DOI: 10.26314/GARR-Conf19-proceedings-08

How to leverage Measurement Lab Internet metrics to get useful insights through Data Science

Stefania Delprete[1], Gianni Spalluto[2], Christian Racca[1]

[1]TOP-IX Consortium, [2] Eutelsat Communications

Abstract. Measurement Lab is an open source project that provides data on Internet performance measurements. TOP-IX Consortium decided to leverage the open data provided by M-Lab to explore the impact of its network on the territory and the factors that might influence the performances of the connectivity providers among its members, in particular FWA operators. This paper presents the data acquisition process and three case studies related to commercial plans comparison, coverage and weather, including the advantages and obstacles of the data used and approach presented.

Keywords. Internet, data science, open data, metrics, performances, weather.

Introduction

How can we leverage the biggest open dataset on Internet metrics to create reports for specific providers? This paper presents a viable approach to acquire, manage and explore the data, and comparing known information or aggregating other external sources.
If validated, this approach might be used to enrich the information that connectivity providers extract from their own networks to offer an objective multi-operator analysis and to represent a good tool to verify (thanks to the dataset openness) the reports created by the operators themselves or other authorities.

1. TOP-IX Consortium and its interest in Internet metrics

TOP-IX (TOrino Piemonte Internet eXchange) is a non-profit consortium founded in 2002 with the aim of creating and managing a neutral hub for the exchange of Internet traffic in North-West Italy. Furthermore, TOP-IX promotes and supports, through the Development Program (DP), technological, engagement and training projects based on the Broadband Internet, Data and People. These activities synergistically promote the growth of the territory.

According to the typical role of an Internet Exchange, TOP-IX operates at layer 2 in OSI model and it is not allowed, by law, to analyze the content of the exchanged traffic over the platform. Furthermore TOP-IX, in its role as a neutral hub, typically does not have the chance to study and track the impact of network events (port saturation, technical faults, other phenomenon) on final end-users.

This study is aimed at expanding the perspective by TOP-IX on its network backbone and at investigating the feasibility of using third-party datasets to provide value added

"services" (such as report, dashboard or more advanced tools) to the stakeholders active on the IX platform.

2. Measurement Lab project and data availability

TOP-IX, since the beginning, demonstrated a strong interest in data collected by network performance analysis tools. During the years TOP-IX activated valuable collaborations with Ookla (1) and Measurement Lab (2) in order to host their tools for performance monitoring.

We started analysing data from the Speedtest by Ookla and, even if it was an insightful starting point to analyse Internet metrics, it couldn't be compared with the amount of variables offered by M-Lab.

Measurement Lab, founded in 2009 to offer a better solution to Internet measurement experiments, collects when available more than 150 variables (3) including: log time, geolocation, browser and operating system, and Internet metrics such as Round-trip delay time (RTT). This kind of approach and transparency attracts interest in analysing data often not available elsewhere.

M-Lab provides a detailed documentation to express the most common network metrics (latency, download and upload speed) applying known formulas in literature on a set of parameters.
For example, from the raw data of a speed test, the download speed in Mbps is expressed by the formula (4):

```
8 * (web100_log_entry.snap.HCThruOctetsAcked /
(web100_log_entry.snap.SndLimTimeRwin +
web100_log_entry.snap.SndLimTimeCwnd +
web100_log_entry.snap.SndLimTimeSnd))
```

We would like to thank M-Lab team, in particular Chris Ritzo and Roberto D'Auria, for their assistance in the interpretation of the data acquisition processes and parameters.

3. Tools for gathering data and managing analysis

Measurement Lab uses Google BigQuery to store the tests results in different Tables. The download Table and upload Table contain test results already filtered based on outliers and possible mistakes. Through a whitelisted email address we were able to access Google BigQuery using standard SQL and explore viable solutions to gather the dataset needed for our research.

After exploring the Python BigQuery library, we did a preliminary analysis. We used a Jupyter Notebook to import the library and directly generate the Pandas DataFrames from a SQL query in order to explore and collect the results in one place.

For our purposes we also decided to write a script in Python, with the assistance of Massimo Santoli, to uniquely match the IP addresses in M-Lab tests and TOP-IX members' ASNs (Anonymous System Numbers).

Through this process we were able to have a clean dataset in .csv format related to selected members of TOP-IX Consortium.

4. Case studies

This section presents the on-going case studies developed by TOP-IX using the dataset and the approach described above. Currently, we decided to focus our attention on the FWA (Fixed Wireless Access) members of the Consortium. The analyses were performed in Jupyter Notebooks using Python and its computational libraries (NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn).

4.1 Profiling connectivity providers and studying correlation between network performance metrics and commercial public plans

The Data Science project started by exploring different FWA operators from an aggregated point view, to understand which parameters could be addressed and to verify the amount of data and test performances over the years.

A small number of providers was selected for a deeper exploration. This analysis has been useful to compare commercial public plans (openly published by operators) and performances measured through the speed tests. Fig. 1 is an example of how much the speed test can mirror the offerings by noticing the visible horizontal lines on precise Upload speed values. This correlation has been studied by provider and regional area, and aggregating by city and IP.

4.2 Role of TOP-IX network as "digital enabler"

TOP-IX has a limited visibility on the network and rarely it can observe the actual impact on the final users. Data from this experiment represents a good proxy for a wider study aimed at exploring, for example, the overall coverage (in terms of connectivity), by the

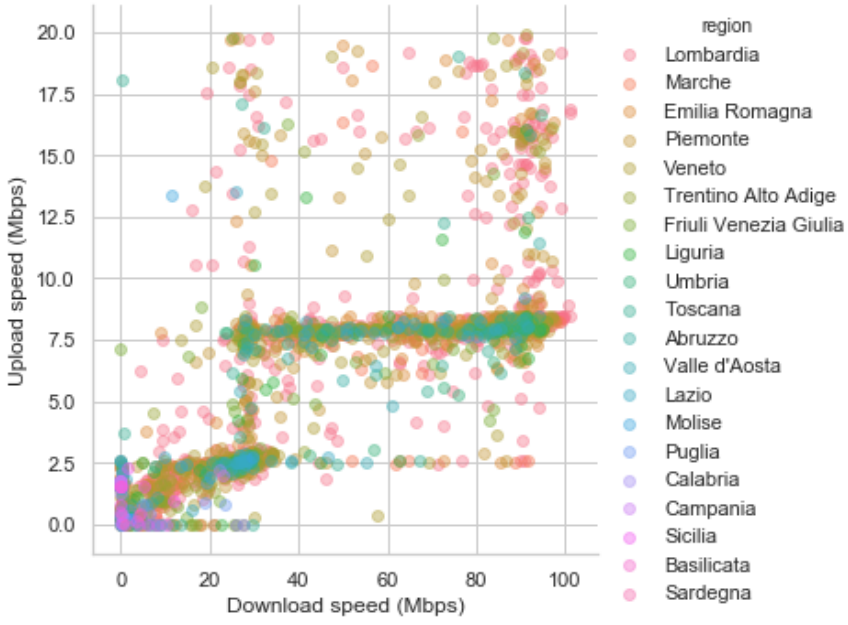

Fig. 1
Correlation of upload and download speed for a provider over different regions and focusing on download speed values below 110 Mbps and upload speed values below 20 Mbps

network operators interconnected to TOP-IX.

The analysis started with an exploration to avoid outliers and understand the most covered areas macroscopically. In a second phase we used QGIS (see an example in Fig. 2) to visualise the data using different layers and parameters to make more visible the area with the best performances.

Interns involved in this study: Domenico Gallo, Christian Bellafemmina

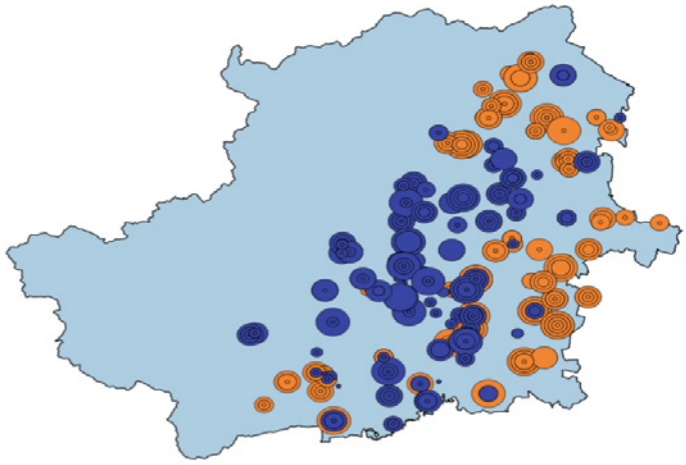

Fig. 2
Two different providers mapped in Turin area, the radius of each point is proportional to the download speed value

4.3 Correlation between performance, weather conditions and other phenomena

The website of the Regional Environmental Protection Agency, ARPA (5), offers, under request, the possibility to access the latest weather data (temperature, precipitation, humidity, wind speed, intensity and direction) organized in .csv datasets. Fig. 3 shows part of the experimental approach comparing one of the weather variable (wind intensity) over time.

The measurements of the selected meteorological variables are organized by daily time bands. The aggregation with available Measurement Lab data allows us to test the possible correlation between network performances and weather conditions. Unfortunately the values of the Pearson correlation coefficient were not relevant to make concrete assumptions for the cases studied until this point.

Interns involved in this study: Gianni Spalluto, Paul Cristian Prisecaru, Adriana Muscau.

Conclusions and further development

The methodology and the use-cases presented above open new possibilities to mix data science and networking. The goal is to get insights from the raw data collected by Measurement Lab as a possible integration and an alternative to “traditional” network monitoring system and official reports not released in real time.

Aggregating datasets from multiple sources might give a better awareness of users’ behaviour to improve network reliability and to offer more customized services from the vendor point of view.

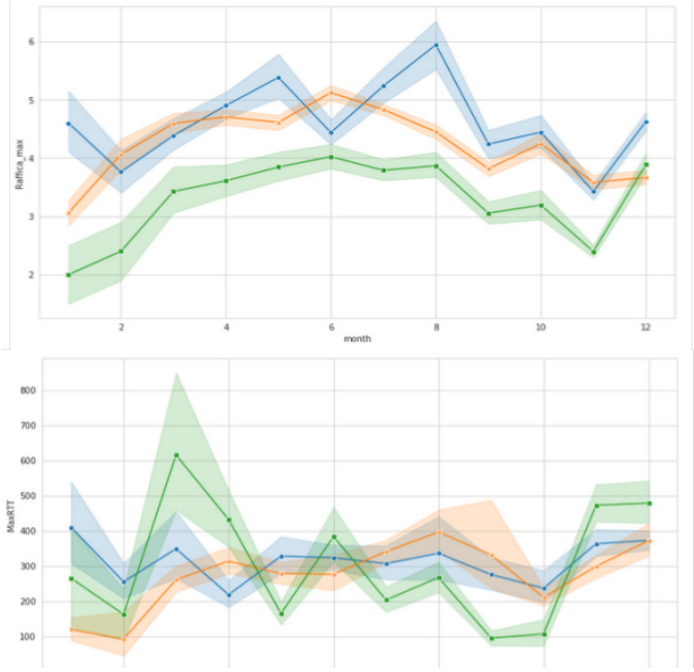

Fig. 3
Example of trends of wind intensity and latency for three different cities over twelve months.

On the other hand we are still working to deal with the imprecision and the gaps in the datasets. For instance both M-Lab and Speedtest by Ookla are using MaxMind as a geolocation reference, this kind of process can lead to incorrect conclusions regarding behaviours based on location.

Furthermore, regarding the integration with weather condition data, for some urban and rural areas, the information is missing or incomplete, or accessible only through private services.

Future developments will include a detailed and precise exploration and pre-processing of the dataset to avoid any misleading conclusions, and further analysis of less common metrics incorporating Machine Learning algorithms to test the possibility to predict performances and stability of the system.

References

Speedtest by Ookla https://www.speedtest.net

Measurement Lab https://www.measurementlab.net

“BigQuery Schema” by Measurement Lab
https://www.measurementlab.net/data/docs/bq/schema

“Calculating Common Metrics for NDT Data” by Measurement Lab
https://www.measurementlab.net/data/docs/bq/ndtmetrics

Weather dataset by ARPA Piemonte http://www.arpa.piemonte.it

Authors

Stefania Delprete - stefania.delprete@top-ix.org

Data Scientist, Python tutor, and BIG DIVE co-organizer at TOP-IX with a background in Theoretical Physics and strong interests in Neuroscience, Human Rights, and Social Change. Stefania has been involved in projects in Italy, Germany and UK on productivity and self-improvement, social aid, and open source as a speaker and tech tutor. She volunteers for PyCon/Pydata conferences, and runs local chapters and events for Mozilla and Rust, Effective Altruism, and MathsJam.

Gianni Spalluto - gianni.spalluto1991@gmail.com

Graduated in Sociology and Applied Social Science, passionate in data analysis and convinced supporter of the Big Data revolution as a catalyst for Social Change and a more equitable society.
Gianni attended IFTS Big Data course and, thanks to internship in TOP-IX, started using Python tools in Data Science projects. He's currently working at Eutelsat group as a human resources analyst where he uses the large amount of data available in the department to support company welfare policies and work life balance of employees.

Christian Racca - christian.racca@top-ix.org

After graduating in Telecommunication Engineering at Politecnico di Torino, Christian joined TOP-IX, working on data streaming and cloud computing, and later on web startups. He has mentored several projects on business models, product development and infrastructure architecture and cultivated relationships with investors, incubators, accelerators and the Innovation ecosystem in Italy and Europe. Currently Christian manages the TOP-IX BIG DIVE program aimed at providing training courses for data scientists, companies, organizations and consultancy projects.