How Good is your Mobile (Web) Surfing? Speed Index Inference from Encrypted Traffic

Sarah Wassermann, Pedro Casas, Michael Seufert, Nikolas Wehner, Joshua Schuler, Tobias Hossfeld

How Good is your Mobile (Web) Surfing? Speed Index Inference from Encrypted Traffic

S. Wassermann†, P. Casas†, M. Seufert∗, N. Wehner∗, J. Schüler∗, T. Hossfeld∗
† AIT Austrian Institute of Technology, ∗ University of Würzburg

ABSTRACT

We address the problem of Web QoE monitoring, in particular Speed Index (SI), from the Internet Service Provider (ISP) perspective, relying on in-network, passive measurements. Given the wide adoption of end-to-end encryption, we resort to machine-learning models to infer the SI of individual web-page loading sessions, using as input only packet-level data. Our study targets the analysis of SI in mobile devices, including smartphones and tablets. To the best of our knowledge, this is the first paper addressing the inference of SI from encrypted network traffic in mobile devices.

CCS CONCEPTS

• **Networks → Network performance evaluation**; **Network measurement**; • **Computing methodologies → Machine learning**;

ACM Reference Format:
S. Wassermann†, P. Casas†, M. Seufert∗, N. Wehner∗, J. Schüler∗, T. Hossfeld∗. 2020. How Good is your Mobile (Web) Surfing? Speed Index Inference from Encrypted Traffic. In *ACM Special Interest Group on Data Communication (SIGCOMM '20 Demos and Posters), August 10–14, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3405837.3411382

1 INTRODUCTION

Web browsing is a paramount Internet service for the end user. The performance of a web service as perceived by the end user can be measured by the corresponding web-browsing Quality of Experience, or Web QoE. From a practical perspective, reliably measuring Web QoE is challenging, especially when the interested party has no access to the application, such as the ISP. The literature on Web-QoE analysis proposes a wide range of objective metrics capturing the performance of web pages, including metrics such as Page Load Time (PLT), Speed Index (SI), Above the Fold Time (AFT), etc. However, these metrics require access to the application layer, which is hidden from the eyes of the ISP by the wide deployment of end-to-end traffic encryption. Therefore, we explore the potential of using machine learning (ML) to infer Web-QoE metrics from encrypted network traffic, for the specific scenario of mobile web browsing. By using controlled page-load experiments, where network data is simultaneously collected with ground-truth Web-QoE metrics such as SI, we build a labeled dataset and train supervised ML models to infer these QoE-related metrics based on network-traffic features, computed from the stream of collected bytes.

SIGCOMM '20 Demos and Posters, August 10–14, 2020, Virtual Event, USA

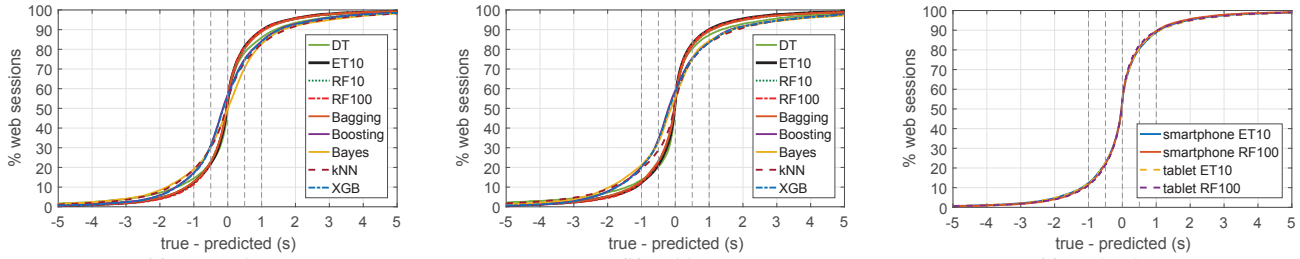
ACM ISBN 978-1-4503-8048-5/20/08…$15.00
https://doi.org/10.1145/3405837.3411382

2 SPEED INDEX INFERENCE

The proposed solution to the mobile-Web-QoE-monitoring problem consists of training supervised machine-learning models to map network-traffic features, extracted from the encrypted network web-page traffic, into relevant Web-QoE metrics. The approach is data-driven, and thus needs datasets containing both the collected traffic traces – the *input* –, and the targeted Web-QoE metric – the *ground truth*. To fully control the generation of such datasets, we conceived a measurement testbed based on multiple private instances of WebPageTest (WPT), a well-known and widely used open-source web-performance analysis tool. Different from previous studies [1–6], which have focused exclusively on desktop browsers and desktop devices (or in some exceptional cases, browser-emulated mobile devices), our measurement testbed consists of three different, non-emulated types of devices, including smartphones, tablets, and desktop (Chrome is used as browser), using WPT agents for Android and Linux. Instead of leveraging in-device WPT traffic-shaping capabilities, devices are connected to the open Internet through independent network emulators, which allows for more realistic network-access performance configurations in terms of bandwidth, latency, packet loss, etc., and avoids further loading the CPU of the devices. This allows for heterogeneity in the generated measurements. Configurations used in the study include access-downlink bandwidth up to 10 Mbps, packet loss rates up to 10%, and RTTs up to 100 ms.

We generated a fully balanced dataset of more than 30.000 web page *loading sessions* (i.e., the loading of a single page), targeting the top 500 websites according to the Alexa top-sites list, from a single vantage point in Europe. The same web pages are visited multiple times for each device type, under the same access-network setups. Without loss of generality, we focus on the inference of one particular Web-QoE metric, the SI, which is today one of the most accepted metrics reflecting Web QoE. In particular, we collect the so-called RUM Speed Index (RUMSI) metric, which is a passive approximation to the SI, computed from the analysis of web-page resource timings.

We treat the inference of the RUMSI metric as a regression problem. To define input features, we follow the rationale behind the computation of the SI metric itself, which considers the whole progress of the page loading. We define the Cumulative-Bytes-Downloaded features $CBD(i)_{\Delta T}$ as the (normalized) cumulative number of bytes downloaded from the first collected byte at time t_0 (time to first byte, TTFB) up to time $t = t_0 + i \times \Delta T$, with $i = 1, \ldots, m$. The *CBD* features track the download progress of the page bytes, using a time resolution ΔT. We take $m = 100$ samples, and consider



(a) Smartphone. (b) Tablet. (c) Multi-device.

Figure 1: RUMSI inference performance, using per device models (a, b) and multi-device models (c).

model	dev	MAE-mAE (ms)	MRE-mRE (%)	PLCC
DT	S	1021 – 372	34 – 14	0.770
	T	1082 – 298	31 – 11	0.731
ET10	S	**788 – 354**	**28 – 13**	**0.859**
	T	**804 – 314**	**25 – 11**	**0.867**
RF10	S	813 – 383	29 – 14	0.856
	T	867 – 357	27 – 12	0.852
RF100	S	**764 – 360**	**27 – 13**	**0.876**
	T	**815 – 334**	**26 – 12**	**0.866**
Bagging	S	820 – 380	29 – 14	0.855
	T	874 – 362	27 – 13	0.853
Boosting	S	1067 – 598	42 – 23	0.834
	T	1206 – 642	43 – 24	0.813
XGB	S	1068 – 601	42 – 23	0.831
	T	1207 – 652	43 – 24	0.811

Table 1: RUMSI-inference performance using ML models for (S)martphone and (T)ablet data.

three different temporal resolutions to compute features, using ΔT = 50 ms, 100 ms, and 500 ms, for a total of 300 *CBD* features. Using different resolutions helps in capturing different phenomena in the traffic-downloading progress, and allows to track different page-load durations, in this case up to 5, 10, and 50 seconds, respectively. We consider 11 additional input features, related to the complete page-loading session; these include: full session duration (first to last packet), downlink/uplink session duration (first to last packet in downlink/uplink direction), total number of packets downlink/uplink/full, total bytes downlink/uplink/full, and session mean throughput downlink/uplink.

3 MOBILE RUMSI INFERENCE

Using the generated dataset and the described network-traffic features, we train multiple regression models to infer the RUMSI metric. We first consider per-device models, and then evaluate a multi-device model, trained on both smartphone and tablet measurements. Table 1 reports the RUMSI inference performance attained by seven different ML models, most of them based on decision trees, for smartphone and tablet devices. These models include single decision tree (DT), multiple types of ensembles using different numbers of trees, such as extremely randomized trees (ET), random forests (RF), bagging trees, and boosting – including XGB optimizations. We assess their performance using 10-fold cross validation and three standard performance metrics for regression problems, including the absolute error (AE), the relative error (RE), and the linear correlation (PLCC). We take both mean (M) and median (m) values for the error metrics, to filter out significantly large errors. Figures 1(a), 1(b) additionally depict the distribution of the inference errors.

RF100 achieves the best inference performance for both smartphone and tablet, with a median absolute error of 360 ms on smartphone and 334 ms on tablet, and a median relative error around 13% for both devices. Absolute inference errors are below 500 ms for more than 60% of the sessions, and more than 80% of the session RUMSI values are inferred with an error below 1 second. Similar performance is realized by smaller ensembles – e.g., RF10, ET10, and bagging, using 10 instead of 100 trees. Given the training-speed improvements attained by the ET10 model, we take it as the underlying model in subsequent evaluations.

Figure 1(c) reports the inference performance achieved by multi-device (MD) models, split by device-type, using both ET10 and RF100 as underlying ML approaches. A single MD model is trained on data from both smartphone and tablet devices. Results for MD models are almost identical to those attained by per-device models, with a slight degradation for smartphone and a slight improvement for tablet. This suggests that proper inference generalization can be achieved by considering device heterogeneity in the training step.

We have extended the analysis to cover mobile applications, both relying on WebView-based apps and on native apps, obtaining similar performance.

REFERENCES

[1] A. Huet, et al., "Revealing QoE of Web Users from Encrypted Network Traffic," in *IFIP Networking*, 2020.
[2] A. Huet, et al., "Web Quality of Experience from Encrypted Packets," in *ACM SIGCOMM Posters and Demos*, 2019.
[3] A. Saverimoutou et al., "A 6-month Analysis of Factors Impacting Web Browsing Quality for QoE Prediction," *Computer Networks*, vol. 164, 2019.
[4] A. S. Asrese et al., "Measuring Web Latency and Rendering Performance: Method, Tools & Longitudinal Dataset," *IEEE Transactions on Network and Service Management*, 2019.
[5] M. Rajiullah et al., "Web Experience in Mobile Networks: Lessons from Two Million Page Visits," in *WWW*, 2019.
[6] E. Bocchi et al., "Measuring the Quality of Experience of Web Users," *ACM SIGCOMM CCR*, vol. 46, 2016.