



Longitudinal, large-scale and unbiased Internet measurements

Flavia Salutari

► To cite this version:

Flavia Salutari. Longitudinal, large-scale and unbiased Internet measurements. Networking and Internet Architecture [cs.NI]. Institut Polytechnique de Paris, 2021. English. NNT : 2021IPPAT023 . tel-03497586

HAL Id: tel-03497586

<https://theses.hal.science/tel-03497586>

Submitted on 20 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Longitudinal, large-scale and unbiased Internet measurements

The users, the Web, the models

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Institut Polytechnique de Paris (IP Paris)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Paris, le 21 septembre 2021, par

FLAVIA SALUTARI

Composition du Jury :

Isabelle Chrisment Professor, LORIA Campus Scientifique	Présidente, Rapporteure
Pedro Casas Senior Scientist, AIT Austrian Institute Of Technology	Rapporteur
Chadi Barakat Senior Researcher, INRIA	Examineur
Tobias Hoßfeld Professor, University of Würzburg	Examineur
Marco Mellia Professor, Politecnico di Torino	Examineur
Philippe Owezarski Director of Research, LAAS-CNRS	Examineur
Dario Rossi Chief Expert, Huawei	Co-directeur de thèse
Mauro Sozio Professor, Télécom Paris	Directeur de thèse

*A U. C.,
per l'amore immenso,
per avermi insegnato il valore dell'inchiostro
e come impugnare la penna nel modo giusto.*

CONTENTS

ABSTRACT	ix
LIST OF FIGURES	xi
LIST OF TABLES	xv
LIST OF ABBREVIATIONS	xvii
1 RÉSUMÉ DE LA THÈSE EN FRANÇAIS	1
2 INTRODUCTION	3
2.1 Context and Motivation	3
2.1.1 Quality of Experience on the Web	4
2.1.2 Machine Learning in modern science	6
2.2 Reading map and Contributions	8
2.2.1 Thesis Organization	8
2.2.2 Publications	10
3 THE USERS AND THE WORLD WIDE WEB	13
3.1 The Wikipedia case	13
3.2 Background and related work	15
3.2.1 Web QoE metrics	17
3.2.2 State of the art limitations	18
3.2.3 Our contribution	18
3.3 User feedback collection	20
3.3.1 Technical aspects of the survey collection	20
3.3.2 Collected features	22
3.3.3 Ethics	24
3.3.4 Validity of the collection methodology	26
3.4 User feedback characterization	27
3.4.1 Aggregate view	27
3.4.2 Temporal breakdown	29
3.4.3 Spatial breakdown	32
3.5 User feedback prediction	35
3.6 Discussion	42
3.7 Conclusions	44
4 THE USERS AND THE CONTROLLED WEB	45
4.1 The Multi-Modality of uPLT	45
4.2 Data Collection	47

4.2.1	Representative Webpage Selection	47
4.2.2	Objective Web Quality Metrics	49
4.2.3	UPLT Crowdsourcing	50
4.3	Understanding Users' Feedback	50
4.3.1	UPLT Distribution Analysis	51
4.3.2	Page Characteristics and uPLT	53
4.3.3	Evaluation of Web Quality Metrics	55
4.4	Discussion	57
4.5	Conclusions	58
5	THE WEB AND THE MODELS	61
5.1	Is the Web diverse enough?: on the fairness of language models	61
5.2	Background and related work	63
5.3	Methodology	64
5.3.1	Corpora for Spoken English	64
5.3.2	Bias in Masked Language Modeling	66
5.4	Quantifying the Bias	68
5.4.1	Measuring the Bias of LMs	68
5.4.2	Bias on AAE Features	72
5.4.3	Bias on Part-of-Speech	75
5.5	Conclusions	75
6	THE MODELS AND THE MACHINES	79
6.1	IP-ID classification via supervised learning	79
6.2	Background and related work	81
6.2.1	Normative reference	81
6.2.2	IP-ID Classification Breakdown	82
6.2.3	IP-ID Based-Inference	83
6.3	Methodology	84
6.3.1	Active probing	84
6.3.2	Features Definition	87
6.3.3	Datasets	88
6.4	IP ID Classification	92
6.4.1	Classification accuracy and validation	92
6.4.2	Robustness	93
6.5	Internet Census	96
6.5.1	Longitudinal Comparison (over the years)	97
6.5.2	Sensitivity Analysis	98
6.5.3	Spatial analysis	99

6.6	Conclusions	101
7	CONCLUSIONS	105
7.1	Summary of our contributions	105
7.1.1	Users' acceptance in the Wild Web	105
7.1.2	User Perceived Page Load Time in controlled experiments	106
7.1.3	The fairness of models trained on the Web	107
7.1.4	Supervised learning to infer machine-generated content	107
7.2	Future work	108
I	APPENDIXES	109
A	APPENDIX CH. 2	111
	BIBLIOGRAPHY	119

ABSTRACT

Today, a world without the Internet is unimaginable. By interconnecting billions of people worldwide and by offering an uncountable number of services, it is now fully embedded in the modern society. Yet, despite technology evolution and development, its pervasiveness and heterogeneity still raise new challenges, such as security concerns, monitoring of the users' Quality of Experience (QoE), care for transparency and fairness.

Accordingly, the goal of this thesis is to shed new light on some of the challenges emerged in recent years. In particular, we provide an in-depth analysis of some of the most prominent aspects of modern Internet. A particular emphasis is given on the World Wide Web, which among all, is undoubtedly one of the most popular Internet applications, and a specific regard to its interaction with machine learning.

The first part of this work studies the Quality of Experience of users' browsing the Web, with measurements led both in the wild and in controlled environments. Our contributions follow with an original analysis of both the *subjective* user feedback and the *objective* QoE metrics, showing how hard it is to build accurate supervised data-driven models capable to predict the user satisfaction, along with an in-depth discussion of the multi-modal nature of the *subjective* user opinions.

In the second part of this work, we analyze and discuss the fairness of state-of-the-art transformer-based language models, which are pre-trained on Web-based corpora and which are typically used to solve a wide variety of Natural Language Processing (NLP) tasks. Here, we question whether the sheer size and heterogeneity of the Web guarantee diversity in the models. The core of our contributions rests in the measure of the bias embedded in the models, that we discuss under different angles.

Finally, the last part of this dissertation addresses the classification of objects generated by machines through some of the simplest state-of-the-art supervised machine learning algorithms. Through a minimally intrusive, robust and lightweight framework, we show that the different behaviors of a field of the IP packet, the IP identification (IP-ID), could be easily classified with few features having high discriminative power. We finally apply our technique to an Internet-wide census and provide an updated view of the adoption of the different implementations in the Internet.

LIST OF FIGURES

Figure 1	Relationship between QoS at different layers and user QoE. . . .	5
Figure 2	Outline of the thesis.	9
Figure 3	Appearance of the Survey in the English Wikipedia (answer order is randomized).	20
Figure 4	Quantile-quantile plot of PLT statistics for different sets	26
Figure 5	Aggregate statistics of navigation timing performance (TTI, RSI and PLT in the figure), conditioned by survey response.	28
Figure 6	Annotation of major Wikipedia-related events occurred during the whole 5-months observation period.	29
Figure 7	Temporal view: daily mean of PLT, TTI and RSI during the observation period.	29
Figure 8	Temporal view: breakdown of daily survey answers among positive, neutral and negative scores.	29
Figure 9	Temporal view: absence of night/day seasonality of survey answers.	31
Figure 10	Temporal view: absence of weekday/weekend seasonality of survey answers.	31
Figure 11	Illustration of spatial breakdown of user scores across <i>page</i> , <i>user</i> and <i>environment</i> features obtained by conditioning each of them over different values and showing on the top x-axis the cardinality of samples for each bar.	33
Figure 12	ROC curves, obtained when averaging the results gathered with a 10-fold cross validation on \mathcal{B} with the <i>PA</i> features set.	38
Figure 13	Classification results, feature subsampling: performance obtained by limiting the (<i>T</i>) total features, 61 (<i>WWW</i>) features, 19 (<i>PA</i>) publicly available features, both with (<i>PA_{w.o.}</i>) and without (<i>PA</i>) outliers filtering.	39
Figure 14	Classification results, dataset subsampling: performance obtained by restricting the attention to (a) Chrome-only browser, (b) Russian population (c) Android OS and (d) top-1000 pages (and combinations thereof).	40
Figure 15	Ranking of the features according to their SHAP values.	41

Figure 16	Relevant snapshots of the <i>www.booking.com</i> rendering process corresponding to the different modes that are visible in the distribution reported in Fig.17. Notice that the “above the fold” content is almost all rendered in (a) and fully rendered in (b). At time (c) a popup arise, inviting users to login in the website.	46
Figure 17	uPLT distribution for <i>www.booking.com</i> , highlighting the issue that users do not agree on a single time instant to identify completion of webpage rendering.	46
Figure 18	Number of webpages per cluster.	48
Figure 19	Component weights for pages with multi-modal uPLT.	51
Figure 20	Ranking of the features according to their SHAP values when predicting uni-modal pages.	54
Figure 21	$ PLT - TTI $ ECDF for uni/multi-modal pages.	55
Figure 22	The ECDF of the utterance length ℓ_u for both AAE and SAE corpora.	66
Figure 23	The difference between the ECDFs of SAE and AAE for the $\Delta P(u)$ measure. When the values are greater than zero the LMs are more biased towards SAE, <i>vice versa</i> otherwise.	70
Figure 24	The difference between the ECDFs of both AAE and SAE for the $CRR(u)$ measure. When the values are greater than zero the LMs are more biased towards SAE, <i>vice versa</i> otherwise.	71
Figure 25	Illustration of Constant, Local, Global, Random and Odd sequences	81
Figure 26	Scenario in which the active probing is performed: only one sender is used to ease the synchronization of packets generation, whilst both the machines are used to receive and collect the stream of IP-IDs generated at the target machine.	85
Figure 27	Sensitivity analysis to external traffic: derivative of the sequence of IP-IDs χ' in the two different scenarios	86
Figure 28	Internet campaign: ECDF of the number of packet replies	90
Figure 29	Manual Ground Truth: Normalized classes occurrences for the training datasets \mathcal{G} and \mathcal{G}'	91
Figure 30	Validation: Confusion Matrix of 20-fold validation over \mathcal{G} done both with Decision Tree and Random Forest Classifiers	92
Figure 31	Validation: Relative importance for the most useful features of the classifier.	93

Figure 32	Robustness: (left) Confusion Matrix of a classifier trained on the real lossless dataset \mathcal{G} and tested on the synthetic lossy dataset $\mathcal{S}_{\text{lossy}}$ with purposefully injected 20% packet losses on each sequence, (right) Confusion Matrix of a classifier trained on the real lossless dataset \mathcal{G} and tested on the dataset where 20% of each sequence is intentionally randomly swapped $\mathcal{S}_{\text{reorder}}$	95
Figure 33	Robustness: Misclassification breakdown of the (local,odd) (14%) for the different loss models.	95
Figure 34	Probing Overhead analysis: Accuracy as a function of the sample set size	96
Figure 35	(a) Internet campaign: Normalized classes occurrences for the training \mathcal{G} and Internet-scale \mathcal{L} dataset; (b) Measured occurrences of Global IP-ID implementations over the years; (c) Breakdown of the classes of \mathcal{L} obtained with both G' and G	97
Figure 36	Breakdown of the classes of \mathcal{L} obtained with both a Decision Tree and a Random Forest Classifier.	98
Figure 37	Normalized classes occurrences for \mathcal{L} and its lighter version when only $N=10$ packets out of 100 are considered.	100
Figure 38	IP-ID census results, shown as a 12th order Hilbert curve, a fractal space-filling curve that allows the mapping of the one-dimensional IPv4 address space into a bi-dimensional image.	100
Figure 39	Standard Deviation of IP-ID classes of IP addresses owned by same AS	102

LIST OF TABLES

Table 2	Summary of recent related work gathering user feedback for Web quality of experience assessment.	16
Table 3	Collected corpus of Wikipedia users' QoE feedback.	21
Table 4	Summary of the features (T/WWW/PA) that are associated to each users' survey response. The mutual information between the survey answer and T/WWW/PA features in the class is reported as a boxplot.	24
Table 5	User feedback prediction: Confusion matrixes, obtained when averaging the results obtained with a 10-fold cross validation on \mathcal{B} with the PA features set.	36
Table 6	User feedback prediction: classification results expressed through several metrics, obtained when averaging the results obtained with a 10-fold cross validation on \mathcal{B} with the PA features set. . .	37
Table 7	Breakdown for $M_{mass} = M_{time}$	52
Table 8	Statistics of uni-modal/multi-modal pages.	53
Table 9	RMSE of (top) uni-modal and (bottom) multi-modal uPLT with Web quality metrics.	56
Table 10	Summary of recent related work.	57
Table 11	Corpora summary: with and without filtering utterances (\mathcal{U}) based on their length. With $\langle \ell_u \rangle$ we indicate the average utterance length; with L , the length of the corpus in number of words, and; with $ \mathcal{T} $, the number of terms (unique words). . . .	65
Table 12	Training data for the tested LMs.	66
Table 13	Example showing the masked token experiment.	67
Table 14	MAE and MSE of $\Delta P(u)$ and $CRR(u)$ measured on AAE and SAE corpora: results obtained through the <i>fill-in-the-blank</i> task with different language models. \dagger signifies that the AAE and SAE expectations are statistically significant according to the Welch's two-tailed t-test (p-value < 0.05). The column d contains their effect size computed according to the Cohen's d. . . .	69
Table 15	A sample of AAE utterances selected based on their syntactical features and their translations to SAE. In brackets the prevalence of the feature over the utterances in the AAE corpus. . . .	72

Table 16	Similar to Table. 14 but calculated over a sample of 50 utterances of AAE and their translated version (AAE ^T) for each feature of AAE.	73
Table 17	Similar to Table. 14 but calculated for t rather than u, for three POS classes.	74
Table 18	MAE and MSE of $\Delta P(t)$ and CRR(t) measured on AAE and SAE corpora: results obtained through the <i>fill-in-the-blank</i> task with different language models, averaging token predictions for each POS class. † signifies that the AAE and SAE expectations are statistically significant according to the Welch’s two-tailed t-test (p-value < 0.05). The column d contains their effect size computed according to the Cohen’s d.	77
Table 19	Summary of related work	82
Table 20	Tabulated expected values for selected features	87
Table 21	Summary of the Datasets.	89
Table 22	Features values for both lossless and lossy synthetic dataset $\mathcal{S}_{\text{lossy}}$ with 20 % losses - local implementation case of IP-ID. . .	94
Table 23	List of features, informing whether each comes from <i>raw</i> data or is <i>derived</i> , and if it is present in the T, WWW or PA set. . . .	111
Table 24	Schema of the Public Available Features.	115

LIST OF ABBREVIATIONS

AAE	African American English
ACR	Absolute Category Ranking
AS	Autonomous System
AI	Artificial Intelligence
ASN	Autonomous System Number
AATF	Approximated Above The Fold
ATF	Above The Fold
AUC	Area Under the Curve
CRR	Complementary Rectiprocal Rank
DOM	Document Object Model
GDP	Gross Domestic Product
GMM	Gaussian Mixture Model
K-NN	K Nearest Neighbor
IP	Internet Protocol
IP-ID	Internet Protocol IDentification
ISP	Internet Service Provider
LM	Language Model
MAE	Mean Absolute Error
MI	Mutual Information
MLM	Masked Language Modeling
MLP	Multi Layer Perceptron
MOS	Mean Opinion Score
MSE	Mean Squared Error
NLP	Natural Language Processing
OS	Operating System
PA	Publicly Available
PLT	Page Load Time

POS	Part Of Speech
QoE	Quality Of Experience
QoS	Quality Of Service
RMSE	Root Mean Squared Error
wRMSE	Weighted Root Mean Squared Error
ROC	Receiver Operating Curve
RUM	Real User Monitoring
RR	Rectiprocal Rank
RSI	Rum SpeedIndex
RTT	Round Trip Time
SAE	Standard American English
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machines
TTC	Time To Click
TTFB	Time To First Byte
TTFP	Time To First Paint
TTI	Time To Interactive
uPLT	User Page Load Time

RÉSUMÉ DE LA THÈSE EN FRANÇAIS

Aujourd'hui, un monde sans Internet est inimaginable. En inter-connectant des milliards de personnes dans le monde et en offrant un nombre incalculable de services, il est désormais pleinement intégré à la société moderne. Pourtant, malgré l'évolution et le développement de la technologie, son omniprésence et son hétérogénéité soulèvent encore de nouveaux défis, tels que les problèmes de sécurité, le contrôle de la qualité d'expérience des utilisateurs (QoE), le souci de transparence et celui d'équité. En conséquence, l'objectif de cette thèse est d'apporter un nouvel éclairage sur certains des défis qui ont émergé ces dernières années. En particulier, nous fournissons une analyse approfondie de certains des aspects les plus importants de l'Internet moderne. Un accent particulier est mis sur le World Wide Web, qui, parmi tous, est sans doute l'une des applications Internet les plus populaires, et un regard spécifique sur son interaction avec l'apprentissage automatique. Nous suivons donc deux directions de recherche principales: premièrement, nous nous concentrons sur l'analyse de la qualité de l'expérience des utilisateurs du Web (Chap. 3 et Chap. 4); deuxièmement, nous mettons l'accent sur l'utilisation de l'apprentissage automatique appliqué aux mesures d'Internet, en étudiant, d'une part, l'impact de son interaction avec le Web (Chap. 5) et, d'autre part, son utilisation pour prédire les objets générés par les machines (Chap. 6).

La première partie de ce travail étudie la qualité de l'expérience de navigation des utilisateurs sur le Web (Web QoE), avec des mesures effectuées à la fois "*in the wild*" et dans des environnements contrôlés. Dans le chapitre 3, nous abordons le problème de l'évaluation de la qualité de l'expérience sur un site web populaire en fonctionnement. Plus précisément, nous le faisons en recueillant l'*acceptance* de Wikipédia par les utilisateurs, soit plus de 62k de réponses au sondage, ce qui représente plus du double des réponses recueillies dans des études similaires à grande échelle sur Wikipédia. Nos contributions continuent avec une analyse originale de l'avis *subjectif* des utilisateurs et des mesures *objectives* de la qualité d'expérience, montrant, d'un côté, une dépendance spatiale entre des caractéristiques collectées, d'un autre côté, la difficulté de construire des modèles supervisés précis, basés sur les données disponibles, capables de prédire la satisfaction des utilisateurs. Dans le chapitre 4, nous nous concentrons plutôt sur la mesure de la qualité de l'expérience dans des expériences contrôlées. En particulier, nous avons mesuré le temps de chargement des pages perçu par les utilisateurs, c'est-à-dire le moment où un utilisateur considère qu'une page web est chargée et prête à être parcourue,

sur 108 pages web via la plateforme Eyeorg. Dans ce chapitre, on trouve une discussion approfondie de la nature multimodale des avis *subjectifs* des utilisateurs.

Dans la deuxième partie de ce travail, nous analysons et discutons l'équité des modèles de langage basés sur des transformateurs de pointe, qui sont pré-entraînés sur des corpus basés sur le Web et qui sont généralement utilisés pour résoudre une grande variété de tâches de traitement du langage naturel (NLP). Dans le chapitre 5 nous nous demandons si la taille et l'hétérogénéité du Web garantissent la diversité des modèles. Le cœur de nos contributions repose sur la mesure du biais intégré dans les modèles, que nous discutons sous différents angles. Nous observons que l'équité des grands modèles, entraînés sur une énorme quantité de contenu Web, est déséquilibrée. Plus précisément, les prédictions faites avec BERT et DistilBERT sur l'anglais américain standard sont jusqu'à 21% plus précises que celles faites sur l'anglais afro-américain. Nous montrons également qu'au contraire, BERT, RoBERTa et DistilRoBERTa présentent un biais opposé, favorisant alors l'anglais afro-américain. Nos résultats soulignent également que les variantes distillées de BERT et RoBERTa, conçues pour être plus légères et entraînées sur une quantité moindre de données, sont les plus justes parmi les sept modèles de langage testés.

Enfin, la dernière partie de cette thèse traite de la classification d'objets générés par des machines à l'aide de certains des plus simples algorithmes d'apprentissage automatique supervisés à l'état de l'art. Dans le chapitre 6, grâce à un framework solide mais peu intrusif, nous montrons que les différents comportements d'un champ du paquet IP, l'identification IP (IP-ID), peuvent être facilement classifiés avec peu de caractéristiques ayant un haut pouvoir discriminatoire. Nous appliquons enfin notre technique à un census à l'échelle de l'Internet et fournissons une vue actualisée de l'adoption de ses différentes implémentations dans l'Internet. En particulier, les résultats du census révèlent que le *global* n'est plus l'implémentation d'IP-ID la plus courante et que, au contraire, d'autres comportements, comme le *local* et le *constant*, sont présents. Du point de vue de la méthodologie, nous constatons que quelques caractéristiques scalaires et un classificateur simple, suffisent pour prédire avec précision les différentes mises en œuvre des IP-ID. Du point de vue des résultats, en revanche, l'application de cette technique fournit une vue actualisée de l'adoption des différentes implémentations d'IP-ID connues.

INTRODUCTION

Contents

2.1	Context and Motivation	3
2.1.1	Quality of Experience on the Web	4
2.1.2	Machine Learning in modern science	6
2.2	Reading map and Contributions	8
2.2.1	Thesis Organization	8
2.2.2	Publications	10

2.1 CONTEXT AND MOTIVATION

Originally conceived to allow the communication among multiple computers on a single network, the Internet is nowadays a worldwide network used by more than 59% of the global population every day [57]. It has become an integral part of the society that today, a world without the Internet is unimaginable. By connecting billions of people worldwide and by enabling an uncountable number of services, the Internet is now a core pillar of the modern information society and a critical piece of the human infrastructure.

At its dawn, the Internet was designed to support very few simple services, which were working over rigidly-specified protocols, and mostly text-based (*e. g.*, electronic mail, remote login, *etc.*). Over the last two decades, along with interconnecting more and more users, it sparked the proliferation of many and diverse applications, significantly increasing the number and the type of activities that users can carry out online. Nowadays, the Internet is a complex and big network, with multiple layers of protocols that interact with each other. It is an evolving system which is constantly changing in operations, size, technologies, and economic relationships, all of which evolve at different time scales.

Among all, the World Wide Web is undoubtedly one of the most successful Internet application. Indeed, nowadays, the Web usage is no longer restricted to sending and reading emails or finding information through search engines but also for watching videos in streaming (*e. g.*, Twitch, Youtube, *etc.*), for building social relations (*e. g.*, Twit-

ter, LinkedIn, *etc.*) and for buying goods (*e.g.*, Amazon, AliExpress, *etc.*). It is then no surprise to find in the Alexa top-10 websites rank of 2021 [1] the predominance of social networks (*baidu.com*, *facebook.com*, *etc.*) and shopping platforms (*tmall.com*, *taobao.com*, *etc.*).

Due to its pervasive and heterogeneous nature, the Internet raised the interest of both industry and academia researchers, who face a constantly renewed interest in exploring and understanding the dynamics interacting in this continuously evolving system. Over the years, to cope with the always increasing Internet demand and popularity, novel network infrastructures, devices and tools, as, for instance, lighter browsers in the case of the Web, have been introduced.

Moreover, along with technology evolution and development, also the habits and the expectations of the users have changed significantly. Indeed, the pervasiveness of the Internet still raises new challenges: traffic growth, security concerns, users' quality of experience, economic interests, transparency and fairness. Of those, Quality of Experience (QoE) embraces many of the others, so that, in this regard, all the Internet players (Domain Name Services, Internet Service Providers, *etc.*) have to keep up using cutting-edge technologies in order to provide a satisfactory QoE to the users. This is crucial since if only one of them is affected by an outage, the entire Internet ecosystem might be affected as well, compromising the final experience of the users.

2.1.1.1 *Quality of Experience on the Web*

Web browsing is one of the most popular applications for both desktop and mobile users. Slow rendering of the websites was due in the 80s to dial-up connections, in the 90s to slow 2G connections and it persists nowadays for a wide range of reasons, including the growingly more complex structure of websites [151] and an increased usage of mobile devices [41, 111]. This sometimes caused to the World Wide Web, since its inception, the title of World Wide "Wait" [156]. Indeed, not surprisingly the first assessment of Web performance can be traced back to the early 90s. Since then, a lot of effort has been devoted to speed up the Web, as well as in designing metrics that can accurately tell whether a webpage loaded fast or not and that can capture well the users' Quality of Experience.

The user-perceived latency, in particular, plays an important role in this regard and has tangible consequences: Amazon claims that 100ms latency penalty results in a 1% sales loss [90], Google affirms that an additional delay of 400ms in search responses reduces search volume by 0.74% [26] and Bing that 500ms of latency decreases the revenue per user by 1.2% [93]. This is why an increasing attention has been given over the

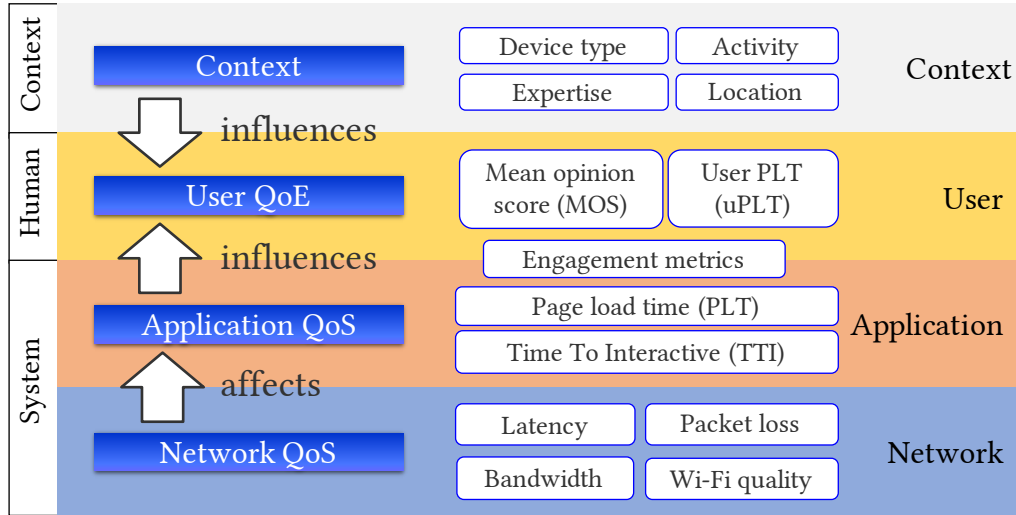


Figure 1: Relationship between QoS at different layers and user QoE.

years to the QoE offered to final users and, more specifically, this is the reason why a lot of effort has been devoted to reduce delays. However, it is still unclear if and by how much a latency reduction translates into a better perceived experience from the user point of view. This, coupled with the manifest aspect of Web QoE heavily impacting revenues for Web-based companies, motivated a proliferation of new metrics proposals and validation studies attempting to better capture human perception on browsing experience. Despite the considerable progresses in this direction, the quality of experience of Web users' remains still largely impenetrable. This difficulty is the result of the fact that users QoE on the Web is affected by several factors. Some of them are often measurable, like those tied to the system (*e. g.*, the network and application Quality of Service metrics), some others more frequently unknown and not trackable, more tied to the context in which the user is located, as the user expectations or expertise. Indeed, the concept of QoE combines user perception, experience, and expectations with Quality of Service (QoS) metrics. Figure 1 shows all the elements which contribute to users' QoE, including the relationship between the system influence factor, *i. e.*, network and application QoS, the context and the human subjective component. These factors are strongly interdependent: network-layer QoS (*e. g.*, the quality of the connection, packet losses, *etc.*) affects and can degrade application-layer QoS metrics, *e. g.*, the time at which the page is loaded (PLT) or when it becomes interactive (TTI) *etc.*, which in turn influences the way in which the user experiences the browsing.

As a consequence, the analysis of Web QoE needs to couple the collection of two type of measurements: the *objective* metrics, which include the system influence factors,

typically automatically collected with the browser, and the *subjective* human feedback, which instead require involving the users.

The *objective* QoS metrics measured from an endpoint at network layer, *e. g.*, latency, packet loss, bandwidth, *etc.*, from a session viewpoint are gathered together as meaningful metrics at application layer. These metrics rely on measurable data (*e. g.*, network, browser events) capturing Web quality [2, 11, 20, 39, 54]. Most of these metrics are available from the browser navigation timing [46] as they pinpoint precise time instants of the page loading progress, as the time when the page becomes interactive (TTI), or the time when the first pixel is painted (TTFP), or can be inferred from packet/flow-level traffic [73, 147] and are easy to include as proxy of user experience.

On the other side, the *subjective* metrics require the user feedback and rely instead on directly collecting responses from users regarding different questions related to Web QoE. Different approaches have been proposed in the literature: collect the Mean Opinion Score (MOS) by averaging answers from a set of many users which have been asked to rate on a 5-scale range their QoE [21], or ask users to comment on a video of the website rendering process [54, 81, 149], or even measure the “user acceptance” of a service [70].

Finally, to map these two kind of measurements two main approaches have been established in literature: expert models, where domain experts specify a closed form function and use *subjective* data to fit model parameters [51, 76, 77], or machine learning models, where instead the *subjective* data is used to train the model [39, 54].

2.1.2 Machine Learning in modern science

Machine learning is a well established subset of Artificial Intelligence (AI) crawling with strong technical and scientific background. Its success in several application domains yielded to a growing demand for systems that can be used by both experts and novices in the field of machine learning. Nowadays, machine learning has become an integrative part of the modern scientific methodology and an essential building block for data science. It offers the possibility to adopt automated procedures for the prediction of different phenomena based on past observations, unraveling underlying patterns present in data and providing insights about the problem under investigation. For instance, as aforementioned, machine learning has been used to build models which map the two different kind of measurements for Web QoE (*objective* metrics and *subjective* user feedback), in order to automatically extract the user-label data from the collected metrics.

Yet, machine learning should ideally not be used as a black-box tool, but rather considered as a methodology, with a rational thought process that is entirely dependent

on the problem under study. Specifically, the use of algorithms should ideally require an adequate understanding of their mechanisms, properties and limitations, in order to better contextualize and interpret their results.

This is particularly relevant in the case of the algorithms which run the risk of replicating and even amplifying human biases, particularly those affecting protected groups. Indeed, models pre-trained on large datasets could encode biases potentially damaging towards marginalized populations and could reinforce existing hegemonic viewpoints [17]. The deployment and wide commercialization of biased algorithms would expose potential ethical implications that should not be understated. This already manifested in recent years in several ways with varying degrees of consequences for the subject group. Notable cases brought to the fore include the gender bias in Amazon online recruitment tools [65], where the AI software penalized any resume that contained the word “*women’s*” in the text, hence downgrading the resumes of women who attended women’s colleges, and the COMPAS algorithm [79] used by some U.S. states to predict whether defendants should be detained or released, which was found to be biased against African-American.

This is crucial in the case of natural language processing (NLP), whose applications are among the most pervasive (*e.g.*, hiring and recruitment, chatbots, conversational systems, *etc.*) and, for which, the presence of a bias would be detrimental and have high impact harmful consequences. In this area, the last three years were characterized by the emergence of several innovative NLP algorithms designed to solve different tasks and engineered to assist the most diverse applications. Among those, groundbreaking transformer-based language models (LMs) have been proposed and gained lots of scientific interest due to their sizable improvements on a wide range of NLP tasks [43, 87, 88, 91, 124]. All the proposed models have been pre-trained trained on Web-based corpora, ranging from user-generated content, as Reddit, to encyclopedia, as Wikipedia, to literary works, as BOOKCORPUS, and news articles, as CC-NEWS.

Actually, given the sheer size and heterogeneity of the Web, one could expect these models to guarantee diversity and be not prone to bias. However, the assumption of considering large amounts of Web text as *representative* of *all* of humanity has already been widely questioned in literature. Several works [34, 74, 137, 144, 160] highlighted the risk that this could have in perpetuating dominant viewpoints, resulting in models that embed stereotypical and derogatory associations along gender, race, ethnicity, and disability status. More specifically, when pre-trained LMs demonstrate a preference depending on the way a group of people speak, *i.e.*, by understanding one group better than the other, we are dealing with a problem of algorithmic bias and, consequently, of fairness.

2.2 READING MAP AND CONTRIBUTIONS

This thesis is organized in six chapters and structured as follows. This first introductory chapter is devoted to provide a global context and present the problems addressed.

During this introductory, we divide the objective of the thesis in three main research questions, enumerated in the following:

- **Q1:** *How accurately can we measure and predict the Quality of Experience of users browsing on the Web?*
- **Q2:** *Does the Web sheer size and heterogeneity ensure the fairness of the models trained on Web-based content?*
- **Q3:** *Are machine learning models better at classification tasks when trained on content generated by machines instead of humans?*

2.2.1 Thesis Organization

In the following, we briefly sum up our contributions and schematize the outline in Figure 2:

- In Chapter 3 we study the users' Quality of Experience on the Wild Web, addressing question **Q1**. We do this by performing a large-scale study of one of the most popular websites in operation, namely Wikipedia, and explicitly asking a fraction of its users for feedback on their actual browsing experience.
 - We show that the analysis of the collected *subjective* users' feedback reveals both expected (*e. g.*, the impact of browser and network connectivity) and surprising findings (*e. g.*, absence of day/night, weekday/weekend seasonality and other temporal dependencies).
 - Also, we leverage user survey responses to build supervised data-driven models to predict user satisfaction which, despite including state-of-the art quality of experience *objective* metrics and a considerable number of features of different type, are still far from achieving accurate results.
 - The collected dataset is made publicly available, hopefully contributing in enriching and refining the scientific community knowledge on Web users' Quality of Experience (QoE).

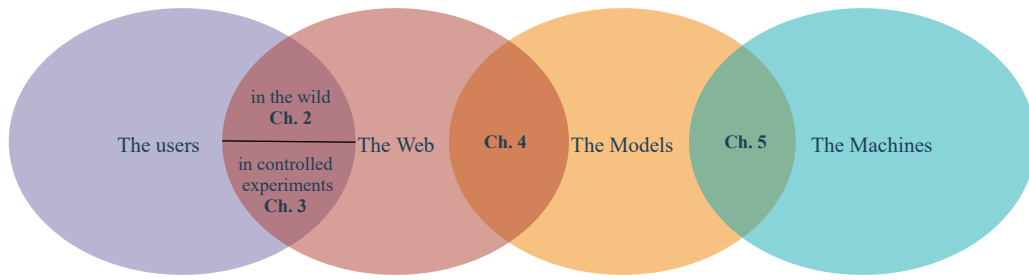


Figure 2: Outline of the thesis.

- In Chapter 4 we continue the study of question **Q1** and focus on QoE when this is measured in controlled environments, conducting experiments diametrically opposed to those shown in Chapter 3. An often implicit assumption made by industrial and academic research communities is that a *single* metric is sufficient to assess whether a webpage loaded fast.
 - In this work we collect and make publicly available a unique dataset which contains webpage features (*e. g.*, number and type of embedded objects) along with both *objective* and *subjective* Web quality of experience metrics. This dataset was collected by crawling over 100 websites—representative of the top 1 M websites in the Wild Web—while crowdsourcing the *subjective* user opinions on the *user perceived page load time* (uPLT).
 - We show that the uPLT distribution is often multi-modal and that, in practice, no more than three modes are present.
 - Our analysis reveals that, for complex webpages, each of the different *objective* QoE metrics proposed in the literature (such as ATF, TTI, PLT, *etc.*) is suited to approximate one of the different uPLT modes.
- In Chapter 5 we investigate the relationship between the Web and machine learning models, addressing question **Q2**. Specifically, we test whether the sheer size and heterogeneity of the Web guarantee also that the models trained therein are bias-free and, consequently, are *inclusive* with respect to different social groups.
 - We study the fairness of state-of-the-art transformer-based language models recently proposed and widely adopted for a plethora of NLP tasks. They have been trained on Web-based content of different size and type.
 - We propose and validate an evaluation technique to assess the quality and the bias of the predictions of 7 language models on transcripts of both spoken African American English and Standard American English.

- Our analysis shows the presence of diverse biases encoded by different state-of-the-art language models, like BERT and RoBERTa, de facto revealing that the heterogeneity of the Web is a feature that does not imply diversity.
- In Chapter 6 we instead focus on Q3 and observe the behavior of some of the simplest state-of-the-art supervised machine learning models when they are trained on content generated by machines. We show how predictions are accurate when it comes to the classification of objects with very pre-determined patterns, in sharp contrast with human-generated data as shown in Chapter 3.
 - We propose a framework to classify the different behaviors of the identification field of the IP packet (IP-ID). This in the past was mostly implemented in the operating systems as a simple packet counter, which allowed to perform a wide range of tasks. However, this behavior has been discouraged over the years for security reasons and other policies, as the use of random values, have been suggested.
 - Despite being only minimally intrusive, our technique is significantly accurate (99% true positive classification), robust against packet losses (up to 20%) and lightweight (few packets suffices to discriminate all the IP-ID behaviors).
 - We then apply our technique to an Internet-wide census, where we actively probe one alive target per each routable /24 subnet, and provide an updated picture of the Internet-wide adoption of the different known IP-ID implementations.
- Finally, in Chapter 7 we summarize and report the main contributions to the research community and conclude the thesis with a discussion of the open issues.

2.2.2 Publications

The content of this dissertation has been partially published in international conferences and journals. In the following we report the list of papers published or under review:

- Salutati, F., Cicalese, D., & Rossi, D., “A closer look at IP-ID behavior in the Wild”. *International Conference on Passive and Active Network Measurement, (PAM’18)*. 2018.
- Salutati, F., Da Hora, D., Dubuc, G., & Rossi, D., “A large-scale study of Wikipedia users’ quality of experience”, *The Web Conference (WWW’19)*. 2019.

- Salutari, F., Da Hora, D., Dubuc, G., & Rossi, D., “Analyzing Wikipedia Users’ Perceived Quality Of Experience: A Large-Scale Study”, *IEEE Transactions on Network and Service Management*. 2020.
- Salutari, F., Da Hora, D., Varvello, M., Teixeira, R., Christophides, V., & Rossi, D., “Implications of the Multi-Modality of User Perceived Page Load Time”, *IEEE Med-ComNet Conference*. 2020.
- (UNDER REVIEW) Salutari, F., Linguaglossa L. & Lipani A., “Quantifying the Bias of Transformer-Based Language Models for African American English in Masked Language Modeling”, *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021.

Moreover, the collaboration with other researchers on different topics resulted in the following contributions:

- (UNDER REVIEW) Bahri, M., Salutari, F., Putina, A. & Sozio, M., “Automated Machine Learning with a Focus on the Unsupervised Learning: a Survey”, *The International Journal of Data Science and Analytics (IJDSA)*. 2021.
- (UNDER REVIEW) Putina, A., Salutari, F., Bahri, M. & Sozio, M., “AutoAD: an Automated Framework for Unsupervised Anomaly Detection”, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*. 2021.

THE USERS AND THE WORLD WIDE WEB

Contents

3.1	The Wikipedia case	13
3.2	Background and related work	15
3.2.1	Web QoE metrics	17
3.2.2	State of the art limitations	18
3.2.3	Our contribution	18
3.3	User feedback collection	20
3.3.1	Technical aspects of the survey collection	20
3.3.2	Collected features	22
3.3.3	Ethics	24
3.3.4	Validity of the collection methodology	26
3.4	User feedback characterization	27
3.4.1	Aggregate view	27
3.4.2	Temporal breakdown	29
3.4.3	Spatial breakdown	32
3.5	User feedback prediction	35
3.6	Discussion	42
3.7	Conclusions	44

3.1 THE WIKIPEDIA CASE

Since its inception, the World Wide Web has sometimes been dubbed as World Wide “Wait” [156]. 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 [33], interactions between network protocols [50], the growingly more complex structure of websites [151], an increased usage of mobile devices [41, 111] and the emergence of new protocols [127]. Yet, whereas the study of Web performance is commonly [41, 50, 98, 111, 127, 150, 151, 163] tackled via simple objective metrics [46], and rather typically the Page Load Time (PLT), the

quality of Web users' experience is still largely impenetrable [29, 81]. As such, a number of alternative metrics that attempt at better fitting the human cognitive process (such as SpeedIndex, user-PLT *etc.*, see Section 3.2) 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, Operating Systems (OSs) and browsers, and are not exempt from cheating so that ingenuity is needed to filter out invalid answers from the labeled dataset [54, 149]. 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 [46] to sophisticated SpeedIndex [2]), 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. Summarizing our main contributions:

- first, we use survey data to deeply characterize user satisfaction along both temporal and spatial dimensions: shortly, we find that on average 85% of users are satisfied and show that user satisfaction does not exhibit seasonality at daily/weekly timescales (which is unexpected) and 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, we release the collected *dataset* as open-source (after having carefully ensured that no sensitive information is leaked in the process, see Section 3.3.3), as we hope this can help the scientific community in refining its understanding of Web users' experience.

In the remainder of this chapter, after overviewing the related work (Section 3.2), we explain the feedback collection process and dataset (Section 3.3), which we dissect under both temporal and spatial angles (Section 3.4) and that we leverage to build a data-driven model of Wikipedia users' quality of experience (Section 3.5). We finally discuss current limitations in Web QoE assessment and possible directions to circumvent them (Section 3.6) and summarize our findings (Section 3.7).

3.2 BACKGROUND AND RELATED WORK

Assessment of Web users' quality of experience can be traced back to [114], that was among the first to adapt classic results of psycho-behavioral studies gathered in the *computer* domain [102] (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 [76, 126] (and models [51]) that encode the Weber-Fechner logarithmic [76, 126] (or exponential [51]) 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 [77] and [48] first pointed out.

Still, with some exceptions [10, 11, 29, 39, 145, 149, 162] most studies still rely on simple metrics such as the Page Load Time (PLT) to assess the expected impact of new Web protocols [50, 127, 151, 163], Web accelerators [98, 150] and devices [111, 115]. 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 [109], and motivated a proliferation of new metrics proposals and validation studies attempting at going beyond PLT. Given that many different definitions of PLT [49] 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 [46].

Table 2: 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 [29]	0 + 120	30	-	-	-	3.6k	Prioritize elements (Above The Fold and user ratings)	Per-user content prioritization
2016 [149]	100 + 1k	100	n.a.	1	1	6k	Side-by-side videos (of the same site)	uPLT metric definition
2017 [21]	147 + 0	25	32	1	1	4k	Controlled browsing experiments	HTTP vs HTTP/2
2017 [162]	28 + 323	28	3	1	1	2.5k	Side-by-side videos or the same website in different protocol settings	HTTP/2 push impact
2017 [54]	0 + 5.4k	500	16	1	1	40k	Side-by-side videos (160 different website pairs)	PSI metric definition
2017 [81]	50 + 0	45	1	1	1	2.2k	Webcam, eye-tracking glasses	Eye gaze, uPLT
2018 [39]	241 + 0	12	n.a.	1	1	9k	Controlled browsing experiments	ATF metric definition
2019 [161]	0 + 50	7	11	1	1	n.a.	User rating of video rendering of Web browsing	QoE-aware networking
2019 [130]	35 + 1.2k	5	3	1	1	10k	User rating of video rendering of Web browsing	QUIC protocol
<i>this study</i>	62k users	46k	3.8k ISPs	45	2.7k	62k	User feedback from real browsing activity	User satisfaction

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

3.2.1 Web QoE metrics

Web QoE metrics fall in two main categories: *objective* and *subjective*. As we are interested in measuring browsing experience on individual pages, *engagement* metrics such as those used in [13, 103] are clearly out of scope.

Objective: As such, objective metrics of interest for Web user QoE rely on measurable data (e.g., network, browser events) capturing Web quality [2, 11, 20, 39, 54]. These metrics can be further categorized in two classes.

On the one hand, there are *tracking* metrics that either *pinpoint precise time instants* and *track specific events* of the W3C Navigation Timing [46]: 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), the Time to The First Byte (TTFB), the time at which the page is fully loaded (PLT)¹ or the time when the Above The Fold (ATF) portion of the page is rendered [27]. Most of these metrics are available from the browser Navigation Timing [46] or can be inferred from packet/flow-level traffic [73, 147] and are easy to include (though not necessarily relevant) as proxy of user experience: for instance, [29] aim at prioritizing delivery of content that is rendered above the fold (further specializing content relevance for each user).

On the other hand, there are the *integration* metrics that are founded on the idea that one page can render faster than another despite finishing loading at the same “time” (e.g., in terms of PLT). These metrics *integrate all events of the waterfall* representing the visual progress of the page, such as SpeedIndex [2] and variants [20, 54, 129], that have received significant attention lately. Denoting with $x(t) \in [0, 1]$ the visual completeness ratio of a page, metrics in the SpeedIndex family are defined as the integral of the residual completion $\int (1 - x(t)) dt$ and differ in the way they express $x(t)$. Initial definitions in this family required capturing movies of the rendering process [2], or to further use similarity metrics SSim [54], making them difficult to use outside a lab environment. To counter this issue, simple approximations such as the ObjectIndex/ByteIndex [20] that merely count the fraction of objects/bytes received (over the total amount), or as the RUM SpeedIndex (RSI) [129] that use areas of rectangles for objects as they are painted on screen (over the total screen size) have been proposed. In this chapter, 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 [29, 81], but they are not covered by this work (cfr. Section 3.6).

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

3.2.2 *State of the art limitations*

At the same time, the above metrics suffer from a limited validation with user feedback.

Subjective: Subjective metrics rely on directly collecting responses from users regarding different questions related to Web QoE. Typical approaches are to crowdsource the validation with A/B testing [54, 149], or by performing experiments on real pages in controlled conditions [39, 109, 112]. Both approaches have their downsides. Controlled experiments with real HTTP server/clients and emulated network conditions for a more faithful and interactive browsing experience, but are harder to scale, topping to few hundreds users and few thousands data points [39]. 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 [106]. 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) [149] or Time To Click (TTC) [54] and is often used as a ground truth of user perception. Yet, when users select a uPLT in [149], 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, [54] employs SpeedIndex and TTC to forecast which among the left or right video was selected by the user at time TTC: the classifier in [54] is accurate in predicting which of the two videos is perceived as fastest by users. Yet, findings in [54] are not informative about whether the user would have been dissatisfied from the slower rendering had s/he actually been browsing.

3.2.3 *Our contribution*

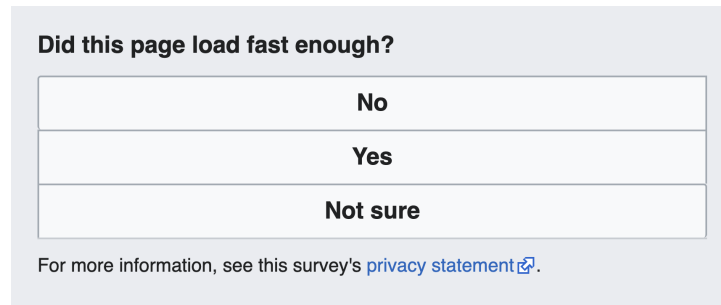
To get beyond the limitations of controlled and crowdsourced experiments just exposed in Section 3.2.2 (e. g., few users involved, lack of real user behavior representativeness and low data heterogeneity), in this work we are the first to query, at scale, Web users

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 this has not been attempted before on the wide and wild Web. Specifically, instead of collecting user feedback on a 5-grade Absolute Category Ranking (ACR) scale, we ask for a slightly more than binary feedback (see Section 3.3), which let us carry on a thorough characterization of user satisfaction (see Section 3.4) and formulate a simple, yet hard, binary classification problem (see Section 3.5). Particularly, we provide a wide variety of classification results but we especially provide a thorough characterization of user QoE along the temporal and spatial dimensions (Section 3.4).

The usefulness of the investigation we carry on and of the models we propose is clear when we consider that recent work such as [52] still employs simple *response times* as a proxy of user satisfaction for Web performance, whereas authors [6] 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 abounds [31, 32], scientific community still misses an established and agreed MOS model for Web performance. At the same time, we point out that the community started adopting slightly more accurate objective models (as in [10, 125, 134] to perform large scale studies), that are inspired by metrics such as SpeedIndex that we consider in this work, although the human component – which is among the main contribution of this work – is generally missing. 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. 2, 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 [139]). 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 devices² on 3,827 ISPs – a significant change with respect to artificial and controlled lab conditions, which make the dataset that we release at [3] of particular interest.

² As inferred from the *User-Agent* header field, after having filtered bots



Did this page load fast enough?

No
Yes
Not sure

For more information, see this survey's [privacy statement](#).

Figure 3: Appearance of the Survey in the English Wikipedia (answer order is randomized).

3.3 USER FEEDBACK COLLECTION

Wikipedia is, according to Alexa [1], 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 3.3.1), that we augment with additional information (Section 3.3.2).

We note that, while this work is not the first in leveraging Wikipedia surveys in general (see *e. g.*, [139]) 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 3.3.3). In this section, we also perform a preliminary assessment of the collection methodology, to confirm the absence of bias in the response process (Section 3.3.4).

3.3.1 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 data of a randomly selected sample \mathcal{T} of page views (less than 1 every 1,000 pageviews). The survey is then 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 3.3.3 and analyze in Section 3.4.3) 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

Table 3: Collected corpus of Wikipedia users’ QoE feedback.

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

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” [70], 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. 3. 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. 3). 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 3.3.4).

The survey is injected in the DOM after the page finished loading (*i. e.*, when the `loadEventEnd` [46] 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).

Overall, users responded as reported in Tab. 3 to about 3.6% of the over 1.7M surveys that have been displayed in the period, for a total of over 62k 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.

3.3.2 Collected features

We enrich the collected corpus with external sources that can be useful for a better understanding of the survey responses (Section 3.4) as well as being instrumental to the purpose of feedback prediction (Section 3.5). A terse summary of the metrics collected (as well as those we plan to release) is reported in Tab. 4. We discuss rationales of the selection for metrics that we make available in the publicly available dataset in Section 3.3.3.

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 at 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 which is a simple, yet non linear, transformation of page size and connection duration, by quantizing 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.

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 [100] 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 [119], 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

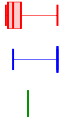
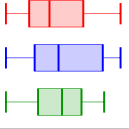



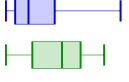
in the same environment, *i. e.*, we normalize the page download speed with respect to the median per-country speed observed in our dataset (in terms of ratio, absolute and relative error).

Additionally, ties between country wealth and network traffic volumes have been established in the literature (particularly, deviation from expected volume [140]): 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 [155]. 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 [126], 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 [55] since arguably the world-level statistics evolve on a relatively long timescale.

Finally, we expect user-home gateways [143] and particularly end-user devices [41, 111] 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 [64] to find techno-economic information about user devices and in particular, collect device CPU, memory and pricing³ 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 [41, 111]. 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 Timing 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

³ 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 4: Summary of the features (T/WWW/PA) that are associated to each users' survey response. 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, <i>etc.</i>	
Performance	32/26/18	PLT, TTI TTFP, RSI, <i>etc.</i>	
User	32/21/0	Device, Browser, editCountBucket, <i>etc.</i>	
Environment	36/12/0	Connection Type, Time, Geolocation, <i>etc.</i>	
Total			
Overall			

timing information): indeed, it can be expected that users of knowingly slow devices be less impatient, which this additional data source could provide.

3.3.3 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.

At first, as discussed in the WWW [132] paper, we proposed a conservative vetting process, that selectively filtered/obscured/aggregated information to select which features could be ultimately released publicly, out of the *total* (*T*) dataset, that in this chapter we denote with (*WWW*): the risk in this case was that perfect unlinkability could not be claimed, since we do not control all *other* sources (*e.g.*, a survey responder wishing to deanonymize himself, well-funded opponents, capable researchers, *etc.*). After a rigorous Wikipedia legal vetting process, we instead choose another option that aggressively

filters/obscures/aggregates features in an extremely conservative manner, considering only features of the performance class, that are not linked to sensitive information. In particular, the publicly available (*PA*) dataset is provided only for the Russian and French Wikipedia.

Method: Specifically, for the conservative (*WWW*) case, we opted for an approach where we transformed 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 quantized 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). 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, *etc.*

Results: As a consequence, comparing results in these two scenarios is useful to see if this loss of information potentially has an impact on the global prediction accuracy, which we assess in Section 3.5: at the same time, from results presented in Tab. 4, we can expect this effect to be rather limited. Indeed, Tab. 4 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. 4 shows that, while we only consider about half of the collected features (*T*), the (*WWW*) features overall have a *higher mutual information* (particularly, note that the 25th percentile, median and 75th percentile are higher in the (*WWW*) feature set). Thus, we conclude that:

- on the one hand, classification results of Section 3.5 are only minimally affected by selecting all (*T*), some (*WWW*) or very few (*PA*) features, so that repeatability of

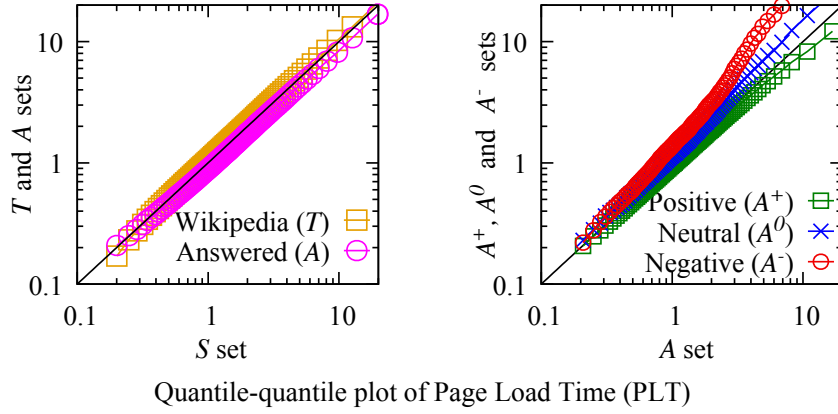


Figure 4: Quantile-quantile plot of PLT statistics for different sets ($\mathcal{T} \supset \mathcal{S} \supset \mathcal{A} = \mathcal{A}^+ \cup \mathcal{A}^0 \cup \mathcal{A}^-$).

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 3.4 would be impossible to reproduce with the available features (PA) set: under this angle, we decide to provide in this work a through spatio-temporal characterization of the collected dataset, as it would hardly be doable otherwise.

3.3.4 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 3.4) 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. 4 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 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. 4, 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 3.5). 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.

3.4 USER FEEDBACK CHARACTERIZATION

We start by analyzing the user feedback along aggregate (Section 3.4.1), temporal (Section 3.4.2) and spatial (Section 3.4.3) viewpoints, including for the time being the neutral answers.

3.4.1 Aggregate view

As previously illustrated in Fig. 4, users' grades exhibit some correlation with performance metrics such as PLT. This is consistent with results reported in Tab. 4, 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. 5 the empirical cumulative distribution functions (ECDFs) of three representative navigation time metrics [46], slicing the dataset depending on the survey answer. Partic-

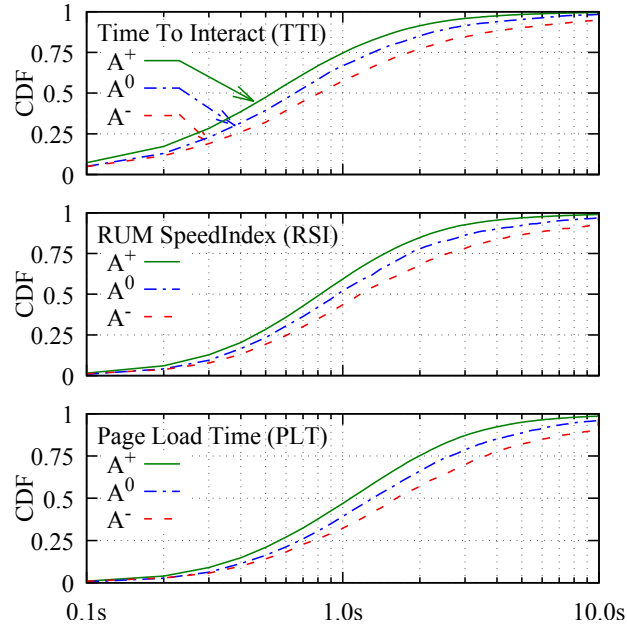


Figure 5: Aggregate statistics of navigation timing performance (TTI, RSI and PLT in the figure), conditioned by survey response.

ularly, 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. 4 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 second, 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.

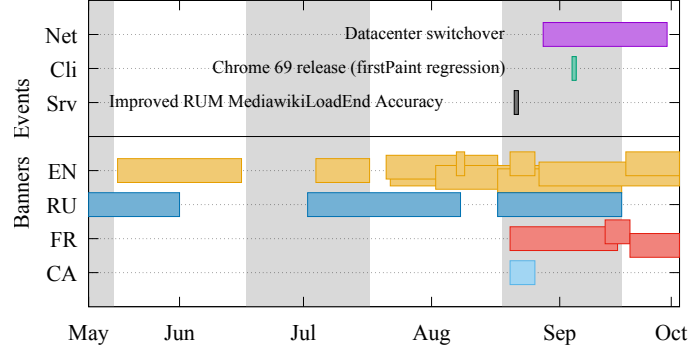


Figure 6: Annotation of major Wikipedia-related events occurred during the whole 5-months observation period.

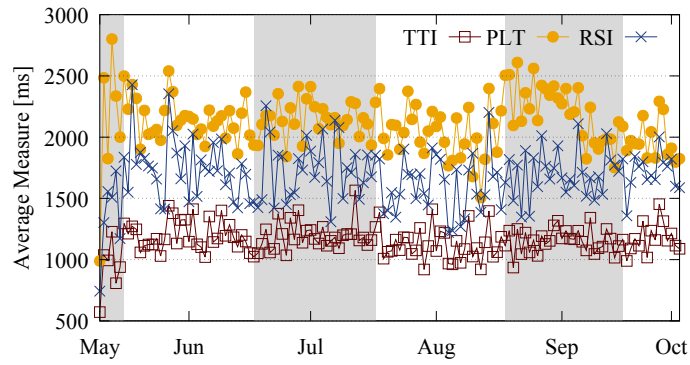


Figure 7: Temporal view: daily mean of PLT, TTI and RSI during the observation period.

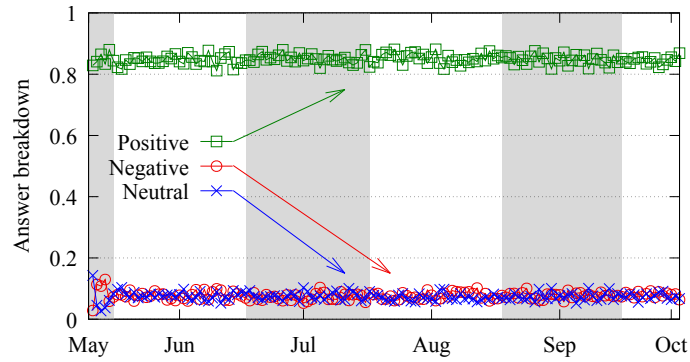


Figure 8: Temporal view: breakdown of daily survey answers among positive, neutral and negative scores.

3.4.2 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. 6, 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. 6, an operational website at scale continuously has events that are generally not available in testbeds (such as those overviewed in Section 3.2), 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 7 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. 8, 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 could be somewhat unexpected since, one could argue events such as, *e. g.*, data center switchover or browser regression to directly affect the objective measurable delay. At the same time, in light of Fig. 8, it appears that the observed level of variability in the PLT/RSI/TTI metrics happen in a range that is not 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. 9 and Fig. 10 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. Plots report the mean (line) and 95% confidence interval (shadowed band) of the metrics of interest.

Top plots in Fig. 9 and Fig. 10 do exhibit a seasonal variation in the *answers volume*. Particularly, in the hour-of-day case in Fig. 9 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.

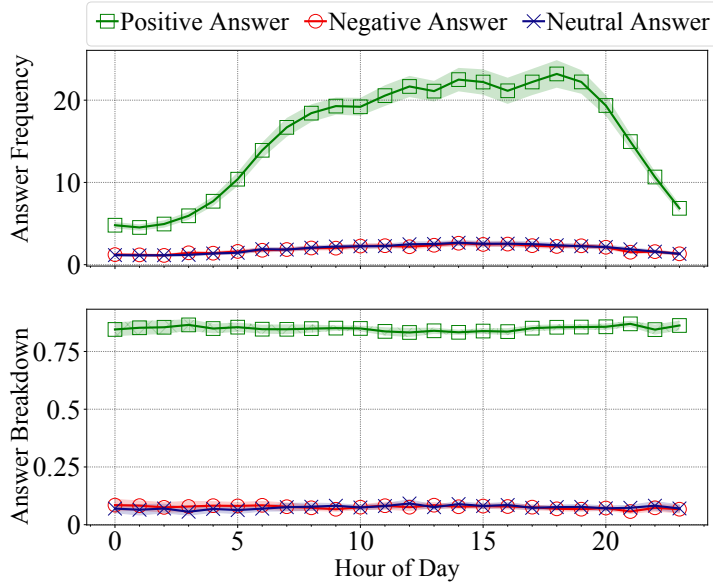


Figure 9: Temporal view: absence of night/day seasonality of survey answers.

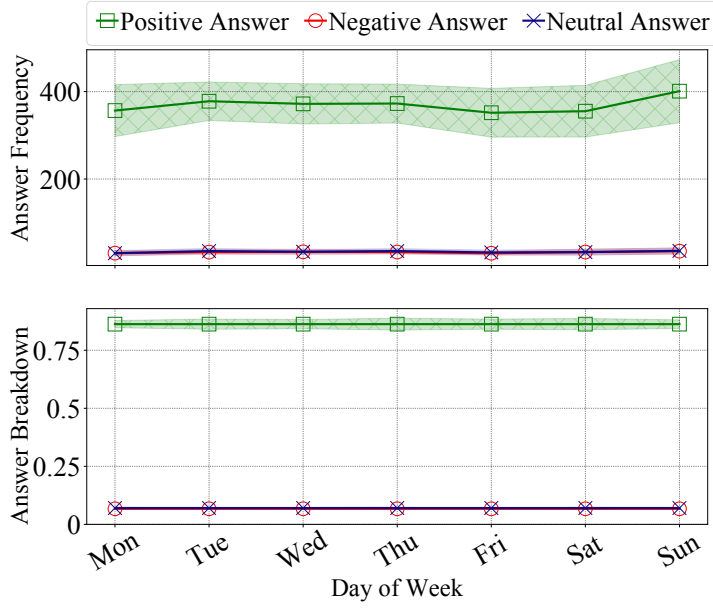


Figure 10: Temporal view: absence of weekday/weekend seasonality of survey answers.

Yet, more interesting is the absence of daily/weekly seasonality in the *answers breakdown*. From the bottom plot of Fig. 9 one clearly gathers the absence of seasonality at 24-hours circadian rhythms, which is somewhat surprising. Indeed, recent work [7] 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 [7]), 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 finer-grained 5-grade ACR scale feedback).

Similarly, from the bottom plot of Fig. 10 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 [85]) does exhibit day-of-week dependence. On the other hand, this is in line with [7] that does not remark a weekly difference in user responsiveness (*i. e.*, weekend and weekdays follow a statistically similar diurnal variability in [7]). 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, confirming the validity of our survey.

3.4.3 *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 the dependency between the user score and the performance class (*e. g.*, 3 out of the 32 collected metrics, namely TTI, RSI and PLT) has already been exposed in Fig. 4 and 5, 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. 11 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



Figure 11: 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.

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, we aggressively censure features that would allow content-linkability, making only two page-related features available in the (WWW) set: namely, the survey viewtime after the page is rendered and the HTML page size. The plot in the top left corner of Fig. 11 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 make linkability complex.

User-related metrics: Among user-related metrics, we select the browser, device and OS families (finner 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. 11.

For the family of browsers, device and OS, we report the top-8 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 [41, 111]). 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 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 Round Trip Time (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. 11 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. 11 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 [41] (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).

3.5 USER FEEDBACK PREDICTION

We continue by disregarding the neutral scores and now build data-driven models that forecast user answers.

Problem formulation: Keeping only negative and positive answers for the user feedback

Table 5: User feedback prediction: Confusion matrixes, obtained when averaging the results obtained with a 10-fold cross validation on \mathcal{B} with the PA features set.

Model	True	Predicted		All
		-	+	
Perceptron	-	0.64	0.42	4494
	+	0.36	0.58	4494
Random Forest	-	0.58	0.41	4494
	+	0.42	0.59	4494
XGBoost	-	0.63	0.41	4494
	+	0.37	0.59	4494
K-NN	-	0.57	0.43	4494
	+	0.43	0.57	4494
SVM	-	0.67	0.44	4494
	+	0.33	0.56	4494

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*.

Reference classification results: Given the class imbalance, we have to preliminarily down-sample the dataset⁴: indeed, given that after discarding the neutral scores 92% of the users are satisfied, a naïve o-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 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 the balanced dataset \mathcal{B} and the complementary dataset, only containing positive answers filtered out in the downsampling as $\overline{\mathcal{B}}$.

⁴ We prefer to avoid the diametrically opposite approach of synthetically generating users score, which is in stark contrast with the very same nature of our survey work.

Table 6: User feedback prediction: classification results expressed through several metrics, obtained when averaging the results obtained with a 10-fold cross validation on \mathcal{B} with the PA features set.

Model	Accuracy	Precision $_{\mathcal{A}^-}$	Recall $_{\mathcal{A}^-}$	Precision $_{\mathcal{A}^+}$	Recall $_{\mathcal{A}^+}$	F1 $_{\mathcal{A}^-}$	F1 $_{\mathcal{A}^+}$
Perceptron	0.60	0.64	0.47	0.58	0.73	0.54	0.65
Random Forest	0.59	0.58	0.61	0.59	0.56	0.60	0.58
XGBoost	0.61	0.63	0.50	0.59	0.71	0.56	0.64
K-NN	0.57	0.57	0.55	0.57	0.59	0.56	0.58
SVM	0.59	0.67	0.34	0.56	0.84	0.45	0.67

Tab. 5 reports 5 confusions matrixes, obtained when training a 20-trees random forest [24], a XGBoost [35] classifier, a Multi Layer Perceptron (MLP) (two layers perceptron: a Rectified Linear Unit (ReLU) and a sigmoid), a K Nearest Neighbor (K-NN) (with $K = 5$) and finally a Support-Vector Machines (SVM) classifier. Results are gathered when considering all the 19 (PA) publicly available features on a 10-fold cross validation with a 90:10 training and testing dataset split, and finally averaging the outcomes of each fold. The entire classification results are reported in Tab. 6 and expressed in terms of accuracy, precision, recall and F1 score for both the positive \mathcal{A}^+ and the negative \mathcal{A}^- answers sets. We obtain, in terms of model accuracy, similar results with both the models, with a slightly higher accuracy when using XGBoost but lower \mathcal{A}^- recall with respect to the Random Forest Classifier. Prediction outcome is clearly deceiving and only slightly better than the naïve baseline, despite the relatively large number of features collected: specifically, with the Random Forest classifier only 61% of the unsatisfied users are correctly captured, with a precision of 0.58. Interestingly, performance on the *remaining dataset* $\overline{\mathcal{B}}$, *i. e.*, the set of positive scores filtered out due to class imbalance, remains consistent with an average accuracy of 0.55.

Fig. 12 presents the Receiver Operating Curve (ROC) plots produced when evaluating the true positive rate by letting the probability threshold vary. We remark that for all the other results we keep the probability threshold fixed to 0.5. This is obtained again when averaging the results of the 10-fold validation with the abovementioned different models. This plot confirms that classification performs just slightly better than random guessing (blue dotted line), regardless of the model adopted, without any surprising result. XGBoost is the model leading to the highest Area Under the Curve (AUC) (0.65), followed by the MLP.

Feature subsampling: We next consider how the classification results change with respect to the above reference (where we consider the collected overall (T) features) when we slice the dataset by keeping only subsets of (T). We repeat the experiments with the same parameters used to obtain the above reference results and the same statistical metrics.

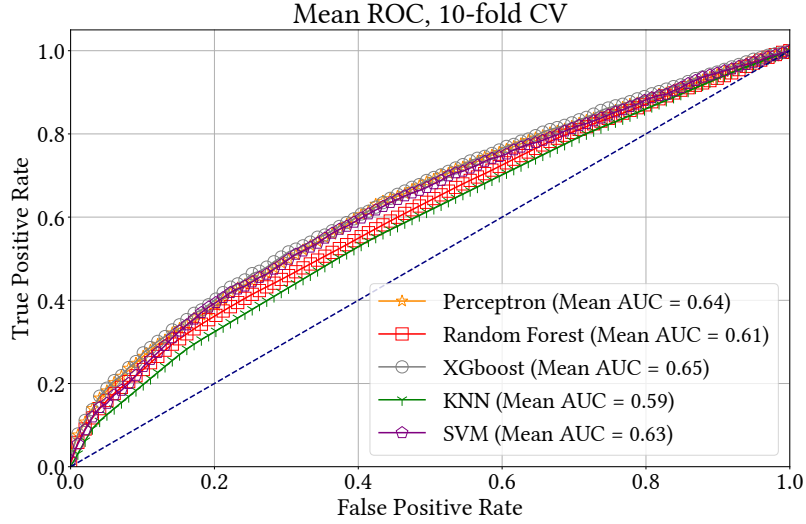


Figure 12: ROC curves, obtained when averaging the results gathered with a 10-fold cross validation on \mathcal{B} with the PA features set.

Fig. 13 reports the classification results, expressed again in terms of accuracy (top), precision (middle) and recall (bottom), when considering the portion of (WWW) features: as expected, since features in the (WWW) set are fewer with respect to (T) but having better mutual information with the survey answers, classification results are practically unaffected. We reduce this set even further by only considering the features in the (PA) set (all belonging to the performance class). In this case, results show a slightly higher, but still very limited specially in the random forest case, reduction of the classification performance: on the one hand, performance-related features consistently rank high in terms of Gini importance, though on the other hand they lack discriminative power for telling user answers apart. We also report the classification results obtained when we remove the outliers from each feature using the three-sigma rule, that we denote with $(PA_{w.o.})$. In this case, we observe that filtering out the outliers do not the performances. The solid black line in top of Fig. 13 shows the fraction of conditioned dataset with respect to the original one. This is evidently equal to 1 in the (T) and (WWW) cases, and instead is decreasing for the (PA) set, where only user records coming from French and Russian wiki are made available, and for the outliers-free $(PA_{w.o.})$ set, where roughly only 75% of the entries are kept. Furthermore, this is as well decreasing when we spatially subsample the dataset in Fig. 14, as we describe next.

Dataset subsampling: We finally spatially condition the dataset, investigating whether classification performance mechanically improves by reducing the heterogeneity in the

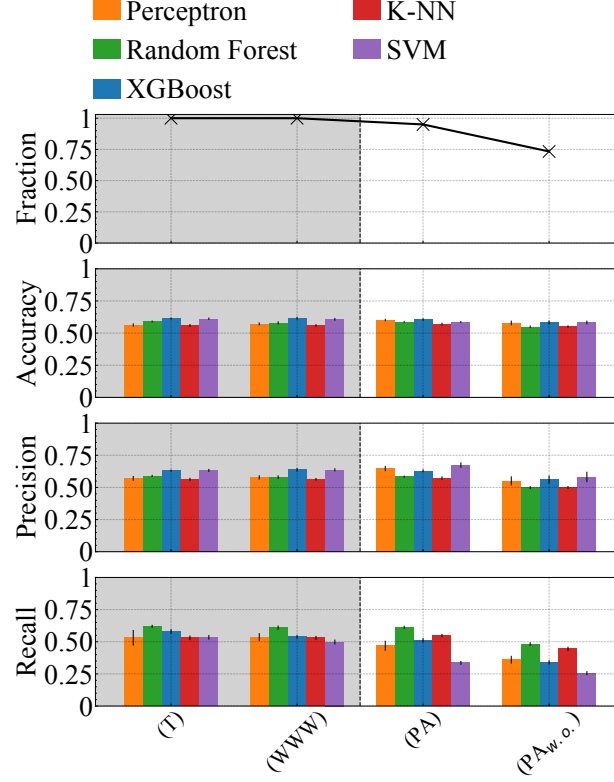


Figure 13: Classification results, feature subsampling: performance obtained by limiting the (T) total features, 61 (WWW) features, 19 (PA) publicly available features, both with $(PA_{w.o.})$ and without (PA) outliers filtering.

dataset, in an attempt to recreate more homogeneous conditions as usually done in the lab studies of Tab. 2.

Particularly, in $(C ; a)$ we use features in (WWW) set and restrict the attention to the most popular browser, namely Chrome, considering both mobile and desktop flavors. In $(WWW ; b)$ we instead restrict to users of the prevalent country, *i. e.*, Russia, and in $(WWW ; c)$ to Android users. We also combine these filters altogether $(WWW ; a ; b)$ and $(WWW ; a ; b ; c)$, and finally consider $(WWW ; d)$ the top-1000 pages in our dataset. We report in Fig. 14 the classification results obtained by running the models on each of the above dataset variants. Clearly, conditioning the dataset implies that a smaller fraction of the original dataset is available (as shown from the decreasing solid black line in top of Fig. 14), which we also have to re-balance: in turn, confidence intervals for the metrics of interest increase for decreasing dataset fractions, which is expected. Yet, it is easy to gather that classification performances are only minimally affected in all the above cases, irrespectively of the portion of features considered, the amount of homogeneity in the data or of the model adopted, so that the state-of-the art quality of experience

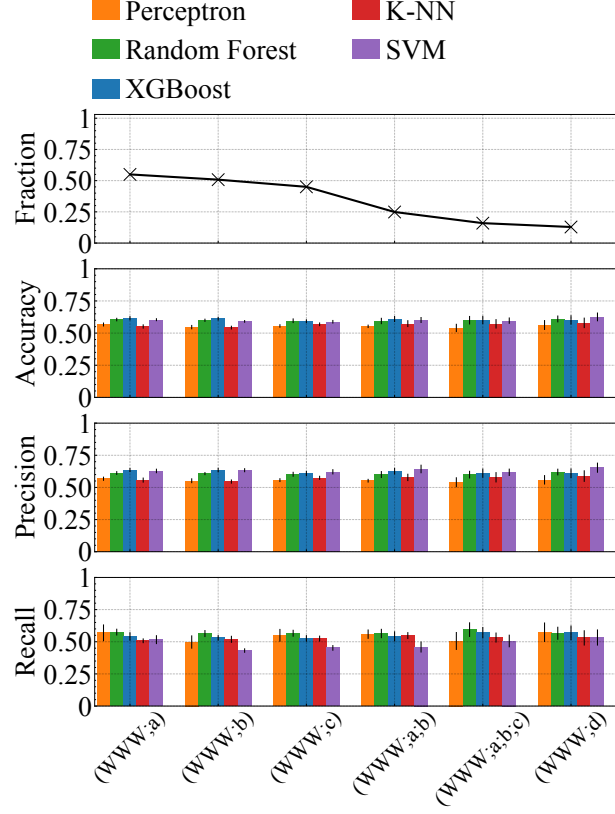


Figure 14: Classification results, dataset subsampling: performance obtained by restricting the attention to (a) Chrome-only browser, (b) Russian population (c) Android OS and (d) top-1000 pages (and combinations thereof).

metrics we collect, enriched with environmental information as described in Section 3.4, are apparently not enough to discriminate among satisfied and unsatisfied users.

Explainability: In step with the current trend towards human interpretable machine learning and model explainability, we leverage SHAP (SHapley Additive exPlanations) [96] to explain which are the relevant features that can help revealing whether a user is satisfied or not. We report in Fig. 15 the top-20 features of (WWW) set, sorted by the sum of the SHAP magnitude values computed for all the samples of the dataset, obtained with a Random Forest Classifier. SHAP values indicate the impact of each feature on the prediction, hence providing a quantitative insight of the importance of each feature for the model. On the one hand, this *summary plot* provides an overview of the most important features for the model. On the other hand, it highlights what is driving the definition of variable importance itself, the feature values. Indeed, the positive x-axis values assess the impact on the model output for predicting the *negative answer* label, whilst the negative ones refer to the *positive answer* label. For the top-five features, the higher their

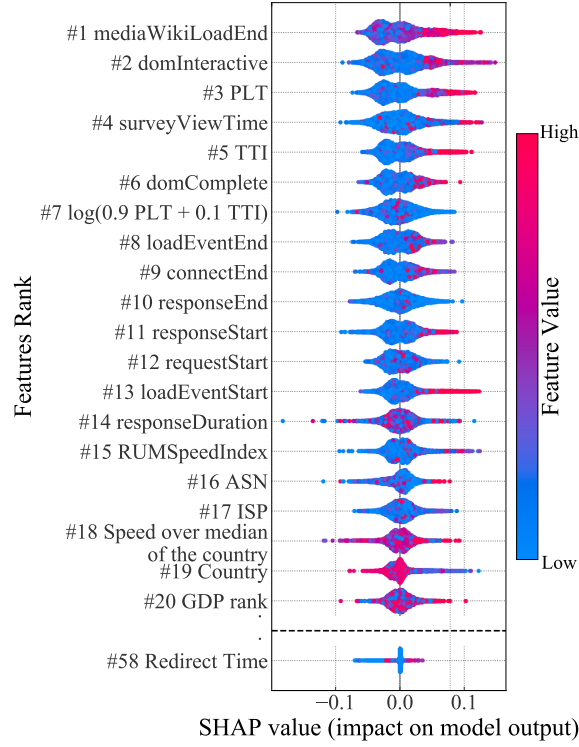


Figure 15: Ranking of the features according to their SHAP values.

values, the larger the impact on the model prediction, indicating that users experiencing longer loading times, and interestingly longer survey appearance times, are more likely associated in the prediction with the *negative answer* label. Moreover, plot in Fig. 15 also highlights the flaws of the model, showing that for none of the features there is a clear abrupt detachment between low and high values around 0, and instead low feature values are almost always implying both positive and negative SHAP values. The main takeaway of this analysis is that the top-15 features belong to the performance class, which is consistent with the *MI* output shown in Sec.3.3.3, which fortunately are also exported in the PA features set. Furthermore, despite the documented influence of HTTP redirects on Web performance results [49], we observe that the redirect time is ranked low in the feature importance scale and thus does not impact much the user feedback prediction.

3.6 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 which we discuss next.

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 discussed in this chapter 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 [12], 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 than in other pages in the Alexa top-100 typically considered in similar studies, see Section 3.2). 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 have currently improved 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 [29, 81]. 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 [110]), 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 [158]. Clearly, further research is needed on whether user-touch can be useful for similar purposes in case of mobile handsets.

Finally, (iii) previous work [39] 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 that 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 [121] but are not readily available on the Hive platform [8]. 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 [41, 111] and confirmed in this work, collecting per-device statistics seems a mandatory step to ameliorate prediction performance, as “com-

putation activities are the main bottleneck when loading a page on mobile browsers” [111]. 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.

3.7 CONCLUSIONS

In this work 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 62k 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.

Contents

4.1	The Multi-Modality of uPLT	45
4.2	Data Collection	47
4.2.1	Representative Webpage Selection	47
4.2.2	Objective Web Quality Metrics	49
4.2.3	UPLT Crowdsourcing	50
4.3	Understanding Users' Feedback	50
4.3.1	UPLT Distribution Analysis	51
4.3.2	Page Characteristics and uPLT	53
4.3.3	Evaluation of Web Quality Metrics	55
4.4	Discussion	57
4.5	Conclusions	58

4.1 THE MULTI-MODALITY OF UPLT

A good Quality of Experience (QoE) on the Web is essential for both content providers and consumers. QoE directly affects end-users' willingness to visit a webpage [150] as well as content providers' business revenues [138]. Both industry (*e. g.*, QUIC, SPDY, and HTTP/2) and academia [29, 113, 128, 150] have made significant effort to design tools and novel protocols to reduce page load times as the main factor that determines Web QoE is how fast a page loads [47].

Originally, quality of user experience on the Web was approximated using simple performance metrics like *time-to-first-byte* (*TTFB*) and the browser *onLoad* event. As modern webpages are composed of hundreds of different objects, these metrics can typically capture only the lower and upper bounds of the user perception on page load time. This limitation has motivated the introduction of a number of recent metrics to better capture user experience on the Web, such as the *Above-the-Fold* (*ATF*) [25] and the *SpeedIndex* [2].

Despite all these efforts, the question regarding how well existing single-valued metrics capture the user perception of page load time remains open. In this regard, the *user*

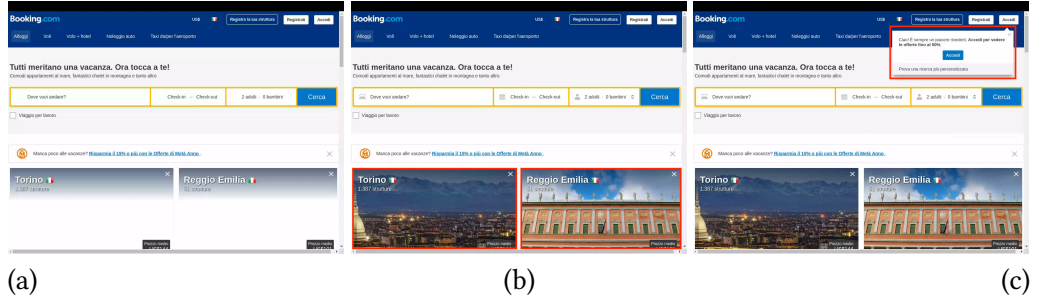


Figure 16: Relevant snapshots of the *www.booking.com* rendering process corresponding to the different modes that are visible in the distribution reported in Fig. 17. Notice that the “above the fold” content is almost all rendered in (a) and fully rendered in (b). At time (c) a popup arise, inviting users to login in the website.

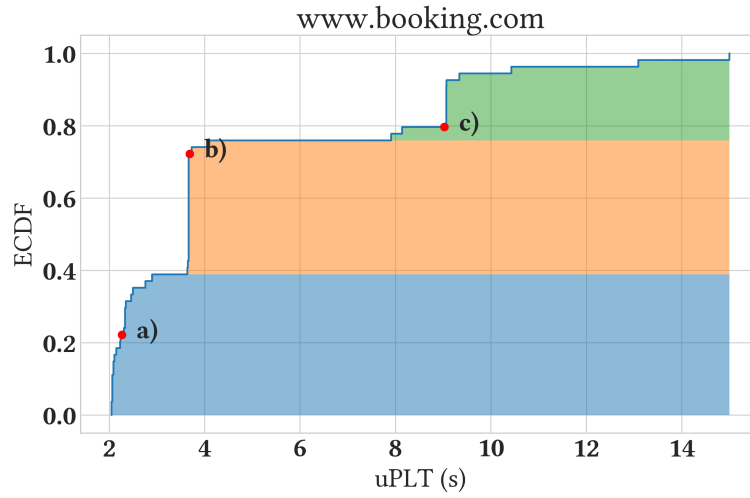


Figure 17: uPLT distribution for *www.booking.com*, highlighting the issue that users do not agree on a single time instant to identify completion of webpage rendering.

perceived Page Load Time (uPLT) is defined as the time when a user considers the webpage to be loaded and ready to browse. With few exceptions, almost the entire previous industrial and academic efforts make the implicit assumption that a *single*-valued metric can capture the uPLT across users – or, equivalently, that the distribution of uPLT of a given page across users is uni-modal. Recent studies have challenged this assumption, showing that users rarely agree on a *single* uPLT [81, 149]. However, the multi-modality of uPLT was not the main focus of these studies, and as such it was not studied in depth.

UPLT multi-modality is rooted in many factors, such as personal preferences with respect to what is considered important on a webpage, *e. g.*, text rather than images, carousels of elements, popups or ads. Fig. 17 illustrates this issue when asking for feedback from 54 recruited participants regarding the uPLT of *www.booking.com*. About 40% of users believe uPLT to be around 2 seconds, another 40% indicates $uPLT \approx 3.7$ seconds

and nearly 20% report a $\text{uPLT} \approx 9.1$ seconds. These uPLT values appear in conjunction with distinct webpage loading events; we report the snapshot of these events in Fig. 16. This example illustrates the challenges of measuring uPLT, and raises questions about which among the numerous objective Web QoE metrics (*e.g.*, PLT, TTFP, ATF [25]) is more suitable as a proxy for these remarkably different user opinions.

To address these questions, we collect a comprehensive data set of webpage features (*e.g.*, number and type of embedded objects) along with both *objective* and *subjective* Web quality metrics. We find that around 50% of the webpages in our study present a multi-modal uPLT distribution and that, in practice, three modes are sufficient to accurately describe uPLT distribution. Moreover, we show that the *number of images* and the *number of objects* in a webpage can help in predicting uPLT modality. To promote cross comparison and enable further studies, we make this dataset publicly available [5].

This chapter is organized as follows. We describe the methodology used to produce the representative set of webpages for our analysis and how we employed the Eyeorg platform [149] to crowdsource uPLT on these pages (Section 4.2). Next, we thoroughly characterize the collected user feedback (Section 4.3), rigorously quantifying violations of the hypothesis that uPLT is uni-modal and finally contrast uPLT modes with objective QoE metrics. Finally, we discuss our findings (Section 4.4) and put our results in perspective with recent related work.

4.2 DATA COLLECTION

To explore the relationship between uPLT and objective Web QoE metrics we need to (i) collect a comprehensive dataset comprising “representative” webpages, and (ii) crowdsource feedback from real users on uPLT. We first devise a novel methodology to identify a limited number (*e.g.*, 100) of webpages to test from the Cisco’s Umbrella top-1M list [72] (Sec. 4.2.1). Second, we automate the collection of webpage characteristics and objective Web QoE metrics from Chrome-based browsers (Sec. 4.2.2). Finally, we conduct an Eyeorg [149] crowdsourced campaign to ask users *when* each webpage finished loading (Sec. 4.2.3). We make the entire dataset collected publicly available [5].

4.2.1 Representative Webpage Selection

A recurring concern in Web performance research is *how* to select a meaningful set of webpages to study. Due to the sheer size of the Web, some sort of sampling needs to be introduced. To study the Web, researchers often resort to the most popular webpages from Alexa or Cisco, or a combination of popular and unpopular webpages. While it

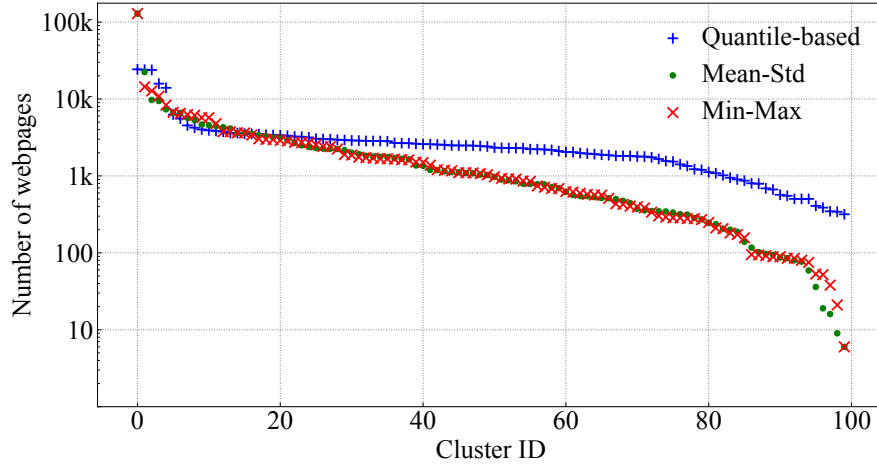


Figure 18: Number of webpages per cluster.

is important to sample popular webpages, since they attract the majority of the traffic, unpopular webpages might have a completely different set of characteristics yielding to different results. In this chapter we argue that popularity should not prime over diversity of webpages, as otherwise the results may lose generality. We therefore opt for a stratified selection of both *popular* and *diverse* pages by clustering them according to the complexity of their HTML content.

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

Clustering: On this set of webpages, we compute six features that, as reported by Butkiewicz et. al [28], are distinctive of the page characteristics and in particular have high correlation with webpage *complexity*: page size (in MB), total number of objects, number of images, CSS, javascript, and number of distinct origins. Then, we rely on K-means to find pages of similar complexity. Given our crowd-measurement budget,

we fix $K = 100$. We experiment with three standard feature normalization techniques: min-max, mean-std, and a quantile-based feature normalization, where we transform the original values of each feature to the quantiles they correspond to in the dataset (e. g., for *amazon.com*, page size: 38%, num imgs: 99%, num domains: 61%).

Fig. 18 reports the number of pages per cluster produced by K-means, ordered by decreasing cluster size. We observe that, due to the wide range of values for page size and number of objects in the dataset (up to 75 and 1,429Mb, respectively), both min-max and mean-std normalization create several “outlier clusters” near the extreme ranges of each feature with very few pages (less than 10), while creating a single overcrowded cluster for simple and small pages. We note that quantile-based normalization results in clusters that represent a sizeable number of pages (the smallest 5 clusters contain between 155 and 347 pages) while at the same time helps in better representing the fine-grain diversity of relatively small pages (there is no single giant cluster). Upon a closer analysis, we put aside 14 clusters that contained a large number of “error pages”. These cluster included regular HTML pages reporting 401, customized 404 pages, pages with valid HTML but no actual content, as well as login pages. We observe that the 5 largest clusters still represent 30% of all pages: therefore, a stratified selection strategy helps avoiding oversampling these pages.

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

Given the fair amount of work involved, this list of “*representative*” webpages is interesting per se, and we make it available [5]. Finally, we add 22 handpicked webpages that we also studied in previous work [39] to obtain a total of 108 sampled webpages.

4.2.2 Objective Web Quality Metrics

For each webpage of the set, we collect the objective Web quality metrics discussed in Section 3.2, notably the TTFP, the TTI, the Approximated Above-The-Fold (AATF) [39] and the PLT. We rely on a Chrome extension [9] to measure all the metrics, since some metrics require the rendered position of all objects in the page, cannot be measured from the HAR file (as opposed to PLT, DOM, TTI, *etc.*) and thus are better measured directly

from the browser. Further, we use FFmpeg to record videos of a webpage rendering process.

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

4.2.3 UPLT Crowdsourcing

We measure uPLT via Eyeorg’s *timeline* experiment [149] where a participant is asked to “scrub” the video of a webpage load until when (s)he considers the page to be ready. We run a single Eyeorg campaign targeting the above 108 webpages and 1,000 paid participants from Figure Eight¹ (total cost: \$120). Each participant evaluates 6 videos—thus generating 6,000 uPLT values or about 54 valid feedbacks per webpage, on average.

In Figure Eight, we request the highest quality participants. As discussed in the Eyeorg paper, we also filter user responses using a mix of their *engagement* (*i. e.*, the time spent on task) and the quality of their opinions using some control questions. Eyeorg implements control questions on top of the *frame selection helper*, a tool that helps the user “rewinding” her uPLT selection if an equal² (earlier) frame is identified. This is needed because, for some users, it can be hard to scrub a video exactly to the earliest point associated with a selected frame. For one video out of six, the frame selection helper suggests the very first video frame as a rewind option. Users that blindly accept this suggestion without noticing the obvious difference between the two frames are considered as potentially distracted, and their responses are discarded. In total, we discard 172 users due to low engagement and due to failing the control questions.

4.3 UNDERSTANDING USERS’ FEEDBACK

In order to provide an in-depth characterization of user feedback, we start our analysis by checking the existence of multiple modes on the uPLT distribution, considering for

¹ <https://www.figure-eight.com/>

² No more than 1% different in a pixel-by-pixel comparison.

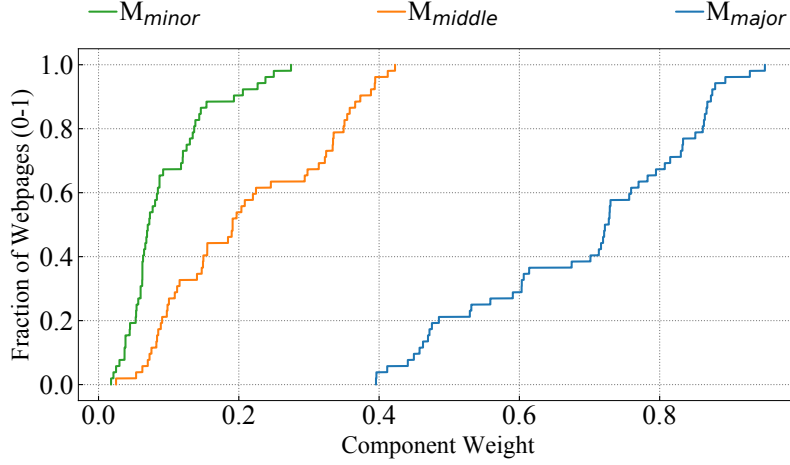


Figure 19: Component weights for pages with multi-modal uPLT.

each webpage the valid uPLT feedbacks. Then, we study the number and parameters of the different modes exhibited by the uPLT distribution for each webpage. Finally, we investigate how the complexity of modern webpages (*e. g.*, number of objects, domains, *etc.*) and user browsing behavior may affect uPLT multi-modality.

4.3.1 UPLT Distribution Analysis

We next analyze the uPLT distributions to inspect the presence of multi-modal behaviors. For this purpose, we rely on a non-parametric statistical test widely used to assess whether a distribution of real-valued random variables, such as the uPLT, is likely to be uni-modal [66]. This test computes the dip statistic as the maximum difference between the empirical cumulative distribution function (ecdf), and the uni-modal distribution function that minimizes that maximum difference. When we perform the dip test, we employ the common threshold $p < 0.05$ to reject the null hypothesis of uni-modality. We find on our set that 56 webpages are likely to exhibit a uni-modal distribution of uPLT and 52 a multi-modal one. By lowering this threshold, the number of likely multi-modal webpages decreases, *e. g.*, when $p < 0.01$ only 42 pages are estimated as multi-modal.

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

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

Table 7: Breakdown for $M_{\text{mass}} = M_{\text{time}}$.

We run the goodness-of-fit Kolmogorov-Smirnov test, with a confidence level of 0.95. The null hypothesis is that the empirical uPLT and the mixture distribution (which we sample to obtain the same number of samples as the empirical one) with $K = 3$ come from the same distribution. The result shows that, for more than 70% of the multi-modal webpages, the null hypothesis is confidently accepted. Hence, for each likely multi-modal webpage, we set $K = 3$ and find the corresponding GMM parameters from its uPLT distribution: mean, standard deviation, and weight of each component.

Fig. 19 shows the weights' distribution of the three components sorted by their mass, $M_{\text{mass}} = [M_{\text{major}}, M_{\text{middle}}, M_{\text{minor}}]$, across the 52 likely multi-modal webpages. We observe that the weight ranges from 0.40 to 0.95 for the major component (M_{major}), which represents on average 69% of users (median 72%), 0.03 to 0.43 for the middle (M_{middle}) component, and 0.02 to 0.27 for the minor component (M_{minor}). For some outlier webpages, such as *booking.com*, users are split into multiple well defined modes of similar size. In the opposite case, there are webpages such as *paperpile.com* where nearly all the users agree on a single uPLT value, with two other smaller modes ($M_{\text{major, mass}} = 0.86$, $M_{\text{middle, mass}} = 0.10$, $M_{\text{minor, mass}} = 0.04$).

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

Tab. 7 shows the percentage of occurrences for each of the 9 couples of $M_{\text{mass}} = M_{\text{time}}$ ($M_{\text{major}} = M_{\text{first}}$, $M_{\text{major}} = M_{\text{second}}$, etc.). This gives us information on which time sorted mode M_{time} is more liable to be the most or least popular one (M_{mass}). The table highlights that the majority of users are more likely to prefer the earliest modes: the major mode M_{major} is indeed equivalent to the earliest mode M_{first} on 33 pages (63% of the whole multi-modal webpages set), it is equal to the second

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

Table 8: Statistics of uni-modal/multi-modal pages.

one M_{second} on 17 pages (33%), and finally it is equivalent to the latest third mode M_{third} on just 2 pages (4%). Reversely, the minor mode M_{minor} tends to rarely coincide with the earliest one M_{first} (8%): it is actually most of the time equal to the latest mode M_{third} (63%) and sparingly to the second one M_{second} (29%). We can finally conclude that the mapping between mass and time sorted modes is such that the *most* popular mode is generally also the earliest in time and vice versa.

4.3.2 Page Characteristics and uPLT

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

Tab. 8 illustrates several statistics (average, standard deviation, 25/50/75th percentile) of the webpage characteristics we considered during our stratified URL selection (see Section 4.2). We can observe that the standard deviation of the size of multi-modal webpages is double with respect to that of uni-modal ones (at the 100% percentile we have 10,338 vs 3,460). We also observe that the mean number of images and of distinct origin domains for multi-modal webpages is respectively 34 and 11 compared to 20 and 7 for uni-modal webpages. This is inline with the intuition that complex webpages are more likely to be multi-modal. We next inspect how prevalent is advertising across these websites by matching the content received against EasyList,³ a list of known advertise-

³ <https://easylist.to/easylist/easylist.txt>

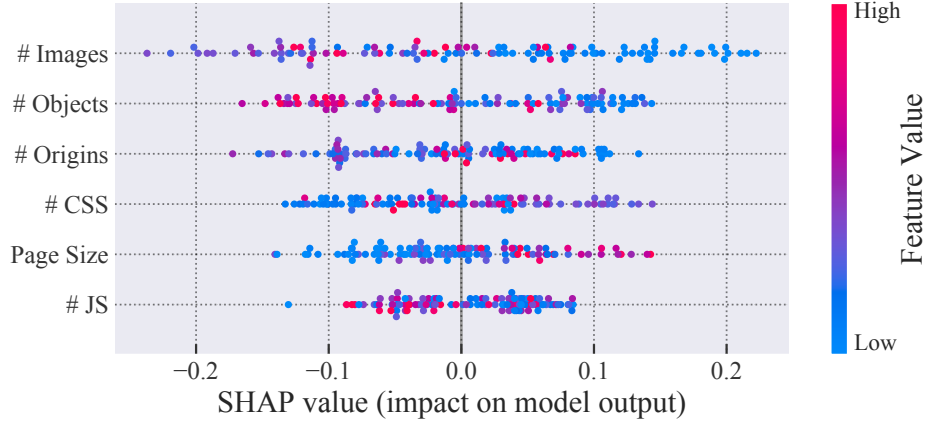


Figure 20: Ranking of the features according to their SHAP values when predicting uni-modal pages.

ment domains. We find that multi-modal websites contain, on average, 5 times more advertisements than uni-modal websites, likely segmenting user opinions on uPLT.

We are now interested in assessing the importance of each of the above features for predicting uPLT multi/uni-modality of webpages. For this task, we train a Random Forest Classifier with 25 estimators, which on a 7-fold cross validation⁴ achieves an average precision of 0.69 and an average recall of 0.68. In line with the current trend towards human interpretable machine learning and model explainability, we leverage SHAP (SHapley Additive exPlanations) [95] to understand which features can better reveal whether a webpage is *uni-modal* or not. We report in Fig. 20 the 6 features, sorted by the sum of the SHAP magnitude values computed for all the webpages. SHAP values capture the effect of removing a feature for a given prediction under all possible combinations of presence or absence of the other features. Hence, they provide a quantitative insight of the importance of each feature for the model. The positive x-axis values assess the impact on the model output for predicting the *uni-modal* class, whilst the negative ones refer to the *multi-modal* class. We can observe that the two most influential features are the number of images and the number of objects present in the webpage. In particular, the lower the values of these features, the higher their SHAP value (up to 0.2 for the number of images and 0.15 for number of objects). In other words, for simple webpages with few images and objects, users more likely agree on a single uPLT, making the uPLT distribution uni-modal. Such effect is less evident for the other webpage properties, where low and high feature values overlap, causing a decrease in the impact

⁴ Seven-fold cross-validation ensures that the validation dataset is at least 15% of the size of the whole dataset.

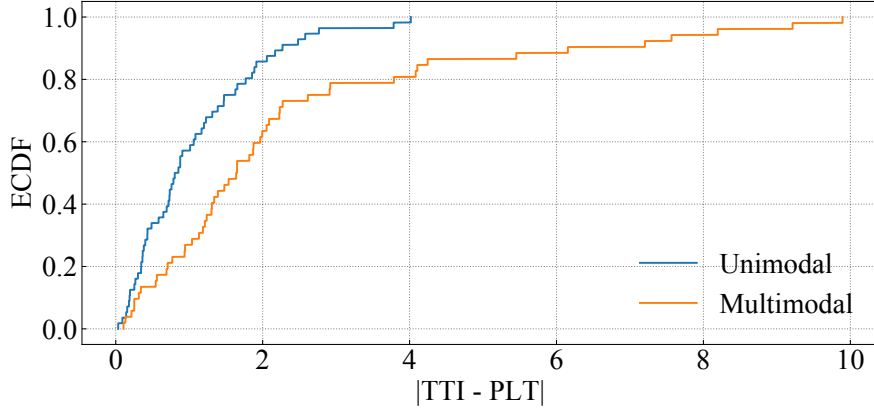


Figure 21: $|PLT - TTI|$ ECDF for uni/multi-modal pages.

factors on model prediction, probably due to the lack of additional data points to train the model. These findings provide valuable insights for designing webpages with more predictable user perception. For instance, we might expect that the uPLT measured via mobile browsers presents a unimodal distribution, as they generally load a simplified version of the webpage. We acknowledge that future studies are needed to further elaborate relevant design guidelines in this direction.

Finally, we quickly investigate if a difference in performance metrics can also explain the multi-modality of uPLT. We check whether the time difference between the early and late events (such as TTI and PLT) of the page loading process provides strong evidence of multi-modality. Fig. 21 shows the ecdf of $|PLT - TTI|$ for webpages we previously categorized as uni-modal or multi-modal. We can observe that multi-modal websites are, overall, characterized by larger $|PLT - TTI|$ differences compared to uni-modal ones. On the other hand, less than 10% of uni-modal pages had a $|PLT - TTI| > 2.2s$. This finding suggests that the rendering of these webpages naturally segment the users: some believe the page loaded as soon as the major part of the page loaded (usually closer to TTI) whereas others wait for all visible images to finish loading to consider the page fully loaded (usually closer to PLT).

4.3.3 Evaluation of Web Quality Metrics

Finally, we investigate to which extent single-valued objective Web quality metrics (see Section 3.2) can approximate the different modes of the uPLT distributions exhibited by the webpages of our study. For each of the 56 webpages showing a uni-modal uPLT behavior, Tab. 9 reports on the top the Root Mean Square Error (RMSE) of the mean of the uPLT distribution $\mu(uPLT)$ with respect to each of the following objective Web perfor-

$\text{RMSE}_{\mu, \text{Metric}}$	TTFP	TTI	AATF	PLT
$\mu(\text{uPLT})$	2.48s	2.35s	1.99s	1.48s

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

Table 9: RMSE of (top) uni-modal and (bottom) multi-modal uPLT with Web quality metrics.

mance metrics: TTFP, TTI, PLT and ATF, computed as Approximated Above-The-Fold (AATF) [39]. Results reveal that PLT is the metric which better approximates (lowest error term) the uPLT for webpages showing a uni-modal behavior of uPLT.

We rely on the weighted RMSE (wRMSE) to assess the quality of approximation of objective Web quality metrics for the 52 webpages with multi-modal uPLT (see the bottom part of Tab. 9). This approach weights the average towards larger components, which is particularly important for better evaluating the error on M_{middle} and M_{minor} . We conduct this analysis from three different perspectives: (i) we compare the wRMSE of the mean of the uPLT distribution $\mu(\text{uPLT})$, as we did for uni-modal webpages, (ii) we examine the three modes sorted by their mass $M_{\text{mass}} = [M_{\text{major}}, M_{\text{middle}}, M_{\text{minor}}]$ (M_{major} is the mode with the largest mass of the distribution), and (iii) by occurring time $M_{\text{time}} = [M_{\text{first}}, M_{\text{second}}, M_{\text{third}}]$ (M_{first} is the earliest).

The results summarized in Tab. 9 show that TTFP and TTI better approximate M_{major} and, not unexpectedly, given the duality shown in Sec. 4.3.1, M_{first} . On the other hand, AATF and PLT better approximate M_{middle} , M_{minor} , M_{second} and M_{third} . The former suggests that, to enhance the uPLT analysis, measuring and optimizing *the last* updates, usually achieved with PLT, is less relevant with respect to the earlier ones, *e.g.*, TTI and TTFP. The latter instead confirms that the users choosing a late uPLT agree on a page to be loaded close to the last two page tracking events. It is also an interesting validation to note that, on uni-modal pages, PLT better matches $\mu(\text{uPLT})$ whereas TTI does that for multi-modal ones.

Table 10: Summary of recent related work.

Year [ref]	Experiments scale	Measurement design	uPLT multi-modality	Metrics evaluation	uPLT modes analysis	Main focus
2012 [48]	n.a.	uPLT crowd-sourcing	No	Yes	No	WQL definition and demonstration
2013 [135]	n.a.	uPLT crowd-sourcing	No	Yes	No	Assessment and Models for Web QoE
2014 [142]	n.a.	n.a.	No	Yes	No	Web QoE overview
2016 [149]	1000 users, 100 webpages	uPLT crowd-sourcing	Yes	No	No	uPLT metric definition
2017 [81]	50 users, 45 webpages	uPLT crowd-sourcing	Yes	No	No	uPLT optimization by tracking user's eye gaze
2017 [54]	5.4k users, 115 webpages	A/B testing by showing side-by-side videos	No	Yes	No	Web browsing QoE assessment
2020 [This work]	1k users, 108 webpages	uPLT crowd-sourcing	Yes	Yes	Yes	uPLT multi-modality analysis and characterization

4.4 DISCUSSION

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

A summary of closely related work is reported in Tab. 10, where we distinguish for each study whether its authors identify or mention the multi-modal trait of uPLT (“*uPLT multi-modality*”) or if they analyze that the uPLT is insufficiently captured by single-valued objective metrics (“*Metrics evaluation*”). For the sake of comparison with our work, we report when available, the experimental settings and the size of the measurements (amount of users and webpages involved). Specifically, we note that previous studies [48, 135, 142] observe that uPLT does not match PLT, while Gao et al. [54] find that, more generally, “commonly used navigation metrics such as *onLoad* and *TTFB* fail to represent majority human perception”. We note that although these works remarked either the multi-modality of uPLT or the difficulty in mapping uPLT to single Web QoE metrics, their focus was not on characterizing the uPLT multi-modal nature. This confirms

that the main hypothesis of our work is in line with the recent empirical observations in Web QoE modeling.

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

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

4.5 CONCLUSIONS

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

selection, which is interesting per se, that ensures our 100 target pages cover the initial 1M set) and repeatable (for which we have already open sourced our dataset [5]).

Whereas this work is far from entirely closing the Web QoE measurement issue, we hope that open sourcing our dataset [5] can help the community into further nailing down the smallest set of relevant Web QoE metrics covering *all* user modes, as opposite to attempting to define yet another *single* Web QoE indicator, that would by definition fail in this task.

Contents

5.1	Is the Web diverse enough?: on the fairness of language models . .	61
5.2	Background and related work	63
5.3	Methodology	64
5.3.1	Corpora for Spoken English	64
5.3.2	Bias in Masked Language Modeling	66
5.4	Quantifying the Bias	68
5.4.1	Measuring the Bias of LMs	68
5.4.2	Bias on AAE Features	72
5.4.3	Bias on Part-of-Speech	75
5.5	Conclusions	75

5.1 IS THE WEB DIVERSE ENOUGH?: ON THE FAIRNESS OF LANGUAGE MODELS

Since their inception [43], transformers-based bidirectional encoder representations language models (LMs) gained lots of scientific interest due to their sizable improvements on a wide range of Natural Language Processing (NLP) tasks. The success of BERT pushed researchers to expand the state-of-the-art by introducing a plethora of model variants with differences in the architecture [124], the size [87, 133, 157] and the training [88, 91]. This resulted in a growing concern of the research community to discuss the potential risks coming from the pervasive adoption of these models [17]. Indeed, several studies highlight that this would hinder an equitable and inclusive access to NLP technologies and have real-world negative consequences in different areas, as education, work and politics [136]. In this context, given the consistent emergence of new LMs trained on Web-based corpora, it is crucial to define to which extent such models are fair and not instead prone to bias.

Actually, given the sheer size and heterogeneity of the Web, one could expect these models to be bias-free. However, already before the explosion of transformer-based LMs, a variety of biases have been identified in standard word embeddings [19, 22]. Recently, some effort has been devoted to highlight the presence of possible biases encoded by

transformer-based LMs along gender, race, ethnicity, and disability status. Yet, whereas the study of such biases is commonly tackled via sentiment analysis and named entity recognition tasks, in this work we take a different approach. Inspired by the frequent scenario occurring in conversational systems, where a word could be unheard or unrecognized by the Automatic Speech Recognition system and would therefore need to be predicted, we measure how token predictions change based on their context. Similar to prior work [86] considering social biases in BERT, we assess the bias for a target token by directly querying the underlying Masked Language model.

In this work we focus on the study of potential bias towards English dialects spoken by underrepresented and historically discriminated groups, such as African American English (AAE). Particularly, AAE slightly differs from *mainstream* English, also known as Standard American English (SAE). In linguistics, these two variants are regarded as two different languages because highly structured with their own phonological, syntactic and morphological rules [62]. However, SAE speakers often believe that AAE is a version of SAE with mistakes and that AAE speakers belong to deficient cultures [122, 154]. While, instead, AAE highlights the regional, societal and cultural environments in which individuals have learned to speak [61].

It is difficult to estimate the number of AAE speakers, since some African Americans may speak a variety that aligns more with SAE and besides, not all AAE speakers are African Americans. Nevertheless, a 2019 census [123] estimates that approximately 13% of the U.S. population is currently African American. This suggests that the fraction of population speaking AAE could be large. Hence, the presence of potential linguistic biases would have discriminatory consequences towards a considerable group of individuals.

For these reasons, we set out to measure the robustness and the quality of 7 transformer-based LMs in the prediction of *missed* words when the input is either SAE or AAE. Here we question whether the heterogeneity of Web content, over which the LMs have been trained, guarantees diversity. We resort to two renowned corpora of spoken SAE and AAE and evaluate the LMs in a Masked Language Modeling (MLM) task. This is a *fill-in-the-blank* task, where we mask and predict a token simulating its absence in every utterance. We next define two metrics to compare the likelihood that the model assigns to the predicted token and to the actual *masked* one, that we use as a proxy of quality and fairness for the model itself.

Specifically, we rigorously quantify the model bias and find that BERT, in both its cased and uncased variants, exposes a non-negligible bias towards SAE (up to 21% more accurate results with respect to AAE). Surprisingly we find this bias to be reversed for RoBERTa and BART models. We additionally observe distilled variants of these LMs to be fairer with respect to their teachers. Finally, our analysis reveals how most of the bias

resides in the AAE structural differences, and identifies the particles, the pronouns and the adpositions as principal parts of speech sources of bias.

This chapter is organized as follows. After overviewing the related work (Section 5.2), we present the corpora and the methodology to measure of the bias (Section 5.3). In Section 5.4 we show the results of the LMs when operated to predict the tokens from AAE and SAE, illustrating that the metrics we define reveal a bias. Finally, we discuss our findings and possible directions (Section 5.5).

5.2 BACKGROUND AND RELATED WORK

The success of transformer-based LMs is down to several factors, among which it is worth mentioning the large architectures and the training done on huge amounts of textual data. Moreover, recently a special effort has been devoted to reducing the size of large LMs, *e. g.*, BERT [43] and RoBERTa [91], by means of various compression techniques as knowledge distillation or quantization. This made possible the emergence of much smaller, but no less accurate, LMs as DistilBERT [133] and DistilRoBERTa [133].

Nevertheless, the emergence and massive spread of all these models raised the interest of the research community towards the potential societal risks linked to the employment of these models for either generating text tasks or as components of classification systems [17]. These works have studied the effects of transferring the stereotypical associations present in the training datasets to LMs, which cause unintended bias towards underrepresented groups. Recent work [160] studies differences in performance for BERT, showing that it often favors the majority group with regards to gender, language, ethnicity, and insurance status, whereas [144] finds racial bias encoded in different models. A significant research effort has been done to show race and gender bias embedded in large models [14, 34, 86, 101, 137]. Authors in [74] highlight instead the presence of topical biases in the words predicted by BERT on sentences mentioning disabilities.

In addition to bias measurement works, researchers have proposed methods to mitigate societal biases with debiasing techniques [80, 89, 148]. As for the bias towards languages, most studies have focused on offensive language and hate speech detection [42, 107, 108], while assessing the bias against dialects spoken by underrepresented groups is quite recent [44]. Whereas the above works mostly focus on the negative sentiment and stereotypical associations towards specific groups in BERT [43], in this work we quantify the linguistic bias towards AAE for 7 different LMs: BERT, RoBERTa, BART [88], DistilBERT and DistilRoBERTa, including both their cased and uncased versions.

These works have proven that the large dimension of the training datasets for state-of-the-art LMs is not synonymous of diversity and, as a consequence, of inclusion [17].

Therefore, in this regard, our analysis is essential to provide a framework to assess, reveal and counteract the existing biases, which we hope will contribute in enriching the scientific community knowledge on this domain.

5.3 METHODOLOGY

Spoken language tends to have incomplete sentences, spontaneous self-corrections and interruptions and its register is more of an informal one, while written language is typically more structured and pre-planned. The features of spoken language are of particular interest for studying conversational systems.

Hence, to capture and provide an accurate and comprehensive account of societal biases embedded in state-of-the-art LMs, we leverage two corpora of spoken English. These are widely used by the linguists because considered a fair representation of their spoken language. We note that, while this work is not the first in studying the presence of societal biases, to the best of our knowledge, this is the first to provide a thorough characterization of it for AAE, across different models tested on a MLM task. In this case, the use of the MLM task would replicate the often occurring scenario where a word could be unrecognized by the Automatic Speech Recognition system which would normally have to infer it given the surrounding context. We summarize LMs performance by means of statistical metrics, which are used to characterize both the bias and the quality of the models.

5.3.1 *Corpora for Spoken English*

For SAE, we leverage the Santa Barbara Corpus of Spoken American English (SBCSAE) [45], which has been already widely adopted for different applications, as the assessment of political risk faced by U.S. firms [67], the measure of grammatical convergence in bilingual individuals [30] and the exploration of new-topic utterances in naturally occurring dialogues [97].

The SBCSAE is the only existing large-scale corpus of naturally occurring spoken interactions from people with different regional origins in USA. It includes conversations from a wide variety of people, differing in gender, occupation and social background, recorded in various real everyday life situations. All the audio recordings are complemented with their transcribed counterparts, which are the ones we use in this work.

The fact that SBCSAE consists of speakers from several regional origins prevents us from crafting the results and unintentionally inducing a bias by comparing AAE with an *academic* version of SAE, which is instead rather different from the commonly spoken

Corpus	Language	$ \mathcal{U} $	$\langle \ell_{\mathcal{U}} \rangle$	L	$ \mathcal{T} $
<i>Original</i>					
CORAAL	AAE	90,493	6.22	563,037	17,214
SBCSAE	SAE	40,838	7.14	291,513	12,324
<i>Preprocessed</i>					
CORAAL	AAE	63,814	8.23	525,067	16,352
SBCSAE	SAE	25,113	8.38	210,430	10,540

Table 11: Corpora summary: with and without filtering utterances (\mathcal{U}) based on their length. With $\langle \ell_{\mathcal{U}} \rangle$ we indicate the average utterance length; with L, the length of the corpus in number of words, and; with $|\mathcal{T}|$, the number of terms (unique words).

English and, hence, far from the purpose of this work. Therefore, we filter out Hispanic and African American speakers (1092 AAE utterances, a negligible number with respect to the size of the corpus) and obtain a corpus of SAE language.

For AAE, we leverage the Corpus of Regional African American Language (CORAAL) [82], which also provides the audio recordings along with their time-aligned orthographic transcription, of particular interest for this work. CORAAL includes 150 sociolinguistic interviews for over a million words. It is periodically updated and is the only publicly available corpus of AAE. As such, it has been used in literature for a plethora of tasks, ranging from dialect specific speech recognition [44] to cross-language transfer learning [71].

In this work, we only focus on the CORAAL:DCB portion, since it is the one comprising the most recent interviews (carried out between 2015 and 2017) and the largest amount of data (more than 500k words). It includes conversations from 48 speakers raised in Washington DC, a city with a long-standing African American population.

For each corpus we define $\mathcal{U} = \{u_1, u_2, \dots, u_n\}$ as the set of all the available utterances, and $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ as the set of all terms (unique words). Since we perform an utterance-level analysis, we first filter out noise. Particularly, we discard both short utterances (composed by just one or two words) and very long ones (greater than 50 words). Therefore, we only keep utterances having a number of words ranging from 3 to 50.

Fig. 22 shows the empirical cumulative distribution function (ECDF) of the utterance length $\ell_{\mathcal{U}}$ for both AAE and SAE corpora. We can see that the two distributions are almost equal, therefore our utterance-level analysis does not introduce any bias.

In Tab. 11 we report a terse summary of the corpora statistics, both before and after having applied the filtering based on the utterance length. Even though the sizes of the two datasets are very different, not only in terms of number of utterances $|\mathcal{U}|$, but also

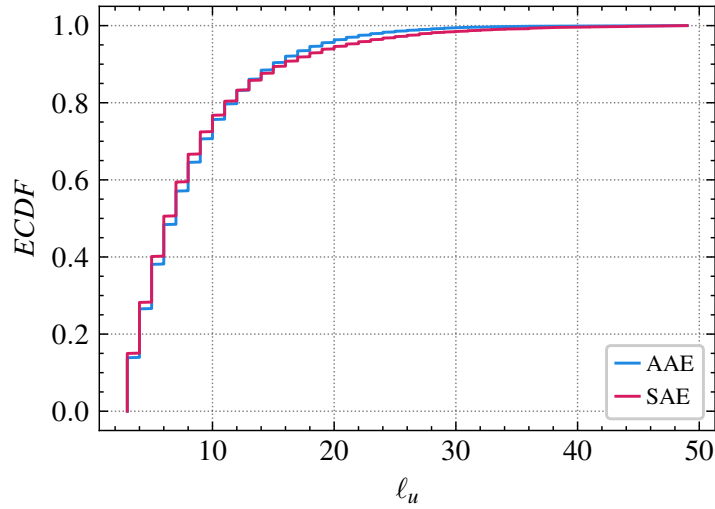


Figure 22: The ECDF of the utterance length ℓ_u for both AAE and SAE corpora.

Model	Training Data
BERT, DistilBERT	BOOKSCORPUS and English Wikipedia (16GB)
RoBERTa, BART	BERT data + CC-NEWS, OPENWEBTEXT and STORIES (160GB)
DistilRoBERTa	OPENWEBTEXT (38GB)

Table 12: Training data for the tested LMs.

in terms of total number of words L and terms $|\mathcal{T}|$, we can see that, after the filtering, the average utterance length $\langle \ell_u \rangle$ is very similar (~ 8 words per utterance).

5.3.2 Bias in Masked Language Modeling

In order to measure the bias in LMs we perform a MLM task. We leverage the transformer-based $\text{BERT}_{\text{base}}$ LM [43] and its recent variants, including $\text{DistilBERT}_{\text{base}}$ [133], in both their cased and uncased flavors, $\text{RoBERTa}_{\text{base}}$ [91], $\text{DistilRoBERTa}_{\text{base}}$ and $\text{BART}_{\text{base}}$ [88]. These LMs have all been pre-trained using a MLM objective, which consists in randomly masking 15% of the tokens using a special [MASK] token. Note that these models are trained on different corpora, summarized in Tab. 12. In practice, they have been instructed to predict a masked token, referred to as a [MASK], given the surrounding context of the sentence.

Therefore, by directly querying the underlying MLM in each LM, we simulate the typical scenario where a conversational system has to infer a *missed* word in an utterance.

original utterance (u)	And I be okay with it .
u with w_1 masked	[MASK] I be okay with it .
u with w_2 masked	And [MASK] be okay with it .
	...
u with w_7 masked	And I be okay with it [MASK]

Table 13: Example showing the masked token experiment.

Specifically, we encode each utterance of the two corpora with the *tokenizer* of the LM considered, then, in turn, we mask each word w_{mask} and finally predict it by feeding the model with only a context of 10 tokens surrounding the masked one w_{mask} . Tab. 13 shows an example, illustrating how the experiment is carried on: (i) we let the LM encode the original utterance \mathbf{u} (the one reported in the table has a length lower than 10 tokens so there is no need for the window), (ii) we mask and predict the first token w_1 , (iii) we iteratively repeat this process until the last token of the utterance is masked.

The LM provides for each run a list of possible terms to *fill-in-the-blank*. In this vocabulary set (\mathcal{T}) we select the predicted term t_p having the highest probability $P(t_p|c)$ and, as such, ranking first in the list $\rho(t_p|c) = 1$, where c is the context surrounding t_p and ρ is the rank of $t|c$ provided by the model. In this notation, a word w is a term t in a context c ($t|c$). We next retrieve from the vocabulary of possible terms \mathcal{T} the corresponding probability $P(t_m|c)$ and the rank $\rho(t_m|c)$ for the actual masked token t_m . The latter provides a measure of how likely the LM will choose t_m as a candidate token to replace the masked one w_{mask} . It is then natural to employ the probabilities difference $\Delta P(t|c)$ as a proxy of the quality of the prediction for a single token, so defined:

$$\Delta P(t|c) = P(t_p|c) - P(t_m|c) = \Delta P(w). \quad (1)$$

We further define for each token $t|c$ the Complementary Reciprocal Rank (CRR) as:

$$\text{CRR}(t|c) = 1 - \rho(t_m|c)^{-1} = \text{CRR}(w). \quad (2)$$

Note that this is the difference between the reciprocal rank (RR) of the predicted token, which is always equal to 1 ($\rho(t_p|c)^{-1} = 1$), and the RR of the masked token.

We then define the probability difference for an utterance by averaging the probability difference for each token in the utterance:

$$\Delta P(\mathbf{u}) = \frac{1}{\ell_{\mathbf{u}}} \sum_{w \in \mathbf{u}} \Delta P(w), \quad (3)$$

with ℓ_u being the length of the utterance in terms of tokens. Similarly, we define the CRR for an utterance as:

$$\text{CRR}(u) = \frac{1}{\ell_u} \sum_{w \in u} \text{CRR}(w). \quad (4)$$

Note that the metrics based on the ranks $\rho(t|c)$ generated by the LMs are necessary to fully capture the bias embedded in the models, as the $\Delta P(t|c)$ alone could be insufficient. This because, the $\Delta P(t|c)$ strongly depends on how the LM assigns the probability. Indeed, the probability distribution of $P(t|c)$ could be more uniform, and consequently would lead, on average, to a smaller $\Delta P(t|c)$, or more skewed, causing instead larger differences $\Delta P(t|c)$. Instead, this effect is not present in CRR that remains unaffected by such differences in the output probability distribution of $P(t|c)$.

5.4 QUANTIFYING THE BIAS

In this section, we first provide an accurate overview of the measured LMs fairness, and then further analyze the discovered biases from different viewpoints. We show how they varies when we take into account the syntactical, grammatical, and lexical patterns typical of AAE language first, and then, when we slice the corpus based on parts of speech.

5.4.1 *Measuring the Bias of LMs*

As described in Section 5.3, we test the fairness of transformer-based LMs by running experiments in a MLM setting. As aforementioned, we use ΔP and CRR as metrics for measuring the quality and the fairness of the models towards the two investigated languages. We are interested in observing the expected behavior of the LMs with respect to each utterance, therefore we consider an aggregate measure of the metrics on a per-utterance level.

Tab. 14 reports an overview of the results of $\Delta P(u)$ and $\text{CRR}(u)$. After having assessed that the difference between the means of AAE and SAE for both $\Delta P(u)$ and $\text{CRR}(u)$ with a Welch’s t-test [152] is significant (p-value < 0.05), we measure their effect size using the Cohen’s d [37]. This is reported in the last two columns of Tab. 14. According to Cohen’s classification there is a *small* effect for both the metrics, and a *medium* effect for BART on $\Delta P(u)$ ($d > 0.5$).

Model	MAE								MSE							
	$\Delta P(u)$				CRR(u)				$\Delta P(u)$				CRR(u)			
	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d
BERT _{cased}	0.217	0.171	21 †	0.417	0.497	0.441	11 †	0.272	0.060	0.040	33 †	0.345	0.289	0.233	20 †	0.262
BERT _{uncased}	0.242	0.198	18 †	0.352	0.494	0.446	10 †	0.232	0.074	0.053	29 †	0.297	0.288	0.238	18 †	0.230
DistilBERT _{cased}	0.113	0.108	5 †	0.081	0.627	0.589	6 †	0.188	0.017	0.016	2 †	0.015	0.436	0.385	12 †	0.203
DistilBERT _{uncased}	0.126	0.118	6 †	0.104	0.578	0.530	8 †	0.222	0.021	0.020	1	0.007	0.380	0.325	15 †	0.223
RoBERTa	0.223	0.261	-15 †	0.368	0.536	0.592	-9 †	0.252	0.061	0.079	-23 †	0.311	0.337	0.396	-15 †	0.225
DistilRoBERTa	0.143	0.153	-7 †	0.137	0.644	0.668	-4 †	0.117	0.026	0.029	-11 †	0.112	0.457	0.487	-6 †	0.115
BART	0.156	0.193	-20 †	0.506	0.613	0.682	-10 †	0.346	0.030	0.043	-31 †	0.447	0.418	0.501	-17 †	0.328

Table 14: MAE and MSE of $\Delta P(u)$ and $CRR(u)$ measured on AAE and SAE corpora: results obtained through the *fill-in-the-blank* task with different language models. † signifies that the AAE and SAE expectations are statistically significant according to the Welch’s two-tailed t-test (p-value < 0.05). The column d contains their effect size computed according to the Cohen’s d.

We summarize the quality of the prediction in the corpora by means of two error measures. We report the Mean Absolute Error (MAE) for each of the two distributions:

$$MAE(\Delta P(u)) = \frac{1}{|u|} \sum_{u \in u} |\Delta P(u)|, \quad (5)$$

$$MAE(CRR(u)) = \frac{1}{|u|} \sum_{u \in u} |CRR(u)|. \quad (6)$$

We also report the Mean Squared Error (MSE), defined as:

$$MSE(\Delta P(u)) = \frac{1}{|u|} \sum_{u \in u} \Delta P(u)^2, \quad (7)$$

$$MSE(CRR(u)) = \frac{1}{|u|} \sum_{u \in u} CRR(u)^2. \quad (8)$$

Indeed, these error measures can be used to quantify the quality of the predicted terms. MAE and MSE closer to 0 correspond to an utterance having more accurately predicted terms. Therefore, in Tab. 14 we highlight the values leading to the smallest error between AAE and SAE. We additionally emphasize the presence of bias by pointing out the percentage of bias change of each LM $\Delta[\%]$. This is always calculated with respect to the model with the largest bias, and when positive the model is biased towards SAE, *vice versa* otherwise.

Three main patterns clearly emerge from Tab. 14. First, BERT and DistilBERT, in both their cased and uncased variants, show a bias towards SAE for all the metrics. Specifically, BERT not only presents a non-negligible bias against AAE but also it is the LM

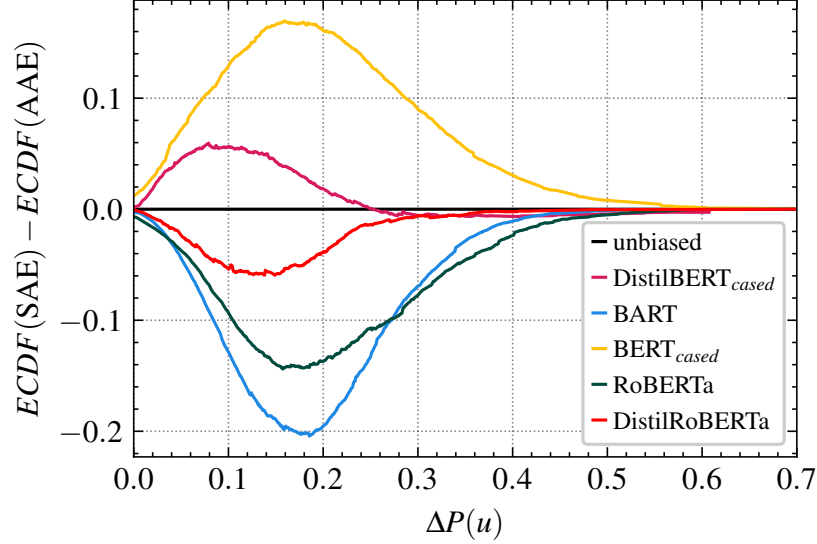


Figure 23: The difference between the ECDFs of SAE and AAE for the $\Delta P(u)$ measure. When the values are greater than zero the LMs are more biased towards SAE, *vice versa* otherwise.

which leads to the highest relative bias. Specifically, notice that the $\text{MAE}(\Delta P(u))$ for SAE is more than 20% lower than AAE, 11% lower for the $\text{MAE}(\text{CRR}(u))$, 33% for the $\text{MSE}(\Delta P(u))$ and 20% for the $\text{MSE}(\text{CRR}(u))$.

Second, DistilBERT, in both its cased and uncased flavors, and DistilRoBERTa, are the models which perform better as regards the average probability difference $\Delta P(u)$. This is true both in terms of MAE and MSE, which are approximately half and one third of the other LMs. On the one hand, this could seem somewhat unexpected since, one could argue that DistilBERT is less accurate than BERT, achieving only 97% of its performance [133]. On the other hand, this is in line with recent work [17] reporting that such LMs sometimes exceed the performance of the original ones. However, as mentioned in Sec. 5.3, it is crucial to also look at the $\text{CRR}(u)$, since a better behavior in terms of $\Delta P(u)$ could in practice just be tied to the fact that the model generates more uniformly distributed probabilities $P(t|c)$ with respect to the others.

Finally, we observe that BART, despite leading to a decent quality of the prediction for AAE ($\text{MAE}(\Delta P(u))$ and $\text{MSE}(\Delta P(u))$ are lower than BERT), shows an opposite trend with respect to BERT and DistilBERT. This reverse unexpected bias towards AAE is also introduced by RoBERTa and DistilRoBERTa. This is somewhat surprising and could probably be ascribable to the type of datasets they have been trained on. Indeed, as shown in Tab. 12, RoBERTa and BART are pre-trained with 1000% more data than BERT. Particularly, by delving into the type of data involved, we discover multiple sources, ranging from English language encyclopedia and literary works (same as BERT), to news articles and Web content. Specifically, RoBERTa, BART and Distil-

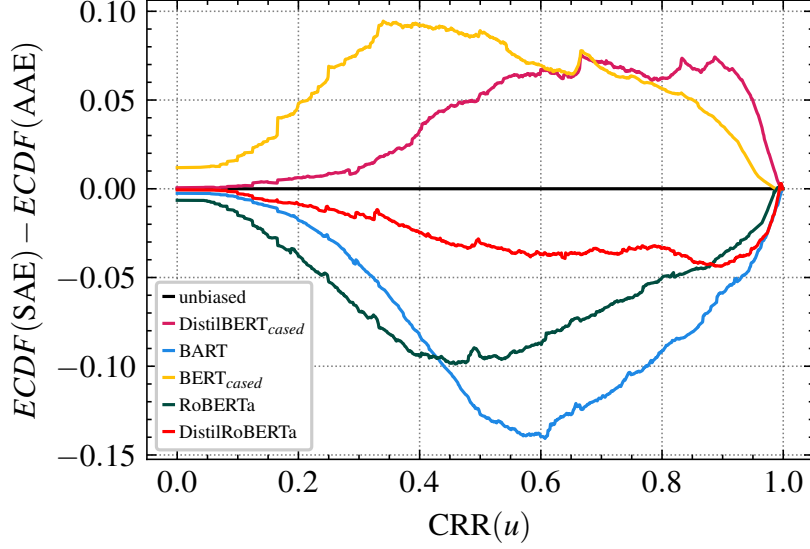


Figure 24: The difference between the ECDFs of both AAE and SAE for the $CRR(u)$ measure. When the values are greater than zero the LMs are more biased towards SAE, *vice versa* otherwise.

RoBERTa leverage OPENWEBTEXT [58], a corpus which includes filtered Web content obtained by scraping the social media platform Reddit, possibly exposing the LMs to a less *standard* American English.

Since Tab. 14 reports only a summary of the distributions of the bias metrics computed on both the datasets, for a better understanding, we show in Fig. 23 the bias measured by subtracting the empirical cumulative distribution functions (ECDFs) of $\Delta P(u)$ of AAE to that of SAE. This figure includes the bias measured for the LMs, reporting, for the sake of simplicity, for BERT and DistilBERT only their *cased* variants. The solid black line at $y = 0$ shows the optimal unbiased LM and, hence, visually separates what is biased against AAE (on the positive y -axis) from what instead is biased against SAE (on the negative y -axis). In this way, we clearly see the behaviors of the LMs leading to the two worst biases, *i. e.*, RoBERTa and BERT_{cased}: they are consistently biased towards one side (BERT_{cased} is always positive, whilst RoBERTa is instead always negative). They both present the maximum bias when $\Delta P(u)$ is close to 0.2 and instead mitigate for larger values.

A similar behavior is observed for the $CRR(u)$. Fig. 24 shows the bias measured by subtracting the empirical cumulative distribution functions (ECDFs) of $CRR(u)$ of AAE to that of SAE. The trend of the models is similar to that shown for $\Delta P(u)$: BERT_{cased} and DistilBERT_{cased} exhibit a consistent bias towards SAE, whilst, on the contrary, RoBERTa, DistilRoBERTa and BART are steadily biased towards AAE.

Original	Translated
Double Negative (0.7%)	
<ul style="list-style-type: none"> • <i>You don't need nothing but you.</i> • <i>I wasn't no lifeguard cause I couldn't swim.</i> • <i>Don't never try to chase another person happiness.</i> • <i>I don't know nobody over there no more.</i> 	<ul style="list-style-type: none"> • <i>You don't need anything but you.</i> • <i>I wasn't a lifeguard because I couldn't swim.</i> • <i>Never try to chase another person's happiness.</i> • <i>I don't know anyone over there anymore.</i>
Copula <i>be</i> (2.8%)	
<ul style="list-style-type: none"> • <i>And I be okay with it.</i> • <i>It depends on where you going to.</i> • <i>All of my friends was from like DC.</i> • <i>Okay, we having a baby.</i> 	<ul style="list-style-type: none"> • <i>And I am okay with it.</i> • <i>It depends on where you are going to.</i> • <i>All of my friends were from DC.</i> • <i>Okay, we are having a baby.</i>
Contractions (4.6%)	
<ul style="list-style-type: none"> • <i>I'm'a ask you.</i> • <i>I ain't coming home.</i> • <i>something gonna happen.</i> • <i>you gonna be there for a couple of hours.</i> 	<ul style="list-style-type: none"> • <i>I'm going to ask you.</i> • <i>I'm not coming home.</i> • <i>something is going to happen.</i> • <i>you will be there for a couple of hours.</i>

Table 15: A sample of AAE utterances selected based on their syntactical features and their translations to SAE. In brackets the prevalence of the feature over the utterances in the AAE corpus.

5.4.2 Bias on AAE Features

We next investigate how results change when we acknowledge the lexical, syntactical, morphological and also phonological rules of AAE. Following AAE grammar [63], we choose to focus on three major syntactical features: (i) the use of *double* negatives, (ii) the different usage of copula *be* and, finally, (iii) the contractions of words and groups of words.

As for (i), we search for the close presence of multiple forms of grammatical negation (which in Standard English are instead understood to resolve to a positive) in all the utterances of the AAE corpus, and find, that 0.7% of the utterances contains such a feature. Concerning (ii), we select the AAE utterances exhibiting the use of the *aspectual be* verb, typically used to denote habitual or iterative meaning (e. g., *I be okay with it* in Tab. 15). Additionally, we also filter on utterances with the verb tense in the *-ing* form where the copula is either omitted (e. g., *It depends on where you going to* in Tab. 15) or left at the base form (e. g., *they be getting mad* in Tab. 15), for a total of 2.8% of utterances. Finally, for (iii) we include those utterances containing not-standard contractions,

Model	MAE				MSE			
	$\Delta P(u)$		CRR(u)		$\Delta P(u)$		CRR(u)	
	AAE	AAE ^T $\Delta[\%]$	d		AAE	AAE ^T $\Delta[\%]$	d	
Double Negative [50 utterances]								
BERT _{cased}	0.202	0.159	21 \uparrow 0.591	0.391	0.334	15 \uparrow 0.493	0.046	0.030 34 \uparrow 0.526
BERT _{uncased}	0.216	0.187	14 0.358	0.404	0.340	16 \uparrow 0.503	0.053	0.041 23 0.319
DistilBERT _{cased}	0.137	0.106	22 \uparrow 0.548	0.506	0.441	13 \uparrow 0.523	0.022	0.014 37 \uparrow 0.504
DistilBERT _{uncased}	0.148	0.117	21 \uparrow 0.485	0.479	0.394	18 \uparrow 0.701	0.025	0.018 27 0.293
RoBERTa	0.202	0.181	10 0.227	0.434	0.383	12 0.328	0.048	0.042 14 0.180
DistilRoBERTa	0.170	0.134	21 \uparrow 0.572	0.581	0.498	14 \uparrow 0.628	0.034	0.020 41 \uparrow 0.567
BART	0.164	0.140	15 \uparrow 0.422	0.534	0.471	12 \uparrow 0.469	0.030	0.023 22 0.368
Copula <i>be</i> [50 utterances]								
BERT _{cased}	0.252	0.184	27 \uparrow 0.691	0.589	0.408	31 \uparrow 1.142	0.074	0.043 42 \uparrow 0.622
BERT _{uncased}	0.287	0.216	25 \uparrow 0.642	0.595	0.417	30 \uparrow 1.009	0.094	0.059 37 \uparrow 0.520
DistilBERT _{cased}	0.134	0.119	11 0.273	0.703	0.540	23 \uparrow 0.910	0.021	0.017 16 0.198
DistilBERT _{uncased}	0.138	0.118	14 \uparrow 0.339	0.678	0.513	24 \uparrow 0.904	0.022	0.017 25 0.344
RoBERTa	0.246	0.211	14 \uparrow 0.403	0.609	0.458	25 \uparrow 0.800	0.069	0.051 26 0.380
DistilRoBERTa	0.169	0.142	16 \uparrow 0.425	0.723	0.554	23 \uparrow 0.947	0.032	0.024 25 0.389
BART	0.161	0.144	11 0.305	0.672	0.556	17 \uparrow 0.672	0.029	0.024 18 0.246
Contractions [50 utterances]								
BERT _{cased}	0.225	0.181	19 \uparrow 0.507	0.470	0.347	26 \uparrow 0.848	0.058	0.040 32 \uparrow 0.436
BERT _{uncased}	0.258	0.205	21 \uparrow 0.605	0.482	0.355	26 \uparrow 0.880	0.075	0.049 34 \uparrow 0.541
DistilBERT _{cased}	0.135	0.114	16 0.381	0.584	0.463	21 \uparrow 0.746	0.022	0.016 28 0.316
DistilBERT _{uncased}	0.140	0.113	19 \uparrow 0.477	0.538	0.410	24 \uparrow 0.799	0.023	0.016 33 0.374
RoBERTa	0.215	0.193	10 0.264	0.500	0.402	20 \uparrow 0.584	0.054	0.043 20 0.242
DistilRoBERTa	0.154	0.130	16 \uparrow 0.436	0.601	0.488	19 \uparrow 0.668	0.027	0.020 28 \uparrow 0.411
BART	0.143	0.136	5 0.117	0.567	0.475	16 \uparrow 0.562	0.023	0.023 1 0.015

Table 16: Similar to Table. 14 but calculated over a sample of 50 utterances of AAE and their translated version (AAE^T) for each feature of AAE.

e. g., I'm'a, ain't or omitting the auxiliary before *gonna*, *e. g., something gonna happen* in Tab. 15. We do not include contractions which are popular in SAE, as *wanna*, *won't*, *aren't*, *etc.* We obtain 4.6% of the utterances in this class. After having properly filtered the utterances corresponding to the specific grammar patterns, we carefully manually validate our selection, by random picking and inspecting 1% of them. We check that the 1% random sampled utterances are actually satisfying the criteria we were looking for. From this manual labeling we double check our selection strategies based on syntactical rules and find that for both the 3 cases these are 99% accurate.

Next, we randomly choose 50 utterances from each AAE case and build a ground truth by *translating* the AAE utterances into a version compliant to SAE, that we define as AAE^T. We keep the translation process as neutral as possible, by preserving the standard officially recognized contractions and by only *adjusting* the selected grammar rules. Tab. 15 reports some examples of the utterances extracted from each AAE grammar case bucket and the corresponding translated ones.

Finally, we repeat the MLM experiments, as described in Section 5.3, on these 150 translated utterances AAE^T and measure the bias. We report the results in Tab. 16. Ac-

Model	MAE								MSE							
	$\Delta P(t)$				CRR(t)				$\Delta P(t)$				CRR(t)			
	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d
Particles [16k (AAE) 6k (SAE) terms], $\sum \Delta[\%] = 66$																
BERT _{cased}	0.113	0.081	29 \uparrow	0.143	0.212	0.192	9 \uparrow	0.054	0.071	0.040	44 \uparrow	0.176	0.177	0.157	11 \uparrow	0.062
BERT _{uncased}	0.126	0.090	29 \uparrow	0.145	0.212	0.191	10 \uparrow	0.059	0.086	0.049	44 \uparrow	0.182	0.176	0.155	12 \uparrow	0.066
DistilBERT _{cased}	0.079	0.070	12 \uparrow	0.062	0.328	0.310	6 \uparrow	0.045	0.030	0.025	16 \uparrow	0.051	0.272	0.259	5 \uparrow	0.036
DistilBERT _{uncased}	0.090	0.073	19 \uparrow	0.099	0.313	0.294	6 \uparrow	0.046	0.040	0.028	29 \uparrow	0.102	0.262	0.248	5 \uparrow	0.038
RoBERTa	0.101	0.114	-11 \uparrow	0.058	0.208	0.238	-13 \uparrow	0.083	0.056	0.062	-10 \uparrow	0.039	0.168	0.194	-13 \uparrow	0.081
DistilRoBERTa	0.099	0.108	-8 \uparrow	0.049	0.354	0.369	-4 \uparrow	0.037	0.043	0.047	-9 \uparrow	0.035	0.300	0.312	-4 \uparrow	0.032
BART	0.079	0.102	-23 \uparrow	0.139	0.261	0.317	-18 \uparrow	0.144	0.032	0.043	-26 \uparrow	0.107	0.212	0.261	-19 \uparrow	0.142
Pronouns [84k (AAE) 31k (SAE) terms], $\sum \Delta[\%] = 59$																
BERT _{cased}	0.182	0.186	-2 \uparrow	0.017	0.349	0.379	-8 \uparrow	0.078	0.101	0.098	3 \uparrow	0.017	0.268	0.288	-7 \uparrow	0.062
BERT _{uncased}	0.186	0.203	-8 \uparrow	0.061	0.326	0.367	-11 \uparrow	0.110	0.110	0.116	-5 \uparrow	0.027	0.246	0.278	-12 \uparrow	0.100
DistilBERT _{cased}	0.139	0.141	-1	0.011	0.554	0.592	-6 \uparrow	0.103	0.051	0.049	4 \uparrow	0.018	0.447	0.480	-7 \uparrow	0.094
DistilBERT _{uncased}	0.090	0.104	-14 \uparrow	0.086	0.404	0.453	-11 \uparrow	0.124	0.034	0.039	-12 \uparrow	0.041	0.319	0.361	-12 \uparrow	0.117
RoBERTa	0.176	0.187	-6 \uparrow	0.045	0.351	0.368	-5 \uparrow	0.044	0.096	0.102	-7 \uparrow	0.034	0.271	0.284	-5 \uparrow	0.039
DistilRoBERTa	0.116	0.123	-5 \uparrow	0.036	0.466	0.481	-3 \uparrow	0.037	0.047	0.051	-7 \uparrow	0.030	0.382	0.393	-3 \uparrow	0.028
BART	0.124	0.166	-25 \uparrow	0.233	0.444	0.520	-15 \uparrow	0.188	0.046	0.067	-32 \uparrow	0.188	0.362	0.428	-16 \uparrow	0.178
Adpositions (prepositions and postpositions) [50k (AAE) 18k (SAE) terms], $\sum \Delta[\%] = 55$																
BERT _{cased}	0.227	0.199	13 \uparrow	0.105	0.507	0.447	12 \uparrow	0.140	0.129	0.105	18 \uparrow	0.108	0.442	0.380	14 \uparrow	0.153
BERT _{uncased}	0.251	0.222	11 \uparrow	0.097	0.499	0.447	11 \uparrow	0.122	0.153	0.127	17 \uparrow	0.107	0.435	0.381	13 \uparrow	0.134
DistilBERT _{cased}	0.103	0.104	-0.3	0.002	0.779	0.753	3 \uparrow	0.073	0.034	0.033	4 \uparrow	0.012	0.730	0.7	3 \uparrow	0.065
DistilBERT _{uncased}	0.140	0.135	4 \uparrow	0.029	0.598	0.562	6 \uparrow	0.084	0.057	0.053	8 \uparrow	0.032	0.532	0.493	7 \uparrow	0.095
RoBERTa	0.199	0.195	2	0.014	0.447	0.408	9 \uparrow	0.090	0.108	0.108	0.4	0.002	0.385	0.344	10 \uparrow	0.100
DistilRoBERTa	0.139	0.143	-3 \uparrow	0.022	0.584	0.542	7 \uparrow	0.099	0.057	0.060	-6 \uparrow	0.026	0.523	0.474	9 \uparrow	0.118
BART	0.154	0.154	0.02	0.000	0.552	0.525	5 \uparrow	0.063	0.062	0.063	-2	0.008	0.485	0.455	6 \uparrow	0.072

Table 17: Similar to Table. 14 but calculated for t rather than u, for three POS classes.

cording to Cohen’s classification there is a prevalent *medium* effect for both the metrics, with the exception of $MSE(CRR(u))$ for the *copula* class, where it is *large*.

At a first glance, we observe that the errors for the set of the AAE utterances in the *copula* class are larger than both the other two classes and the whole AAE corpus (reported in Tab. 14). More in general, we observe that, on average, both the three classes, and therefore, all the 150 AAE utterances, come with a less accurate average prediction with respect to the overall AAE corpus. We observe instead that the translated utterances AAE^T are better predicted with respect to AAE surprisingly for all the seven LMs.

Notably, we observe that for the translated utterances in the *double negative* class, the four metrics are always smaller (and hence sign of better performance) than those measured for the SAE corpus. This is somewhat unexpected since we observed for RoBERTa and BART an opposite bias on SAE. However, we remind that the SAE corpus, *SBC-SAE*, is made up of conversations collected from people with different regional origins. Consequently, despite the effort we make in trying not to excessively standardize the utterances during the translation process, we could be generating sentences which are free from regional bias and consequently “*cleaner*” than those found in the SAE corpus.

5.4.3 Bias on Part-of-Speech

Finally, we investigate to which extent the POS tags are tied to the measured bias towards AAE or SAE. To produce these results, we preliminary tag the tokens independently generated by each language model with the NLTK [116] POS-tagger. Next, we group by the 12 main tags of the universal tagset [118] and compute the MAE and the MSE on the term-level measurements $\Delta P(t)$ and $CRR(t)$.

Indeed, rather than averaging across the tokens in one utterance, we consider all the terms t belonging to a given POS tag. Tab. 17 reports the results obtained for the top-3 POS featuring the highest cumulative bias, computed by summing the absolute bias $|\Delta[\%]|$ introduced by each LM and measured with the $MAE(CRR(t))$: the particles (e. g., *to, up, out, etc.*), the pronouns (e. g., *you, it, my, etc.*) and the adpositions (e. g., *like, of, with, etc.*). The results for the rest of the POS are reported in Tab. 18. In order to trust the results of the POS-tagger we manually check the correctness of 100 tokens for each class and language. We find that the accuracy is 100% for the *pronouns*, 99% for the *adpositions* and 92% for the *particles*. Also in this case, we measure the effect for both the metrics, and find that, according to the 6-grade Cohen’s classification scale, it is *very small*.

Interestingly, for the *particles* class, one can notice the same pattern reported in Tab. 14. Particularly, DistilBERT_{cased} is the LM which performs better in terms of $\Delta P(t)$ and, DistilRoBERTa the one that leads to the lowest bias. Conversely, BERT is the model that shows the highest bias towards SAE: it is up to 29% more accurate with respect to AAE for $MAE(\Delta P(t))$. BART presents the opposite largest bias in favor of AAE: 23% (18%) more for the MAE of $\Delta P(t)$ ($CRR(t)$) on the *particles* class. It is also interesting to note that DistilBERT also at a token-level analysis presents better values for $\Delta P(t)$ rather than $CRR(t)$.

Quite surprisingly, we discover a bias presented by all the tested LMs towards AAE in the *pronouns* class. This holds for both the $\Delta P(t)$ and the $CRR(t)$ and is revealed with both the error measures, with the exception of BERT and DistilBERT *cased* for the MSE of $\Delta P(t)$. This result deserves further investigation.

5.5 CONCLUSIONS

In this chapter we proposed a methodology for the evaluation of the fairness of transformer-based language models. We assess and analyze the bias for two corpora, one of the spoken SAE and one of the AAE. By directly querying the underlying MLM in seven LMs, we study the quality and the bias of their predictions under several angles. The focus

is then narrowed down to the partitioning of the measured bias, by selecting the AAE corpus according to its language features. Additionally, we assess the impact of the POS tags on the found bias.

In a nutshell, results presented in this chapter suggest that different models embed diverse biases. Particularly, the most popular state-of-the-art LMs, namely BERT and DistilBERT, show a non-negligible bias towards SAE (quality of the predictions up to 21% more accurate than AAE). Instead, BART, RoBERTa and DistilRoBERTa exhibit an opposite bias. Our experiments reveal also that the distilled variants of BERT and RoBERTa are the fairest among the seven tested LMs.

Yet, despite this work provides a first insightful snapshot of linguistic bias embedded in different LMs, it opens a number of research questions. First, can fairer prediction outcomes be achieved with an ensemble learner of LMs embedding opposite biases, as, for instance, BERT_{cased} and BART? Second, our results give insights on how the bias could be consistently mitigated with more inclusive corpora, by taking into account AAE features. Finally, a special care could be put in the analysis of the distilled LMs, narrowing the gap on the causes which lead them to fairer predictions with respect to their teacher models, with a particular emphasis on the Web-based corpora used for training.

Model	MAE								MSE							
	$\Delta P(t)$				CRR(t)				$\Delta P(t)$				CRR(t)			
	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d	AAE	SAE	$\Delta[\%]$	d
Verbs [118k (AAE) 46k (SAE) terms], $\sum \Delta[\%] = 38$																
BERT _{cased}	0.171	0.148	14 \uparrow	0.097	0.413	0.378	8 \uparrow	0.082	0.091	0.073	20 \uparrow	0.097	0.353	0.320	9 \uparrow	0.082
BERT _{uncased}	0.193	0.169	13 \uparrow	0.091	0.409	0.381	7 \uparrow	0.068	0.111	0.090	19 \uparrow	0.098	0.350	0.323	8 \uparrow	0.069
DistilBERT _{cased}	0.103	0.094	9 \uparrow	0.055	0.508	0.463	9 \uparrow	0.104	0.036	0.033	8 \uparrow	0.029	0.444	0.401	10 \uparrow	0.103
DistilBERT _{uncased}	0.119	0.106	11 \uparrow	0.074	0.471	0.435	8 \uparrow	0.085	0.047	0.041	13 \uparrow	0.051	0.409	0.375	8 \uparrow	0.083
RoBERTa	0.206	0.198	4 \uparrow	0.031	0.457	0.449	2 \uparrow	0.020	0.110	0.103	6 \uparrow	0.035	0.383	0.374	2 \uparrow	0.023
DistilRoBERTa	0.167	0.159	5 \uparrow	0.036	0.551	0.535	3 \uparrow	0.039	0.074	0.069	6 \uparrow	0.029	0.475	0.457	4 \uparrow	0.045
bart	0.137	0.141	-3 \uparrow	0.021	0.554	0.557	-1	0.006	0.049	0.051	-4 \uparrow	0.017	0.479	0.477	0.4	0.004
Conjunctions [20k (AAE) 8k (SAE) terms], $\sum \Delta[\%] = 25$																
BERT _{cased}	0.269	0.274	-2	0.019	0.543	0.557	-3 \uparrow	0.036	0.147	0.153	-3	0.024	0.442	0.457	-3 \uparrow	0.042
BERT _{uncased}	0.313	0.322	-3 \uparrow	0.032	0.535	0.576	-7 \uparrow	0.113	0.191	0.191	0	0.000	0.418	0.464	-10 \uparrow	0.139
DistilBERT _{cased}	0.119	0.118	1	0.007	0.712	0.733	-3 \uparrow	0.062	0.039	0.037	5	0.020	0.625	0.652	-4 \uparrow	0.079
DistilBERT _{uncased}	0.147	0.145	1	0.009	0.674	0.699	-4 \uparrow	0.070	0.058	0.054	6 \uparrow	0.026	0.578	0.610	-5 \uparrow	0.091
RoBERTa	0.237	0.234	1	0.010	0.529	0.518	2	0.025	0.126	0.126	0	0.000	0.438	0.426	3 \uparrow	0.033
DistilRoBERTa	0.137	0.132	4 \uparrow	0.028	0.676	0.656	3 \uparrow	0.054	0.050	0.048	3	0.014	0.597	0.581	3 \uparrow	0.044
BART	0.157	0.151	8 \uparrow	0.034	0.666	0.645	3 \uparrow	0.060	0.055	0.053	4	0.018	0.565	0.548	3 \uparrow	0.049
Determiners [41k (AAE) 19k (SAE) terms], $\sum \Delta[\%] = 23$																
BERT _{cased}	0.194	0.176	9 \uparrow	0.066	0.354	0.360	-2	0.016	0.116	0.096	17 \uparrow	0.095	0.283	0.287	-2	0.014
BERT _{uncased}	0.193	0.192	0.7	0.004	0.328	0.351	-7 \uparrow	0.059	0.120	0.111	7 \uparrow	0.038	0.260	0.279	-7 \uparrow	0.057
DistilBERT _{cased}	0.122	0.116	5 \uparrow	0.032	0.524	0.552	-5 \uparrow	0.070	0.047	0.042	11 \uparrow	0.044	0.436	0.464	-6 \uparrow	0.073
DistilBERT _{uncased}	0.127	0.120	6 \uparrow	0.037	0.498	0.516	-4 \uparrow	0.044	0.053	0.047	12 \uparrow	0.049	0.412	0.431	-4 \uparrow	0.049
RoBERTa	0.181	0.169	7 \uparrow	0.049	0.342	0.339	1	0.009	0.106	0.093	12 \uparrow	0.063	0.272	0.268	2	0.013
DistilRoBERTa	0.144	0.136	5 \uparrow	0.038	0.515	0.509	1	0.015	0.064	0.059	8 \uparrow	0.036	0.433	0.426	2 \uparrow	0.019
BART	0.130	0.146	-12 \uparrow	0.088	0.419	0.430	-3 \uparrow	0.026	0.052	0.061	-15 \uparrow	0.074	0.343	0.347	-1	0.010
Adjectives [37k (AAE) 14k (SAE) terms], $\sum \Delta[\%] = 19$																
BERT _{cased}	0.199	0.194	2 \uparrow	0.020	0.631	0.645	-2 \uparrow	0.033	0.100	0.090	10 \uparrow	0.051	0.576	0.588	-2 \uparrow	0.028
BERT _{uncased}	0.228	0.225	1	0.011	0.644	0.662	-3 \uparrow	0.042	0.124	0.112	10 \uparrow	0.055	0.592	0.605	-2 \uparrow	0.032
DistilBERT _{cased}	0.097	0.106	-9 \uparrow	0.064	0.709	0.702	1	0.016	0.031	0.034	-10 \uparrow	0.033	0.658	0.647	2 \uparrow	0.026
DistilBERT _{uncased}	0.106	0.119	-11 \uparrow	0.082	0.699	0.706	-1	0.019	0.036	0.042	-14 \uparrow	0.054	0.647	0.651	-0.6	0.009
RoBERTa	0.197	0.207	-5 \uparrow	0.044	0.614	0.636	-3 \uparrow	0.052	0.095	0.098	-3	0.018	0.557	0.580	-4 \uparrow	0.054
DistilRoBERTa	0.118	0.130	-9 \uparrow	0.072	0.705	0.688	3 \uparrow	0.044	0.043	0.048	-12 \uparrow	0.049	0.652	0.633	3 \uparrow	0.048
BART	0.185	0.204	-9 \uparrow	0.089	0.685	0.731	-6 \uparrow	0.118	0.078	0.089	-13 \uparrow	0.073	0.622	0.670	-7 \uparrow	0.120
Adverbs [44k (AAE) 17k (SAE) terms], $\sum \Delta[\%] = 14$																
BERT _{cased}	0.205	0.194	6 \uparrow	0.044	0.460	0.461	-0.1	0.001	0.115	0.103	10 \uparrow	0.056	0.398	0.402	-1	0.008
BERT _{uncased}	0.229	0.220	4 \uparrow	0.032	0.454	0.460	-1	0.015	0.140	0.126	9 \uparrow	0.055	0.390	0.401	-3 \uparrow	0.026
DistilBERT _{cased}	0.099	0.095	5	0.029	0.531	0.542	-2 \uparrow	0.025	0.034	0.031	10 \uparrow	0.034	0.475	0.490	-3 \uparrow	0.035
DistilBERT _{uncased}	0.113	0.107	5 \uparrow	0.030	0.496	0.510	-3 \uparrow	0.031	0.044	0.039	12 \uparrow	0.045	0.437	0.455	-4 \uparrow	0.043
RoBERTa	0.214	0.216	-1	0.006	0.483	0.499	-3 \uparrow	0.038	0.116	0.114	1	0.008	0.414	0.433	-5 \uparrow	0.048
DistilRoBERTa	0.140	0.130	7 \uparrow	0.054	0.561	0.567	-1	0.016	0.058	0.050	15 \uparrow	0.063	0.495	0.505	-2 \uparrow	0.025
BART	0.221	0.238	-7 \uparrow	0.072	0.620	0.645	-4 \uparrow	0.067	0.106	0.115	-7 \uparrow	0.046	0.539	0.560	-4 \uparrow	0.052
Nouns [104k (AAE) 41k (SAE) terms], $\sum \Delta[\%] = 12$																
BERT _{cased}	0.217	0.192	11 \uparrow	0.098	0.661	0.655	0.8 \uparrow	0.013	0.114	0.091	20 \uparrow	0.113	0.609	0.606	0.6	0.008
BERT _{uncased}	0.258	0.224	13 \uparrow	0.124	0.707	0.693	2 \uparrow	0.035	0.147	0.113	24 \uparrow	0.146	0.660	0.645	2 \uparrow	0.036
DistilBERT _{cased}	0.103	0.104	-0.3	0.002	0.779	0.753	3 \uparrow	0.073	0.034	0.033	4 \uparrow	0.012	0.730	0.7	3 \uparrow	0.065
DistilBERT _{uncased}	0.117	0.110	6 \uparrow	0.044	0.741	0.724	2 \uparrow	0.045	0.041	0.036	11 \uparrow	0.039	0.693	0.676	2 \uparrow	0.043
RoBERTa	0.203	0.208	-2 \uparrow	0.021	0.655	0.663	-1 \uparrow	0.018	0.096	0.097	-2	0.009	0.604	0.613	-2 \uparrow	0.022
DistilRoBERTa	0.128	0.132	-3 \uparrow	0.023	0.714	0.706	1 \uparrow	0.022	0.048	0.050	-4 \uparrow	0.014	0.665	0.656	1 \uparrow	0.023
BART	0.156	0.161	-3 \uparrow	0.025	0.718	0.733	-2 \uparrow	0.040	0.058	0.060	-2	0.009	0.668	0.684	-2 \uparrow	0.041

Table 18: MAE and MSE of $\Delta P(t)$ and CRR(t) measured on AAE and SAE corpora: results obtained through the *fill-in-the-blank* task with different language models, averaging token predictions for each POS class. \uparrow signifies that the AAE and SAE expectations are statistically significant according to the Welch’s two-tailed t-test (p-value < 0.05). The column d contains their effect size computed according to the Cohen’s d.

Contents

6.1	IP-ID classification via supervised learning	79
6.2	Background and related work	81
6.2.1	Normative reference	81
6.2.2	IP-ID Classification Breakdown	82
6.2.3	IP-ID Based-Inference	83
6.3	Methodology	84
6.3.1	Active probing	84
6.3.2	Features Definition	87
6.3.3	Datasets	88
6.4	IP ID Classification	92
6.4.1	Classification accuracy and validation	92
6.4.2	Robustness	93
6.5	Internet Census	96
6.5.1	Longitudinal Comparison (over the years)	97
6.5.2	Sensitivity Analysis	98
6.5.3	Spatial analysis	99
6.6	Conclusions	101

6.1 IP-ID CLASSIFICATION VIA SUPERVISED LEARNING

The IP identification (IP-ID) is a 16 (32) bits field in the IPv4 (v6) header [120]. Originally, along with the fragment offset, the IP-ID was used to assist packet segmentation and reassembly and it was unique per each combination of source, destination and protocol. Yet, with technology evolution and the adoption of the MTU path discovery [104], IP fragmentation becomes less common nowadays, so that the last normative reference [146] allows IP-ID of atomic datagrams to be non-unique. As a consequence, IP-ID fields values are determined by the specific implementation of the Operating System [105]. In particular, the majority of research work focus their attention on the *global*

counter implementation, which used to be the most common implementation about a decade ago [153]. However, due to recent evolution of the standards [59, 146], a wider range of behaviors can be expected nowadays. Over time, different behaviors have been observed such as *global* and *per-flow* counters, *pseudo-random* sequences and *constant* values [15], as well as odd behaviors such as those due to load balancing [36] middle-boxes, or host implementations using the wrong endianness [105]. Given that some of the above implementations maintain state at the IP level, the IP-ID field has been of invaluable help to infer a wealth of information concerning the network. Particularly, by leveraging inference from *global* IP-ID implementation, researchers have been able to count the number of hosts behind NATs [15, 105], or even assess the traffic they generate [36, 75] and finally expose censorship in the Internet [18, 94, 105, 117].

Given this context, and in particular the emergence of new IP-ID behaviors, it is important to define methods to classify them, as well as using these methods to quantify the prevalence of IP-ID implementation in the current Internet. To summarize our main contributions:

- we design and implement a lightweight framework to classify the full range of IP-ID behaviors, based on a handful of ICMP packets;
- we carefully validate our method against two datasets comprising the replies from about 1,855 sample hosts, chosen in different manners, for which we build a ground-truth by manual inspection and against multiple synthetic datasets, tailor-made to test robustness against various forms of shortfalls;
- we apply the methodology to an Internet-wide campaign, where we classify one alive target per each routable /24 subnet, gathering a full blown picture of the IP-ID adoption in the wild.

Specifically, whereas the *global* counter (18% of occurrences in our measurement) implementation was the most common a decade ago [153], we find that other behaviors (*constant* 34% and *local* counter 39%) are now prevalent. We also find that security recommendations expressed in 2011 [59] are rarely followed (*random*, 2%). Finally, our census quantifies a non marginal number of hosts (7%) showing evidence of a range of behaviors, that can be traced to poor or non-standard implementations (*e. g.*, bogus endianness, non-standard increments) or network-level techniques (*e. g.*, load balancing, or exogenous traffic intermingled to our probes confusing the classifier). To make our findings useful to a larger extent, we make all our dataset and results available at [131].

This chapter is structured as follows. After overviewing the related work (Section 6.2), we describe the methodology and illustrates the workflow and the datasets involved (Section 6.3). We next show the performance of the *supervised classification* approach chosen

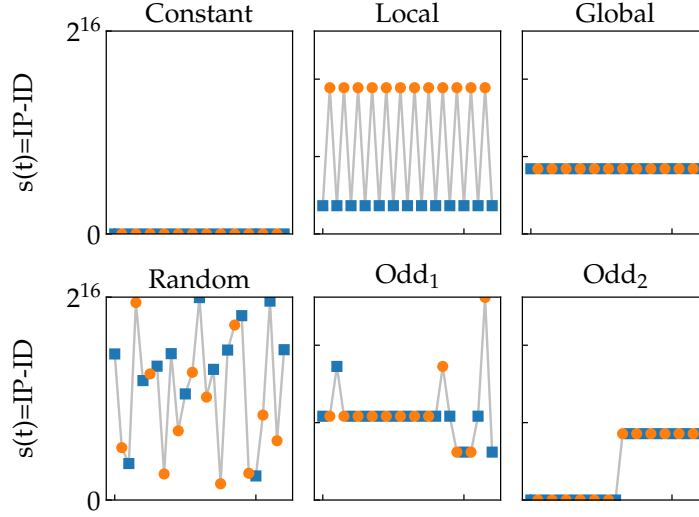


Figure 25: Illustration of Constant, Local, Global, Random and Odd sequences

in the following order: system validation, robustness assessment and probing overhead analysis (Section 6.4). Then, we present the results of the classifier when operated in the wild and put in perspective our findings with those of the previous works (Section 6.5). Finally, we summarize the main outcomes (Section 6.6).

6.2 BACKGROUND AND RELATED WORK

6.2.1 Normative reference

The IP identification (IP-ID) field identifies the unique fragments of a packet and it is used to handle the re-assembling process. First documented in the early 80s by RFC791 [120] its use has been updated in several RFCs [23, 53, 59, 60, 146, 153]. Whereas [120] does not fully specify the IP-ID behavior (*i. e.*, it only states that each packet must have a unique IP-ID for the triplet of source, destination and protocol), different behaviors (namely *global*, *local* and *random*, illustrated in Fig. 25) are detailed in 2006 by RFC4413 [153]. In 2008, RFC5225 [53] observed that some hosts set the IP-ID to *zero*: at the time of [53], this was a not legal implementation as the field was supposed to be unique.

Yet, in 2012 [105] observed that the actual IP-ID implementation depends on the specific Operating System (OS) and versions¹. In 2013, RFC6864 [146] updated the specifications by affirming that the IPv4 ID uniqueness applies to only non-atomic datagrams: in other words, if the don't fragment (DF) bit is set, fragmentation and reassembly are not necessary and hence devices may set the IP-ID to zero. At the same time, concern has

¹ In particular [105] reports Windows and FreeBSD to use a *global* counter, Linux and MacOS to use *local* counters and OpenBSD to use *pseudo-random* IP-IDs.

Table 19: Summary of related work

Work	Year	Features	Census	Classes down (%)	Break- down (%)	Methodology	Scope of the work
[99]	2003	Δ IP-ID	no (5000 target routers)	70% remaining between constant (equal to 0) and counters with increment by 2.	global, 30%	Analysis of replies to active probing (ICMP requests)	Packet reordering and losses diagnosis.
[36]	2005	Δ IP-ID	no (50 target web-servers)	38% global		Analysis of replies to active probing (HTTP requests)	Discover the amount of load balanced servers, measure the traffic generated by a server.
[69]	2013	-	no	57% global, 14% local, 9% constant, 20% <i>mixed</i> IP-ID, 1% random/other (^a)		-	Off-path DNS cache poisoning attacks and defense against them through DNSSEC validation.
[117]	2017	IP-ID acceleration	no	16% global		TCP SYN-ACK from multiple vantage points	Reveal Internet censorship.

^a Due to the rounding done by the authors [69], the sum of all the percentages is 101%

been raised about security problems following the predictability of IP-ID sequences [56, 60, 69, 84]. In particular, in 2012 RFC6274 [59] discouraged the use of a *global* counter implementation for many security issues, such as stealth port scan to a third (victim) host, and in 2016 RFC7739 [60] addressed concerns concerning IPv6-specific implementations. In light of the recent evolution of the standards, a re-assessment of IP-ID usage in the wild is thus highly relevant.

6.2.2 IP-ID Classification Breakdown

In the last decade, to the best of our knowledge, few research works have provided a complete picture of the breakdown of the existing IP-IDs behaviors. That is what makes the comparison of the results of this work with the previous ones with the purpose of analysing the temporal changes on the IP-ID popularity an hard task.

Specifically, the sole quantitative assessment of IP-ID behavior over multiple classes dates back to 2013. This is limited to 271 Top Level Domains TLDs probed by [69] (whose main aim is to propose practical poisoning and name-server blocking attacks on standard DNS resolvers, by off-path, spoofing adversaries). In particular, the 2013 study finds 57% *global*, 14% *local*, 9% *constant*, 1% *random/other*. Additionally, [69] suggests that

20% of DNS TLD exhibit evidence of “two or more sequential sequences mixed up, probably due to multiple machines behind load balancer”.

The remaining works concentrate instead on assessing the popularity of just the *global* implementation being it only the focus of their studies, proving once again the relevance of a Internet-wide list comprising IP addresses generating IP-ID with the aforementioned behavior. Namely, in 2003, [99] reported that 70% (over 5000 probed targets) were using an IP-ID counter (*global* or *local* implementation); in 2005, [36] reported that 38% (over 150 hosts) used a *global* IP-ID; in 2006, [153] affirms the *global* implementation to be the most common assignment policy (among 3 behaviors).

6.2.3 IP-ID Based-Inference

Additionally, the IP-ID has been exploited for numerous purposes in the literature. Notably, IP-ID side-channel information helped to discover load balancing server [36], count hosts behind NAT [15, 105], measure the traffic [36, 75] and detect router alias [16, 83, 141]. More recently, [94] leverages IP-ID to detect router aliases, or infer router up time [18] and to reveal Internet censorship [117], refueling interest in the study of IP-ID behavior. Whereas the above work [15, 36, 75, 117, 141] mostly focus only on the *global* IP-ID behavior, in this work we not only consider all *expected* IP-ID behavior, but additionally quantify *non-standard* behaviors: in particular, we provide a methodology to accurately classify IP-ID behaviors, that we apply to the Internet at large, gathering a picture of the relative popularity of each IP-ID behavior. In terms of methodologies, authors in [99] use ICMP timestamp and IP-ID to diagnose paths from the source to arbitrary destinations and find reordering, loss, and queuing delay. In [78], the authors identify out-of-sequence packets in TCP connections that can be the result of different network events such as packet loss, reordering or duplication. In [36], they use HTTP requests from two different machines toward 150 target websites, to discover the number of load-balancing server. Authors in [117] use TCP SYN-ACK from multiple vantage points to identify connectivity disruptions by means of IP-ID fields, which then they use as a building block of a censorship detection framework.

In this chapter, we leverage ICMP traffic (spoofing IP addresses to craft sequences of packets that are precisely interleaved when they hit the target under observation) to build an accurate, robust and lightweight IP-ID classification technique.

6.3 METHODOLOGY

To provide an accurate and comprehensive account of IP-ID behavior in the wild, we need (i) a reliable classifier, able to discriminate among the different typical and anomalous IP-ID behaviors. At the same time, to enable Internet coverage, (ii) the classifier should rely on features with high discriminative power, extracted from the data gathered through an active probing technique that is as lightweight as possible. In this section we illustrate the practical building blocks and their theoretical foundations, that our classification framework builds upon.

IP-ID classes: From the host perspective, several IP-ID behaviors are possible as depicted in Fig. 25. The image shows the sequences of 25 IP-ID samples sent from 2 different host (orange and blue) where the packets are sent alternatively to the target. The different behaviors depicted are, from left to right: (i) *constant* counters are never incremented (and for the most part are equal to 0x0000); (ii) *local* or per-host counters that are incremented at each new packet arrival for that flow (mostly by 1 unit, 99.7% of the times in our large scale measurements): as a consequence, while the orange or blue per-host sub-sequences are monotonically increasing, the aggregate sequence alternates between the two; (iii) *global* counters are incremented by 1 unit at each new packet arrival for any flow: thus, the sequence s is monotonically increasing (90.3% of the times by 1 unit, 4.7% by 2 units and 4.6% by 3 units), and the orange or blue per-host sub-sequences are monotonically increasing but at a faster rate (by 2 units); (iv) *random* IP-IDs are extracted according to a pseudo-random number generator. Finally, a special mention is worth for the class of (v, vi) *odd* IP-ID behaviors, that are not systematically documented in the literature and that arise for several reasons (including bugs, misconfiguration, non-standard increments, unforeseen interaction with other network apparatuses, *etc.*) and for which we report two different samples occurring in real experiments.

6.3.1 Active probing

To gather the above described sequences, our measurement technique relies on active probing. We craft a tool able to send and receive ICMP packets, running at two vantage points (VP) with public IP addresses in our campus network. Specifically, we send a stream of N ICMP echo requests packets in a *back-to-back* fashion, which forces the target machine to generate consecutive ICMP echo replies: thus, assuming for the time being that no packet were lost, we gather a stream of N IP-IDs samples for that target.

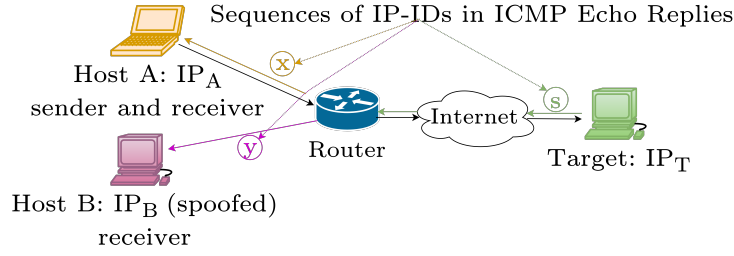


Figure 26: Scenario in which the active probing is performed: only one sender is used to ease the synchronization of packets generation, whilst both the machines are used to receive and collect the stream of IP-IDs generated at the target machine.

Sending packets back-to-back is necessary to reduce the noise in the IP-IDs stream sequence: if probe packets were spaced over time, the sequence could be altered by exogenous traffic hitting the target (*e. g.*, in case of global counter). As a result, the sequence would depend on the (unknown) packet arrival rate in between two consecutive probe packets, likely confusing the classifier. In this way, the use of back-to-back packets reduces as much as possible the interference with some possible extra exogenous traffic hitting the same destination, that could otherwise alter the sequences. A second observation is that, whereas a single vantage point may be sufficient to distinguish among constant, random and global counters, it would fail to discriminate between global vs local counters. However, sending packets from two different VPs is not advisable, due to the difficulty in precisely synchronizing the sending patterns so that packets from different hosts alternate in the sequence [99].

Therefore, a better alternative is to receive packets on two different VPs, x and y , but shift the packet generation process to only one of them, as x , and use it as sender: by letting x spoof the address IP_y of the colluding receiver y , it is possible to generate a sequence of *back-to-back packets* that are also *perfectly interleaved* as depicted in Fig. 25. Fig. 26 shows the scenario in which the experiments are carried out. It provides information about how the hosts are involved and the kind of data collected: there are two receivers but only one real sender, and the information gathered at the two vantage points regards the sequences of IP-IDs generated by the target machine. To validate our assumptions, we carry on additional experiments on a preliminary testbed to test the sensitivity of the algorithm to external traffic hitting the target. In these experiments:

- we send UDP CBR traffic with Iperf at $TX_{rate} = 10\text{Mbps}$ and vary the packet size over time (in particular decrease), so that we increase the packet rate during the experiment (to control the IP-ID generation);
- in one experiment, we send ICMP Echo Request packets with an inter-packet gap of $\Delta t_{interpacketgap} = 10\text{ms}$ and collect the IP-ID sequence x , for which we derive the derivative series (gray color line);

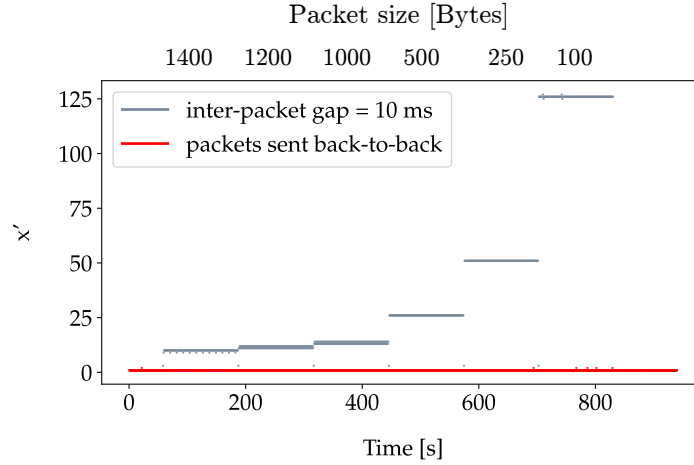


Figure 27: Sensitivity analysis to external traffic: derivative of the sequence of IP-IDs x' in the two different scenarios

- in the other experiment, we send ICMP packets back-to-back and again measure the growth of IP-ID in the sequence (red color line).

Even though the experiments are simple, the results are very telling: the plots in Fig. 27 show the derivative of the sequence of IP-IDs x' , which basically just counts the amount of exogenous packets in between two consecutive ICMP probes, in both scenarios of the experiments. For instance, when packets are 100 Bytes long, in $\Delta t_{\text{interpacketgap}} = 10\text{ms}$ it is expected to have $\frac{TX_{\text{rate}} \cdot \Delta t_{\text{interpacketgap}}}{\text{packet size}} = 125$ packets slipping in between two probes, which actually happens. This would clearly jeopardize the classifier. Conversely, in the experiments carried out in our lab, back-to-back packets leave no possibility to the other UDP packets to intermingle and confuse the classifier. These experiments suggest that sending packets back-to-back is a good strategy, although we do not feel results to be conclusive for all the devices available in the network (*e. g.*, router, setup, shaper, *etc.*). However, even in case the reality was not as nice as our experimental lab results (which is likely to be the case), at the same time this affects at most some of the *odd* behaviors, which already are a tiny (7%) fraction of the overall cases. Indeed, it is very unlikely that the amount of real traffic is so perfectly varying between probes to erroneously confuse a classifier to believe that a *global* sequence is a *random* one just due to exogenous traffic. Very high information entropy of those sequences is not a side-product of some variable traffic, but truly coming from a random number generator (it is pretty well known that is hard to generate good pseudo-random sequences, and the arrival rate is surely not a source of perfect entropy).

To overcome reordering, packet loss and duplication events, we additionally control the sequence number in the stream of generated probe packets.

Feature	Constant	Local	Global	Random	Odd
$H(x)$	0	$\log_2 \frac{N}{2}$	$\log_2 \frac{N}{2}$	$\leq \log_2 \frac{N}{2}$	-
$H(s)$	0	$\leq \log_2 N$	$\log_2 N$	$\leq \log_2 N$	-
$H(x')$	0	0	0	$\leq \log_2 \frac{N}{2}$	-
$H(s')$	0	1	0	$\leq \log_2 N$	-
$\mathbb{E}[x']$	0	1	2	$\frac{(2^{16}-1)}{2}$	-
σ_x	0	$\sqrt{\frac{(N^2-4)}{48}}$	$\sqrt{\frac{(N^2-4)}{12}}$	$\frac{(2^{16}-1)}{\sqrt{12}}$	-
σ_s	0	$\leq \frac{(2^{16}-1)}{\sqrt{12}}$	$\sqrt{\frac{(N^2-1)}{12}}$	$\frac{(2^{16}-1)}{2}$	-
σ'_x	0	0	0	$\frac{(2^{16}-1)}{\sqrt{12}}$	-
σ'_s	0	$ x_1 - y_1 - \frac{1}{2} $	0	$\frac{(2^{16}-1)}{\sqrt{12}}$	-

Table 20: Tabulated expected values for selected features

6.3.2 Features Definition

As anticipated, to build a robust classifier we need to *manually* define a set of tailor-made features able to discriminate among the different IP-IDs implementations. The experiment and the measurements can be formalised as follows: we send N packets to a given target t , with the source address field alternating between consecutive requests, whose replies are sent back to our two vantage points x and y : we indicate with s the aggregated sequence comprising the N IP-IDs sent back by t , as we receive it at the edge of our network². By abuse of language, we indicate with x and y the sub-sequences (each of length $N/2$) of IP-IDs, sent back by t and received by the homonymous host. From these sequences x, y and s we further construct derivative series x', y' and s' by computing the discrete differences between consecutive IP-IDs (*i. e.*, $x'_i = x_i - x_{i-1}$). We summarize these series with few scalar features by computing the first $\mathbb{E}[X] = \frac{1}{N} \cdot \sum_i^N x_i$ and second moments of the IP-ID series, $\sigma = \sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}$ as well as their information entropy, defined as the expected value of the information content $H(X) = \mathbb{E}[\mathbb{I}(X)] = -\sum_i^N p_i \log_2 p_i$

where p_i is the provability that the discrete random variable X takes the x_i value.

Specifically, we consider the mean $\mathbb{E}[X]$ of the derivative series x' and y' , the entropy $H(X)$ and the standard deviation of s, x and y and of their derivatives s', x' and y' . Actually, for each feature we can derive an *expected value* in the ideal³ case (so that no

² Notice that packet losses and reordering may let us receive less than N packets, or receive packets in a slight different order than what sent by the target.

³ Sequences from well behaving hosts that have no software bug or malicious behavior, and that are neither affected by losses nor reordering

expected values is reported for the odd class) that we summarize in Tab. 20. For the sake of brevity, we report in Tab. 20 only once the expectations of the features of the subsequences x and y , given that they are conceptually equivalent. Intuitively, we expect the mean of the constant sequence to be unknown, but that of its derivative to be null. Similarly, the derivative of a global counter would have a value of 1 (2) for the aggregate sequence s (subsequences x and y). The entropy of the sequence is expected to increase from the minimum of a constant sequence equal to $\mathbb{H}(X) = -1 \log_2(1) = 0$

to the maximum of $\mathbb{H}(X) = -N \cdot \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N$ occurring when all the N elements of the series are different. Consequently, by considering the global and local implementations, we can observe that the entropy for the sequences x and y of length $\frac{N}{2}$ is expected to be maximum $\mathbb{H}(x_{\text{global}}) = \mathbb{H}(y_{\text{global}}) = \log_2 \frac{N}{2}$. Accordingly, in the global implementation, the sequence s is made up of two not-overlapping sequences, leading to an expected maximum entropy of $\mathbb{H}(s_{\text{global}}) = \log_2 N$. Differently, in the local implementation this is true only when the two counters do not overlap, otherwise this remains only an upper bound. A similar observation can be done for the entropy expectations for the random sequences, in which the presence of duplicate values would reduce the entropy. For the local implementation, the sequences x', y' , derivatives of two independent counters, are constant thus the entropy, as said, is expected to be 0. On the other hand, the derivative s' of the aggregate sequence s is made up of two alternating values, corresponding to the two offsets:

$$s'_{\text{local}}(n) = \begin{cases} \theta_1 = y_1 - x_0 & \text{if } n \text{ even} \\ \theta_2 = x_2 - y_1 & \text{if } n \text{ odd} \end{cases} \quad (9)$$

Both θ_1 and θ_2 are repeated for $\frac{N}{2}$ times, so each one occurs with a probability of $\frac{1}{2}$. The entropy becomes $\mathbb{H}(s'_{\text{local}}) = -2 \cdot \frac{1}{2} \log_2 \frac{1}{2} = 1$. Conversely, being the expected derivative sequence of a global counter always equal to $s'_{\text{global}} = 1$, as a result the entropy becomes $\mathbb{H}(s'_{\text{global}}) = 0$.

Similarly, the other expectation values can be easily derived by analogy.

6.3.3 Datasets

In this work, we collect four different datasets, that we use in the different stages of the work alternatively to make the classifier learn the classification function, *i. e.*, as training dataset, and to evaluate performances as testing dataset.

Table 21: Summary of the Datasets.

Name	Type	Description	Properties	Size [Targets]	URL
\mathcal{L}	Real Measurements	Large scale measurements dataset comprising the IP-ID sequences received from the portion of hitlist [68] providing response rate $\geq 80\%$	Presence of presence of odd behaviors of the IP-ID, possibility of losses or out-of-order packets	2,5 M	[131]
\mathcal{G}	Real Measurements	Manually labeled dataset containing the IP-IDs contained in the replies of a set of IP addresses sampled uniformly from the hitlist to guarantee class balance	Targets chosen to provide IP-prefix level and class balance, presence of odd behaviors, used for training and classification of \mathcal{L}	2 k	[131]
\mathcal{G}'	Real Measurements	Manually labeled dataset containing the IP-IDs from the replies of a set of IP addresses where 75% of it belong to the same IP/8 subnet	Targets chosen to provide IP-prefix level imbalance, presence of odd behaviors, used for validation of performances	2 k	[131]
$\mathcal{S}_{\text{ideal}}$	Synthetic	Dataset manually designed to intentionally contain the four possible IP-ID implementations in the ideal case evenly distributed emulating the replies collected through real measurements	Lossless, absence of odd behaviors, used for validation of performances	20 k	[131]
$\mathcal{S}_{\text{lossy}}$	Lossy Synthetic	Dataset manually designed to intentionally contain the four possible IP-ID implementations spoiled with four different flavour of losses ($\mathcal{S}_{\text{lossy}} = \cup(\mathcal{S}_{\text{unif}}, \mathcal{S}_{\text{hole}}, \mathcal{S}_{\text{extr}}, \mathcal{S}_{\text{equi}})$) evenly distributed	Lossy, absence of odd behaviors, used for testing resilience to losses	20 k	[131]
$\mathcal{S}_{\text{reorder}}$	Synthetic	\mathcal{G} dataset spoiled when of 20% of each IP-ID sequence is intentionally randomly swapped	Used for testing resilience to sequence alteration due to out-of-order packets	20 k	[131]

Large scale census \mathcal{L} : The first dataset is made up of real measurements coming from a large scale measurement campaign and includes the replies coming from a subset of a hitlist of alive IP addresses [68]. We avoid putting stress on the infrastructure carrying a full Internet census: as we aim at providing an accurate picture of the *relative* popularity of IP-ID implementations on the Internet, it suffices to collect measurements for portion of targets, namely 1 alive IP/32 host per each /24 prefix. For this reason, for the targets selection, we rely on the public available hitlist regularly published by [68], comprising 16 millions of targets IP/32. The hitlist contains targets for all /24, including those who have never been replying to the probing: excluding them from our target list, leaves us with approximately 6 millions of potential targets. To reduce the amount of probe traf-

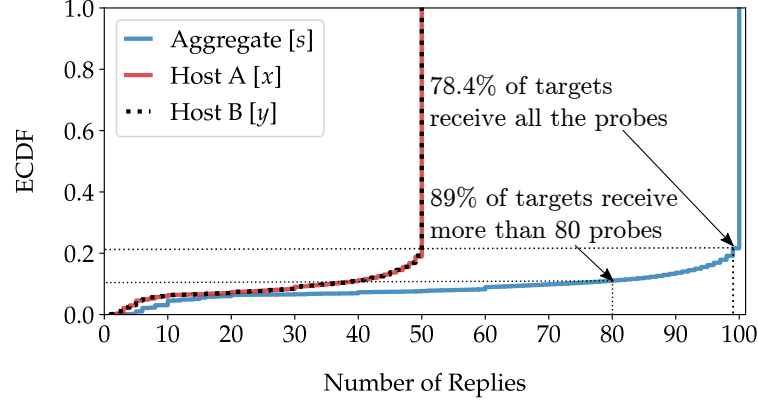


Figure 28: Internet campaign: ECDF of the number of packet replies

fic, we decide to be conservative: we preliminary probe the 6 millions potential targets sending two ICMP echo requests, and include in our final target list the approximately 3,2 million responsive hosts (in line with [40, 159]). We send a batch of $N = 100$ back-to-back probe packets to each target, but otherwise probe at a low average rate, so that we complete a /24 census in about 3 days. Fig. 28 shows the empirical cumulative distribution function (ECDF) of the received packets at our VPs. We observe that we receive almost all the replies from most of the targets: the 90% (80%) of the targets answer to more than 40 (all) packets per each host, corresponding to a 20% (0%) loss scenario. A large plateau in the CDF also indicates that the distribution is bi-modal, *i. e.*, the remaining hosts generally reply with very few packets (*e. g.*, 10 or less per each VP or over 90% loss rate). This suggests that future campaigns could be safely conducted with a smaller $N' < N$. To provide accurate classification results, in light of our robustness analysis done with synthetic dataset and whose results are shown in Sec. 6.4.2, we limit our attention to the 2,588,148 hosts for which we have received at least $N = 80$ packets.

Ground Truth \mathcal{G} and \mathcal{G}' :

The second real dataset is \mathcal{G} , made of IP-ID sequences for which we manually construct a ground truth. For this purpose, we extract the replies from a subset of targets of \mathcal{L} which satisfy some pre-established requirements. We include in this dataset only the 1,855 hosts from which we receive 100% of the replies, and perform the manual inspection of each of the sequences. We repeat the process twice, with two very different choices of the ground-truth datasets: \mathcal{G} sampled uniformly from the hitlist paying attention to guarantee class balance and \mathcal{G}' where about 75% samples belong to the same IP/8 subnet. Interestingly, when performing the manual labelling, we find a small but non marginal fraction (about 7%) of sequences that are hard to classify: a deeper investigation reveals these odd behaviors to be due to a variety of reasons – including

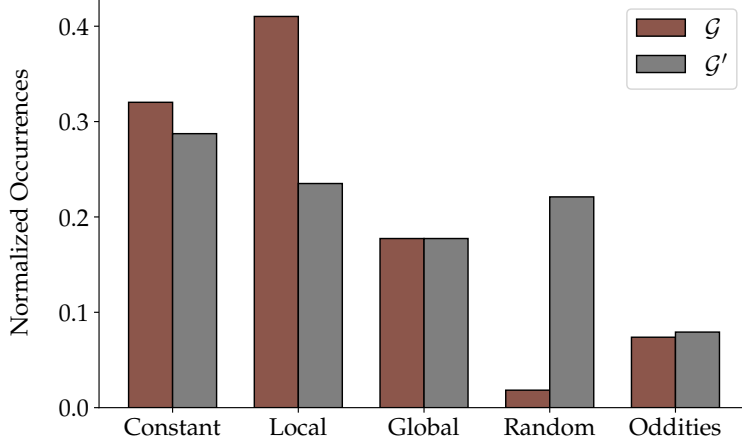


Figure 29: Manual Ground Truth: Normalized classes occurrences for the training datasets \mathcal{G} and \mathcal{G}'

per-packet IP-level load balancing, wrong endianness, non standard increments in the global counter, *etc.* While we cannot completely rule out interference of exogenous traffic altering our IP-ID sequences, lab experiments suggest that the use of back-to-back packets lessen its impact, as described before in Sec. 6.3.1. Nevertheless, these samples provide a useful description of the odd class, that would otherwise have been difficult to define. In Fig. 29 we report the breakdowns of the two datasets \mathcal{G} and \mathcal{G}' .

Syntethic Datasets:

In order to assess the robustness of our classifier against packet losses, we rely on two more datasets which are made up by synthetic sequences, from which we can derive the features useful in the classification process. While for simple loss patterns (*e. g.*, uniform i.i.d. losses) it is still possible to analytically derive expected values in closed form, for loss models where losses are correlated, this becomes significantly more difficult. As such, we opt for an experimental assessment of classification accuracy in presence of different synthetic loss models, that we apply to synthetic ideal sequences contained in dataset \mathcal{S}_{ideal} by purposely discarding a part of the sequences. Specifically, we consider: (i) a *uniform* i.i.d. loss model; (ii) a *hole* model where, starting from a random point in the sequence, 20% of consecutive samples are removed; (iii) an *extreme* model where we remove 20% of the initial values (or equivalently the final 20% of the sequence); and finally (iv) an *equidistant* model where losses start at a random point and are equally spaced over the sequence. We apply these loss models to obtain a synthetic lossy dataset \mathcal{S}_{lossy} . Specifically, for each loss model we generate 5,000 loss sequence pattern, for an overall of 20,000 test cases. In order to deeper investigate the reordering phenomena effect on the performances of the classifier, we manually disrupt the sequences contained

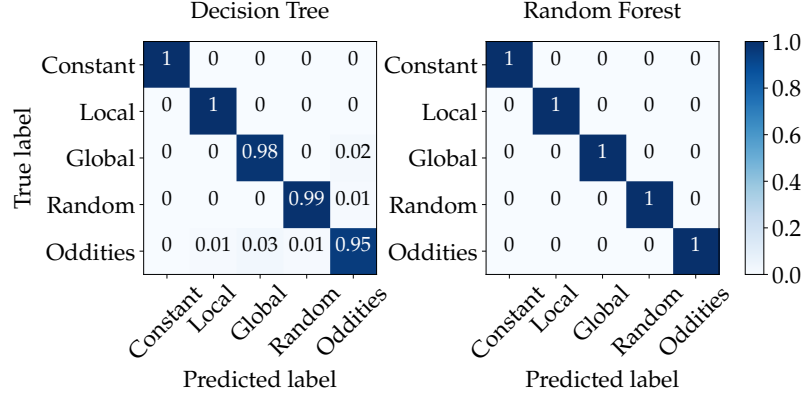


Figure 30: Validation: Confusion Matrix of 20-fold validation over \mathcal{G} done both with Decision Tree and Random Forest Classifiers

in \mathcal{G} . Specifically, we impose the swapping of 20% on the IP-IDs contained in the series x, y collected for each IP address in \mathcal{G} and build a new rigged dataset $\mathcal{S}_{\text{reorder}}$.

A summary of all the datasets with their description and properties is reported in Tab. 21.

6.4 IP ID CLASSIFICATION

From the values tabulated in Tab 20, we expect classifiers that use this set of features to be able to fully discriminate the set of IP-ID well-defined behaviors under ideal conditions. However, as we shall see, unexpected behavior may arise in the Internet, due to a variety of reasons, which are hard to capture in general. We thus opt for a *supervised classification* approach, which allows to learn a predictive model with decision trees (DTs), based on the above features. Additionally, we investigate to what extent the classifier is robust against losses and reordering, and finally assess the minimum number of samples N needed to achieve a reliable classification.

6.4.1 Classification accuracy and validation

We first train and validate our classifier using the the real dataset \mathcal{G} of IP-ID sequences for which we have manually constructed a ground truth. Note that, for the moment we train the classifier only over the dataset \mathcal{G} , but later we will show the independence of the model from this choice.

We assess the classification accuracy over \mathcal{G} with a 20-fold cross-validation, whose results are reported in Fig. 30 as a confusion matrix: we can observe that the classifier is extremely accurate, with 100% true positive in the constant and local classes, 99%,

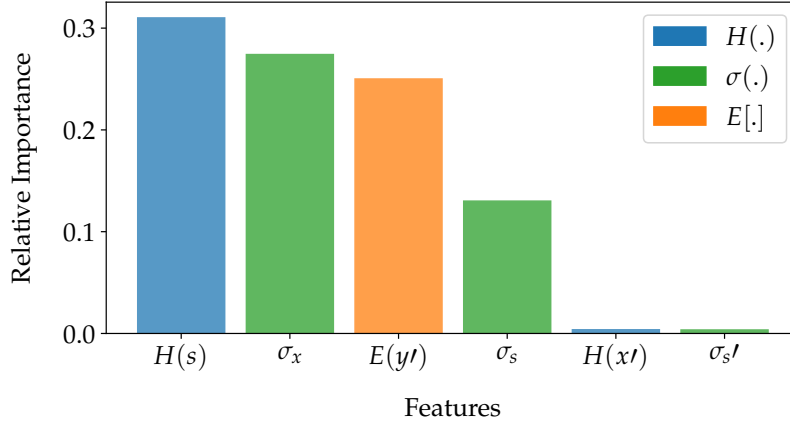


Figure 31: Validation: Relative importance for the most useful features of the classifier.

for the random and 98% for the global class. The worst case is represented by 95% true positive for the odd class (that represent only 7% of the samples): these very few misclassifications are erroneously attributed to local, global or random classes, and additional series definition (*e. g.*, to compensate for wrong endianness) could help reducing if needed. For completeness, we compare the results obtained with the Decision Tree [92] with the ones achieved with Random Forest. Results, shown in Fig. 30, again as a confusion matrix, show that the small misclassification gaps introduced by the Decision Tree are fully filled when using a Random Forest Classifier, which leads to 100% classification accuracy for all the classes.

Additionally, Fig. 31 reports the importance for the most useful features of the classifier. Four main takeaways can be gathered from the picture: first, just four features are necessary for a full discrimination, which is reasonable as the cardinality of the classes to discriminate is small; second, as expected features that measure the dispersion (entropy and standard deviation) are prevalent; third, both original and derivative sequences are useful in the detection; fourth, subsequence metrics are highly redundant (*i. e.*, $H(x) = H(y)$, $\sigma_x = \sigma_y$, *etc.*).

6.4.2 Robustness

Before operating the classifier on the data collected in the wild, we study its robustness both to packet losses and to the number of packets employed in the experiments.

It is fundamental to test the robustness of the features to losses in a controlled scenario, in order to emulate the real measurements, in which events such as packet losses or out-of-order arrivals are not so rare.

Table 22: Features values for both lossless and lossy synthetic dataset $\mathcal{S}_{\text{lossy}}$ with 20 % losses - local implementation case of IP-ID.

	$\mathcal{S}_{\text{ideal}}$	$\mathcal{S}_{\text{lossy}}$			
Feature	Lossless	Uniform	Hole	Extremal	Equidistant
$H(s)$	6.64	64	6.64	6.64	6.64
$H(x')$	0	0.84	0.17	0	0.78
$E[y']$	1	1.25	1.25	1	1.26
$\sigma(x)$	32.64	38.75	29.68	18.02	20.92
$\sigma(s)$	$10.97 \cdot 10^3$	$11.09 \cdot 10^3$	$1072 \cdot 10^3$	$11.03 \cdot 10^3$	$11.05 \cdot 10^3$
$\sigma(s')$	$16.29 \cdot 10^3$	$16.01 \cdot 10^3$	$16.6 \cdot 10^3$	$16.15 \cdot 10^3$	$16.4 \cdot 10^3$

Robustness to Losses:

For the previously shown six features we evaluate their values in the lossy synthetic $\mathcal{S}_{\text{lossy}}$ sequences and tabulate the results averaged over the dataset, respecting the IP-ID and loss type partitioning, in order to compare with the ones evaluated for the lossless sequences in $\mathcal{S}_{\text{ideal}}$. In Tab. 22 we report those values evaluated for the simulated local implementations. We compare the features evaluated for the lossless dataset $\mathcal{S}_{\text{ideal}}$ with those of the lossy $\mathcal{S}_{\text{lossy}}$ with uniform random, hole, extremal and equidistant losses. As a whole, results obtained with the synthetic dataset $\mathcal{S}_{\text{lossy}}$ do not significantly diverge from the ones obtained with $\mathcal{S}_{\text{ideal}}$, proving the strength of the features and their robustness to change and alteration of the original sequences. Specifically, $H(s)$, which turns out to be the most important feature, as shown in Fig. 31, does not vary in presence of any kind of losses, while $H(x')$ can vary more depending on the flavour of the loss.

Given these results, we next assess the robustness of the classifier against packet losses, which may introduce distortion in the features. Since, as previously described, the expected values in the ideal conditions are significantly apart, we expect the classifier to be resilient to a high degree of losses. Without loss of generality, we consider an extreme case where only 80 out of 100 samples are correctly received (*i. e.*, a 20% loss rate) by exploiting the lossy synthetic dataset $\mathcal{S}_{\text{lossy}}$.

We want to assess the accuracy of the previously validated model, *i. e.*, the one trained on the real lossless dataset \mathcal{G} over $\mathcal{S}_{\text{lossy}}$. Results of these experiments are reported in Fig 32 and Fig 33. In particular, the confusion matrix reported in the left side of Fig 32 shows the aggregated results over all loss models: we can observe that most of the classes have a true positive classification of 99% or 100% even in presence of 20% packet losses, and irrespectively of the actual loss pattern.

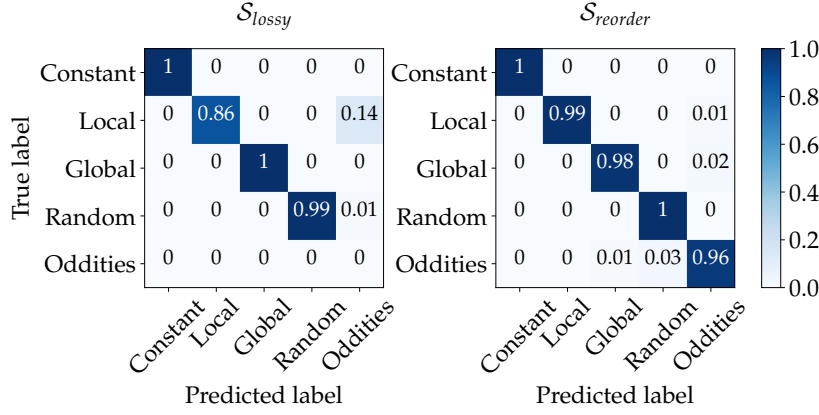


Figure 32: Robustness: (left) Confusion Matrix of a classifier trained on the real lossless dataset \mathcal{G} and tested on the synthetic lossy dataset \mathcal{S}_{lossy} with purposefully injected 20% packet losses on each sequence, (right) Confusion Matrix of a classifier trained on the real lossless dataset \mathcal{G} and tested on the dataset where 20% of each sequence is intentionally randomly swapped $\mathcal{S}_{reorder}$.

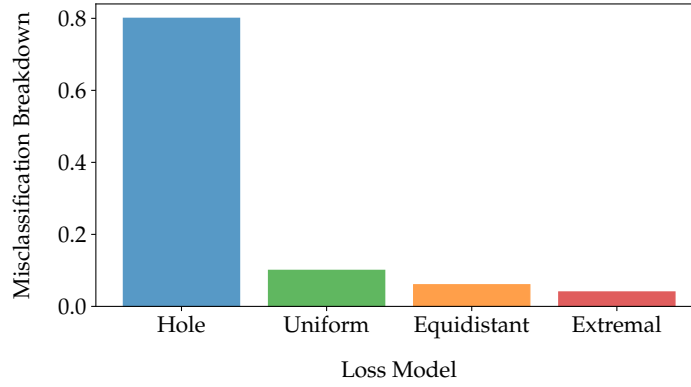


Figure 33: Robustness: Misclassification breakdown of the (local,odd) (14%) for the different loss models.

Additionally, we observe that in the case of the *local class*, only 86% of the sequences are correctly classified, whereas 14% of the local sequences in presence of heavy losses are erroneously classified as being part of the “odd” behavior class. Fig 33 dig further the reasons of this discrepancy, showing that the misclassification mostly happens for the *hole* loss model, while in the other cases is a very rare event. Recalling the odd behavior early shown in the plot of Fig. 25, we notice that this model induces a gap in the sequence, which is possibly large enough to be statistically similar to cases such as load balancing, where the sequence alternates among multiple counters. Overall, we find the classifier to be robust to very high loss rates and, with a single exception, also invariant to the actual loss pattern – which is a rather desirable property to operate the classifier into a real Internet environment. To investigate the effect of the presence of out-of-order packets received at the vantage point and of the reordering phenomena, we

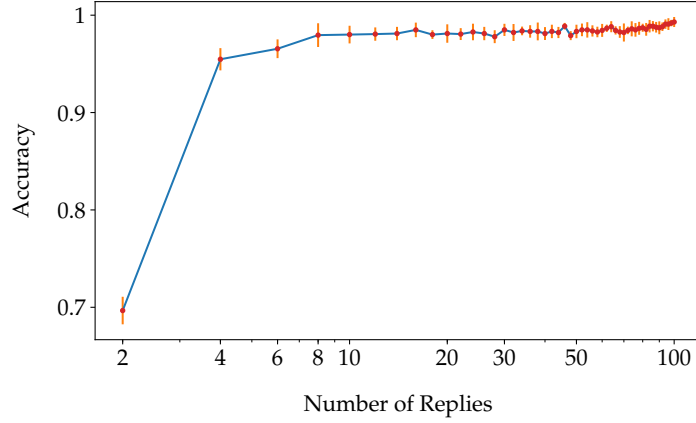


Figure 34: Probing Overhead analysis: Accuracy as a function of the sample set size

perform again the classification, with the decision tree classifier still trained on \mathcal{G} but tested on $\mathcal{S}_{\text{reorder}}$. We use again the confusion matrix as graphical way to highlight the quality of the classification. Results of these experiments are shown in the right matrix of Fig 32: we can observe that reordering does not affect at all constant and random labels classification and that the classifier is strong in recognizing the local and global behaviors leading to respectively 1% and 2% false positive misclassification.

Probing Overhead: We finally assess how large the number of samples N needs to be to have accurate classification results. In principle, features tabulated in Fig 20 are diverse enough so that we expect high accuracy even for very small values of N .

To assess this experimentally, we take the real lossless dataset \mathcal{G} and only consider that we have at our disposal only $N' < N$ out of the $N = 100$ samples gathered in the experiment. For each value of N' , we perform a 20-fold cross validation, training and validating with N' samples. We start from a minimum of $N' = 10$ (*i. e.*, 5 packets per host) up to the maximum of $N = 100$ (*i. e.*, 50 probes per host) samples. Fig 34 clearly shows that accuracy is already very high⁴ at 0.95 when $N' = 4$ and exceeds 0.99 when $N = 100$.

6.5 INTERNET CENSUS

The last step of the analysis consists in using the previously trained classifier over \mathcal{G} to classify the IP-ID behaviors present in the dataset \mathcal{L} . In this section, we first show the results of the classification and, then, we put them in perspective with those obtained

⁴ Notice that even in the extreme case with as few as $N' = 2$ packets, random and constant classification are correctly labeled, whereas the remaining global vs local cannot be discriminated, yielding to 0.70 accuracy in the \mathcal{G} set.

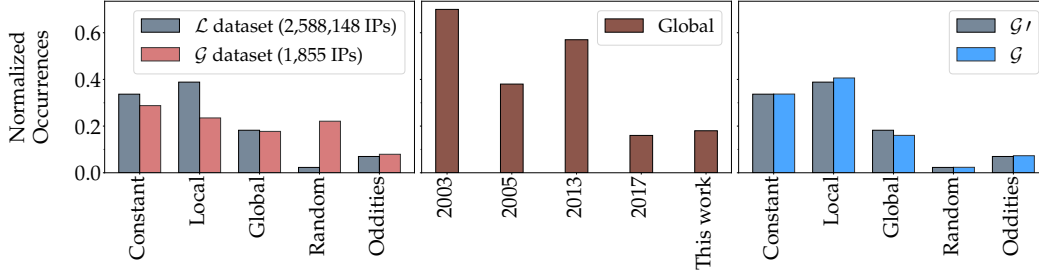


Figure 35: (a) Internet campaign: Normalized classes occurrences for the training \mathcal{G} and Internet-scale \mathcal{L} dataset; (b) Measured occurrences of Global IP-ID implementations over the years; (c) Breakdown of the classes of \mathcal{L} obtained with both \mathcal{G}' and \mathcal{G}

by related work. Then, we deeper investigate some aspects to see how different boundary conditions affect classification performances, as the impact of different training set choices or of the number of probe packets on the performances of the classification. Finally, we perform a spatial analysis and we deepen the analysis of the odd behaviors.

6.5.1 Longitudinal Comparison (over the years)

We apply our classifier in the wild, specifically on the previously introduced dataset \mathcal{L} (Sec. 6.3.3), made with the data collected through a large scale Internet measurement campaign. We observe that, while our classifier is able to perform a very accurate classification even with few samples, we need to deal with loss rates, which is unknown a priori. Hence, even though our probing overhead analysis in Sec. 6.4.2 revealed high accuracy for few number of samples, we prefer for the time being to use a simple and conservative approach and select $N = 100$ samples, being very accurate also in presence of very high loss rates. We apply the classification to batches of 100,000 hosts, and for each class c , we compute the relative breakdown of the class in that batch $\hat{n}_c = n_c / \sum_i n_i$, evaluating the confidence intervals of \hat{n}_c over the different batches. Results are reported in Fig. 35 (a), where we additionally report the breakdown in our \mathcal{G} training set comprising just 1,855 population samples: it can be seen that while \mathcal{G} has no statistical relevance for the census, it is not affected by class imbalance and thus proves to be a good training set.

Results are particularly interesting to put in perspective with current literature knowledge. Specifically, past work [36, 56, 99, 153] consistently reported the global counter to be more widespread: in 2003, [99] 70% ; in 2005, [36] 38%; in 2006, [153] affirms the global implementation to be the most common assignment policy; in 2013, [69] 57%. On the contrary, we find that only 18% (over 2,5 million targets) are still using global counter implementation: this in line with 2017 results that reports slightly more than 16% global IP-IDs [117] (whose main aim is to detect censorship in the Internet).

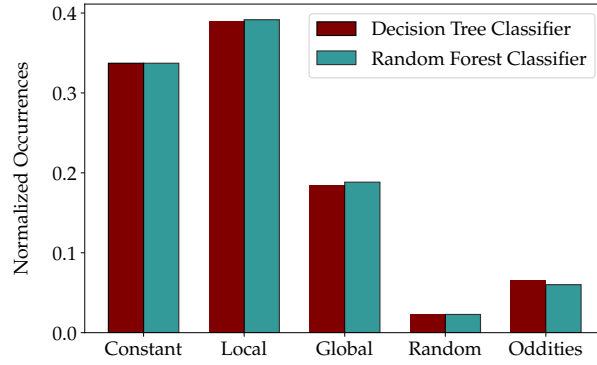


Figure 36: Breakdown of the classes of \mathcal{L} obtained with both a Decision Tree and a Random Forest Classifier.

This decreasing trend, summarized in Fig. 35 (b), is possibly affected by the comparably smaller population size of early studies. However, we believe this to be rooted into OS-level changes in IP-ID policy implementations: *e.g.*, Linux and Solaris, which previously adopted a global counter, for security reasons later moved to a local counter implementation [59].

By comparing our results with the only one providing the occurrences of both the normatives-compliant IP-ID behaviors and some odd practices [69], the 2013 study (our census) finds 57% (18%) global, 14% (39%) local and 9% (34%) constant IP-IDs, which testify of a significant evolution. Additionally, recalling that [69] suggests that 20% of DNS TLD generate *mixed* IP-IDs, we find out that this is much larger than the 7% fraction of the larger “odd” class (including but not limited to load balance) that we find in this work. Finally, despite 2012 recommendations [59], the percentage of random IP-ID sequence was (and remains) limited 1% (2%).

For completeness and in light of what showed in Sec. 6.4.1, we compare the results obtained with the Decision Tree those of the Random Forest. From the outcomes reported in Fig. 36 we can observe that no statistical difference is present in the two cases.

6.5.2 Sensitivity Analysis

Training Set Choice: In order to prove the independence of the results from the choice of the training dataset we exploit the second manually validated dataset \mathcal{G}' , which satisfies the previously described requirements and it is purposely biased, as it contains 75% of the samples from the same /8, which is something not desirable from a IP coverage point of view.

We then use these two datasets to classify the IP-ID behaviors in the whole large scale dataset \mathcal{L} covering the all the responsive IP addresses of the full hitlist. Results, shown in Fig. 35 (c) confirms indeed the validity of our methodology since, statistically, there are only slight differences between the occurrences breakdown when the classifier is trained on \mathcal{G} or \mathcal{G}' . Both datasets yield to consistent results ensuring the independence of the model from the training dataset and proving that as long as the behaviors are balanced the IP-prefix level imbalance is irrelevant.

Lightweight Census: Additionally, we may want to further investigate how the classification results change when we have a fewer number of packets building the IP-ID series that we aim at classify. This is important since we want to avoid injecting useless traffic in the network. Similarly to what previously described in Sec. 6.4.2, we take the measurements dataset \mathcal{G} and only consider that we have at our disposal only the first $N' = 10 < N$ out of the $N = 100$ samples gathered in the full experiment. Given that in this case we are only looking at a small portion of the collected series, we may expect that in this case we can have a loss in terms of amount of *oddities* really present in the dataset, and behaviors like the one depicted in Fig. 25 might not be correctly classified, simply due to lack of information about it. In fact, in this case, it is possible that the *jump* of the IP-ID counter occurs later in the sequence, so all the features are evaluated on a resembling simple counter. What practically happens in reality confirms the expectations: about half of the *oddities* are spread between the *global* and *local* implementations. What is instead more surprising is the substantial decrease for the population *random* class. This might be due again to the lack of fundamental information to correctly classify those behaviors. Conversely, constant behaviors are easy to be identified even with a bunch of few packets. These results show that to correctly detect *random* and *odd* IP-ID classes more care is needed: more data might be required to spot the proper behavior of the series. Whilst, for the other classes, few packets are more than enough to correctly classify them.

6.5.3 Spatial analysis

Odd class: As already mentioned, during the manual labelling phase we discovered some targets setting the IP-ID in not-standard unexpected manners, which may be ascribable to different causes, and that we named as *odd* behaviors. In this section we try to investigate a bit more the *odd* class, trying to figure out whether we can re-map some of those IP addresses in other classes or not. The first analysis we perform consist in

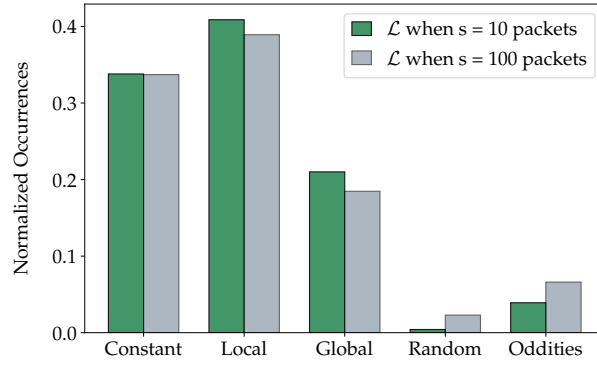


Figure 37: Normalized classes occurrences for \mathcal{L} and its lighter version when only $N=10$ packets out of 100 are considered.

converting the interpretation of the bytes contained in the IP-ID IPv4 header field to little endian and try to perform again the classification to check if results change. We focus only on the 172,679 IP addresses in \mathcal{L} previously classified as *odd* and perform byte swapping to each IP-ID value of the x, y series. Then, we re-build the dataset with the new features and operate the classifier trained on \mathcal{G} on it. Results show that no meaningful change has occurred, since, except for a negligible amount of IP addresses becoming *global*, almost all the IP-ID series remain *odd*.

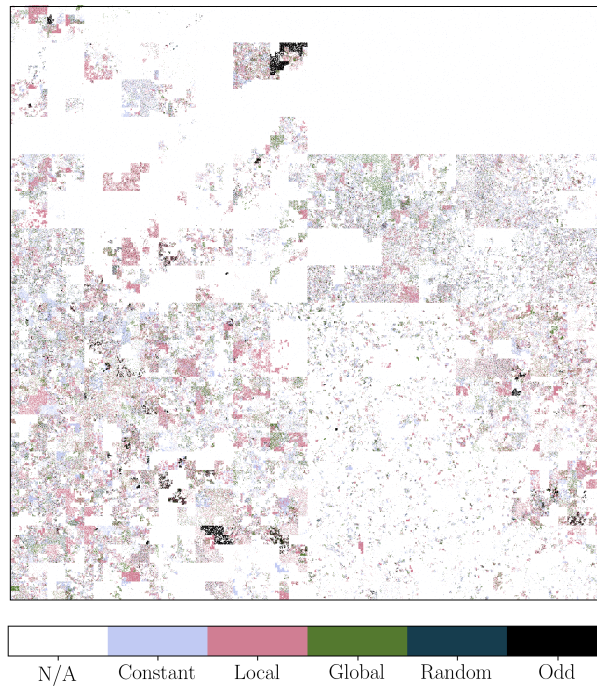


Figure 38: IP-ID census results, shown as a 12th order Hilbert curve, a fractal space-filling curve that allows the mapping of the one-dimensional IPv4 address space into a bi-dimensional image.

Results on the IPv4 address space: Next, we want to visualize the IP-ID classes distribution on the IPv4 address space. A functional way to graphically perform this task is through a 12th-order Hilbert curve, a fractal space-filling curve which allows the mapping of the one-dimensional IPv4 address space into a bi-dimensional image. The use of Hilbert curves to compactly represent Internet-wide characteristics was first popularized by the Xkcd comic [4] and then used ever since. Each pixel in the image depicted in Fig. 38 represents a single /24 prefix block and its color can range among six different hues. Five of these refer to the five IP-ID classes and are respectively assigned to the pixel if one representative address of that /24 network is part of our analysis, *e. g.*, it belongs to \mathcal{L} , and the model has classified it as the corresponding color label. On the contrary if there are no IP addresses in \mathcal{L} belonging to that /24 the associated pixel is coloured white. From the image it is clearly possible to highlight some easily distinguishable *islands* of close IP addresses which implement the IP-ID in the same way. However, this is not an exhaustive result to assess that the hosts whose IP addresses belong to the same prefix block generates IP-ID in the same manner.

AS aggregation: Finally, we inspect the spatial aggregation of the IP addresses per Autonomous System. We perform this by querying Team Cymru whois database [38] and collecting from there information about the 49189 ASes of the the IP addresses present in our dataset \mathcal{L} . We focus only on the 32994 ASes owning at least two IP addresses of the list, discarding in this way 16k IP addresses. We evaluate then the standard deviation σ of the IP-ID classes of the IP addresses belonging to the same AS. We find out, as shown in Fig. 39, that 29% of the ASes own IP addresses from whom we collected packets containing the IP-ID generated in the same way (standard deviation $\sigma = 0$). This result is not telling much if considered alone, and since the most popular class is about 40% of the total this could just be equal to a random clustering of the IP addresses.

6.6 CONCLUSIONS

This chapter presents, to the best of our knowledge, the first systematic study of the prevalence of different IP-ID behaviors in the current IPv4 Internet (extending this work to IPv6 is a future, necessary, work). In this work, we find evidence that local and constant implementations of the IP-ID are prevalent: this is in contrast with common knowledge [36, 53, 69, 99, 105, 153], from which the global counter was expected, even in recent times, to be the most popular IP-ID implementation.

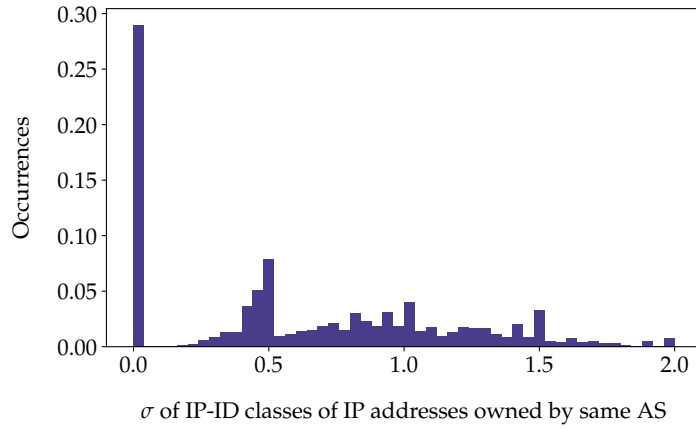


Figure 39: Standard Deviation of IP-ID classes of IP addresses owned by same AS

Summary and Perspectives: We proposed a framework to robustly classify the different IP-ID behaviours with only a handful of IP packets.

The data collection block relies on an experimental testbed comprising one sender and two receivers, which collect the IP packets and, specifically, the information related to the IP-ID field. The sender sends a burst of packets, minimizing the impact of external traffic and purposely exploiting spoofing to precisely alternate addresses in the sequence.

To classify the different IP-ID classes, we trained and validated different classifiers on datasets gathered from real measurements and additionally tested in the presence of controlled losses to assess its robustness. Training of the model required manual validation of thousands of sequences: during this phase, we also discovered some odd behaviour, not documented in any of the previous RFCs, and which may be attributed to different causes.

In instances where odd behaviour was previously reported, our classifier is the first to automatically and correctly label such instances, making it easier to perform large-scale analysis over the Internet. Moreover, classification only requires a handful of packets, making the methodology extremely lightweight.

Experimental results show that the majority of hosts adopt local IP-IDs (39%) or a constant counter (34%) of which:

- A fraction of global counters (18%) is significantly lower than expected;
- A non-marginal number of hosts have an odd behaviour (7%);
- Random IP-IDs are only slightly more than an exception (2%).

This outcome provides a picture of Internet-wide adoption of the different IP-ID implementations. Indeed, we gather that the 18% breakdown of the global implementation in 2017 is three times lower with respect to the 57% reported in 2013 [69]. While the quantitative reduction is in line with the statistics reported by recent work that leverages global IP-ID behaviour to detect censorship in the Internet [117], one could have expected the decrease in global implementation to be compensated by an increase of random IP-IDs, which is not the case.

Contributions: Our first contribution is to devise an accurate, lightweight and robust classifier: accuracy of the classifier follows from a principled definition of the statistical features used to succinctly describe the IP-ID sequence; robustness is a consequence of this choice, as features remains wide apart even under heavy losses.

Our second contribution is to carry on a manual investigation effort for a moderate size dataset coming from real Internet measurements: this valuable ground truth allow us to adopt a supervised classification techniques to train a model able not only to detect well-defined behaviors, but also to correctly recognize a wide range of odd behaviors.

Finally, all our datasets, including the testing with manual ground truth, as well as the results of our census, are publicly available at [131]: we hope that the former can assist scientists to build and test new techniques for IP-ID classification, whereas the latter provides practitioners with readily usable lists of the hosts with global IP-ID implementations for their inference. Specifically, the available readily usable list of the approximate half million hosts with global IP-ID implementations global implementations [131] can make work such as [15, 36, 117, 141] still possible. Moreover, by updating and consolidating the scattered knowledge [36, 56, 99, 117, 153] of IP-ID prevalence, this work contributes in refining the current global Internet map.

CONCLUSIONS

Contents

7.1	Summary of our contributions	105
7.1.1	Users' acceptance in the Wild Web	105
7.1.2	User Perceived Page Load Time in controlled experiments	106
7.1.3	The fairness of models trained on the Web	107
7.1.4	Supervised learning to infer machine-generated content .	107
7.2	Future work	108

Over recent years, the technological development and pervasiveness of the Internet and, in parallel, the rise and massive spread of machine learning algorithms exposed researchers to a number of different new challenges.

This manuscript offered a walk through different studies aimed at providing an overview of some of the most prominent aspects of modern Internet. Here we followed two main research directions: first, we focused on the analysis of Web users' Quality of Experience (Ch. 3 and Ch. 4); second, we gave emphasis on the use of machine learning applied to Internet measurements, studying, on one side, the impact of its interaction with the Web (Ch. 5); and, on the other, its use to predict objects generated by machines (Ch. 6).

We next summarize the achievements of this thesis work and discuss possible future research directions.

7.1 SUMMARY OF OUR CONTRIBUTIONS

In the first part of this thesis we focused on the measurement and on the analysis of the Quality of Experience of users browsing the Web.

7.1.1 *Users' acceptance in the Wild Web*

In Chapter 3 we tackled the problem of assessing the quality of experience on a popular website in operation. Specifically, we do this by gathering the user acceptance of Wikipedia, over 62k user answers, more than twice the survey responses collected in similar large-scale Wikipedia studies. We collect either positive (84.8%), neutral (7.7%)

or negative (7.5%) experiences, each one associated with over 100 features. The collected dataset is interesting per se, as it is particularly heterogeneous, comprising 59k distinct IP addresses, 45 browsers, 3.8k ISPs and 2.7k hardware devices. A portion of it, including 19 features which ensures that no sensitive information allowing to deanonymize Wikipedia visitors is present, is made available to the research community as we hope this can help in refining the understanding of Web users' experience.

This chapter presents some important results. First, we observe that the concerned Wikipedia users are consistently satisfied, and that not only user answers unexpectedly do not exhibit seasonality at circadian or weekly timescales but also they are not affected by network-related events, which are typically inducing measurable delay changes.

Second, we find evidence of spatial dependency across many of the collected features: particularly, we observe that user scores are influenced by user-level expertise and equipment as well as network and country-level characteristics.

Finally, we observe that supervised data-driven models of user experience, used to predict the user scores still falls short from attaining satisfactory performance in operational settings. This occurs despite when they include performance metrics considered to be the state-of-the art of Web QoE and even when reducing the variance and heterogeneity of the data.

7.1.2 *User Perceived Page Load Time in controlled experiments*

In Chapter 4 we focused instead on measuring Quality of Experience in controlled experiments. Particularly, we crowdsource the user perceived page load time, the time when a user considers a webpage to be loaded and ready to browse, on 108 webpages via the Eyeorg platform. Through the Eyeorg's timeline experiment, participants are shown a video of webpage load and asked to scrub it until when (s)he considers the page to be ready. Similarly to Chapter 3, also in this work we gather both *objective* and *subjective* Web quality metrics. This chapter showed two main results.

First, we observe that half of the webpages involved in our study present a multi-modal uPLT distribution and that, in practice, three modes are sufficient to accurately describe the uPLT distribution.

Second, we find that the number of images and the number of objects in a webpage can help in predicting the uPLT modality.

In spirit with the current trends toward research reproducibility, also this dataset is made publicly available.

* * *

In the second part of the thesis we presented some results related to the interaction of machine learning models with (i) Web and (ii) machine-generated content.

7.1.3 *The fairness of models trained on the Web*

In Chapter 5 we question the assumption that the Web sheer size and heterogeneity ensure the fairness of the models trained on Web-based content. Here, we propose a methodology for the evaluation of the fairness of state-of-the-art transformer-based language models. This chapter showed two main results, related to the fact that different models encode diverse biases when used for the prediction of terms of both Standard American English and African American English.

First, we observe that the fairness of large models, trained on huge amount of Web-based content, is unbalanced. Specifically, the predictions done with BERT and DistilBERT on Standard American English are up to 21% more accurate with respect to those done on African American English. We show also how instead BART, RoBERTa and DistilRoBERTa exhibit an opposite bias, favouring then African American English.

Second, results highlight that the distilled variants of BERT and RoBERTa, designed to be lighter and trained on a lower amount of data, are the fairest among the seven tested language models.

7.1.4 *Supervised learning to infer machine-generated content*

In Chapter 6 we observe instead how state-of-the-art machine learning algorithms behave when they are trained on content generated by machines, *i. e.*, the IP identification (IP-ID) field of the IPv4 header. Despite being only minimally intrusive and fairly lightweight, the proposed technique is significantly accurate and unveils two main findings, the first one more related to the methodology, the second one more tied instead to the classification results.

From the methodology perspective, we find that few scalar features and a simple classifier, as a decision tree, are enough to accurately predict the different IP-IDs implementations. This is in sharp contrast with the results shown in Chapter 3. Here the classification of objects with very pre-determined behaviors, depending on the specific implementation of the OS, leads to very accurate predictions.

From the point of view of the results, instead, the application of this technique to an Internet-wide census provides an updated view of the adoption of the different known IP-ID implementations in the wild. Particularly, the results of the census reveals that the *global* is no longer the most common IP-ID implementation and that, instead, other be-

haviors, as *local* and *constant*, are present. This is particularly relevant, being the *global* implementation of invaluable help to infer a wealth of information concerning the network. Releasing all our datasets and results publicly, including two manually labeled ground truth datasets and a list of the approximate half million hosts with the IP-ID implementations, can make works relying on this IP-ID class still possible.

7.2 FUTURE WORK

There are several ways in which the results discussed in this thesis could be extended. We next present some perspectives that could be accomplished as future work.

As concerns the Web QoE, different directions can be further explored. With regards to the experiment carried out in Chapter 3, we must point out that there exist other QoE influence factors that we did not include in the analysis, such as content and context factors that are known to affect user QoE. For instance, the sentiment linked to the topic and the content of the page or more informative indicators about the context in which the measurements are carried out, as the earlier user browsing experience, heavily impact QoE. However, they are hard to capture. Moreover, adding the knowledge of whether the rendered element is under the user gaze, using mouse-movements as a proxy of eye gaze activity, can help further refining QoE metric in the spatial direction. Clearly, further research is needed on whether user-touch can be useful for similar purposes in case of mobile handsets.

As for the bias embedded in models trained on Web based corpora, different debiasing techniques can be thought and implemented. Besides traditional debiasing approaches, as those based on the loss function modification, research directions which rely on ensemble methods could be explored. Ensemble learning might lead to fairer prediction outcomes, by combining language models which embed opposite biases, as, for instance, BERT_{cased} and BART in our analysis. Moreover, distilled language models need further investigations. Particularly, a special emphasis should be given on the study of the causes which lead them to have fairer predictions with respect to their teacher models. This could pose new research questions, such as the comparison of the cost-benefit analysis of large base models with respect to the smaller distilled counterparts.

Part I

APPENDIXES

APPENDIX CH. 2

Table 23: List of features, informing whether each comes from *raw* data or is *derived*, and if it is present in the T, WWW or PA set.

Feature Class	Feature Name	Raw	Derived	T	WWW	PA
Page	recvfrom	✓		✓		
	revision	✓		✓		
	wiki	✓		✓	✓	✓
	seqid	✓		✓		
	schema	✓		✓		
	pageid	✓		✓		
	pagetitle	✓		✓		
	skin	✓		✓		
	survey code name	✓		✓		
	revid	✓		✓		
	transfersize	✓		✓		
	page size quantized		✓	✓	✓	
	page size plugin		✓	✓		
	tot num. objects		✓	✓		
	survey viewtime	✓		✓		
	connectEnd	✓		✓	✓	✓
	connectStart	✓		✓	✓	✓
	dnsLookup	✓		✓	✓	✓
	domComplete	✓		✓	✓	✓
	domInteractive	✓		✓	✓	✓
	fetchstart	✓		✓	✓	✓

firstpaint	✓		✓	✓	✓
gaps	✓		✓	✓	✓
loadEventEnd	✓		✓	✓	✓
loadEventEtart	✓		✓	✓	✓
mediawikiLoadEnd	✓		✓	✓	✓
redirectcount	✓		✓	✓	
redirecting	✓		✓	✓	✓
requestStart	✓		✓	✓	✓
responseEnd	✓		✓	✓	✓
responseEtart	✓		✓	✓	✓
rsi	✓		✓	✓	✓
secureConnectionStart	✓		✓	✓	✓
unload	✓		✓	✓	✓
connectduration		✓	✓	✓	
responseduration		✓	✓	✓	
plt		✓	✓	✓	
tti		✓	✓	✓	
country speed avg		✓	✓		
speed		✓	✓		
speed quantized		✓	✓	✓	
country speed ratio		✓	✓		
country speed delta		✓	✓		
country speed relative		✓	✓		
speed over median per country		✓	✓	✓	
$(PLT \cdot 0.9) + (TTI \cdot 0.1)$		✓	✓	✓	
log(RSI)		✓	✓		
IP	✓		✓		
uuID	✓		✓		
editcountbucket	✓		✓	✓	
isloggedin	✓		✓	✓	

	istablet	✓		✓	✓	
	platform	✓		✓	✓	
	userlanguage	✓		✓	✓	
	browser family	✓		✓		
	device family	✓		✓		
	os family	✓		✓		
	browser family sanitized		✓	✓	✓	
	device family sanitized		✓	✓	✓	
	os family sanitized		✓	✓	✓	
	RAM device		✓	✓	✓	
	price device		✓	✓	✓	
	mobileview		✓	✓		
	webhostMobile		✓	✓		
	browserMobile		✓	✓		
	osMobile		✓	✓		
	isMobile		✓	✓		
	ASN		✓	✓	✓	
	ISP		✓	✓	✓	
	PLT _{decile1}		✓	✓	✓	
	PLT _{decile2}		✓	✓	✓	
	PLT _{decile3}		✓	✓	✓	
	PLT _{decile4}		✓	✓	✓	
	PLT _{decile5}		✓	✓	✓	
	PLT _{decile6}		✓	✓	✓	
	PLT _{decile7}		✓	✓	✓	
	PLT _{decile8}		✓	✓	✓	
	PLT _{decile9}		✓	✓	✓	
	count		✓	✓		
	datetime	✓		✓		
	year	✓		✓	✓	

month	✓		✓	✓	
day	✓		✓		
hour	✓		✓	✓	
country code	✓		✓		
namespaceid	✓		✓		
webhost	✓		✓		
city	✓		✓		
continent	✓		✓		
country	✓		✓	✓	
country code	✓		✓		
latitude	✓		✓		
longitude	✓		✓		
postal code	✓		✓		
subdivision	✓		✓		
timezone	✓		✓		
action	✓		✓		
isanon	✓		✓		
isoversample	✓		✓		
mediawiki version	✓		✓		
mobilemode	✓		✓	✓	
effective connection type	✓		✓	✓	
day of week		✓	✓	✓	
day of year		✓	✓	✓	
day of month		✓	✓	✓	
week of year		✓	✓	✓	
GDP		✓	✓		
GDP rank		✓	✓	✓	
GDP percapita		✓	✓		
Device Price over GDP		✓	✓		
Device Price over GDP per capita		✓	✓	✓	

Environment

GDP rank BIN		✓	✓		
unix timestamp		✓	✓		
unix timestamp normalized to minutes		✓	✓		
datetime normalized to hour		✓	✓		✓ ¹

Table 24: Schema of the Public Available Features.

¹ provided in a separate dataset together with the user survey answer, where the order of the entries has been randomly shuffled in order to prevent user deanonymization

wiki	Which wiki the request was on (ruwiki, cawiki, eswiki, frwiki or enwikivoyage)
unload ²	The time spent on unload (unloadEventEnd - unloadEventStart).
redirecting ⁵	Time spent following redirects.
fetchStart ⁵	The time immediately before the user agent starts checking any relevant application caches.
dnsLookup ⁵	Time it took to resolve names (domainLookupEnd - domainLookupStart).
secureConnectionStart ⁵	The time immediately before the user agent starts the handshake process to secure the current connection.
connectStart ⁵	The time immediately before the user agent start establishing the connection to the server to retrieve the document.
connectEnd ⁵	The time immediately after the user agent finishes establishing the connection to the server to retrieve the current document.
requestStart ⁵	The time immediately before the user agent starts requesting the current document from the server, or from relevant application caches or from local resources.
responseStart ⁵	The time immediately after the user agent receives the first byte of the response from the server, or from relevant application caches or from local resources.
responseEnd ⁵	The time immediately after the user agent receives the last byte of the current document or immediately before the transport connection is closed, whichever comes first.
loadEventStart ⁵	The time immediately before the load event of the current document is fired.
loadEventEnd ⁵	The time when the load event of the current document is completed.
mediawikiLoadEnd	(Mediawiki-specific.) The time at which all ResourceLoader modules for this page have completed loading and executing.

domComplete ⁵	The time immediately before the user agent sets the current document readiness to "complete".
domInteractive ⁵	The time immediately before the user agent sets the current document readiness to "interactive".
gaps ⁵	The gaps in the Navigation Timing metrics. Calculated by taking the sum of: domainLookupStart - fetchStart, connectStart - domainLookupEnd, requestStart - connectEnd and loadEventStart - domComplete.
firstPaint ³	The time when something is first displayed on the screen.
rsi ⁴	RUMSpeedIndex. Estimate of the SpeedIndex value based on ResourceTiming data. Now moved to the RUMSpeedIndex Event-Logging schema, but was collected as part of the NavigationTiming schema at the time of the study.

² metrics coming from the browsers' implementation of the NavigationTiming API

³ firstPaint comes from the **Paint Timing API** or vendor-specific implementations predating the standards.

⁴ RUMSpeedIndex is a compound metric combining several NavigationTiming and ResourceTiming (**level 1** and **level 2**) metrics into a single score. It's a 3rd-party FLOSS library found here: <https://github.com/WPO-Foundation/RUM-SpeedIndex>

BIBLIOGRAPHY

- [1] [Online]. Available from: <https://www.alexa.com/topsites>.
- [2] [Online]. Available from: <https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index>.
- [3] [Online]. Available from: <https://webqoe.telecom-paristech.fr/data/>.
- [4] [Online]. Available from: <https://xkcd.com/195/>.
- [5] [Online]. Available from: <https://bit.ly/2VW2Gnd>. 2019.
- [6] E. Aguiar et al. “Video quality estimator for wireless mesh networks.” In: *2012 IEEE 20th International Workshop on Quality of Service*. 2012.
- [7] Tim Althoff et al. “Harnessing the web for population-scale physiological sensing: A case study of sleep and performance.” In: *Proc. The World Wide Web Conference (WWW)*. 2017.
- [8] *Apache Hive*. [Online]. Available from: <https://hive.apache.org/>.
- [9] Approximate ATF chrome extension. [Online]. Available from: <https://chrome.google.com/webstore/detail/approximate-atf/eedmonedcfjniaagehchbkdolbobmfhb>.
- [10] A. S. Asrese et al. “Measuring Web Latency and Rendering Performance: Method, Tools, and Longitudinal Dataset.” In: *IEEE Transactions on Network and Service Management* 16.2 (2019), pp. 535–549.
- [11] Alemnew Asrese et al. “Measuring Web Quality of Experience in Cellular Networks.” In: *Proc. Passive and Active Measurement Conference (PAM)*. 2019.
- [12] Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. “Designing and Deploying Online Field Experiments.” In: *Proc. The World Wide Web Conference (WWW)*. 2014.
- [13] Athula Balachandran et al. “Modeling Web Quality-of-experience on Cellular Networks.” In: *Proc. ACM MOBICOM*. ACM, 2014.
- [14] Christine Basta, Marta R. Costa-jussà, and Noe Casas. “Evaluating the Underlying Gender Bias in Contextualized Word Embeddings.” In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2019.
- [15] Steven M Bellovin. “A technique for counting NATted hosts.” In: *Proc. Internet Measurement Workshop (IMW)*. 2002.

- [16] Adam Bender, Rob Sherwood, and Neil Spring. “Fixing ally’s growing pains with velocity modeling.” In: *Proc. ACM Internet Measurement Conference (IMC)*. 2008.
- [17] Emily M. Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 2021.
- [18] Robert Beverly et al. “Measuring and characterizing IPv6 router availability.” In: *Proc. Passive and Active Measurement Conference (PAM)*. 2015.
- [19] Su Lin Blodgett and Brendan O’Connor. “Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English.” In: *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*. 2017.
- [20] Enrico Bocchi, Luca De Cicco, and Dario Rossi. “Measuring the Quality of Experience of Web Users.” In: *ACM SIGCOMM Workshop on Internet-QoE’16*. 2016.
- [21] Enrico Bocchi et al. “The Web, the Users, and the MOS: Influence of HTTP/2 on User Experience.” In: *Proc. Passive and Active Measurement Conference (PAM)*. 2017.
- [22] Tolga Bolukbasi et al. “Man is to Computer Programmer as Woman is to Home-maker? Debiasing Word Embeddings.” In: *Proc. Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc., 2016, 4356–4364.
- [23] R. Braden. *RFC 1122, Requirements for Internet Hosts – Communication Layers*. 1989.
- [24] Leo Breiman. “Random Forests.” In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 1573-0565.
- [25] J. Brutlag, Z. Abrams, and P. Meenan. “Above the fold time: Measuring web page performance visually.” In: *Velocity: Web Performance and Operations Conference*. 2011.
- [26] Jake Brutlag. *Speed matters for Google web search*. 2009.
- [27] Jake Brutlag, Zoe Abrams, and Pat Meenan. *Above the fold time: Measuring Web page performance visually*. [Online]. Available from: <http://conferences.oreilly.com/velocity/velocity-mar2011/public/schedule/detail/18692>.
- [28] Michael Butkiewicz, Harsha V Madhyastha, and Vyas Sekar. “Understanding website complexity: measurements, metrics, and implications.” In: *Proc. ACM Internet Measurement Conference (IMC)*. ACM. 2011, pp. 313–328.

- [29] Michael Butkiewicz et al. "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, pp. 439–453.
- [30] Rena Torres Cacoullos and Catherine E Travis. *Bilingualism in the Community: Code-switching and Grammars in Contact*. Cambridge University Press, 2018.
- [31] P. Casas et al. "When YouTube Does not Work—Analysis of QoE-Relevant Degradation in Google CDN Traffic." In: *IEEE Transactions on Network and Service Management* 11.4 (2014), pp. 441–457.
- [32] P. Casas et al. "Next to You: Monitoring Quality of Experience in Cellular Networks From the End-Devices." In: *IEEE Transactions on Network and Service Management* 13.2 (2016), pp. 181–196.
- [33] Vint Cerf et al. "BufferBloat: what's wrong with the internet?" In: *Communications of the ACM* 55.2 (2012), pp. 40–47.
- [34] Rakesh Chada. "Gendered Pronoun Resolution using BERT and an Extractive Question Answering Formulation." In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2019.
- [35] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In: *Proc. of the SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.
- [36] Weifeng Chen, Yong Huang, Bruno Ribeiro, et al. "Exploiting the IPID field to infer network path and end-system characteristics." In: *Proc. Passive and Active Measurement Conference (PAM)*. 2005.
- [37] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.
- [38] Team Cymru. *IP to ASN mapping*. [Online]. Available from: <http://www.team-cymru.org/IP-ASN-mapping.html>.
- [39] Diego Da Hora et al. "Narrowing the gap between QoS metrics and Web QoE using Above-the-fold metrics." In: *Proc. Passive and Active Measurement Conference (PAM)*. 2018.
- [40] Alberto Dainotti et al. "Lost in Space: Improving Inference of IPv4 Address Space Utilization." In: *IEEE Journal on Selected Areas in Communications (J-SAC)* (2016).
- [41] Mallesh Dasari et al. "Impact of Device Performance on Mobile Internet QoE." In: *Proc. ACM Internet Measurement Conference (IMC)*. 2018.

- [42] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. "Racial Bias in Hate Speech and Abusive Language Detection Datasets." In: *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, 2019.
- [43] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.
- [44] Rachel Dorn. "Dialect-Specific Models for Automatic Speech Recognition of African American Vernacular English." In: *Proceedings of the Student Research Workshop Associated with RANLP 2019*. INCOMA Ltd., 2019.
- [45] John W Du Bois et al. *Santa barbara corpus of spoken american english*. 2000. URL: <https://www.linguistics.ucsb.edu/research/santa-barbara-corpus/>.
- [46] Zhiheng Wang (Ed.) "Navigation Timing." In: *W3C Recommendation*. 2012.
- [47] S. Egger et al. "Waiting times in quality of experience for web based services." In: *2012 Fourth International Workshop on Quality of Multimedia Experience*. 2012, pp. 86–96.
- [48] Sebastian Egger-Lampl et al. "Time is Bandwidth? Narrowing the Gap between Subjective Time Perception and Quality of Experience." In: *Proc. IEEE International Conference on Communications (ICC)*. 2012.
- [49] Theresa Enghardt, Thomas Zinner, and Anja Feldmann. "Web Performance Pitfalls: Methods and Protocols." In: *Proc. Passive and Active Measurement Conference (PAM)*. 2019.
- [50] Jeffrey Erman et al. "Towards a SPDY'ier Mobile Web?" In: *Proc. ACM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*. 2013.
- [51] Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia. "A generic quantitative relationship between quality of experience and quality of service." In: *IEEE Network* 24.2 (2010), pp. 36–41.
- [52] P. A. Frangoudis, L. Yala, and A. Ksentini. "CDN-As-a-Service Provision Over a Telecom Operator's Cloud." In: *IEEE Transactions on Network and Service Management* 14.3 (2017), pp. 702–716.
- [53] K. Sandlund G. Pelletier. *RFC 5225, RObust Header Compression Version 2 (ROHCv2): Profiles for RTP, UDP, IP, ESP and UDP-Lite*. 2008.

- [54] Qingzhu Gao, Prasenjit Dey, and Parvez Ahammad. “Perceived Performance of Top Retail Webpages In the Wild: Insights from Large-scale Crowdsourcing of Above-the-Fold QoE.” In: *Proc. of the Workshop on QoE-based Analysis and Management of Data Communication Networks*. ACM. 2017, pp. 13–18.
- [55] *GDP per capita by country*. [Online]. Available from: <https://github.com/secure411dotorg/GDP-per-Capita-by-Country>.
- [56] Yossi Gilad and Amir Herzberg. “Fragmentation considered vulnerable.” In: *ACM TISSEC* (2013).
- [57] *Global digital population as of January 2021*. 2021.
- [58] Aaron Gokaslan and Vanya Cohen. *Openwebtext corpus*. 2019.
- [59] F. Gont. *RFC 6274, Security Assessment of the Internet Protocol Version 4*. 2011.
- [60] F. Gont. *RFC 7739, Security Implications of Predictable Fragment Identification Values*. 2016.
- [61] Paul C Gorski. *Reaching and teaching students in poverty: Strategies for erasing the opportunity gap*. Teachers College Press, 2017.
- [62] Lisa J. Green. “Introduction.” In: *African American English: A Linguistic Introduction*. Cambridge University Press, 2002, pp. 1–11.
- [63] Lisa J. Green. “Syntax part 1: verbal markers in AAE.” In: *African American English: A Linguistic Introduction*. Cambridge University Press, 2002, 34–75.
- [64] *GSM arena*. [Online]. Available from: <https://www.gsmarena.com>.
- [65] Isobel Asher Hamilton. *Why It’s Totally Unsurprising That Amazon’s Recruitment AI Was Biased against Women*. [Online]. Available from: <https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10>. 2018.
- [66] J. A. Hartigan and P. M. Hartigan. “The Dip Test of Unimodality.” In: *The Annals of Statistics* 13.1 (Mar. 1985), pp. 70–84.
- [67] Tarek A Hassan et al. “Firm-level political risk: Measurement and effects.” In: *The Quarterly Journal of Economics* 134.4 (2019), pp. 2135–2202.
- [68] J. Heidemann et al. “Census and Survey of the Visible Internet.” In: *Proc. ACM Internet Measurement Conference (IMC)*. 2008.
- [69] Amir Herzberg and Haya Shulman. “Fragmentation considered poisonous, or: One-domain-to-rule-them-all. org.” In: *IEEE CCNS*. 2013.

- [70] Tobias Hossfeld et al. “QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS.” In: *Quality and User Experience* 1.1 (2016), p. 2.
- [71] Jocelyn Huang et al. “Cross-Language Transfer Learning, Continuous Learning, and Domain Adaptation for End-to-End Automatic Speech Recognition.” In: *arXiv preprint arXiv:2005.04290* (2020).
- [72] Dan Hubbard. *Cisco Umbrella 1M*. (2016). [Online]. Available from: <https://umbrella.cisco.com/blog/blog/2016/12/14/cisco-umbrella-1-million>. 2016.
- [73] Alexis Huet et al. “Web Quality of Experience from Encrypted Packets.” In: *ACM SIGCOMM Posters and Demos*. 2019.
- [74] Ben Hutchinson et al. “Social Biases in NLP Models as Barriers for Persons with Disabilities.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- [75] *Idle scanning and related IPID games*. <https://nmap.org/book/idlescan.html>.
- [76] ITU-T. *Estimating end-to-end performance in IP networks for data application*. Recommendation. 2014.
- [77] ITU-T. *QoE factors in web-browsing*. Recommendation. 2014.
- [78] Sharad Jaiswal et al. “Measurement and classification of out-of-sequence packets in a tier-1 IP backbone.” In: *IEEE/ACM TON* (2007).
- [79] Angwin Julia et al. *Machine Bias*. [Online]. Available from: <https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10>. 2016.
- [80] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. “End-to-End Bias Mitigation by Modelling Biases in Corpora.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- [81] Conor Kelton et al. “Improving User Perceived Page Load Times Using Gaze.” In: *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*. 2017, pp. 545–559.
- [82] Tyler Kendall and Charlie Farrington. *The corpus of regional african american language*. 2018. URL: <http://lingtools.uoregon.edu/coraal/>.
- [83] Ken Keys et al. “Internet-scale IPv4 alias resolution with MIDAR.” In: *IEEE/ACM TON* (2013).
- [84] Amit Klein. *OpenBSD DNS Cache Poisoning and Multiple O/S Predictable IP ID Vulnerability*. Tech. rep. 2007.

- [85] Farshad Kooti et al. “Evolution of Conversations in the Age of Email Overload.” In: *Proc. The World Wide Web Conference (WWW)*. 2015.
- [86] Keita Kurita et al. “Measuring Bias in Contextualized Word Representations.” In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2019.
- [87] Zhenzhong Lan et al. “Albert: A lite bert for self-supervised learning of language representations.” In: *Proceedings of the 2020 International Conference on Learning Representations*. 2020.
- [88] Mike Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- [89] Paul Pu Liang et al. “Towards Debiasing Sentence Representations.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- [90] J. Liddle. *Amazon Found Every 100ms of Latency Cost Them 1% in Sales*. <http://blog.gigaspace.com/amazon-found-every-100ms-of-latency-cost-them-1-in-sales>.
- [91] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach.” In: *arXiv preprint arXiv:1907.11692* (2019).
- [92] Wei-Yin Loh. “Classification and regression trees.” In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2011).
- [93] Steve Lohr. “For impatient web users, an eye blink is just too long to wait.” In: *New York Times* (2012).
- [94] Matthew Luckie, Robert Beverly, William Brinkmeyer, et al. “Speedtrap: internet-scale IPv6 alias resolution.” In: *Proc. ACM Internet Measurement Conference (IMC)*. 2013.
- [95] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” In: *Proc. Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2017, pp. 4765–4774.
- [96] Scott M Lundberg et al. “Explainable AI for Trees: From Local Explanations to Global Understanding.” In: *arXiv preprint arXiv:1905.04610* (2019).

- [97] Alex Luu and Sophia A Malamud. “Non-topical coherence in social talk: A call for dialogue model enrichment.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 2020, pp. 118–133.
- [98] Yun Ma et al. “Measurement and Analysis of Mobile Web Cache Performance.” In: *Proc. The World Wide Web Conference (WWW)*. 2015.
- [99] Ratul Mahajan et al. “User-level internet path diagnosis.” In: *ACM SIGOPS Operating Systems Review* (2003).
- [100] *MaxMind*. [Online]. Available from: <https://www.maxmind.com/>.
- [101] Chandler May et al. “On Measuring Social Biases in Sentence Encoders.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019.
- [102] Robert B Miller. “Response time in man-computer conversational transactions.” In: *Proc. AFIPS Fall Joint Computer Conference*. ACM. 1968.
- [103] Ben Miroglio et al. “The Effect of Ad Blocking on User Engagement with the Web.” In: *Proc. The World Wide Web Conference (WWW)*. 2018.
- [104] Jeffrey C Mogul and Steven E Deering. *RFC 1191, Path MTU discovery*. 1990.
- [105] Sophon Mongkolluksamee, Kensuke Fukuda, and Panita Pongpaibool. “Counting NATted hosts by observing TCP/IP field behaviors.” In: *Proc. IEEE ICC*. 2012.
- [106] Yashar Moshfeghi and Joemon M. Jose. “On Cognition, Emotion, and Interaction Aspects of Search Tasks with Different Search Intentions.” In: *Proc. The World Wide Web Conference (WWW)*. 2013.
- [107] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. “Hate speech detection and racial bias mitigation in social media based on BERT model.” In: *PLOS ONE journal*. Vol. 15. 8. Public Library of Science, Aug. 2020, pp. 1–26.
- [108] Hamdy Mubarak et al. *Arabic Offensive Language on Twitter: Analysis and Experiments*. 2020. arXiv: [2004.02192](https://arxiv.org/abs/2004.02192) [CS . CL].
- [109] Fiona Fui-Hoon Nah. “A study on tolerable waiting time: how long are web users willing to wait?” In: *Behaviour & Information Technology* 23.3 (2004), pp. 153–163.
- [110] Vidhya Navalpakkam et al. “Measurement and Modeling of Eye-mouse Behavior in the Presence of Nonlinear Page Layouts.” In: *Proc. The World Wide Web Conference (WWW)*. 2013.
- [111] Javad Nejati and Aruna Balasubramanian. “An In-depth Study of Mobile Browser Performance.” In: *Proc. The World Wide Web Conference (WWW)*. 2016.

- [112] Ravi Netravali et al. “Mahimahi: a lightweight toolkit for reproducible web measurement.” In: *ACM SIGCOMM Computer Communication Review*. Vol. 44. 4. ACM. 2014, pp. 129–130.
- [113] Ravi Netravali et al. “Polaris: Faster Page Loads Using Fine-grained Dependency Tracking.” In: *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*. 2016.
- [114] Jakob Nielsen. *Response Times: The 3 Important Limits*. [Online]. Available from: <https://www.nngroup.com/articles/response-times-3-important-limits/>.
- [115] Ashkan Nikraves et al. “Mobilyzer: An Open Platform for Controllable Mobile Network Measurements.” In: *Proc. ACM MobiSys*. 2015.
- [116] “NLTK: The Natural Language Toolkit.” In: Association for Computational Linguistics, 2002.
- [117] Paul Pearce et al. “Augur: Internet-Wide Detection of Connectivity Disruptions.” In: *IEEE Symposium on Security and Privacy (SP)*. 2017.
- [118] Slav Petrov, Dipanjan Das, and Ryan McDonald. “A universal part-of-speech tagset.” In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. European Language Resources Association (ELRA), 2012.
- [119] Ingmar Poesse et al. “IP Geolocation Databases: Unreliable?” In: *ACM SIGCOMM Computer Communication Review* 41.2 (2011).
- [120] J. Postel. *RFC 791, Internet protocol*. 1981.
- [121] *Prometheus*. [Online]. Available from: <https://prometheus.io>.
- [122] Geoffrey K Pullum. “African American Vernacular English is not standard English with mistakes.” In: *The workings of language: From prescriptions to perspectives* (1999), pp. 59–66.
- [123] U.S. Census Bureau QuickFacts. *United States Census, QuickFacts statistics on U.S. Population Origin*. 2019. URL: <https://www.census.gov/quickfacts/fact/table/US/PSTo45219>.
- [124] Alec Radford et al. “Language models are unsupervised multitask learners.” In: *OpenAI blog* 1.8 (2019), p. 9.
- [125] Mohammad Rajiullah et al. “Web Experience in Mobile Networks: Lessons from Two Million Page Visits.” In: *Proc. The World Wide Web Conference (WWW)*. 2019.
- [126] Peter Reichl et al. “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.

- [127] Sanae Rosen et al. “Push or Request: An Investigation of HTTP/2 Server Push for Improving Mobile Performance.” In: *Proc. The World Wide Web Conference (WWW)*. 2017.
- [128] Vaspoul Ruamviboonsuk et al. “VROOM: Accelerating the Mobile Web with Server-Aided Dependency Resolution.” In: *Proc. of the Conference of the ACM Special Interest Group on Data Communication*. ACM. 2017, pp. 390–403.
- [129] *RUMSpeedindex*. [Online]. Available from: <https://github.com/WPO-Foundation/RUM-SpeedIndex>.
- [130] Jan Ruth et al. “Perceiving QUIC: Do Users Notice or Even Care?” In: *Proc. ACM Internet Measurement Conference (IMC)*. 2019.
- [131] Flavia Salutari, Danilo Cicalese, and Dario Rossi. *Datasets*. [Online]. Available from: <https://perso.telecom-paristech.fr/drossi/dataset/IP-ID/>.
- [132] Flavia Salutari et al. “A Large-Scale Study of Wikipedia Users’ Quality of Experience.” In: *Proc. The World Wide Web Conference (WWW)*. 2019, 3194–3200.
- [133] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” In: *NeurIPS Energy Efficient Machine Learning and Cognitive Computing Workshop*. 2019.
- [134] 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.
- [135] Raimund Schatz et al. “From Packets to People: Quality of Experience as a New Measurement Challenge.” In: *Data Traffic Monitoring and Analysis*. Jan. 2013, pp. 219–263.
- [136] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. “Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- [137] Emily Sheng et al. “The Woman Worked as a Babysitter: On Biases in Language Generation.” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.

- [138] *Shopzilla: faster page load time = 12 percent revenue increase*. [Online]. Available from: <http://www.strangeloopnetworks.com/resources/infographics/web-performance-andecommerce/shopzilla-faster-pages-12-revenue-increase/>. 2016.
- [139] Philipp Singer et al. “Why We Read Wikipedia.” In: *Proc. The World Wide Web Conference (WWW)*. 2017.
- [140] Chris Smith-Clarke and Licia Capra. “Beyond the Baseline: Establishing the Value in Mobile Phone Based Poverty Estimates.” In: *Proc. The World Wide Web Conference (WWW)*. 2016.
- [141] Neil Spring et al. “Measuring ISP topologies with Rocketfuel.” In: *IEEE/ACM TON* (2004).
- [142] Dominik Strohmeier et al. “Web Browsing.” In: *Quality of Experience: Advanced Concepts, Applications and Methods* (2014). Ed. by Sebastian Möller and Alexander Raake, pp. 329–338.
- [143] Srikanth Sundaresan et al. “Broadband Internet Performance: A View from the Gateway.” In: *Proc. ACM SIGCOMM*. 2011.
- [144] Yi Chern Tan and L. Elisa Celis. “Assessing Social and Intersectional Biases in Contextualized Word Representations.” In: *Proc. Neural Information Processing Systems (NeurIPS)*. 2019, pp. 13209–13220.
- [145] G. Tangari et al. “Tackling Mobile Traffic Critical Path Analysis With Passive and Active Measurements.” In: *Proc. Traffic Monitoring and Analysis Workshop (TMA)*. 2019, pp. 105–112.
- [146] J. Touch. *RFC 6864, Updated Specification of the IPv4 ID Field*. 2013.
- [147] Martino Trevisan, Idilio Drago, and Marco Mellia. “PAIN: A Passive Web performance indicator for ISPs.” In: *Computer Networks* 149 (2019), pp. 115 –126. ISSN: 1389-1286.
- [148] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. *Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance*. 2020. arXiv: 2005.00315 [CS . CL].
- [149] Matteo Varvello et al. “EYEORG: A Platform For Crowdsourcing Web Quality Of Experience Measurements.” In: *Proc. ACM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*. 2016.
- [150] Xiao Sophia Wang, Arvind Krishnamurthy, and David Wetherall. “Speeding up Web Page Loads with Shandian.” In: *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*. 2016, pp. 109–122.

- [151] Xiao Sophia Wang et al. "How Speedy is SPDY?" In: *Proc. of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*. 2014.
- [152] Bernard L Welch. "The generalization of student's' problem when several different population variances are involved." In: *Biometrika* 34.1/2 (1947), pp. 28–35.
- [153] Mark A West and Stephen McCann. *RFC 4413, TCP/IP field behavior*. 2006.
- [154] Rebecca Wheeler and Julia Thomas. "And "Still" the Children Suffer: The Dilemma of Standard English, Social Justice, and Social Access." In: *JAC* (2013), pp. 363–396.
- [155] *World Bank*. [Online]. Available from: <https://data.worldbank.org/>.
- [156] *World Wide Wait*. [Online]. Available from: <https://www.economist.com/science-and-technology/2010/02/12/world-wide-wait>. 2010.
- [157] Canwen Xu et al. "BERT-of-Theseus: Compressing BERT by Progressive Module Replacing." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 7859–7869.
- [158] Brit Youngmann and Elad Yom-Tov. "Anxiety and Information Seeking: Evidence From Large-Scale Mouse Tracking." In: *Proc. The World Wide Web Conference (WWW)*. 2018.
- [159] Sebastian Zander, Lachlan LH Andrew, and Grenville Armitage. "Capturing ghosts: Predicting the used IPv4 space by inferring unobserved addresses." In: *Proc. ACM Internet Measurement Conference (IMC)*. 2014.
- [160] Haoran Zhang et al. "Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings." In: Association for Computing Machinery, 2020.
- [161] Xu Zhang et al. "EzE: 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.
- [162] Torsten Zimmermann, Benedikt Wolters, and Oliver Hohlfeld. "A QoE Perspective on HTTP/2 Server Push." In: *Proc. ACM SIGCOMM, Internet-QoE Workshop*. 2017.
- [163] Torsten Zimmermann et al. "How HTTP/2 Pushes the Web: An Empirical Study of HTTP/2 Server Push." In: *Proc. IFIP Networking*. 2017.

Titre : Mesures d'Internet à large echelle, longitudinale et sans biais

Mots clés : Mesures d'Internet, Qualité d'expérience, Apprentissage

Résumé : Aujourd'hui, un monde sans Internet est inimaginable. En interconnectant des milliards de personnes dans le monde et en offrant un nombre incalculable de services, il est désormais pleinement intégré à la société moderne. Pourtant, malgré l'évolution et le développement de la technologie, son omniprésence et son hétérogénéité soulèvent encore de nouveaux défis, tels que les problèmes de sécurité, le contrôle de la qualité d'expérience des utilisateurs (QoE), le souci de transparence et celui d'équité. En conséquence, l'objectif de cette thèse est d'apporter un nouvel éclairage sur certains des défis qui ont émergé ces dernières années. En particulier, nous fournissons une analyse approfondie de certains des aspects les plus importants de l'Internet moderne. Un accent particulier est mis sur le World Wide Web, qui, parmi tous, est sans doute l'une des applications Internet les plus populaires, et un regard spécifique sur son interaction avec l'apprentissage automatique. La première partie de ce travail étudie la qualité de l'expérience de navigation des utilisateurs sur le Web, avec des mesures effectuées à la fois "*in the wild*" et dans des environnements contrôlés. Nos contributions continuent avec une analyse originale de l'avis *subjectif* des utilisateurs et des mesures *objectives* de la qualité d'expérience, montrant la difficulté de construire des modèles supervisés

précis, basés sur des données, capables de prédire la satisfaction des utilisateurs, ainsi qu'une discussion approfondie de la nature multimodale des avis *subjectifs* des utilisateurs. Dans la deuxième partie de ce travail, nous analysons et discutons l'équité des modèles de langage basés sur des transformateurs de pointe, qui sont pré-entraînés sur des corpus basés sur le Web et qui sont généralement utilisés pour résoudre une grande variété de tâches de traitement du langage naturel (NLP). Nous nous demandons ici si la taille et l'hétérogénéité du Web garantissent la diversité des modèles. Le cœur de nos contributions repose sur la mesure du biais intégré dans les modèles, que nous discutons sous différents angles. Enfin, la dernière partie de cette thèse traite de la classification d'objets générés par des machines à l'aide de certains des plus simples algorithmes d'apprentissage automatique supervisés à l'état de l'art. Grâce à un framework solide mais peu intrusif, nous montrons que les différents comportements d'un champ du paquet IP, l'identification IP (IP-ID), peuvent être facilement classifiés avec peu de caractéristiques ayant un haut pouvoir discriminatoire. Nous appliquons enfin notre technique à un census à l'échelle de l'Internet et fournissons une vue actualisée de l'adoption de ses différentes implémentations dans l'Internet.

Title : Longitudinal, large-scale and unbiased Internet measurements

Keywords : Internet Measurements, Quality of Experience, Machine learning

Abstract : Today, a world without the Internet is unimaginable. By interconnecting billions of people worldwide and by offering an uncountable number of services, it is now fully embedded in the modern society. Yet, despite technology evolution and development, its pervasiveness and heterogeneity still raise new challenges, such as security concerns, monitoring of the users' Quality of Experience (QoE), care for transparency and fairness. Accordingly, the goal of this thesis is to shed new light on some of the challenges emerged in recent years. In particular, we provide an in-depth analysis of some of the most prominent aspects of modern Internet. A particular emphasis is given on the World Wide Web, which among all, is undoubtedly one of the most popular Internet applications, and a specific regard to its interaction with machine learning.

The first part of this work studies the Quality of Experience of users' browsing the Web, with measurements led both in the wild and in controlled environments. Our contributions follow with an original analysis of both the *subjective* user feedback and the *objective* QoE metrics, showing how hard it is to build accurate supervised data-driven models capable to predict the user satisfaction, along with an in-depth

discussion of the multi-modal nature of the *subjective* user opinions.

In the second part of this work, we analyze and discuss the fairness of state-of-the-art transformer-based language models, which are pre-trained on Web-based corpora and which are typically used to solve a wide variety of Natural Language Processing (NLP) tasks. Here, we question whether the sheer size and heterogeneity of the Web guarantee diversity in the models. The core of our contributions rests in the measure of the bias embedded in the models, that we discuss under different angles.

Finally, the last part of this dissertation addresses the classification of objects generated by machines through some of the simplest state-of-the-art supervised machine learning algorithms. Through a minimally intrusive, robust and lightweight framework, we show that the different behaviors of a field of the IP packet, the IP identification (IP-ID), could be easily classified with few features having high discriminative power. We finally apply our technique to an Internet-wide census and provide an updated view of the adoption of the different implementations in the Internet.