IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2019, accepted October 6, 2019, date of publication October 23, 2019, date of current version November 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949051

Crowdsensed Performance Benchmarking of Mobile Networks

VACLAV RAIDA, (Student Member, IEEE), PHILIPP SVOBODA, (Senior Member, IEEE), MARTIN LERCH, AND MARKUS RUPP, (Fellow, IEEE)
Institute of Telecommunications, Technische Universität Wien, 1040 Vienna, Austria

Corresponding author: Vaclav Raida (vaclav.raida@gmail.com)

This work was supported in part by the ITC, in part by TU Wien, and in part by A1 Telekom Austria AG, and in part by the Austrian FFG, Bridge Project, under Grant 871261.

ABSTRACT In recent years, the boom of mobile-network-measurement apps has stimulated the growth of publicly available datasets, which contain up to billions of measurements conducted by anonymous end-users. Although the crowdsensing has proven its value in different areas, in the context of performance monitoring in cellular mobile networks, convincing applications are scarce. A portion of already published research focuses on modeling and predicting achievable throughput as a function of observed signal indicators. Due to the system complexity and a large number of possible predictors, the performance of these models is low since the throughput varies greatly even at a constant signal power level. In this paper, we introduce a new method for evaluating an empirical network-centric upper bound on the throughput-performance of different cellular mobile networks. Based on simulations and reference measurements, we propose a model function that characterizes throughput as a function of signal power. The critical point that increases the quality of fit when processing the crowdsourced measurements is the removal of the system-inherent high noise level by finding a method to exclude biased samples systematically. We apply our method to crowdsourced measurements provided by several national regulatory bodies, and benchmark the performance of LTE networks of different mobile operators in different countries.

INDEX TERMS Crowdsensing, performance, mobile, cellular, network, crowdsourcing, measurements, operator, benchmarking, iteratively reweighted least squares, LTE, 5G, throughput, signal strength, RSRP.

I. INTRODUCTION

Crowdsourcing is all around us: Data shared by car drivers help to monitor the road traffic, detect jams and accidents [1]. Users from the entire Earth contribute to online encyclopedias [2], collect geographic information [3], and label photos or scanned texts [4]. Smartphone applications can gather seismic data [5]. Even a questionnaire or an opinion poll are examples of crowdsourcing.

In the context of cellular mobile networks, the term ''mobile crowdsensing'' (MCS) denotes the measurements that end-users perform in a distributed manner on their user equipment (UE). For the researchers, the main advantage is the access to millions of measurement results without needing to perform their own tests. The disadvantages are the limited knowledge of circumstances under which the tests were undertaken (indoor/outdoor, limited/unlimited tariff, cell load, user's cross-traffic, base station handover) and the inability to control when and where they will be executed. Researchers attempt to overcome crowdsourcing's disadvantages and exploit its data to extract meaningful information about cellular mobile networks and their users.

Recently, multiple MCS apps and platforms have evolved, measuring users' throughput, latency, signal power, and other parameters. State of the art in the processing of these datasets is unsatisfactory as the majority of the platforms built their business model on publishing reports and summaries or selling aggregated data. The actual measurement data and the analysis approach stay hidden from the public.

Efforts of national regulatory bodies to publicly offer their own tools and to share the raw measurement data enable researchers to access common datasets and thus to achieve reproducibility and comparability of their results.

The associate editor coordinating the review of this manuscript and approving it for publication was Young Jin Chun.

Nevertheless, the sensing process of the measurement tools is not optimized in any sense (Section I-A.1), unknown factors that cannot be neglected are not reported (Section I-A.2), and machine learning techniques fail in reaching overambitious goals due to underrepresented categories (Section I-A.3).

In this paper, we show that by suitable data aggregation, we are able to overcome the outlined limitations and extract a metric that allows operator benchmarking. Based on simulations and reference measurements, we develop a method that utilizes crowdsourced measurements for benchmarking peak throughput performance of mobile networks. Our network-centric benchmark provides the view of the network as a whole and reflects differences due to planning strategy and technology limits.

A. CHALLENGES

1) MEASUREMENT TOOLS ARE PURELY HEURISTIC

In the rich offer of various MCS performance monitoring tools, there is a lack of a measurement and evaluation methodology. Each tool employs its own heuristic test duration and number of parallel flows/connections. It is thus unclear if the results of different tools produce comparable results and how to achieve specific goals such as operator benchmarking (benchmarking = fair comparison).

Despite efforts to reduce the data volume required per measurement [6], [7], the current tests take several seconds, consuming tens of megabytes of users' data. The long test duration limits the spatial resolution of measurements conducted by mobile users, making it often impossible to link the measured performance to a specific location.

2) THE CROWD IS BIASED

The high test price decreases users' motivation to collect more samples. Moreover, customers deliberately choose lower tariffs to save their money. Users may more likely run a test if they are unsatisfied with their connection. Cell load varies throughout the day, allowing different maximum throughputs at different times (nonstationary process).

Although there exist measurement methods that estimate cell load [8] or avoid tariff limits [9], such solutions are not implemented in the current tools. Further unknown factors, such as indoor/outdoor environment or presence of user's own cross-traffic may remain undetected.

As a way out of this, published benchmarks usually restrict themselves only to the country-wide comparison of means or medians [10]–[12] of the key performance indicators (KPIs), e.g., LTE throughput or latency.

3) MY BIG DATA WILL ALWAYS BE TOO SMALL

At first glance, in order to provide more insight, it seems reasonable to employ all available features for the training of a model that predicts a quantity of interest and hope that our, e.g., deep neural network learns to compensate all undesired effects. The following example—throughput prediction in LTE—enlightens why such an approach fails.

TABLE 1. A non-exhaustive list of throughput-relevant LTE features.

Bins	Predictor	Combinations	
24	Hour of day	$1.9 \cdot 10^9$	$30 \cdot 10^{12}$
2	Workday / weekend		
98	RSRP (1 dB bins)		
35	RSRQ (1 dB bins)		
128	RS-SINR (0.5 dB bins)		
3	Mobile network operator (MNO)		
30	Device model		
4	Indoor / outdoor / in car / in train		
5	3 frequency bands + $2 \times$ CA$^{(a)}$		
800	Location: Austria, 100 km^2 bins		

$^{(a)}$Three frequency bands with two carrier aggregation (CA) modes. The first band is an access layer. The two CA possibilities are: "The first & the second band," or "the first & the third band."

Table 1 contains selected LTE-throughput relevant features and the number of possible values they can take.[1] Considering the first seven features—as Cainey et al. did in [13]—yields nearly two billions unique combinations, which by far exceeds the 130 296 samples that the authors had at disposal and explains why the results are unconvincing (see Section I-B). With most of the bins empty, even imbalanced data learning techniques [16] cannot help.

Many features that cannot be neglected are missing in the list—e.g., user's velocity or virtual operators. We already mentioned unknown factors—tariff limits or the presence of cross-traffic. The final nail in the coffin is spatial binning: For example, splitting Austria into 10 km $\times$ 10 km squares (too coarse already since moving just a few meters can cause significant performance differences) adds a new dimension with more than 800 independent bins.

Even the most massive crowdsourced data set that is known to us—Tutela [17], gathering billions of LTE measurements every day worldwide—is not large enough to fill all the different multidimensional bins (30 trillion in our example in Table 1). In Section I-B, we give more details about Tutela's usage of the data.

Does it even make sense to compare, e.g., UE models, given that features corresponding to their measurements barely coincide? We do not think so, and we suspect that the various model coefficients derived in [13] for different models are merely a random outcome.

4) SO WHAT IS POSSIBLE...

When the brute-force learning approach fails, we need to fall back to domain knowledge, reducing the number of possible unique multidimensional bins: There are dependencies between RSRP, RSRQ, and cell load [8], [18]. The cell load could be represented as a cyclo-stationary process depending on day hour and weekday. Tariff-limits can be detected [19].

[1]The relevant indicators are addressed in Section III-A. The first seven features correspond to predictors employed in [13]. The number of bins for each signal metric (RSRP, RSRQ, RS-SINR) we took from [14]. The other counts (operators, UE models, bands) we choose small enough (smaller than we found in an actual dataset [15]) to get a conservative lower bound.

In a restricted time range, RSRP could be modeled as time-invariant based on measurements in [20].

We try to find a trade-off between coarse per operator per country benchmarking (Section I-A.2) and between explaining everything (Section I-A.3). Only reasonably posed questions can be answered with crowdsourced measurements. For comparing UE models, lab tests may be a better choice.

B. RELATED WORK

Ganti *et al.* [21] coined the term mobile crowdsensing (MCS), offering several examples of its applications. Guo *et al.* [22] further extended the concept to mobile crowdsensing and computing (MCSC), in which the mobile nodes are used not only for data collections but also for their processing.

The major difficulty in crowdsensing is that we often lack enough measurements in desired locations and time points. To tackle this problem, many researchers [23]–[29] introduce various cost functions and optimize reward and recruitment strategies that motivate users to perform tests in locations of interest.

Tutela [17] chose a different approach - motivating app developers instead of end-users. Any app developer can gain money by including Tutela's software development kit (SDK) into his/her application. The SDK then collects anonymized network data (signal power, throughput, latency, operator, radio channel) from the users.

Kousias *et al.* [30] developed a classifier for identifying different MNOs and applied it to the RTR open data. Khatouni *et al.* [31] used a MONROE platform to perform controlled measurements in 3G and 4G cellular mobile networks and found only a weak correlation between received signal strength indicator and goodput.

Jungermann [32], in his online article, benchmarked Nordic operators based on OpenSignal's crowdsourced measurements of LTE downlink (DL) throughput and LTE coverage (in % of time). Linder *et al.* [33] reconstructed performance maps from Bredbandskollen's crowdsourced measurements, and compared the LTE DL throughputs of different Swedish MNOs in different location bins using Welch's t-test.

Cainey *et al.* [13] modeled the logarithm of LTE DL throughput as a linear function of combined several predictors. They claimed that reference signal received power (RSRP) is an inadequate measure of the signal. Uludağ and Korçak [34] modeled energy consumption and LTE DL throughput using ordinary least squares, robust regression methods, and other machine learning techniques (decision trees, random forests). Their model in all cases yielded a coefficient of determination (R^2) worse than 28%. They explained this as a result of unpredictable events, e.g., download activities of nearby UEs.

From what Tutela publicly revealed [35], it seems that the current root-cause analysis exploiting their giant dataset employs (after nailing down the problem of efficiently storing and quickly querying billions of database entries) mainly manual exploration of measurements—spatially binning the data into hexagonal cells, plotting histograms of different quantities in selected areas, filtering by operators and frequency bands, looking where and when a concentration of specific parameters deviates from what is usual.

C. OUR CONTRIBUTION

We propose—based on simulations and reference measurements—a method for benchmarking different mobile network operators in different countries with crowdsourced measurements provided by several national regulatory bodies. As discussed in Section I-A.4, it is essential to introduce a certain level of data aggregation to avoid empty bins and to bridge the gap between too coarse approach and between the trial to explain everything.

From the crowdsourced measurements, we extract a signal-power-dependent upper bound of throughput (downlink and uplink) for a given operator in a given area. Such a bound gives the throughput of a single user in an unloaded cell (with no tariff limits, at minimum interference during off-peak hours) and can be seen as a network-centric metric: We do not compare individual users or devices, but we characterize the maximum throughput that an operator is able to offer with its technology in selected area.

With enough data at hand, our approach can compare individual cells and identify performance differences caused by channel bandwidth (BW), time division duplex/frequency division duplex, open/closed loop spatial multiplexing, and the presence of carrier aggregation or different MIMO/SIMO/MISO[2] orders.

With an LTE link-level simulator, we simulate different channel models, MIMO orders, and spatial multiplexing strategies. We conclude that the throughput versus signal-to-noise-ratio (SNR) in a one-user scenario can be approximated by a logistic function. We verify our conclusion by measurements in an empty reference cell.

Deriving the same characteristics from the crowdsourced tests can be challenging due to the presence of throughput-degraded measurements. The key point is to consider only nondegraded tests, which serve as a basis for fitting our model function that removes the remaining outliers caused by, e.g., automatized testing at a constant signal power level.

Finally, we utilize crowdsourced measurements in cellular mobile networks (with a focus on LTE) to characterize and benchmark throughput performance of different MNOs. The size of the publicly available datasets is not sufficient to inspect smaller areas. Therefore we perform operator benchmarking only on the country-level. Higher resolution can be achieved with, e.g., Tutela's dataset.

Despite the conclusion of Cainey *et al.* [13] that RSRP is an inadequate measure of signal quality, we reached a coefficient of determination $R^2 > 90\%$, outperforming the model introduced by Uludağ and Korçak [34] ($R^2 < 28\%$). Previously published approaches tried to explain all the data

[2]S = single, M = multiple, I = input, O = output.

employing linear models. Our method excludes throughput-degraded tests and adopts a nonlinear model.

Although we focus on LTE and RSRP, the method can be used for other technologies and their signal strength indicators, possibly considering different model functions.

D. PAPER OUTLINE

Section II briefly introduces iteratively reweighted least squares (IRLS) [36]. We present our simulations and reference measurements. We then compare the results with logistic functions, which we later apply to the crowdsourced measurements.

Section III explains the concept and intuition behind our method for benchmarking throughput performance in cellular mobile networks based on crowdsourced tests. We clarify our motivation for modeling throughput as a function of signal strength and for excluding throughput-degraded tests.

Section IV gives more details on the crowdsourcing smartphone apps and introduces the datasets we used. We propose specific implementation of the concept: We consider p-percentile of throughput in every signal-strength bin in order to avoid throughput degradation, then we employ IRLS to fit the model function, and thus remove remaining outliers.

Section V evaluates our method and benchmarks different MNOs. We discuss a graphical representation, introduce a scalar performance indicator that allows for time-series inspection, and we analyze the obtained fit quality depending on the number of samples and time interval duration. Regulatory bodies often benchmark operators based on the throughput median; thus, we compare our scalar performance indicator with medians. Finally, we summarize numerical results.

II. MODELING THROUGHPUT VS SIGNAL STRENGTH

In this section, we propose a logistic function as a model for the throughput of a single user in an empty cell. We briefly summarize the IRLS algorithm to clarify how we fitted the model function to our data. To validate our model choice, we simulate and measure the throughput in a single-user-scenario.

A. MODEL FUNCTION

The following are the minimum requirements for a model characterizing the throughput as a function of signal strength (or SNR). The curve should be:

- Continuous—a small change in signal strength should cause a small change in throughput.
- Monotonically increasing—with better signal strength we expect higher throughput.
- Bounded—throughput cannot be smaller than zero and larger than technology allows.

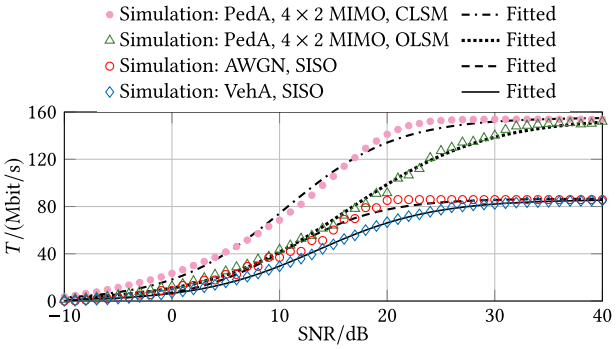

FIGURE 1. DL throughput vs SNR in different single-user scenarios simulated by Vienna LTE link level simulator. Black lines correspond to logistic functions fitted to the simulated data.

A straightforward choice is the logistic function:[3]

$$f(x,\boldsymbol{\theta}) = \frac{\alpha}{1+\mathrm{e}^{-\beta(x-x_0)}}, \quad \boldsymbol{\theta} = (x_0,\alpha,\beta)^{\mathsf{T}}. \tag{1}$$

Fitted logistic functions are depicted in Fig. 1 and 3.

1) OTHER MODELS

A different steepness near the lower and upper asymptote can be achieved by the generalized logistic function:

$$f(x,\boldsymbol{\theta}) = A + \frac{K-A}{\left(C+Q\mathrm{e}^{-Bx}\right)^{1/\nu}}. \tag{2}$$

Another possibility is the Weibull cumulative distribution function $F(x) = 1-\exp\left\{-(\lambda x)^k\right\}$, just scaled and shifted:

$$f(x,\boldsymbol{\theta}) = L\left(1-\mathrm{e}^{-(\lambda(x-x_0))^k}\right). \tag{3}$$

In contrast to the symmetric function (1), both equations (2) and (3) allow for higher flexibility. However, as they require more parameters to be optimized, the numerical fitting algorithm did not converge in many cases. Therefore, we use only (1) for the automatic evaluation.

B. ITERATIVE FITTING

Compared to the simulations and reference measurements, in the case of crowdsourced data, we need to deal with more outliers. To preserve the reasonable samples and to down weight the outliers, we apply the IRLS [36]:

$$\boldsymbol{\theta}^{(j+1)} = \arg\min_{\boldsymbol{\theta}} \sum_l w_l\left(\boldsymbol{\theta}^{(j)}\right) \underbrace{\left(y_l - f(x_l,\boldsymbol{\theta})\right)^2}_{r_l^2(\boldsymbol{\theta})}, \tag{4}$$

with "Tukey's biweight" weight function [37]

$$w_l(\boldsymbol{\theta}) = \begin{cases} \left(1-u_l^2(\boldsymbol{\theta})\right)^2, & |u_l(\boldsymbol{\theta})| \le 1, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where $u_l(\boldsymbol{\theta}) = r_l(\boldsymbol{\theta})/\left(c\cdot s(\boldsymbol{\theta})\sqrt{1-h_l}\right)$ are adjusted residuals with tuning constant $c = 4.685$, leverages h_l and estimate of the standard deviation of the

[3]The x denotes the independent variable. The vector $\boldsymbol{\theta}$ summarizes all parameters of the model function that we are going to fit.

error $s(\boldsymbol{\theta}) = \mathrm{MAD}(\boldsymbol{\theta})/0.6745$.[4] The first iteration is initialized with $\boldsymbol{\theta}^{(0)}$, which is based on non-weighted LS: (4) with unit weights. The algorithm repeats (4) and (5) until the convergence criterion $\left\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(j-1)}\right\| < \varepsilon$ is met or until the maximum number of iterations is reached. We then obtain the final estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(J)}$. For more details see MATLAB's implementation [38], which is based on Dumouchel and O'Brien [39]. The optimization problem (4)—in general nonlinear—is solved numerically using a trust region reflective algorithm [40], [41].

The IRLS algorithm assigns higher weights to reasonable samples that are closer to the model function and lower weights to outliers. Outliers with residuals larger than a certain threshold obtain zero weights. After calculating new weights, the fit is repeated.

We measure the goodness of the fit by the coefficient of determination $R^2 \in [0, 1]$ (often expressed as percentage):

$$R^2 = 1 - \frac{\sum_l w_l \, (y_l - f_l)^2}{\sum_l w_l \, (y_l - \bar{y})^2}, \tag{6}$$

where $\bar{y} = \left(\sum_l w_l y_l\right) / \left(\sum_l w_l\right)$ is the weighted average and $:= w_l := w_l\left(\boldsymbol{\theta}^{(J-1)}\right)$ are the weights used in the last J-th iteration of (4). The term R^2 measures how well the model explains the variations of the fitted data: 1 = the model perfectly explains all the data, 0 = the model performs same as the weighted mean $\bar{y}$.

The advantage is that R^2 is scale invariant, i.e., multiplying the y_l values by a constant does not impact R^2 (this is not the case of, e.g., MSE), which is helpful since different MNOs or different technologies (LTE/HSPA+) can reach a different maximum throughput. For more details see Kvalseth [42] and Willet and Singer [43].

C. MODEL VERIFICATION: SIMULATIONS

To simulate the throughput T of a single user in different scenarios, we employ the Vienna LTE link level simulator [44], [45]. We simulate different channels (AWGN, PedA, VehA—see simulator documentation); different MIMO orders (SISO, 2 × 1 MISO, 1 × 2 SIMO, 2 × 2 MIMO and 4 × 2 MIMO); and open-loop spatial multiplexing (OLSM) vs closed-loop spatial multiplexing (CLSM). Simulations are carried out for LTE 2600 MHz (band 7), DL, 20 MHz channel BW (100 resource blocks), 100 frames, normal cyclic prefix, 15 kHz subcarrier spacing, and perfect channel estimation.

Four examples, together with fitted logistic functions, are illustrated in Fig. 1. In this figure, we show the best obtained fit (VehA, SISO) and the worst fit (PedA, 4 × 2 MIMO, CLSM). The logistic function is symmetric, and thus offers lower flexibility than, e.g., generalized logistic function in (2). Nevertheless, even in the worst observed case, we reach $R^2 > 99\%$ and root-mean-square error RMSE < 3.18 Mbit/s.

Note: In the case of AWGN SISO we can notice individual steps due to different modulation and coding schemes (see Fig. 8.4 in Rupp *et al.* [44]). It is not feasible to model and fit a function that would reproduce these steps; the logistic function smooths them out.

[4] $\mathrm{MAD}(\boldsymbol{\theta}) = \mathrm{med}\{|r_l(\boldsymbol{\theta}) - \mathrm{med}\{r_l(\boldsymbol{\theta})\}|\}$ is the median of the absolute deviations of the residuals. The value of c gives coefficient estimates that are approximately 95% as statistically efficient as the ordinary least squares. The constant 0.6745 makes the standard deviation estimate unbiased for the normal distribution. The calculation of leverages is based on the QR decomposition of Jacobian matrix.

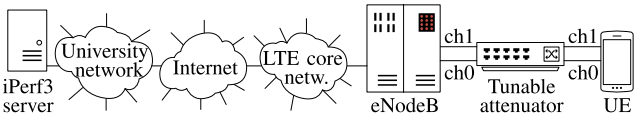

FIGURE 2. Diagram of our measurement setup. Although we have full control of the cell itself, the traffic still travels through live LTE core network and the Internet up to iPef3 measurement server at TU Wien, Institute of Telecommunications (Vienna, Austria).

D. MODEL VERIFICATION: CONTROLLED MEASUREMENTS

As a sanity check and to complement our simulations, we perform reference measurements in an LTE test-cell of one of the Austrian MNOs. In our simulations, SNR is well defined (see the simulator documentation). Although "reference signal-signal to noise and interference ratio" (RS-SINR) is specified in 3GPP's physical layer measurements [46],[5] it is not included in the crowdsourced open data. Therefore, we represent the measured throughput as a function of RSRP. We address RSRP in more detail in Section III-A.

The two UE antenna ports are connected with cables over an attenuator directly to the eNodeB's channel 0 (ch0) and channel 1 (ch1),[6] see Fig. 2. Thus, the UE is the only active user in the cell. By varying the attenuation levels we obtain DL and UL (uplink) throughputs for different RSRP levels. Noise and interference power P_Z is approximately equal to −96 dBm.

Network layer throughput and RSRP are monitored by the NEMO software [47], and the traffic is generated by iPerf3 [48]. We use the following hardware: UE – Samsung Galaxy Note 4 [49], LTE Cat. 4; eNodeB – Nokia, LTE 800 MHz, band 20, channel BW 20 MHz.

Results are shown in Fig. 3. Also in this case, the logistic function proves to be a good model. The step near −108.51 dBm, 113.19 Mbit/s in DL is due to change of modulation and coding scheme as discussed in the Section II-C.

III. BENCHMARKING METHOD: CONCEPT

A fair benchmarking of MNOs is possible only if the measurements are conducted under comparable conditions. Drive tests measure the throughputs of different MNOs in the same

[5] RS-SINR was introduced in December 2015, Version 13.0.0.

[6] Channel matrix $\boldsymbol{H} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

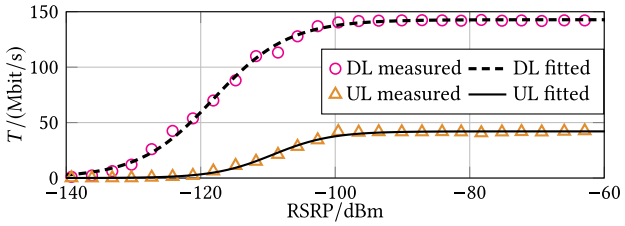

FIGURE 3. DL and UL throughput measured in an LTE test-cell. Black lines represent logistic functions fitted to the measurement data.

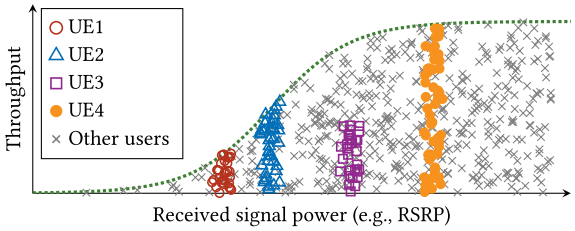

FIGURE 4. An illustration of throughput vs signal-power measurements in a crowdsourcing scenario. Whereas UE1, UE2 and UE4 reach the maximum throughput at some point, the UE3 experiences throughput degradation (e.g., tariff limit or less receive antennas).

locations at the same time points, whereas times and locations of crowdsourced measurements barely coincide.

A. THROUGHPUT AS A FUNCTION OF SIGNAL POWER

As already mentioned, neither SNR nor SINR are provided in the open data. Therefore, we characterize throughput as a function of signal power. In LTE, the indicator conveying average power of cell-specific reference signals [50] is called RSRP (see 3GPP [8], [46] for more details).

Keeping RSRP and all system parameters unchanged, the variation in noise + interference power P_Z modifies the SINR. If static users in different locations with a stable RSRP experienced enough different P_Z levels, we would observe a broad spread of their throughput values (as illustrated in Fig. 4).

The RSRP is based on the DL signal; however, there is no UL signal strength indicator available in the open data. Although our method works well for UL throughput versus RSRP, we should be cautious with its interpretation.[7]

B. EXTRACTING NETWORK PERFORMANCE

To achieve a particular maximum throughput, a certain SINR is necessary but not sufficient. The challenge in the processing of crowdsourced measurements is that there are many possible causes, which prevent a user from reaching full throughput for a given channel. Although it is possible to detect tariff limits [19] and to a certain extent also cell load caused by other users [8], we lack solutions for detecting the user's own cross-traffic or anomalies in the core network.

[7] DL and UL differ in modulation [50]. In case of frequency division duplex, the center-frequencies are different.

TABLE 2. We analyzed three major MNOs in Austria, Slovenia and Slovakia considering LTE tests from given time ranges.

Country	Regulator	Considered time range	MNO
Austria (AT)	RTR	2014-12-01 – 2018-01-01	A, B, C
Slovenia (SL)	AKOS	2015-07-01 – 2018-01-01	D, E, F
Slovakia (SK)	RÚ	2017-01-01 – 2018-01-01	G, H, I

Finding an exhaustive list of all degradation causes may not be possible; there can always be some that exceed our imagination. Therefore, we need to abstract from specific causes to allow a generic solution. By selecting only the best measurements (highest throughput) for every RSRP level, we can obtain a throughput characteristic that reflects technology limit (given by coding and modulation scheme, MIMO order, spatial multiplexing strategy, quality of network planning) rather than the performance of limited tariffs or users watching streamed videos during their measurements.

C. SELECTION OF BEST PERFORMANCE

Let us again consider Fig. 4. By selecting only the best tests for every RSRP, we obtain the upper boundary of the whole dataset (green dotted line). In the illustration, UE1, UE2, and UE4 are able to reach the maximum throughput corresponding to their max. SINR (min. noise and interference achieved during the off-peak hours at given RSRP). Although UE3 experiences higher RSRP than UE2, his/her maximum throughput is smaller – the possible causes of throughput degradation are cross-traffic, tariff limit, or only one (instead of two) receive antenna.

In the real dataset, in contrast to our example, we observe only a few repeated measurements from each of many unique users. The idea is, however, the same: **By considering only the best throughput for every signal strength level, we obtain performance corresponding to measurements conducted at the lowest cell load and lowest interference power (during off-peak hours), which are not degraded by any tariff limit or users' own cross-traffic.** We are convinced that this boundary represents a suitable performance benchmark that is capable of comparing which MNO uses more advanced technology, has better network planning, or better hardware configuration.

The current size of the dataset is not sufficient to characterize individual cells. We can evaluate only large areas. The limitation is that we detect only the best technology: if 4 × 4-MIMO is in use, we cannot detect the performance of 2 × 2-MIMO. We measure the performance of 20 MHz channel BW, but not the performance of 10 MHz channel BW. However, in principle, it should be possible to match the results of individual cells with that of the simulation results (Fig. 1), and find whether a base station uses CLSM or OLSM, 2 or 4 transmit antennas; and whether it is capable of carrier aggregation or offers only 20 MHz or 10 MHz channels.

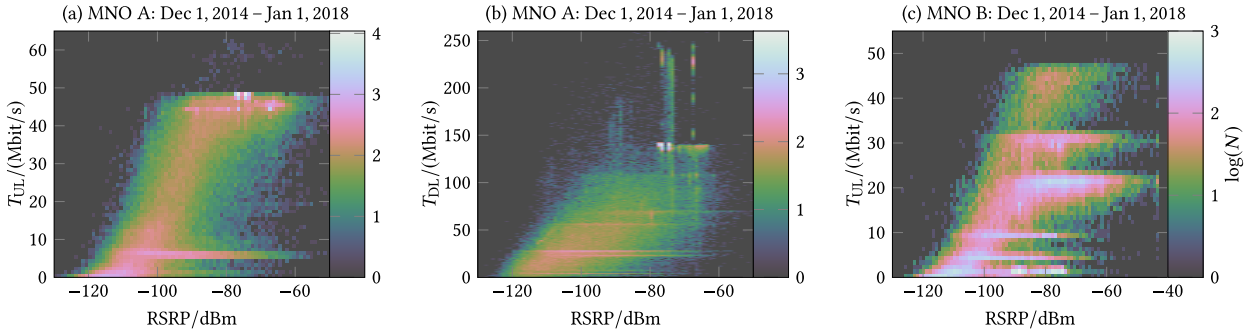

FIGURE 5. Histograms of (s_i, T_i) pairs. All three cases visualize different sets of LTE tests, i.e., s_i represents RSRP. (a) UL tests of MNO A. (b) DL tests of MNO A. (c) UL tests of MNO B. Bin size is 1 dBm × 1 Mbit/s.

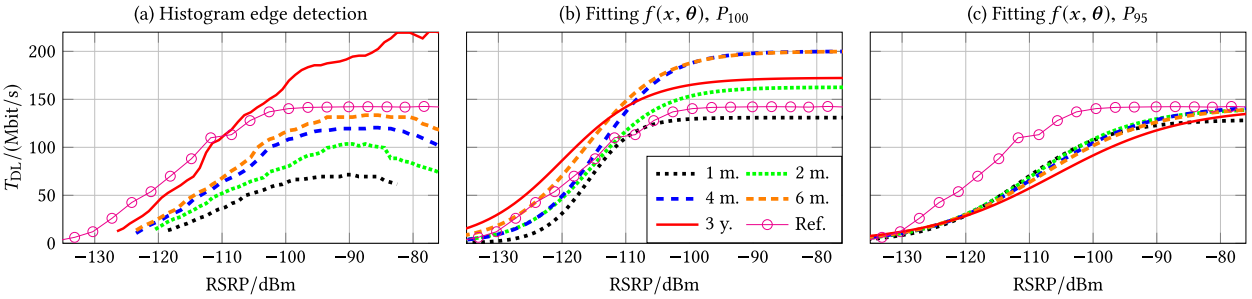

FIGURE 6. Boundary lines characterizing throughput performance of MNO A in LTE DL for different sizes of time interval $\mathcal{I}$, i.e., for different numbers of tests $|\mathcal{K}|$. (a) Detection of histogram edge. (b) Function fitting approach based on 100-percentile. (c) Function fitting based on 95-percentile.

IV. BENCHMARKING METHOD: REAL DATA

A. CROWDSOURCED DATASETS

There are many mobile applications that users can download and trigger a measurement by pressing a single button [51], [52], [53]. Each test usually consists of throughput measurements in both directions (DL, UL) and a latency test. Additional tasks vary among different applications. Some of them perform quality-of-service measurements (loading a web page, DNS lookup, playing a video), or collect additional information (GPS location, signal strength).

Measurement results are uploaded to a centralized database. In some cases, the data are free of charge and publicly available. We use open data provided by the regulatory bodies of Austria (RTR) [54], Slovenia (AKOS) [55] and Slovakia (RÚ) [15].

B. NOTATION AND FILTERING

Measurement applications of all three open data sources (RTR, AKOS, RÚ) are based on the same open-source framework Open-RMBT [56]. Therefore, the accessible measurement results have a similar structure. We consider the following entries: unique test ID k; timestamp t; SIM MCC-MNC (mobile country code, mobile network code); network MCC-MNC; technology (2G, 3G, 4G); network type (LTE, HSPA+, EDGE,...); mean network layer throughput[8] T in both directions (UL and DL).

[8]I.e., including transport layer headers. By inspecting the Open-RMBT's source code [56], we found out that it uses class TrafficStats [57] and its methods getUidRxBytes, getUidTxBytes, getTotalRxBytes, getTotalTxBytes.

Additionally, LTE tests contain RSRP. Only the Austrian open data provide the user ID for every test and the RSRQ (reference signal received quality) for LTE tests.

We identify every MNO by its MCC-MNC. To select only tests that have been performed in the users' home network (no roaming), we require the SIM MCC-MNC equal to the network MCC-MNC. Let $\mathcal{K}$ be the set of all IDs corresponding to tests of a single operator and unique technology (LTE only, HSPA+ only,...) in a given time interval $t \in \mathcal{I}$. To ease notation, we omit these filter criteria, and write just $k \in \mathcal{K}$.

The test denoted by a unique ID $k \in \mathcal{K}$ is the tuple

$$\text{test}_k = \left(k, t_k, s_k, T_{\text{UL},k}, T_{\text{DL},k}, \text{MNO}_k\right)$$

with timestamp t_k, signal level s_k (RSRP in LTE), mean throughput in UL $T_{\text{UL},k}$, mean throughput in DL $T_{\text{DL},k}$. In every country, we select the three MNOs with the highest number of tests (see Table 2).

C. VISUAL INSPECTION: 2D-HISTOGRAM

Our first approach was based on 2D-histogram of pairs (s_k, T_k) $\forall k \in \mathcal{K}$ (see examples in Fig. 5). The boundary line that characterizes the best throughput for every signal strength level is obtained by detecting the upper edge in the histogram image. To remove outliers, (smooth the boundary line), a median filter is applied on the histogram before the edge-detection.

Although this method does not require knowledge of any model, it has two disadvantages. First, different bin sizes, median filter sizes, and edge detection threshold values lead to different boundary lines. Likewise, it is unclear how to

choose these parameters and how to interpret the resulting differences. Second, (and even more severe), increasing the number of tests (choosing $\mathcal{K}' \supset \mathcal{K}$, e.g., by increasing time range $|\mathcal{I}'| > |\mathcal{I}|$) shifts the boundary line up [see Fig. 6 (a)].

D. MODEL-BASED PERFORMANCE INSPECTION

Instead of a 2D-histogram, we partition the data in one dimension only (we split the tests into different signal strength columns). In every column, we calculate the p-percentile of throughput and then fit a model function to these percentiles.

1) SIGNAL STRENGTH BINS

Let $b_1 < b_2 < \cdots < b_{L+1}$ denote bin edges for signal strength values, and let $x_l = (b_l + b_{l+1})/2$ be their corresponding bin centers; $l = 1, \ldots, L$. Set $\mathcal{K}$ can be split into L pairwise disjoint "columns" (or bins)

$$\mathcal{C}_l = \{k \in \mathcal{K}\colon\ b_l \le s_k < b_{l+1}\}, \quad l = 1, \ldots, L,$$

The tests with $s_k < b_1$ or $s_k \ge b_{L+1}$ are excluded, $\bigcup_{l=1}^{L} \mathcal{C}_l \subseteq \mathcal{D}$. In what follows, we will consider only non-empty columns: $l \in \mathcal{L}$ with $\mathcal{L} = \{l \in \{1, \ldots, L\}\colon\ |\mathcal{C}_l| > 0\}$.

2) THROUGHPUT PERCENTILES

For every $l \in \mathcal{L}$, we calculate the p-percentile (P_p) of the throughput (whether T_{UL} or T_{DL} will always be clear from context):

$$y_l = P_p of \{T_k\colon\ k \in \mathcal{C}_l\}.$$

By "P_p of some set", we mean the p-percentile calculated using the linear interpolation between adjacent ranks.

Fig. 7 depicts an example of throughput p-percentiles y_l based on 1 dB bins with centers x_l for different values of p. The choice $p = 100$ is the most literal implementation of our idea that the network's throughput performance should be characterized by throughput nondegraded tests.

However, $p = 100$ still leads to severe inconsistencies when the number of tests $|\mathcal{K}'| > |\mathcal{K}|$ is increased [Fig. 6 (b)]. By removing a single percentile, $p = 99$, the curve in Fig. 7 drops significantly and appears to be less susceptible to outliers (especially in the interval $[-130, -110]$ dBm).

By further decreasing the p-value, we achieve smoother curves and more consistent results for different time ranges of considered tests [Fig. 6 (c)]. On the other hand, compared to [Fig. 6 (a) and (b)], we move further away from the reference measurements.

The interpretation is still more straightforward than that in the case of the 2D-histogram. To benchmark different MNOs, we compare their upper bounds on the throughput achieved by the lower p percent of tests for every signal strength level.

3) CURVE FITTING

Even with the choice of $p = 90$, there are some apparent outliers left, especially in columns containing fewer tests

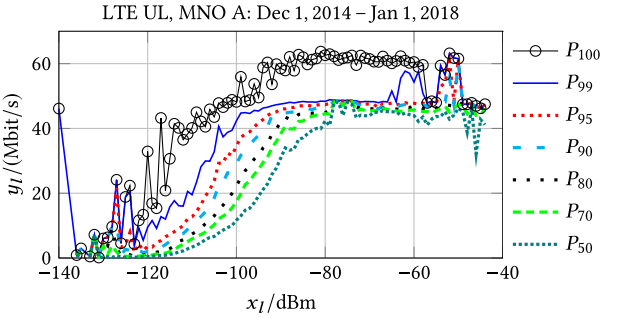

FIGURE 7. LTE UL tests of MNO A [same as Fig. 5 (a)]. Throughput percentiles y_l for RSRP columns with centers x_l at $-140, -139, \ldots, -40$ dBm for different values of p.

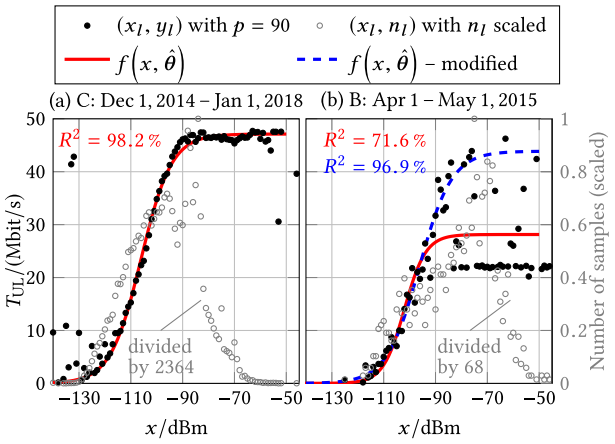

FIGURE 8. LTE tests (x = RSRP). Examples of function $f(x, \hat{\theta})$ fitted to (x_l, y_l). We denote $n_l = |\mathcal{C}_l|$ the number of samples in the l-th column. In plot (b) the fit is heavily impacted by tariff limited tests or by tests conducted in frequency band with 10 MHz BW. The modified fit (blue dashed) is performed after removing all (x_l, y_l) with $x_l > -90$ dBm $\wedge$ $y_l < 35$ Mbit/s, and it should characterize performance of 20 MHz BW with no tariff limits.

[e.g., in Fig. 8 (a) below -120 and above -60 dBm]. To remove the remaining outliers, we characterize the p-percentile throughput by a model function $f(x, \boldsymbol{\theta})$ fitted to data points (x_l, y_l). In our implementation, we use the logistic function (1), which we introduced in Section II-A.

In Fig. 8 (a), the counts of tests in every column $n_l = |\mathcal{C}_l|$ are not uniformly distributed. This suggests that the most extreme outliers are based on just a few samples, indicating a lack of data. At first glance, it appears reasonable to employ weighted least squares and assign the weights according to counts n_l.

However, although all y_l in the region $x_l \in [-140, -130]$ dBm are based on similar counts n_l, the higher throughput values are less trustworthy because they are not physically achievable at such poor signal strength levels. Therefore, we do not consider counts n_l and utilize IRLS (Section II-B) instead.

V. RESULTS AND DISCUSSION

In the following section, we discuss the consistency over time and the visualization of the fits. Subsequently, we show that the goodness of the fit can be improved by increasing

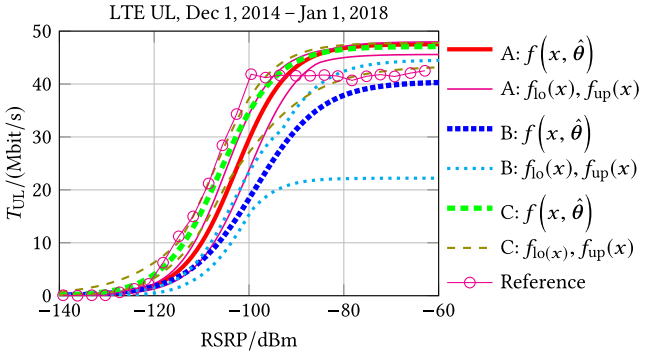

FIGURE 9. Fits f of MNOs A, B, C based on a three-year interval and functions f_{lo}, f_{up} based on six-month subintervals.

the number of tests. Finally, we compare the results of the different MNOs.

A. FIT CONSISTENCY AND VISUALIZATION

To inspect variations of the fits over time, we split the data into six-month intervals (6-m.i.) 2014-12-01 – 2015-06-01, …, 2014-12-01 – 2017-12-01, obtaining seven different fits for every MNO. Displaying fits of all subintervals in a single plot can be confusing; thus, we plot the fit $f\left(x,\hat{\boldsymbol{\theta}}\right)$ based on all samples with $t \in \mathcal{I}$, and then boundaries $f_{\mathrm{lo}}(x,)$, $f_{\mathrm{up}}(x)$ based on fits $f_i\left(x,\hat{\boldsymbol{\theta}}_i\right)$ obtained from pairwise disjoint subintervals $\mathcal{I}_i$, with $\bigcup_{i=1}^{I}\mathcal{I}_i = \mathcal{I}$:

$$f_{\mathrm{lo}}(x) = \min_{i\in\{1,\ldots,I\}} f_i\left(x,\hat{\boldsymbol{\theta}}_i\right), \tag{7}$$

$$f_{\mathrm{up}}(x) = \max_{i\in\{1,\ldots,I\}} f_i\left(x,\hat{\boldsymbol{\theta}}_i\right). \tag{8}$$

An example is depicted in Fig. 9. We choose $|\mathcal{I}_i|$ = six months to assure $R^2 > 0.9$ for all fits (details are given in Section V-C).

B. PERFORMANCE INDICATOR DEFINITION

It is not clear in Fig. 9, which interval produces which fit $f_i\left(x,\hat{\boldsymbol{\theta}}_i\right)$. Every fit is fully characterized by n parameters (for logistic function $n = 3$). To inspect the dependency of the network performance on time, we need to compress all n parameters to a single indicator. For a non-negative, continuous, monotonically increasing function, the area under the curve is a reasonable indicator (although it cannot preserve all information contained in $\boldsymbol{\theta}$):

$$A(\boldsymbol{\theta}) = \frac{1}{\mathrm{dBW}\cdot\mathrm{Mbit/s}} \cdot \int_a^b f(x,\boldsymbol{\theta})\,\mathrm{d}x. \tag{9}$$

The normalization term in front of the integral is included, such that A is dimensionless: $[A] = 1$.

Area A does not capture all the details. It may be that MNO 1 performs better at lower signal strengths and worse at higher signal strengths, as compared to MNO 2, albeit both have the same area A. In order to discover such cases, we have to fall back to the representation in Fig. 9.

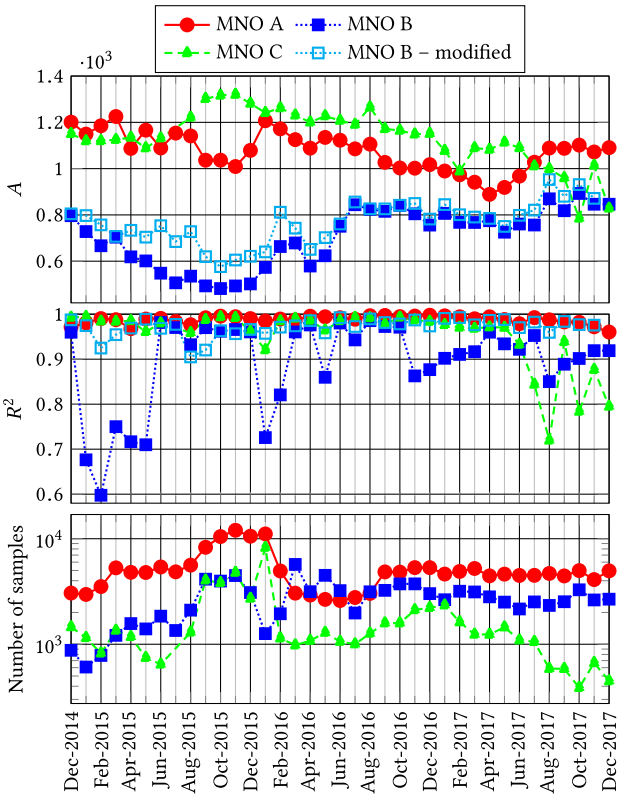

FIGURE 10. Network performance as a function of time, evaluated for one-month intervals; LTE UL, $p = 90$. Plots show area indicator A, the corresponding coefficient of determination R^2, and the number of samples on which every fit is based. "MNO B – modified" stands for fits that are calculated after excluding outliers $x_l > -90\,\mathrm{dBm} \wedge y_l < 35\,\mathrm{Mbit/s}$ [see also Fig. 8 (b)].

In the following, we apply $a = -140\,\mathrm{dBm}$ for the lower bound. For the upper bound, we use $b = -80\,\mathrm{dBm}$ because most of the fits are already close to the higher asymptote at this signal strength level. This means that $f(x,\hat{\boldsymbol{\theta}})$ is nearly constant for $x > b$. If we chose a larger b, then area A would be more impacted by the upper asymptote α and less by the shift x_0 and steepness β.

C. QUALITY OF FIT

This section analyzes the performance at different time intervals. Fig. 10 depicts A, R^2, and the number of samples of three AT MNOs based on 1-m.i. In agreement with Fig. 9, we see that MNO C performs slightly better than MNO A most of the time. Both perform better than MNO B. In all intervals of MNO A, the R^2 is larger than 96%. The fits of MNO C show $R^2 > 92\%$ except during the last six months, in which the number of samples drops below 700. This may be due to the low fit quality.

Fits of MNO B have poorer quality than those of MNO A and C. The distribution in Fig. 5 (c) is strongly multimodal and $p = 90$, is not sufficient to eliminate the lower modes in Fig. 8 (b). As a workaround, we manually exclude all (x_l, y_l) with $x_l > -90\,\mathrm{dBm} \wedge y_l < 35\,\mathrm{Mbit/s}$, and obtain the modified fits [Fig. 8 (b), blue dashed], achieving $R^2 > 90\%$ in all subintervals (Fig. 10). Area A is still smaller than MNO

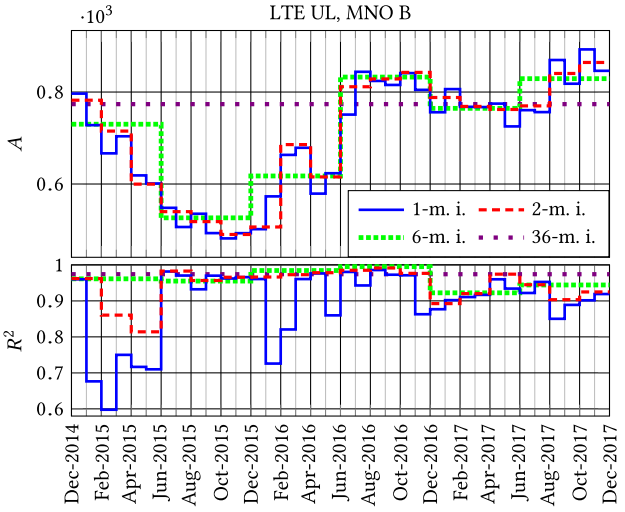

FIGURE 11. Area indicator A and coefficient of determination R^2 for different interval lengths. Increasing the interval duration leads to better fit (higher R^2).

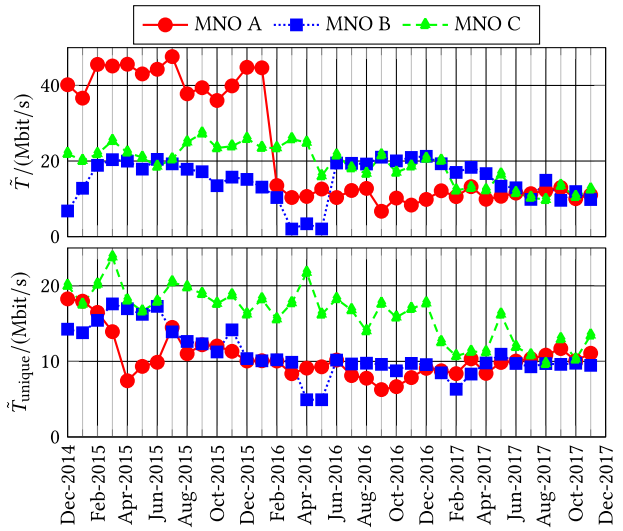

FIGURE 12. LTE UL. Throughput median $\tilde{T}$ considering all tests (upper plot) and throughput median $\tilde{T}_{\text{unique}}$ after excluding repeated tests, i.e., considering the first test of every user, (lower plot).

A and C—the upper asymptote is lifted up, but the whole logistic function of MNO B is still right-shifted relative to A and C.

How large should the time intervals be for our time-series analysis? Longer time intervals accommodate more tests, which usually leads to more accurate percentile estimates and also to a better fit. On the other hand, we lose time resolution. This trade-off is illustrated in Fig. 11.

It is not clear how large the dataset should be for a good fit. Each MNO has a different (usually multimodal) distribution. We can start with short intervals, and then prolong them until R^2 decreases below the desired value. For MNO B, which has the poorest fit, we get $R^2 < 60\%$ for 1-m.i., $R^2 > 80\%$ for 2-m.i., and $R^2 > 90\%$ for 6-m.i. (Fig. 11). Extending the time intervals leads to smoothing with preference of higher values of A.

D. STABILITY IN COMPARISON WITH MEDIAN

We compare area A with a median that is often reported by regulatory bodies as a benchmark of operators' performance. AT open data also includes (besides test IDs) user IDs. User ID allows detecting repeated tests of a single user. The statistics published by RTR [58] is calculated after removing repeated tests. The median $\tilde{T}$ (Fig. 12, upper plot) based on all tests of MNO A between December 2014 and January 2016 is higher by $\tilde{3}0$ Mbit/s than that after January 2016. If we remove repeated tests ($\tilde{T}_{\text{unique}}$, lower plot), then this sudden jump disappears.

Removing the repeated tests is an attempt to make a fairer comparison, that is more resilient against systematic bias (e.g., MNO injecting high-throughput tests conducted in an unloaded cell to improve its position in the published statistics). However, a new user ID can be generated by deleting the app's data, which can be easily automatized to obtain a new user ID for every test. Moreover, SL and SK open data contain no user IDs.

The throughput median can be manipulated by injecting many high-throughput tests at arbitrary signal strength levels ["hot-spots" in Fig. 5 (a) and (b)], several high columns between -80 and -60 dBm in plot (b). To influence our method, one would need to inject many tests at many different signal strength levels.[9]

Comparing A in Fig. 10 with $\tilde{T}_{\text{unique}}$ in Fig. 12, we see that performance of MNO C is more stable (less noisy) using our method. The metric $\tilde{T}_{\text{unique}}$ also suggests—contradicting our observations—that MNO B and A perform comparably poor.

E. NUMERICAL RESULTS

In the same way, as for the Austrian MNOs, we repeated the whole analysis for Slovenian and Slovakian MNOs. We summarize the results in Table 3 that shows A, R^2 and $\tilde{T}$ for fits based on all available tests in given time ranges, for both UL and DL.

To numerically indicate variations of 6-month subinterval fits, we always included the lowest and the highest 6-m.i. value expressed as a percentage of the value, which we base on all the samples: The value A (based on all samples) is accompanied by $+A_{\text{rel,max}}/-A_{\text{rel,min}}$ %, where $A_{\text{rel,max}} = \frac{\max_i\{A_i\}-A}{A} \cdot 100$ and $A_{\text{rel,min}} = \frac{A-\min_i\{A_i\}}{A} \cdot 100$. The value A_i is based on samples from i-th subinterval. We represent the medians $\tilde{T}$ in the same manner. The R^2 values are already expressed as percentages, therefore $R^2_{\text{rel,max}} = \max_i\{R_i^2\} - R^2$ and $R^2_{\text{rel,min}} = R^2 - \min_i\{R_i^2\}$.

If all subinterval quantities Q_i are smaller (or larger) than Q, than table contains 0 instead of $Q_{\text{rel,max}}$ (or $Q_{\text{rel,min}}$). For AT MNOs we had six 6-m.i., for SL five 6-m.i. and for SK only two 6-m.i. Therefore the variation percentages for SK

[9]Every signal strength column yields one value y_l regardless of the number of samples in the column. The fit is based on all y_l values. If some y_l differs too much from its neighbors, then it gets (due to IRLS) a lower weight assigned.

TABLE 3. Evaluation of fits in LTE UL and DL based on data from time ranges in table 2.

ISP	Tests /10^3	LTE UL $A/10^3$	LTE UL $R^2/\%$	LTE UL $\tilde{T}/(\mathrm{Mbit/s})$	LTE DL $A/10^3$	LTE DL $R^2/\%$	LTE DL $\tilde{T}/(\mathrm{Mbit/s})$
A	191.0	**1.09** +8.4/−12.1 %	99.1 +0.8/−0.3	**24.6** +82.3/−63.2 %	**3.3** + 3.3/−11.1 %	95.4 +3.5/− 2.7	**51.4** +105.5/−38.4 %
B	98.4	**0.78** +6.5/−32.6 %	97.4 +2.1/−5.2	**15.9** +28.4/−67.3 %	**2.2** +24.4/−32.6 %	94.7 +4.0/− 0.8	**32.0** + 46.5/−44.0 %
C	60.3	**1.23** +6.3/−12.4 %	98.2 +1.1/−4.1	**21.4** +17.2/−50.3 %	**3.5** + 6.3/−25.2 %	95.9 +2.0/−10.8	**41.8** + 13.0/−41.9 %
D	14.0	**0.93** +8.4/−13.3 %	97.7 +1.0/−5.2	**9.4** +32.2/−26.5 %	**2.6** +11.8/−13.5 %	96.0 +0.0/−15.0	**23.7** +36.5/−14.4 %
E	11.7	**0.65** +9.7/− 3.7 %	97.6 +0.2/−2.5	**5.6** +68.5/−11.6 %	**2.0** + 9.5/−18.7 %	91.0 +0.0/−15.0	**24.7** +16.4/−13.6 %
F	5.2	**0.52** +8.4/− 9.3 %	99.1 +0.2/−8.5	**11.3** +23.5/−25.0 %	**1.4** +21.2/−13.4 %	97.7 +0.0/−22.8	**19.6** +18.3/−11.0 %
G	8.8	**0.40** +5.8/− 0.6 %	99.7 +0.0/−0.5	**14.0** + 2.4/− 1.0 %	**1.4** + 0.3/− 4.8 %	98.0 +0.0/− 1.6	**26.5** +12.9/− 8.6 %
H	2.7	**0.62** +0.0/− 3.4 %	96.6 +0.0/−1.1	**10.3** + 0.6/− 5.4 %	**2.5** +17.9/−23.1 %	88.0 +0.0/−14.9	**25.2** +24.1/−10.2 %
I	1.5	**1.24** +0.0/− 3.1 %	98.0 +0.0/−0.6	**19.9** + 0.9/− 3.2 %	**4.1** + 7.1/−26.2 %	72.9 +0.1/−32.3	**42.9** + 7.2/−14.6 %

For each MNO, we show the total number of tests on which the fit is based, the area indicator A, coefficient of determination R^2 and median $\tilde{T}$. Each of the quantities $A, R^2, \tilde{T}$ is followed by the maximum and minimum 6-m.i. value (expressed as a percentage).

MNOs are grayed out in the table to indicate that these values are not very informative.

All fits in LTE UL have $R^2 > 90\%$. In LTE DL, the fit quality is worse for some MNOs (especially in SK, where the fitting performs poorly probably due to the insufficient number of samples).

In LTE UL our metric A shows smaller variations than $\tilde{T}$ in case of all AT and SL ISPs. In LTE DL this is the case only for the AT ISPs. The other cases, with higher variations of A, are also the cases with poorer fit quality R^2, indicating that we either have not enough samples or that we should use a better model function.

The area indicator A leads to a ranking that is utterly different from the classification based on the median $\tilde{T}$ (different shades of gray in Table 3). We would like to emphasize that MNO benchmarking is a sensitive issue and should not be reduced to a simple comparison of throughput medians (or other quantiles), which is, unfortunately, state of the art among national regulatory bodies.

VI. CONCLUSION

With our thought experiment, we have demonstrated that in such a complex system as a cellular mobile network, with many known and unknown features, we face the problem of underrepresentation even with the largest known crowdsourced dataset. It might never be possible to make certain types of predictions (e.g., predicting throughput of a specific device model at a given time and location based on all available features).

By allowing a certain level of data aggregation, we avoided the underrepresentation problem, yet, at the same time, we were able to provide a more informative picture than is often presented in public reports [10]–[12]. We proposed a new network-centric metric for network throughput benchmarking: We model the throughput of an unloaded cell as a function of signal strength, and by calculating the p-percentile, we exclude throughput-degraded tests—thus keeping only the measurements conducted under the best conditions for every signal strength level. The IRLS algorithm automatically removes the remaining outliers (caused by, e.g., automatized testing) by iteratively fitting a chosen model function with no need for manual preprocessing (e.g., setting cut-off thresholds for signal strength and throughput).

Previous studies investigating the dependency of throughput on signal strength [31], [13], [34] have not been so successful because they used linear models and aimed to explain all the measurements without reflecting the possibility of throughput degradation. Uludağ and Korçak [34] achieved $R^2 < 28\%$, our method reaches $R^2 > 90\%$.

We proposed several different graphical representations suitable for inspecting various aspects. We evaluated three open data sources (AT, SL, and SK regulators) and compared a total of nine MNOs (three from every country) in LTE DL and UL. Furthermore, we simulated and measured throughput vs RSRP in an unloaded cell to baseline the results obtained from crowdsourced data.

National regulators often benchmark mobile operators based on the throughput percentile regardless of the signal strength. Our method is more resilient to systematic bias (injected high-throughput tests), leading to different MNO ranking (shown in Table 3).

Our proposed method allows benchmarking MNOs at different signal strength levels (one MNO may perform better at lower signal strengths, another at higher—due to, e.g., frequency vs time division duplex). Aside from the performance metric, our method provides a reliability indicator (R^2), which tells how good the fit is and how much we can trust the result. Finally, considering subintervals, we can inspect the stability of the metric / observe network changes (e.g., technology upgrade).

REFERENCES

[1] Waze Mobile. (2019). *Free Driving Directions, Traffic Reports & GPS Navigation App by Waze*. [Online]. Available: https://www.waze.com/
[2] (2019). *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia
[3] (2019). *OpenStreetMap*. [Online]. Available: https://www.openstreetmap.org/about
[4] Wikipedia. (2019). *reCAPTCHA*. [Online]. Available: https://en.wikipedia.org/wiki/ReCAPTCHA

[5] F. Finazzi. (2019). *Earthquake Network Project—Earthquake Network*. [Online]. Available: http://wp.earthquakenetwork.it/
[6] V. Raida, P. Svoboda, and M. Rupp, "Constant rate ultra short probing (CRUSP) measurements in LTE networks," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Chicago, IL, USA, Aug. 2018, pp. 1–5.
[7] V. Raida, P. Svoboda, M. Kruschke, and M. Rupp, "Constant rate ultra short probing (CRUSP): Measurements in live LTE networks," in *Proc. IEEE ICC*, May 2019, pp. 1–6.
[8] V. Raida, M. Lerch, P. Svoboda, and M. Rupp, "Deriving cell load from RSRQ measurements," in *Proc. Netw. Traffic Meas. Anal. Conf. (TMA)*, Jun. 2018, pp. 1–6.
[9] M. Rindler, P. Svoboda, and M. Rupp, "FLARP, fast lightweight available rate probing: Benchmarking mobile broadband networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.
[10] Tutela. (2019). *Austria: 3 Lead Download Speeds as A1 Top Upload and Latency*. [Online]. Available: https://www.tutela.com/blog/austria-3-lead-download-speeds-as-a1-top-upload-and-latency
[11] OpenSignal. (2018). *Telekom Scores a 4G Clean Sweep of Germany's Regions*. [Online]. Available: https://www.opensignal.com/blog/2018/06/05/telekom-scores-a-4g-clean-sweep-of-germanys-regions
[12] Ookla. (2016). *Germany Speedtest Report*. [Online]. Available: https://www.speedtest.net/reports/germany/
[13] J. Cainey, B. Gill, S. Johnston, J. Robinson, and S. Westwood, "Modelling download throughput of LTE networks," in *Proc. IEEE 39th Conf. Local Comput. Netw. Workshops (LCN Workshops)*, Sep. 2014, pp. 623–628.
[14] *Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for Support of Radio Resource Management*, document TS 36.133, Version 15.0.0, 3GPP, Sep. 2017.
[15] Merač Internetu. (2018). *Úrad pre Reguláciu Elektronických Komunikácií a Poštovných Služieb*. [Online]. Available: https://www.meracinternetu.sk/en/opendata
[16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[17] Tutela. (2019). *Our Methodology—Tutela*. [Online]. Available: https://www.tutela.com/methodology
[18] P. Kanuparthy and C. Dovrolis, "ShaperProbe: End-to-end detection of ISP traffic shaping using active methods," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf. (IMC)*, 2011, pp. 473–482. doi: 10.1145/2068816.2068860.
[19] V. Raida, P. Svoboda, and M. Rupp, "Lightweight detection of tariff limits in cellular mobile networks," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–7.
[20] V. Raida, P. Svoboda, and M. Rupp, "Repeatability for spatiotemporal throughput measurements in LTE," in *Proc. 89th VTC Spring*, Apr. 2019, pp. 1–5.
[21] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 32–39, Nov. 2011.
[22] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, and X. Zhou, "Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 7:1–7:31, Aug. 2015. [Online]. Available: http://doi.acm.org/10.1145/2794400
[23] F. Ma, X. Liu, A. Liu, M. Zhao, C. Huang, and T. Wang, "A time and location correlation incentive scheme for deep data gathering in crowdsourcing networks," *Wireless Commun. Mobile Comput.*, vol. 2018, Jan. 2018, Art. no. 8052620.
[24] Y. Wang, Z. Cai, G. Yin, Y. Gao, X. Tong, and G. Wu, "An incentive mechanism with privacy protection in mobile crowdsourcing systems," *Comput. Netw.*, vol. 102, pp. 157–171, Jun. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128616300883
[25] C.-K. Tham and T. Luo, "Quality of contributed service and market equilibrium for participatory sensing," in *Proc. IEEE Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, May 2013, pp. 133–140.
[26] S. Reddy, D. Estrin, and M. Srivastava, "Recruitment framework for participatory sensing data collections," in *Proc. 8th Int. Conf. Pervasive Comput. (Pervasive)*. Berlin, Germany: Springer-Verlag, 2010, pp. 138–155. doi: 10.1007/978-3-642-12654-3_9.
[27] Z. Song, C. H. Liu, J. Wu, J. Ma, and W. Wang, "QoI-aware multitask-oriented dynamic participant selection with budget constraints," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4618–4632, Nov. 2014.
[28] F. Zhang, B. Jin, H. Liu, Y.-W. Leung, and X. Chu, "Minimum-cost recruitment of mobile crowdsensing in cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–7.
[29] F. Campioni, S. Choudhury, K. Salomaa, and S. G. Akl, "Improved recruitment algorithms for vehicular crowdsensing networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1198–1207, Feb. 2019.
[30] K. Kousias, C. Midoglu, O. Alay, A. Lutu, A. Argyriou, and M. Riegler, "The same, only different: Contrasting mobile operator behavior from crowdsourced dataset," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–6.
[31] A. S. Khatouni, M. Mellia, M. A. Marsan, S. Alfredsson, J. Karlsson, A. Brunstrom, O. Alay, A. Lutu, C. Midoglu, and V. Mancuso, "Speedtest-like measurements in 3G/4G networks: The monroe experience," in *Proc. 29th Int. Teletraffic Congr. (ITC)*, vol. 1, Sep. 2017, pp. 169–177.
[32] F. Jungermann. (2016). *Crowdsourced 4G Experience: Benchmarking Nordic Operators*. [Online]. Available: https://tefficient.com/crowdsourced-4g-experience-benchmarking-nordic-operators/
[33] T. Linder, P. Persson, A. Forsberg, J. Danielsson, and N. Carlsson, "On using crowd-sourced network measurements for performance prediction," in *Proc. 12th Annu. Conf. Wireless On-Demand Netw. Syst. Services (WONS)*, Jan. 2016, pp. 1–8.
[34] K. Uludağ and O. Korçak, "Energy and rate modeling of data download over LTE with respect to received signal characteristics," in *Proc. 27th Int. Telecommun. Netw. Appl. Conf. (ITNAC)*, Nov. 2017, pp. 1–6.
[35] Tutela. (2019). *Tutela Explorer—Extreme Analysis and Visualisation Platform*. [Online]. Available: https://www.tutela.com/explorer
[36] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Commun. Statist.-Theory Methods*, vol. 6, no. 9, pp. 813–827, 1977. doi: 10.1080/03610927708827533.
[37] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, vol. 16, no. 2, pp. 147–185, 1974.
[38] MathWorks. (2018). *Nonlinear Regression—MATLAB Nlinfit*. [Online]. Available: https://de.mathworks.com/help/stats/nlinfit.html
[39] W. Dumouchel and F. O'Brien, "Integrating a robust option into a multiple regression computing environment," in *Proc. 21st Symp. Interface Comput. Sci. Statist.*, 1989, pp. 297–302.
[40] T. F. Coleman and Y. Li, "An interior, trust region approach for nonlinear minimization subject to bounds," *SIAM J. Optim.*, vol. 6, no. 2, pp. 418–445, 1996.
[41] T. F. Coleman and Y. Li, "On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds," *Math. Programm.*, vol. 67, no. 2, pp. 189–224, Mar. 1994.
[42] T. O. Kvålseth, "Cautionary note about R^2," *Amer. Statist.*, vol. 39, no. 4, pp. 279–285, 1985. [Online]. Available: http://www.jstor.org/stable/2683704
[43] J. B. Willett and J. D. Singer, "Another cautionary note about R^2: Its use in weighted least-squares regression analysis," *Amer. Statist.*, vol. 42, no. 3, pp. 236–238, 1988. [Online]. Available: http://www.jstor.org/stable/2685031
[44] M. Rupp, S. Schwarz, and M. Taranetz, *The Vienna LTE-Advanced Simulators: Up and Downlink, Link and System Level Simulation* (Signals and Communication Technology), 1st ed. Singapore: Springer, 2016.
[45] S. Schwarz, J. C. Ikuno, M. Simko, M. Taranetz, Q. Wang, and M. Rupp, "Pushing the limits of LTE: A survey on research enhancing the standard," *IEEE Access*, vol. 1, pp. 51–62, 2013.
[46] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer; Measurements*, document TS 36.214, Version 15.2.0, 3GPP, 2018.
[47] Keysight Technologies. (2018). *Nemo Handy Handheld Measurement Solution*. [Online]. Available: https://www.keysight.com/en/pd-2767485-pn-NTH00000A/nemo-handy
[48] J. Dugan. (2016). *iPerf—The TCP, UDP and SCTP Network Bandwidth Measurement Tool*. [Online]. Available: https://iperf.fr
[49] Wikipedia. (2018). *Samsung Galaxy Note 4*. [Online]. Available: https://en.wikipedia.org/wiki/Samsung%5FGalaxy%5FNote%5F4
[50] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, document TS 36.211, Version 15.2.0, 3GPP, 2018.
[51] (2019). *Apps | Opensignal*. [Online]. Available: https://opensignal.com/apps
[52] Ookla LLC. (2019). *Speedtest by Ookla—The Global Broadband Speed Test*. [Online]. Available: http://www.speedtest.net/
[53] Alladin IT. (2018). *The Alladin Nettest*. [Online]. Available: https://nettest.alladin.at/home
[54] Rundfunk und Telekom Regulierungs-GmbH. (2019). *RTR—NetTest*. [Online]. Available: https://www.netztest.at/en/Opendata
[55] AKOS. (2018). *AKOS Test Net*. [Online]. Available: https://www.akostest.net/en/opendata

[56] GitHub. (2018). *Alladin-It/Open-Rmbt*. [Online]. Available: https://github.com/alladin-IT/open-rmbt
[57] GitHub. (2018). *TrafficStats | Android Developers*. [Online]. Available: https://developer.android.com/reference/android/net/TrafficStats.html
[58] Rundfunk und Telekom Regulierungs-GmbH. (2019). *RTR—NetTest*. [Online]. Available: https://www.netztest.at/en/Statistik

VACLAV RAIDA received the Dipl.Ing. degree in telecommunications from Technische Universität Wien (TU Wien), in 2017. He is currently a Project Assistant with the Institute of Telecommunications, TU Wien, focusing on performance measurements and data analysis in the context of cellular mobile networks.

PHILIPP SVOBODA (SM'15) received the Dr.Ing. degree in electrical engineering from Technische Universität Wien (TU Wien). He is currently a Senior Scientist with TU Wien, with a research focus on the performance aspects of mobile cellular technologies. He is currently examining the feasibility of using crowdsourcing to conduct performance measurements on 4G and 5G mobile networks. His research aims to establish a common framework for evaluating the performance of mobile networks, guaranteeing reliable and fair connectivity for end-users.

MARTIN LERCH received the Dipl.Ing. degree in telecommunications from Technische Universität Wien (TU Wien). He is currently with the Institute of Telecommunications, TU Wien, developing testbeds and measurement methodologies for controlled and reproducible wireless experiments at high velocities.

MARKUS RUPP received the Dipl.Ing. degree from the University of Saarbrücken, Germany, in 1988, and the Dr.Ing. degree from the Technische Universität Darmstadt, Germany, in 1993. Until 1995, he held a postdoctoral position at the University of California at Santa Barbara, Santa Barbara, CA, USA. From 1995 to 2001, he was with the Wireless Technology Research Department, Nokia Bell Labs, Holmdel, NJ, USA. Since 2001, he has been a Full Professor of digital signal processing in mobile communications with TU Wien.