A Machine Learning Application for Latency Prediction in Operational 4G Networks

Ali Safari Khatouni[1,2], Francesca Soro[2], Danilo Giordano [2]
[1]Dalhousie University, Canada ali.safari@dal.ca
[2]Politecnico di Torino, Italy firstname.lastname@polito.it

***Abstract*—Measuring performance on Internet is always challenging. When it comes to the mobile networks, the variety of technology characteristics coupled with the opaque network configuration make the performance evaluation even a more difficult task. Latency is one of the aspects having the largest impact on the performance and on the end users' Quality of Experience. In this paper, we present a machine learning approach that, exploiting real mobile network data on the end user, try to predict the latency in a real operational network. We consider a large-scale dataset with more than 238 million latency measurements coming from 3 different commercial mobile operators. The presented methodology flattens the RTT values into several bins, turning the latency prediction problem to a multi-label classification problem. Then, three well-known supervised algorithms are exploited to predict the latency. The obtained results highlight the importance of representative dataset from operational network. It calls for further improvements on the algorithm selection, tuning, and their predictive capabilities.**

I. Introduction

We are witnessing two major changes in the Internet. On the one hand, smartphones and mobile broadband (MBB) networks have revolutionized the way people interact, offering access to web, video, messaging applications in mobility with a capacity similar to wired networks. On the other hand, this dependency on the MBB network poses severe challenges for mobile network operators. In this scenario, monitoring the network to analyze the intertwining of technologies, protocols, setups and website design is crucial. Hence, designing scientifically sound measurement campaigns is very important. Not surprisingly the research community has put a lot of effort into measuring the benefits of new technologies [1], [2], [3]. Previous works mainly focused on specific measurement setups, and often on specific angles of this complex ecosystem [4], [5], [6], [7]. In this paper, we use an open dataset from the MONROE platform[1] to study and predict the network latency, i.e., the Round Trip Time (RTT) in a real operational 4G network.

In MONROE platform, RTT is measured continuously (every 1 second) by running ICMP Ping[2] toward dedicated servers in several locations from four European countries. Several physical information such as the frequency, the radio station, signal strength, etc. are captured by nodes and stored in open access database along with the RTT experiments. Measuring and predicting latency is an important task, specifically for MBB networks, authors in [8] and [9] show that latency plays a key role when it comes to the users' web browsing choices (e.g., abandoning a page before it is fully loaded). Such behaviour is known to have a significant impact on economic revenues, but up to now, companies can only attain to few rules [10] to try and reduce the latency impact. From the network analyst perspective, a more detailed and technical set of metrics are needed to define users' perceived QoE [11]. Furthermore, authors in [4] highlight the impact of low-level measurable radio network characteristics on the user QoE during web browsing. An overview of various end-to-end delay prediction methods is offered in [12]; they illustrate different methodologies to estimate latency, e.g., *Queueing Network Modelling*, *System Identification*, *Time Series Approach*, and *Neural Networks*. Authors [13] focus on TCP performance in a Mobile Ad Hoc network and present a machine learning technique called *Experts Framework*.

Our work differs from the others as we use a large dataset from a real operational network, we leverage different machine learning solutions to model the latency of 4G networks and predict the RTT. Firstly, we present our dataset to characterize different features, showing the RTT temporal evolution, and their distributions. Secondly, we use the Random Forest algorithm to select the most important features in RTT prediction problem among all available features. Thirdly, we leverage three different classifiers to predict the latency from the device toward a dedicated server by using selected features.

To validate our model, we used a real dataset composed by more than 200M RTT measurements, collected in 45 different locations connected with three different mobile operators. Our results show that the use of feature selection techniques can significantly reduce the amount of data (more than 60%) by minimizing loss of information to predict RTT. We also present basic algorithm cannot obtain high accuracy in RTT prediction.

Our work poses itself as preliminary work to emphasize the challenges and the importance of latency prediction in operational mobile networks. As a result, we are using this work as a starting point to design a tailored-algorithm suitable for MBB networks, in order to avoid simplistic and superficial solutions.

II. Measurements design

In the following, we describe the experiment design and the considered dataset.

[1] https://www.monroe-project.eu/access-monroe-platform/

[2] https://github.com/MONROE-PROJECT/Experiments/tree/master/experiments/ping

978-3-903176-15-7 © 2019 IFIP

TABLE I: Statistics on the dataset collected in Italy.

# Nodes	# Operators	# RTT Measurements	# Metadata
45	3	238 M	37 M

A. *Measurement Infrastructure*

We rely on the MONROE [14] platform, the first European open access platform for multi-homed [15] and large-scale mobile measurements on commercial mobile providers. The MONROE platform covers 4 countries in Europe (Italy, Norway, Spain, and Sweden) with 100 nodes equipped with Ethernet, WIFI and 3G/4G interfaces with commercial subscriptions.

In more details, nodes integrate single-board computers with three 3G/4G mobile broadband modems using regular commercial subscriptions. Nodes operate both under mobility and in stationary scenarios. This allows to perform passive measurements on the network of three mobile operators in each country, guaranteeing the repeatability of the experiments under the same conditions (e.g., same hardware, software, configuration, and geographical location).

MONROE allows us to access the information about the network, time and location for experiments, as well as metadata from the mobile modems, including, for example, cell ID, signal strength, and link technology for each network provider [3]. The platform is also instrumented to regularly run baseline experiments (e.g., HTTP download, Ping, passive network traffic measurements, etc.). Experiment results are stored in the project database and publicly available for researchers. In the following, we provide details about the experimental setup and describe the metrics we considered in this work. MONROE runs regular ICMP Ping (84 Byte payload) experiments to measure Round Trip Time (RTT) from nodes toward a dedicated server. After each measurement, the system logs all collected data to a central repository.

The platform enables us to measure multiple operators at the same location and using the identical device and software, thus, limiting potential sources of bias from our experiment. However, this still leaves us space to understand the impact of the different operator configurations, the radio coverage, and all the time-varying parameters that may impact the latency (See. Fig. 1).

For this, we explore a series of metadata that characterize the specific context in which each experiment runs. Especially, we focus on the access network context parameters: This includes parameters from the Radio Access Technology (more specifically for 4G technology) such as radio status during the experiment (RSRQ, RSRP, RSSI) and RTT against the target server (measured via ping).

B. *Dataset Description*

In this paper, we use collected data from 45 MONROE nodes, measuring 3 operators in Italy. Nodes operate in a different scenario such as university campuses, dense urban areas, countryside, aboard trams, and buses. The data is collected in more than of 3 months, from 1st of January 2018 to 15th of April 2018. This dataset focuses on RTT measurements. In total, we collected latency for about 238 million samples, as detailed in tab:dataset.

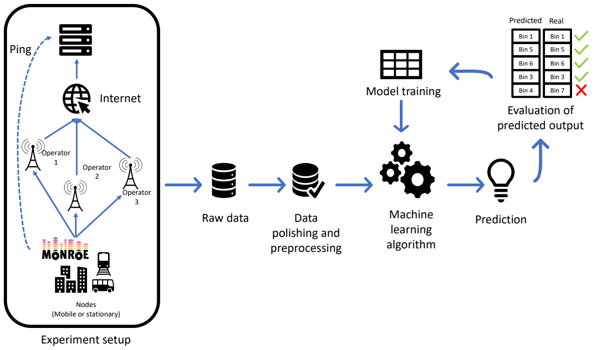

Fig. 1: Experimental setup and latency prediction methodology.

As a first step, a preprocessing phase was necessary to identify and remove anomalies and inconsistent samples from the dataset. Despite all the care we took in processing collected data to ensure that they are extensive, representative and with limited biases, some limitations need to be pointed out. It should be noted, that the use of commercial mobile subscriptions, despite being coherent with what end customers can get, on the other hand, limit our knowledge on the specific configuration each MBB is using (e.g., in terms of QoS, or presence of proxies commonly found in MBB networks [16]). We cannot guarantee that our results generalize for all possible subscription for these operators, nor for different locations. We explicitly opted to a specific scenario, i.e., location, in our tests.

III. Methodology

In this section, we provide a brief description of our proposed machine learning solutions. Firstly, the great effort in data cleaning based on domain knowledge should be pointed out, subsequently, feature selection was operated, then the processed data is feed to the classification and evaluation phase. Each classifier was trained with a subset of the input data containing also the respective label (the RTT class); they are hence tested with the rest of the input dataset, unlabelled.

- **Feature Selection:** Frequently, the data is represented with a large number of features which causes high complexity of the model and overfitting: model fits the parameters too closely to the particular observations in the training dataset but does not generalize well to real data (unseen dataset). However, feature selection algorithms automatically select a subset of features that are most relevant to the problem to improve computational efficiency or reduce the generalization error of the model by removing irrelevant features or noise, which can be useful

[3] The description and detail information about metadata is available on: https://github.com/MONROE-PROJECT/UserManual

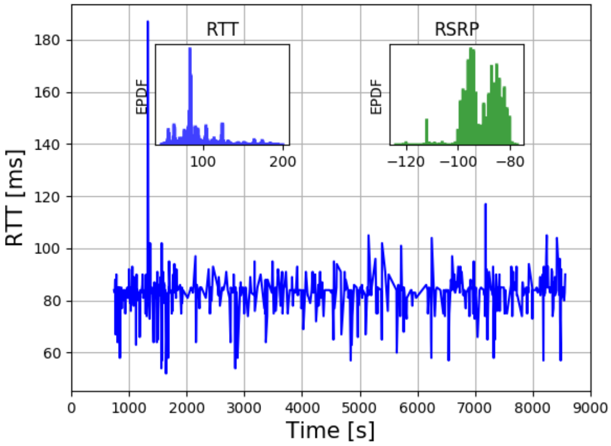

Fig. 2: An example of the temporal evolution of RTT, distribution of RTT in all dataset and distribution of RSRP.

for algorithms that do not support regularization [17], [18]. In this study, we rely on Random Forest as a feature selection algorithm. It exploits several decision trees to assign a label at each input record [19].

- **Classification:** In this work, we consider three different classifiers in order to understand which would suits better for our dataset and achieve better results. These are i) Logistic Regression (LR), ii) Support Vector Machines (SVMs), and iii) Decision Tree (DT). K-fold ($k = 10$) cross-validation technique for model selection is used: first, the dataset is divided in k folds, secondly, $k - 1$ folds are used as training set and 1 is used for the test set. Finally, this process is repeated k times until all folds are used 1 time as a test set (See. Fig. 1).
- **Performance metrics:** Three metrics are used to illustrate the performance of the classifiers, i.e., precision, recall, and f1-score. i) Precision: it defines as $\frac{TP}{TP+FP}$ where TP (True Positive) is the number of true positives and FP (False Positive) the number of false positives. It shows the ability of the classifier not to label as positive a sample that is negative. ii) Recall: it defines as $\frac{TP}{TP+FN}$ where FN (False Negative) the number of false negatives. It indicates the ability of the classifier to find all the positive samples. iii) f1-score: It is the average of the precision and recall ($2 * \frac{(Precision*Recall)}{Precision+Recall}$).

IV. Results

In this section, we present the characteristics of the selected features from the input dataset and the prediction results. Figure 2 shows the RTT (y-axis) temporal evolution for a specific node and operator. As depicted from Figure 2, during less than 3 hours the RTT value changes from 40 ms up to 190 ms. Two embedded figures inside Figure 2 illustrates Empirical Probability Distribution Functions (EPDF) of the RTT (left side) and RSRP (right side). These two figures indicate that there is no obvious correlation between just RTT and RSRP (Pearson correlation coefficient = -0.07 confirms our visual observation.).

TABLE II: Feature importance.

Feature	Importance	Definition
rsrp	0.893814	Reference Signal Received Power (LTE)
rsrq	0.041941	Reference Signal Received Quality (LTE)
rssi	0.024896	Received Signal Strength Indicator
timestamp	0.019767	Entry timestamp
band_3 band_7 band_20	0.005337 0.002701 0.002690	Band corresponding to the frequency used (e.g., 3, 7 or 20 in Europe)
cid_1 ... cid_n	0.002659 0.000310 0.000153	Cell ID
devicesubmode_0 devicesubmode_2	0.000089 0.000087	Connection submode for 3G connections
frequency_800 frequency_1800 frequency_2600	0.000081 0.000053 0.000012	Frequency in MHz (e.g., 700, 800, 900, 1800 or 2600 in Europe)
imei_1 imei_2	0.000012 0.000009	International Mobile Station Equipment Identity
imsi_1 imsi_2	0.000008 0.000002	International Mobile Subscriber Identity
imsimccmnc_op1 imsimccmnc_op2	0.000002 0.000001	Mobile Country and Network Code
ipaddress_True	0.000000	Assigned IP address
lac_1	0.000000	Cell Local Area Code
nwmccmnc_op1 nwmccmnc_op2	0.000000 0.000000	MCC and MNC as read from network

Typically, applications are not sensitive to small variations of RTT, we thus divide the latency in bins of 50 milliseconds length (11 bins in total). Then, we define three classes based on the bin value: (i) low ($< 100ms$), (ii) medium (100-200ms), and (iii) high ($> 200ms$). Our goal is to predict the RTT bin by having available physical features without any active measurement. Since we analyze data from 45 nodes under varying settings, coverage, and locations, we start with a subset of the whole data. We focus on a specific node with more than one million experiments. We consider domain knowledge to clean the dataset. Then, one-hot encoding is used to convert categorical features. The idea behind this approach is to create a new dummy feature for each unique value in the nominal feature column (e.g., $band \in 3, 7, 20$ convert to three new features band_3, band_7, and band_20).

Table II presents all features in our dataset and their importance computed by means of Random Forest (RF) with 1000 estimators. We use RF to measure feature importance as the averaged impurity decrease computed from all decision trees in the forest without making any assumptions whether our data is linearly separable or not. We see that the RSRP, despite the absence of visible correlation with the quantity to be predicted, turns out to be the most discriminating feature in the dataset based on the average impurity decrease in all trees. By observing the impact of the Radio Access Technology (3G and 4G/LTE) used to access the internet, we notice that, as ex-

TABLE III: Detail classification accuracy.

bin	precision	recall	f1-score	support
0	0.03	0.03	0.03	33
1	0.71	0.70	0.71	132453
2	0.58	0.58	0.58	86220
3	0.12	0.13	0.13	9635
4	0.04	0.04	0.04	1268

pected, the better performance offered by 4G/LTE technology clearly benefits the RTT. Coverage and time are among the most prominent causes of RTT degradation. However, their impact is not linear. For the rest of the paper, we use features with importance greater than 0.0001.

Finally, we consider three supervised machine learning algorithms, SVM, DT, and LR to predict RTT class from selected features. The model is validated by k-fold cross-validation in combination with grid search for fine-tuning the performance of a machine learning model by varying its hyperparameters values. To this end, we exploit nested cross-validation, Varma et al. [20] indicate the true error of the estimate is almost unbiased relative to the test set when nested cross-validation is used. The nested cross-validation performance of the SVM, DT, and LR models are 0.664 +/- 0.100 percent, 0.743 +/- 0.004, and 0.609 +/- 0.076, respectively. DT is notably better than the performance of the others. Table III shows the performance of the DT for the first 5 bins.

V. Conclusion

This paper presented a study of latency prediction on commercial mobile carriers using the data collected by MONROE open measurement infrastructure. The novelty of the study stands in the design of a machine learning application to analyze the sheer volume of data, nearly 238 million RTT measurements of 3 different operational MBB networks. Our results show the importance of data prepossessing and feature selection but considered basic algorithms did not provide high accuracy in this context.

The presented approach was designed to be scientifically sound and reproducible, and the exploited dataset is fully accessible by the community. It is ongoing research and we are working on further improvements on the algorithm selection, tuning, and their predictive capabilities.

Acknowledgements

This work was partially funded by the *MONROE* Alliance and the SmartData@PoliTO lab.

References

[1] J. Sommers and P. Barford, “Cell vs. wifi: on the performance of metro area mobile connections,” in *Proceedings of the 2012 ACM conference on Internet measurement conference*, pp. 301–314, ACM, 2012.

[2] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck, “An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance,” in *Proc. of SIGCOMM*, 2013.

[3] O. Fagbohun, “Comparative studies on 3g,4g and 5g wireless technology,” vol. 9, pp. 133–139, 01 2014.

[4] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, and H. Yan, “Modeling web-quality of experience on cellular networks,” in *Proceedings of the 20th annual international conference on Mobile computing and networking, MobiCom’14*, pp. 213–224, ACM, 2014.

[5] U. Goel, M. Steiner, M. P. Wittie, M. Flack, and S. Ludin, “Http/2 performance in cellular networks,” in *ACM MobiCom*, 2016.

[6] D. Baltrunas, A. Elmokashfi, A. Kvalbein, and [U+FFFD] Alay, “Investigating packet loss in mobile broadband networks under mobility,” in *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, pp. 225–233, May 2016.

[7] M. Xu, Q. Wang, and Q. Lin, “Hybrid holiday traffic predictions in cellular networks,” in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, pp. 1–6, April 2018.

[8] S. Souders, “The performance of web applications – one-second wonders for winning or losing pcustomers..” https://goo.gl/ErNLcm, Nov 2008. [Online; accessed 07-Aug-2018].

[9] E. S. (Bing) and J. B. (Google), “Performance related changes and their user impact.” https://goo.gl/hCr7ka, Jul 2009. [Online; accessed 07-Aug-2018].

[10] S. Souders, “High-performance web sites,” *Communications of the ACM*, vol. 51, no. 12, pp. 36–41, 2008.

[11] E. Bocchi, L. De Cicco, and D. Rossi, “Measuring the quality of experience of web users,” *ACM SIGCOMM Computer Communication Review*, vol. 46, no. 4, pp. 8–13, 2016.

[12] M. Yang, X. R. Li, H. Chen, and N. S. Rao, “Predicting internet end-to-end delay: an overview,” in *System Theory, 2004. Proceedings of the Thirty-Sixth Southeastern Symposium on*, IEEE, 2004.

[13] B. A. Nunes, K. Veenstra, W. Ballenthin, S. Lukin, and K. Obraczka, “A machine learning approach to end-to-end rtt estimation and its application to tcp,” in *Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on*, pp. 1–6, IEEE, 2011.

[14] O. Alay, A. Lutu, M. Peón-Quirós, V. Mancuso, T. Hirsch, K. Evensen, A. Hansen, S. Alfredsson, J. Karlsson, A. Brunstrom, A. Safari Khatouni, M. Mellia, and M. Ajmone Marsan, “Experience: An Open Platform for Experimentation with Commercial Mobile Broadband Networks,” *Proc. ACM MobiCom ’17*, pp. 70–78, 2017.

[15] B. M. Sousa, K. Pentikousis, and M. Curado, “Multihoming management for future networks,” *Mob. Netw. Appl.*, vol. 16, pp. 505–517, Aug. 2011.

[16] A. S. Khatouni, M. Mellia, M. A. Marsan, S. Alfredsson, J. Karlsson, A. Brunstrom, O. Alay, A. Lutu, C. Midoglu, and V. Mancuso, “Speedtest-like measurements in 3g/4g networks: The monroe experience,” in *Teletraffic Congress (ITC 29), 2017 29th International*, vol. 1, pp. 169–177, IEEE, 2017.

[17] S. Khalid, T. Khalil, and S. Nasreen, “A survey of feature selection and feature extraction techniques in machine learning,” in *2014 Science and Information Conference*, pp. 372–378, Aug 2014.

[18] F. Ferri, P. Pudil, M. Hatef, and J. Kittler, “Comparative study of techniques for large-scale feature selection,” in *Pattern Recognition in Practice IV* (E. S. GELSEMA and L. S. KANAL, eds.), vol. 16 of *Machine Intelligence and Pattern Recognition*, pp. 403 – 413, North-Holland, 1994.

[19] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.

[20] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, pp. 91–91, Feb 2006. 1471-2105-7-91[PII].