# On Mobile Traffic Distribution over Cellular Backhauling Network Nodes

Sandesh Uppoor*†, Cezary Ziemlicki*, Stefano Secci†, Zbigniew Smoreda*
*Orange Labs, 92794 Issy-les-Moulineaux, France. Email: firstname.lastname@orange.com
†Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, 75005 Paris, France. Email: firstname.lastname@upmc.fr

***Abstract*—The rapid growth of mobile traffic and the emergence of advanced mobile services and infrastructures are shifting significant attention toward the cellular network back-hauling infrastructure. At this network segment, there is a growing interest in understanding spatio-temporal mobile traffic distributions at different network levels, in order to better define flexible networking solutions for forthcoming smart 5G infrastructures including, for instance, mobile edge computing features. In this work we study these aspects and characterize the load on cellular access networks using real-world anonymized subscriber data, from the Lyon metropolitan area in France, providing statistical distribution to the research community. We find that the traffic distribution at Node-B level is best fit by a Weibull distribution, and that at the radio network aggregation it is best fit by a hybrid Weibull-Pareto distribution.**

## I. Introduction

Global mobile data traffic grew 69% in 2014 and is expected to reach 24.3 Exabytes per month by 2019 [1]. These numbers are expected to further swell due to the fact that handheld devices are not the sole contributor for mobile data demand.

Besides simple users, smart cities and industrial automation solutions are increasingly relying on a large variety of connected objects, constantly collecting data about the user, the network, and the environment, back-hauled via the cellular access infrastructure, in order to support advanced mobile services, data analytics and business intelligence. In order to meet the rising requirements of these novel applications, in terms of service access latency and reliability, forthcoming 5G infrastructures are being designed, integrating high mobile service provisioning flexibility offered by novel network technologies. For instance, in Mobile Edge Computing (MEC) [4], cloud server surrogates are provided next to access and aggregation cellular network nodes as Cloudlets [2], [3] to narrow the gap between the target population and popular services and contents, thereby minimizing cloud access delays and increasing mobile service availability. Thanks to cloud and virtualization technologies, also the network can be made smarter, adapting the service provisioning behavior to the user mobility and consumption dynamics. In this context, understanding the spatiotemporal user and data demand dynamics is a crucial first step toward the design of advanced 5G infrastructures.

With the objective to support the research community in the evaluation of novel 5G services and infrastructures, we unintrusively collect and anonymize data from the Orange cellular network and characterize the data traffic distribution over the cellular back-hauling network nodes, determining the statistical laws behind. We worked on a dataset collected from the Lyon metropolitan area network in France.

Before we proceed any further, for the sake of readability, let us clarify the terminology used in this paper. A *Telecommunication cell* or *Site* is a geographic coverage where a subscriber mobile device automatically connects to the telecom infrastructure. To meet the growing cellular demand, these cells are further divided into smaller cells, also called *sectors*, with a unique *cell identifier* (CI) (thus making the system and network planning more complicated to visualize). For the ease of illustration, we group all the sectors with same antenna coordinates as one entity known as Base Station (BS). In the real-world cellular architecture these sectors are grouped with a Location Area Code (LAC) to minimize the control message overhead, paging and for faster localization of the user to deliver data content or other services. The number of cells or sectors in a LAC is purely based on the popularity of the geographical area and the population residing in that area; urban region can have cells grouped into more LACs than the outskirts, and a single LAC can cover a few smaller cities.

We perform the data traffic analysis on two cellular back-hauling architecture levels as shown in Fig. 1: 1) BS node level, through which the subscriber connects; 2) Radio Network (RN) aggregation node level, which aggregates multiple BSs toward the core network, involving Radio Network Controller (RNC) in 3G and Base Station Controller (BSC) in the 2G. It is worth mentioning that we easily found that the core network nodes, further aggregating traffic towards the Internet and the Public Switch Telephone Network (PSTN), exhibit similar behavior as RN level and is therefore omitted in the following.

In Sec. II, we present the dataset we use for this analysis, with a description about data collection and data preparation procedures. In Subsection III-A, we discuss the variability in user density over time in BS level and RN level. Subsection III-B presents the data traffic consumption behavior and modeling.

## II. Dataset

In the last few years, many operators offer telecom network data for data analytics research. The most common datasets are Call Detail Record (CDR) datasets, which are limited to service usage information like duration of calls, caller, called user information, SMS and MMS information.

CDR datasets are nowadays considered too primitive datasets for 5G research as they do not convey useful information about data volumes. Nowadays, the explosion of mobile Over the Top (OTT) services (e.g., WhatsApp, Skype) [5] has

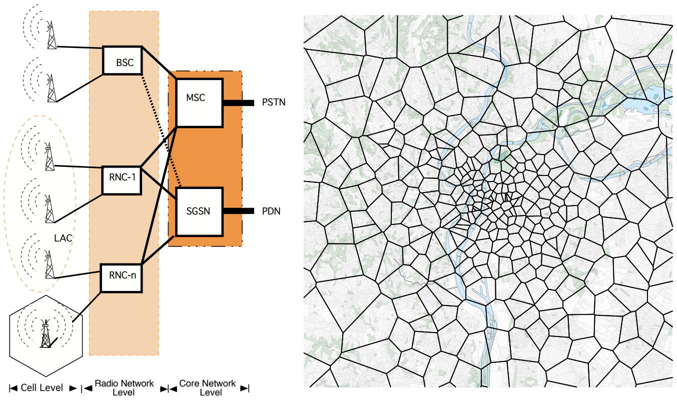


Fig. 1. Basic representation of cellular architecture (left), and cell layout resulting from the Voronoi tessellation (right).

strongly influenced the way the cellular infrastructure is used, and 5G data will represent voice calls only very marginally as opposed to 2G services to which most CDRs refer to.

This change in service usage behavior mandates datasets to include data traffic consumption information. One of the popular solution to fill this gap has been to consider Voice, SMS, MMS service requests as simply *calls*. Though there are datasets that provide both number of data service request and consumption volume, they are either at small scale or lack spatial details of data consumption. In this work, we present our analysis and discussion on data consumption behavior based on a dataset that is rich in data service requests, representative of 3G and also 4G traffic, with spatial signatures and volume information. We describe in the following the *Data Collection* and *Dataset Preparation* procedures.

### A. Data Collection

To prepare the dataset used in this analysis, we collect data from two probes in the Orange mobile cellular French network, namely the *Cigale* and *Otarie* probes. The Cigale probe is placed between the RNC and the core network with as main objective to capture the detailed user connectivity. Hence this data describes the time, serving cell, service demand and delivery messages associated to a user. The placement of Cigale probe does not allow to accurately record the amount of data consumed by a user in a particular cell. To that end, we use data from the Otarie probe, which is placed between the SGSN (Serving GPRS Support Node) and the packet data network, and which offers detailed data consumption including application type information. The data collected from the Otarie probe includes number of bytes consumed by a user every 6 minutes for different application types like Web, Mail, News, VoIP, Chat, Streaming etc. Contrarily, though Otarie probe is rich in data consumption information, it lacks precise user localization within a LAC.

The collected data covers the Lyon urban region and outskirts with approximately 1.5 million users for 17 days from April 14 to April 30, 2014. As these data are continuously collected and sized in tera bytes, and need a high performance computing BigData platform - we used the Orange Hadoop Corporate Infrastructure (OHCI), with Hive [6] and Pig [7] as processing Languages.

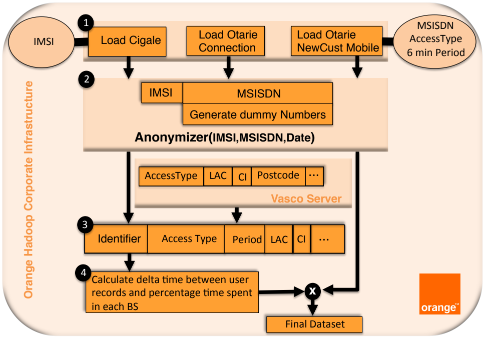

IMSI
1
Load Cigale
Load Otarie Connection
Load Otarie NewCust Mobile
MSISDN AccessType 6 min Period
2
IMSI
MSISDN
Generate dummy Numbers
Anonymizer(IMSI,MSISDN,Date)
AccessType
LAC
CI
Postcode
...
Vasco Server
3
Identifier
Access Type
Period
LAC
CI
...
4
Calculate delta time between user records and percentage time spent in each BS
X
Final Dataset
Orange Hadoop Corporate Infrastructure
orange™

Fig. 2. Dataset preparation procedure using Orange Hadoop Corporate Infrastructure.

### B. Dataset Preparation

The collected raw data does not allow us to carry out any analysis straight away. The data need to be cleaned first, and then data from the two probes are to be merged in order to obtain both spatial and consumption precision. Hence we adopt a four step process as shown in Fig. 2.

**Step 1**: In this step, we load the collected data to the OHCI. In Cigale data, a subscriber is identified by his International Mobile Subscriber Identity (IMSI), but their Mobile Station International Subscriber Directory Number (MSISDN) identifies user in the Otarie probes. To find a mapping between these two identifiers, we managed specific mapping tables.

**Step 2**: The raw data is then fed to an *Anonymizer* that runs SHA-1 algorithm and calculates a unique *identifier* for each user from a string composed of IMSI, MSISDN and date. However, not all IMSIs in the Cigale dataset have a corresponding MSISDN in the mapping table. This gap is addressed by using an incremental number. As a result of this step each user has a unique identifier every day.

**Step 3**: Each record in the Cigale data also provides information about LAC and CI. With this knowledge we can add additional information like postal codes, antenna orientation, latitude and longitude by retrieving them from a third database that maintains up-to-date information about the serving cells in France. We also filter out many records in this step like records with unknown LAC, IMSI and erroneous time-stamp.

**Step 4**: In this step, the 24 hours day window is represented with 6 minute slots, and the user records from Step 3 are mapped to these slots. This process is done because the user data consumption from the Otarie probe is available with 6 minutes granularity. Now the two datasets are merged when user identifier and the time slot matches. In case the user travels over many cells within this 6 minutes window in Cigale data, the percentage of time spent in each cell is multiplied by the total data consumption in that 6 minutes from Otarie data.

The final dataset preserves all the originality of the Cigale data along with data consumption volume - to the best of our knowledge this is the first dataset that captures mobility and data consumption volume at high granularity.

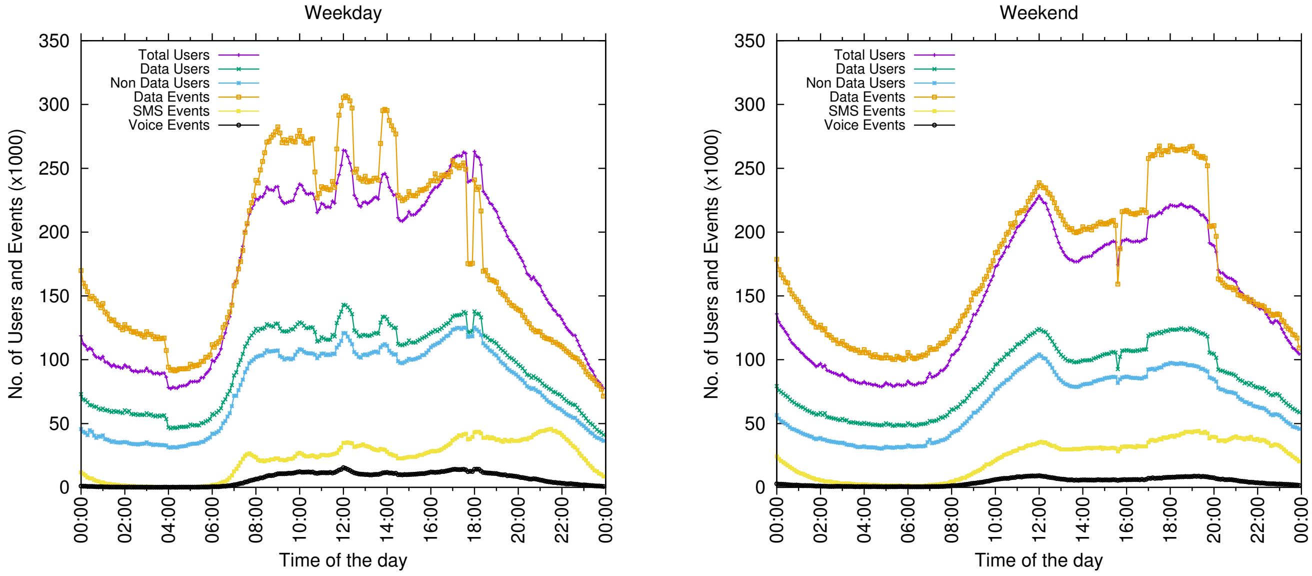
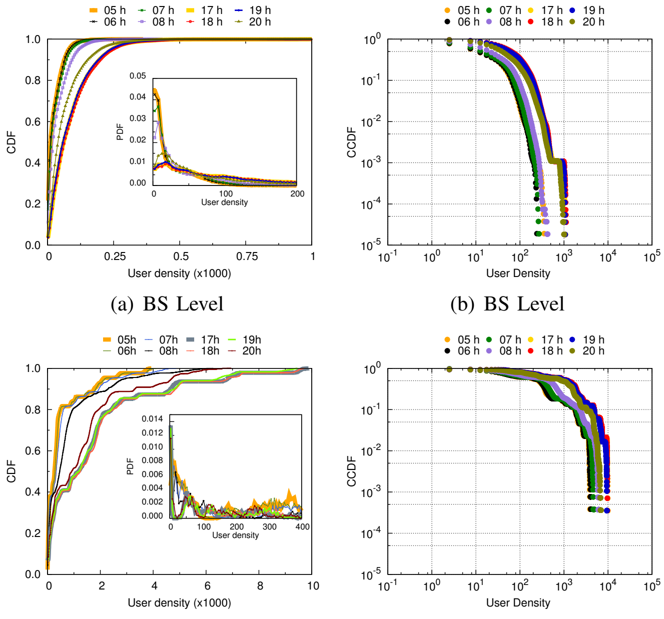


Fig. 3. Variability in number of users and service request over a typical Weekday (left) and Weekend (right).



(a) BS Level

(b) BS Level

(c) RN Level

(d) RN Level

Fig. 4. User density distributions (linear-linear and logarithmic-logarithmic scales).

## III. Analysis

In this section, we discuss the user and service usage dynamics from the designed dataset. In Fig. 3, we present the overall statistics of the user density and volume of service requested on week days and on weekend days. An important user density and service demand can be observed between 8 am and 8 pm because of various activities in which people engages, including home to office commutes, leisure, fitness, shopping etc. Hence during these times, the use of hand held devices are more frequent than in other times of the day. In Fig. 3 (right), the curve has a lazy start because of the week-off; a case when the daily activity is cut short by the home-office commutes. These overall statistics do not provide much clarity on the actual data consumption load on the cellular network levels; hence we present a detailed analysis in the rest of this Section.

In sub-section III-A, we look into the user dynamics. In sub-section III-B, we discuss the total data traffic demand dynamics and highlight few popular applications. Generalizing the data consumption behavior is very crucial (and a starting point for advanced cloud based applications and 5G architectures), hence we discuss the best-fit statistical distribution. We conclude the section with a discussion on spatiotemporal variation of data traffic load over the back-hauling network nodes.

### A. User density

To better understand the varying user density as presented in Fig. 3, we account the user density over geographical sites over different times of the day. Real-world cellular site boundaries are very complex to calculate because of non-deterministic radio wave propagation. Therefore in our analysis, we mimic the telecom sites by running the Voronoi tessellation algorithm on the base station locations in the study region, i.e, Lyon city area, as shown in Fig. 1 (right). The number of users connected to a BS represents the user density at a Voronoi site or cellular site. We see that the city center has smaller sites than in the outskirts of the city, as single large cell cannot handle the large amount of service request posed by the dense population in those region.

Fig. 4 shows the cumulative distribution function (CDF) and the probability density function (PDF) of the user density at both the BS and the RN levels. These distribution are calculated on samples that are the aggregation of user density at each cell during each minute. The CDF curve representing 5 am shows that 95% of the samples have approximately 100 users/min/BS and at 5 pm have approximately 250 users/min/BS, thus defining the min-max boundaries. Apart from these values, justification for 5% of the user density is not clear, so we visualize the complementary cumulative distribution function (CCDF) with logarithmic scales in Fig. 4(b). We see that the extreme values of user density can even swell to 1000 users/min/BS, which is acceptable as there are cells covering a very large area in the city outskirts or because of a social event in the city center.

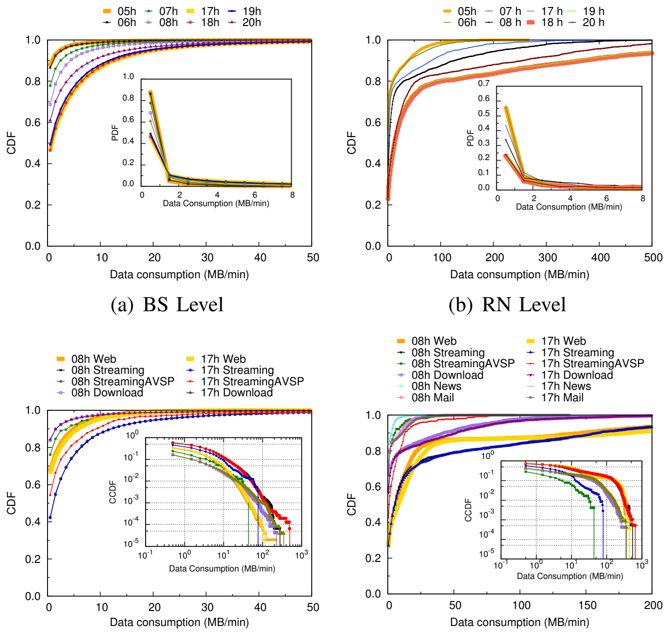
At the RN level, the numbers are in few thousands given the fact that approximately 1000 cell sectors are connected to a single radio network element like RNC or BSC in the study region. Similar user load distribution as seen in BS level could be expected at the RN level, where instead the distributions are not very smooth. This is because as we go up the hierarchy the number of statistical samples on which the distribution is calculated reduces leading to a step like curve. Other reason behind this is due to drastic change in the user densities in few RNCs, which is a direct implication of probe downtime. This situation occurs when the probes fail to match the velocity and volumes of data arriving from the network element. Though we eliminate such cases in our samples, a small effect on the result is still evident but do not affect the overall result or information that it conveys.

Looking back at Fig. 3, we see that density of users who subscribe for data package seems to be close to that of basic cellular service users. In addition, comparing the number of data traffic requests vs SMS or Voice dictates a major shift in how the cellular network is being used with respect to what it was initially designed for. Hence user density shall not be individually used either for network planning nor for decision making to include smart techniques to improve content delivery times. This motivates us to look deeper into spatiotemporal dynamic of data traffic consumption.

### B. Mobile Data Traffic

We first present the CDF and PDF of total data consumed in Mega Bytes (MB) at BS and RN levels in Fig. 5 at different times of the day. Similar to the user density discussed earlier, total data traffic at cellular sites increases at two notable active peaks of the day i.e., 5 am to 8 am (morning peak) and 5 pm to 8 pm (evening peak). In 95% of the samples, the data traffic increases from less than 5 MB/min/BS in the morning peak and reaches the maximum of 15 - 20 MB/min/BS during the evening peak hours, and it tends to fall back as seen in the curve at 8 pm. This result follows the same pattern as in Fig. 3, users are more active during the day time, hence reflected in the data consumption behavior. The PDF in the inset plot explains that the majority of the data consumption samples every minute falls below 2 MB/min/BS.

In case of RN level, as depicted in Fig. 5(b), 75% data traffic routed is approximately 10 - 50 MB/min/RNC at 5 pm, 6 pm and 7 pm as compared to a negligible amount at the morning peak time. It is interesting to see that approximately 25% of the data traffic at RN level is greater than 100 MB/min/RNC and that the distribution takes a different shape thereafter. This heavy tailed behavior is due to significant number of smaller cells in the Lyon city area with highest concentration of mobile data subscribers at evening peaks and also due to the popular content type; it is worth noting that this is a common characteristic of any dense metropolitan area network of a large European city, with population, services, universities and many industrial and commercial activities concentrated hierarchically as one reaches the centroids of the metropolitan area. To further investigate this aspect, we present the CDF and CCDF of popular application types.



(a) BS Level

(b) RN Level

(c) BS Level

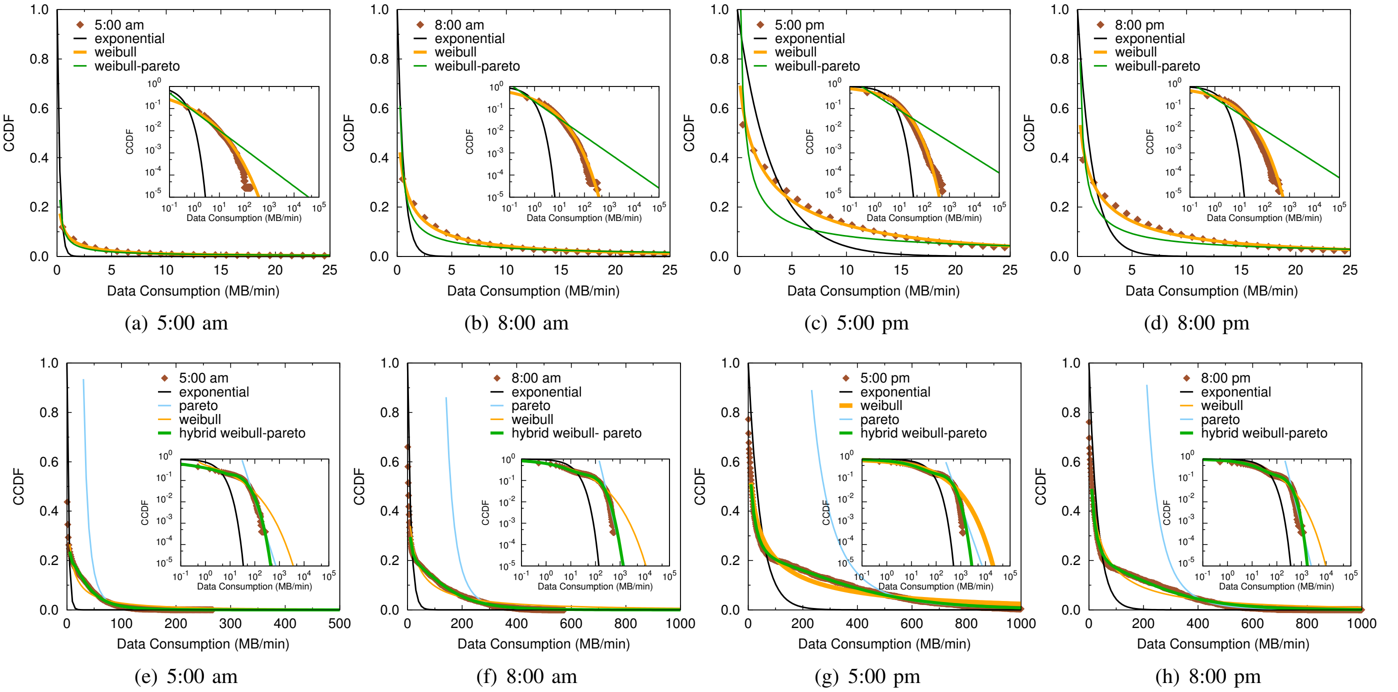
(d) RN Level

Fig. 5. Distributions of total and per application type data consumption in BS level and RN level.

In Fig. 5(c) and Fig. 5(d), we plot the per application type distributions at the two peaks of the day. Surprisingly the content consumption behavior per application at morning peak closely matches to that in the evening peak. Video streaming and web are the most popular among the data subscribers in Lyon and this content consumption even soar to 500 MB per minute at a few base stations. In case of RN level the heavy-tail consumption behavior seen in Fig. 5(b) is because of video streaming and web traffic. This somehow confirms the already mentioned global report from cisco - *Mobile video traffic exceeded 50 percent of total mobile data traffic at the end of 2012 and grew to 55 percent by the end of 2014*. These numbers are accelerated by new techniques in content delivery like Adaptive Video Streaming Protocol (AVSP) and Auto-play mode of advertising video, and are expected to further grow with mobile edge computing.

On the other hand, two more applications that contribute to total data traffic are unveiled, namely ‘News’ and ‘Mail’ in RN level. Though the consumption is less compared to that of streaming, it was hidden in case of BS level analysis. These data traffic demand being very essential inputs for network planning or system planning in general, it is interesting to see if these demands follow some statistical distributions. So we proceed to make an attempt to find a best statistical function for this empirical distribution.

Distribution fittings were performed using the nonlinear least-squares (NLLS) Marquardt-Levenberg algorithm



(a) 5:00 am

(b) 8:00 am

(c) 5:00 pm

(d) 8:00 pm

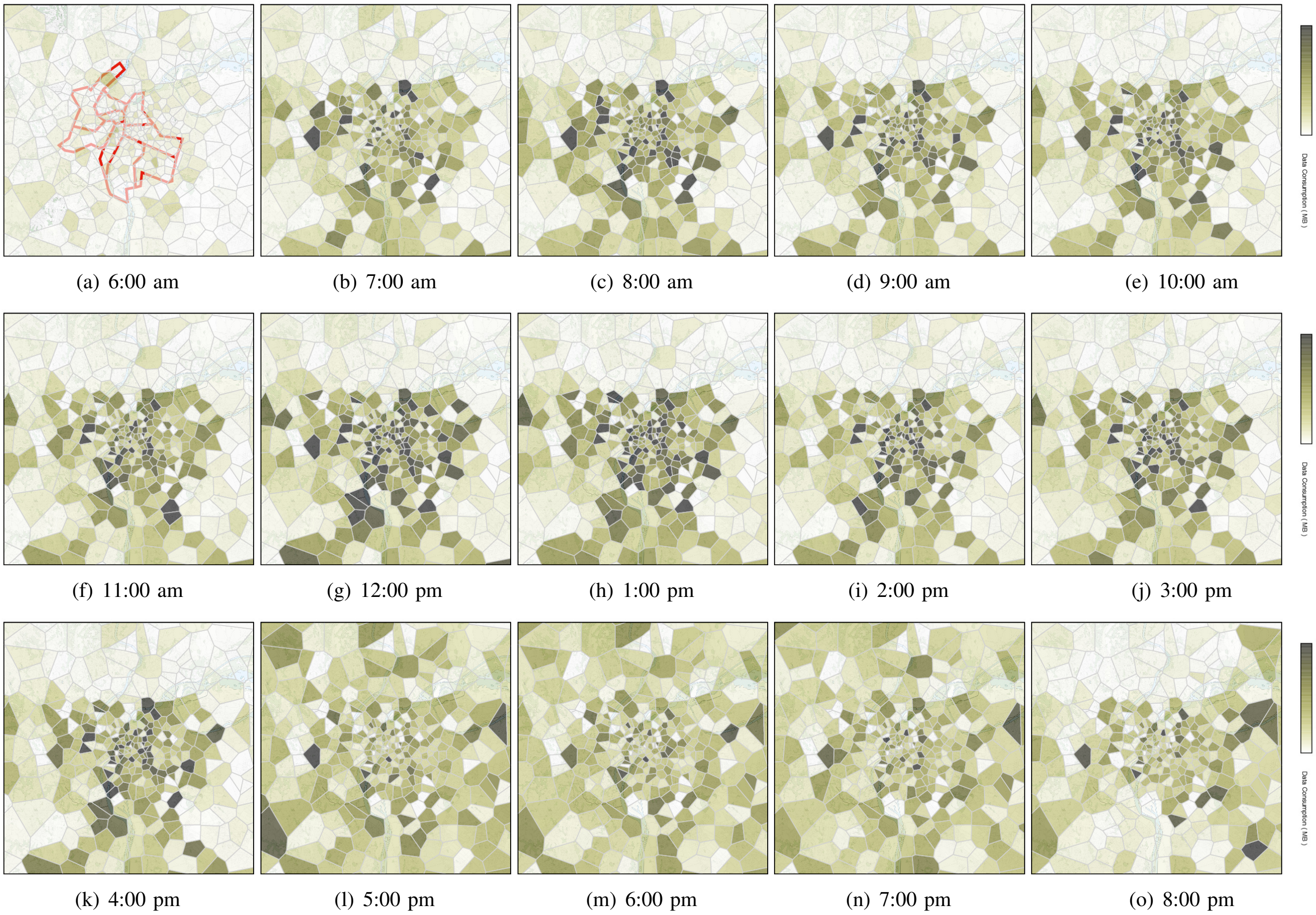
(e) 5:00 am

(f) 8:00 am

(g) 5:00 pm

(h) 8:00 pm

Fig. 6. Statistical distribution fittings on the total data consumption on BS level (a to d) and RN level (e to h). Inset plots are in logarithmic-logarithmic scales.



(a) 6:00 am

(b) 7:00 am

(c) 8:00 am

(d) 9:00 am

(e) 10:00 am

(f) 11:00 am

(g) 12:00 pm

(h) 1:00 pm

(i) 2:00 pm

(j) 3:00 pm

(k) 4:00 pm

(l) 5:00 pm

(m) 6:00 pm

(n) 7:00 pm

(o) 8:00 pm

Fig. 7. Spatiotemporal data consumption during a working day.

(see [10]) - implemented in Gnuplot and Matlab. In Fig. 6, we present the linear-linear plots showing the total data traffic consumption at BS level and their best-fit statistical distribution. To examine the nature of the heavy tail distribution, we show the inset plot with logarithmic scales. After evaluating fitting errors of various statistical distributions in the literature, we determined that the Weibull distribution [8] is the closest one, fitting the CCDF in both the data traffic peak periods of the day, in particular when compared to Weibull-Pareto and exponential distributions that are other quite close distributions. The CCDF of Weibull distribution is defined in $(2)$, and the parameter values that can closely mimic the empirical data consumption behavior at different times as seen in Fig. 6 (a to d) are presented in Table I. The Root Mean Square Residual (RMSR) error gives an idea of the average error during the fitting process (it is commonly known as the quality of the fit).

$$F_{W_{\alpha,\beta}}(x) = 1 - e^{-\left(\frac{x}{\beta}\right)^{\alpha}} \tag{1}$$

$$CCDF_{W_{\alpha,\beta}}(x) = 1 - F_{W_{\alpha,\beta}}(x) \tag{2}$$

TABLE I. Parameters of the fitting Weibull distribution.

| Time | $\alpha$ | $\beta$ | RMSR |
|---|---|---|---|
| 5 am | 2.875e-02 | 2.574e-01 | 8.270e-04 |
| 8 am | 3.826e-01 | 3.538e-01 | 1.813e-03 |
| 5 pm | 2.133e+00 | 4.685e-01 | 4.096e-03 |
| 8 pm | 8.569e-01 | 3.782e-01 | 3.864e-03 |

In case of RN level, the head of the distribution is best fit by the Weibull distribution, but the heavy tail is instead best fit by the Pareto distribution, hence the combination of these two distributions best defines the data traffic consumption behavior. This distribution is coined as Hybrid Weibull-Pareto [9] as in $(4)$, $x \mapsto \rho^x$ is the function that maps $x$ to $\rho^x$, $W_{\alpha,\beta}$ and $P_{\gamma,\delta}$ are random variables distributed according to standard Weibull and Pareto functions as in $(1)$ and $(3)$. In Table II, we present the best-fit parameter values for the Hybrid Weibull-Pareto distribution at different times. Besides the plots, interested readers can obtain the empirical distribution data by directly contacting the authors.

$$F_{P_{\gamma,\delta}}(x) = 1 - \left(\frac{\delta}{\delta + x}\right)^{\gamma} \tag{3}$$

$$H_{\alpha,\beta,\gamma,\delta,\rho} = H_{x \mapsto \rho^x}\left(W_{\alpha,\beta}, P_{\gamma,\delta}\right) \tag{4}$$

TABLE II. Parameters of the fitting hybrid Weibull-Pareto distribution.

| Time | $\alpha$ | $\beta$ | $\gamma$ | $\delta$ | $\rho$ | RMSR |
|---|---|---|---|---|---|---|
| 5 am | 1.15e+00 | 2.49e-01 | 1.30e+02 | 9.81e+00 | 1 | 2.63e-03 |
| 8 am | 5.94e+00 | 4.07e-01 | 8.84e+02 | 1.35e+01 | 1 | 3.25e-03 |
| 5 pm | 2.110e+01 | 5.231e-01 | 4.737e+03 | 2.572e+01 | 1 | 5.16e-03 |
| 8 pm | 1.71e+01 | 5.79e-01 | 1.72e+04 | 1.20e+02 | 1 | 7.99e-03 |

After identifying the popular applications among the users in Lyon and generalizing the data consumption behavior, we now look into the total data consumption spatially as portrayed in Fig. 7. The color shades in the Voronoi cells represent the volume of data consumed every hour. The dark shades signals high data consumption whereas lighter ones the contrary. We highlight the Lyon urban region with red boundaries in Fig. 7 (a). We see that as the time approaches the morning peak, most of the cells in the city center and those that cover the freeways indicate high data consumption. This behavior can be expected as morning peaks are popular for home-office commutes. At noon and at 1 pm, we see a slightly higher consumption as it corresponds to the lunch break. The time between 5 pm to 7 pm shows an evenly distributed data consumption after the office hours. The outlook from the spatial analysis is that data consumption is tightly time bounded - as a result, knowing the exact data traffic content can readily reduce the delivery delay. The behavior in the evening peak in the outskirts of the city is surprising and can be associated with a higher latency than in the city center. This scenario showcases the arena where mobile cloud architectures could be advantageous.

## IV. Conclusion

We unintrusively and anonymously collected real-world data from a telecom operator spanning a representative European metropolitan area network and studied the user density and data consumption load at the base station level and at the radio network level. We found that most of the traffic consumption behavior is shaped by few application types, mainly video streaming and web. Later this empirical consumption behavior was mapped to best fit statistical functions.

## Acknowledgment

The authors would like to thank Alexandre Linte, in-charge of Orange Hadoop Corporate Infrastructure for his help during Big-Data processing, Marco Fiore and Razvan Stanica from Inria for their useful suggestions. This work was funded by the ANR ABCD project (contract nb: ANR-13-INFR-005).

## References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019, *White Paper*, Feb. 2015.

[2] M. Satyanarayanan, P. Bahl, R. Caceres, N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol.8, no.4, pp.14-23, Oct.-Dec. 2009.

[3] A. Ceselli, M. Premoli, S. Secci, "Cloudlet Network Design Optimization," *IFIP Networking*, Toulouse, France, May, 2015.

[4] Mobile-Edge Computing, *Introductory Technical White Paper*, Issue 1, Sept. 2014. https://portal.etsi.org

[5] Consumer OTT VoIP Outlook: 2013 to 2018 by Ovum. Available from: http://fortune.com/2014/06/23/telecom-companies-count-386-billion-in-lost-revenue-to-skype-whatsapp-others/

[6] A. Thusoo, J. Sen Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, R. Murthy. "Hive: a warehousing solution over a map-reduce framework," *VLDB Endowment*, vol.2, no.2, pp. 1626-1629, Aug. 2009.

[7] C. Olston, B. Reed, U. Srivastava, R. Kumar, A. Tomkins, 2008. "Pig Latin: A Not-So-Foreign Language for Data Processing," *ACM SIGMOD international conference on Management of data*, pp. 1099-1110. Jun. 2008.

[8] W. Weibull, "A Statistical Theory of the Strength of Materials," *Generalstabens litografiska anstalts förlag*, 1939.

[9] S. Karpinski, E.M. Belding-Royer, K.C. Almeroth, "Living on the Edge: The Distribution of Flows Across Mobile Nodes in Large Wireless Networks," *Technical Report*, 2005.

[10] T. Strutz, 2010. "Data Fitting and Uncertainty (A practical introduction to weighted least squares and beyond)," *Vieweg+Teubner. ISBN 978-3-8348-1022-9.*