Telecommunication Systems (2024) 87:437–453
https://doi.org/10.1007/s11235-024-01196-5

ORIGINAL PAPER

Statistical analysis and characterization of signaling and user traffic of a commercial multi-band LTE system

Cesar Vargas Anamuro[1] · Alberto Blanc[2] · Xavier Lagrange[2]

Accepted: 24 June 2024 / Published online: 5 July 2024

Abstract
The fourth-generation (4 G) cellular network is currently the dominant mobile technology used worldwide. The analysis of the network behavior can help forecast the traffic and thus improve the network. The characterization of mobile user behavior can be useful in evaluating the emerging concepts. We collected traffic traces on a commercial multi-band 4 G cell to better understand and model the network and user behavior. We evaluate the network utilization, the number of connections, and the uplink and downlink cell achieved throughput. In addition, we analyze and identify statistical models that describe the connection inter-arrival time, connection duration, and connection size. The results show daily and weekly patterns that depend not only on the time of day but also on the frequency band. We find that the frequency band and time of day have a limited impact on user behavior. On the contrary, the connection inter-arrival time strongly depends on the frequency band.

Keywords 4 G · LTE analyzer · Measurement · Multi-band · Statistical analysis · Characterization

1 Introduction

In recent years, the use of the fourth-generation (4 G) network has experienced significant growth due to the emergence of new services and applications. The 4 G mobile network is currently the dominant mobile technology used worldwide, representing around 58% of global connections during 2021 [1].

Thanks to measurements collected in a production cellular network, we can fit statistical models to different aspects of user and network behavior. These statistical models can be used, among other applications, to perform more realistic simulations, forecast network anomalies, predict cellular traffic [2], and optimize the network. For instance, Mobile Network Operators (MNOs) turn off some frequency bands to reduce energy consumption when network utilization is low [3]. On the other hand, the behavior of users, in general, is independent of the cellular technology used. Therefore, the characterization of user behavior can help optimize the current cellular technology and for better planning of emerging technologies.

Researchers can rarely access the measurements that MNOs collect to monitor and manage their networks. There are few datasets available on the internet [4, 5]. These datasets were collected more than 5 years ago when 4 G was not the dominant cellular technology [6]. However, these databases are still used by the scientific community [7]. Due to the lack of datasets and to collect specific network information, some studies focus on developing passive platforms for Long-Term Evolution (LTE) monitoring (LTE sniffers). Passive LTE sniffers based on Software-Defined Radios (SDRs) and open-source were presented in [8, 9]. The authors of [10] implemented an LTE sniffer based on field-programmable gate array (FPGA) hardware and proprietary software.

To address the lack of recent datasets, we developed an LTE analysis platform based on the open source library srsRAN (formerly srsLTE [11]). With this tool, we collected passive measurements inside a multi-band LTE cell of a French MNO in 2022. Our LTE analyzer captures all the mes-

This study was made when the co-author was a post-doc at IMT Atlantique.

✉ Xavier Lagrange
xavier.lagrange@imt-atlantique.fr

Cesar Vargas Anamuro
c.vargasanamuro@fr.merce.mee.com

Alberto Blanc
alberto.blanc@imt-atlantique.fr

[1] Mitsubishi Electric R&D Centre Europe, 1 Allee de Beaulieu, 35700 Rennes, France

[2] ADOPNET Team, IMT Atlantique/IRISA, 2 rue de la chataigneraie, 35576 Cesson-Sévigné, France

sages broadcasted by the eNodeB (eNB) in clear text, such as system information and downlink control information (DCI) messages. We conceived our LTE analyzer to capture data over several days and minimize information loss. Furthermore, we minimized the modifications of srsRAN software to facilitate the integration of new versions of srsRAN. Note that our LTE analyzer does not collect any private user information.

Using four LTE analyzers in parallel, we monitored the control messages of the four operating frequency bands (800 MHz, 1800 MHz, 2100 MHz, and 2600 MHz) within a cell of an MNO for 1 week. We verify that we missed no more than 0.03% of the subframes in each frequency band during the monitored period. The source code has been made publicly available at https://gitlab.imt-atlantique.fr/cvargasa/lte_analyzer, and the datasets used for this study can be provided upon request to 4gdataset@imt-atlantique.fr.

This paper provides two main contributions. First, we analyze the network characteristics (number of connections, network utilization, and downlink and uplink cell achieved throughput). This analysis includes the effects of time of day and frequency band. Second, we analyze and identify the statistical models that best characterize the user connections (inter-arrival time, connection duration, and connection size). We use two statistical tests (Kolmogorov-Smirnov and Cramer-Von Mises) to evaluate the goodness-of-fit of the proposed models. Moreover, to the best of our knowledge, this is the first study that analyzes a multi-band LTE cell.

The remainder of this paper is organized as follows. Section 2 discusses related works. In Sect. 3, we give an overview of the control channel in LTE. Section 4 describes the LTE analyzer platform implemented for this study. The measurement campaign details are explained in Sect. 5. Section 6 presents the analysis and characterization of the monitored 4 G network. The analysis and modeling of LTE connections are presented in Sect. 7. In Sect. 8, we briefly explain how our study can be extended. Conclusions and future work are given in Sect. 9.

2 Related work

As mentioned above, similar studies exist for older cellular technologies. For instance, using real data from 3 G cellular networks, the authors of [12–14] analyze the characteristics of cellular networks (cell throughput, number of users, traffic volume). The daily pattern shown in these studies coincides with our results in that the off-peak hour is around 5:00.[1] Some studies aim to characterize the behavior of 4 G cellular networks based on data collected by MNOs [8, 10, 15–17]. The time of day effects on network utilization and the number of active terminals on a university campus are shown in [8]. The authors of [10] presented the measurements made in different LTE cells. Among other results, the authors show the variation over time in the number of active users and the downlink network utilization. These studies focus on the overall behavior of the cellular network. However, the displayed results are limited to a few hours. In [16] authors investigate the network data traffic, showing the daily patterns of traffic and the number of active users in an eNB. None of these previous studies consider the effects of the frequency band in the analysis. The authors of [17] analyze the number of active terminals connected to a tri-sector eNB, where each sector has three frequency bands. They show that the number of active terminals has a periodic behavior that varies depending on the frequency band. However, the results are limited only to analyzing the number of active terminals. To the best of our knowledge, there is no study of the characteristics of 4 G networks that considers both the effects of the time of day and the frequency band.

As far as statistical models are concerned, most studies focus on the user inter-arrival time [16, 18–21]. Traditionally, researchers model the inter-arrival time as an exponential distribution [18, 22] since this assumption simplifies the mathematical models. A statistical model of the inter-arrival time based on real data from a public mobile 2 G network is presented in [19]. The authors show that the inter-arrival time can be modeled according to an exponential distribution as expected. The authors of [20] model the inter-arrival time. In this case, the distribution that better fitted the empirical distribution was the Erlang-3,k. Unlike 2 G mobile networks, 4 G provides multiple services such as web browsing, video streaming, data transfer, and voice calls. The connection inter-arrival time in commercial 4 G networks is investigated in [16, 21]. The authors of [21] suggest that the random access inter-arrival times in a terminal can be modeled as a mixture of exponential distributions. Unlike this work, we focus on the connection inter-arrival time at the eNB level. From two 30-minute data sets obtained from an eNB on a typical weekday afternoon, the authors of [16] showed that the inter-arrival time is exponentially distributed. However, the authors validate their hypothesis using only the Quantile-Quantile plots, a graphical check of goodness-of-fit. Note that the datasets used in previous studies were collected before 2016. Using data collected in 2022, a year in which the 4 G network is the dominant cellular technology in Europe, we aim to investigate whether the proposed models are still valid. In addition, we extend the previous studies to multi-band cell scenarios.

Few studies have attempted to model the connection duration and size using cellular measurements. The authors of [14] model of the duration and average throughput of a user session. However, this study is based on data from 3 G networks. The authors of [15] analyze the behavior of users in

[1] In this paper, we use a 24-hour clock notation.

terms of duration and data volume of a session. This study uses measurements from a 4G network. Nevertheless, no statistical model is proposed. The authors of [23] show the effects of time of day on session size by analyzing the Transmission Control Protocol (TCP) flows obtained from the core of a 4G network. In contrast to these studies, our approach is focused on the Radio Resource Control (RRC) connection level. Furthermore, to the best of our knowledge, in the literature, no statistical models have been proposed to characterize the connection duration and size in 4G networks.

3 Control channel in LTE

In LTE, downlink and uplink transmissions are structured into subframes of 1-millisecond duration. Each downlink subframe is divided into a control region and a data region.

The radio interface is structured in physical resource blocks (PRBs), each consisting of 12 consecutive subcarriers in the frequency domain (i.e., 180 kHz). Therefore, the number of PRBs per subframe depends on the system bandwidth, varying between 6 (for 1.4 MHz) and 100 (for 20 MHz). PRBs are shared among the terminals connected to the same cell.

A Physical Cell Identity (PCI) is assigned to each cell. It is locally unique and is used to compute the specific scrambling code that is applied to transmissions in the cell. Its role is to ensure that the neighbor transmissions have the same spectral characteristics as noise.

The control region always starts at the first Orthogonal Frequency Division Multiplex (OFDM) symbol of each 1-ms subframe. The size varies dynamically depending on the traffic situation, occupying one, two, or three OFDM symbols (two, three, or four for bandwidths less than 10 MHz). The control region is occupied by several physical control channels, among which the Physical Control Format Indicator Channel (PCFICH) and the Physical Downlink Control Channel (PDCCH) are the concern of our study. The PCFICH indicates the size of the control region (in OFDM symbols). The PDCCH is used to transport the DCI, which carries information about downlink scheduling assignments and uplink scheduling grants for a single terminal or a group of terminals.

3.1 Downlink control information in LTE

In addition to information about the location of the PRBs allocated to a terminal, DCI messages contain information such as the modulation and coding scheme (MCS) index, information related to spatial multiplexing, and power control information for the uplink. A DCI message is generally addressed to a single terminal. However, there are also DCIs addressed to all terminals in the cell. There are different types of DCI formats for different usages [24, Chapter 10]: format 0 is used for the uplink grant, and formats 1/1A/1C/2/2A are used for the downlink resource assignment.

3.2 Resources of the control channel in LTE

A PDCCH carries a DCI message, usually addressed to a specific terminal. Since the eNB must share resources among multiple terminals within the cell, multiple simultaneous PDCCH can be transmitted within each subframe.

The eNB uses a Radio Network Temporary Identifier (RNTI) to identify the destination of a PDCCH. The RNTI has a length of 16 bits and thus a range of values between 0 and 65535. There are different types of RNTIs; the most commonly used are: C-RNTI (between 1 and 65523) which is used to identify a specific terminal, P-RNTI (65534) for paging messages, SI-RNTI (65535) for system information, RA-RNTI (between 1 and 60) for random access responses.

When a PCDDH is transmitted, the eNB does not explicitly indicate the RNTI of the receiver. Instead, the cyclic redundancy check (CRC) of the PDCCH is scrambled by the RNTI of the addressed terminal.

The terminal uses a procedure named blind decoding to decode the control messages sent by the eNB. It descrambles the CRC using its RNTI and performs the CRC check. If the check result is successful, the terminal decodes the DCI message.

Once the RRC connection is established, the base station assigns an RNTI to the terminal (see Sect. 3.4 and Fig. 3). Hence, an RNTI is a unique identifier of the RRC connection within the cell. In every sub-frame, the base station informs connected terminals of the downlink scheduling assignment (and the MCS index) for the current subframe n and the uplink scheduling grant (and the MCS index) for the subframe $n+4$. As an example, Fig. 1 shows the resource allocation within a cell in which five terminals are connected to the network. In the PDCCH region of subframe n, the base station transmits five DCI f_i, where i represents the DCI format. In subframe n, the PRBs 0–4, 7–10, and 11–13 in the downlink are reserved for terminals 51, 52, and 53, respectively. In subframe $n+4$, the PRBs 2–5 and 10–13 in the uplink are reserved for terminals 51 and 53, respectively.

3.3 Multi-band LTE system

MNOs deploy multi-band systems (different frequency bands in one cell) to increase network capacity. The coverage of each frequency band depends on the MNO strategy. Generally, a low-frequency band such as 800 MHz has wider coverage, and thus it is used at the cell edge. On the other hand, a high-frequency band such as 2600 MHz exhibits a shorter range, and thus it is used to provide coverage to terminals located near the base station. The base station indicates

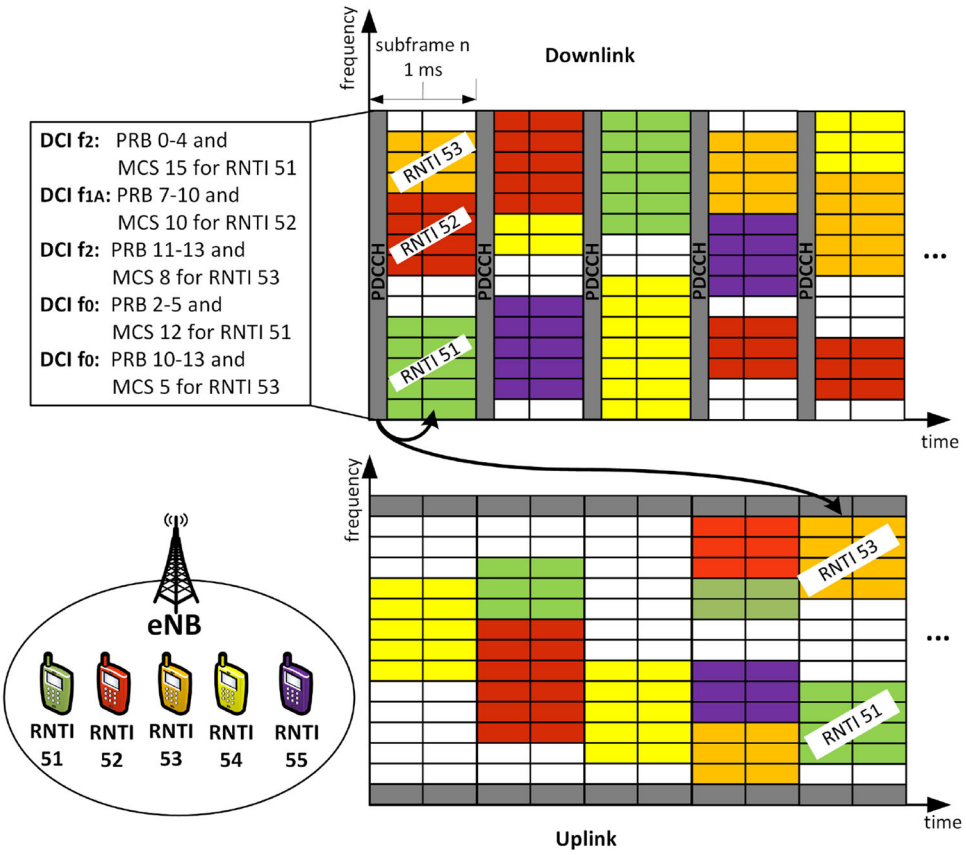


Fig. 1 PRB allocation, network with five active terminals

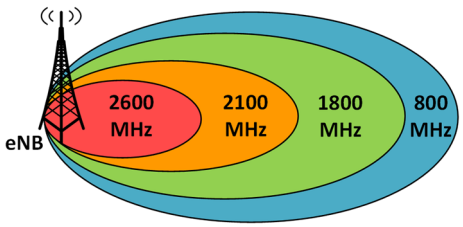


Fig. 2 Multi-band LTE system with four frequency bands, 2600 MHz, 2100 MHz, 1800 MHz, and 800 MHz

the selection priority of each frequency band in the system information block 3 (SIB3). Figure 2 shows a multi-band LTE system with four different frequency bands, where the priority order from highest to lowest is 2600 MHz, 2100 MHz, 1800 MHz, and 800 MHz. Due to the priority order, terminals near the base station, within the coverage area of the four frequency bands, select the 2600 MHz band. The terminals located at the cell edge (or indoor) can only use the 800 MHz band since they are not within the coverage area of the other bands.

In a multi-band LTE system, it is important to analyze all the frequency bands used in the cell to compute the total number of connections in the cell, the cell achieved throughput, and the utilization of radio resources within the cell. In addition, analyzing all the frequency bands allows us to identify aspects that do not depend on the frequency band and aspects in which the frequency band has an influence.

3.4 RNTI management

This section illustrates the evolution of the type of RNTI used by a terminal during the connection. Fig 3 shows the sequence of messages exchanged between the terminal and the eNB. Synchronization is the first procedure that a terminal must perform to connect with the eNB. The terminal then decodes the Master Information Block (MIB), which contains the system information. The eNB broadcasts the MIB using the physical layer channel called Physical Broadcast Channel (PBCH). The remainder of the system information is found in the SIBs, which are periodically transmitted by the eNB. The terminal uses the SI-RNTI to decode SIB messages. The eNB also broadcasts paging messages that the terminal must decode using the P-RNTI.

If the terminal has any information to transmit, or if the network requests it, it performs the random access procedure by sending the random access preamble message. RA-RNTI is used to decode the DCI corresponding to the eNB response (Random Access Response message). The random access response contains, among other parameters, the temporary C-RNTI (TC-RNTI), the preamble sent by the terminal, and

the timing advance. TC-RNTI is used to complete the contention process, and it becomes C-RNTI if the connection is successful. C-RNTI is a unique and temporary identification used by a terminal within a cell.

Once the connection is established, the terminal will use the C-RNTI until the end of the connection. There are three possible scenarios:

- The connection is established to transmit only signaling messages (see case 1 of Fig 3). This type of connection is very short (e.g., tracking area update).
- The connection is established to transmit user data, and it is terminated due to a handover (see case 2 of Fig 3). The duration of this type of connection is variable and depends on the period the terminal camps on a given cell. This scenario corresponds to a UE that moves and a long RRC connection.
- The connection is established to transmit user data and is terminated due to user inactivity (see case 3 of Fig 3). The RRC inactivity timer parameter fixes the maximum inactivity allowed in the cell. This parameter is not transmitted to the terminals. This scenario corresponds to either a fixed UE (or with low mobility) or a short RRC connection.

When the connection is terminated, the eNB can reassign a C-RNTI to another connection.

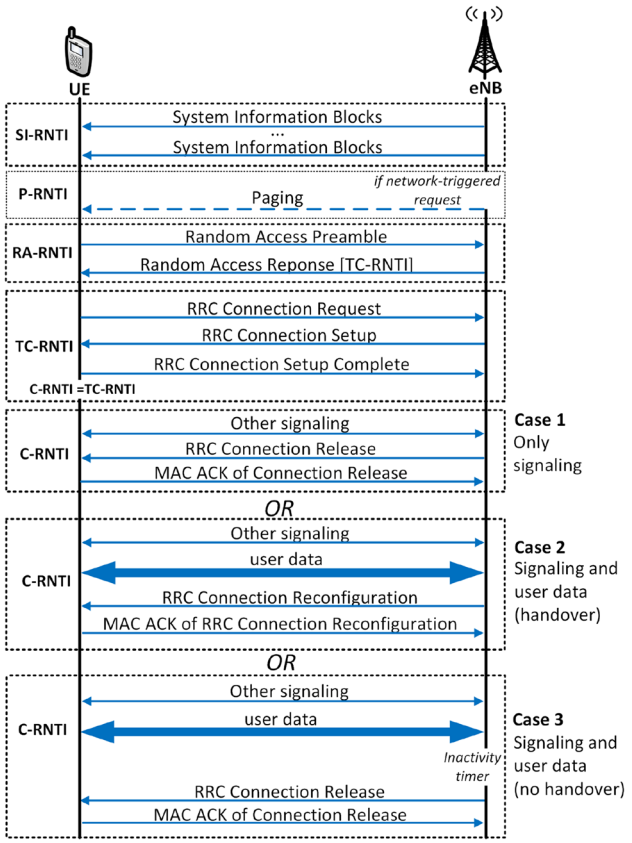


Fig. 3 Sequence of messages during a connection in LTE. A UE can establish a connection to transmit only signaling messages or signaling and user data

4 LTE analyzer

4.1 Architecture overview

We developed an LTE analyzer based on srsRAN, an open-source 4G and 5G library, which can be used to build a mobile network using SDRs. srsUE is one of the components offered by the srsRAN library; it can emulate the main functionalities of a UE. Moreover, it supports different radio front ends, including Ettus USRP B210 and BladeRFx40. We strived to minimize the modifications of srsRAN and to develop specific functions in a very limited number of modules, to simplify porting them to a new version of the standard srsRAN software.

The objective of our LTE analyzer is to decode all the messages transmitted in clear text by the eNB, including system information, random access responses, and DCI messages addressed to UEs within the cell. Our goal is not real-time processing and visualization; rather, we want to monitor the network over long periods (weeks). The LTE analyzer consists of an SDR connected to a computer executing the modified srsUE, as shown in Fig. 4. The analyzer returns two files, one containing the system information of

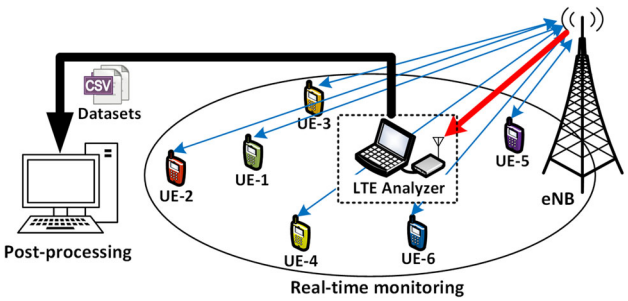


Fig. 4 Overview of the measurement scenario: data is collected in real-time, locally recorded, and then processed offline

the monitored cell and another containing the dataset (one-millisecond granularity) of DCIs transmitted by the eNB.

4.2 Monitoring

As mentioned in [24], a terminal only searches DCI messages in a specific search space of the control region, which depends on the RNTI and the current subframe number. We modified the srsUE to monitor all the control messages inside the cell. The main modifications are: (i) making srsUE a non-intrusive LTE analyzer, (ii) the implementation of a new function that

decodes the whole DCI messages, and (iii) minimizing the loss of subframes.

Every subframe (every millisecond), our LTE analyzer looks for DCI messages transmitted by the eNB. If it finds one or more DCIs, it saves in a comma-separated values (CSV) file the DCI payload and the timestamp (in milliseconds). Otherwise, it saves only a timestamp. Poor channel quality between our LTE analyzer and the eNB can cause synchronization loss and thus lost subframes. In this case, our LTE analyzer does not save any timestamp. Therefore, the total number of missing subframes equals the number of missing timestamps in the CSV file.

4.3 Post-processing

Post-processing consists in generating a new CSV file where each row is the decoded information of a DCI message. The columns of this CSV file contain, among other information, a Unix timestamp in milliseconds, the RNTI associated with the DCI message, the DCI message format, the number of allocated PRBs, the MCS index of the first transport block, the transport block size (TBS) of the first transport block, the MCS index of the second transport block, and the TBS of the second transport block. The TBS is obtained as a function of the MCS index and the number of PRBs.

The RNTI is a unique connection identifier in the cell. However, the same RNTI value can be reused t seconds after the connection release for a different UE, where t is greater than the RRC inactivity timer. Therefore, DCI messages containing the same RNTI are part of the same connection only if the inactivity period is less than the RRC inactivity timer. The inactivity period is the time difference between two consecutive DCIs containing the same RNTI.

5 Measurement dataset

We performed three one-week measurement campaigns in December 2021, January 2022, and February 2022. The weeks selected do not include school vacations, holidays, or lockdown periods. The results shown in this paper correspond to the last measurement campaign. However, we obtain similar qualitative results in the other measurement campaigns.

The measurements were collected in a commercial three-sector eNB of the SFR MNO in the city center of Rennes. Rennes is a medium-sized city in north-western France. The metropolitan area includes 50 km^2 with 220,488 inhabitants and more than 66,000 students. It counts approximately 1014 cells covered by the four French MNOs (SFR, Orange, Bouygues, and Free). The locations of base stations for each MNO are available on a public website (www.cartoradio.fr). We monitor a sector in which there are apartment buildings,

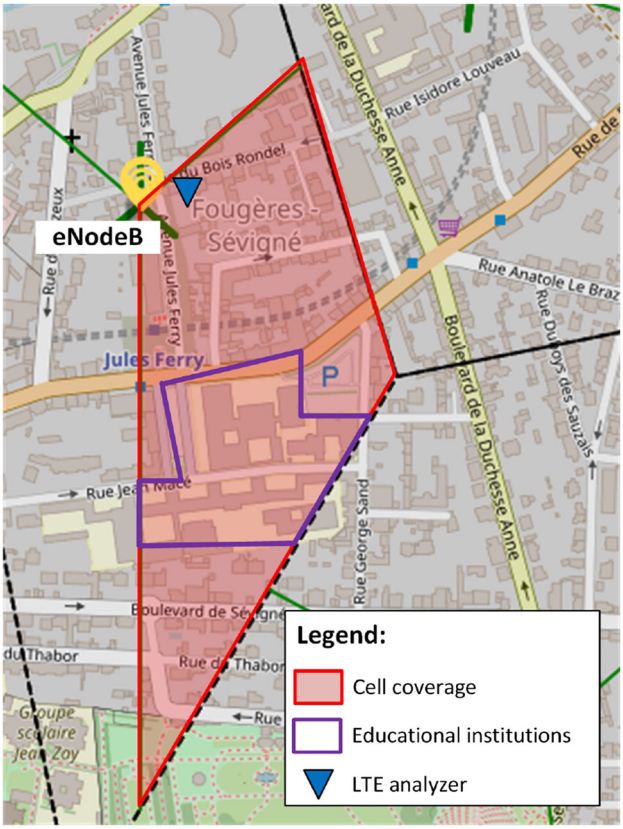


Fig. 5 Monitored cell, SFR three-sector eNB in Rennes, France. The LTE analyzer is located 30 ms from the eNB

educational institutions, and shops (see Fig. 5). The sets of LTE analyzers were located less than 30 meters from the eNB. Note that the terms sector and cell are used interchangeably in this work.

We verified that the MNO operates in four different frequency bands in the monitored sector. Hence, we used four LTE analyzers in parallel. Table 1 summarizes the cell information parameters obtained during the measurement campaign. The analysis period is between Saturday, February 19, and Friday, February 25, 2022, corresponding to 6048×10^6 subframes. We verify that we missed at most 0.03% of the subframes in each frequency band. All four frequency bands use the same PCI, confirming that the LTE analyzers monitored the same sector. During a week of monitoring, the 800 MHz, 1800 MHz, 2100 MHz, and 2600 MHz bands generated CSV files of 40 GB, 70 GB, 30 GB, and 25 GB, respectively. We analyze all these datasets.

6 LTE network analysis

In this section, we examine characteristics related to the use of the network (network utilization, the connection setup rate, and the cell achieved throughput). Unlike previously

Table 1 Information of the monitored cell

DL frequency (MHz)	806	1835	2117.8	2627.5
EARFCN	6300	1501	78	2825
Band (MHz)	800	1800	2100	2600
PCI	127	127	127	127
Duplex mode	FDD	FDD	FDD	FDD
Number of PRBs	50	100	75	75
Number of sub-frames $\times 10^6$	6048	6048	6048	6048
Missing sub-frames	0.02%	0.03%	0.01%	0.01%
Dataset size (GB)	40	70	30	25

published results, our approach captures information at the physical layer level by monitoring the PDCCH region. This approach allows us to obtain additional information, such as the number of allocated PRBs. Moreover, we present quantitative results for each frequency band in uplink and downlink. This can be interesting to develop algorithms of load adaptation and switch on/off frequency bands.

6.1 PRB utilization on the downlink

Given that signaling and user data share the same radio resources and are encrypted after the RRC connection setup, it is not straightforward to differentiate between them. Signaling messages exchanged before the RRC connection setup can be easily identified because they have specific RNTIs. This includes system information messages, paging messages, and random access responses, which are bound to SI-RNTI, P-RNTI, and RA-RNTI, respectively (see Sect. 3.2). The objective of this section is thus to analyze how many PRBs are used for signaling outside RRC connections and compare it with the number of PRBs used for data and RRC signaling. We compute the network utilization, defined as the number of PRBs used for a given function divided by the number of PRBs available in the frequency band (see Table 1). Our analysis considers only downlink resources.

Figure 6 compares the 1-hour moving average of the PRB utilization for each RNTI type. We remark that the amount of resources used to transmit the system information (SI-RNTI) does not show daily patterns and is almost time-invariant. On the contrary, we can observe diurnal utilization patterns for the other RNTI types (C-RNTI, P-RNTI, and RA-RNTI). There is a significant difference between working hours and early morning. Furthermore, network usage decreases on weekends (February 19 and 20). The PRB utilization peak hours are between 10:00 and 12:00 and the off-peak hours are between 4:00 and 6:00.

Table 2 summarizes the average PRB utilization by type of RNTI during peak and off-peak hours. This analysis does not consider the weekend. During peak hours, in the 1800 MHz band, the average utilization of the system information (SI-RNTI), paging information (P-RNTI), and random access responses (RA-RNTI) represent respectively 1/170, 1/1150, and 1/4100 of the resources used by user data (C-RNTI).

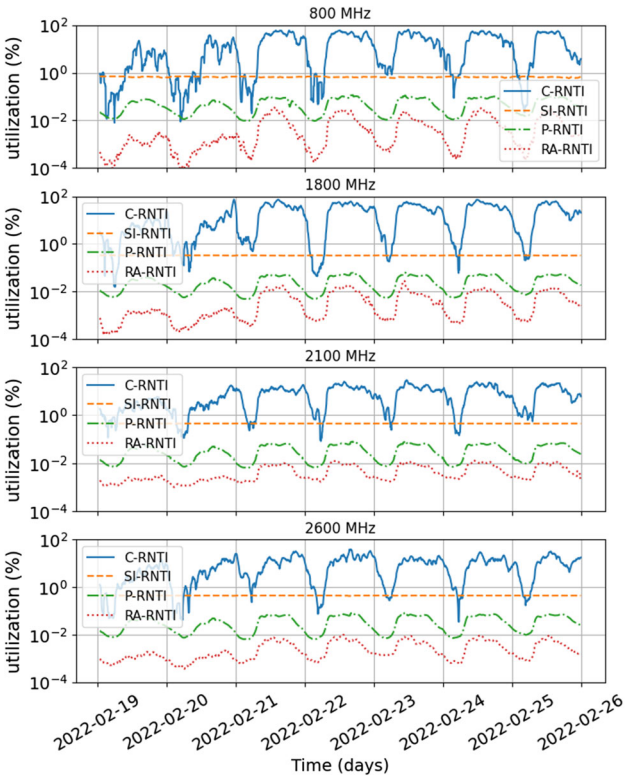


Fig. 6 PRB utilization (in log scale) by RNTI type, user data (C-RNTI), system information messages (SI-RNTI), paging messages (P-RNTI), and random access responses (RA-RNTI). Monitoring begins on Saturday, February 19, 2022

As far as user data is concerned, during peak hours, the average utilization (1-hour moving average) is 50%, 55%, 19%, and 15% of the available resources in the 800 MHz, 1800 MHz, 2100 MHz, and 2600 MHz bands, respectively. Note that there are peaks of utilization greater than 90%.

The 1800 MHz and 800 MHz bands are the most used, while the least used is the 2600 MHz band. The 1800 MHz band is four times more occupied than the 2600 MHz band. In a commercial multi-band LTE system, the utilization of

Table 2 Average PRB utilization on downlink

Band (MHz)	RNTI	4:00–6:00		10:00–12:00	
		Mean (%)	Std. Dev. (%)	Mean (%)	Std. Dev. (%)
800	C-RNTI	0.612	1.270	50.327	12.577
	SI-RNTI	0.665	0.039	0.653	0.018
	P-RNTI	0.013	0.002	0.087	0.014
	RA-RNTI	0.0004	0.0004	0.021	0.007
1800	C-RNTI	0.378	0.775	55.489	11.639
	SI-RNTI	0.329	0.003	0.323	0.007
	P-RNTI	0.006	0.001	0.048	0.008
	RA-RNTI	0.001	0.0002	0.014	0.006
2100	C-RNTI	0.738	1.139	19.286	7.928
	SI-RNTI	0.441	0.001	0.438	0.001
	P-RNTI	0.009	0.001	0.060	0.011
	RA-RNTI	0.002	0.0003	0.011	0.002
2600	C-RNTI	0.466	0.799	15.531	5.764
	SI-RNTI	0.441	0.002	0.439	0.002
	P-RNTI	0.009	0.001	0.066	0.010
	RA-RNTI	0.001	0.0003	0.007	0.002

a given frequency band depends mainly on the number of terminals camping on that frequency band. The number of terminals is related to the coverage area, which depends on the MNO strategy (see Sect. 3.3). The 2600 MHz and 2100 MHz bands offer a small coverage area since only terminals with good channel quality can camp on these frequency bands. On the other hand, the 1800 MHz and 800 MHz bands provide a wider coverage area and thus higher network utilization. The average utilization in the off-peak hours falls to 0.6%, 0.3%, 0.7%, 0.4% in the 800 MHz, 1800 MHz, 2100 MHz, and 2600 MHz bands, respectively.

To summarize, the total signaling (excluding RRC) is less than 1% of the total capacity. This validates studies that considered medium to high load conditions in which the signaling load is neglected compared to user data.

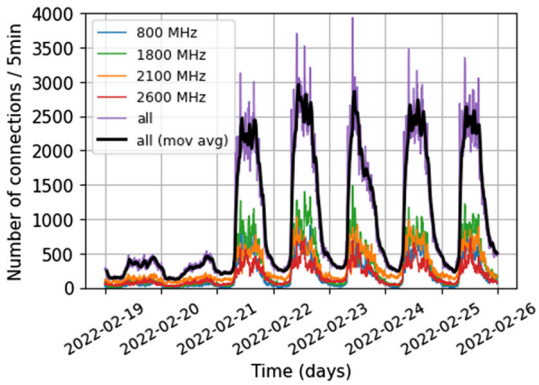


Fig. 7 Number of connections per 5 min. The 800 MHz, 1800 MHz, 2100 MHz, and 2600 MHz bands have a bandwidth of 10 MHz, 20 MHz, 15 MHz, and 15 MHz, respectively. Monitoring begins on Saturday, February 19, 2022

6.2 RRC connection set up rate

This section illustrates the daily and weekly patterns in the number of RRC connections for each frequency band. A connection can start and end anytime, depending on the user. To represent the variation of the number of connections over time, we split the time into 5-minutes intervals. Figure 7 shows the number of connections starting between T and $T + 5$ for each band as well as the total, where T represents a time instant in minutes. The black lines represent the moving average of the total number of connections over a time window of 1 h.

There is an important difference between weekdays and weekends (February 19 and 20). These results could be due to the presence of two educational establishments within the monitored cell. We also observed a significant difference between working hours and early mornings. On weekdays, there are peaks of more than 3500 connections per 5 min, that is, more than 11.6 connections per second per cell. On the other hand, on the weekend, the peaks are around 500 connections per 5 min, that is, 1.6 connections per second.

Figure 8 shows the daily pattern of the total number of connections (i.e., adding the four bands) per 5 min. We compute the moving average over a window of 1 h (i.e., averaging 12 consecutive five-minute periods). On weekdays, the least active hours are between 4:00 and 6:00, with between 200 and 300 connections per five-minute interval. On the other hand, the hours with more connections are between 10:00 and 12:00, with between 2200 and 3000 connections per

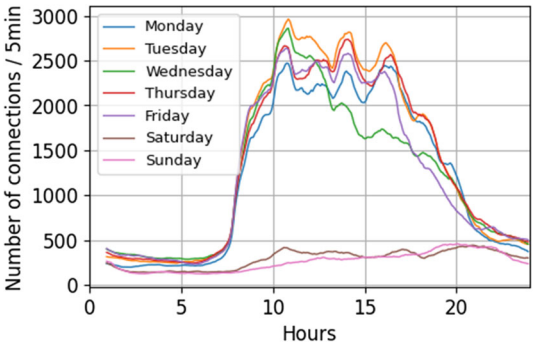


Fig. 8 Daily number of connections per 5 min, 1-hour moving average

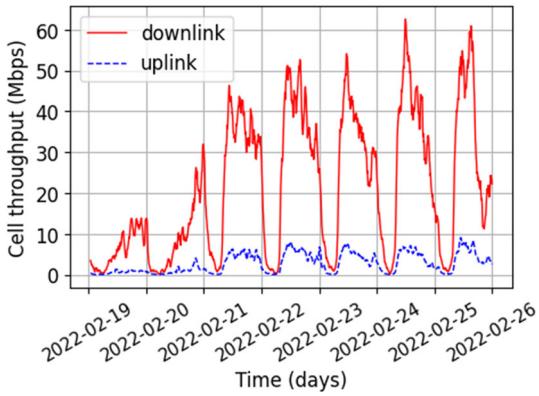


Fig. 9 UL and DL cell achieved throughput, 1-hour moving average. Monitoring begins on Saturday, February 19, 2022

five-minute interval. We observe that the daily trend during weekdays is similar (except Wednesday), especially between midnight and 10:00. Some peaks are repeated every day at the same time. These peaks could be due to daily routines in educational institutions, such as starting courses, breaks, and lunchtime.

The weekday daily pattern can be divided into four well-defined periods: (i) a stable period with few connections between midnight and 07:00, (ii) a period with a rapid increase followed by a peak between 7:00 and 10:00, (iii) a period with significant activity with a downward trend between 10:00 and 18:00, (iv) a period where the number of connections drops off rapidly between 18:00 and midnight.

6.3 Cell achieved throughput

The analysis carried out in the previous sections illustrates the utilization of the cellular network and the number of connections. However, these parameters do not indicate the amount of traffic (in bits) passing through the eNB. This section analyzes the daily pattern of the cell achieved throughput in both directions (uplink and downlink). Additionally, we analyze the ratio between uplink and downlink cell achieved throughput. The cell achieved throughput refers to the number of bits per second flowing through the eNB. The number of bits in a 1-ms subframe is the sum of the transport block sizes in a given direction.

Figure 9 displays the cell achieved throughput (i.e., adding the four bands) in downlink and uplink considering a 1-hour moving average. We can distinguish a diurnal pattern during the weekdays despite high variation over time. On the weekend, we observe a significant reduction in the achieved throughput. On weekdays between 10:00 and 20:00, the average downlink achieved throughput varies between 30 and 63 Mbps, while uplink achieved throughput varies between 3 and 9 Mbps. During off-peak hours (5:00–6:00), the average downlink achieved throughput is less than 5 Mbps, and the average uplink achieved throughput is less than 2 Mbps. Note that we present the results considering a 1-hour moving average. However, by reducing the window size, we can see instants in which the average cell achieved throughput reaches values greater than 25 Mbps and 50 Mbps in uplink and downlink, respectively.

Table 3 presents the downlink achieved throughput of two periods of the day averaged over 1 week of monitoring (excluding the weekend). During peak hours (10:00–12:00) and off-peak hours (4:00–6:00), the average downlink achieved throughput is approximately seven times the average uplink achieved throughput. On the other hand, the average downlink achieved throughput during peak hours is at least 45 times that during off-peak hours.

Figure 10 shows the ratio between uplink and downlink traffic. Although the uplink and downlink traffic have a daily and weekly pattern, we observe that the uplink to downlink ratio does not have a daily pattern. The ratio varies in most cases between 0.1 and 0.2, with some peaks outside this range. The mean ratio during the week is 0.15, and the standard deviation is 0.17. This means that the downlink traffic is approximately seven times as much as the uplink traffic. Excluding peaks greater than 0.2, the mean ratio is 0.11, and the standard deviation is 0.03. That is, the downlink traffic is approximately eight times as much as the uplink traffic.

7 Modeling LTE connections

In the literature, there is a lack of statistical models that describe user connections based on data collected in commercial cellular networks. In this section, we identify the statistical models that best characterize user connections. Such models can be used for optimization and more realistic simulations of cellular networks.

Table 3 Average cell achieved throughput

Direction	4:00–6:00		10:00–12:00	
	Mean (Mbps)	Std. Dev. (Mbps)	Mean (Mbps)	Std. Dev. (Mbps)
DL	1.09	0.98	49.87	9.93
UL	0.16	0.20	7.17	1.73

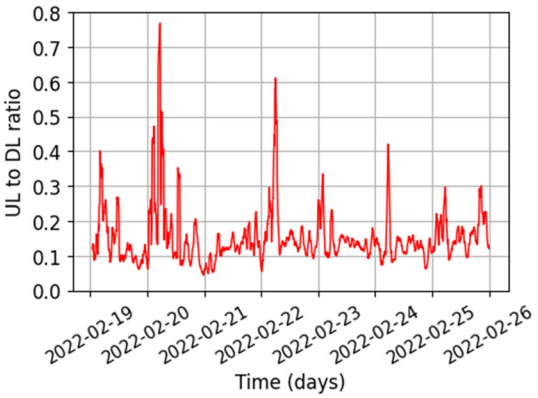


Fig. 10 UL to DL cell achieved throughput ratio, 1-hour moving average. Monitoring begins on Saturday, February 19, 2022

Table 4 Theoretical distributions

Distribution	Cumulative distribution function
Exponential	$F(x) = 1 - \exp(-\lambda x)$
Weibull	$F(x) = 1 - \exp(-(x/\beta)^{\alpha})$
Lognormal	$F(x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{\ln(x)-\mu}{\sigma\sqrt{2}}\right]$
Pareto	$F(x) = 1 - (x_m/x)^{\alpha}$
Beta	$F(x) = I_x(\alpha, \beta)$
Normal	$F(x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{x-\mu}{\sigma\sqrt{2}}\right]$
Rayleigh	$F(x) = 1 - \exp(-x^2/2\sigma^2)$

7.1 Statistical modeling approach

To determine whether we can model the quantities mentioned above with well-known distributions, we proceed as follows: first, we estimate the parameters of each distribution with the standard Maximum Likelihood Estimates (MLE) method [25, 26]; second, we evaluate the goodness of fit with the Kolmogorov-Smirnov [25, 27] and the Cramer-Von Mises [28, 29] tests, with a confidence level of 95%. We use the Scipy library [30] to perform both parameter estimation and evaluation tests. Table 4 gives the cumulative distribution function (CDF) of the list of the distributions that we considered. Note that the exponential distribution is a particular case of the Weibull distribution when $\alpha = 1$ and $\beta = 1/\lambda$.

Since the connection arrival process can be presumed to be stationary within short periods [16, 20], we split the dataset into shorter periods. Moreover, the analysis of short periods allows us to evaluate the variation of user behavior as a function of the time of day. In each period, we fit the data to a given distribution, and then the distribution is validated using the goodness-of-fit tests. If a distribution passes the tests in more than 90% of the periods, we consider it a valid distribution. This analysis considers only the connections established on working days since the use of terminals on weekends shows a different pattern than on working days.

7.2 Connection inter-arrival time

In this section, we study the connection inter-arrival time, which is the time difference between the start of two consecutive connections.

Let I be a random variable representing the connection inter-arrival time in seconds. Working hours (8:00–20:00) are divided into periods of 20 min. During non-working hours, we consider 2-hour periods since there are few connections per minute. Each frequency band has 210 periods (42 per weekday). Therefore a total of 840 periods were fitted. The average number of connections per period is 2132. Table 5 summarizes the results of the curve fitting validation. The results show that Weibull, Beta, and exponential distributions best fit the data. The Weibull distribution passes the Kolmogorov-Smirnov test and the Cramer-Von Mises test in 97% and 96% of the periods, respectively, while the exponential distribution in 92% and 90% of the periods, respectively. The results also reveal that Weibull is the best fit in most cases. Thus we can model the connection inter-arrival time as follows:

$$I \sim \text{Weibull}(\alpha, \beta)$$

The mean α parameter (see Table 4) is 0.99, and the coefficient of variation (the ratio of standard deviation to mean) is 0.03. This means that the α parameter of the Weibull distribution is almost constant and approximately equal to 1; thus, the inter-arrival time can be modeled as an exponential distribution. The mean β is 1.29, and the coefficient of variation is 2.03. In other words, the β presents higher variability. This can be explained by the fact that the number of users varies depending on the time of day and the frequency band since each one has a specific coverage area.

As an example, Fig. 11 compares the empirical distributions and the fitted Weibull distribution during the peak hour of a typical working day (Tuesday, February 22) between

Table 5 Inter-arrival time, percentages of goodness-of-fit tests passed with a confidence level of 95%

Distribution	Cramer-Von Mises		Kolmogorov-Smirnov		Valid
	Passed (%)	Best-fit (%)	Passed (%)	Best-fit (%)	
Exponential	91.55	17.26	90.00	18.81	✓
Weibull	97.02	49.17	96.43	49.64	✓
Lognormal	0.48	0.24	0.48	0.24	×
Pareto	0	0	0	0	×
Beta	96.43	31.19	95.95	28.45	✓
Normal	0	0	0	0	×
Rayleigh	0	0	0	0	×

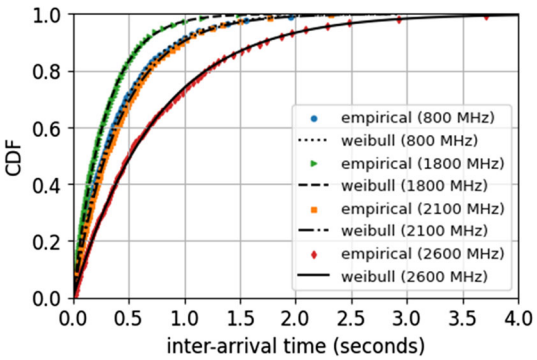


Fig. 11 Comparison of empirical data and fitted curves of inter-arrival time (Weibull distribution). Thursday, February 22, 2022, 11:00–11:20

11:00 and 11:20, i.e., the connections that start between 11:00 and 11:20. The four fitted Weibull distributions pass the Cramer-Von Mises and Kolmogorov-Smirnov tests. From the figure, we can see that the fitted curves closely match the empirical data. Furthermore, the fitted curves have different parameters depending on the frequency band. In the 1800 MHz band, the inter-arrival times are shorter, while in the 2600 MHz band, they are longer. This is because, during peak hours, the 1800 MHz and 2600 MHz bands receive the highest and lowest number of users, respectively. This is consistent with the fact that generally, only terminals close to the eNB camp on the higher frequency band.

Table 6 summarizes the statistical analysis of the empirical data, as well as the results of the curve fitting. The mean inter-arrival time is less than 0.5 s in the 800 MHz, 1800 MHz, and 2100 MHz bands, and approximately 0.7 s in the 2600 MHz band. Note that the α parameter of the Weibull distribution is approximately 1.

7.3 Connection duration

In this study, we define the duration of a connection as the time between the establishment of the connection (RRC connection setup message) and the release of the allocated resources. Hence, the connection duration is the time difference between the first and the last DCI message received by the terminal.

Figure 12 shows the empirical CDF (ECDF) of the connection duration at four different times of the day for each frequency band. For the sake of simplicity, the figure depicts only four representative one-hour periods. We can see that in 90% of cases, the duration of the connections is less than 2 min.

In the four bands, the duration of a large percentage of connections is between 10 and 11 s, similar to what was previously reported [15]. The authors explain that this is due to the inactivity timer parameter set by the MNO. In many connections, the terminal activity is very short (less than 1 s); however, the connection is maintained until the inactivity timer expires. Hence there is a peak of connections around the inactivity timer parameter. Connections shorter than the inactivity parameter can correspond to signaling exchanges or handovers. The figure shows that the inactivity timer parameter is set to 10 s in the monitored cell.

Between 10:00 and 23:00 (green, orange, and red curves), the percentage of short connections (less than 10 s) reaches almost 40% in the 800 MHz band, while in the 1800 MHz, 2100 MHz, and 2600 MHz bands is less than 20%. These results are consistent because this type of connection is mainly due to handovers. Mobile operators generally use low-frequency bands such as 800 MHz to cover the cell edge; thus, there is a greater probability that the terminals change serving cell. In all frequency bands, we notice that the percentage of short connections decreases significantly between 4:00 and 5:00 (blue curves). The reason is that there is less user mobility in the early morning and therefore fewer handovers.

Between 10:00 and 23:00 (green, orange, and red curves), we observe that between 20 and 30% of connections are between 10 and 11 s long. These percentages are even higher from 4:00 to 5:00. Without taking handovers into account, a duration between 10 and 11 s means that the terminal is active for less than 1 s and then remains inactive for 10 s, which triggers a connection release. This typically corresponds to a terminal receiving a message without the user taking further

Table 6 Connection inter-arrival time model, estimated parameters during peak hour

Band (MHz)	Mean (s)	Std. Dev. (s)	Median (s)	Weibull parameters
800	0.40	0.41	0.27	$\alpha = 0.97,\ \beta = 0.40$
1800	0.28	0.28	0.20	$\alpha = 1,\ \beta = 0.28$
2100	0.43	0.43	0.29	$\alpha = 0.99,\ \beta = 0.42$
2600	0.74	0.75	0.50	$\alpha = 0.99,\ \beta = 0.74$

Thursday, February 22, 2022, 11:00–11:20

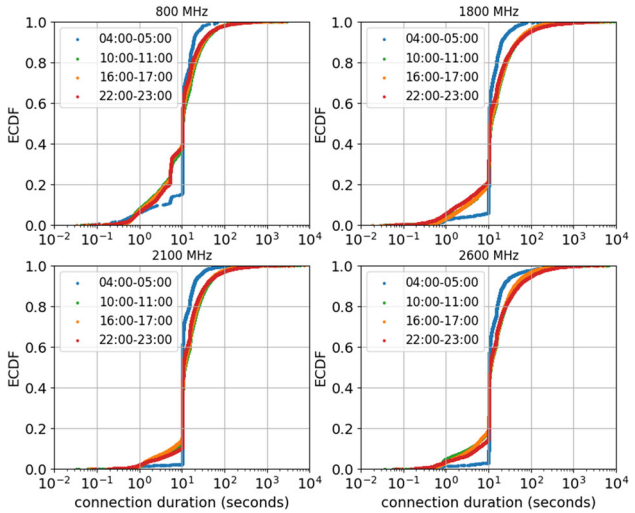


Fig. 12 Empirical CDF of the connection duration (x-axis, log scale), considering only working days

action. It could also be due to background traffic, in other words, the autonomous exchange of data between the terminal and the network, which is generally characterized by the transmission of small packets (less than 1000 Bytes [31]) and thus a short activity duration. As most people sleep between 4:00 and 5:00, the part of background traffic compared to the user-generated traffic is more significant than in the activity periods [32]. This trend is clearly visible for every frequency band in Fig. 12.

The figure reveals that, in the periods 10:00–11:00, 16:00–17:00, and 22:00–23:00, the distributions of the connections with a duration greater than 11 s are very similar in the four frequency bands. The results suggest that during working hours, the ECDF of the connections with a long duration (longer than the inactivity timer) is almost independent of the frequency band and the time-of-day effects.

7.3.1 Modeling connection duration

We model the duration of the user-initiated connections, that is, without considering connections established by background traffic. We assume that a connection is user-initiated if the terminal remains active for at least 6 s. Since the inactivity timer is set to 10 s by the eNB, we only model the distribution of connections longer than 16 s. It should be noted that user-initiated connections terminated in less than 16 s due to a handover are not considered in the analysis.

Let D be a random variable that represents the duration in seconds of long user-initiated connections (duration greater than the inactivity timer parameter plus 6 s). We find the best fitting distribution for the connection duration at different times of the day. Finding a distribution with a good fit is difficult because few connections are longer than 16 s during non-working hours. Hence, we only consider working hours (i.e., between 8:00 and 20:00). We analyze 36 different 20-minute periods for each working day. There are 180 periods for each frequency band; thus, we fit a total of 720 periods, with an average of 668 connections per period. Since we only consider connections longer than 16 s, an additional location parameter is included in the model. The location parameter shifts the distribution on the x-axis. Table 7 summarizes the results, revealing that, in 99.44% and 98.47% of periods, the lognormal distribution passes the Cramer-Von Mises test and the Kolmogorov-Smirnov test, respectively. Furthermore, it is the best-fitting distribution in 55.69% and 55.97% of the cases according to the Cramer-Von Mises test and the Kolmogorov-Smirnov test, respectively. The Pareto is also a valid distribution. Nevertheless, the α parameter (see Table 4) of the fitted curves varies between 1.1 and 1.6; therefore, the variance is infinite.

We can conclude that lognormal distribution is the best candidate to model the duration of long user-initiated connections. From the parameters of the fitted curves, we note that the location parameter varies between 15.5 and 16.1. To reduce the number of parameters of the statistical model, we fix the location parameter of the lognormal distribution to 15.7 s. This value provides the best results when we evaluate theoretical distributions. The results are minimally affected. The lognormal distribution passes the Cramer-Von Mises test and the Kolmogorov-Smirnov test in 99.86% and 99.58% of periods, respectively. Thus we have:

$$D = 15.7 + X,$$

where $X \sim \text{Log-}\mathcal{N}(\mu, \sigma)$.

Table 7 Connection duration, percentages of goodness-of-fit tests passed with a confidence level of 95%

Distribution	Cramer-Von Mises		Kolmogorov-Smirnov		Valid
	Passed (%)	Best-fit (%)	Passed (%)	Best-fit (%)	
Exponential	0	0	0	0	×
Weibull	6.81	0.28	8.61	0.69	×
Lognormal	99.44	55.69	98.47	55.97	✓
Pareto	94.03	43.75	91.39	42.50	✓
Beta	7.78	0.28	6.53	0.42	×
Normal	0	0	0	0	×
Rayleigh	0	0	0	0	×

The lognormal distribution has two parameters, the shape parameter σ (standard deviation of the logarithm of the distribution) and the scale parameter $m = \exp(\mu)$, which is the median of the distribution. The average and the coefficient of variation of m are 11.59 and 0.16, respectively. The average and the coefficient of variation of σ are 1.46 and 0.05, respectively. This means that the variability of the lognormal distribution parameters that model the connection duration is low between 8:00 and 20:00. In other words, the time of day and the frequency band do not significantly affect the distribution of the duration of the user-initiated connections. These results can be explained by the fact that the activity time of a terminal depends exclusively on the user. However, when a terminal changes its serving cell (handover), the connection in a given frequency band is interrupted before the user completes the activity. This reduces the connection duration in the monitored frequency band. The handover occurrence probability varies according to the frequency band. For this reason, there is a slight difference between the frequency bands.

Figure 13 compares the empirical distributions and the fitted lognormal distribution during the peak hour (i.e., between 11:00 and 11:20) of a typical working day (Tuesday, February 22). Recall that the model is only for connections longer than 16 s. The four fitted lognormal distributions pass the Cramer-Von Mises and the Kolmogorov-Smirnov tests. We observe that theoretical models are a good fit for the empirical data.

Table 8 presents the statistical indicators of the actual data as well as the parameters of a lognormal curve fitting for each frequency band. The parameters of the four lognormal distributions are similar. For any frequency band, the median is less than 30 s. In the 1800 MHz and 2100 MHz bands, some connections last up to 49 min and 63 min, respectively.

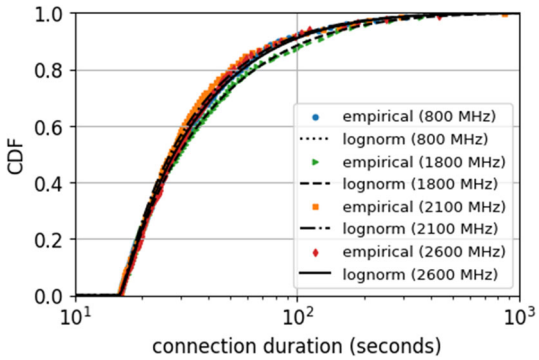


Fig. 13 Comparison of empirical and fitted curves of the connection duration (lognormal distribution). Thursday, February 22, 2022, 11:00–11:20

7.4 Connection size

In each DCI message, the base station indicates the RNTI that identifies the connection, the MCS index, and the number of PRBs allocated to the terminal (see Sect. 3.2). From the MCS index and the number of PRBs, we obtain the transport block size (in bits). There can be up to two transport blocks per DCI in the case of multiple-input multiple-output (MIMO). The connection size is the sum of the transport block sizes used during the connection. Figure 14 plots the ECDF of the downlink connection size (in Bytes) of different periods of the working days. For the sake of simplicity, the figure displays only four representative one-hour periods.

We note that the curves have similar shapes for the four frequency bands in a given period, especially in periods 10:00–11:00 and 16:00–17:00. On the other hand, comparing the periods, we observe that the curves of the periods 10:00–11:00 and 16:00–17:00 are very similar, with median and 90th percentile of approximately 6 kB and 200 kB, respectively. In the period 4:00–5:00, the connections are smaller than in the other periods, with a median and 90th percentile of approximately 1.5 kB and 30 kB, respectively. Therefore, the results suggest that during working hours (i.e., between 8:00 and 20:00), the frequency band and time of day have little influence on the connection size distribution. These results are consistent with the fact that the connection size depends mainly on user behavior and not cellular technology. User behavior usually varies depending on the day (weekdays

Table 8 Connection duration model, estimated parameters during peak hour

Band (MHz)	Mean (s)	Std. Dev. (s)	Median (s)	Lognormal parameters
800	50.42	90.92	27.91	$\mu = 2.48,\ \sigma = 1.46$
1800	62.56	155.39	29.35	$\mu = 2.61,\ \sigma = 1.52$
2100	53.52	155.55	27.04	$\mu = 2.41,\ \sigma = 1.46$
2600	56.30	135.85	28.12	$\mu = 2.51,\ \sigma = 1.44$

Thursday, February 22, 2022, 11:00–11:20

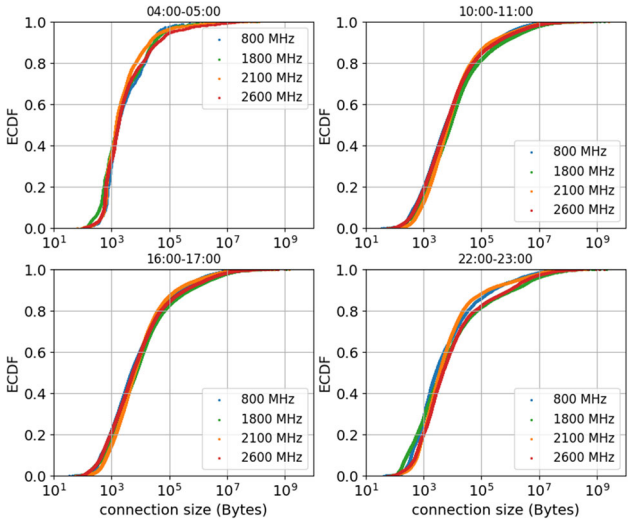


Fig. 14 Empirical CDF of the downlink connection size (x-axis, log scale), considering only working days

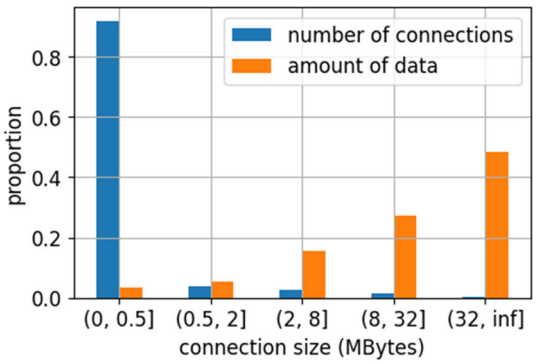


Fig. 15 Amount of data transmitted in downlink and number of connections for different connection sizes, considering only working days between 8:00 and 20:00

or weekends) and day periods (working hours, night, early morning).

7.4.1 Modeling connection size

We apply the same methodology presented in Sect. 7.3.1 to model the connection size.

Figure 15 shows the number of connections and the amount of data in MB that these connections represent for a given connection size range. There are more than 1.4×10^6 connections in the analyzed period, and the total amount of data transmitted is greater than 1100 GB. Most connections (i.e., 92%) have a size smaller than 0.5 MB. Nevertheless, these small connections represent an insignificant amount of data (i.e., 4%) compared to the total data transmitted during the analyzed period. These connections are probably due to background traffic, signaling exchanges, or connections from terminals that camp few seconds on the monitored cell. We also observed a few connections larger than 32 MB; however, these connections represent 48% of the total data. Given that the connections smaller than 0.5 MB represent a small fraction of the total data exchanged, we consider only connections larger than 0.5 MB when modeling the connection size. Let N be a random variable that denotes the size in MBytes of connections greater than 0.5 MB. A location parameter is included in the model to shift the distributions to take into account the fact that we ignore small connections.

We model only connections during working hours (between 8:00 and 20:00) since few connections are greater than 0.5 MB during non-working hours. We analyze one-hour periods to have a sufficient number of connections per period. There are 12 periods for each day and 60 periods for each frequency band; thus, we fit a total of 240 periods. For each curve fitting, there are, on average, 515 connections. Results of the curve fitting validation are listed in Table 9. We note that lognormal distribution passes the Cramer-Von Mises and Kolmogorov-Smirnov test in 99.58% and 100% of the periods, respectively.

Results suggest that lognormal is the best candidate to model the connection size. Using a similar argument to the one presented in Sect. 7.3.1, we fix the location parameter of the lognormal distribution to 0.4 MB. The lognormal distribution passes the Cramer-Von Mises test and the Kolmogorov-Smirnov test in 98.75% and 98.33% of periods, respectively. Thus we have:

$$N = 0.4 + X,$$

where $X \sim \text{Log-}\mathcal{N}(\mu, \sigma)$.

We analyze the parameters of the fitted lognormal curves. The average and the coefficient of variation of the shape parameter σ are 1.55 and 0.05, respectively. This means that this parameter is almost constant. The average and the coef-

Table 9 Downlink connection size, percentages of goodness-of-fit tests passed with a confidence level of 95%

Distribution	Cramer-Von Mises		Kolmogorov-Smirnov		Valid
	Passed (%)	Best-fit (%)	Passed (%)	Best-fit (%)	
Exponential	0	0	0	0	×
Weibull	26.25	1.25	27.50	1.67	×
Lognormal	99.58	86.25	100	88.33	✓
Pareto	78.75	12.50	74.58	15	×
Beta	11.25	0	7.08	0	×
Normal	0	0	0	0	×
Rayleigh	0	0	0	0	×

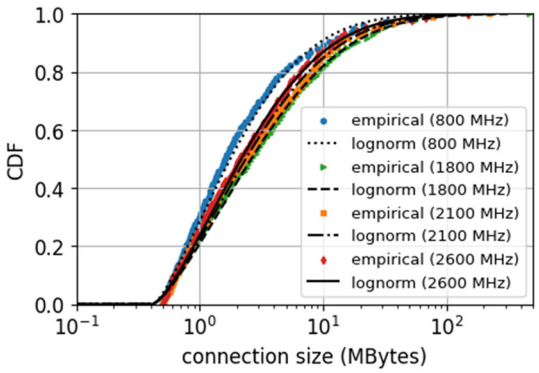


Fig. 16 Comparison of empirical and fitted curves of the downlink connection size (lognormal distribution). Thursday, February 22, 2022, 11:00–12:00

ficient of variation of the scale parameter $m = \exp(\mu)$ are 1.94 and 0.20, respectively. This means that m has a higher variability. This is because the distributions of the 800 MHz band are generally slightly shifted to the left of the other distributions. This could be because the 800 MHz band is used at the cell edge, and thus there is a higher handover probability and a shorter connection. Considering only the 1800 MHz, 2100 MHz, and 2600 MHz bands, the scale parameter's average and the coefficient of variation are 2.08 and 0.15, respectively. These results indicate that the time of day and the frequency band do not significantly affect the distribution of the connection size between 8:00 and 20:00, especially in the 1800 MHz, 2100 MHz, and 2600 MHz frequency bands. Like the connection duration, the connection size depends mainly on user behavior. However, when a handover occurs, the connection terminates, and the connection size is smaller than expected. The handover probability varies according to the frequency band. The 800 MHz band has a higher handover probability since it is used at the cell edge. This would explain why the connection sizes are smaller in the 800 MHz band.

Figure 16 compares empirical data and the fitted lognormal curves during the peak hour of a typical working day (Tuesday, February 22, between 11:00 and 12:00). Recall that the model is only for connections greater than 0.5 MB. The fitted lognormal distribution passes the Cramer-Von Mises test and the Kolmogorov-Smirnov test in the four frequency bands. The figure shows that the fitted curves and the real data distributions match well. The statistical indicators and the parameters of the fitted curves are presented in Table 10. The value of the μ parameter in the 800 MHz band is smaller compared to the other bands. That is, connection sizes in this band are smaller. The value of the σ parameter is similar in the four bands.

8 Towards a 5 G analyzer

The development of 5 G networks will enable new use cases (virtual reality), new types of deployments (industrial), and new models will need to be established. However, in the early stages of deployment, 5 G complements 4 G by providing higher bit rates. Currently, no specific applications for the 5 G network are used on a large scale. Hence, we believe the proposed models are also valid for 5 G networks since user behavior does not vary depending on the cellular technology used.

Table 10 Connection size model, estimated parameters during peak hour

Band (MHz)	Mean (MB)	Std. Dev. (MB)	Median (MB)	Lognormal parameters
800	5.19	13.36	1.61	$\mu = 0.31,\ \sigma = 1.46$
1800	9.55	42.69	2.66	$\mu = 0.79,\ \sigma = 1.61$
2100	7.16	17.88	2.35	$\mu = 0.67,\ \sigma = 1.55$
2600	6.74	17.73	2.08	$\mu = 0.57,\ \sigma = 1.51$

Thursday, February 22, 2022, 11:00–12:00

We are extending our tool to monitor 5G networks. As in 4G, 5G uses a DCI to inform one terminal or group of terminals about the resource allocation for the uplink or downlink. A 24-bit CRC is appended to the DCI payload. The last 16-bit of the CRC is masked with a specific RNTI as in 4G. A 16-bit RNTI is assigned to a terminal that establishes a connection with the network. Unlike LTE, 5G introduces the control resource set (CORESET), a set of physical time/frequency resources and parameters used to carry DCI. Furthermore, there is a new aggregation level in addition to the four already existing in LTE. This means there will be more variables to consider, increasing the complexity of the search function and requiring more powerful computing resources. Moreover, when beamforming is used, a greater number of analysis points are needed. In other words, more analyzers are required.

9 Conclusion and future work

In this paper, we have analyzed the behavior of a 4G network using real data collected inside an LTE cell of a French MNO. To this end, we have first developed a platform (LTE analyzer) to collect all the control messages transmitted by the eNB in a monitored cell. We have then carried out a one-week measurement campaign using four LTE analyzers in parallel (one for each frequency band).

We have analyzed the resulting traces better to understand the behavior of a 4G cell. We have considered several parameters: network utilization, number of connections, and uplink (and downlink) cell achieved throughput. We have confirmed that signaling messages represent a small fraction of the traffic: excluding RRC messages, they use less than 1% of the cell capacity. As expected, these parameters vary depending on the day (weekend or not) and time of day. Interestingly, some also depend on the frequency band, namely the inter-arrival time distribution between two consecutive connections. On the contrary, connection duration and connection size are almost independent of the frequency band and the time of day during working hours. We have modeled the connection inter-arrival time, the connection duration, and the connection size. The results suggest that the Weibull distribution is the distribution that best models the inter-arrival time, but the exponential distribution can also be used. The connection duration and the connection size can be modeled as shifted lognormal distributions.

We plan to extend this work by monitoring other MNOs and conducting longer measurement campaigns. These new campaigns will include cellular traffic from 5G networks. To achieve this, our team is developing a dedicated sniffer specifically designed for capturing information within 5G networks. Furthermore, we will integrate packet-level statistical analysis.

Author contributions All authors contributed to the study conception and design. Data collection and analysis were performed by Cesar Vargas Anamuro. The first draft of the manuscript was written by Cesar Vargas Anamuro and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors have not disclosed any funding.

Declarations

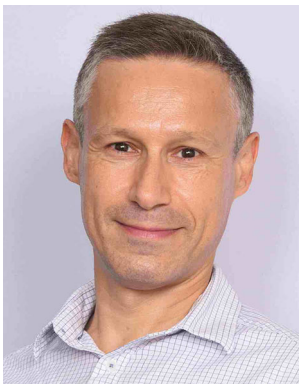
Conflict of interest The authors declare no conflict of interest.

References

1. The mobile economy. (2022). *White paper, GSMA, Intelligence*.
2. Zeng, Q., Sun, Q., Chen, G., Duan, H., Li, C., & Song, G. (2020). Traffic prediction of wireless cellular networks based on deep transfer learning and cross-domain data. *IEEE access, 8*, 172387–172397.
3. Elayoubi, S.-E., Saker, L., & Chahed, T. (2011). Optimal control for base station sleep mode in energy efficient radio access networks. In *2011 Proceedings IEEE INFOCOM*, (pp. 106–110). IEEE.
4. Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., & Lepri, B. (2015). A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific data, 2*(1), 1–15.
5. Chen, X., Jin, Y., Qiang, S., Hu, W., & Jiang, K. (2015). Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale. In *2015 IEEE International Conference on Communications (ICC)*, (pp. 3585–3591). IEEE.
6. Cisco visual networking index. (2019). *Global mobile data traffic forecast update, 2017–2022*. Cisco: White paper.
7. Zhang, C., Dang, S., Shihada, B., & Alouini, M.-S. (2021). Dual attention-based federated learning for wireless traffic prediction. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, (pp. 1–10). IEEE.
8. Kumar, S., Hamed, E., Katabi, D., & Erran Li, L. (2014). LTE radio analytics made easy and accessible. *ACM SIGCOMM Computer Communication Review, 44*(4), 211–222.
9. Falkenberg, R., & Wietfeld, C. (2019). FALCON: An accurate real-time monitor for client-based mobile network data analytics. In *2019 IEEE Global Communications Conference (GLOBECOM)*, (pp. 1–7). IEEE.
10. Um, J., Kim, I., & Park, S. (2021). Implementation of platform for long-term evolution cell perspective resource utilization analysis. *ETRI Journal, 43*(2), 232–245.
11. Gomez-Miguelez, I., Garcia-Saavedra, A., Sutton, P.D., Serrano, P., Cano, C., & Leith, D. J. (2016). srsLTE: An open-source platform for LTE evolution and experimentation. In *Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*, (pp. 25–32).
12. Paul, U., Subramanian, A.P., Buddhikot, M.M., & Das, S. R. (2011). Understanding traffic dynamics in cellular data networks. In *2011 Proceedings IEEE INFOCOM*, (pp. 882–890). IEEE.
13. Nan, E., Chu, X., Guo, W., & Zhang, J. (2013). User data traffic analysis for 3G cellular networks. In *2013 8th International Conference on Communications and Networking in China (CHINACOM)*, (pp. 468–472). IEEE.
14. Laner, M., Svoboda, P., Schwarz, S., & Rupp, M. (2012). Users in cells: A data traffic analysis. In *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, (pp. 3063–3068). IEEE.
15. Foddis, G., Garroppo, R. G., Giordano, S., Procissi, G., Roma, S., & Topazzi, S. (2014). LTE traffic analysis and application behavior

characterization. In *2014 European Conference on Networks and Communications (EuCNC)*, (pp. 1–5). IEEE.
16. Polaganga, R. K., & Liang, Q. (2015). Self-similarity and modeling of LTE/LTE-A data traffic. *Measurement, 75*, 218–229.
17. Narmanlioglu, O., Zeydan, E., Kandemir, M., & Kranda, T. (2017). Prediction of active UE number with Bayesian neural networks for self-organizing LTE networks. In *2017 8th International Conference on the Network of the Future (NOF)*, (pp. 73–78). IEEE.
18. Kim, T., Ko, K. S., & Sung, D. K. (2015). Prioritized random access for accommodating M2M and H2H communications in cellular networks. In *2015 IEEE Globecom Workshops (GC Wkshps)*, (pp. 1–6). IEEE.
19. Jedrzycki, C., & Leung, V. C. (1996). Probability distribution of channel holding time in cellular telephony systems. In *Proceedings of Vehicular Technology Conference-VTC*, (vol. 1, pp. 247–251). IEEE.
20. Barcelo, F., & Sánchez, J. I. (1999). Probability distribution of the inter-arrival time to cellular telephony channels. In *1999 IEEE 49th Vehicular Technology Conference (Cat. No. 99CH36363)*, (vol. 1, pp. 762–766). IEEE.
21. Foddis, G., Garroppo, R., Giordano, S., Procissi, G., Roma, S., & Topazzi, S. (2016). Modeling RACH arrivals and collisions for human-type communication. *IEEE Communications Letters, 20*(7), 1417–1420.
22. Lagrange, X., & Godlewski, P. (1995). Teletraffic analysis of a hierarchical cellular network. In *1995 IEEE 45th Vehicular Technology Conference. Countdown to the Wireless Twenty-First Century*, (vol. 2, pp. 882–886). IEEE.
23. Garcia, J., Alfredsson, S., & Brunstrom, A. (2015). Examining TCP short flow performance in cellular networks through active and passive measurements. In *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, (pp. 7–12).
24. Dahlman, E., Parkvall, S., & Skold, J. (2011). *4G LTE/LTE-Advanced for Mobile Broadband*. Oxford: Academic press.
25. Barceló, F., & Jordán, J. (2000). Channel holding time distribution in public telephony systems (PAMR and PCS). *IEEE Transactions on vehicular technology, 49*(5), 1615–1625.
26. Sandoval, R. M., Garcia-Sanchez, A.-J., Molina-Garcia-Pardo, J.-M., Garcia-Sanchez, F., & Garcia-Haro, J. (2016). Radio-channel characterization of smart grid substations in the 2.4-GHz ISM band. *IEEE Transactions on Wireless Communications, 16*(2), 1294–1307.
27. Collonge, S., Zaharia, G., & Zein, G. E. (2004). Influence of the human activity on wide-band characteristics of the 60 GHz indoor radio channel. *IEEE Transactions on Wireless Communications, 3*(6), 2396–2406.
28. Feng, H., & Shu, Y. (2007). Statistical analysis of packet inter-arrival times in wireless LAN. In *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, (pp. 1888–1891). IEEE.
29. Abdi, A., & Kaveh, M. (2011). A comparative study of two shadow fading models in ultrawideband and other wireless systems. *IEEE transactions on wireless communications, 10*(5), 1428–1434.
30. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods, 17*(3), 261–272.
31. LTE. (2012). *Radio Access Network (RAN) enhancements for diverse data applications*. Technical report 36.822, 3rd Generation Partnership Project (3GPP). Version 11.0.0.
32. Meng, L., Liu, S., & Striegel, A. D. (2014). Characterizing the utility of smartphone background traffic. In *2014 23rd International Conference on Computer Communication and Networks (ICCCN)*, (pp. 1–5). IEEE.

Cesar Vargas Anamuro received the Telecommunications Engineering Diploma degree from the National University of Engineering of Lima, Peru, in 2012, the M.S. degree in Information Technology and the Ph.D. degree in computer science from IMT Atlantique, France, in 2015 and 2020, respectively. He was working as a Post-Doctoral Researcher at IMT Atlantique and is now with Mitsubishi Electric R&D Centre Europe. His research interests are focused on wireless communication and data analysis.



Alberto Blanc received a "laurea" in computer engineering in 1998 from the "Politecnico di Torino," Italy, and a M. S. and a Ph.D. in electrical engineering from the University of California, San Diego, in 2004 and 2006 respectively. Between 2006 and 2010 he was a post-doctoral researcher at INRIA and Orange Labs in Sophia Antipolis, France. Since 2010, he is an associate professor at IMT Atlantique (formerly Telecom Bretagne), with the "Network Systems, Cybersecurity and Digital Law" department. His research interests include Quality of Service guarantees, network measurement and modeling.

Xavier Lagrange received the Engineering Degree from Ecole Centrale des Arts et Manufactures, Paris, France, in 1984 and the Ph.D. degree from TELECOM Paris in 1998. Since 2002, he has been professor in IMT Atlantique at the "Network Systems, Cybersecurity and Digital Law". He leads research group ADOPNET (Advanced Technologies for Operated Networks) at IRISA (Institut de Recherche en Informatique et Systemes Aléatoires). His domain of interest includes radio resource allocation, RAN architectures, medium access control and performance analysis of 4G-6G cellular networks.