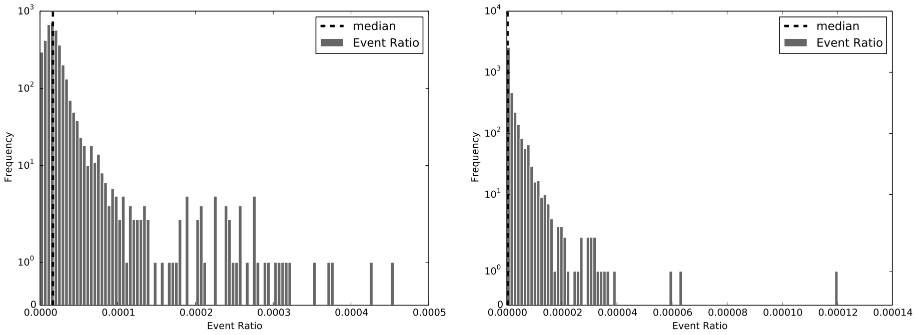
Data Smoothing The time resolution we applied for this analysis is 5 minutes. We could have chosen smaller intervals, but then occasional random disconnections of the Twitter stream, network anomalies or other random errors, would have a more significant effect on the results. Hence, we have chosen this general smoothing. Furthermore, we can observe increased activity at every full hour and also at every half hour. The latter does not weight as much as the first. To address the artificially generated bursts in the data set, we applied an exponential weighted moving average (EWMA) with a half-life period of 15 minutes ($\alpha \approx 0.206$) and used this value for further calculations. The frequency of the analyzed tweets also shows a typical variation during a day.

5 Evaluation

For determining a baseline of reports about service outages or service restrictions like increased response time, we need to apply an appropriate metric. Due to the nature of tweet distribution that is not unique across a day, we cannot simply count the number of event occurrences and use this value as a baseline. Since reports of events underlie the same daily rhythm as the firehose and are correlated with the biological rhythm of the sensors, we have to normalize the event count to the current activity. This is achieved by using the proportion between all events and all non-event tweets as metric. Hence, the ratio of event occurrences in a specific time interval in comparison to the total number of tweets in the same time interval qualifies as the desired metric. The simple formula for the wanted score in time interval n is

$$\text{event-ratio}_n = \frac{|\text{events}|_n}{|\text{EWMA}(\text{tweets})|_n - |\text{events}|_n}. \tag{1}$$

To address temporal spikes in the total tweet number, we applied the smoothing described in the preceding section to the total number of tweets and Figure 2(a) show the distribution of the event ratio in our data set for the outage event defined above. Figure 2(b) shows the event ratio distribution for *slow* events, for that we changed Condition 2 to the term *slow* only.



(a) Ratio distribution for event *down*   (b) Ratio distribution for event *slow*

Fig. 2. Event ratios of the analyzed data set. The right most spikes, that constitute a multiple of the average ratio, are related to Amazon's S3 outage during the capture time of the data set.

The mean value for the *down report* ratio can be identified as 2.25×10^{-5}, while the median is 1.66×10^{-5}. Using the median as a baseline seems appropriate, as there have been major outages taken place during the capturing of the data set. The mean value for the *slow report* ratio is 1.58×10^{-6}, while the median is 2.56×10^{-7}. The significant difference between median and mean is a clear indicator for outliers, that can be confirmed by the event ratio histogram that shows occurrences of event ratios that are multiples of the median. Table 2 lists an excerpt from the top 50 event ratios and informs about the causing event, that we identified by textually analyzing the specific period.

Table 2. Excerpt from top 50 highest event ratio intervals. Events manually evaluated.

#	Time	Event Ratio	Median Multiplier	Causing Event
1	28-Feb-17 18:30	0.000456	27.5	Amazon S3 Outage
2	28-Feb-17 18:10	0.000424	25.6	Amazon S3 Outage
3	28-Feb-17 18:15	0.000374	22.5	Amazon S3 Outage
4	28-Feb-17 18:00	0.000371	22.4	Amazon S3 Outage
5	28-Feb-17 18:25	0.000352	21.2	Amazon S3 Outage
16	27-Feb-17 11:30	0.000275	16.6	Error message on hilton.com
24	02-Mar-17 10:50	0.000246	14.8	Vainglory game server maintenance
37	02-Mar-17 12:25	0.000204	12.3	Booking problem on qatarairways.com
42	27-Feb-17 12:10	0.000188	11.3	Booking problem on klm.com
50	02-Mar-17 12:35	0.000141	8.5	Amazon S3 Outage

6 Conclusion and Future Work

We have been able to define a global baseline for *down report* and *slow report* frequencies. Therefore, there are two main contributions of this work to mention: 1. A practical system for monitoring the overall internet web QoE is feasible and can be implemented using Twitter analysis. This fulfills Requirement 1 that most likely has not been addressed by any existing work. 2. Not only outages of web services, but also degradation of web service quality can be detected. This fulfills Requirement 2 that has not been completely covered by existing publications. The remaining requirements have been matched by using humans as sensors.

The presented primary findings about the feasibility of using social media posts for gaining internet wide insights to QoE aspects in real-time denote an important step towards more detailed analysis of affected networks, domains and technologies, constituting a necessary requirement for novel approaches to improve overall network and internet security, e.g., as suggested in [18]. As follow up research questions, we are already investigating whether root causes of drops in QoE can be identified by using additional information contained in tweets, for instance, analyzing geographical origin of the complaints might lead to insights about regional problems and using the contained information about which client software was used to create the tweet might give further hints on whether mobile or fixed networks or both are affected by drops in perceived web service quality. Furthermore, mapping complaints to specific web services in an automated fashion seems to become feasible, while still being a complex problem. This would allow to drill down the QoE measurements to individual domains and accordingly to underlying networks and technologies.

References

1. Ahmad, A., Floris, A., Atzori, L.: QoE-aware service delivery: A joint-venture approach for content and network providers. In: Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on. pp. 1–6. IEEE (2016)
2. Al-Shammari, S., Al-Yasiri, A.: Defining a Metric for Measuring QoE of SaaS Cloud Computing. Proceedings of PGNET pp. 251–256 (2014)
3. Augustine, E., Cushing, C., Dekhtyar, A., McEntee, K., Paterson, K., Tognetti, M.: Outage detection via real-time social stream analysis: Leveraging the power of online complaints. In: Proceedings of the 21st international conference companion on World Wide Web. pp. 13–22. ACM (2012)
4. Belshe, M.: More Bandwidth Doesn't Matter (much) (2010)
5. Casas, P., Seufert, M., Schatz, R.: YOUQMON: A System for On-line Monitoring of YouTube QoE in Operational 3G Networks. ACM SIGMETRICS Performance Evaluation Review 41(2), 44–46 (2013)
6. Dong, M., Kimata, T., Sugiura, K., Zettsu, K.: Quality-of-Experience (QoE) in emerging mobile social networks. IEICE Transactions on Information and Systems E97D(10), 2606–2612 (2014)
7. Egger, S., Hossfeld, T., Schatz, R., Fiedler, M.: Waiting times in quality of experience for web based services. In: 2012 Fourth International Workshop on Quality of Multimedia Experience. pp. 86–96 (2012)
8. Egger, S., Reichl, P., Hossfeld, T., Schatz, R.: 'Time is bandwidth'? Narrowing the gap between subjective time perception and Quality of Experience. IEEE International Conference on Communications pp. 1325–1330 (2012)
9. Fiedler, M., Hossfeld, T., Tran-Gia, P.: A generic quantitative relationship between quality of experience and quality of service. IEEE Network 24(2), 36–41 (2010)
10. Hecht, S.: The visual discrimination of intensity and the Weber-Fechner law. The Journal of general physiology 7(2), 235–267 (1924)
11. Hei, X.H.X., Liu, Y.L.Y., Ross, K.: IPTV over P2P streaming networks: the mesh-pull approach. IEEE Communications Magazine 46(February), 86–92 (2008)
12. Ho, T., Hock, D., Tran-gia, P., Tutschku, K., Fiedler, M.: Testing the IQX Hypothesis for Exponential Interdependency between QoS and QoE of Voice Codecs iLBC and G.711. In: Proceedings of the 18th ITC Specialist Seminar on Quality of Experience. pp. 105–114 (2008)
13. Hossfeld, T., Biedermann, S., Schatz, R., Platzer, A., Egger, S., Fiedler, M.: The memory effect and its implications on Web QoE modeling. 2011 23rd International Teletraffic Congress (ITC) pp. 103–110 (2011)
14. International Telecommunication Union: P.800: Methods for subjective determination of transmission quality. Tech. rep., International Telecommunication Union (1996)
15. International Telecommunication Union: G.1030: Estimating end-to-end performance in IP networks for data applications. Tech. rep., International Telecommunication Union (2005)
16. International Telecommunication Union: E.800: Definitions of terms related to quality of service. Tech. rep., International Telecommunication Union (2008)
17. International Telecommunication Union: E.804: QoS aspects for popular services in mobile networks. Tech. rep., International Telecommunication Union (2014)
18. Kergl, D., Roedler, R., Dreo Rodosek, G.: Detection of Zero Day Exploits Using Real-Time Social Media Streams. In: Pillay, N., Engelbrecht, A.P., Abraham, A., Du Plessis, M.C., Snášel, V., Muda, A.K. (eds.) Advances in Intelligent Systems and Computing. vol. 7, pp. 405–416. Springer International Publishing (2016)

19. Kergl, D., Roedler, R., Seeber, S.: On the endogenesis of Twitter's Spritzer and Gardenhose sample streams. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). pp. 357–364. IEEE (2014)
20. Ketykó, I., De Moor, K., Joseph, W., Martens, L., De Marez, L.: Performing QoE-measurements in an actual 3G network. In: Broadband Multimedia Systems and Broadcasting (BMSB), 2010 IEEE International Symposium on. pp. 1–6. IEEE (2010)
21. Kilkki, K.: Quality of Experience in Communications Ecosystem. Journal Of Universal Computer Science 14(5), 615–624 (2008)
22. Liotou, E., Elshaer, H., Schatz, R., Irmer, R., Dohler, M., Passas, N., Merakos, L.: Shaping QoE in the 5G ecosystem. Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on pp. 1–6 (2015)
23. Mok, R.K.P., Chan, E.W.W., Chang, R.K.C.: Measuring the quality of experience of HTTP video streaming. In: Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on. pp. 485–492 (2011)
24. Mok, R.K., Chang, R.K., Li, W.: Detecting low-quality workers in QoE crowdtesting: A worker behavior based approach. IEEE Transactions on Multimedia XX(X), 1–1 (2016)
25. Moorsel, A.V.: Metrics for the Internet Age : Quality of Experience and Quality of Business Metrics for the Internet Age : Quality of Experience and Quality of Business 1 . Quantitative Evaluation in the Internet Age : What is Different ? Perspective 34, 26–31 (2001)
26. Motoyama, M., Meeder, B., Levchenko, K., Voelker, G.M., Savage, S.: Measuring online service availability using twitter. In Proceedings of the 3rd conference on Online social networks (WOSN'10) (2010)
27. Qiu, T., Feng, J., Ge, Z., Wang, J., Xu, J., Yates, J.: Listen to Me if You Can: Tracking User Experience of Mobile Network on Social Media. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. pp. 288–293. ACM, Melbourne, Australia (2010)
28. Rengaraju, P., Lung, C.H., Yu, F.R., Srinivasan, A.: On QoE monitoring and E2E service assurance in 4G wireless networks. IEEE Wireless Communications 19(4), 89–96 (2012)
29. Schatz, R., Hossfeld, T., Janowski, L., Egger, S.: From packets to people: Quality of experience as a new measurement challenge. In: Biersack, E., Callegari, C., Matijasevic, M. (eds.) Data Traffic Monitoring and Analysis, pp. 219–263. Springer (2013)
30. Streijl, R.C., Winkler, S., Hands, D.S.: Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Systems 22(2), 213–227 (2016)
31. Takahashi, a., Hands, D., Barriac, V.: Standardization activities in the ITU for a QoE assessment of IPTV. IEEE Communications Magazine 46(2), 78–84 (2008)
32. Venkataraman, M., Chatterjee, M.: Inferring video QoE in real time. IEEE Network 25(1), 4–13 (2011)
33. Wamser, F., Casas, P., Seufert, M., Moldovan, C., Tran-Gia, P., Hossfeld, T.: Modeling the YouTube stack: From packets to quality of experience. Computer Networks 109, 211–224 (2016)
34. Zimmer, M., Proferes, N.J.: A topology of Twitter research: disciplines, methods, and ethics. Aslib Journal of Information Management 66(3), 250–261 (2014)