Analyzing Wikipedia Users' Perceived Quality Of Experience: A Large-Scale Study

Flavia Salutari, *Telecom Paris*, Diego Da Hora, *Telecom Paris*, Gilles Dubuc, *Wikimedia Foundation*, Dario Rossi, *Huawei Technologies, Co. Ltd*, *Senior Member, IEEE*

***Abstract*—The Web is one of the most successful Internet applications. Yet, the quality of Web users' experience is still largely impenetrable. Whereas Web performance is typically studied with controlled experiments, in this work we perform a large-scale study of a real site, Wikipedia, explicitly asking (a small fraction of its) users for feedback on the browsing experience. The analysis of the collected feedback reveals that 85% of users are satisfied, along with both expected (e.g., the impact of browser and network connectivity) and surprising findings (e.g., absence of day/night, weekday/weekend seasonality) that we detail in this paper. Also, we leverage user responses to build supervised data-driven models to predict user satisfaction which, despite including state-of-the art quality of experience metrics, are still far from achieving accurate results (0.62 recall of negative answers). Finally, we make our dataset publicly available, hopefully contributing in enriching and refining the scientific community knowledge on Web users' QoE.**

***Index Terms*—Quality of Experience, Network Performance Analysis, World Wide Web, Network Measurements.**

I. Introduction

SINCE its inception, the World Wide Web has sometimes been dubbed as World Wide "Wait" [1]. Slow rendering of the websites happened due to dial-up connections in the 80s, slow 2G connections in the 90s and so on, but it also persists nowadays for several reasons including unexpected sources of latencies [2], interactions between network protocols [3], the growingly more complex structure of websites [4], an increased usage of mobile devices [5], [6] and the emergence of new protocols [7]. Yet, whereas the study of Web performance is commonly [3], [4], [5], [8], [9], [10], [7], [6] tackled via simple objective metrics [11], and rather typically via the Page Load Time (PLT), the quality of Web users' experience is still largely impenetrable [12], [13]. As such, a number of alternative metrics that attempt at better fitting the human cognitive process (such as SpeedIndex, user-PLT etc., see Section II) have been proposed as a proxy of users' Quality of Experience (QoE), whose monitoring is important for both Over The Top (OTT) operators to keep users engaged as well as for Internet Service Providers (ISP) to lower user churn.

At the same time, studies involving more advanced metrics are typically validated with rather small-scale experiments, either with a small number of volunteers, or by relying on crowdsourcing platforms to recruit (cheap) labor and produce a dataset labeled with user opinion. Often, *videos* of websites rendering process are used (as opposite to actual browsing), with possibly very specific instruction (e.g., such as in A/B testing, by clicking on the fastest of two rendering processes) that are however rather different from the cognitive process in action during the typical user browsing activities. Additionally, such tests are carried on a limited number of fixed conditions, with a small heterogeneity of devices, OSs and browsers, and are not exempt from cheating so that ingenuity is needed to filter out invalid answers from the labeled dataset [14], [15]. Finally, because these tests are carried on a limited number of pages, it is possible to evaluate computationally costly metrics, such as those that require processing the visual rendering of the website, which would hardly be doable in the World "Wild" Web. Our aim is instead to take a completely different approach and perform a large-scale study of a popular website in operation, by explicitly asking a fraction of users for feedback on their actual browsing experience. Clearly, the approach is challenging but it opens the possibility to gather more relevant user-labels, as they are issued from *real users of a real service*, as opposite to crowdworkers payed to play a game (e.g., find which video completes first as in A/B testing). We do so by launching a measurement campaign over Wikipedia, that has gathered over $62k$ survey responses in nearly 5 months. We complement the collection of user labels with objective metrics concerning the user browsing experience (ranging from simple PLT [11] to sophisticated SpeedIndex [16]), and harvest several data sources to further enrich the dataset with several other informations (ranging from technical specification of the user device to techno-economic aspects tied to the user country) so that each user survey answer is associated with over 100 features. This work extends [17], that limitedly focused on forecasting user experience, with the following main contributions:

- First, we use survey data to characterize user satisfaction along temporal and spatial dimensions: shortly, we find that user satisfaction does not exhibit seasonality at daily/weekly timescales (which is unexpected) and we document evidence of spatial dependency across many of the collected features (e.g., network access, browsing equipment, country wealth, etc.).
- Second, we use labels to build data-driven models of user experience: despite including performance metrics considered to be the state-of-the art in user quality of experience, we find that the model still falls short from attaining satisfactory performance in operational settings.
- Third, in spirit with the current trends toward research reproducibility, after carefully removing of sensitive information (see Section III-C), we release the collected dataset at [23], hopefully helping the scientific commu-

TABLE I
SUMMARY OF RECENT RELATED WORK GATHERING USER FEEDBACK FOR WEB QUALITY OF EXPERIENCE ASSESSMENT.

Year [ref]	Scale/heterogeneity						Experimental settings	Main focus
	Lab+CW[1]	Pages	Network[2]	Sw[3]	Hw[4]	Samples		
2015 [13]	0 + 120	30	-	-	-	3.6*k*	Prioritize elements (Above The Fold and user ratings)	Per-user content prioritization
2016 [14]	100 + 1*k*	100	n.a.	1	1	6*k*	Side-by-side videos (of the same site)	uPLT metric definition
2017 [18]	147 + 0	25	32	1	1	4*k*	Controlled browsing experiments	HTTP vs HTTP/2
2017 [19]	28 + 323	28	3	1	1	2.5*k*	Side-by-side videos of the same website in different protocol settings	HTTP/2 push impact
2017 [15]	0 + 5.4*k*	500	16	1	1	40*k*	Side-by-side videos (160 different website pairs)	PSI metric definition
2017 [12]	50 + 0	45	1	1	1	2.2*k*	Webcam, eye-tracking glasses	Eye gaze, uPLT
2018 [20]	241 + 0	12	n.a.	1	1	9*k*	Controlled browsing experiments	ATF metric definition
2019 [21]	0 + 50	7	11	1	1	n.a.	User rating of video rendering of Web browsing	QoE-aware networking
2019 [22]	35 + 1.2*k*	5	3	1	1	10*k*	User rating of video rendering of Web browsing	QUIC protocol
this study	62*k users*	46*k*	3.8*k ISPs*	45	2.7*k*	62*k*	User feedback from real browsing activity	User satisfaction

[1]Crowdworkers, [2]Number of controlled network conditions, [3]Software browser, [4]Hardware device

nity in refining its understanding of Web users' QoE.

In the remainder of this paper, after overviewing the related work (Section II), we explain the feedback collection process and dataset (Section III), which we dissect under both temporal and spatial angles (Section IV) and that we leverage to build a data-driven model of Wikipedia users' quality of experience experience (Section V). We finally discuss current limitations in Web QoE assessment and possible directions to circumvent them (Section VI) and summarize our findings (Section VII).

II. BACKGROUND

Assessment of Web users' quality of experience can be traced back to [24], that was among the first to adapt classic results of psycho-behavioral studies gathered in the *computer* domain [25] (in turn inspired by work by Weber and Fechner in the late 1800s), to the *computer-network* domain. This knowledge was later embedded into standards ITU-T G1030 [26], [27] (and models [28]) that encode the Weber-Fechner logarithmic [26], [27] (or exponential [28]) relationship between a stimulus (e.g., a delay) and its perceived impact (e.g., nuisance for Web users). However, while logarithmic models are valid for simple waiting tasks (e.g., file downloads), the case of interactive Web browsing is knowingly much more complex, as ITU-T G1031 [29] and [30] first pointed out.

Still, with some exceptions [13], [19], [14], [20], [31], [32], [33] most studies still rely on simple metrics such as the Page Load Time (PLT) to assess the expected impact of new Web protocols [3], [4], [7], [10], Web accelerators [9], [8], [21], and devices [34], [5]. While reducing delay is clearly a desirable objective, it is however unclear if (and by how much) a latency reduction translate into a better perceived experience, which is the ultimate goal of the above studies. In other words, while the importance of *delay* in human perception is agreed upon, the exact relationship between the Web response time and user satisfaction appear much less clear than it appeared to be [35], and motivated a proliferation of new metrics proposals and validation studies attempting at going beyond PLT. Given that many different definitions of PLT [36] are used in the literature, we specify that in this work we denote PLT as the time elapsed between the `fetchStart` and `loadEventStart` browser events defined by W3C Navigation Timing [11].

A. Web QoE metrics

As we are interested in measuring browsing experience on individual pages, *engagement* metrics such as those used in [37], [38] are clearly out of scope. As such, objective metrics of interest for Web user QoE can be divided in two classes. On the one hand, there are metrics that either *pinpoint precise time instants*: notable examples include the time at which the Document Object Model (DOM) is loaded or becomes interactive (TTI), the time at which the first element is painted (TTFP) or the time when the Above The Fold (ATF) portion of the page is rendered [39] etc. Most of these metrics are available from the browser navigation timing [11] or can be inferred from packet/flow-level traffic [40], [41] as proxy of user experience. For instance, [13], [21] aim at prioritizing delivery of content that is rendered above the fold, either arbitrating among sessions [21] or further specializing content relevance for each user [13].

On the other hand, there are metrics that *integrate all events of the waterfall* representing the visual progress of the page, such as SpeedIndex [16] and variants [42], [15], [43], that have received significant attention lately. Initial definitions in this family required capturing movies of the rendering process [16], or to further use similarity metrics SSim [15], making difficult to use them outside a lab environment. To counter this issue, simple approximations such as the ObjectIndex/ByteIndex [42] that merely count the fraction of objects/bytes received (over the total amount), or as the RUM SpeedIndex (RSI) [43] that use areas of rectangles for objects as they are painted on screen (over the total screen size) have been proposed. In this paper, we use RSI, which is among the most advanced Web QoE metrics considered to be the current industry standard. Finally, while we are aware that more complex approaches involving the spatial dimension (i.e., eye gaze) also exist [13], [12], we prefer to leave them for future work (cfr. Section VI).

B. State of the art limitations

At the same time, the above metrics suffer from a limited validation with user feedback. Typical approaches are to crowdsource the validation with A/B testing [15], [14], or by performing experiments on real pages in controlled

conditions [20], [35], [44]. Both approaches have their downsides. Controlled experiments with real HTTP server/clients and emulated network conditions enable for a more faithful and interactive browsing experience, but are harder to scale, topping to few hundreds users and few thousands data points [20]. A/B tests try to circumvent this limit, but introduce other limitations. First and foremost, A/B testing is hardly representative of Web browsing activity, since crowdworkers are instructed to select which among two videos, that they are passively screening side-by-side and that correspond to two different Web rendering processes, appears to finish first – whereas it is known that even for a simple Web browsing task such as information seeking, already different types of searches are rather different from the user standpoint in terms of cognition, emotion and interaction [45]. In other words, these experiments inform us that humans can perceive differences in these rendering processes, however they fail to signify if these perceptible rendering changes would impact the user satisfaction through the course of a normal browsing session.

The time at which users consider the process finished is denoted as user-perceived-PLT (uPLT) [14] or Time To Click (TTC) [15] and is often used as a ground truth of user perception. Yet, when users select a uPLT in [14], they are proposed with similar frames at earlier times, which has the beneficial effect of clustering answers and make uPLT more consistent at the price of possibly inducing a bias. Similarly, [15] employs SpeedIndex and TTC to forecast which among the left or right video was selected by the user at time TTC: the classifier in [15] is accurate in predicting which of the two videos is perceived as fastest by users. Yet, findings in [15] are not informative about whether the user would have been dissatisfied from the slower rendering had s/he actually been browsing.

C. Our contribution

To get beyond the limitations of controlled and crowdsourced experiments just exposed in Section II-B (e.g, few users involved, lack of real user behavior representativeness and low data heterogeneity), in this work we are the first to query, at scale, Web users for their feedback on the quality of their browsing experience. We remark that this approach is rather common with VoIP services (e.g., Skype, Hangouts often ask for Mean Opinion Score (MOS) rating at the end of the call), but to the best of our knowledge, with the exception of our preliminary work presented in [17] that this work extends, this has not been attempted before on the wide and wild Web. Specifically, we ask users for a slightly more than binary feedback (i.e., acceptance, see Section III), which let us carry on a thorough characterization of user satisfaction (see Section IV) and formulate a simple, yet hard, binary classification problem (see Section V). Particularly, our preliminary work [17] focused on the classification view, that is only briefly touched here (Section V), and of which we provide a more extensive view in a companion technical report [46]. This paper instead extends [17] by an orthogonal and thorough characterization of user QoE along the temporal and spatial dimensions (Section IV).

The usefulness of the investigation we carry on and of the models we propose is clear when we consider that recent work such as [47] still employs simple *response times* as a proxy of user sastisfaction for Web performance, whereas authors [48] go at a deep level to investigate performance of Video application, considering a more involved Pseudo-Subjective Quality Assessment (PSQA) involving a Random Neural Network (RNN). It is thus clear that, whereas models for video quality abunds [49], [50], scientific community still misses an established and agreed MOS model for Web performance. On the one hand, to perform large scale studies, the community started adopting more accurate objective models (as in [51], [52], [32]), that are inspired by metrics such as SpeedIndex that we consider in this work. On the other hand, we point out that, the human component is generally missing in large scale studies, which is among the main contribution of this work. Particularly, our collection effort allows us to perform a large scale study *across the human dimension*, to levels that were previously unprecedented.

Compared to recent literature, compactly summarized in Tab. I, we are the first to involve a large number of real users ($62k$ from $59k$ distinct IP addresses) accessing a diverse set of pages ($46k$ Wikipedia pages, which are more likely similar among them than the set of different websites used in other studies), gathering over $62k$ user responses overall (more than twice the survey responses collected in similar large-scale Wikipedia studies [53]). Particularly, whereas most of the studies involving lab volunteers & crowdworkers employ a single browser and hardware (since crowdworkers are shown videos rendered with a single browser and hardware combination) on a relatively small set of synthetic controlled network conditions (1–32), in our dataset we observe 45 distinct browsers software used on over 2,716 hardware devices[1] on 3,827 ISPs – a significant change with respect to artificial and controlled lab conditions, which make the dataset that we release at [23] of particular interest.

III. User feedback collection

Wikipedia is, according to Alexa [54], the 5th most popular website, with over 1 billion monthly visitors, that spend over 4 minutes over 3 pages on average per day on the site. We engineer a survey that is triggered after the page ends loading and collects user feedback (Section III-A), that we augment with additional information (Section III-B).

We note that, while this paper is not the first in leveraging Wikipedia surveys in general (see e.g., [53]) this is the first to gather user feedback on quality of Web browsing experience from operational websites, for which we believe releasing the dataset can be valuable for the community. To make sharing of the dataset possible, we take special care into making user and content deanonymization as hard as possible, without hurting the dataset informative value as much as possible (Section III-C). In this section, we also perform a preliminary assessment of the collection methodology, to confirm the absence of bias in the response process (Section III-D).

[1] As inferred from the *User-Agent* header field, after having filtered bots.

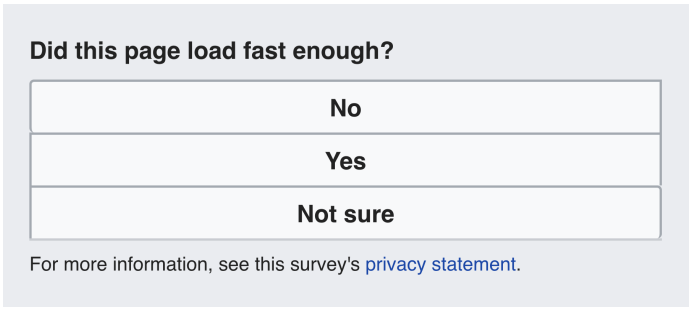


Fig. 1. Appearance of the Survey in the English Wikipedia (answer order is randomized).

A. Technical aspects of the survey collection

Due to limitations in Wikimedia's caching infrastructure, the survey is injected into the page via client-side code. Wikimedia continuously collects navigation timing performance of a randomly selected sample $\mathcal{T}$ of page views (less than 1 every 1,000 pageviews), the survey is displayed to a randomly selected sub-sample $\mathcal{S}$ of this population (less than 1 every 1,000 of the pageviews with navigation timing information) and only part of the surveys do receive an answer $\mathcal{A}$. Since $\mathcal{A} \subset \mathcal{T}$, several features (that we detail in Section III-C and analyze in Section IV-C) related to page loading performances are also available for pages sampled in the survey responses.

The survey appears on Russian, French and Catalan Wikipedias, as well as English Wikivoyage, and it is displayed in the appropriate language to the viewer. We collect the survey on mobile & desktop version of the site (but not on the mobile app). The goal of the survey is to assert whether there are Quality of Experience issues that a significant fraction of users consider to be problematic, and that Wikipedia should thus deal with. Since it is well known that "results that are only based on user ratings do not reflect user acceptance" [55], instead of asking users a 5-grade Absolute Category Ranking (ACR) score, the survey explicitly asks for user *acceptance*, i.e., users can respond with a *positive*, *neutral* or *negative* experience. For the sake of completeness, a snapshot of the survey question as it is rendered for English readers is reported in Fig. 1. To avoid biasing user answers, we randomize the order of survey answers and we avoid priming effect by refraining to explain/formulate specific survey goal (e.g., collect data to make Wikipedia faster) prior of the answer (survey purpose and data collection policies are available through the "privacy statement" hyperlink shown in Fig. 1). Similarly, neutral feedback is meant for, e.g, users that have no honest opinion, as well as users who were not paying attention during the rendering, or users that do not understand the question, etc. to avoid biasing the results (Section III-D).

The survey is injected in the DOM after the page finished loading (i.e., when `loadEventEnd` [11] fires). In order to give the survey visibility, it is consistently inserted in the top-right area of the wiki article, ensuring that it typically appears above the fold. However, as the users can freely browse the page before the survey appears, it might be out of sight when it's injected in the DOM, which is why we also record the time elapsed between the `loadEventEnd` and the moment the user sees the survey. Also, users that are shown the survey are free *not* to respond to the survey, or might as well respond very late (e.g., possibly browsing to other tabs in the meanwhile).

TABLE II
COLLECTED CORPUS OF WIKIPEDIA USERS' QOE FEEDBACK.

Period	May 24th – Oct 15th		
No. of survey requests	$\lvert\mathcal{S}\rvert =$	1746799	
No. of survey answers	$\lvert\mathcal{A}\rvert =$	62740	$\lvert\mathcal{S}\rvert/\lvert\mathcal{A}\rvert = 3.6\%$
No. of positive answers	$\lvert\mathcal{A}^+\rvert =$	53208	$\lvert\mathcal{A}^+\rvert/\lvert\mathcal{A}\rvert = 84.8\%$
No. of neutral answers	$\lvert\mathcal{A}^0\rvert =$	4838	$\lvert\mathcal{A}^0\rvert/\lvert\mathcal{A}\rvert = 7.7\%$
No. of negative answers	$\lvert\mathcal{A}^-\rvert =$	4694	$\lvert\mathcal{A}^-\rvert/\lvert\mathcal{A}\rvert = 7.5\%$

Overall, as reported in Tab. II users responded to about 3.6% of the over 1.7M surveys that have been displayed, for a total of over 62*k* answers: 84.8% of the users respond positively to the survey with an almost equal split of the remaining answers to a neutral (7.7%) or negative (7.5%) grades.

B. Collected features

We augment the collected corpus with external sources that are instrumental to the understanding of the survey responses (Section IV) as well as assist their prediction (Section V). Due to lack of space, we merely report in Tab. III a terse summary of all the metrics collected (T), as well as those we used in the preliminary conference version (WWW) and finally those that we make publicly available (PA); the full list is detailed in a companion technical report [46]. Later, we also discuss rationales of the selection for PA metrics (Section III-C).

Page: For each page, we record 15 features that concerns it (e.g., its URL, revision ID, size, etc.) and that thus are critical from a privacy point of view. We additionally record the time lapse after which the survey is shown to users, which is instead innocuous.

Performance: Since $\mathcal{S} \subset \mathcal{T}$, then all the 32 navigation-timing performance-related metrics (such as DOM, PLT, TTI, TTFP, connection duration, number of HTTP redirects and their duration, DNS wait time, SSL handshake time, etc.) are also collected. Finally, we compute the page download speed (quantized it in steps of 100Kbps). These informations are specific to page views, and are less critical to be shared.

User: The 32 collected user-related metrics include the browser, device and OS families. Additionally, we know whether users are logged in Wikipedia, if they are accessing Wikipedia through a tablet device and the number of edits that users have made (coarse bins). These informations are of course highly critical and cannot be publicly released: hence, we believe it is interesting for this paper to discuss them in detail (Section IV-C).

Environment: The 36 environmental collected features pertain time, network, geolocation and techno-economic aspects. With the exception of time information, which are directly available from the survey query, we extensively use external data sources to extract environmental features.

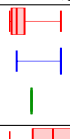
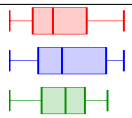
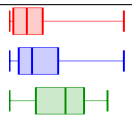
As for the network, we leverage MaxMind [56] for IP to ASN and ISP mappings and for geolocation at country (and city) granularity. ISP and ASN mappings are potentially interesting as it can be expected that performances (for the same access technology) vary across ISPs (access technology is also available for about 2/3 of the samples). Concerning geolocation, whereas databases are known not to be reliable for city-level geolocation of server addresses [57], they are generally sufficiently accurate for resolving customer IP addresses, and especially when only ISO-3166-2 country-level precision is required. Country-level precision also allows us to relatively compare performances across users in the same environment, i.e., we normalize the page download speed with respect to the median per-country speed observed in our dataset (in terms of ratio, absolute and relative error).

Additionally, ties between country wealth and network traffic volumes have been established in the literature (particularly, deviation from expected volume [58]): it is thus worth investigating whether there also exist ties between wealth and users' impatience. We use the Gross Domestic Product (GDP) information made available by the World Bank Open Data project [59]. The per-country economic features we consider (namely, per-country GDP, country GDP rank, per-country per-capita GDP, etc.) are expressed in terms of Geary-Khamis dollars, which relate to the purchasing power parity, i.e, how much money would be needed to purchase the same goods and services in two countries. The rationale in so doing is that, albeit Web users perception is tied to psychophysics laws [27], there may be environmental conditions that tune this law differently in each country. For instance, a fixed amount of delay (the stimulus) may have a smaller perceptual value to users of countries with poor Internet access which GDP-related features might capture: e.g., in other words, one can expect users in a high-GDP country to have better average performance and thus be more impatient than users from a low-GDP one. In particular, we use the 2012 per-country dataset provided by [60] since arguably the world-level statistics evolve on a relatively long timescale.

Finally, we expect user-home gateways [61] and particularly end-user devices [5], [6] to have a direct impact on the overall performance. As such, we complement the ISP-level view with a device-level information. Particularly, we harvest the Web [62] to find techno-economic information about user devices and in particular, collect device CPU, memory and pricing[2] information. Intuitively, this information complements the per-country GDP information as, e.g., there may be further perceptual differences between users with a costly smartphone in low-GDP vs high-GDP countries. We recognize that device CPU and memory specs are only an *upper-bound* of the achievable performance, as it is the mixture of applications installed and running on a device that determine the amount of *available* CPU and RAM resources, from which user perception will be ultimately affected [5], [6]. Missing this information on a per-sample basis, we attempt to at least construct the per-device statistics, by considering navigation timing information of a large representative sample of Wikipedia users. Particularly, we consider the month of August 2018 during which we observe over 30 million navigation time samples from 29,336 different devices, including all 2,716 devices in our survey. We then construct *deciles* of per-device performance (e.g., of page load time and similar timing information): indeed, it can be expected that users of knowingly slow devices be less impatient, which this additional data source could provide.

[2]Note that we collect pricing information at the time of our query, and not at the time when the device was actually bought; we also ignore price differences among countries, and per-ISP offer bundles.

TABLE III
SUMMARY OF THE FEATURES (T/WWW/PA) THAT ARE ASSOCIATED TO EACH USERS' SURVEY RESPONSE (FULL DETAILS IN [46]). THE MUTUAL INFORMATION BETWEEN THE SURVEY ANSWER AND T/WWW/PA FEATURES IN THE CLASS IS REPORTED AS A BOXPLOT.

Class	T/WWW/PA	Sample features	MI(x,y)
Page	15/2/1	Wiki, Page size, Survey viewtime, etc.	
Performance	32/26/18	PLT, TTI TTFP, RSI, etc.	
User	32/21/0	Device, Browser, editCountBucket, etc.	
Environment	36/12/0	Connection Type, Time, Geolocation, etc.	
Overall	115/61/19	Total WWW [17] paper Publicly Available	

C. Ethics

The dataset we collect contains obviously sensitive information allowing to deanonymize Wikipedia visitors (such as IP addresses, version of their browser and handsets), as well as linking them to the content they visited (e.g., page, revision ID, time of their visit, etc.). Despite the dataset release policy explicitly forbids user deanonymization, in the interest of respecting personal privacy we have to obscure information so to render user deanonymization as hard as possible, while still allowing meaningful information to be extracted from the data – which we detail here.

Datasets: Specifically, whereas releasing the totally collected (T) feature set is clearly impossible, we defined two different subsets of features (denoted as WWW and PA), that were scrutinized by the Wikimedia legal team. While at the time at which the preliminary conference version of this paper was published [17] the legal vetting process was still ongoing, a decision was made to release a publicly available (PA) dataset. Conversely, the legal team decided that it was not possible to release the dataset WWW used in [17], since perfect unlinkability could not be claimed (as we do not

control all *other* sources, e.g., a survey responder wishing to deanonymize himself, well-funded opponents, capable researchers, etc.) which could threaten Wikipedia user privacy.

Features: In the publicly available *(PA)* set, only performance metrics are considered, that are not linkable to any property related to time-of-day, user, content, geography, device, page, etc. This is clearly a very conservative approach, in the interest of protecting user sensitive data. In particular, the PA dataset is provided only for the Russian and French Wikipedia, which host the largest fraction of survey responses: this still provides useful data, without risking to exposing individual users. In the conference paper [17], we additionally use an intermediate dataset *(WWW)* where we selectively filtered/obscured/aggregated features. As it is relevant to contrast performance of T, PA and WWW, it is useful to briefly describe the WWW dataset as well. In particular, WWW dataset transforms data in a non bijective way (e.g., IP to ASN and ISP mappings that provide network-related properties, while preventing user deanonymization at the same time), or aggregated at a sufficiently coarse grain (e.g., country-level geolocation; obfuscation of browser major/minor version; aggregation of unpopular devices, etc.). For the same reason, we decided to aggregate time-related information at a coarse-grain (hour-level) and drop most content-related features (e.g., page ID). We quantize the page size with a resolution of 10KB, to also make it hard to reverse-engineer which page was visited. We maintained most of the navigation timing related performance features, that have the highest mutual information, which we obfuscated wherever necessary (e.g., given that with precise PLT and download speed one could easily reverse engineer the page size, and thus the content, we quantize the download speed in steps of 100Kbps).

Implications: Comparing T/WWW/PA features sets is thus useful to understand the implications of this lossy feature selection process on the quality of the released dataset. While the prediction accuracy is the object of Section V, from properties presented in Tab. III, we can expect the feature selection/transformation process to have a limited effect. Indeed, Tab. III reports the number of features that are collected overall *(T)* vs those that would have been available under a conservative *(WWW)* vetting process and the publicly available ones *(PA)*. For each class (first column), the table reports the number of T/WWW/PA features (second column), and additionally reports boxplots of the mutual information $MI(x, y)$ between features in the class and the survey answer (last column). MI expresses the amount of information (in bits) that can be obtained about the survey answers through the observed variable. Tab. III shows that, while we only consider a rather small subset of the total collected features *(T)*, the *(WWW)* and *(PA)* features have a *higher mutual information* (particularly, note that the median MI is higher in the *(WWW)* and *(PA)* feature sets). Thus, we conclude that:

- on the one hand, classification results of Section V are only minimally affected by selecting all *(T)* , some *(WWW)* or very few *(PA)* features, so that repeatability of the QoE study is not affected by the vetting process: under this angle, it is fortunate that features belonging to the performance class, which are those exhibiting the highest mutual information with the user grade, are also the ones made available, being the least critical to share.
- on the other hand, the type of study we conduct in Section IV would be impossible to reproduce with the available features *(PA)* set: under this angle, we decide to provide in this paper a through spatio-temporal characterization of the collected *(T)* dataset.

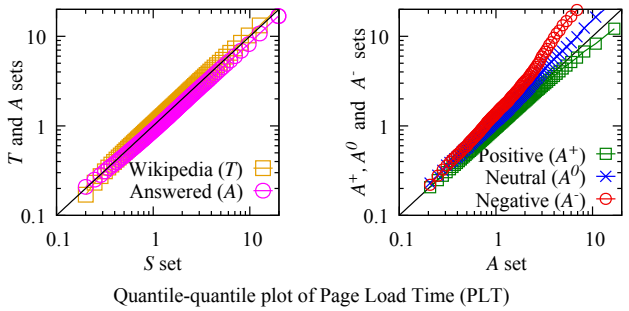


Fig. 2. Quantile-quantile plot of PLT statistics for different sets ($\mathcal{T} \supset \mathcal{S} \supset \mathcal{A} = \mathcal{A}^+ \cup \mathcal{A}' \cup \mathcal{A}^-$).

D. Validity of the collection methodology

Despite our care in engineering the survey questioning process, we cannot exclude a-priori the existence of bias in the user survey answer process. For instance, users might refrain to answer when the page loading experience was positive, and be more willing to express their opinion in case of bad experience, which would lead to under-estimate the user satisfaction.

To assess whether our survey collection methodology yields to such (or other) biases, we compare three sets of page view experiences, namely (i) the set $\mathcal{T}$ where we record navigation timing information from the browser (ii) the set $\mathcal{S}$ where users have been *shown* the survey (iii) the set $\mathcal{A}$ where users have actually *answered* to the survey. Finally, we further slice the set of answered surveys $\mathcal{A}$ according to the answer in three additional datasets with (iv) positive $\mathcal{A}^+$, (v) neutral $\mathcal{A}^0$ and (vi) negative $\mathcal{A}^-$ grades.

Among the numerous features we collect, without loss of generality we now limitedly consider the Page Load Time (PLT) distribution. Since $\mathcal{S} \subset \mathcal{T}$ is selected with uniform random sampling, by construction we have that $\mathcal{S}$ and $\mathcal{T}$ are statistically equivalent as far as individual features, such as PLT, are concerned. However, in case where users *decision* to answer to the survey (irrespectively of the actual *grade* that we consider in Section IV) would be biased by the performance of the page, then the PLT statistics should differ among the set of displayed $\mathcal{S}$ vs answered $\mathcal{A}$ surveys. The left-side of Fig. 2 reports a quantile-quantile (QQ)-plot of the empirical PLT distribution, using quantiles of $\mathcal{S}$ on the x-axis and $\mathcal{T}, \mathcal{A}$ on the y-axis, from which one can clearly remark the absence of such bias.

Conversely, one would expect that, shall the PLT affect the actual grading of the browsing experience, then PLT statistics

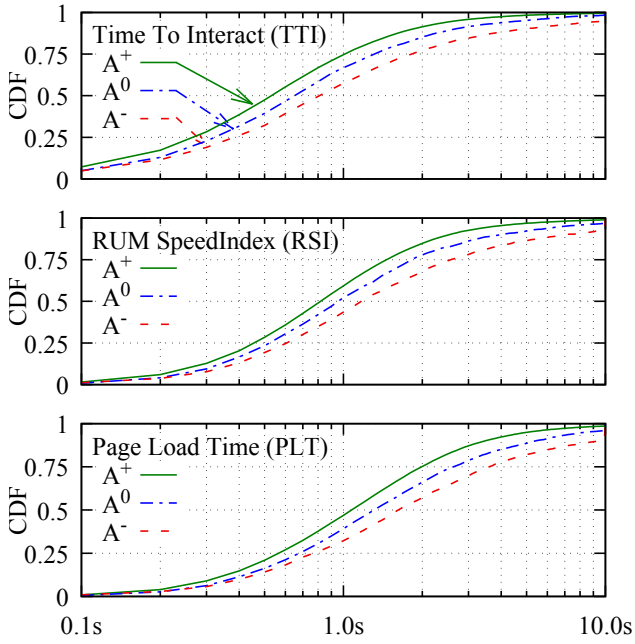


Fig. 3. Aggregate statistics of navigation timing performance (TTI, RSI and PLT in the figure), conditioned by survey response.

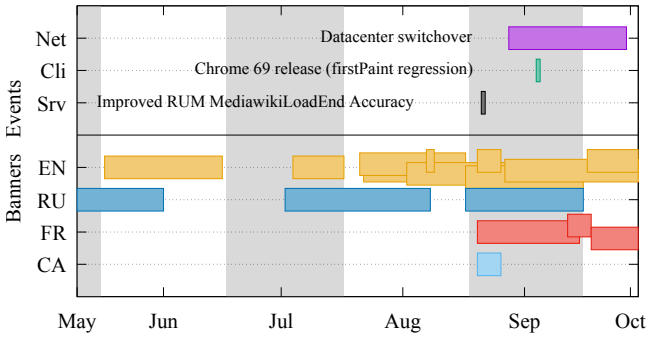


Fig. 4. Annotation of major Wikipedia-related events occurred during the whole 5-months observation period.

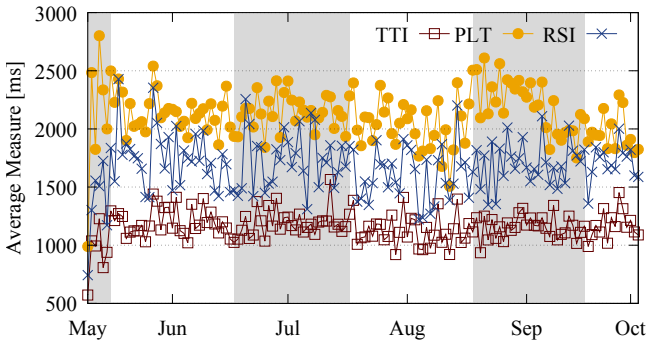


Fig. 5. Temporal view: daily mean of PLT, TTI and RSI during the period.

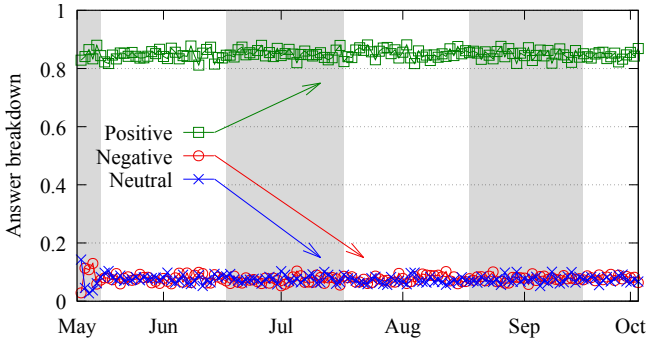


Fig. 6. Temporal view: breakdown of daily survey answers among positive, neutral and negative scores.

should differ among the $\mathcal{A}^+ \cup \mathcal{A}^0 \cup \mathcal{A}^- = \mathcal{A}$ sets. This is shown in the right-side of Fig. 2, comparing the quantiles of the answer set $\mathcal{A}$ on the x-axis to its per-grade slices on the y-axis. Several remarks are in order. First, it can clearly be seen that browsing experience with negative scores fall above the equality line, confirming as expected that the set of negatively rated pages $\mathcal{A}^-$ contains pages with longer download time compared to the positive $\mathcal{A}^+$ and neutral $\mathcal{A}^0$ sets. Second, similar considerations hold for neutral (slightly above) and positive (slightly below) answers, although they are less visible – in part, this is due since positive grades represent the bulk of the answers $|\mathcal{A}^+|/|\mathcal{A}| = 84.8\%$, for which the PLT statistics of $\mathcal{A}^+$ and $\mathcal{A}$ are mechanically more similar (we will take care of class imbalance when appropriate later on in Section V). Third, we notice that the QQ-plots of positive, neutral and negative answers overlap for quantiles corresponding to low and moderate PLT values, indicating as expected that the PLT alone cannot fully capture user perception.

IV. User feedback characterization

We start by analyzing the user feedback along aggregate (Section IV-A), temporal (Section IV-B) and spatial (Section IV-C) viewpoints, including for the time being neutral answers.

A. Aggregate view

As previously illustrated in Fig. 2, users' grades exhibit some correlation with performance metrics such as PLT. This is consistent with results reported in Tab. III, further showing that metrics in the performance class have the highest mutual information with user answers. We now consider other performance indicators beyond PLT, and depict in Fig. 3 the empirical cumulative distribution functions (ECDFs) of three representative navigation time metrics [11], slicing the dataset depending on the survey answer. Particularly, the figure includes the Time To Interact (TTI), the RUM SpeedIndex (RSI) and the Page Load Time (PLT), although we point out that results qualitatively hold for other metrics such as Time to The First Paint (TTFP). These are the most widely used metrics to express Web users quality of experience, and are among the metrics with the highest mutual information with the survey answer (namely TTI=0.032, RSI=0.024, PLT=0.04).

Two takeaways clearly emerge from the picture. First, as expected order relationships that were early shown in Fig. 2

for PLT are maintained for the TTI and RSI ECDFs, in the sense that TTI, RSI and PLT for page views having a positive score are smaller (the distribution is shifted to the left) with respect to TTI, RSI and PLT for neutral (middle curves) or negative (right curves) scores.

Second, scores are hardly separable along any of the TTI, RSI or PLT metrics: notice for instance that 75% of positive (57% negative) pages have a TTI up to 1 sec, and that similar considerations hold for RSI$\leq$1s (59% positive vs 43% negative) and PLT$\leq$1s (47% vs 32%). This raises the need for additional metrics beyond those related to performance timing, which hopefully can further assist the prediction of user scores.

B. Temporal breakdown

At a glance: We next present the daily amount of user answers over the whole 5-months period, with annotation of different Wikipedia-related events. Such events, some of which are reported in Fig. 4, are of different nature and include, e.g., the injection of banners for fundraising or the call for volunteering contributions to Wikipedia content; network-related events such as data center switchover/switchback; browser-related event such as new versions that introduce known regression in performance metrics (e.g., Chrome 69 release that introduces a `firstPaint` regression); back-end events and deployment of new features (e.g., RUM metric "MediawikiLoadEnd" improved). As it can be seen from Fig. 4, an operational website at scale continuously has events that are generally not available in testbeds (such as those overviewed in Section II), that thus sample very narrow and specific conditions that are not representative of real deployments.

Yet, these operational changes appear to have only a moderate effect on browser timing metrics: Fig 5 shows that events and banner campaigns do not alter in a significant fashion the evolution of PLT/RSI/TTI metrics, that are intrinsically variable at a daily timescale. Particularly, from Fig. 6, one can notice that the daily fraction of positive, neutral and negative answers remains remarkably steady over the observation period, with a stationary fraction of about 85% satisfied users.

On the one hand, this is somewhat unexpected since one could have argued that events such as, e.g., data center switchover or browser regression, would directly affect the objective measurable delay. At the same time, in light of Fig. 6, it appears that the observed level of variability in the PLT/RSI/TTI metrics happen in a range that is not large enough to affect human perception – or in other words that the measured delay changes do not necessarily harm user QoE.

Seasonality: We next study if user scores follow classic night/day and weekday/weekend effect. The first circadian timescale is intrinsic to variation in human cognitive capability throughout the day, whereas the second can possibly reflect a change in the environment (work/leisure), which not only affects the environment (e.g., user mood) but also possibly the devices used to access the service (e.g., company vs personal). Fig. 7 and Fig. 8 report the raw answer frequency (top plots) as well as the breakdown of users scores (bottom plots) at hour-of-day and day-of-week aggregation granularities respectively.

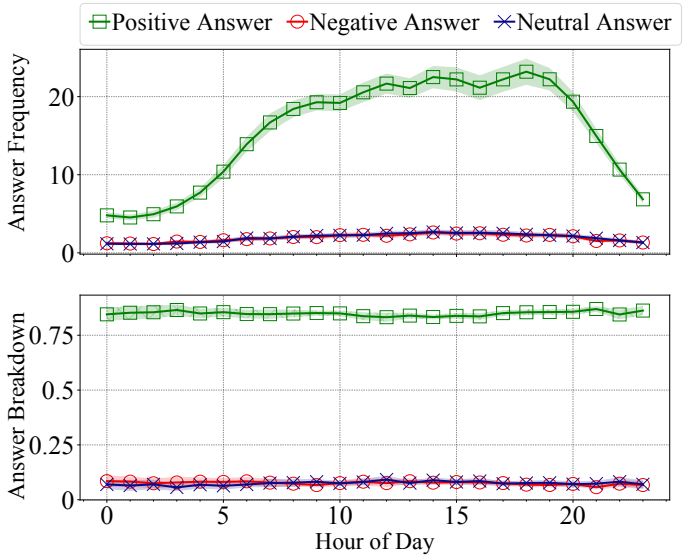


Fig. 7. Temporal view: absence of night/day seasonality of survey answers.

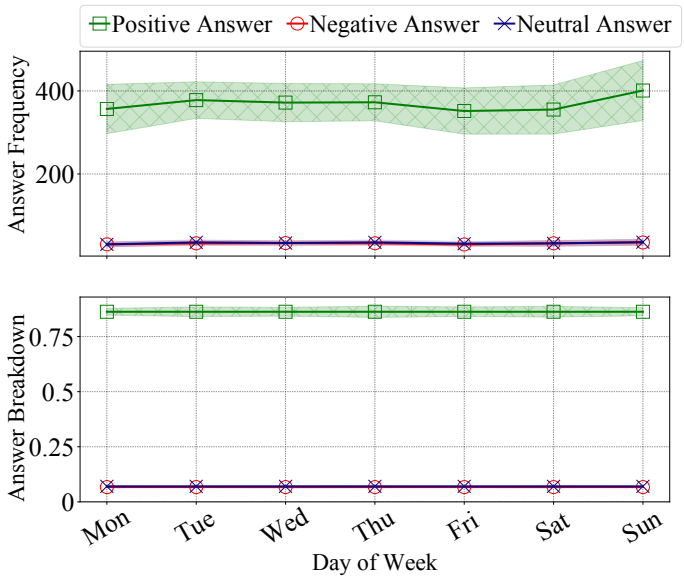


Fig. 8. Temporal view: absence of weekday/weekend seasonality.

Plots report the mean (line) and 95% confidence interval (shadowed band) of the metrics of interest.

Top plots in Fig. 7 and Fig. 8 do exhibit a seasonal variation in the *answers volume*. Particularly, in the hour-of-day case in Fig. 7 the volume is merely correlated with the volume of users activity, which as expected follows a seasonal pattern with lower night-time activity that is preserved by our random sampling. In the day-of-week case one can notice a slight increase in the answer frequency on Sundays, which in our dataset is due to a combination of (i) a slightly higher traffic volume on some Sundays over the 5-months period, (ii) as well as a higher propensity to answer the survey on Sunday, especially during some weeks of September.

Yet, more interesting is the absence of daily/weekly sea-

sonality in the *answers breakdown*. From the bottom plot of Fig. 7 one clearly gathers the absence of seasonality at 24-hours circadian rhythms, which is somewhat surprising. Indeed, recent work [63] that leverages wearable devices to infer user activity and correlate it to Web user responsiveness (i.e., keystroke and click times in the Bing search engine), do show that users have worse responsiveness (i.e., higher keystroke and click delays) especially after wakeup and at night-time, whereas their response times are significant faster during daytime. In turn, from daily variability in user responsiveness, one could have expected a higher tolerance to, e.g., slow websites performance, that however does not appear in our results. One likely reason is that the largest discrepancy between maximal and minimal user click time is on average of about 1 second during the day (see Fig. 2(b) in [63]), which may not be enough to trigger perceptual changes so important to affect the *acceptability* of the page rendering process (whereas they could have appeared had our survey involved a 5-grade ACR scale feedback).

Similarly, from the bottom plot of Fig. 8 one again gathers the absence of seasonality over a weekly timescale. On the one hand, this is somewhat unexpected since human behavior on computer networks (such as personal communication [64]) does exhibit day-of-week dependence. On the other hand, this is in line with [63] that does not remark a weekly difference in user responsiveness (i.e., weekend and weekdays follow a statistically similar diurnal variability in [63]). Under this light, and given the absence of time-of-day dependence on user website acceptability, the absence of day-of-week seasonality is less striking.

Additionally, we gather that, despite the propensity to answer the survey may change over typical human timescales, the answer itself may be more tied to the perceived performance, further confirming the validity of our survey.

C. Spatial breakdown

Overall, our dataset comprises 115 features from 4 main classes. We now investigate how the score breakdown is affected by some representative features in each class. Particularly, since features of the *performance class* are publicly available, and since since their dependency with the user score has already been exposed in Fig. 2 and 3, in this section we further dig into *page, user and environment*-related features. Specifically, whereas lab studies have rather poor diversity in terms of handsets, browser software, and geographical diversity, the collected dataset allows to peek at Web users' QoE under each of these angles.

Fig. 9 reports, for 15 cherry-picked features in the dataset, the breakdown of positive/negative scores (neglecting neutral answers for the sake of simplicity). For each subplot, we condition over different values of the feature and visually report the positive/negative breakdown as stacked bars. For categorical features without a natural ordering, the bars are ordered in increasing satisfaction rates. In case of numerical features, the natural ordering is otherwise preserved (so that breakdown is not monotonously increasing). On each subplot, the top x-axis report the cardinality of samples for each bar, and the bottom label reports the feature name and is further annotated with the mutual information value.

Page-related metrics: Particularly, page related features are filtered so that only the wiki from which the request was issued (*ruwiki* or *frwiki*) is available in *(PA)* and the page size is additionally available in *(WWW)*. The plot in the top left corner of Fig. 9 reports the variation on scores as a function of the HTML page size. It can also be seen that breakdown is very similar irrespectively of the HTML page size, with the exception of smaller pages, that have a slightly higher negative scores (which deserves further attention). Thus, in our dataset the page size only plays a minor role in the user feedback, which can be expected since Wikipedia pages tend to be relatively small. Concerning the smallest bin of pages up to 10KB, notice that it comprises 7.8% of the over 46k pages (i.e., a bag of 3.6k pages) confirming that a 10KB granularity makes linkability of the *(WWW)* dataset complex.

User-related metrics: Among user-related metrics, which are not available in *(PA)*, we select the browser, device and OS *families* (finer grain information is precluded from sharing), and report whether users are logged in Wikipedia (binary flag), if they are accessing Wikipedia through a tablet device (binary flag) and the number of edits that users have made (coarse bins). These features are reported in the top row (and the first two features in the second row) of Fig. 9.

For the family of browsers, device and OS, we report the most popular and aggregate all others into a "other" bin. Interestingly, from the browser family one can notice a remarkable discrepancy of users score breakdown for different browsers. Particularly, one can observe "mobile" versions of popular browsers to have poorer scores than their "laptop/desktop" counterpart: in this case, one cannot easily disambiguate whether poor scores are tied to bad implementation of the browser, or to bad performance of the mobile device (a nevertheless very likely cause [5], [6]). Considering only laptop/desktop browsers, we have that Safari (1st), Opera (2nd) and Chrome (3rd) are on the podium, with Firefox (4th) a close next.

It is also interesting to observe that, whereas users scores quite clearly differ among browsers, the amount of mutual information is still relatively low (comparable to the HTML page size) – which is due to the fact that browsers are not equally represented in the dataset, with Chrome and Chrome mobile taking up over 50% of the samples in our dataset. Similarly, score breakdown is remarkably different across devices, yet the number of devices is so large (over 2.7k) and the categories either too precise (as for the different XiaoMi models) or too coarse (iPhone and iPad do not unfortunately report the model version, which mixes old and new devices in a single bin) resulting in a very low mutual information.

Score breakdown per OS confirms that users score are better on laptop/desktop. However class imbalance across OSs makes it so that a simple binary indicator (isTablet) has a higher predictive power with respect to more precise labels (e.g., twice as much as the OS and browsers family).

Next, concerning user experience on Wikipedia, we notice that readers (0 edit) are more likely to provide a negative

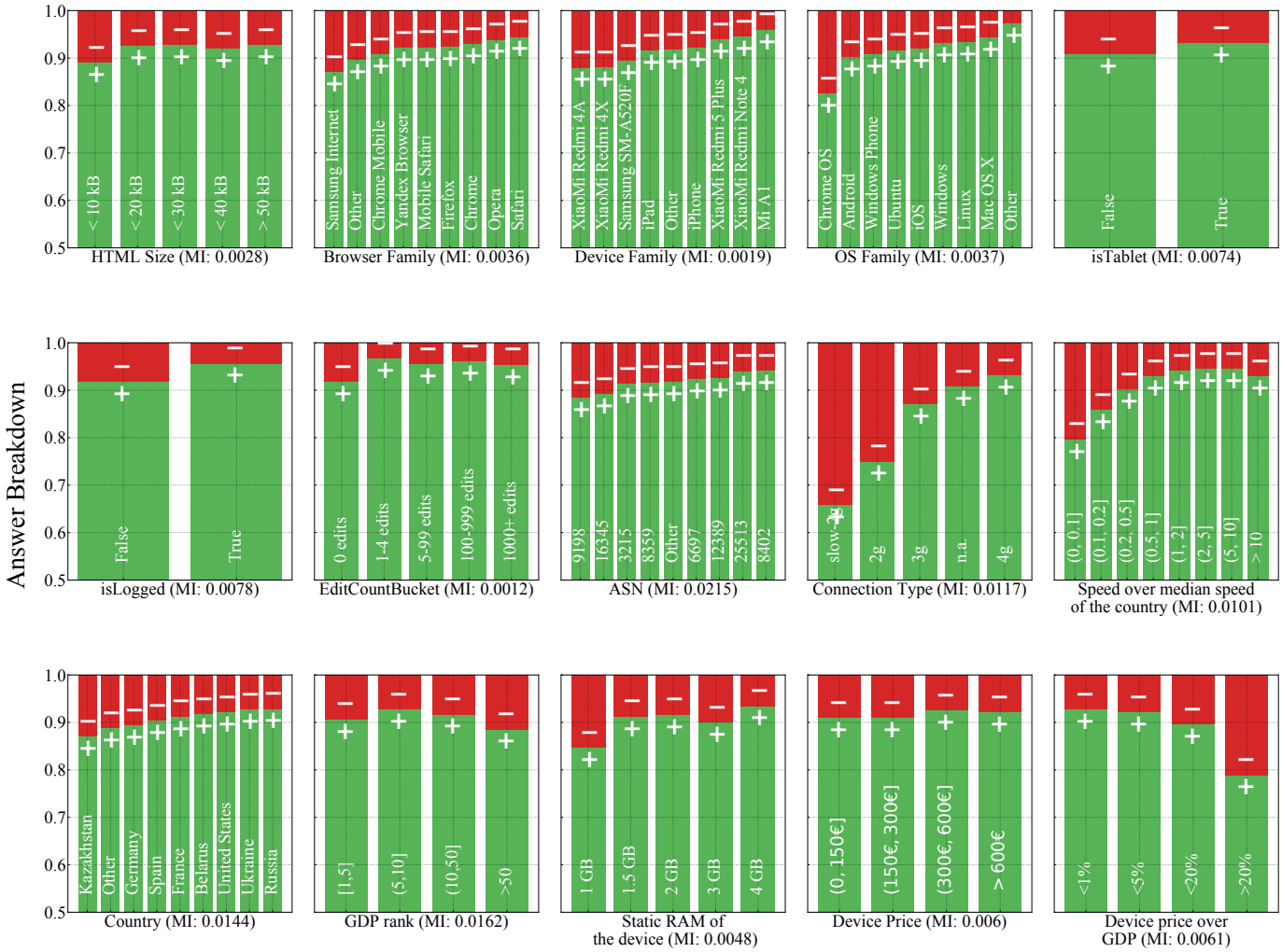


Fig. 9. Illustration of spatial breakdown of user scores across *page*, *user* and *environment* features obtained by conditioning each of them over different values and showing on the top x-axis the cardinality of samples for each bar.

answer than writers (from 1 to over 1000 edits). This is somewhat surprising since whereas our survey population is mostly European, logged editors are always directed to US servers, incurring in higher latency. The higher fraction of positive answers can be due, on the one hand to the fact that higher RTT delay may be masked from warm-up caches for the page they are editing, or on being more accustomed with (and thus more adapted to) Wikipedia service. At the same time, given that most (97%) of Wikipedia users are readers, the knowledge of the edit counts is irrelevant for predicting user satisfaction – so that even in this case a simple binary information such as whether the user is logged in has more predictive power (high *MI*).

Environment-related metrics: Features in the environment class include network-related and per-country information, reported in the middle and bottom column of Fig. 9 respectively. Network information is represented by ASN, connection type and speed information (particularly, we report in the picture the ratio of the download speed to the median speed in the country observed in our dataset). We see that all have a clear impact on the user scores, with consistent differences across ASN, very strong differences across connection type (although there are only very few 2G and 3G connections in our dataset, thus a low MI) and strong difference on the relative connection speed. Interestingly, concerning the latter one can notice that the ratio of negative scores decreases for increasing speed, and finally exhibits a slight decrease again for users having $10\times$ the median speed in the country – likely well equipped and possibly more impatient users.

In terms of country-level information, bottom-row plots in Fig. 9 inspect the country name and its GDP rank. Two phenomena appear: on the one hand, we observe that users living in countries with poor GDP (high rank) consistently report poor performance (likely tied to poorer infrastructures); on the other hand, we observe that users of wealthy countries, that have comparably better performance (e.g., higher rates), also possibly report negative scores, but possibly due to different reasons (e.g., tied to higher user impatience due to higher expectations).

We finally consider further information concerning the user device (such as RAM and price harvested from the

Web), which we report to the median per-country per-capita GDP. We gather that, whereas poor maximum RAM (1GB) is symptomatic of bad performance, scores are strikingly similar across a range of device prices: as performances are likely different across devices [6] (which in part justifies the price difference), this seems to suggest that owners of cheap devices are prepared to be more tolerable in spite of poorer performance. However, if we do take into account the relative wealth of the country by normalizing the price tag over the per-capita GDP, we see that there is a negative correlation with user scores (possibly, expectations of users owning a pricey device in a lower-GDP country are also higher, and users are more likely to report bad performance as negative experience).

V. User feedback prediction

We continue by disregarding the neutral scores and build data-driven models that forecast user answers. We point out that feedback prediction was the main focus in the conference version of this paper [17]. We recall that the exact composition of the feature set *(PA)* that was finally agreed for public release by the Wikimedia legal team has been finalized after the publication of the preliminary work [17]. It is important to assess that the results achieved with the datasets used in [17] also hold for the publicly released dataset [23]. Due to space constraints, we limitedly focus on this consistency aspect in this section. However, for the sake of completeness we point out that a significantly more developed section (including a wider range of classification techniques, more extensive feature sampling, outlier filtering and dataset conditioning experiments) is available in an extended technical report [46] for the interested reader.

A. Problem formulation

Keeping only negative and positive answers for the user feedback prediction analysis is a simplification which directly stems from the structure of our survey, and allows to turn the problem into a binary classification one. This simple formulation enables immediate and intuitive statements of performance objective, that we express in terms of the classic information retrieval metrics.

Clearly, from an operational standpoint a *conservative* estimation of user satisfaction is preferable. Indeed, the service operator wants to avoid that a malfunctioning service that is truly affecting user experience goes undetected, as when the ratio of dissatisfied users increases above a given level this can prompt alert to repair or ameliorate the service. In our settings, conservative prediction results translate into *maximizing the recall of negative scores*.

B. Classification results

Given the class imbalance, we have to preliminarily downsample the dataset[3]: indeed, given that after discarding the neutral scores 92% of the users are satisfied, a naïve 0-R classifier that just learns the relative frequency of the scores and systematically answers with the majority class, would achieve 0.92 accuracy – but would entirely miss negative scores, having thus a null $\mathcal{A}^-$ recall. Hence, a more appropriate baseline for recall of unsatisfied users requires performing a stratified undersampling, i.e., keep only a portion of the positive scores, equal to the size of the negative ones, to obtain a balanced dataset. We denote by $\mathcal{B}$ the balanced dataset and by $\overline{\mathcal{B}}$ the complementary dataset, only containing positive answers filtered out in the downsampling. For the sake of brevity, we limitedly report results from 10-fold cross validation of 20-trees random forest [65], as similar results hold for different classification models [46]. Tab. IV reports three confusion matrixes, each one highlighting the average accuracy, precision and recall of the unsatisfied users $\mathcal{A}^-$. Notice that the cardinality of $\mathcal{B}$ in the first two cases is the same and equal to 9388, whilst for the *(PA)* set is lower, given that only the French and Russian wikis are included, and equal to 8988. To verify the consistency (or discrepancies) of the previously published results [17] with respect to the dataset finally made available, we contrast results gathered over the full set comprising all the 115 collected features *(T)* vs the 61 features of the *(WWW)* set [17] and the 19 publicly available (PA) feature set [23].

TABLE IV
User feedback prediction: Confusion matrixes for Random forest classifiers, 10-fold cross validation. Comparison of all collected features (T) vs (WWW) vs (PA).

Set (card.)	True	Predicted -	Predicted +	All	
$\lvert T\rvert =$ 115	-	0.59	0.40	4694	Accuracy ($\mathcal{B}$) = 0.59
	+	0.41	0.60	4694	Accuracy ($\overline{\mathcal{B}}$) = 0.55
					$\mathcal{A}^-$ Precision = 0.59
					$\mathcal{A}^-$ Recall = 0.62

Set (card.)	True	Predicted -	Predicted +	All	
$\lvert WWW\rvert =$ 61	-	0.58	0.41	4694	Accuracy($\mathcal{B}$) = 0.58
	+	0.42	0.59	4694	Accuracy ($\overline{\mathcal{B}}$) = 0.55
					$\mathcal{A}^-$ Precision = 0.58
					$\mathcal{A}^-$ Recall = 0.61

Set (card.)	True	Predicted -	Predicted +	All	
$\lvert PA\rvert =$ 19	-	0.58	0.41	4494	Accuracy($\mathcal{B}$) = 0.59
	+	0.42	0.59	4494	Accuracy ($\overline{\mathcal{B}}$) = 0.55
					$\mathcal{A}^-$ Precision = 0.58
					$\mathcal{A}^-$ Recall = 0.61

We obtain very similar results on all datasets, with marginal accuracy degradation for different feature sets. On the one hand, recalling the mutual information statistics presented early in Table III, this is not surprising as features in the *(PA)* set are among those having the highest amount of mutual information with the class label (i.e., the user answer). On the other hand, these results are clearly deceiving and only slightly better than the naïve baseline. This holds despite the relatively large number of features collected: for features in the T set, only 0.62 of the unsatisfied users are correctly captured (0.61 in WWW and PA), with a precision of 0.59 (0.58 in WWW and PA). Interestingly, performance on the complement $\overline{\mathcal{B}}$, i.e., the set of positive scores filtered out due to class imbalance, remains consistent with an average accuracy of 0.55. In the extended technical report [46], we employ also other models for the prediction of user feedback, namely Multi Layer Perceptron, XGBoost, *K* Nearest Neighbor and Support-

[3]We prefer to avoid the opposite approach of *synthetically* generating users score, which is in stark contrast with the very same nature of our survey work.

Vector Machines classifiers. We show that, on average, the accuracy remains practically unchanged and highlight that Random Forest performs better with respect to the other algorithms in terms of $\mathcal{A}^-$ recall (0.61) and F1 score (0.60).

While *qualitatively* deceiving for what their accuracy is concerned, we believe the implication of these negative results is worth sharing with the community – especially in light on how to ameliorate the state of the art in Web QoE estimation, which we discuss next.

VI. Discussion

This work is the first to leverage user feedback from real browsing sessions in operational settings. As any new work, there are a number of limits, which requires community-level-efforts, that we discuss here.

Collection and validation methodologies: We remark that this work is the first to collect user feedback from real users in real browsing activity, from an operational deployment. This is in stark contrast with most lab research, where volunteers or crowdworkers are exposed to a very limited heterogeneity (e.g., single device/browser), are not carrying on a browsing activity (e.g., A/B testing uses videos) and are not asked about their satisfaction but about other metrics as a proxy (e.g., which video finished first?). We argue that lab/crowdsourcing experiments and collection in the wild should *coexist*.

On the one hand, we stress that while A/B testing is a necessary step, it is however not sufficient. Survey data in this paper seems to suggest that metrics that are considered as state-of-the art for Web QoE, seems to be ultimately poorly correlated with the experience of real Wikipedia users. In turn, it also follows that lab/crowdsourcing experiments should diversify the type of user feedback: e.g., the fact that a user is able to notice which video finishes first (which uPLT metrics attempt to model), does not imply that he would grade that Web rendering process as positive (or the rendering corresponding to the other video as negative).

On the other hand, we are aware that part of the challenges in real-world experiments comes from diversity and variance: it follows that surveys such as those we are carrying on should be kept *running continuously*, as it is commonplace for VoIP applications that regularly poll their users for a QoE opinion. Operating continuously would lower barriers for further experiments [66], empower website operators with a very relevant performance indicator for their service, informing them in near-real time about impact of new features deployment. Additionally, long-time surveys allow to collect significant volumes of data to keep ameliorating models for user prediction in spite of high variance and heterogeneity. Moreover, there exist other QoE influence factors that we did not include in this study, like the sentiment linked to the topic and the content of the page or more information about the context in which the measurement is carried out, as the earlier user browsing experience. These undoubtedly have an important impact, that is however hard to capture.

RSI: not needed, or not enough?: Concerning Web user QoE metrics, this study seems to suggest a poor discriminative power of the RUM SpeedIndex (RSI) so as to predict users scores, at least for Wikipedia users. In part, this may be due to the structure of Wikipedia pages (where, e.g., text may be more prevalent that in other pages in the Alexa top-100 typically considered in similar studies, see Section II). This nevertheless raises the question so as to whether it is possible to (i) design metrics that are better fit to the spatial structure of the page, or (ii) metrics capable of better weighting the focus of user attention, and at the same time (iii) raises questions about the accuracy vs generality of QoE metrics.

As for (i), we are currently improving the system to also collect navigation timing statistics for specific elements that are believed to be important for Wikipedia, such as the "time to the top image". This is a good compromise between collecting the whole waterfall (which is impossible in operational settings) and could yield to metrics that are website-specific (losing generality), but better correlated with user experience (gaining discriminative power).

As for (ii), we are aware that more complex approaches involving spatial dimension (i.e., eye gaze) also exist [13], [12]. However, including the spatial dimension in the user perception is hard to capture in the lab, and challenging in the wild: a good starting point would be to leverage mouse-movements as a proxy of eye gaze activity (which are known to be strongly correlated [67]), and that can help further refining QoE metric in the spatial direction (e.g., by adding the knowledge of whether the rendered element is under the user gaze). Additionally, mouse-movements can capture user anxiety which further reduces the user viewport [68]. Clearly, further research is needed on whether user-touch can be useful for similar purposes in case of mobile handsets.

Finally, (iii) previous work [20] already has pointed out a tension between accuracy vs generality of QoE metrics and models: on the one hand, it seems rather challenging to capture the rich diversity of over one billion pages with a single QoE model, so that it may be tempting to develop website-specific models, as it is our focus here; on the other hand, it may be possible to develop models for groups of websites sharing similarities in their underlying structure (e.g., picture-dominant vs text-dominant sites; interactive vs static pages; etc.), which remains an open question to date.

Per-server vs per-device statistics: In this work, we did not explicitly leverage time-series of server-related operational metrics, as these are gathered live at minute-timescale on Prometheus [69] but are not readily available on the Hive platform [70]. At the same time, the raw load on during the considered period appears too low in practice to have an impact so significant to affect user satisfaction.

Conversely, given that mobile browsers performances are significantly dependent on the handsets, as already shown in [6], [5] and confirmed in this work, collecting per-device statistics seems a mandatory step to ameliorate prediction performance, as "computation activities are the main bottleneck when loading a page on mobile browsers" [5]. Unfortunately, average per-device performance we considered in this work are not telling enough, as they merely report the resource *upper-*

bound (i.e., CPU and RAM capacity) as opposite to the *actual state* of the device (i.e., free RAM and available CPU cycles) corresponding to the page view that the user answered about – which could hopefully ameliorate prediction performance.

VII. Conclusions

In this paper we engineer, collect, analyze and predict user survey scores pertaining to the quality of their Web browsing experience. Out of over 1.7 million queries, we gather over 62*k* answers corresponding to either positive (84.8%), neutral (7.7%) or negative (7.5%) experiences. Associated to each answer, we collect 115 features, part of which we make publicly available taking care of rendering user deanonymization and content-linkability as hard as possible.

The main takeaways in our analysis are that users are consistently satisfied, and that scores do not exhibit seasonality at circadian or weekly timescales, which is unexpected. Quite surprisingly, scores are also not affected by network-related events (e.g. data center switchover) happening during the period, nor by Wikipedia-related events (e.g., banner campaigns that alter the page rendering) nor by known browsers events (e.g., Chrome 69 first paint regression). Additionally, we find that scores are, as expected, heavily influenced by user-level expertise and equipment (e.g., device, OS and browser), as well as network and country-level characteristics (including access technologies, ISP and economical factors). Interestingly, scores are not affected by the Wikipedia page size, nor by the device price (unless economical factors are also weighted in).

Concerning user score prediction, perhaps the most important (and equally disturbing) takeaway is that it is surprisingly hard to predict even a very coarse-grained indication of user satisfaction. This can be tied in part to the lack of more informative indicators in our dataset (such as *content* and *context* factors that are known to affect user QoE), and also raises a number of interesting questions and challenges for the whole community.

Acknowledgments

We thank Tilman Bayer, Leila Zia and Nuria Ruiz from the Wikimedia Foundation and Saint Johann from the Russian Wikipedia community, whose input and help shaped this work. We also thank the Associate Editor and anonymous Reviewers for their valuable comments and suggestions.

References

[1] "World wide wait," [Online]. Available from: https://www.economist.com/science-and-technology/2010/02/12/world-wide-wait, Feb. 2010.

[2] V. Cerf, V. Jacobson, N. Weaver, and J. Gettys, "Bufferbloat: what's wrong with the internet?" *Communications of the ACM*, vol. 55, no. 2, pp. 40–47, 2012.

[3] J. Erman, V. Gopalakrishnan, R. Jana, and K. K. Ramakrishnan, "Towards a spdy'ier mobile web?" in *Proc. ACM CoNEXT*, 2013.

[4] X. S. Wang, A. Balasubramanian, A. Krishnamurthy, and D. Wetherall, "How speedy is spdy?" in *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*, 2014.

[5] J. Nejati and A. Balasubramanian, "An in-depth study of mobile browser performance," in *Proc. The World Wide Web Conference (WWW)*, 2016.

[6] M. Dasari, S. Vargas, A. Bhattacharya, A. Balasubramanian, S. R. Das, and M. Ferdman, "Impact of device performance on mobile internet qoe," in *Proc. ACM Internet Measurement Conference*, 2018.

[7] S. Rosen, B. Han, S. Hao, Z. M. Mao, and F. Qian, "Push or request: An investigation of http/2 server push for improving mobile performance," in *Proc. The World Wide Web Conference (WWW)*, 2017.

[8] X. S. Wang, A. Krishnamurthy, and D. Wetherall, "Speeding up web page loads with shandian," in *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*, 2016.

[9] Y. Ma, X. Liu, S. Zhang, R. Xiang, Y. Liu, and T. Xie, "Measurement and analysis of mobile web cache performance," in *Proc. The World Wide Web Conference (WWW)*, 2015.

[10] T. Zimmermann, J. Ruth, B. Wolters, and O. Hohlfeld, "How http/2 pushes the web: An empirical study of http/2 server push," in *Proc. IFIP Networking*, 2017.

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

[12] C. Kelton, J. Ryoo, A. Balasubramanian, and S. R. Das, "Improving user perceived page load time using gaze," in *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*, 2017.

[13] 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. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*, 2015.

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

[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. ACM SIGCOMM, Internet-QoE Workshop*, 2017.

[16] [Online]. Available from: https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index.

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

[18] 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.

[19] T. Zimmermann, B. Wolters, and O. Hohlfeld, "A qoe perspective on http/2 server push," in *Proc. ACM SIGCOMM, Internet-QoE Workshop*, 2017.

[20] D. Da Hora, A. 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.

[21] X. Zhang, S. Sen, D. Kurniawan, H. Gunawi, and J. Jiang, "E2e: Embracing user heterogeneity to improve quality of experience on the web," in *Proceedings of the ACM Special Interest Group on Data Communication*. ACM, 2019, pp. 289–302.

[22] J. Ruth, K. Wolsing, K. Wehrle, and O. Hohlfeld, "Perceiving quic: Do users notice or even care?" in *Proc. ACM Internet Measurement Conference*, 2019.

[23] [Online]. Available from: https://webqoe.telecom-paristech.fr/data/.

[24] J. Nielsen, "Response times: The 3 important limits," [Online]. Available from: https://www.nngroup.com/articles/response-times-3-important-limits/.

[25] R. B. Miller, "Response time in man-computer conversational transactions," in *Proc. AFIPS Fall Joint Computer Conference*. ACM, 1968.

[26] ITU-T, "Estimating end-to-end performance in ip networks for data application." Feb. 2014.

[27] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The logarithmic nature of qoe and the role of the weber-fechner law in qoe assessment," in *Proc. IEEE International Conference on Communications (ICC)*, 2010.

[28] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[29] ITU-T, "Qoe factors in web-browsing," Feb. 2014.

[30] S. Egger, 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.

[31] G. Tangari, D. Perino, A. Finamore, M. Charalambides, and G. Pavlou, "Tackling mobile traffic critical path analysis with passive and active measurements," in *Proc. Traffic Monitoring and Analysis Workshop (TMA)*, June 2019, pp. 105–112.

[32] A. S. Asrese, S. J. Eravuchira, V. Bajpai, P. Sarolahti, and J. Ott, "Measuring web latency and rendering performance: Method, tools, and longitudinal dataset," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 535–549, June 2019.

[33] A. Asrese, E. Walelgne, V. Bajpai, A. Lutu, O. Alay, and J. Ott, "Measuring web quality of experience in cellular networks," in *Proc. Passive and Active Measurement Conference (PAM)*, 2019.

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

[35] F. F. Nah, "A study on tolerable waiting time: how long are web users willing to wait?" *Behaviour & Information Technology*, vol. 23, no. 3, pp. 153–163, 2004.

[36] T. Enghardt, T. Zinner, and A. Feldmann, "Web performance pitfalls: Methods and protocols," in *Proc. Passive and Active Measurement Conference (PAM)*, 2019.

[37] A. Balachandran, V. Aggarwal, Shobha, H. Yan *et al.*, "Modeling web quality-of-experience on cellular networks," in *Proc. ACM MOBICOM*. ACM, 2014.

[38] B. Miroglio, D. Zeber, J. Kaye, and R. Weiss, "The effect of ad blocking on user engagement with the web," in *Proc. The World Wide Web Conference (WWW)*, 2018.

[39] 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.

[40] 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.

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

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

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

[44] R. Netravali, A. Sivaraman, K. Winstein, S. Das, A. Goyal, and H. Balakrishnan, "Mahimahi: a lightweight toolkit for reproducible web measurement," in *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4. ACM, 2014, pp. 129–130.

[45] Y. Moshfeghi and J. M. Jose, "On cognition, emotion, and interaction aspects of search tasks with different search intentions," in *Proc. The World Wide Web Conference (WWW)*, 2013.

[46] F. Salutari, D. D. Hora, G. Dubuc, and D. Rossi, "Analyzing wikipedia users perceived quality of experience: A large-scale study (extended version)," Extended Techical Report available from: https://webqoe.telecom-paristech.fr/papers/, Dec. 2019.

[47] P. A. Frangoudis, L. Yala, and A. Ksentini, "Cdn-as-a-service provision over a telecom operatorâs cloud," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 702–716, Sep. 2017.

[48] E. Aguiar, A. Riker, A. Abelem, E. Cerqueira, and M. Mu, "Video quality estimator for wireless mesh networks," in *2012 IEEE 20th International Workshop on Quality of Service*, 2012.

[49] P. Casas, A. D'Alconzo, P. Fiadino, A. Bar, A. Finamore, and T. Zseby, "When youtube does not workâanalysis of qoe-relevant degradation in google cdn traffic," *IEEE Transactions on Network and Service Management*, vol. 11, no. 4, pp. 441–457, Dec. 2014.

[50] P. Casas, M. Seufert, F. Wamser, B. Gardlo, A. Sackl, and R. Schatz, "Next to you: Monitoring quality of experience in cellular networks from the end-devices," *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, pp. 181–196, June 2016.

[51] M. Rajiullah, A. Lutu, A. S. Khatouni, M.-R. Fida, M. Mellia, A. Brunstrom, O. Alay, S. Alfredsson, and V. Mancuso, "Web experience in mobile networks: Lessons from two million page visits," in *Proc. The World Wide Web Conference (WWW)*, 2019.

[52] A. Saverimoutou, B. Mathieu, and S. Vaton, "Web view: Measuring monitoring representative information on websites," in *2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, 2019.

[53] P. Singer, F. Lemmerich, R. West, L. Zia, E. Wulczyn, M. Strohmaier, and J. Leskovec, "Why we read wikipedia," in *Proc. The World Wide Web Conference (WWW)*, 2017.

[54] [Online]. Available from: https://www.alexa.com/topsites.

[55] 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*, vol. 1, no. 1, p. 2, 2016.

[56] "Maxmind," [Online]. Available from: https://www.maxmind.com/.

[57] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "IP geolocation databases: Unreliable?" *ACM SIGCOMM CCR*, vol. 41, no. 2, 2011.

[58] C. Smith-Clarke and L. Capra, "Beyond the baseline: Establishing the value in mobile phone based poverty estimates," in *Proc. The World Wide Web Conference (WWW)*, 2016.

[59] "World bank," [Online]. Available from: https://data.worldbank.org/.

[60] "Gdp per capita by country," [Online]. Available from: https://github.com/secure411dotorg/GDP-per-Capita-by-Country.

[61] S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, and A. Pescape, "Broadband internet performance: A view from the gateway," in *Proc. ACM SIGCOMM*, 2011.

[62] "Gsm arena," [Online]. Available from: https://www.gsmarena.com.

[63] T. Althoff, E. Horvitz, R. W. White, and J. Zeitzer, "Harnessing the web for population-scale physiological sensing: A case study of sleep and performance," in *Proc. The World Wide Web Conference (WWW)*, 2017.

[64] F. Kooti, L. M. Aiello, M. Grbovic, K. Lerman, and A. Mantrach, "Evolution of conversations in the age of email overload," in *Proc. The World Wide Web Conference (WWW)*, 2015.

[65] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[66] E. Bakshy, D. Eckles, and M. S. Bernstein, "Designing and deploying online field experiments," in *Proc. The World Wide Web Conference (WWW)*, 2014.

[67] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola, "Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts," in *Proc. The World Wide Web Conference (WWW)*, 2013.

[68] B. Youngmann and E. Yom-Tov, "Anxiety and information seeking: Evidence from large-scale mouse tracking," in *Proc. The World Wide Web Conference (WWW)*, 2018.

[69] "Prometheus," [Online]. Available from: https://prometheus.io.

[70] "Apache hive," [Online]. Available from: https://hive.apache.org/.

Flavia Salutari is a PhD candidate at Telecom Paris. She received her MSc degree in Information and Communications Technologies for Smart Societies from Politecnico di Torino, Italy in 2017.

Diego Da Hora is a Software Engineer at Google in Brazil. He received a MSc in Computer Science from Federal University of Minas Gerais in 2012, and a PhD in Computer Science from Universite Pierre et Marie Curie in 2017.

Gilles Dubuc has been a member of the Performance Team at the Wikimedia Foundation since its inception in 2015, where he currently focuses on bridging the gap between performance metrics and real user performance perception.

Dario Rossi is a Chief Expert at Huawei. He has been a Full Professor at Telecom Paris and Ecole Polytechnique and holder of Cisco's Chair NewNet Paris. He has coauthored 9 patents and over 150 papers in leading conferences and journals, that received 8 best paper awards, a Google Faculty Research Award (2015) and an IRTF Applied Network Research Prize (2016). He is a Senior Member of IEEE and ACM.