Predicting Quality of Experience from RSRP and Throughput

Vivek Adarsh, Alex Ermakov and David Weinflash
Department of Computer Science, UC Santa Barbara
{vivek, aermakov, dweinflash}@cs.ucsb.edu

ABSTRACT

With over 80% of global IP traffic attributed to video [5], measuring quality of experience (QoE) across streaming services becomes imperative for service providers. However, such evaluations are often difficult on mobile broadband, particularly due to the inherent wireless characteristics associated with cellular networks. Investigating quality of experience usually involves implementing complex systems to capture usability at the end-user device level. This is both resource and time intensive. In this study, we provide the first look into QoE prediction on mobile broadband using commonly reported metrics such as signal strength and throughput measurements. To validate our idea we collect a comprehensive dataset that contains ground truth measurements from over sixteen locations across the continental United States. Using several state-of-the-art machine learning algorithms we show that it is possible to predict quality of experience with an accuracy of 87%, while minimizing computing resources.

CCS CONCEPTS

• **Networks → Network performance analysis**; *Network measurement*.

KEYWORDS

LTE, Quality of Experience, Network Modelling

ACM Reference Format:
Vivek Adarsh, Alex Ermakov and David Weinflash. 2020. Predicting Quality of Experience from RSRP and Throughput. In *Proceedings of UCSB (ML For Networked Systems).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/xxxxxx.xxxxxx

ML For Networked Systems, CS 293N, Winter '20

ACM ISBN 978-1-4503-6948-0/19/10…$15.00
https://doi.org/10.1145/xxxxxx.xxxxxx

1 INTRODUCTION

Quality of Experience (QoE) has become an increasingly important measurement as service providers aim to maximize each individual customers' experiences in order to grow their business. As a result, identifying the key factors that influence user experience has become an increasingly important task. In this study, we focus on exploring the link between key radio measurements and user experience in the context of video streaming sessions. In order to quantify network performance, we introduce Edain, a comprehensive monitoring suite to accurately ascertain network level and application level measurements. We deploy Edain in a diverse array of environments to collect ground truth network measurements from sixteen different locations across the United States. Finally, in order to accurately predict user experience from radio and network level measurements, we utilize a comprehensive array of machine learning algorithms ranging from Random Forest and ARIMA to popular neural network models such as recurrent neural networks in order to predict user experience on video streaming services. The novelty in our approach can be attributed to the easily-accessible metrics used (RSRP, auto reported on the device and throughput, that can be gathered through popular speed checker applications) for prediction of quality of experience (QoE).

2 BACKGROUND

Quality of Experience (QoE) is an increasingly important metric for today's modern service providers. Discovering and optimizing the factors that determine user experience has thus become an essential goal for network operators. As a result, we focus on exploring the link between key radio measurements and user experience in the context of video streaming sessions.

2.1 LTE Coverage and Usability

RSRP: Reference signal received power (RSRP), is defined as the linear average over the power contributions (in Watts) of the resource elements that carry cell-specific reference signals within the considered measurement frequency bandwidth. Theoretically, it varies between -44dBm to -140dBm, with -44dBm indicating the best quality signal strength. Estimation of received signal strength plays a vital role in many control plane operations, including inter- and intra- eNodeB handovers [4, 7, 14, 15, 17]. Precise detection of RSRP plays a crucial role in these handovers, as well as several diagnostic methods in LTE networks. For instance, Anas et al. [3] evaluate the performance of RSRP handovers in LTE. They observe that a handover margin of 2dB to 6dB (RSRP) leads to an optimal number of handovers without sacrificing much of uplink SINR (for a specific range of user velocity). The effect of RSRP measurement bandwidth on the accuracy of handovers is studied in [13, 14]. From a telecom provider's perspective, this suggests a need for up-to-date, accurate RSRP space-maps for improving service quality.

Several prior works examine the relationships between RSRP and SINR [1, 11, 12, 21], but little work explores the correlation between RSRP and usability at the user end. While the FCC publishes LTE coverage maps for the United States using RSRP measurements reported by the telecoms [8], targeted studies about the quality of experience as a function of radio measurements remain unexplored.

2.2 Quality of Experience

Quality of Experience (QoE) has received much attention over the past years and has become a prominent issue for delivering services and applications. A significant amount of research has been devoted to understanding, measuring, and modelling QoE for a variety of media services. Moreover, with many different offered to the emerging consumer base, identifying the root causes of QoE impairments and finding effective solutions for meeting the end users' requirements and expectations in terms of service quality is a challenging and complex problem. Thus, we turn our focus to on-demand video streaming in order to examine the change in user experience under varying network conditions.

In this study, we investigate the effect of radio measurements such as RSRP and QoS metrics such as throughput on quality of experience metrics, in particular buffer size during video streaming sessions.

3 EDAIN: NETWORK MONITORING SUITE

We develop Edain, a comprehensive network monitoring suite, to quantify network performance. Edain provides an extensive set of features to measure QoS and QoE metrics at the client. At the time of writing, Edain has been used in over sixteen locations across the United States to compare the performance of mobile broadband under varying network conditions.

3.1 Implementation

Edain's functionality ranges from computing network level (throughput, latency and packet traces) to application level (on-demand video streaming (YouTube)) and page load time measurements. We measure cellular performance by tethering the phones with the laptops running Edain. We ensure that the cellular plans on all our devices have unlimited data and hot-spot enabled to effectively achieve the same level of performance as we would on the mobile device. Edain was developed for Linux, keeping ease of deployment in mind. Edain is agnostic to network type and provides the flexibility to deploy it on either wired, Wi-Fi or cellular environments. Development of an integrated smartphone app was impractical as the level of unification achieved for various application-level measurements (YouTube, Skype, etc.) was simply not possible on smartphone operating systems, given the walled access to iOS ecosystem and recent restrictions introduced in Android APIs [2, 22].

Latency: Edain's `rtt_out` function automates the collection of round-trip times by initiating pings through Hping3 [26] to a server hosted on an AWS instance (Virginia). We configure Hping3 to use TCP packets instead of ICMP. The ping duration is capped at 120 seconds with one-second intervals between each ping. Edain computes the average latency using two different sessions - one before the throughput tests (described below) and one after. This captures the vagaries introduced during a long throughput measurement session that may or may not prompt a cellular provider to tune the level of service provided soon-after [16]. We observe an average round-trip time of 61ms with a standard deviation of ±3 ms.

Throughput: To calculate the achieved throughput Edain initiates iPerf threads to download a specified file from the same AWS instance as the latency test. The measurement is repeated 10 times and results are saved at the client side. Further, Edain logs the packet traces at the client throughout the iPerf tests in order to compute second-order metrics such as packet loss.

Page Load Time: Load times are initiated through the `plt_stream` function. Edain automates the loading of Web pages using Selenium [24]. For our measurements, we use the Tranco Top 25 list [23]. To evaluate load times. Edain logs the navigation timings of a Web page starting from *navigationStart* through the *loadEventEnd* event [27]. These instances of event timings help in a finer grain analysis of page load times. We set Edain to run `plt_stream` three times for better estimation of load times. Browser cache is automatically wiped out after each Web page load to reflect true load time for the next iteration.

Video Streaming: YouTube: Examination of QoE metrics from on-demand video streaming services is a challenging problem, particularly because of encrypted traffic, as demonstrated by prior work [10, 19, 20, 25]. Since three-quarters of global IP traffic is dominated by video, user experience for streaming services then becomes critical on mobile broadband. We built the *video_stream* function into Edain that enables logging QoE metrics from YouTube videos. Our design is inspired from eMIMIC testbed [19]. It uses passive network measurements to estimate key video QoE metrics for encrypted HTTP-based Adaptive Streaming (HAS) sessions. Using packet headers from network traffic to model a HAS session, video QoE metrics are estimated, such as start-up latency, achieved resolution and buffer size. The function logs samples at one-second resolution. To ensure uniformity across all our datasets, we play a 180-second video in loop for three runs, for every location and cellular operator. Video resolution preference is set to auto so that the client takes care of any required resolution switches, which is indicative of the network conditions.

4 DATASET

To test our proposed study, we perform a targeted measurement campaign to collect network performance measurements from sixteen different locations across the United States. These include eight datasets from the state of New Mexico, six from San Diego, CA and two from the city of San Francisco, gathering measurements from four major telecom operators in the US: AT&T, Sprint, T-Mobile and Verizon. We cover a cumulative distance of over 200 miles in New Mexico over a period of five days beginning May 28, 2019. Table 1 shows the locations of ground measurements and their descriptive labels we use for this analysis. On our campaign, we cover two American Indian reservations near Santa Fe: Santa Clara and Ohkay Owingeh Pueblos. In both the Pueblos, tribal leadership permitted us to collect additional measurements in residential zones. We select these areas of New Mexico for their mix of tribal and non-tribal demographics; tribal lands tend to have the highest coverage over-statements and the most limited cellular availability within the United States [6, 9]. Additional locations in New Mexico include a weave of non-tribal rural, semi-urban and micropolitan regions. Finally, we collect datasets from San Diego and San Francisco to capture network performance in a more urban, metropolitan setting. We also gather cellular performance under varying network conditions such as during heavy congestion and times of under-utilization (when no network overload is likely to occur). These ground measurements provide an important comparison point for actual coverage and user experience particularly since, combined, they are representative of a wide spectrum of network availability and usability.

In our measurement campaign, we record signal strength readings from four Motorola G7 Power (XT1955-5) phones, each running Android Pie (9.0.0). We collect measurements using the Network Monitor application [18]. An external GlobalSat BU-353-S4 GPS connected to an Ubuntu Lenovo ThinkPad laptop gathered geolocation measurements, which we matched to the appropriate ground measurement by timestamp. We outfitted each phone with a SIM card from one of the four top cellular providers in the region: Verizon, T-Mobile, AT&T, and Sprint. The phones recorded signal strength every second through the areas of study.

Table 1: Summary of dataset locations

Location	County	State	Cluster	Remarks
SC_A01	Rio Arriba	New Mexico	Tribal, Rural	–
SC_B01	Rio Arriba	New Mexico	Tribal, Rural	–
SC_B02	Rio Arriba	New Mexico	Tribal, Rural	–
SC_C01	Rio Arriba	New Mexico	Semi-Urban	–
SC_C02	Rio Arriba	New Mexico	Tribal, Rural	–
OO_D01	Rio Arriba	New Mexico	Tribal, Rural	–
AR_D02	Rio Arriba	New Mexico	Non-Tribal, Rural	–
SF_D03	Santa Fe	New Mexico	Non-Tribal, Rural	–
ADM	San Diego	California	Urban	Congested
ADM Base	San Diego	California	Urban	–
CWF	San Diego	California	Urban	Congested
CWF Base	San Diego	California	Urban	–
FMR	San Diego	California	Urban	Congested
FMR Base	San Diego	California	Urban	–
AIS	San Francisco	California	Urban	Congested
AIS Base	San Francisco	California	Urban	–

5 EVALUATION

5.1 Analyzing the dataset

To collect cellular performance data, we instantiate two devices operating on same provider. We run YouTube and RSRP capture tests on one of the device while concurrently gathering throughput data on the other. YouTube tests compiles QoE metrics such as achieved video quality, playback mode (playing or buffering) and buffer size at one-second granularity. Throughput tests run on a separate device in order to mitigate any interference that might arise from active monitoring while running YouTube tests on the same device. RSRP tests do not introduce any interference while running Edain. For throughput tests, we fill up the pipeline by initiating iPerf test for a file size of 500 MB, which is usually takes longer than Edain to complete YouTube measurements. Our curating and pre-processing, our dataset consists of over 24,000 individual datapoints.

In this study, we aim to to predict buffer size using RSRP and throughput measurements for the LTE network provider. To do so, we first explore the correlation between the RSRP and buffer level. To do so, we analyse and plot the cross correlation of between these two distributions. Cross correlation is a measure of similarity of two series as a function of the displacement of one relative to the other. In a time series dataset such ours, the displacement is in time lags (one second is one time lag). This is also known as a sliding dot product or sliding inner-product. It is commonly used for

Table 2: Accuracy and runtime of the models used to perform buffer size prediction.

Model	Accuracy	Precision	Recall	MAE	MSE	RMSE	Training Time	Prediction Time
Random Forest (Quantile Binning)	49%	53%	49%	–	–	–	45.565s	1.920s
Random Forest (Uniform Binning)	**87%**	**88%**	**87%**	–	–	–	44.44s	2.05s
ARIMA	–	–	–	**2.45**	43.44	6.59	–	–
LSTM (Quantile Binning)	78%	78%	78%	–	–	–	92.233s	0.170s
LSTM (Uniform Binning)	**80%**	**80%**	**80%**	–	–	–	76.12s	0.17s
ADA Boost	35%	21%	35%	–	–	–	1.587s	0.135s
ADA Boost Regression	–	–	–	3.89	25.46	5.04	0.21s	0.001s
Extra Randomized Tree	76%	76%	76%	–	–	–	0.008s	0.005s
Extra Randomized Tree Regression	–	–	–	2.64	30.92	5.56	0.022s	0.0006s
Bagging	**87%**	**87%**	**87%**	–	–	–	0.603s	0.006s
Bagging Regression	–	–	–	2.38	18.03	4.24	0.618s	0.006s
Boosting	**88%**	**88%**	**88%**	–	–	–	29.14s	0.025s
Boosting Regression	–	–	–	2.22	14.81	3.85	2.001s	0.002s
Naive Bayes	31%	18%	31%	–	–	–	0.003s	0.002s
KNN	15%	13%	15%	–	–	–	0.013s	0.094s
KNN Regression	–	–	–	22.37	718.47	26.8	0.012s	0.271s
SVM	17%	5%	17%	–	–	–	10.09s	1.383s
Decision Trees	80%	80%	80%	–	–	–	0.104s	0.0004s
Decision Trees Regression	–	–	–	2.59	29.99	5.47	0.102s	0.0006s

searching a long signal for a shorter, known feature. Figure 1 reveals that the highest correlation between RSRP and buffer level is within ±5 time lags. In addition, we observe that the system is not a perfectly dynamic system, as illustrated by asymmetric peaks on either side of lag 0. Hence, for further exposition of non-linear dependencies between RSRP and buffer level, we deploy machine learning algorithms to compute the best-fit model.

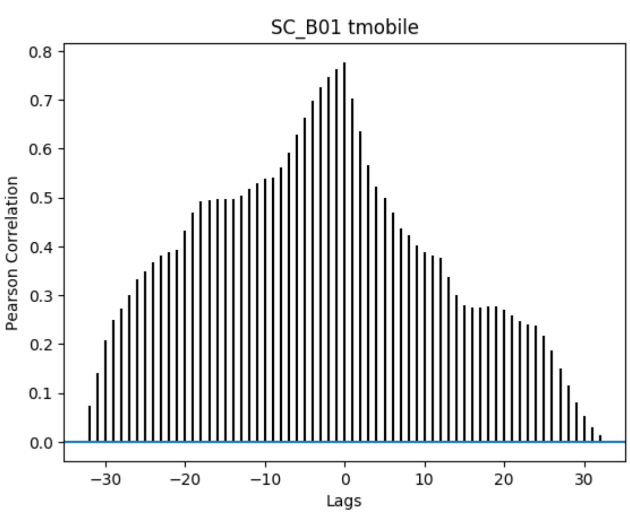

Figure 1: Cross correlation between RSRP and buffer level.

In our analysis, we deploy a comprehensive array of machine learning algorithms, commonly referred to as an evaluation sweep. We examine simpler models such as Random Forest, ARIMA, etc. as well popular neural network model like recurrent neural network (RNN). Table 2 presents the results we obtain after applying each of the machine learning models on the curated dataset. We achieve significant accuracy using simpler models as opposed to complex neural networks (for instance, RNN). Random Forest was able to produce an accuracy of 87% as opposed to 80% in RNNs, with much faster training times.

5.2 Random Forest

Non-parametric methods were surprisingly effective in predicting Quality of Experience from radio and network level measurements. Specifically, our random forest model was especially effective in segmenting network level measurements using both a quantile binning and uniform binning approach. By making no assumptions about the functional relationship between network measurements and Quality of Experience, our random forest model is able to make informed and unbiased predictions. What is more, by aggregating a large group of decision trees - where each tree uses a different subset of features - our random forest model effectively prevents overly powerful features from influencing the model, thus de-correlating the ensemble of trees and allowing other features to maintain predictive power. For our model, we use uniform binning of 12 timestamps consisting of 1000 estimators (decision trees). RF with uniform binning produces the best accuracy, precision and recall of 87%, 88% and 87% respectively.

5.3 Autoregressive Integrated Moving Average: ARIMA

Similar to recurrent neural networks, ARIMA integrates temporal structures and previous observations in sequential data sets to forecast time series data. Unlike recurrent neural networks, however, ARIMA provides a simple, off-the-shelf solution with an intuitive design to quickly make predictions. By constructing a linear regression model using lagged observations, ARIMA provides a statistical analysis of network and radio level measurements and outputs predictions on Quality of Experience in a minimal amount of time. With a focus on short-term dependencies, ARIMA also offers a unique perspective on our time-series data, as our recurrent neural network LSTM model primarily focuses on long-term dependencies. Given its usability and seamless integration, ARIMA provides the ideal benchmark for our more complex models and allows efficient analysis of our time series data. Our model employs a 5-second time window that results in a peak mean absolute error (MAE) of just over 2 seconds. State otherwise, our model can predict the buffer size with an average error margin of about 2 seconds.

5.4 Recurrent Neural Networks: LSTM

Given their innate ability to recognize temporal patterns in data, recurrent neural networks are especially well suited in our task of predicting Quality of Experience from observed network and radio measurements. Indeed, such networks have been implemented successfully in other similar sequential data domains, where the internal memory components of recurrent neural networks have helped to advance the state of the art from natural language processing to time series prediction. Similar to other successful applications, our neural network was enhanced by following the design of a Long Short-Term Memory (LSTM) network. By utilizing the architecture of a LSTM, long-term dependencies in the data were successfully established, effectively avoiding the vanishing gradient problem and allowing our model to utilize previous patterns in the data to make future predictions. This implementation uses 3 layers of LSTM nodes followed by 2 dense layers. Each LSTM layer is succeeded by 20% dropouts that retain model accuracy and prevent overfitting. We notice a total of 54,284 trainable parameters. The dataset was split into training and test set in the ratio of 70:30. We trained the RNN model containing 208 neurons for 200 iterations. Not surprisingly, we obtain an appreciable accuracy, precision and recall of 80%. The only drawback of using an RNN is the relative increase in computing resources and longer training times, which we emphasize may not matter in our case given the size of dataset we operate with.

6 CONCLUSION

In this work, we take an in-depth look at predicting buffer level from radio and throughput measurements. We first curate our dataset to parse ground truth throughput value for each buffer-size datapoint. These parsed dataset is then synced across corresponding RSRP values. We perform a detailed analysis to better understand the dependency of buffer level in YouTube streaming with throughput and radio measurements. Finally, we apply various state-of-the-art machine learning models to predict future buffer level given adequate knowledge of current and previous values of RSRP and throughput. Results show that computationally inexpensive models such as Random Forest and ARIMA can be used to predict the level of remaining buffer with high accuracy, precision and recall. Our study incorporates clear comparison between aforementioned simpler models and recurrent neural networks (RNN). We observe that simpler models perform noticeably better than compute-intensive LSTM networks. Our framework can be implemented to provide information about user experience by easy-to-acquire metrics such as RSRP (auto reported by device) and throughput (commonly used speed checker applications).

REFERENCES

[1] Farhana Afroz, Ramprasad Subramanian, Roshanak Heidary, Kumbesan Sandrasegaran, and Solaiman Ahmed. 2015. SINR, RSRP, RSSI and RSRQ measurements in long term evolution networks. *International Journal of Wireless & Mobile Networks* (2015).

[2] Efthimios Alepis and Constantinos Patsakis. 2017. Trapped by the UI: The Android case. In *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 334–354.

[3] Mohmmad Anas, Francesco D Calabrese, Preben E Mogensen, Claudio Rosa, and Klaus I Pedersen. 2007. Performance evaluation of received signal strength based hard handover for UTRAN LTE. In *65th Vehicular Tech. Conf.* IEEE.

[4] Danish Aziz and Rolf Sigle. 2009. Improvement of LTE Handover Performance through Interference Coordination. In *69th Vehicular Tech. Conf.* IEEE.

[5] Cisco. 2020. Cisco Annual Internet Report (2018–2023). (2020). https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[6] Federal Communications Commission. 2019. Broadband Deployment Report. https://www.fcc.gov/reports-research/reports/broadband-progress-reports/2019-broadband-deployment-report. (May 2019).

[7] Konstantinos Dimou, Min Wang, Yu Yang, Muhammmad Kazmi, Anna Larmo, Jonas Pettersson, Walter Muller, and Ylva Timner. 2009. Handover within 3GPP LTE: Design Principles and Performance. In *70th Vehicular Tech. Conf.* IEEE.

[8] FCC. 2020. LTE Coverage by Number of Providers - YE 2018. (2020). https://www.fcc.gov/reports-research/maps/lte-coverage-number-providers-ye-2018/

[9] Federal Communications Commission. 2018. Broadband Deployment Report. https://www.fcc.gov/reports-research/reports/broadband-progress-reports/2018-broadband-deployment-report. (February 2018).

[10] Craig Gutterman, Katherine Guo, Sarthak Arora, Xiaoyang Wang, Les Wu, Ethan Katz-Bassett, and Gil Zussman. 2019. Requet: Real-time QoE Detection for Encrypted YouTube Traffic. In *Proceedings of the 10th ACM Multimedia Systems Conference*. 48–59.

[11] Christoph Ide, Bjoern Dusza, and Christian Wietfeld. 2013. Performance of channel-aware M2M communications based on LTE network measurements. In *24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE.

[12] Christoph Ide, Robert Falkenberg, Dennis Kaulbars, and Christian Wietfeld. 2016. Empirical analysis of the impact of LTE downlink channel indicators on the uplink connectivity. In *83rd Vehicular Tech. Conf.* IEEE.

[13] Janne Kurjenniemi and Tero Henttonen. 2008. Effect of measurement bandwidth to the accuracy of inter-frequency RSRP measurements in LTE. In *19th International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE.

[14] Janne Kurjenniemi, Tero Henttonen, and Jorma Kaikkonen. 2008. Suitability of RSRQ measurement for quality based inter-frequency handover in LTE. In *Intern. Symposium on Wireless Communication Sys.* IEEE.

[15] Peter Legg, Gao Hui, and Johan Johansson. 2010. A Simulation Study of LTE Intra-frequency Handover Performance. In *72nd Vehicular Tech. Conf.* IEEE.

[16] Yuanjie Li, Haotian Deng, Chunyi Peng, Zengwen Yuan, Guan-Hua Tu, Jiayao Li, and Songwu Lu. 2016. icellular: Device-Customized Cellular Network Access on Commodity Smartphones. In *13th USENIX Symposium on Networked Systems Design and Implementation NSDI 16)*. 643–656.

[17] Cheng-Chung Lin, Kumbesan Sandrasegaran, Huda Adibah Mohd Ramli, and Riyaj Basukala. 2011. Optimized Performance Evaluation of LTE Hard Handover Algorithm with Average RSRP Constraint. *CoRR* abs/1105.0234 (2011). http://arxiv.org/abs/1105.0234

[18] Benoit Lubek. [n. d.]. Network Monitor. ([n. d.]). https://github.com/caarmen/network-monitor

[19] Tarun Mangla, Emir Halepovic, Mostafa Ammar, and Ellen Zegura. 2018. emimic: Estimating HTTP-based Video QoEe Metrics from Encrypted Network Traffic. In *2018 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 1–8.

[20] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*. 197–210.

[21] Chester Sungchung Park and Sungkyung Park. 2016. Analysis of RSRP measurement accuracy. *Communications Letters* 20, 3 (2016), 430–433.

[22] Sai Teja Peddinti, Igor Bilogrevic, Nina Taft, Martin Pelikan, Ulfar Erlingsson, Pauline Anthonysamy, and Giles Hogben. 2019. Reducing Permission Requests in Mobile Apps. In *Proceedings of ACM Internet Measurement Conference (IMC 19)*.

[23] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2018. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156* (2018).

[24] Sagar Shivaji Salunke. 2014. *Selenium Webdriver in Python: Learn with Examples*. CreateSpace Independent Publishing Platform.

[25] Roei Schuster, Vitaly Shmatikov, and Eran Tromer. 2017. Beauty and the Burst: Remote Identification of Encrypted Video Streams. In *26th USENIX Security Symposium USENIX Security 17)*. 1357–1374.

[26] Kali Tools. 2019. Hping3 Package. (2019).

[27] W3C. 2019. Navigation Timing. (2019). https://www.w3.org/TR/navigation-timing/#processing-model