Where Are The (Cellular) Data?

Maryam Amini, Razvan Stanica, Catherine Rosenberg

Where Are The (Cellular) Data?

MARYAM AMINI, University of Waterloo, Canada
RAZVAN STANICA, Univ Lyon, INSA Lyon, Inria, CITI, France
CATHERINE ROSENBERG, University of Waterloo, Canada

New generations of cellular networks are data oriented, targeting the integration of machine learning (ML) and artificial intelligence solutions. Data availability, required to train and compare ML-based networking solutions, is therefore becoming an important topic and a significant concern. Operators do collect data, but they rarely share it because of privacy concerns. This paper starts by reviewing the few publicly available cellular datasets, which created bursts of innovation with their release. The scarcity of such data is so acute that researchers are collecting network data using their own tools, developed in-house and covered by the second part of this survey.

CCS Concepts: • **Networks** → *Network measurement*; Mobile networks; • **General and reference** → Surveys and overviews.

Additional Key Words and Phrases: Cellular networks, Datasets, Measurement tools

1 INTRODUCTION

Data is not a new subject in cellular networks. Data are and have always been a critical part of the network for both research and operation purposes. Up until now, data have been used for different tasks ranging from the most traditional ones, like billing and network monitoring, to more advanced operations like network optimization, reconfiguration and user mobility management. From another point of view, data are also critical for the research community to validate their models and evaluate their solutions. With the subject of data availability in mind, this paper focuses on data collection and datasets as opposed to data analysis or usage [54].

The importance of data in cellular networks is highlighted by the fact that they will play a critical role in enhancing the performance of the fifth generation (5G) of networks. Specifically, Enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC) are the three generic 5G services which are enabled through novel technologies like New Radio (NR) [92], massive Multiple Input Multiple Output (MIMO) [88], dual connectivity [43], etc. The 3rd generation partnership project (3GPP) has recently published its Release 17 [42] with specific expectations for these services. With strict performance requirements for various key performance indicators (KPIs), mobile network operators (MNOs) need to deploy, organize and monitor their infrastructure and services in a much smarter way than before. Therefore, machine learning (ML) and artificial intelligence (AI) will play a critical role in enhancing the performance of the network. However, ML and AI techniques are data intensive, and proper designing, training, and evaluating them relies on large datasets.

The architecture and business model in 5G are more open than those of previous generations, and this has resulted in the emergence of new stakeholders [44]. This broader model is enabled through a set of partnerships and open-source projects among new players and established actors from the mobile network community. For instance, the Next-Generation Radio Access Networks (NG-RANs) [38] aims to bring flexibility, scalability, smartness, and interoperability through disaggregating network functions, programmable interfaces, and an open architecture. This new paradigm is attracting many new players to the realm of cellular networks, which is a potential economic game-changer. O-RAN [46] is one of the biggest projects based on this concept of next-generation RANs. Many companies and small-to-medium enterprises (SMEs) are now working on different architectural designs and solutions to test and integrate the general 3GPP requirements of NG-RANs. To train and evaluate their AI-based solutions, these new players require *real* data.

Therefore, by sharing data, MNOs would empower all the new stakeholders with the right tools to test their technical solutions, which would eventually contribute to increasing the overall quality of the network.

However, the truth about datasets that are derived from real scenarios and contain a useful set of features is that they are extremely rare. The lack of open datasets has limited the abilities of researchers in academia and SMEs to develop new solutions for cellular networks. Unfortunately, many research groups working on ML-based network management solutions overlook a highly important topic: data accessibility. While it is tempting to use synthetic data to train and evaluate such solutions, since this eliminates most of the hassle of finding the right *real* data, this never completely captures the complexity that can be observed in the real world. Using these synthetic datasets as a training or testing source limits the capacity of AI/ML models to adapt to new environments. Indeed, since synthetic datasets are only representative of the major trends observed in reality, deploying a model trained on such datasets in a real environment will raise major convergence and accuracy challenges.

Moreover, not all datasets contain the same type of information. Indeed, there are hundreds of KPIs exposed by different cellular network equipment, placed in various locations of the mobile network. These KPIs can be monitored and logged by different network probes, either hardware or software, and they can represent different layers of the protocol stack and different interfaces. This means that even the few existing datasets are heterogeneous, and they can not be easily combined to produce a large training set for ML-based algorithms [82]. To further exacerbate the problem, the ML solutions themselves might exploit different features from the data, limiting even more the datasets available for training.

Faced with this lack of publicly available datasets, more and more researchers are trying to use software tools and capture the data they seek by themselves. However, at least for now, such data collection has only taken place on the user side, over very limited populations, or in some very specific situations on the server side. We find this issue to be surprising, and even disturbing, in a field where AI and ML have been intensively promoted in the last few years. Therefore, the purpose of this paper is to survey: *i)* the existing open datasets in the cellular networks community, and *ii)* the tools available to capture cellular network data.

To the best of our knowledge, this paper is the first at presenting a comprehensive organization of publicly available cellular network datasets and data collection software tools. In short, our list of contributors is as follows:

- First, we introduce a list of publicly available datasets in three main categories, namely, user-side, network-side, and server-side datasets. We discuss the available datasets in each group. Moreover, in Table 1, we provide a summary of the main features of each dataset, including the collection type, data granularity, the type of software and the devices used for the data collection campaign, and some related applications to each dataset.
- Second, we provide a list of open software available for data collection. We have grouped these software tools into two main groups. Software tools in the form of mobile applications that help capture data on the user equipment (UE) side, and some open RAN and core network solutions that enable deploying a cellular network and hence capturing data on the network side. Moreover, in the network-side category of open software tools, we have included a subcategory for open platforms, which represent network infrastructures that provide researchers with an interface to test and evaluate their new algorithms.

In the remainder of this paper, we will first discuss the various data sources in cellular networks in Section 2. In Section 3, we provide a review of the publicly available datasets that have been used in the cellular network literature, with a particular focus on the features of the datasets. Our goal for Section 4 is to review the data collecting tools

available to researchers and the information they provide. Finally, in Section 5, some of the challenges of using available datasets and generating data are identified, followed by a conclusion.

2 WHAT IS USEFUL DATA?

As depicted in Fig. 1, there are many entities in a cellular network, each of them capable of generating data. Based on the entity responsible for collecting data, the datasets can be classified as: *i)* user-side datasets, *ii)* network-side datasets, *iii)* server-side datasets, and *iv)* network topology datasets.

Since cellular networks are deployed to serve mobile users, a significant part of the data are coming from the mobile terminal devices, owned by the users. The advantage is that both network-level and service-level KPIs can be collected on the UE side. However, these user-side datasets can only provide a limited and indirect view on the state of the cellular infrastructure. They are also usually collected on a limited number of devices, a few tens or, in the best case, hundreds. This highly impacts the interest and the validity of these data collection campaigns.

On the MNO side, there are numerous entities in the RAN and core network where data can be collected. Indeed, cellular infrastructure can generate a huge amount of data, from different sources such as the base stations, the mobility management entity, the billing database, the user plane packet gateway, etc. This leads to a different type of problem, since every MNO equipment and interface exposes tens of KPIs. Selecting the adequate location in the network and the right KPIs to log for a given data-driven networking problem is a complicated task. This is about to become even more complex in 5G, where a supplementary dimension in this metrology process is added through the virtualisation of network functions, no longer associated with precise physical equipment.

The services situated beyond the core network, and mainly on the Internet, represent the third source of data. There are numerous routers and servers residing on this side of the network, which may be used to collect information by content and service providers. This solution is interesting for studies focused on a specific service, and major industrial actors clearly make significant usage of this type of data to technically and financially optimise their products. However, datasets collected on the server side only provide information relative to the upper layers and are agnostic of anything taking place on the mobile operator side.

Finally, MNOs operate networks on radio spectrum conditionally allocated by governments, and the infrastructure they deploy is considered as public information in many countries. Therefore, cellular network topology information is often released as open data by governmental agencies.

With all these possible data sources, it is fair to explore in more detail what makes a dataset useful in the context of cellular network research. As a matter of fact, even when data are available, it is not always appropriate for the task of interest.

To better exemplify this, let us consider for instance a user mobility prediction task in two different scenarios. In the first one, a well-known publicly available dataset, like *roma/taxi* [59] or *eplf/mobility* [106] is used to address the problem of predicting the next location of a UE. This prediction can be helpful in a number of networking problems, such as handover acceleration [98], quality of service (QoS) enforcement [128], or resource allocation [62]. Both datasets exemplified above are captured on the UE-side and contain spatio-temporal information of users, so they seem appropriate for the task. However, the issue with these datasets, which are in fact very popular in the topic of mobility prediction, is their exclusive observation of users in driving conditions. Therefore, training an ML model with these datasets will result in a model that is not able to fully recognize all the different mobility patterns.

In a second scenario, imagine a set of Call Detail Records (CDRs) [54] used as the training data for a mobility prediction model. CDRs contain detailed information about any network service usage of the users, collected at the

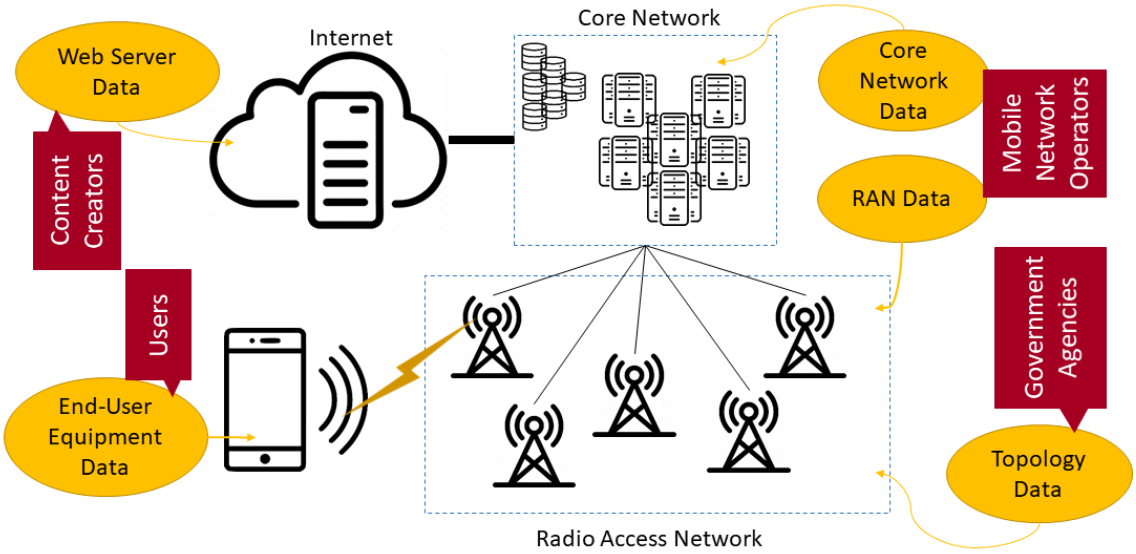

Fig. 1. Possible locations for data collection probes.

operator end. In this case, a considerably broader range of users are captured in the dataset, representative this time of all transportation modes. However, CDRs are very coarse in time granularity [81], since they are collected only when a user is actually active on the cellular network. Therefore, using a set of CDRs that are recorded with an average granularity of 15 minutes may not be the best option for the next-location prediction of a high-speed user [119].

To summarize, cellular network research explores a large number of problems, each coming with its own specificity and requirements in terms of data. This means that there is no such thing as a one-size-fits-all dataset. Each problem necessitates specific information, and even when data are available, it might not be suitable for each and every study.

Table 1. A Summary on Cellular Networks Publicly Available Datasets

Reference	Technology	Date	Collection Type	Granularity	Software	Device	Multi Operator	UE Mobility	PHY Measurements	Dataset Size	Related Applications
[73]	Bluetooth	2005	User side	5 mins	BlueAware	Nokia 6600	NA	NA	no	285 MB	mobility
[59]	GPS	2014	User side	15 secs	NA	NA	NA	yes	no	1.61 GB	mobility
[106]	GPS	2009	User side	10 secs	NA	NA	NA	yes	no	380 MB	mobility
[132]	2G/3G/4G	2015	User side, Server side	1 sec	YoMoApp	Android devices	yes	NA	no	NA	YouTube streaming, QoE monitoring and prediction
[70]	LTE/WiFi	2014	User side, Server side	1 min	Cell vs. WiFi	Android devices	yes	NA	yes	393 MB	comparing LTE and WiFi, DASH
[67]	LTE	2016	User side	Message level	tcpdump	NEXUS 5	NA	NA	no	2 GB	MPTCP
[58]	3G/4G	2016	User side	10 secs	BW Collector	2 Android devices	NA	yes	no	1.5 MB	study of adaptive bit rate algorithms, QoE analysis, mobility
[131]	LTE	2016	User side	1 secs	NA	Huawei P8 Lite	NA	yes	no	384 KB	study of adaptive bit rate algorithms, QoE analysis, mobility
[86]	3G/4G	2019	User side	NA	Viprinet	2 Cameras & a mic	yes	yes	yes	729 KB	telemedicine, remote monitoring
[80]	2G/3G/4G	2017	User side	4 secs	Carle Profiler	NEXUS 5 NEXUS 5X	yes	yes	no	837 KB	telemedicine, remote monitoring
[108]	LTE	2018	User side	1 sec	G-Net track	Samsung J5	yes	yes	yes	19.8 MB	mobility and handover prediction, DASH
[95]	LTE/non-LTE	2018	User side	100 ms	LTE Signal Logger	NEXUS 5X	yes	yes	yes	582 MB	network performance analysis, network optimization, DASH
[83]	LTE	2020	User side	1 sec	NA	Huawei Modem E392	no	yes	yes	451 KB	RF planning, radio channel analysis
[111]	LTE	2020	User side	2 secs, 60 mins	Android API, ipref	LG K4	no	yes	yes	36.94 MB	network coverage and performance analysis
[118]	LTE	2020	User side	500 ms	Keysight NEMO	Samsung Galaxy Note 4	no	yes	yes	130.9 MB	network coverage and performance analysis
[117]	LTE	2020	User side	500 ms	Keysight NEMO	NA	yes	yes	yes	9.86 MB	mobility, network coverage and performance analysis
[107]	5G	2020	User side	1 sec	G-NetTrack Pro	Samsung S10	no	yes	yes	23.5 MB	mobility, 5G mmWave, network performance analysis
[100]	5G	2020	User side	1 sec	Android API	Samsung Galaxy S10	yes	yes	no	913 KB	mobility, 5g mmWave, application and network performance analysis
[101]	5G	2020	User side	1 sec	Android API	Samsung Galaxy S10	no	yes	yes	1.6 MB	mobility, 5g mmWave, application and network performance analysis
[122]	5G/LTE	2021	User side	1 min, 10 secs	FCC, SigCap	Google Pixel 2, Pixel 3, Pixel 5	yes	yes	yes	6.4 MB	mobility, 5G mmWave, network performance analysis
[102]	5G	2022	User side	1 sec	Accuver XCAL	3 Samsung Galaxy S21 Ultra 5G	yes	yes	yes	79.9 MB	5G mmWave, beam management, network performance analysis
[48]	5G/LTE	2022	User side, Network side	2 secs, 8 ms	Huawei API	Huawei LTE Modem E3772, Huawei CPE PRO 2	no	NA	yes	69.22 MB	YouTube streaming, cloud gaming, network performance analysis
[130]	5G/LTE	2022	User side, Server side	1 sec	G-Net Track Pro	Samsung Galaxy S21 5G, Samsung Galaxy S8	no	yes	yes	4.05 MB	YouTube streaming, QoE analysis, network performance analysis
[125]	5G O-RAN	2022	Network side	Message level	srsRAN	USRP B210	no	no	yes	26.49 MB	O-RAN, energy efficiency, computing power
[91]	2G/3G/4G	2016	User side	Message level	Mobile Insight	Android devices	yes	NA	yes	NA	network performance analysis, 5G mmWave, security, energy efficiency
[97]	Bluetooth	2019	User side	2.34 meters, 1.64 meters	NA	BQ Aquaris X5 plus, Samsung Galaxy S6 & A5	NA	NA	yes	451 KB	indoor localization and fingerprinting
[68]	OFDM	2021	User side	5 millimeter	NA	4 USRP devices	NA	NA	yes	17.04 GB	indoor localization, Massive MIMO analysis
[77]	LTE	2021	User side	1 centimeter	srsLTE	USRP B200mini	no	yes	yes	1.89 GB	indoor localization, signal processing
[105]	5G	2023	Network side	80 ms	NA	NA	no	yes	yes	1.13 GB	direction and time of arrival estimation, indoor positioning and navigation
[64]	WiFi	2022	Network side	10 ms	Nexmon	Asus RT-AC86U	NA	yes	yes	2.2 GB	device-free localization, environment sensing, CSI obfuscation
[129]	RIS	2023	NA	NA	MATLAB API	Pico Technology PicoVNA 106	NA	NA	yes	1.2 GB	reconfigurable intelligent surface
[85]	3G	2015	Network side	15 mins	NA	NA	no	NA	no	10.8 GB	mobility, urban structure, traffic planning, energy efficiency
[61]	2G/3G	2018	Network side	NA	NA	NA	no	NA	no	738.7 MB	mobility
[26]	5G/4G/3G/WiFi	since 2013	Server side	User level	RTR-NetTest	NA	yes	no	yes	NA	network performance analysis, comparing LTE and WiFi
[66]	5G/4G/3G	since 2012	Server side	User level	FCC-ST	NA	yes	no	yes	NA	network performance analysis
[94]	social networks	2012	Server side	User level	Twitter, Facebook, Google+	NA	yes	no	no	845 MB	UE communication pattern detection, social networks
[135]	streaming	since 2019	Server side	User level	Puffer	NA	yes	no	no	NA	study of adaptive bit rate algorithms, QoE monitoring
[123]	streaming	2020	Server side	User level	Last FM, Deezer	NA	yes	no	no	15.3 MB	UE communication pattern detection, social networks
[90]	email	2007	Server side	User level	NA	NA	yes	no	no	1.6 MB	UE communication pattern detection
[87]	email	2004	Server side	User level	NA	NA	yes	no	no	1.7 GB	UE communication pattern detection, spam detection
[121]	P2P	2002	Server side	User level	Gnutella	NA	yes	no	no	520 KB	P2P and multi-agent networks

3 PUBLICLY AVAILABLE DATASETS

Out of the four main groups of datasets on cellular networks discussed in Section 2, publicly available datasets mainly belong to the user-side category. In rare occasions, network-side and server-side datasets have also been published. It is the objective of this section to discuss all these approaches, also summarized in Table 1.

Regarding the final data category discussed in Section 2, network topology datasets are published by government agencies and National Regulatory Authorities (NRAs) in numerous countries. For instance, the Federal Communications Commission (FCC) [12] provides information on the U.S. licensing of radio frequency spectrum to the general public. Body of European Regulators for Electronic Communications (BEREC) [33] is an agency in charge of regulating the telecom market in Europe that frequently issues documents and reports on such developments. The Canadian Radio-television and Telecommunications Commission (CRTC) [9] and also the government of Canada's online spectrum management system [29] provide online access to the spectrum management and authorization data used for communications and broadcasting. Moreover, the Government of Canada's data portal [16] contains datasets on the coverage, availability, regulatory, financial, and subscription information. Since these datasets are country specific and largely available, we do not further discuss them in this section.

On top of the publicly available datasets, the last years have shown the emergence of a number of actors collecting mobile network data for commercial purposes. These actors can be divided in three categories. The first group is the one of data brokers, such as Tutela [32] or Predict.io [36], who propose software development kits (SDKs) which can be integrated in any mobile network application. In this case, whenever an application integrating the SDK is used on a UE, data are logged and some measurements are conducted. The second type of actor, such as RootMetrics [25] in the U.S., conduct actual field tests, using a diversity of UEs and targeting multiple KPIs. Finally, MNOs themselves have business entities providing services based on the data they collect in their network. For example, in France, two MNOs provide such services: Flux Vision by Orange [35] and Geostatistics by SFR [37]. However, since our focus is on public datasets, which can be used by the community to train ML models, these commercial (and generally very expensive) alternatives are not further discussed. In the following, we begin by presenting the user-side data collection campaigns, followed by network-side and, finally, server-side public datasets.

3.1 User-side datasets

Multiple datasets related to cellular networks have been generated, collected, and published throughout the years on the user-side.

Each of these datasets provides different kinds of measurements on different layers of the protocol stack. 3GPP has defined the measurement control concept and has classified the UE-reported measurements into different categories, i.e. inter/intra-frequency, inter-system, traffic volume, quality, and UE internal measurements [41]. In short, Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), Received Signal Strength Indicator (RSSI), and signal-to-noise and interference ratio (SINR) are defined as the main reported physical layer (PHY) measurements by the UE. Furthermore, with 5G NR, new PHY measurements are introduced. For instance, Reference Signal Time Difference (RSTD) and Uplink (UL) Angle of Arrival (AoA) are specially designed for the new positioning services. It is worth noting that Channel Quality Indicator (CQI) is a report from the UE to the base station rather than a measurement itself. Also, in contrast to the other measurements described above, the process by which the UE computes the CQI is vendor-specific and has not been standardized by 3GPP.

3GPP has also introduced other measurements for higher layers in [39, 40]. In these documents, detailed information on measuring different performance indicators has been provided. The only measurement that is performed by the UE for the layer two (L2) is the packet delay. Other measurements like the packet loss rate, the number of active/inactive users, etc., are all performed by the base station in a per UE or an aggregated manner. In higher layers, this aggregation is done over a group of Network Functions (NFs), for instance, the latency along a network slice.

We may reasonably assume that a wide range of variables influences the quality and interpretation of each measurement. For instance, throughput is a widely used indicator of the network performance. However, the base station transmit power, the number of active users in the cell, or the interference from the adjacent cells play a significant part in the achieved throughout by each UE [112]. Raida *et al.* [110] have shown that in a perfect condition, where there are no active UEs in the cell other than the measuring UE, and there is no interference from other cells, RSRP can be an indicator of the achievable throughout by the UE. Moreover, in [109] a thorough investigation on the impact of different parameters on the UE achievable throughout has been conducted. Therefore, having knowledge of the measurement campaign setup and data collection process can be very beneficial in understanding or unveiling the spatial or temporal behavior of these measurements. In this survey, we do not intend to analyze the datasets. Our main focus is discussing the availability of datasets, and in the case of unavailability, how to generate the desired datasets. Interested readers are encouraged to read [55, 99] to get insight on the mobile data analysis process.

BlueAware [73] may be among the very first publicly available mobile datasets. This dataset was originally generated to capture user context patterns. Multiple traces containing information about the communication, location, and social activities of 100 volunteers were published within this dataset. Although Bluetooth was used as the primary technology in this work, this dataset contains spatiotemporal information of mobile nodes, which can represent the movement of mobile users in a cellular network. This kind of information is of great value and can be used for various topics related to mobility such as resource allocation, network planning and optimization, user experience, traffic management, and location-based services [133].

Some datasets widely used in the literature do not even contain any cellular network KPIs, but simple user mobility information in the form of global positioning system (GPS) traces. To give a few representative, but not exhaustive examples, user mobility traces can be found in *roma/taxi* [59] and *eplf/mobility* [106]. These datasets do not contain any cell related information, but represent trajectory information of taxis in different cities around the world. One possible application of these datasets is location prediction in cellular networks [126]. These datasets are especially useful in ultra-dense networks, where the overlapping and dense deployment of cells makes mobility management very challenging [72]. The dataset in [59] includes GPS trajectories of 320 taxis in Rome, Italy. The dataset features include an identifier, timestamp, and location of each taxi with a granularity of 15 seconds. The dataset used in [106] is another taxi mobility traces collected in San Francisco with GPS coordinates, timestamps, and the occupancy of the cabs captured every 10 seconds. These are very helpful features that can be used for training models to predict the next location or trajectories of high-speed vehicular users.

Of course, the most relevant user-side datasets for cellular network research are collected using cellular UEs. Some of these datasets were not collected for generic purposes, but in very specific contexts. For example, *Cell vs. WiFi* [70] was a study conducted in 2014 to compare smartphone application performance under different network connectivity options: WiFi, LTE, and multi-path TCP. This study managed to collect almost 10 GB of data from 750 users over 180 days in 16 different countries. The measurements focused on link throughput and signal strength for cellular and WiFi networks, as well as on some other metrics allowing to compare single-path and multi-path TCP: DNS lookup time, ping round trip time (RTT), uplink/downlink goodput. This dataset could be used to study the relative performance of

LTE and WiFi networks under different conditions, or to identify areas where one type of network may be more suitable than the other. Also, it can be used to evaluate and optimize different rate adaptation techniques for Dynamic Adaptive Streaming over HTTP (DASH). In the same context of multi-path TCP performance evaluation, the dataset published in [67] investigates the behavior of different applications through logs collected with *tcpdump* on a proxy server with a dozen of clients using Android smartphones over a period of 7 weeks. This dataset is a great resource for understanding the performance of MPTCP in cellular networks, and how congestion affects the performance of cellular networks.

In the absence of bandwidth-specific measurements, Bokani et al. [58] and Hooft et al. [131] launched separate data measurement campaigns in order to generate comprehensive bandwidth datasets. [58] provides 2 datasets with spatiotemporal information and bandwidth measurements on the UE side under driving conditions for 3G and 4G networks. A similar dataset to those in [58] was published by Hooft et al. [131] in Ghent, Belgium, containing spatiotemporal information of the measuring UE and the obtained DL throughput under six different mobility patterns. These datasets can be leveraged for analysis of the mobility of UEs in an urban area. Moreover, the bandwidth traces are essential for the study of rate-adaptive video streaming algorithms, particularly in real-world settings.

One of the main 5G technologies that help augmenting the network capacity and coverage is massive MIMO. In a massive MIMO system, a large number of antennas are employed at the base station, and the signals from these antennas are carefully coordinated to enhance the communication link between the base station and the UE. Channel estimation is one of the main challenges of massive MIMO systems, and the availability of large datasets with Channel State Information (CSI) is greatly beneficial for massive MIMO solutions. Arnold et al. [47], have discussed this issue and proposed a channel sounder architecture that can measure multi-antenna and multi-subcarrier CSI at different frequency bands, antenna geometries, and propagation environments. This channel sounder enables the generation of massive MIMO datasets for analyzing the channel properties. Moreover, [76] discusses a new paradigm for generating 5G datasets for high-precision positioning in integrated sensing and communication systems. This model generates detailed CSI data [75] for UEs based on features of massive MIMO channels.

There are a few datasets available with CSI KPIs and measurements [64, 68, 77, 97]. However, not all these datasets are collected on a cellular network, though. In fact, [97] has been collected over a Bluetooth low-energy network. This dataset contains information on the locations as well as the measured RSS values of the UEs at certain locations. [64] contains a comprehensive collection of CSI measurements obtained from the access point side of a WiFi system in an indoor environment. This dataset has been utilized to study CSI-based device-free localization, PHY-level privacy, and investigation of different obfuscation systems[63]. The extraction of CSI information from the access point was done using Nexmon [127], a C-based firmware capable of extracting CSI from OFDM-modulated WiFi frames (802.11a/(g)/n/ac) on a per-frame basis. Both [68, 77] are datasets collected in massive MIMO systems. [68] is collected using a 64-antenna massive MIMO testbed located in KU Leuven. These measurements contain four different antenna array topologies, and each topology has a dataset with a size of 252,004 samples, which makes it one of the largest massive MIMO datasets. Gassner et al. have introduced an open CSI-specific dataset in [77]. This dataset is collected over an LTE network, with CSI-specific KPIs like RSSI, RSRQ, RSRP. Datasets with PHY information of massive MIMO systems are helpful in understanding the propagation conditions and determining the optimal deployment of antennas. They are especially useful for the study of sensing, location-based services, and the problem of indoor localization. The authors of [104] have investigated the joint problem of the estimation of the angle of arrival and delay using uplink SRS signals in a 5G network. For this study, a dataset [105] was generated using commercial gNB, and UE devices. This dataset contains both angle and delay information of the direct and reflected paths for the uplink samples in an indoor environment. In the context of PHY measurements, there is a scarcity of Reconfigurable Intelligent Surfaces (RIS)-based datasets.

The recent open dataset presented in [129] which provides measurements for 30 different scenarios with multiple antenna arrangements, positioning and spectral reflection onto the RIS partially fills that gap. To mitigate the influence of multipath propagation, these measurements were conducted within an anechoic chamber located in Germany.

Cellular network measurements during mobility conditions are much prized in the research community. In this sense, both [86] and [80] collect datasets focused on the cellular communications of moving ambulances. The general context of these studies is that of clinical multimedia communication, enabling pre-hospital transfer of images and sound from patients. Cellular networks provide the best infrastructure for this communication. The dataset in [86] provides spatio-temporal information of the vehicles, as well as the RSSI at specific locations. These measurements help to detect the areas with coverage holes on the different trajectories taken by an ambulance. With QoS requirements of clinical multimedia communications in mind, the dataset collected in [80] presents features like the uplink and downlink throughput, as well as spatio-temporal information and speed of the vehicle. Unlike [86], this dataset is confined to upper layer indicators and does not include channel quality measurements. Both these datasets have the potential to support a wide range of telemedicine applications like patient diagnosis or treatment planning. When it comes to the context of cellular networks, these datasets are great resources for training ML-based mobility models and improving the remote communication between the healthcare professional and the patient.

A more detailed mobility-oriented analysis is proposed in [108], where 4G data are collected for five different mobility patterns: static, pedestrian, car, tram and train. The 135 traces, coming from an unknown number of users, present an average duration of 15 minutes, and they are collected via an Android application called *G-NetTrack*. The dataset contains information on the quality of the cellular channel, cell settings, user location, and throughput measurements. The use-cases suggested by the authors focus on evaluating the performance of different HTTP Adaptive Streaming (HAS) algorithms and on handover analysis with user spatio-temporal information. In addition to the data traces, which have a temporal granularity of 1 second, a synthetic dataset obtained from the ns-3 simulation of a 4G network with 100 users was also generated. This synthetic dataset provides finer network-side measurements, with a time granularity of 250 ms, to complement the dataset collected from real users.

An Android application capable of collecting cellular network data with a granularity of 100 ms is developed in [95] and used to collect around 500 traces from 3G and 4G networks in multiple cities around the world, using a crowdsensing approach. To achieve such a fine granularity, the application only measures spatio-temporal information and reference signal metrics, which only require listening to the cellular broadcast channel. RSRP and RSRQ are the main collected metrics. The traces provided with this dataset can be used to study, analyze, and optimize the performance of cellular networks. Moreover, since the dataset contains CSI measurements, it can be used for analyzing different DASH techniques under cellular network conditions.

To study the temporal dynamics of KPIs in an LTE network, Raida *et al.* [111] collected week-long measurements, in collaboration with an Austrian operator. The resulting dataset [114] has been used to expose distributions and correlations between different network metrics. For instance, the authors conclude that RSRP does not show any diurnal patterns, while RSRQ and throughput do exhibit such patterns. In other data collection campaigns conducted by the same team [117, 118], two open datasets were generated [115, 116] with measurements from LTE networks using *Keysight NEMO* devices. These special probing devices equipped with NEMO software allow a time granularity of 500 ms and expose a significant number of KPIs. To cite just a few, cell-specific signal measurements, channel state information, physical link adaptation information (such as transport block size, physical resource block utilization, modulation and coding schemes), and throughput are among the information provided in the two datasets. The differences in the two datasets come from the fact that [115] provides 90 hours of LTE downlink measurements in an empty and controlled

room, while [116] provides traces from train drives capturing information from three Austrian operators. All these datasets contain detailed information on rank, modulation and coding scheme (MCS), transport block sizes, cyclic redundancy check (CRC) failures, and hybrid automatic repeat request (HARQ) retransmissions. Therefore, they are especially valuable in topics like network coverage and performance analysis, resource allocation, physical resource block (PRB) scheduling, and sub-band CQI computation.

Once the first 5G deployments started emerging around the world, data collection targeted these networks and, in the cases discussed below [100–102, 107, 122, 130], they have even been made available to the public.

The dataset in [107] is collected using *G-NetTrack Pro*, also used above for 4G scenarios by [108]. The data are coming from the 5G networks of two major Irish operators, across two different services (video streaming and file download), and two mobility patterns. The features included in the dataset are channel quality metrics, cell information, user spatio-temporal information, and throughput KPIs. Much like in their previous work [108], the authors have also generated a synthetic dataset with the 5G mmWave module of ns-3 to complement the collected data. These traces, which are produced from an operational 5G network, can be used for network performance prediction, revealing correlations between different KPIs or validating results of projects conducted on mmWave-5G.

With mmWave deployments being an important novelty in 5G, several data collection campaigns focused on this technology. For example, Narayanan et al. [100, 101] have conducted two studies on 5G mmWave performance. First, *5Gophers* [100] contains KPIs of the 5G networks of three major carriers in U.S., in Minneapolis and Chicago. These measurements include user spatio-temporal information, cell ID, 5G service status, and network KPIs like the throughout or RTT. The experiment includes two different mobility scenarios. Several scenarios were studied using these traces, e.g., the 5G performance for stationary UEs, mobile UEs, and multiple applications performance over 5G networks. One of the main findings of this study was the inefficiency of performance prediction in mmWave-5G networks based on UE location. In a follow-up study [101], the authors collect more 5G data in mobile scenarios. More precisely, they focus on the UE location, speed, direction, signal strength and throughput across three different mobility patterns. This dataset is further utilized in an ML framework, *Lumos5G*, to produce a dynamic 5G map of the city of Minneapolis. Both these datasets are focused on 5G mmWave performance. They are also great resources for analyzing the performance of mmWave subjected to various mobility patterns or how the environment affects mmWave performance. Other relevant areas of study using these datasets include the analysis of various applications and the investigation of mobility and handover.

Rochman *et al.* [122] conducted a study to compare the channel quality and network performance metrics of 5G NR with those of the legacy network. They collected data from three U.S. operators in the region of Hutchinson Field, Chicago, and also made a detailed study on the performance of the mmWave deployment of one specific operator in downtown Miami. In this project, multiple Android applications, various frequency bands, and technologies were used to generate a comprehensive dataset [2]. Since this dataset provides both LTE and 5G, it can be used to investigate the differences between the performance of the two networks in bandwidth utilization, latency, and throughout. Moreover, the Miami dataset also includes a mixture of 5G mmWave and 4G measurements. This dataset can be beneficial for the study of mmWave performance.

In another study, collaborators of [100] have conducted a 5G measurement campaign project in Chicago to gain insights into the real-world performance of 5G mmWave deployments. This dataset, which is called *5GBeam*, provides detailed 5G measurements, including UE's spatio-temporal information, and beam-specific metrics. This dataset is collected in two different locations with different characteristics, namely an open-field baseball park, and a downtown area. The collaborators have also used different mobility patterns to enhance the understanding of the propagation

properties of 5G mmWave technology. The beam-specific metrics provided in the dataset were collected using a professional 5G test tool called *Accuver XCAL*[1].

With the advent of 5G deployments, it has become crucial to examine the users' QoE, particularly for conventional services such as eMBB. It is essential to understand whether 5G can meet the demands of these services before exploring new service possibilities. Two datasets address this [130] and [48]. They have been made available publicly to analyze the performance of YouTube on 5G networks. Both datasets provide detailed channel-level and QoE metrics. However, they differ in their data generation approaches. While, the contributors of [130] opted to utilize a French Mobile Network Operator (MNO) and collected data on commercial phone devices, employing G-Net Track Pro, the authors of [48] have taken the approach of deploying a small 5G cellular network and collecting the data both on the UE and the network side. This 5G testbed is built using commercial software and hardware like *Amarisoft* for the gNB-side and Huawei modems to perform the UE tasks. This way the authors have been able to collect performance metrics not only on the UE side but also on the network side. However, the usefulness of such a dataset is questionable in view of the size of the testbed.

Finally, in a recent publication, Lozano et al. [124] describe their work on setting up an O-RAN compliant testbed and introduce the datasets [125] they have collected over this testbed. The project's main objective is to gain a comprehensive understanding of the energy consumption and performance of the next generation of RANs, where various elements from different vendors are integrated to work together seamlessly. The accompanying datasets are categorized into three distinct groups: Firstly, to study the computing usage of the RAN. Secondly, to examine the energy consumption of the RAN, offering information on energy consumption profiles and efficiency metrics. And, finally, to explore the joint impact of network and edge service configurations on energy consumption and system performance.

3.2 Network-side datasets

While operators continuously collect data, for billing, network management and even legal requirements, they rarely openly share it. This is mainly due to potential privacy problems, since operator data might reveal very detailed user information. Therefore, the numerous research papers exploiting and analysing MNO data [54] are usually the result of non disclosure agreements signed between operators and different research groups. These datasets are not public, and they can not be used by the community, for example to train pertinent ML models. However, there are a few major exceptions to this rule, discussed below.

Operator data usually take the form of CDR. These records contain detailed information about any network service usage of the users, such as voice call duration, the caller and callee identifiers, source and destination of short messages (SMS), as well as the identifier of the cell used by the UE to access these services. The time granularity of CDR is not fixed, it is dictated by the user activity, as information is only logged in case of user communication. This also means that this data source is biased towards very active users, who are over-represented in the corresponding datasets [119]. Despite this, CDRs have long been a rich source of information for operators, albeit with very limited access for the research community.

Nokia Mobile Data Challenge (MDC) was among the very first campaigns to collect and share mobile data with the research community. In [89], the organizers of the event discuss the challenges they faced to hold the campaign, such as inviting volunteers and distributing them devices, collecting the data over a year (2011-2012), handling the privacy concerns, and then creating a contest with the participation of the research community. MDC was a great indicator of

[1]available at https://www.accuver.com/

the power of network-side information. Although planning the challenge, collecting the data and the competition itself took almost three years, it was well-received by the community and more than 100 submissions of multi-disciplinary ideas exploiting the available data were accepted by the committee.

In the line of the huge success of Nokia MDC, Orange also held a data challenge called *Data for Development* (D4D). This took the form of a competition with two editions, one in 2012 [56] and one in 2014 [69], with the objective of finding concrete applications of mobile phone data in the development of infrastructure, health and environment sectors in African countries. The open challenge provided the contestants with CDR of over 5 million users in the Ivory Coast [56] and more than 9 million users in Senegal [69]. While the D4D challenge raised a huge interest in the research community, with hundreds of submissions for each edition, the datasets were shared by Orange under strict non-disclosure agreements and only for the duration of the challenges. This is mainly a consequence of privacy concerns from the MNOs who, faced with numerous studies trying to deanonymize mobile data [74], prefer to keep control of their datasets. Very recently, another data challenge, the *NetMob 2023* Data Challenge [93], was held with the collaboration of Orange. This 4G dataset is based on captured traffic in 20 metropolitan areas in France for 77 consecutive days in 2019. For this challenge not only the spatio-temporal information of eNodeB infrastructure, but also information on data usage for different classes of services was released. Unfortunately, this dataset is not publically available, and it was only available for a brief time, under non-disclosure agreement.

However, in two rare occasions, CDRs have been made publicly available, namely CDRs available through the *Telecom Italia Big Data Challenge* [85], and from the *Metropolitan Region of Rio de Janeiro* [61]. These datasets do not contain any per-user information. Instead, the CDRs are aggregated over a geographical region, which reduces the privacy risks.

In 2014, Telecom Italia in association with EIT ICT Labs, SpazioDati, MIT Media Lab, Northeastern University, Polytechnic University of Milan, Fondazione Bruno Kessler, University of Trento, and Trento RISE organized a Big Data Challenge. A large collection of various datasets on the urban life of people in the city of Milan and the province of Trentino was published [85]. Simultaneous access to multiple sources of information has made this dataset ideal for tackling various problems such as energy consumption [57], urban structure [71], traffic planning [136], and many more. The aggregated CDRs published as part of this dataset contain features on incoming/outgoing calls, sent/received SMSs, and Internet usage of Telecom Italia customers. All recorded transactions are between Telecom Italia users, which accounts for almost 34% of the population in the studied areas [50].

Another available aggregated dataset represents mobility information extracted from a set of CDRs from 2.9 million mobile users in the Metropolitan Region of Rio de Janeiro (MRRJ), Brazil that was collected during 2014 [61]. The CDRs initially contain voice call information on the entire MRRJ and four smaller neighboring municipalities. In [49], the original CDRs are used to extract spatio-temporal modeling of the human mobility in the studied region. The output of this study includes estimated origin-destination (OD) matrices, accompanied by real commuting data extracted from individual interviews, that are used for validating the extracted patterns from the CDRs. Despite the fact that the mobility data provided by this dataset is neither precise nor granular, it is a valuable resource for urban transportation planning due to the large population it covers during a lengthy timeframe.

As it can be noticed, the publication of network-side datasets, with several open challenges organised at the beginning of the last decade, has practically stopped in the last years. This is, beyond any doubt, the consequence of more stringent privacy regulations. Afraid of negative publicity, MNOs are reluctant to share even aggregated datasets. Moreover, the few available network-side datasets are CDRs, collected in the core network for billing purposes, and they do not contain any fine grained information related to the radio access network, e.g., traces per logical channel. MNOs of course collect this lower level data, which represents the basis of their network management strategies. However, these

radio-level KPIs are not normalised and they are vendor-specific, meaning that it is difficult even for operators to build consistent datasets. Fine grained RAN data, with millisecond resolution required to follow the activity on the logical channels, is also massive and difficult to share because of its size.

3.3 Server-side datasets

For performance monitoring purposes, service providers continuously collect data regarding their users and their interactions with the proposed services. However, as in the case of the MNOs, these datasets are rarely shared with the research community, because of user privacy and business reasons. Moreover, most services nowadays are generic web-based, available to users through all kind of communication technologies. This means that server-side datasets are not necessarily representative of the behavior of cellular users or of the performance of cellular networks.

With service providers reluctant to share their data, several datasets were collected independently using the application programming interface (API) proposed by the service. This is mainly the case for social networks, which allow exploring users interactions through their APIs. For example, the contacts and communities (*i.e.*, circles) of specific users were collected in 2012 by McAuley and Leskovec [94] for three different social networks: Twitter (1000 users), Google+ (133 users) and Facebook (10 users). Together with surveys conducted by the authors, these datasets were used to study the structure of human communities. Similar graphs of user interactions are collected in [123] for two music streaming services: 7 thousand LastFM users and more than 28 thousand Deezer users, in 2020. In these datasets, the music preferences of the users and their interactions (follows, likes) are recorded.

While these datasets are mainly used for sociological studies, they can also help for networking purposes, such as in the prediction of a user communication patterns. This is the case not only for social networks, but for other services, such as email. As a matter of fact, several email datasets are available in the literature. Metadata from more than 3 million emails from a European research institution, including timestamp, sender and recipient of the email, were collected in 2007 in [90]. Moreover, one of the best known email datasets, known as the *Enron corpus* [87], also contains the actual content of the emails. This dataset has been released in 2004 during the legal investigation of the Enron corporation and it has been widely used to evaluate anti-spam filters [84].

Another class of services where data can be collected without the direct involvement of a service provider is represented by services based on peer-to-peer (P2P) networks. In these networks, crawlers can be used to explore the peers in the network, such as demonstrated in [121], where more than 400 thousand nodes of the Gnutella file sharing system are mapped, based on traces collected in 2002. For each discovered peer, the dataset contains the IP address, port number, number of shared files and their total size.

Setting up the service and starting data collection on their own servers is an additional choice for researchers seeking access to server-side data. However, in these kinds of settings, the generated datasets are susceptible to a lack of user diversity. If the research team is successful in promoting their service to a large number of users, this issue will not emerge; otherwise, the dataset will be severely constrained to the activities of a small group of users. For instance, in order to capture the server-side data on video streaming services, researchers in [135] set up *Puffer* [24], a free and open-source live TV streaming website. At the time of this writing, Puffer is an ongoing project with more than 150 thousand users and more data are added to the publicly available dataset every day since January 2019. These measurements contain information on the video chunk like the size, timestamp or coding, congestion window size, RTT, and client-side information like the client's buffer status, or acknowledgment status of the chunk.

As explained, these datasets do not specifically distinguish cellular network users from WiFi and desktop users. An important exception is *YoMoApp* [132], a fairly large dataset on YouTube Quality of Experience (QoE) in cellular

networks. The contributors of this project have developed an Android application and a cloud-based dashboard[2] that allows access to raw data measurements and log files of individual devices to their owners, as well as an aggregated and anonymized public dataset. Data were collected in from more than 3 thousand YouTube sessions over the course of five years (2010-2015). The measured features vary from application layer KPIs like buffer status or playtime, to network information like TCP segment length. This dataset is excellent for understanding the behavior of YouTube streaming, monitoring and prediction of QoE, and data transmission optimization. A further study on the dataset [134] revealed that, over time, the QoE of YouTube videos has improved on cellular networks.

While data collection at server level can be a tempting source for individuals seeking to get access to mobile data, NRAs around the world are also interested in collecting server-side crowdsourced data to assess and benchmark the performance of different MNOs for a long time. In a technical report [52], *BEREC* has specified how to implement a QoS measurement software tool that can be used by NRAs. Moreover, in the Annex to [51], *BEREC* has also introduced some of the measurement tools and platforms that are used by different NRAs throughout Europe to assess Internet quality. Generally, the scope of NRAs covers all the aspects of the telecommunications industry, including wired and wireless communications, the Internet, broadband, cable, and satellites. Therefore, most of these crowdsourcing tools/platforms and the data collected by them are not specifically designed for cellular networks or cannot differentiate if a sample is captured on a cellular network or not. That being said, the software tools and the data collected by some NRAs include detailed information on the network performance as well as the type of access technology. In some cases, the data are made available to the public.

For instance, since 2012, the American FCC has been researching mobile broadband services performance. As part of this study, the collected data have become available to the public. The aggregated dataset [66] is collected, periodically enriched and released based on measurements from the users on the FCC website. These measurements contain information on the wireless performance parameters like upload and download speed, latency, packet loss, and signal strength, the types of UE device used, and the tested operating system versions.

Several European NRAs have opened up access to their Operator benchmarking and crowdsourcing data. First, the Austrian regulatory authority for broadcasting and telecommunications has implemented a crowdsourcing architecture called *RTR-NetTest* [26]. This implementation does web-based or software-based, UE-initiated, speed tests to capture the QoS of broadband Internet access provided by both fixed and mobile networks. As part of the objective of this project, the anonymized measurements are published as open data[3]. These measurements contain various information about the network performance such as uplink and downlink throughput, delay, and signal strength, plus the spatiotemporal and cell information. However, the number of samples captured on a mobile network is reduced. Adopting the same measurement methodology as RTR-NetsTest, NRAs in Slovenia (Test Net by AKOS) [7], Slovakia (MobileTest by SPECURE) [30], and Czechia (netMetr by CZ.NIC, discontinued in April 2022, access to their open data is still available) [19] made their own measurement platforms and made their collected data publicly available.

4 TOOLS FOR DATA COLLECTION

After reviewing the publicly available datasets in Section 3, it becomes clear that the scarcity the community experiences in terms of data limits the development of ML/AI-based solutions. Although they are the best placed to collect and share large datasets, it is unlikely that mobile operators or service providers will publicly open data usable to train ML models to the community, because of legal (user privacy protection) and business (commercial secret) reasons. The

[2]accessible at http://yomoapp.de/dashboard
[3]available at https://www.netztest.at/en/Opendata

Table 2. A Summary on Available Software Tools

Software Tool	Collection Type	Software/Hardware Requirements	Network Connection Technology	Granularity	Root Access Requirement
FCC Speed Test [65]	User side	Android/iOS Operating System	Cellular (2G to 5G), WiFi	5 secs	No
tPacketCapture [1]	User side	Android Operating System	Cellular, WiFi	Message level	No
tcpdump [31]	User side	Android Operating System	Cellular, WiFi	Message level	Yes
G-NetTrack [15]	User side	Android Operating System	Cellular (2G to 5G)	1 sec	No
SigCap [27]	User side	Android Operating System	Cellular (5G, LTE), WiFi	10 secs	No
NetMonitor [3]	User side	Android Operating System	Cellular (2G to 5G)	1 sec	No
MobileInsight [91]	User side	Android Operating System	Cellular	Message level	Yes
Open BTS [60]	Network side	Range Networks SDR1, USRP	2G/3G	NA	NA
free5GC [13]	Network side	NA	5GC	NA	NA
open5G core [4]	Network side	NA	5GC	NA	NA
Open5GS [21]	Network side	NA	EPC/5GC	NA	NA
OpenAirInterface [103]	Network side	USRP, Blade RF, Lime SDR	5G-NSA/5G-SA	NA	NA
srsRAN [78]	Network side	USRP, Blade RF, Lime SDR	5G-NSA/5G-SA	NA	NA

most conspicuous path towards network data availability appears to be the collection of large datasets by independent studies. This section discusses a series of software tools currently available to research teams willing to conduct such an independent data collection campaign. We divide these tools in two categories: *i)* user-side tools, and *ii)* network-side tools.

4.1 User-side Tools

There are a number of smartphone applications available on digital distribution stores that detect cellular networks, capture cell information, or even perform basic signal measurements. Almost all of the following tools work with the Android operating system, where some of them extract measurements from the chipset and require root access grants, and others do the task using the Android API.

First of all, the United States FCC has an official Speed Test (ST) application [65] that measures broadband mobile network performance. Throughput, latency, jitter, packet loss, cell identifiers, signal strength, and context information for 5G and LTE networks are captured through this application. FCC-ST is only available in the U.S. for both Android and iOS devices. An aggregated dataset [66] is collected, periodically enriched and released based on measurements from the users on the FCC website.

Packet analyzers are common for TCP/IP level data logging on desktop operating systems, and especially on Linux. *tPacketCapture* [1] and *tcpdump* [31] are Android packet capturing tools, with the latter requiring rooted devices. These tools are used to collect classical pcap traces, which can be later replayed using any packet analyzer.

Several tools have been designed specifically for smartphones and cellular networks. *G-NetTrack* [15] is a monitoring tool for 2G to 5G networks. This Android application records context information, cell measurements such as RSRP, RSRQ, signal-to-noise ratio (SNR), channel quality indication (CQI), network performance metrics, location and route information for drive tests, voice/SMS/data tests, indoor mode measurement, and dual SIM support. G-NetTrack does not require any root access and uses the Android API to perform the measurements. The achievable time granularity of the measurements is one second. *SigCap* [27] is another software tool working directly with the Android API. It captures WiFi and LTE/5G cellular networks measurements. The application collects user context information, WiFi-related information, plus cellular-related information such as cell ID, signal strength, RSRP and RSRQ. Similarly, *NetMonitor* [3] is a cellular network monitoring tool that captures channel quality, cell information, data throughput, and location

information and exports them in log files. This tool has the particularity of measuring channel quality KPIs not only on the serving cell, but also on neighboring cells. Moreover, dual-SIM devices are supported by this tool.

To get even more detailed traces, *Mobile Insight* [91] is an Android application that helps with cellular network logging on the chipset of mobile devices. This software tool only supports Qualcomm chipsets and needs root access. It can expose protocol messages in both the control plane and (below IP) data plane. This tool is designed to collect fine-grained network data, it offers an API to enable others to build and extend their own frameworks, and also shares a large dataset collected by MobileInsight users, or as the contributors say: *"for the community and by the community"* [17]. MobileInsight fully supports 3G/LTE networks. In a recent update, a version that supports 5G NR has been released.

4.2 Network-side tools

For decades, running a cellular network and collecting traces from it was only possible for mobile operators. These tasks used to require significant expertise and specialized equipment, not available to other organisations. However, two recent trends democratized cellular network experimental campaigns. First of all, the virtualisation of mobile network functions allows running a cellular network on regular computers. Second, open and easy to use implementations of cellular network software became available in the last few years. Therefore, it is largely possible nowadays for a research team to set up a private mobile network for experimental purposes and collect cellular data.

4.2.1 Open-source RAN and Core Solutions. The first open-source implementation of the cellular network protocols was realized by the *OpenBTS* project [60]. Focused on GSM technology and voice communications, OpenBTS allows running a self contained cellular network, including the GSM core network, soft switches and a telephony system, on a single computer.

More alternatives appeared with the development of the 5G technology, virtualized by design. Three fully operational 5G core network (CN) implementations are freely available. The *free5GC* project [13] is compatible with the 3GPP Release 15, and it is led by the National Chiao Tung University. The *Open5GCore* effort [4] is conducted by Fraunhofer FOKUS and the Technical University of Berlin. It is fully compatible with 3GPP Release 16. Finally, *Open5GS* [21] which also complies with the 3GPP Release 16 and implements 5G and LTE core networks.

The two most complete software tools available today, covering both RAN and CN implementations, are *OpenAirInterface* [103] and *srsRAN* [78]. Both these solutions can integrate software defined radios and enable the deployment of an entire cellular network, providing Internet access to off-the-shelf smartphones. The performance of OpenAirInterface and srsRAN was compared and discussed in detail by Gringoli *et al.* [79]. They were used in numerous experimental studies on 5G, but we are not aware of any extensive dataset collected using these tools.

4.2.2 Open Platforms. An open platform is a network infrastructure created to let researchers and developers test and assess new algorithms and applications in a real-world operating network. These platforms are typically built on top of existing networks and are intended to be versatile and flexible, enabling researchers to easily experiment with various configurations and scenarios. An alternative to building a cellular network from scratch using the current open-source network-side tools is to leverage the open experimental platforms. Open platforms often include tools and resources for data collection. Therefore, they can play a significant role in generating data for the research community.

There are a number of open 5G platforms, available for researchers to use. In this subsection, we introduce some, that have been used by the community. But, it is important to note that these are just a few examples. Depending on the country and region, different initiatives and projects might be in place to operate such platforms.

Colosseum [96, 120] is one of the most powerful open testbeds, that allows large-scale experiments intended for the next generation of wireless systems. With 256 programmable Software Defined Radios (SDRs), operated by the Institute for the Wireless Internet of Things at Northeastern University, Colosseum allows large datasets to be created on specific environments and conditions defined by the users for AL/ML training and testing. *Arena* [53], is the other open-access wireless testbed, housed at Northeastern University and operated by Wireless Networks and Embedded Systems (WiNES) laboratory. Arena supports a variety of applications such as MIMO schemes, multi-hop, and ad hoc networks, spectrum sensing and secure wireless communications through 24 SDRs, 12 servers, and 64 antennas.

The Platform for Advanced Wireless Research (*PAWR*) [22] project in the US is focused on enabling experiments with new wireless systems. At the time of this writing, there are 3 main operating *PAWD* platforms available and one under construction.

- *the Platform for Open Wireless Data-driven Experimental Research (POWDER)* [23] in Salt Lake City: This platform, which is operated by University of Utah, Rice University, and Salt Lake City, provides a collection of SDRs and antennas that enable experiments on OpenRAN, massive MIMO, RF monitoring, and over-the-air operations.
- *Cloud Enhanced Open Software-Defined Mobile Wireless Testbed (COSMOS)* [11]: is a large-scale experimental platform that is set up in a heavily populated area of New York City. mmWAVE, optical switching technologies, and edge computing capabilities are supported by COSMOS using programmable SDRs, antenna modules, optical transport network, core, and edge cloud.
- *Aerial Experimentation and Research Platform for Advanced Wireless(AERPAW) [6]*: located in North Carolina, *AERPAW* is the first wireless experimental platform designed for the application of unmanned aerial vehicles (UAVs) for 5G technologies and beyond. *AERPAW* provides the hardware, software, and flight capabilities to run experiments related to 5G communications by UAVs.
- *Wireless Living Lab for Smart and Connected Rural Communities (ARA) [8]*: Finally, *ARA* is the fourth *PAWD* platform and under construction at the time of this writing. *ARA* is surrounded by the farms and rural communities of the city of Ames, IA, aiming to feature applications in precision agriculture. The *ARA* deployment is expected to be fully deployed by 2024 and will be connected to the other *PAWR* platforms.

The Virginia Tech COgnitive Radio NETwork (*CORNET*) [10] is a large-scale testbed with 48 remotely accessible SDR nodes that are used for research and education purposes. The USRPs that were installed in the Wireless @ Virginia Tech research labs in 2008 served as the foundation for the creation of this testbed. Using open-source software and flexible hardware, the CORNET nodes facilitate research and education on topics like cognitive radio (CR) and dynamic spectrum access (DSA). CORNET offers a wide range of experimental research and educational tools, including an FCC experimental license agreement for several frequency bands.

OneLab [20] is a European partnership between five research institutions with the aim of developing testbeds used for network computer communications available to both enterprise and scientific researchers. This consortium provides access to many testbeds around Europe like *Future Internet Testing Facility* FIT [14], or w-iLab [34]. OneLab is administered from the NOC (network operations centre) located at the LIP6 Laboratory at Sorbonne Université. Each testbed is equipped with specific hardware and software to support various applications.

The *MONROE* project [18], also known as the *Measuring Mobile Broadband in Europe* project, is a research project funded by the European Commission Horizon 2020 research and innovation program. With a focus on 4G and 5G technologies, the project seeks to enhance the monitoring and analysis of mobile broadband performance throughout Europe. *MONROE* aims to improve the measurement and analysis of mobile broadband performance in Europe, with a

focus on 4G and 5G technologies. It involves a consortium of research institutions and industry partners to develop new measurement methods and tools for evaluating mobile broadband performance and to provide insights and recommendations for industry stakeholders on how to improve the quality and availability of mobile broadband services. *MONROE* consists of 150 mobile nodes that have been deployed on trains and buses, and 450 nodes scattered across Europe to collect data under various scenarios [45].

The Scientific Large-scale Infrastructure for Computing/Communication Experimental Studies (*SLICES*) [28] is a distributed infrastructure built to support large-scale, experimental research on networking protocols, radio technologies, services, data collection, parallel and distributed computing. *SLICES* is built on previous European experiences like *Fed4Fire+* [5] which was a project operated under the European Union's Horizon Program, providing open, and accessible testbed facilities to researchers, with over 20 testbeds located in Greece, Spain, Belgium, France, and Switzerland, for wired, and wireless applications like 5G, IoT, Big data, SDN, and cognitive radio. *SLICES* will make a completely virtualized, remotely accessible Europe-wide research infrastructure available. At the time of this writing, this project is in the preparation phase and is expected to start its operation by 2024.

5 CONCLUSION

Cellular networks have been around for decades and 5G technology is currently being deployed, with 6G on the horizon. With machine learning solutions being considered as the next industry game changer, cellular data are critical for their development and training. In other scientific fields making extensive use of AI/ML solutions (e.g., image processing), large datasets are available to the research community, fostering research. Meanwhile, the lack of data availability in the wireless networks community is clearly blocking innovation in the field and represents a frustrating experience for researchers.

In this paper, we surveyed the different types of data that can be collected in a cellular network and discussed the available datasets. In most cases, researchers have access to network topology data, provided by national agencies. However, this is contextual data at best, and it is not enough to properly train ML models.

User-side data represent most of the datasets available in the literature. Their advantage is that they can be easily tailored for different scenarios, including specific KPIs when needed and customisable granularity. On the downside, these datasets are generally smaller and heterogeneous, which is problematic when trying to use them in the design of an ML solution.

Although network-side probes are the most promising data sources, they present some bias, especially with respect to active and inactive users [119]. However, these datasets are very rare in practice, since MNOs have legal and business constraints that dissuade them from sharing this data. In fact, MNOs sometimes share datasets with research collaborators under non disclosure agreements. This practice does not seem to be enough in the context of AI/ML applications.

Server-side datasets are also quite rare in the field, because of the same legal and business reasons as MNO data. Moreover, server-side data present an important inconvenience through the fact that they do not generally distinguish between cellular, WiFi and desktop users. In the case of the open crowdsourcing datasets that are published by different NRAs to benchmark the performance of MNOs, there are some factors that impact the quality of measurements [113]. Different traffic shaping strategies may be applied to UEs depending on their data plan limit, which will affect the value of throughput at the UE-side and invalidate the process of comparing the performance of MNOs. Additionally, UEs can be wary of using too much bandwidth with their limited mobile data and avoid running measurement tests all together.

We argue that, if machine learning solutions are really expected to be integrated in the cellular network architecture, non-aggregated and finely granular data availability is essential. Since operators and service providers are unlikely to share the required data, the research community needs to take control of the issue. In this sense, we provide a survey of the software tools available for cellular data collection. These tools take the form of smartphone applications, for data collection on the user side, and open-source RAN and CN platforms, for data collection on the network side. Overall, in the absence of publicly available datasets, open platforms appear to be a promising substitute for large datasets that are collected on commercial networks. With access to such a platform, massive measurement campaigns can be launched to collect data from an end-to-end network. Especially, in the eye of 5G technology, open 5G platforms are expected to be built based on open standards and protocols, white-box hardware, and interoperable interfaces. This approach will promote generating high volumes of data that are not hardware-specific and can be more representative of the actual performance of the applications.

REFERENCES

[1] 2015. *tPacketCapture.* Retrieved August 5, 2022 from https://play.google.com/store/apps/details?id=jp.co.taosoftware.android.packetcapture
[2] 2021. *GrantPark.* Retrieved August 5, 2022 from https://people.cs.uchicago.edu/~muhiqbalcr/grant-park-may-jun-2021/CSV-GP.zip
[3] 2021. *NetMonitor Pro.* Retrieved August 5, 2022 from https://play.google.com/store/apps/details?id=ru.v_a_v.netmonitorpro&hl=en_CA&gl=US&showAllReviews=true
[4] 2021. *Open5GCore.* Retrieved August 5, 2022 from https://www.open5gcore.org
[5] 2022. *ABOUT FED4FIRE+.* Retrieved December 13, 2022 from https://www.fed4fire.eu/
[6] 2022. *AERPAW.* Retrieved December 5, 2022 from https://aerpaw.org/
[7] 2022. *AKOS TestNet - Open Data.* Retrieved December 1, 2022 from https://www.akostest.net/en/opendata
[8] 2022. *ARA.* Retrieved December 5, 2022 from https://arawireless.org/
[9] 2022. *Canadian Radio-television and Telecommunications Commission.* Retrieved August 4, 2022 from https://crtc.gc.ca/eng/home-accueil.htm
[10] 2022. *Cognitive Radio Network Testbed.* Retrieved December 13, 2022 from https://cornet.wireless.vt.edu/index.html
[11] 2022. *COSMOS.* Retrieved December 5, 2022 from https://cosmos-lab.org/
[12] 2022. *Federal Communications Commission.* Retrieved August 4, 2022 from https://www.fcc.gov
[13] 2022. *free5GC.* Retrieved August 5, 2022 from https://www.free5gc.org
[14] 2022. *FUTURE INTERNET TESTING FACILITY.* Retrieved December 13, 2022 from https://fit-equipex.fr
[15] 2022. *G-NetTrack.* Retrieved August 4, 2022 from https://gyokovsolutions.com/g-nettrack/
[16] 2022. *Innovation, Science and Economic Development Canada.* Retrieved August 4, 2022 from https://open.canada.ca/en/open-data
[17] 2022. *MobileInsight.* Retrieved August 4, 2022 from http://mobileinsight.net/
[18] 2022. *MONROE- Measuring Mobile Broadband Networks in Europe.* Retrieved December 19, 2022 from https://www.monroe-project.eu
[19] 2022. *NetMetr - Open Data.* Retrieved December 1, 2022 from https://www.netmetr.cz/en/open-data.html
[20] 2022. *OneLab, FUTURE INTERNET TESTBEDS.* Retrieved December 18, 2022 from https://onelab.eu
[21] 2022. *Open5GS.* Retrieved August 5, 2022 from https://open5gs.org
[22] 2022. *Platforms for Advanced Wireless Research.* Retrieved December 4, 2022 from https://advancedwireless.org/
[23] 2022. *POWDER.* Retrieved December 4, 2022 from https://powderwireless.net/
[24] 2022. *Puffer.* Retrieved August 5, 2022 from https://puffer.stanford.edu
[25] 2022. *RootMetrics.* Retrieved January 11, 2023 from https://www.rootmetrics.com/en-US/home
[26] 2022. *RTR - NetTest.* Retrieved November 23, 2022 from https://www.netztest.at/en/
[27] 2022. *SigCap.* Retrieved August 5, 2022 from https://appdistribution.firebase.google.com/pub/i/5b022e1d936d1211
[28] 2022. *Slices-RI.* Retrieved December 13, 2022 from https://www.slices-ri.eu/
[29] 2022. *Spectrum Management System.* Retrieved August 5, 2022 from http://sms-sgs.ic.gc.ca/eic/site/sms-sgs-prod.nsf/eng/home
[30] 2022. *SPECURE - Open Data.* Retrieved December 1, 2022 from https://www.meracinternetu.sk/en/opendata
[31] 2022. *tcpdump.* Retrieved August 5, 2022 from https://www.androidtcpdump.com/
[32] 2022. *Tutela, Crowdsourced data for the mobile industry.* Retrieved January 11, 2023 from https://www.tutela.com
[33] 2022. *What is BEREC?* Retrieved November 22, 2022 from https://www.berec.europa.eu/en/berec/what-is-berec
[34] 2022. *Wireless Testlab and OfficeLab.* Retrieved December 18, 2022 from https://doc.ilabt.imec.be/ilabt/wilab/
[35] 2023. *Orange Flux Vision.* Retrieved January 11, 2023 from https://www.orange-business.com/fr/produits/flux-vision
[36] 2023. *Predict.io.* Retrieved January 11, 2023 from https://www.predict.io
[37] 2023. *SFR Geostatistics.* Retrieved January 11, 2023 from https://www.sfrbusiness.fr/relation-client/sfr-geostatistics/

[38] 3GPP. 2021. *NG-RAN; Architecture description.* Technical specification (TS) 38.401. 3rd Generation Partnership Project (3GPP). https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3219 Version 16.7.0.

[39] 3GPP. 2022. *Management and orchestration; 5G performance measurements.* Technical specification (TS) 28.552. 3rd Generation Partnership Project (3GPP). https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3413 Version 17.1.0.

[40] 3GPP. 2022. *NR; Layer 2 measurements.* Technical specification (TS) 38.314. 3rd Generation Partnership Project (3GPP). https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3671 Version 17.1.0.

[41] 3GPP. 2022. *NR; Physical layer measurements.* Technical specification (TS) 38.215. 3rd Generation Partnership Project (3GPP). https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3217 Version 17.2.0.

[42] 3rd Generation Partnership Project. 2022. Technical Specification Group Services and System Aspects; Release 17 Description.

[43] Mamta Agiwal, Hyeyeon Kwon, Seungkeun Park, and Hu Jin. 2021. A Survey on 4G-5G Dual Connectivity: Road to 5G Implementation. *IEEE Access* 9 (2021), 16193–16210.

[44] Petri Ahokangas, Marja Matinmikko-Blue, Seppo Yrjölä, Veikko Seppänen, Heikki Hämmäinen, Risto Jurva, and Matti Latva-aho. 2019. Business Models for Local 5G Micro Operators. *IEEE Transactions on Cognitive Communications and Networking* 5, 3 (2019), 730–740.

[45] Özgü Alay, Andra Lutu, Rafael García, Miguel Peón-Quirós, Vincenzo Mancuso, Thomas Hirsch, Tobias Dely, Jonas Werme, Kristian Evensen, Audun Hansen, et al. 2022. MONROE: Measuring mobile broadband networks in Europe. In *Building the Future Internet through FIRE*. River Publishers, 155–187.

[46] ORAN Alliance. 2018. *O-RAN: Towards an open and smart RAN.* https://www.o-ran.org/resources

[47] Maximilian Arnold, Jakob Hoydis, and Stephan ten Brink. 2019. Novel massive MIMO channel sounding data applied to deep learning-based indoor positioning. In *SCC 2019; 12th International ITG Conference on Systems, Communications and Coding*. VDE, 1–6.

[48] Carlos Baena, Oswaldo Sebastián Peñaherrera-Pulla, Lourdes Camacho, Raquel Barco, and Sergio Fortes. 2022. E2E dataset of Video Streaming and Cloud Gaming services over 4G and 5G. https://doi.org/10.21227/k0w8-qz67

[49] Matheus HC Barboza, Ricardo de S Alencar, Julio C Chaves, Moacyr AHB Silva, Romulo D Orrico, and Alexandre G Evsukoff. 2021. Identifying human mobility patterns in the Rio de Janeiro metropolitan area using call detail records. *Transportation research record* 2675, 4 (2021), 213–221.

[50] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific data* 2, 1 (2015), 1–15.

[51] BEREC. 2014. *Annex of Monitoring quality of Internet access services in the context of net neutrality.* Report 14-117. Body of European Regulators for Electronic Communications (BEREC). https://www.berec.europa.eu/sites/default/files/files/document_register_store/2014/9/BoR%20%2814%29%20117_Annex%201_NN_QoS_Monitoring_Report_final.pdf

[52] BEREC. 2017. *Net neutrality measurement tool specification.* Report 17-179. Body of European Regulators for Electronic Communications (BEREC). https://www.berec.europa.eu/sites/default/files/files/document_register_store/2017/10/BoR_%2817%29_179_NN_measurement_tool_specification_-_Prep_for_publication_-_Clean.pdf

[53] Lorenzo Bertizzolo, Leonardo Bonati, Emrecan Demirors, Amani Al-Shawabka, Salvatore D'Oro, Francesco Restuccia, and Tommaso Melodia. 2020. Arena: A 64-antenna SDR-based ceiling grid testing platform for sub-6 GHz 5G-and-Beyond radio spectrum research. *Computer Networks* 181 (2020), 107436.

[54] Vincent D. Blondel, Adeline Decuyper, and Gautier Krings. 2015. A Survey of Results on Mobile Phone Datasets Analysis. *EPJ Data Science* 4, 10 (2015), 1–55.

[55] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ data science* 4, 1 (2015), 10.

[56] Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. 2012. Data for Development: The D4D Challenge on Mobile Phone Data. *arXiv preprint arXiv:1210.0137* (2012).

[57] Andrey Bogomolov, Bruno Lepri, Roberto Larcher, Fabrizio Antonelli, Fabio Pianesi, and Alex Pentland. 2016. Energy Consumption Prediction Using People Dynamics Derived from Cellular Network Data. *EPJ Data Science* 5, 13 (2016).

[58] Ayub Bokani, Mahbub Hassan, Salil S Kanhere, Jun Yao, and Garson Zhong. 2016. Comprehensive mobile bandwidth traces from vehicular networks. In *Proceedings of the 7th international conference on multimedia systems*. 1–6.

[59] Lorenzo Bracciale, Marco Bonola, Pierpaolo Loreti, Giuseppe Bianchi, Raul Amici, and Antonello Rabuffi. 2014. CRAWDAD dataset roma/taxi (v. 2014-07-17). Downloaded from https://crawdad.org/roma/taxi/20140717. https://doi.org/10.15783/C7QC7M

[60] David A. Burgess and Harvind S. Samra. 2008. *The Open BTS Project.* Technical Report. Kestrel Signal Processing.

[61] Julio Cesar Chaves. 2018. Ordinary mobility detected by call detail records along 2014. https://doi.org/10.7910/DVN/LAWIYW

[62] Soukaina Cherkaoui, Ines Keskes, Herve Rivano, and Razvan Stanica. 2016. LTE-A Random Access Channel Capacity Evaluation for M2M Communications. In *IFIP Wireless Days (WD)*. Toulouse, France.

[63] Marco Cominelli, Francesco Gringoli, and Renato Lo Cigno. 2022. AntiSense: Standard-compliant CSI obfuscation against unauthorized Wi-Fi sensing. *Computer Communications* 185 (2022), 92–103.

[64] Marco Cominelli, Francesco Gringoli, and Renato Lo Cigno. 2022. *CSI-based Device-free Localization and Obfuscation.* https://doi.org/10.5281/zenodo.5885636

[65] Federal Communications Commission. 2020. *Measuring Mobile Broadband.* Retrieved August 4, 2022 from https://www.fcc.gov/general/measuring-mobile-broadband-performance

[66] Federal Communications Commission. 2022. *Measuring Broadband America Mobile Data.* Retrieved August 4, 2022 from https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-broadband-america-mobile-data#data-dictionaries-and-information

[67] Quentin De Coninck, Matthieu Baerts, Benjamin Hesmans, and Olivier Bonaventure. 2016. CRAWDAD dataset uclouvain/mptcp_smartphone (v. 2016-03-04). Downloaded from https://crawdad.org/uclouvain/mptcp_smartphone/20160304. https://doi.org/10.15783/C7VG6H

[68] Sibren De Bast and Sofie Pollin. 2021. Ultra Dense Indoor MaMIMO CSI Dataset. https://doi.org/10.21227/nr6k-8r78

[69] Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D. Blondel. 2014. D4D-Senegal: The Second Mobile Phone Data for Development Challenge. *arXiv preprint arXiv:1407.4885* (2014).

[70] Shuo Deng, Ravi Netravali, Anirudh Sivaraman, and Hari Balakrishnan. 2014. WiFi, LTE, or both? Measuring multi-homed wireless Internet performance. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. 181–194.

[71] Rex W. Douglass, David A. Meyer, Megha Ram, David Rideout, and Dongjin Song. 2015. High Resolution Population Estimates from Telecommunications Data. *EPJ Data Science* 4, 4 (2015).

[72] Merim Dzaferagic, Nicola Marchetti, and Irene Macaluso. 2021. Minimizing the Signaling Overhead and Latency Based on Users' Mobility Patterns. *IEEE Systems Journal* 15, 1 (2021), 77–84.

[73] Nathan Eagle and Alex (Sandy) Pentland. 2005. CRAWDAD dataset mit/reality (v. 2005-07-01). Downloaded from https://crawdad.org/mit/reality/20050701. https://doi.org/10.15783/C71S31

[74] Marco Fiore, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Francoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica. 2020. Privacy in Trajectory Micro-Data Publishing: A Survey. *Transactions on Data Privacy* 13, 2 (2020), 91–149.

[75] Kaixuan Gao, Huiqiang Wang, and Hongwu Lv. 2021. CSI Dataset towards 5G NR High-Precision Positioning. https://doi.org/10.21227/jsat-pb50

[76] Kaixuan Gao, Huiqiang Wang, Hongwu Lv, and Wenxue Liu. 2022. Towards 5G NR high-precision indoor positioning via channel frequency response: A new paradigm and dataset generation method. *IEEE Journal on Selected Areas in Communications* (2022).

[77] Arthur Gassner, Claudiu Musat, Alexandru Rusu, and Andreas Burg. 2021. OpenCSI: An Open-Source Dataset for Indoor Localization Using CSI-Based Fingerprinting. *arXiv e-prints* (2021), arXiv–2104.

[78] Ismael Gomez-Miguelez, Andres Garcia-Saavedra, Paul D. Sutton, Pablo Serrano, Cristina Cano, and Doug J. Leith. 2016. srsLTE: An Open-source Platform for LTE Evolution and Experimentation. In *Proceedings of the ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization (WiNTECH)*. New York, NY, USA.

[79] Francesco Gringoli, Paul Patras, Carlos Donato, Pablo Serrano, and Yan Grunenberger. 2018. Performance Assessment of Open Software Platforms for 5G Prototyping. *IEEE Wireless Communications* 25, 5 (2018).

[80] Mohammad Hosseini, Yu Jiang, Ali Yekkehkhany, Richard R Berlin, and Lui Sha. 2017. A mobile geo-communication dataset for physiology-aware dash in rural ambulance transport. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. 158–163.

[81] Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Carlo Ratti, and Guy Pujolle. 2014. Estimating Human Trajectories and Hotspots through Mobile Phone Data. *Computer Networks* 64 (2014), 296–307.

[82] Hamza Awad Hamza Ibrahim, Omer Radhi Aqeel Al Zuobi, Maran A. Al Namari, Gaafer Mohamed Ali, and Ali Ahmed Alfaki Abdalla. 2016. Internet Traffic Classification using Machine Learning Approach: Datasets Validation Issues. In *Conference on Basic Sciences and Engineering Studies (SGCAC)*. Khartoum, Sudan.

[83] Agbotiname Lucky Imoize, Kehinde Orolu, and Aderemi Aaron-Anthony Atayero. 2020. Analysis of key performance indicators of a 4G LTE network based on experimental data obtained from a densely populated smart city. *Data in brief* 29 (2020), 105304.

[84] Biju Issac, Wendy Japutra Jap, and Jofry Hadi Sutanto. 2009. Improved Bayesian Anti-Spam Filter Implementation and Analysis on Independent Spam Corpuses. In *International Conference on Computer Engineering and Technology (ICCET)*. Singapore.

[85] Telecom Italia. 2015. Telecommunications - SMS, Call, Internet - TN. https://doi.org/10.7910/DVN/QLCABU

[86] Anders Johansson, Magnus Esbjörnsson, Per Nordqvist, Stig Wiinberg, Roger Andersson, Bodil Ivarsson, Bengt Eksund, and Sebastian Möller. 2019. Dataset on multichannel connectivity and video transmission carried on commercial 3G/4G networks in southern Sweden. *Data in brief* 25 (2019), 104192.

[87] Bryan Klimt and Yiming Yang. 2004. Introducing the Enron Corpus. In *International Conference on Email and Anti-Spam (CEAS)*. Mountain View, CA, USA.

[88] Erik G. Larsson, Ove Edfors, Fredrik Tufvesson, and Thomas L. Marzetta. 2014. Massive MIMO for Next Generation Wireless Systems. *IEEE Communications Magazine* 52, 2 (2014), 186–195.

[89] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. 2012. *The mobile data challenge: Big data for mobile computing research.* Technical Report.

[90] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (2007), 2–42.

[91] Yuanjie Li, Chunyi Peng, Zengwen Yuan, Jiayao Li, Haotian Deng, and Tao Wang. 2016. Mobileinsight: Extracting and analyzing cellular network information on smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 202–215.

[92] Shao-Yu Lien, Shin-Lin Shieh, Yenming Huang, Borching Su, Yung-Lin Hsu, and Hung-Yu Wei. 2017. 5G New Radio: Waveform, Frame Structure, Multiple Access, and Initial Access. *IEEE Communications Magazine* 55, 6 (2017), 64–71.

[93] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. 2023. The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography. *arXiv preprint arXiv:2305.06933* (2023).

[94] Julian McAuley and Jure Leskovec. 2012. Learning to Discover Social Circles in Ego Networks. In *International Conference on Neural Information Processing Systems (NIPS)*. Lake Tahoe, NE, USA.

[95] Britta Meixner, Jan Willem Kleinrouweler, and Pablo Cesar. 2018. *4G/LTE Channel Quality Reference Signal Traces Data Set.* https://doi.org/10.5281/zenodo.1220256

[96] Tommaso Melodia, Stefano Basagni, Kaushik R Chowdhury, Abhimanyu Gosain, Michele Polese, Pedram Johari, and Leonardo Bonati. 2022. Tutorial: Colosseum, the World's Largest Wireless Network Emulator. (2022).

[97] Germán Martín Mendoza-Silva, Miguel Matey-Sanz, Joaquín Torres-Sospedra, and Joaquín Huerta. 2019. BLE RSS measurements dataset for research on accurate indoor positioning. *Data* 4, 1 (2019), 12.

[98] Abdelrahim Mohamed, Oluwakayode Onireti, Seyed Amir Hoseinitabatabaei, Muhammad Imran, Ali Imran, and Rahim Tafazolli. 2015. Mobility Prediction for Handover Management in Cellular Networks with Control/Data Separation. In *IEEE International Conference on Communications (ICC)*. London, UK.

[99] Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. 2015. Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials* 18, 1 (2015), 124–161.

[100] Arvind Narayanan, Eman Ramadan, Jason Carpenter, Qingxu Liu, Yu Liu, Feng Qian, and Zhi-Li Zhang. 2020. A First Look at Commercial 5G Performance on Smartphones. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 894–905. https://doi.org/10.1145/3366423.3380169

[101] Arvind Narayanan, Eman Ramadan, Rishabh Mehta, Xinyue Hu, Qingxu Liu, Rostand A. K. Fezeu, Udhaya Kumar Dayalan, Saurabh Verma, Peiqi Ji, Tao Li, Feng Qian, and Zhi-Li Zhang. 2020. Lumos5G: Mapping and Predicting Commercial MmWave 5G Throughput. In *Proceedings of the ACM Internet Measurement Conference* (Virtual Event, USA) *(IMC '20)*. Association for Computing Machinery, New York, NY, USA, 176–193. https://doi.org/10.1145/3419394.3423629

[102] Arvind Narayanan, Muhammad Iqbal Rochman, Ahmad Hassan, Bariq S. Firmansyah, Vanlin Sathya, Monisha Ghosh, Feng Qian, and Zhi-Li Zhang. 2022. A Comparative Measurement Study of Commercial 5G mmWave Deployments. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*. 800–809. https://doi.org/10.1109/INFOCOM48880.2022.9796693

[103] Navid Nikaein, Mahesh K. Marina, Saravana Manickam, Alex Dawson, Raymond Knopp, and Christian Bonnet. 2014. OpenAirInterface: A Flexible Platform for 5G Research. *ACM SIGCOMM Computer Communication Review* 44, 5 (2014).

[104] Mengguan Pan, Peng Liu, Shengheng Liu, Wangdong Qi, Yongming Huang, Xiaohu You, Xinghua Jia, and Xiaodong Li. 2022. Efficient joint DOA and TOA estimation for indoor positioning with 5G picocell base stations. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–19.

[105] Mengguan Pan, Shengheng Liu, Peng Liu, Wangdong Qi, Yongming Huang, Wang Zheng, Qihui Wu, and Markus Gardill. 2022. 5G CFR/CSI dataset for wireless channel parameter estimation, array calibration, and indoor positioning. https://doi.org/10.21227/k2f0-k132

[106] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. 2009. CRAWDAD dataset epfl/mobility (v. 2009-02-24). Downloaded from https://crawdad.org/epfl/mobility/20090224. https://doi.org/10.15783/C7J010

[107] Darijo Raca, Dylan Leahy, Cormac J Sreenan, and Jason J Quinlan. 2020. Beyond throughput, the next generation: a 5G dataset with channel and context metrics. In *Proceedings of the 11th ACM Multimedia Systems Conference*. 303–308.

[108] Darijo Raca, Jason J Quinlan, Ahmed H Zahran, and Cormac J Sreenan. 2018. Beyond throughput: a 4G LTE dataset with channel and context metrics. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 460–465.

[109] Dipl-Ing Vaclav Raida. 2021. *Data-Driven Estimation of Spatiotemporal Performance Maps in Cellular Networks.* Ph. D. Dissertation. University of Bologna Vienna.

[110] Vaclav Raida, Martin Lerch, Philipp Svoboda, and Markus Rupp. 2018. Deriving cell load from RSRQ measurements. In *2018 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 1–6.

[111] Vaclav Raida, Philipp Svoboda, Martin Koglbauer, and Markus Rupp. 2020. On the Stability of RSRP and Variability of Other KPIs in LTE Downlink-An Open Dataset. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.

[112] Vaclav Raida, Philipp Svoboda, Martin Lerch, and Markus Rupp. 2019. Repeatability for spatiotemporal throughput measurements in LTE. In *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*. IEEE, 1–5.

[113] Vaclav Raida, Philipp Svoboda, and Markus Rupp. 2018. Lightweight detection of tariff limits in cellular mobile networks. In *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 1–7.

[114] Vaclav Raida, Philipp Svoboda, and Markus Rupp. 2020. 2020-GLOBECOM-LTE-DL-static-rural-and-urban-outdoor-dataset. https://doi.org/10.23728/B2SHARE.F5DF0B362B7E473C825776EC166E658F

[115] Vaclav Raida, Philipp Svoboda, and Markus Rupp. 2020. 2020-GLOBECOM-LTE-DL-static-urban-indoor-dataset. https://doi.org/10.23728/B2SHARE.2367746915E34D1CA4BDA84193F8056B

[116] Vaclav Raida, Philipp Svoboda, and Markus Rupp. 2020. 2020-VTC-Fall-RP-DTW-dataset. https://doi.org/10.23728/B2SHARE.B7F93ADA6C6F46B296CD3C39665982C4

[117] Vaclav Raida, Philipp Svoboda, and Markus Rupp. 2020. Modified dynamic time warping with a reference path for alignment of repeated drive-tests. In *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. IEEE, 1–6.

[118] Vaclav Raida, Philipp Svoboda, and Markus Rupp. 2020. Real World Performance of LTE Downlink in a Static Dense Urban Scenario-An Open Dataset. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.

[119] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. 2012. Are Call Detail Records Biased for Sampling Human Mobility? *ACM SIGMOBILE Mobile Computing and Communications Review* 16, 3 (2012), 33–44.

[120] Dipankar Raychaudhuri, Ivan Seskar, Gil Zussman, Thanasis Korakis, Dan Kilper, Tingjun Chen, Jakub Kolodziejski, Michael Sherman, Zoran Kostic, Xiaoxiong Gu, et al. 2020. Challenge: COSMOS: A city-scale programmable testbed for experimentation with advanced wireless. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[121] Matei Ripeanu, Adriana Iamnitchi, and Ian T. Foster. 2002. Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. *IEEE Internet Computing Journal* 6, 1 (2002), 50–57.

[122] Muhammad Iqbal Rochman, Vanlin Sathya, Norlen Nunez, Damian Fernandez, Monisha Ghosh, Ahmed S Ibrahim, and William Payne. 2021. A Comparison Study of Cellular Deployments in Chicago and Miami Using Apps on Smartphones. *arXiv preprint arXiv:2108.00453* (2021).

[123] Benedek Rozemberczki and Rik Sarkar. 2020. Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models. In *ACM International Conference on Information & Knowledge Management (CIKM)*. Virtual Event.

[124] J Xavier Salvat, Jose A Ayala-Romero, Lanfranco Zanzi, Andres Garcia-Saavedra, and Xavier Costa-Perez. 2023. Open Radio Access Networks (O-RAN) Experimentation Platform: Design and Datasets. *IEEE Communications Magazine* (2023).

[125] J. Xavier Salvat Lozano, Jose A. Ayala-Romero, Lanfranco Zanzi, Andres Garcia-Saavedra, and Xavier Costa-Perez. 2022. O-RAN experimental evaluation datasets. https://doi.org/10.21227/64s5-q431

[126] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T. Campbell. 2011. NextPlace: A Spatio-Temporal Prediction Framework for Pervasive Systems. In *International Conference on Pervasive Computing*. San Francisco, CA, USA.

[127] Matthias Schulz, Daniel Wegemer, and Matthias Hollick. 2017. *Nexmon: The C-based Firmware Patching Framework*. https://github.com/seemoo-lab/nexmon

[128] Wee-Seng Soh and Hyong S. Kim. 2003. QoS Provisioning in Cellular Networks based on Mobility Prediction Techniques. *IEEE Communications Magazine* 41, 1 (2003), 86–92.

[129] Simon Tewes, Markus Heinrichs, Kevin Weinberger, Rainer Kronberger, and Aydin Sezgin. 2023. A comprehensive dataset of RIS-based channel measurements in the 5GHz band. https://doi.org/10.21227/zzx0-tp88

[130] Raza Ul Mustafa, Christian Esteve Rothenberg, and Chadi Barakat. 2022. YouTube goes 5G: Benchmarking YouTube in 4G vs 5G. https://doi.org/10.21227/h00h-ew92

[131] J. van der Hooft, S. Petrangeli, T. Wauters, R. Huysegems, P. R. Alface, T. Bostoen, and F. De Turck. 2016. HTTP/2-Based Adaptive Streaming of HEVC Video Over 4G/LTE Networks. *IEEE Communications Letters* 20, 11 (2016), 2177–2180.

[132] Florian Wamser, Michael Seufert, Pedro Casas, Ralf Irmer, Phuoc Tran-Gia, and Raimund Schatz. 2015. YoMoApp: A tool for analyzing QoE of YouTube HTTP adaptive streaming in mobile networks. In *2015 European Conference on Networks and Communications (EuCNC)*. IEEE, 239–243.

[133] Fang Wang, Yong Li, Zhaocheng Wang, and Zhixing yang. 2016. Social-Community-Aware Resource Allocation for D2D Communications Underlaying Cellular Networks. *IEEE Transactions on Vehicular Technology* 65, 5 (2016), 3628–3640.

[134] Sarah Wassermann, Pedro Casas, Michael Seufert, and Florian Wamser. 2019. On the analysis of YouTube QoE in cellular networks through in-smartphone measurements. In *2019 12th IFIP Wireless and Mobile Networking Conference (WMNC)*. IEEE, 71–78.

[135] Francis Y Yan, Hudson Ayers, Chenzhi Zhu, Sadjad Fouladi, James Hong, Keyi Zhang, Philip Levis, and Keith Winstein. 2020. Learning in situ: a randomized experiment in video streaming. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. 495–511.

[136] Paolo Zanini, Haipeng Shen, and Young Truong. 2016. Understanding Resident Mobility in Milan through Independent Component Analysis of Telecom Italia Mobile Usage Data. *The Annals of Applied Statistics* 10, 2 (2016).