Implications of the Multi-Modality of User Perceived Page Load Time

Flavia Salutari
Telecom Paris
flavia.salutari@telecom-paris.fr

Diego da Hora
*Telecom Paris**
dahora@google.com

Matteo Varvello
Brave Software
varvello@brave.com

Renata Teixeira
INRIA
renata.teixeira@inria.fr

Vassilis Christophides
INRIA
vassilis.christophides@inria.fr

Dario Rossi
Huawei Technologies, Co. Ltd
dario.rossi@huawei.com

***Abstract*—Web browsing is one of the most popular applications for both desktop and mobile users. A lot of effort has been devoted to speedup the Web, as well as in designing metrics that can accurately tell whether a webpage loaded fast or not. An often implicit assumption made by industrial and academic research communities is that a *single* metric is sufficient to assess whether a webpage loaded fast. In this paper we collect and make publicly available a unique dataset which contains webpage features (*e.g.,* number and type of embedded objects) along with both *objective* and *subjective* Web quality metrics. This dataset was collected by crawling over 100 websites—representative of the top 1 M websites in the Web—while crowdsourcing 6,000 user opinions on *user perceived page load time* (uPLT). We show that the uPLT distribution is often multi-modal and that, in practice, no more than three modes are present. The main conclusion drawn from our analysis is that, for complex webpages, each of the different objective QoE metrics proposed in the literature (such as AFT, TTI, PLT, etc.) is suited to approximate one of the different uPLT modes.**

***Index Terms*—Web Performance, Quality Of Experience, Measurements.**

I. Introduction

A good Quality of Experience (QoE) on the Web is essential for both content providers and consumers. QoE directly affects end-users' willingness to visit a webpage [1] as well as content providers' business revenues [2]. Both industry (*e.g.,* QUIC, SPDY, and HTTP/2) and academia [3], [4], [5], [1] have made significant effort to design tools and novel protocols to reduce page load times as the main factor that determines Web QoE is how fast a page loads [6].

Originally, quality of user experience on the Web was approximated using simple performance metrics like *time-to-first-byte (TTFB)* and the browser `onLoad` event. As modern webpages are composed of hundreds of different objects, these metrics can typically capture only the lower and upper bounds of the user perception on page load time. This limitation has motivated the introduction of a number of recent metrics to better capture user experience on the Web, such as *Above-the-Fold (ATF)* [7] and *SpeedIndex* [8].

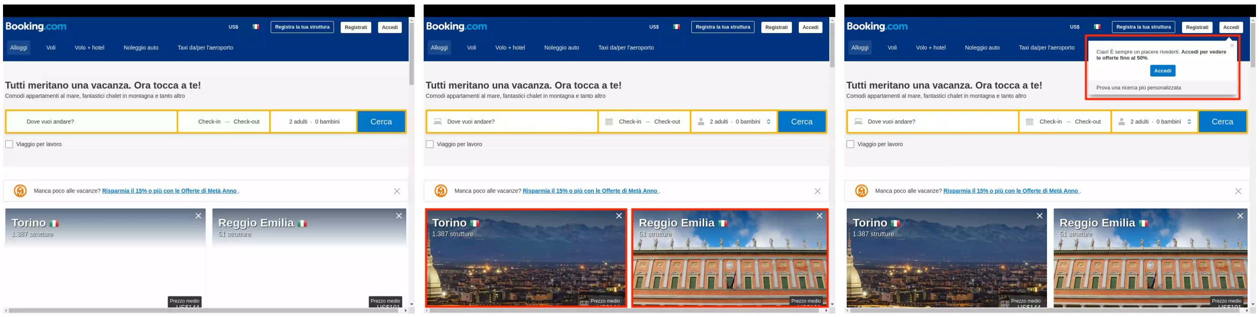
Despite all these efforts, the question regarding how well existing single-valued metrics capture the user perception of page load time remains open. We define *user perceived Page Load Time (uPLT)* as the time when a user considers the webpage to be loaded and ready to browse. With few exceptions, almost the entire previous industrial and academic efforts make the implicit assumption that a *single*-valued metric can capture the uPLT across users – or, equivalently, that the distribution of uPLT of a given page across users is uni-modal. Recent studies have challenged this assumption, showing that users rarely agree on a *single* uPLT [9], [10]. However, the multi-modality of uPLT was not the main focus of these studies, and as such it was not studied in depth.

UPLT multi-modality is rooted in many factors, such as personal preferences with respect to what is considered important on a webpage, *e.g.,* text rather than images, carousels of elements, popups or ads. Fig. 2 illustrates this issue when asking for feedback from 54 recruited participants regarding the uPLT of *www.booking.com.* About 40% of users believe uPLT to be around 2 seconds, another 40% indicates uPLT≈ 3.7 seconds and nearly 20% report a uPLT≈ 9.1 seconds. These uPLT values appear in conjunction with distinct webpage loading events; we report the snapshot of these events in Fig. 1. This example illustrates the challenges of measuring uPLT, and raises questions about which among the numerous objective Web QoE metrics (*e.g.,* PLT, TTFP, ATF [7]) is more suitable as a proxy for these remarkably different user opinions.

To address these questions, we collect a comprehensive data set of webpage features (*e.g.,* number and type of embedded objects) along with both *objective* and *subjective* Web quality metrics. We find that around 50% of the webpages in our study present a multi-modal uPLT distribution and that, in practice, three modes are sufficient to accurately describe uPLT distribution. Moreover, we show that the *number of images* and the *number of objects* in a webpage can help in predicting uPLT modality. To promote cross comparison and enable further studies, we make this dataset publicly available [11].

This paper is organized as follows. After overviewing the related work (Sec. II), we describe the methodology

[*] Current affiliation: Google Inc.



Tutti meritano una vacanza. Ora tocca a te!

Comodi appartamenti al mare, fantastici chalet in montagna e tanto altro

Manca poco alle vacanze? Risparmia il 15% o più con le Offerte di Metà Anno .

Ciao! È sempre un piacere rivederti. **Accedi per vedere le offerte fino al 50%**.

Accedi

Prova una ricerca più personalizzata

(a) (b) (c)

Fig. 1: Relevant snapshots of the *www.booking.com* rendering process corresponding to the different modes that are visible in the distribution reported in Fig.2. Notice that the “above the fold” content is almost all rendered in (a) and fully rendered in (b). At time (c) a popup arise, inviting users to login in the website.

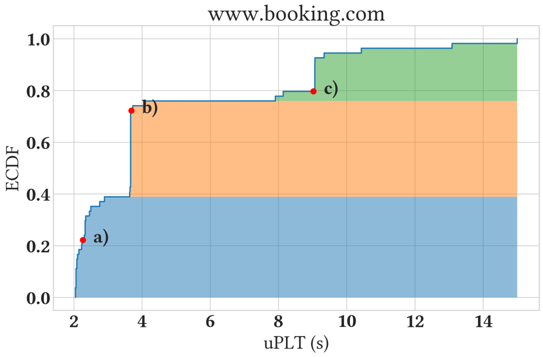


Fig. 2: uPLT distribution for *www.booking.com*, highlighting the issue that users do not agree on a single time instant to identify completion of webpage rendering.

used to produce the representative set of webpages for our analysis and how we employed the Eyeorg platform [10] to crowdsource uPLT on these pages (Sec. III). Next, we thoroughly characterize the collected user feedback (Sec. IV), rigorously quantifying violations of the hypotesis that uPLT is uni-modal and finally contrast uPLT modes with objective QoE metrics. Finally, we discuss our findings (Sec. V) and put results of this work in perspective with recent related work.

II. Related Work

Web QoE metrics fall in two main categories: *objective* and *subjective*. Our work span both categories since we crowdsource user feedback (*subjective*), which we use to benchmark *objective* Web quality metrics. We briefly survey previous work related to each category.

Objective – These metrics rely on measurable data (*e.g.,* network, browser events) capturing Web quality [12], [13], [14], [15], [8]. objective metrics of interest for Web user QoE can be further categorized in two classes: metrics that (i) *track specific events* of the W3C Navigation timing [16] or (ii) *integrate a residual completion function* over the full set of events.

On the one hand, notable examples of *tracking* metrics are the Time to The First Byte (TTFB), the Time to The First Paint (TTFP), the time at which parsing of the Document Object Model (DOM) is completed, the Time To Interaction (TTI), the time at which the content Above-the-Fold (ATF) is rendered [17], and the time at which the page was fully loaded (PLT).[1] Most of these metrics are directly available from the browser Navigation Timing [16] or can be inferred from packet/flow-level traffic [18], [19] as proxy of user experience.

On the other hand, *integration* metrics are instead founded on the idea that one page can render faster than another despite finishing loading at the same “time” (*e.g.,* in terms of PLT). These metrics introduce a function that *integrate all events of the waterfall* representing the visual progress of the page, and approximate the rendering process by using visual or byte-level completion ratios over time. SpeedIndex [8] and its variants [20], [13], [15] fall in this class of metrics.

Subjective – Subjective metrics rely on directly collecting responses from users regarding different questions related to Web QoE.

Different approaches have been proposed in the literature. Bocchi et al. [21] asked volunteers to rate page load performance on a testbed using the 5-grade Absolute Category Ranking (ACR) score. Salutari et al. [22] collected “user acceptance” [23] by asking Wikipedia users about their experience, using a binary satisfied/unsatisfied rating. These studies rate the user experience for each individual website access, but prove challenging when aggregating opinions across users, specially when measurements are conducted in the wild.

In contrast, SpeedPerception [15], Eyeorg [10], and WebGaze [9] ask users to comment on a video of the website rendering process. SpeedPerception proposes to crowdsource the user validation with A/B testing, asking users about which page loaded faster, while Eyeorg and WebGaze ask users to directly report the point in time

[1] PLT corresponds to a browser’s `onload` event, which indicates that all of the objects in the document are in the DOM, and all the images, scripts, links and sub-frames have finished loading.

when they consider that the page finished loading. This approach provides a consistent experience to all participants, regardless of their network connectivity and device configurations, making it easier to aggregate and interpret results. In this paper, we leverage Eyeorg [10] for the collection of user opinions.

Finally, ties between objective and subjective Web QoE metrics have been established in the literature [21], [14], [15], [24], [25], [26]. These studies attempt to capture the "wisdom of the crowd" by aggregating the subjective feedback using the mean or the median, while implicitly assuming uni-modality of the underlying user opinion distribution. In this work, we challenge this common assumption and instead of attempting to define a *single best* metric, we evaluate the value and complementarity of these metrics in capturing the perception of different user classes.

III. Data Collection

To explore the relationship between uPLT and objective Web QoE metrics we need to (i) collect a comprehensive dataset comprising "representative" webpages, and (ii) crowdsource feedback from real users on uPLT. We first devise a novel methodology to identify a limited number (*e.g.,* 100) of webpages to test from the Cisco's Umbrella top-1M list [27] (Sec. III-A). Second, we automate the collection of webpage characteristics and objective Web QoE metrics from Chrome-based browsers (Sec. III-B). Finally, we conduct an Eyeorg [10] crowdsourced campaign to ask users *when* each webpage finished loading (Sec. III-C). We make the entire dataset collected publicly available [11].

A. Representative Webpage Selection

A recurring concern in Web performance research is *how* to select a meaningful set of webpages to study. Due to the sheer size of the Web, some sort of sampling needs to be introduced. To study the Web, researchers often resort to the most popular webpages from Alexa or Cisco, or a combination of popular and unpopular webpages. While it is important to sample popular webpages, since they attract the majority of the traffic, unpopular webpages might have a completely different set of characteristics yielding to different results. In this paper we argue that popularity should not prime over diversity of webpages, as otherwise the results may lose generality. We therefore opt for a stratified selection of both *popular* and *diverse* pages by clustering them according to the complexity of their HTML content.

Initial hitlist: We crawl all URLs from Cisco's Umbrella top 1-million list, which became popular in the research community after Alexa became paywalled, on August 2018. The Umbrella list is generated by tracking the total number of worldwide DNS requests. The main advantage of this approach is that this gives us insights not only on popular top level domains (e.g. `wikipedia.org`), but also on popular actual pages with content (e.g. `https://en.wikipedia.org/wiki/Main_Page`). However, the Umbrella list also contains URLs that are the target of automated DNS requests (*i.e.,* not associated to an actual user request) notably for ads and analytic services. We additionally find that some webpages on the list are either no longer valid or implement access control. By discarding URLs that either never responded to our request, or returned non-HTML content (*e.g.,* JSON or XML) we obtain 317,000 *valid* HTML pages.

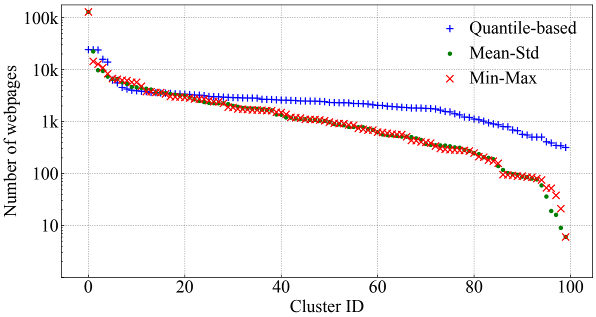


Fig. 3: Number of webpages per cluster.

Clustering: On this set of webpages, we compute six features that, as reported by Butkiewicz et. al [28], are distinctive of the page characteristics and in particular have high correlation with webpage *complexity*: page size (in MB), total number of objects, number of images, CSS, javascript, and number of distinct origins. Then, we rely on K-means to find pages of similar complexity. Given our crowd-measurement budget, we fix $K = 100$. We experiment with three standard feature normalization techniques: min-max, mean-std, and a quantile-based feature normalization, where we transform the original values of each feature to the quantiles they correspond to in the dataset (*e.g.,* for `amazon.com`, page size: 38%, num imgs: 99%, num domains: 61%).

Fig. 3 reports the number of pages per cluster produced by K-means, ordered by decreasing cluster size. We observe that, due to the wide range of values for page size and number of objects in the dataset (up to 75 and 1,429Mb, respectively), both min-max and mean-std normalization create several "outlier clusters" near the extreme ranges of each feature with very few pages (less than 10), while creating a single overcrowded cluster for simple and small pages. We note that quantile-based normalization results in clusters that represent a sizeable number of pages (the smallest 5 clusters contain between 155 and 347 pages) while at the same time helps in better representing the fine-grain diversity of relatively small pages (there is no single giant cluster). Upon a closer analysis, we put aside 14 clusters that contained a large number of "error pages". These cluster included regular HTML pages reporting 401, customized 404 pages, pages with valid HTML but no actual content, as well as login pages. We observe that the 5 largest clusters still represent

30% of all pages: therefore, a stratified selection strategy helps avoiding oversampling these pages.

Stratified selection: From each of the remaining 86 clusters, we manually pick one webpage for user evaluation in Eyeorg. We do this by choosing a popular webpage according to the ranking, *i.e.,* which is simultaneously (i) the closest to the centroid, (ii) in English language, and that (iii) does not contain offensive or adult content (*e.g.,* porn, gambling), in order to avoid exposing crowdsource participants to upsetting content.

Given the fair amount of work involved, this list of "representative" webpages is interesting per se, and we make it available [11]. Finally, we add 22 handpicked webpages that we also studied in previous work [14] to obtain a total of 108 sampled webpages.

B. Objective Web Quality Metrics

For each webpage of the set, we collect the objective Web quality metrics discussed in Sec. II. We rely on a Chrome extension [29] to measure all the metrics, since some metrics require the rendered position of all objects in the page, cannot be measured from the HAR file (as opposed to PLT, DOM, TTI, etc.) and thus are better measured directly from the browser. Further, we use `FFmpeg` to record videos of a webpage rendering process.

We instrument a stock version of Chrome (v68.0.3440.84) with the above extension and attempt to load the 108 selected webpages, consecutively. For each load, we set a maximum duration of 15 seconds and also record *webm* videos of the rendering process. This is needed since we then plan to crowdsource users responses with Eyeorg. Since headless Chrome currently does not support extensions, we leverage the X virtualframe buffer *Xvfb* to allow remote execution without the need for a physical monitor. We measure each webpage 5 times, ensuring warm DNS caches, and a clean browser profile at the beginning of each run. We then select the experiment (video and set of performance metrics) with the median PLT among the 5 repetitions.

C. UPLT Crowdsourcing

We measure uPLT via Eyeorg's *timeline* experiment [10] where a participant is asked to "scrub" the video of a webpage load until when (s)he considers the page to be ready. We run a single Eyeorg campaign targeting the above 108 webpages and 1,000 paid participants from Figure Eight[2] (total cost: $120). Each participant evaluates 6 videos—thus generating 6,000 uPLT values or about 54 valid feedbacks per webpage, on average.

In Figure Eight, we request the highest quality participants. As discussed in the Eyeorg paper, we also filter user responses using a mix of their *engagement* (*i.e.,* the time spent on task) and the quality of their opinions using some control questions. Eyeorg implements control questions on top of the *frame selection helper*, a tool that helps the user "rewinding" her uPLT selection if an equal[3] (earlier) frame is identified. This is needed because, for some users, it can be hard to scrub a video exactly to the earliest point associated with a selected frame. For one video out of six, the frame selection helper suggests the very first video frame as a rewind option. Users that blindly accept this suggestion without noticing the obvious difference between the two frames are considered as potentially distracted, and their responses are discarded. In total, we discard 172 users due to low engagement and due to failing the control questions.

IV. Understanding Users' Feedback

In order to provide an in-depth characterization of user feedback, we start our analysis by checking the existence of multiple modes on the uPLT distribution, considering for each webpage the valid uPLT feedbacks. Then, we study the number and parameters of the different modes exhibited by the uPLT distribution for each webpage. Finally, we investigate how the complexity of modern webpages (e.g., number of objects, domains, etc.) and user browsing behavior may affect uPLT multi-modality.

A. UPLT Distribution Analysis

We next analyze the uPLT distributions to inspect the presence of multi-modal behaviors. For this purpose, we rely on a non-parametric statistical test widely used to assess whether a distribution of real-valued random variables, such as the uPLT, is likely to be uni-modal [30]. This test computes the dip statistic as the maximum difference between the empirical cumulative distribution function (ecdf), and the uni-modal distribution function that minimizes that maximum difference. When we perform the dip test, we employ the common threshold $p < 0.05$ to reject the null hypothesis of uni-modality. We find on our set that 56 webpages are likely to exhibit a uni-modal distribution of uPLT and 52 a multi-modal one. By lowering this threshold, the number of likely multi-modal webpages decreases, *e.g.,* when $p < 0.01$ only 42 pages are estimated as multi-modal.

For the webpages found to be likely multi-modal, we model their uPLT distributions with a Gaussian mixture model (GMM), *i.e.,* a weighted sum of K independent Gaussian distributions. The question that naturally arises is how many Gaussian components K have to be considered per webpage. By letting the parameter K of the GMM range from 2 to 10, we observe that the GMM accurately models the uPLT distribution for $K \geq 3$. However, we find that even for $K = 3$ some webpages have small modes (34 webpages have at least one of the three components with weight lower than 0.05).

We run the goodness-of-fit Kolmogorov-Smirnov test, with a confidence level of 0.95. The null hypothesis is that the empirical uPLT and the mixture distribution (which we sample to obtain the same number of samples as the

[2] https://www.figure-eight.com/

[3] No more than 1% different in a pixel-by-pixel comparison.

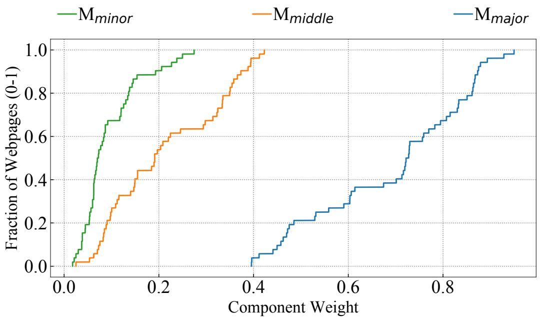


Fig. 4: Component weights for pages with multi-modal uPLT.

M_{time} M_{mass}	M_{first}	M_{second}	M_{third}
M_{major}	63%	33%	4%
M_{middle}	29%	38%	33%
M_{minor}	8%	29%	63%

TABLE I: Breakdown for $M_{mass} = M_{time}$.

Page Feature	μ	σ	25%	50%	75%
Size [MB]	854/970	862/1,687	176/136	564/439	1,382/1,145
# Objects	53/81	47/50	17/44	46/76	72/106
# JS	15/21	14/14	5/9	10/19	22/28
# Images	20/34	27/31	4/12	10/26	22/44
# CSS	13/16	10/15	5/6	12/10	16/24
# Domains	7/11	8/9	3/6	4/8	9/14

TABLE II: Statistics of uni-modal/multi-modal pages.

empirical one) with $K = 3$ come from the same distribution. The result shows that, for more than 70% of the multi-modal webpages, the null hypothesis is confidently accepted. Hence, for each likely multi-modal webpage, we set $K = 3$ and find the corresponding GMM parameters from its uPLT distribution: mean, standard deviation, and weight of each component.

Fig. 4 shows the weights' distribution of the three components sorted by their mass, $M_{mass} = [M_{major}, M_{middle}, M_{minor}]$, across the 52 likely multi-modal webpages. We observe that the weight ranges from 0.40 to 0.95 for the major component (M_{major}), which represents on average 69% of users (median 72%), 0.03 to 0.43 for the middle (M_{middle}) component, and 0.02 to 0.27 for the minor component (M_{minor}). For some outlier webpages, such as *booking.com*, users are split into multiple well defined modes of similar size. In the opposite case, there are webpages such as *paperpile.com* where nearly all the users agree on a single uPLT value, with two other smaller modes ($M_{major,mass}$ = 0.86, $M_{middle,mass}$ = 0.10, $M_{minor,mass}$ = 0.04).

The uPLT components can alternatively be sorted by occurring time, in such a way that the user opinion is split among M_{time} = [M_{first}, M_{second}, M_{third}] on multi-modal webpages (so that M_{first} refers to the earliest in time and M_{third} to the latest one). By analyzing the modes defined in these two distinct ways, we can check when each M_{mass} mode coincides with each M_{time} mode.

Tab. I shows the percentage of occurrences for each of the 9 couples of $M_{mass} = M_{time}$ ($M_{major} = M_{first}$, $M_{major} = M_{second}$, etc.). This gives us information on which time sorted mode M_{time} is more liable to be the most or least popular one (M_{mass}). The table highlights that the majority of users are more likely to prefer the earliest modes: the major mode M_{major} is indeed equivalent to the earliest mode M_{first} on 33 pages (63% of the whole multi-modal webpages set), it is equal to the second one M_{second} on 17 pages (33%), and finally it is equivalent to the latest third mode M_{third} on just 2 pages (4%). Reversely, the minor mode M_{minor} tends to rarely coincide with the earliest one M_{first} (8%): it is actually most of the time equal to the latest mode M_{third} (63%) and sparingly to the second one M_{second} (29%). We can finally conclude that the mapping between mass and time sorted modes is such that the *most* popular mode is generally also the earliest in time and vice versa.

B. *Page Characteristics and uPLT*

Given that half of the webpages exhibit a multi-modal uPLT, we investigate which of their characteristics (*e.g.,* number of objects, images, domains, etc.) may cause a split of users' feedback with respect to when the page is loaded. For example, ads heavy webpages might be (at least) bi-modal since some users consider the page to be loaded before ads are shown, while some others would wait for the whole content to be retrieved and displayed.

Tab. II illustrates several statistics (average, standard deviation, 25/50/75th percentile) of the webpage characteristics we considered during our stratified URL selection (see Sec. III). We can observe that the standard deviation of the size of multi-modal webpages is double with respect to that of uni-modal ones (at the 100% percentile we have 10,338 vs 3,460). We also observe that the mean number of images and of distinct origin domains for multi-modal webpages is respectively 34 and 11 compared to 20 and 7 for uni-modal webpages. This is inline with the intuition that complex webpages are more likely to be multi-modal. We next inspect how prevalent is advertising across these websites by matching the content received against EasyList,[4] a list of known advertisement domains. We find that multi-modal websites contain, on average, 5 times more advertisements than uni-modal websites, likely segmenting user opinions on uPLT.

We are now interested in assessing the importance of each of the above features for predicting uPLT multi/uni-modality of webpages. For this task, we train a Random Forest Classifier with 25 estimators, which on a 7-fold cross validation[5] achieves an average precision of 0.69

[4] https://easylist.to/easylist/easylist.txt

[5] Seven-fold cross-validation ensures that the validation dataset is at least 15% of the size of the whole dataset.

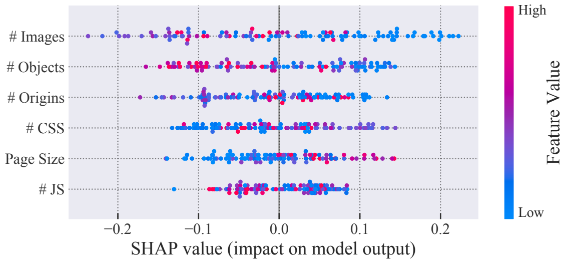


Fig. 5: Ranking of the features according to their SHAP values when predicting uni-modal pages.

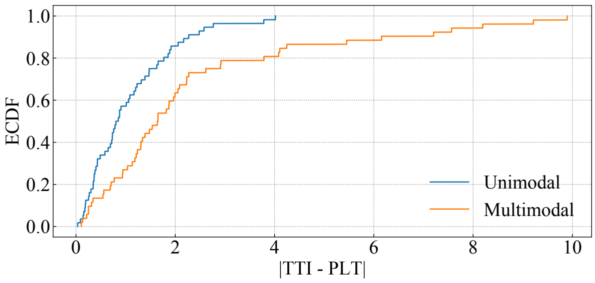


Fig. 6: $|PLT - TTI|$ ECDF for uni/multi-modal pages.

$RMSE_{\mu,Metric}$	TTFP	TTI	AATF	PLT
$\mu(uPLT)$	2.48s	2.35s	1.99s	**1.48s**

$wRMSE$	TTFP	TTI	AATF	PLT
$\mu(uPLT)$	3.10s	**2.45s**	2.57s	2.64s
M_{major}	2.03s	**1.84s**	2.54s	3.27s
M_{middle}	4.89s	4.33s	**4.23s**	4.29s
M_{minor}	9.36s	8.69s	8.67s	**7.96s**
M_{first}	**1.44s**	1.81s	2.80s	3.79s
M_{second}	4.60s	3.74s	**3.72s**	3.65s
M_{third}	11.71s	10.89s	10.39s	**9.14s**

TABLE III: RMSE of (top) uni-modal and (bottom) multi-modal uPLT with Web quality metrics.

and an average recall of 0.68. In line with the current trend towards human interpretable machine learning and model explainability, we leverage SHAP (SHapley Additive exPlanations) [31] to understand which features can better reveal whether a webpage is *uni-modal* or not. We report in Fig. 5 the 6 features, sorted by the sum of the SHAP magnitude values computed for all the webpages. SHAP values capture the effect of removing a feature for a given prediction under all possible combinations of presence or absence of the other features. Hence, they provide a quantitative insight of the importance of each feature for the model. The positive x-axis values assess the impact on the model output for predicting the *uni-modal* class, whilst the negative ones refer to the *multi-modal* class. We can observe that the two most influential features are the number of images and the number of objects present in the webpage. In particular, the lower the values of these features, the higher their SHAP value (up to 0.2 for the number of images and 0.15 for number of objects). In other words, for simple webpages with few images and objects, users more likely agree on a single uPLT, making the uPLT distribution uni-modal. Such effect is less evident for the other webpage properties, where low and high feature values overlap, causing a decrease in the impact factors on model prediction, probably due to the lack of additional data points to train the model. These findings provide valuable insights for designing webpages with more predictable user perception. For instance, we might expect that the uPLT measured via mobile browsers presents a unimodal distribution, as they generally load a simplified version of the webpage. We acknowledge that future studies are needed to further elaborate relevant design guidelines in this direction.

Finally, we quickly investigate if a difference in performance metrics can also explain the multi-modality of uPLT. We check whether the time difference between the early and late events (such as TTI and PLT) of the page loading process provides strong evidence of multi-modality. Fig. 6 shows the ecdf of $|PLT - TTI|$ for webpages we previously categorized as uni-modal or multi-modal. We can observe that multi-modal websites are, overall, characterized by larger $|PLT - TTI|$ differences compared to uni-modal ones. On the other hand, less than 10% of uni-modal pages had a $|PLT - TTI| > 2.2s$. This finding suggests that the rendering of these webpages naturally segment the users: some believe the page loaded as soon as the major part of the page loaded (usually closer to TTI) whereas others wait for all visible images to finish loading to consider the page fully loaded (usually closer to PLT).

C. Evaluation of Web Quality Metrics

Finally, we investigate to which extent single-valued objective Web quality metrics (see Sec. II) can approximate the different modes of the uPLT distributions exhibited by the webpages of our study. For each of the 56 webpages showing a uni-modal uPLT behavior, Tab. III reports on the top the Root Mean Square Error (RMSE) of the mean of the uPLT distribution $\mu(uPLT)$ with respect to each of the following objective Web performance metrics: TTFP, TTI, PLT and ATF, computed as Approximated Above-The-Fold (AATF) [14]. Results reveal that PLT is the metric which better approximates (lowest error term) the uPLT for webpages showing a uni-modal behavior of uPLT.

We rely on the weighted RMSE (wRMSE) to assess the quality of approximation of objective Web quality metrics for the 52 webpages with multi-modal uPLT (see the bottom part of Tab. III). This approach weights the average towards larger components, which is particularly important for better evaluating the error on M_{middle} and M_{minor}. We conduct this analysis from three different perspectives: (i) we compare the $wRMSE$ of the mean of the uPLT distribution $\mu(uPLT)$, as we did for uni-modal webpages, (ii) we examine the three modes sorted by their

TABLE IV: Summary of recent related work.

Year [ref]	Experiments scale	Measurement design	uPLT multi-modality	Metrics evaluation	uPLT modes analysis	Main focus
2012 [24]	n.a.	uPLT crowdsourcing	No	Yes	No	WQL definition and demonstration
2013 [32]	n.a.	uPLT crowdsourcing	No	Yes	No	Assessment and Models for Web QoE
2014 [26]	n.a.	n.a.	No	Yes	No	Web QoE overview
2016 [10]	1000 users, 100 webpages	uPLT crowdsourcing	Yes	No	No	uPLT metric definition
2017 [9]	50 users, 45 webpages	uPLT crowdsourcing	Yes	No	No	uPLT optimization by tracking user's eye gaze
2017 [15]	5.4k users, 115 webpages	A/B testing by showing side-by-side videos	No	Yes	No	Web browsing QoE assessment
2020 [This work]	1k users, 108 webpages	uPLT crowdsourcing	Yes	Yes	Yes	uPLT multi-modality analysis and characterization

mass M_{mass} = [M_{major}, M_{middle}, M_{minor}] (M_{major} is the mode with the largest mass of the distribution), and (iii) by occurring time M_{time} = [M_{first}, M_{second}, M_{third}] (M_{first} is the earliest).

The results summarized in Tab. III show that TTFP and TTI better approximate M_{major} and, not unexpectedly, given the duality shown in Sec. IV-A, M_{first}. On the other hand, AATF and PLT better approximate M_{middle}, M_{minor}, M_{second} and M_{third}. The former suggests that, to enhance the uPLT analysis, measuring and optimizing *the last* updates, usually achieved with PLT, is less relevant with respect to the earlier ones, *e.g.,* TTI and TTFP. The latter instead confirms that the users choosing a late uPLT agree on a page to be loaded close to the last two page tracking events. It is also an interesting validation to note that, on uni-modal pages, PLT better matches $\mu(uPLT)$ whereas TTI does that for multi-modal ones.

V. Discussion

Only few among recent works highlighted the existence of possible multi-modal behaviors for the uPLT. However, none of them deepened the study of the uPLT multi-modality or further explored the existence of these underlying different user behaviors, by carrying out the analysis of user feedback under this angle.

A summary of closely related work is reported in Tab. IV, where we distinguish for each study whether its authors identify or mention the multi-modal trait of uPLT ("*uPLT multi-modality*") or if they analyze that the uPLT is insufficiently captured by single-valued objective metrics (*"Metrics evaluation"*). For the sake of comparison with our work, we report when available, the experimental settings and the size of the measurements (amount of users and webpages involved). Specifically, we note that previous studies [24], [32], [26] observe that uPLT does not match PLT, while Gao et al. [15] find that, more generally, "commonly used navigation metrics such as *onLoad* and *TTFB* fail to represent majority human perception". We note that although these works remarked either the multi-modality of uPLT or the difficulty in mapping uPLT to single Web QoE metrics, their focus was not on characterizing the uPLT multi-modal nature. This confirms that the main hypothesis of our work is in line with the recent empirical observations in Web QoE modeling.

In this paper, we go beyond related work by (i) evaluating the fraction of uni-modal versus multi-modal pages according to a rigorous statistical test, (ii) thoroughly characterizing the different uPLT modes, and finally (iii) mapping between the different uPLT modes and the Web QoE metrics proposed in the literature. Specifically, our analysis shows that (i) the uPLT distribution is uni-modal for approximately half of the webpages in our dataset, for which a simple PLT indicator (measured via the browser `onLoad` event) is a good estimator of user perception. We also show that, among classical indicators of webpage complexity, the *number of objects* and the *number of images* are good indicators for uPLT modality. We then show that (ii) multi-modal webpages are, in practice, never characterized by more than three modes. The most prevalent mode represents no less than 40% of users (69% on average, 72% median) in our dataset. We also observe that the earliest and most popular modes tend to match.

Finally, we demonstrate that (iii) we can approximate the earliest and most popular mode by TTFP and TTI, whereas metrics such as ATF and PLT better approximate the other modes. These findings can be summarized in the following rule of thumb for measuring Web QoE using existing metrics. On the one hand, given that user browsing statistics are likely to exhibit multi-modality, one metric is generally not sufficient to faithfully capture user perception. On the other hand, the whole spectrum of user perception seems to be captured by relatively few user modes, so that a small number of metrics are good at capturing uni-modal (e.g., where PLT or AATF will suffice) as well as multi-modal behavior (e.g., where additionally TTI should be measured for increased representativeness).

VI. Conclusions

In this paper, we have asked a very simple but yet important and challenging question: *to which extent users agree on a single time for when a page is loaded?* This question is important because, traditionally, Web quality metrics (*e.g.,* PLT and SpeedIndex) are conceived to produce a unique time indicator, implicitly assuming that user opinions would statistically converge to a single value. This question is also challenging, because of the sheer size of the Web coupled with the complexity to collect and understand user opinions. We show that for around half of the webpages considered, the uPLT distribution is multi-modal and that instead, for simple webpages users more likely agree on a single uPLT. We point out our results are representative (as per the stratified sampling selection, which is interesting per se, that ensures our 100 target pages cover the initial 1M set) and repeatable (for which we have already open sourced our dataset [11]).

Whereas this paper is far from entirely closing the Web QoE measurement issue, we hope that open sourcing our dataset [11] can help the community into further nailing down the smallest set of relevant Web QoE metrics covering *all* user modes, as opposite to attempting to define yet another *single* Web QoE indicator, that would by definition fail in this task.

Acknowledgements

This project was supported by a cooperation between Huawei Technologies France SASU and Telecom Paris (Grant no. YBN2018125164) and by the ANR Project No ANR-15-CE25-0013-01 (BottleNet). We thank Liz Izhikevich for suggesting the dip test.

References

[1] X. S. Wang, A. Krishnamurthy, and D. Wetherall, "Speeding up web page loads with shandian," in *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. USENIX Association, 2016, pp. 109–122. [Online]. Available: https://www.usenix.org/conference/nsdi16/technical-sessions/presentation/wang

[2] "Shopzilla: faster page load time = 12 percent revenue increase," http://www.strangeloopnetworks.com/resources/infographics/web-performance-andecommerce/shopzilla-faster-pages-12-revenue-increase/, 2016.

[3] M. Butkiewicz, D. Wang, Z. Wu, H. V. Madhyastha, and V. Sekar, "Klotski: Reprioritizing web content to improve user experience on mobile devices," in *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. USENIX Association, 2015, pp. 439–453. [Online]. Available: https://www.usenix.org/conference/nsdi15/technical-sessions/presentation/butkiewicz

[4] R. Netravali, A. Goyal, J. Mickens, and H. Balakrishnan, "Polaris: Faster page loads using fine-grained dependency tracking," in *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. USENIX Association, 2016. [Online]. Available: https://www.usenix.org/conference/nsdi16/technical-sessions/presentation/netravali

[5] V. Ruamviboonsuk, R. Netravali, M. Uluyol, and H. V. Madhyastha, "Vroom: Accelerating the mobile web with server-aided dependency resolution," in *Proc. of the Conference of the ACM Special Interest Group on Data Communication*. ACM, 2017, pp. 390–403.

[6] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler, "Waiting times in quality of experience for web based services," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, July 2012, pp. 86–96.

[7] J. Brutlag, Z. Abrams, and P. Meenan, "Above the fold time: Measuring web page performance visually," in *Velocity: Web Performance and Operations Conference*, 2011.

[8] Google, "Speed Index," https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index.

[10] M. Varvello, J. Blackburn, D. Naylor, and K. Papagiannaki, "EYEORG: a platform for crowdsourcing web quality of experience measurements," in *Proc. ACM CoNEXT*, 2016.

[9] C. Kelton, J. Ryoo, A. Balasubramanian, and S. R. Das, "Improving user perceived page load times using gaze." in *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2017, pp. 545–559.

[11] 2019, https://bit.ly/2VW2Gnd.

[12] A. S. Asrese, E. A. Walelgne, V. Bajpai, A. Lutu, Ö. Alay, and J. Ott, "Measuring web quality of experience in cellular networks," in *Proc. Passive and Active Measurement Conference (PAM)*, D. Choffnes and M. Barcellos, Eds. Cham: Springer International Publishing, 2019, pp. 18–33.

[13] E. Bocchi, L. De Cicco, and D. Rossi, "Measuring the quality of experience of web users," in *ACM SIGCOMM Workshop on Internet-QoE'16*, 2016.

[14] D. N. da Hora, A. S. Asrese, V. Christophides, R. Teixeira, and D. Rossi, "Narrowing the gap between qos metrics and web qoe using above-the-fold metrics," in *Proc. Passive and Active Measurement Conference (PAM)*, 2018, pp. 31–43.

[15] Q. Gao, P. Dey, and P. Ahammad, "Perceived performance of top retail webpages in the wild: Insights from large-scale crowdsourcing of above-the-fold qoe," in *Proc. of the Workshop on QoE-based Analysis and Management of Data Communication Networks*. ACM, 2017, pp. 13–18.

[16] Z. W. (Ed.), "Navigation timing," in *W3C Recommendation*, Dec. 2012.

[17] J. Brutlag, Z. Abrams, and P. Meenan, "Above the fold time: Measuring web page performance visually," [Online]. Available from: http://conferences.oreilly.com/velocity/velocity-mar2011/public/schedule/detail/18692.

[18] A. Huet, Z. Ben Houidi, S. Cai, H. Shi, J. Xu, and D. Rossi, "Web quality of experience from encrypted packets," in *ACM SIGCOMM Posters and Demos*, 2019.

[19] M. Trevisan, I. Drago, and M. Mellia, "Pain: A passive web performance indicator for isps," *Computer Networks*, vol. 149, pp. 115 – 126, 2019.

[20] "Rum speedindex," [Online]. Available from: https://github.com/WPO-Foundation/RUM-SpeedIndex.

[21] E. Bocchi, L. De Cicco, M. Mellia, and D. Rossi, "The Web, the Users, and the MOS: Influence of HTTP/2 on User Experience," in *Proc. Passive and Active Measurement Conference (PAM)*, 2017.

[22] F. Salutari, D. D. Hora, G. Dubuc, and D. Rossi, "A large-scale study of wikipedia users' quality of experience," in *The Web Conference (WWW'19)*, May 2019.

[23] T. Hossfeld, P. E. Heegaard, M. Varela, and S. Moller, "Qoe beyond the mos: an in-depth look at qoe via better metrics and their relation to mos," *Quality and User Experience*, 2016.

[24] S. Egger-Lampl, P. Reichl, T. Hoßfeld, and R. Schatz, "Time is bandwidth? narrowing the gap between subjective time perception and quality of experience," in *Proc. IEEE International Conference on Communications (ICC)*, 2012.

[25] T. Hoßfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, "The memory effect and its implications on web qoe modeling," in *Proc. International Teletraffic Congress (ITC)*. International Teletraffic Congress, 2011, pp. 103–110.

[26] D. Strohmeier, S. Egger, A. Raake, T. Hoßfeld, and R. Schatz, "Web browsing," pp. 329–338, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-02681-7_22

[27] D. Hubbard, "Cisco umbrella 1m. (2016)," 2016. [Online]. Available: https://umbrella.cisco.com/blog/blog/2016/12/14/cisco-umbrella-1-million

[28] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Understanding website complexity: measurements, metrics, and implications," in *Proc. ACM Internet Measurement Conference (IMC)*. ACM, 2011, pp. 313–328.

[29] Approximate ATF chrome extension, https://chrome.google.com/webstore/detail/approximate-atf/eedmonedcfjniaagehchbkdolbobmfhb.

[30] J. A. Hartigan and P. M. Hartigan, "The dip test of unimodality," *Ann. Statist.*, vol. 13, no. 1, pp. 70–84, 03 1985. [Online]. Available: https://doi.org/10.1214/aos/1176346577

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[32] R. Schatz, T. Hoßfeld, L. Janowski, and S. Egger-Lampl, *From Packets to People: Quality of Experience as a New Measurement Challenge*, 01 2013, pp. 219–263.