This is the author's accepted version of the article. The final version published by IEEE is O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda and M. Fiore, "France Through the Lens of Mobile Traffic Data", IEEE Network Traffic Measurement and Analysis Conference (TMA), 2023, pp. 1-4, doi: 10.23919/TMA58422.2023.10198971

France Through the Lens of Mobile Traffic Data

Orlando E. Martínez-Durive*†§, Sachit Mishra*†§, Cezary Ziemlicki‡,
Stefania Rubrichi‡, Zbigniew Smoreda‡ and Marco Fiore*
*IMDEA Networks Institute, Spain,†Universidad Carlos III de Madrid, Spain, ‡Orange Innovation, France.
{orlando.martinez,sachit.mishra,marco.fiore}@imdea.org, {cezary.ziemlicki,stefania.rubrich,zbigniew.smoreda}@orange.com

***Abstract*—Mobile usage data have shown unprecedented potential for data-driven research in various fields such as demography, sociology, geography, urban studies, criminology, and engineering. However, the lack of reference datasets limits research methods, results, verifiability, and reproducibility of outcomes hindering innovation opportunities. We release a novel mobile usage dataset offering a rare opportunity for the multidisciplinary research community to access rich mobile data of the spatiotemporal consumption of mobile applications in a developed country. The generation process of the dataset forms a new quality standard, leading to information about the demands generated by 68 popular mobile services, geo-referenced at a high resolution of 100×100 m^2 over 20 metropolitan areas in France and monitored during 77 consecutive days in 2019.**

I. Introduction

The surge in the number of mobile devices and Internet services is generating an enormous amount of their usage data, which provides a unique untapped opportunity to study and discovery of new knowledge about human behaviors. The usage data collected in production mobile networks is already proving an invaluable proxy to analyze the habits and behavior of large populations in developed cities or countries, complementing and in some cases replacing traditional sources such as surveys or censuses that are expensive and time-consuming to manage and collect. Examples of substantial utility of mobile network data for research that spans across a plethora of domains can unlock analyses of mobility patterns [1]–[5] and social interactions [6], explorations of transportation systems [7] estimates of static and dynamic population density [8]–[11], predictions of poverty [12], [13], socioeconomic inequality [14], [15] or digital divides [16].

Despite their wide impact, mobile network data are hard to come by. The sensitivity of the information they provide, the concerns for the privacy of the data subjects, and questions on the advantage they could provide to market competitors are all reasons why mobile network operators and service providers typically regard the data as confidential and are not prone to share them with the research community. This limits access to mobile network data, curbing innovation as well as preventing verifiability and reproducibility of the research results whenever permission to use some data is granted under restrictive Non-Disclosure Agreements.

§Equal contributors. The work of O.M., S.M., and M.F. was supported by NetSense (Network Sensing), grant no. 2019-T1/TIC-16037 funded by Comunidad de Madrid. S.R., C.Z., and Z.S. work was supported by CoCo5G (Traffic Collection, Contextual Analysis. Data-driven Optimization for 5G), grant no. ANR-22-CE25-0016, funded by the French National Research Agency (ANR).

In this paper, we present a novel mobile traffic dataset with the following salient features:

- We impart information about the data traffic generated by the mobile devices attached to a modern 4G cellular network, which has been for the past ten years the vastly predominant way of accessing wireless network services. Prior available datasets only focus on Call Detail Records (CDRs) capturing events associated with voice calls and text messages that are sparse and irregular in time.
- The dataset captures mobile data traffic in a developed country like France, thus offering a different perspective than earlier datasets originating from developing countries. Also, the data spans 20 metropolitan areas in France, offering the possibility of generalizing analyses and juxtaposing results across heterogeneous urbanization levels and population densities.
- Unlike any other datasets previously available to the research community through *open challenges*, our dataset adds the novel dimension of mobile services (*i.e.*, 68 major mobile application as shown in Figure 1); this opens up significant opportunities for understanding apps usage and its impact on various research domains.
- The unique creation process of this dataset is a major step beyond the legacy approach of using Voronoi tessellations as a proxy for antenna coverage and results in a dataset of unprecedented spatial accuracy. Specifically, mobile traffic information is mapped onto more than $870,000$ high-resolution regular grids whose individual elements span 100×100 m^2 each.

II. Data sources

The dataset generation hinges on open-source geospatial data and extensive measurements from Orange, a major mobile operator in France with a relatively evenly market share of 30% across France. This provides a solid statistical basis for downstream analysis that can be generalized to the entire local population. Next, we discuss the different data sources and the ethical standards of data collection and processing.

A. Metropolitan areas

We select 20 French metropolises, an administrative zone employed by national and local administrations to carry out joint planning of educational, cultural, economic, and social initiatives. Each metropolis embodies a set of neighboring *communes* which are local administrative zones in France thus includes dense urban, suburban, and rural areas.

B. Mobile network traffic

We use mobile traffic collected over France for 77 consecutive days, from March 16, 2019, to May 31, 2019. The dataset relies on two types of traffic-related information:

1) Service-level traffic volumes: The traffic measurements were performed by Orange using passive measurement probes tapping at the Gi, SGi and Gn interfaces connecting the Gateway GPRS Support Nodes (GGSNs) and the Packet Data Network Gateways (PGWs) of the of Long Term Evolution (LTE) Evolved Packet Core (EPC) network to external public data networks (PDNs). This monitoring strategy allows capturing all 4G traffic traversing the mobile network serving the whole country. The probes run dedicated proprietary classifiers that allow associating individual TPC and UDP traffic flows to the corresponding mobile applications generating them, for purposes that include network monitoring, traffic engineering, and research activities.

2) Traffic flow to eNodeB association: To assign traffic volumes to specific base stations, we resort to Network Signaling Data (NSD) captured by probes monitoring the LTE S1 interface connecting eNodeBs (4G base stations) to the Mobility Management Entity (MME). NSD events allow the association of each traffic flow to the exact sequence of its servicing eNodeBs thus accurately allocating the correct fraction of the total volume of data traffic in the flow to each serving eNodeBs. We update the flows association to eNodeBs at every 15 minutes, hence the whole traffic dataset describes mobile applications traffic records at each eNodeB with that same temporal granularity. We consider 68 mobile applications showing their demand in the selected metropolitan areas.

C. Coverage dataset

For the spatial mapping of the traffic information to the geographical space, we employ coverage information for each eNodeB in the dataset, which was computed using a commercial radio-frequency signal propagation tool. Coverage is encoded as association probabilities into 2-D matrices of dimension 600×600, where each cell represents a square of $10,000$ m^2. Hence, a total area of 60 km^2 is represented. The matrix cells contain the probability $p(i \mid \ell)$, which explains the likelihood that user equipment (UE) would connect to the i^{th} base station while being present at ℓ^{th} location cell. We can also invert this probability information by leveraging Bayes' theorem, *i.e.*, $p(\ell \mid i)$ allowing us to estimate the UE's location or, more generally, any kind of metadata (*i.e.*, traffic). We illustrate the $p(\ell \mid i)$ matrix in Figure 2.

III. Methodology

This section describes the steps to create a time series of traffic maps for an arbitrary application in a given area. These maps contain 4G mobile traffic volumes at a spatial resolution of 100×100 m^2 with a time granularity of 15 minutes. Formally, let us denote by $\mathcal{T}_a^i(t)$ the mobile traffic generated by application a at eNodeB $i \in I$ during time slot $t \in T$, where I is the set of all base stations and T denotes the whole system observation period. Firstly, we calculate $\mathcal{M}_a^i(t)$, a traffic map at eNodeB i of application a at time t and then finally calculate $\mathcal{M}_a(t)$, which is a traffic map for application a at time t.

To achieve our goal, we follow a four-step process as described in Figure 2. **Process (A)** consists of extracting the coverage matrix with the information distribution matrix $\mathcal{P}(\ell \mid i)$ from the coverage dataset and using Bayes' theorem. In **Process (B)**, we take time series of mobile traffic for all selected applications. **Process (C)** is the multiplication phase, where we multiply the traffic data with the coverage matrix to map the space with the traffic information, Figure 2 shows an example where the traffic data for any timestamp t is multiplied by the coverage matrix $\mathcal{P}(\ell \mid i)$. Finally, in the **Process (D)**, all traffic maps from each eNodeB at a given time t are summed to obtain a traffic map for any region.

$$\mathcal{M}_{\mathrm{a}}(t) = \sum_{i=1}^{I} \mathcal{M}_{\mathrm{a}}^{i}(t) \tag{1}$$

We repeat the above steps for all $t \in T$ considering chosen applications, to obtain the time series of traffic maps for each metropolis, a city definition described in section II-A. The resultant dataset is aggregated over all UEs both in space, at eNodeB where we adhere to article 89 of GDPR [17].

IV. Qualitative Analysis

This section illustrates the correctness of the generated dataset based on the intrinsic characteristics of mobile data, such as temporal and spatial usage patterns.

Temporal analysis of the mobile traffic dataset has shown unprecedented potential for tracking daily activity patterns, as people tend to use different applications depending on the time of day, e.g., during office hours, people are more inclined to use work applications during the day, while at night, people tend to use entertainment application. Figure 3(a,b) shows the time series analysis of two distinct mobile applications i.e. Netflix and LinkedIn in Paris, used for entertainment and business purposes respectively. These applications show different usage patterns in their time series, i.e. LinkedIn shows a high traffic peak in the early morning hours until the afternoon, and then the traffic values start to decrease for both the UL and the DL. At the same time, Netflix shows a high traffic peak in the late evening, as people are usually at home at these times. We can observe an almost similar pattern of app usage in the city of Bordeaux by comparing Figure 3(c,d). This pattern of mobile app usage is not limited to Paris. An app is used in the same way in different geographic locations, as seen in Figures 3(a,c) and 3(b,d). We show the time series of Netflix usage in Paris and Bordeaux over one week, indicating a consistent usage trend in these cities.

Spatial analysis leads to an exclusive hypothesis about the more frequented application in different areas of the city based on its usage. Figure 4 presents the average traffic maps for two apps: Netflix and LinkedIn, on different days of the week in Paris. The comparison is made in two dimensions: (1) from left to right, to show that the different apps have different impacts on the region on weekdays or weekends. Netflix, for example,

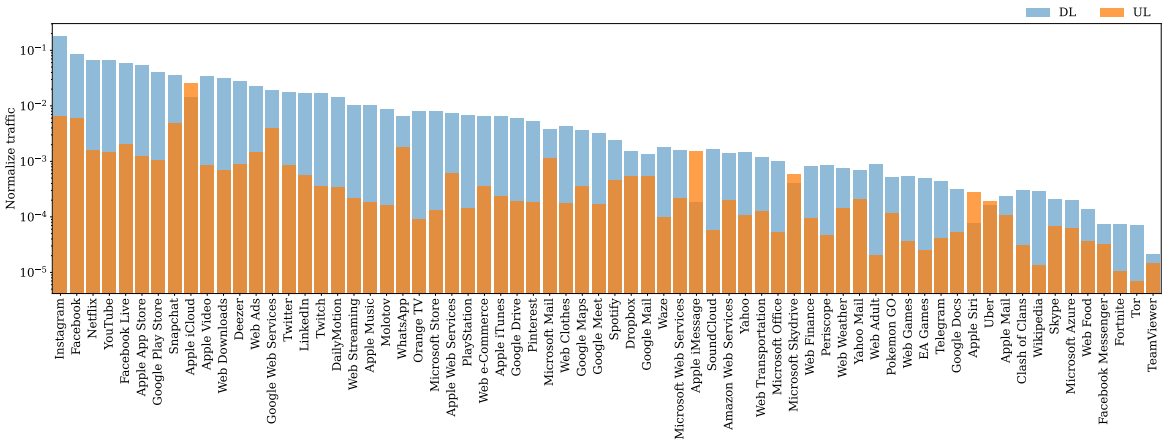


Fig. 1: The normalized proportion of uplink (UL) and downlink traffic (DL) for the 68 mobile applications.

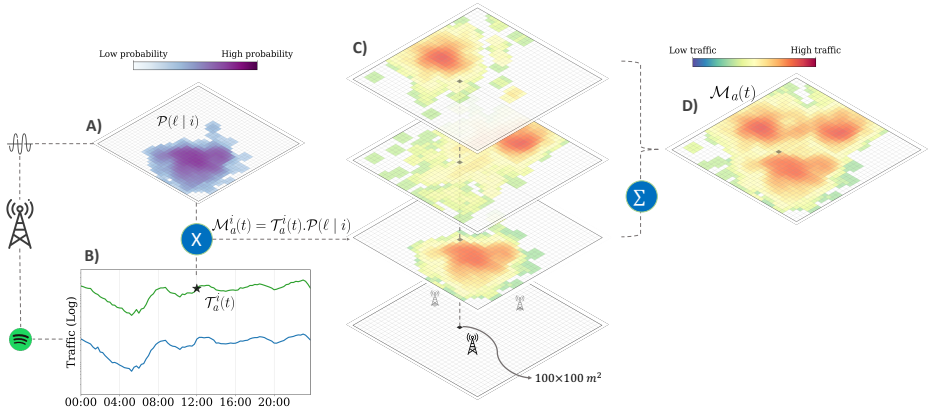
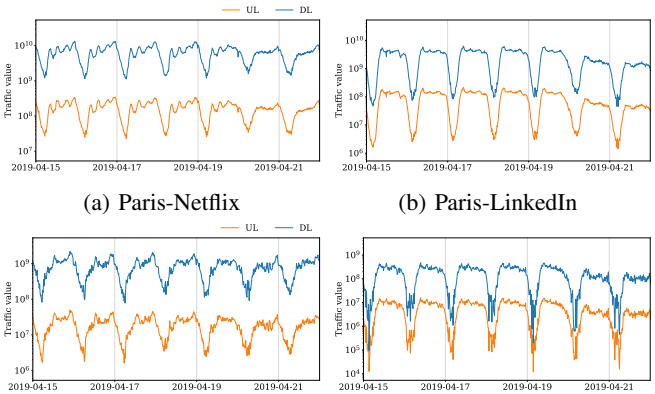


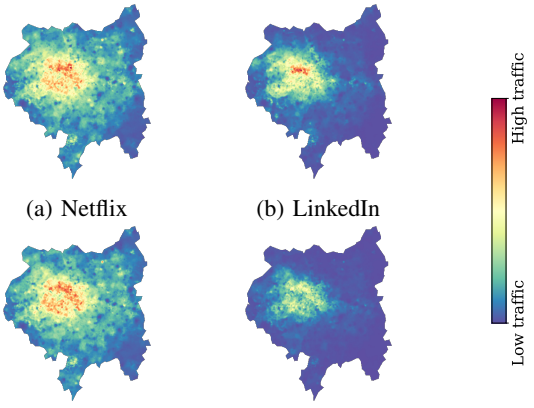
Fig. 2: Methodology for computing the traffic maps; (a) Coverage matrix for an eNodeB i, (b) Traffic time series for Spotify, (c) Multiplication of the coverage matrix with the traffic time series, (d) Summation of all traffic maps.



(a) Paris-Netflix (b) Paris-LinkedIn

(c) Bordeaux-Netflix (d) Bordeaux-LinkedIn

Fig. 3: Traffic time series for Netflix and LinkedIn in Paris and Bordeaux during the same week.



(a) Netflix (b) LinkedIn

(c) Netflix (d) LinkedIn

Fig. 4: Spatial traffic maps for two applications in Paris, on Mondays (top) and Sundays (bottom).

is widely distributed across the region in both lines. On the other hand, LinkedIn has a high concentration in the part of the city where there are many large offices and workplaces. Also in approach (2), i.e., the top-down comparison, we observe a different trend among weekdays and weekend for the same application, e.g., Netflix is more actively used on weekends than on weekdays, while LinkedIn shows an opposite pattern.

V. Conclusion

We present the generation process of a new dataset of mobile network traffic. The result is a high-resolution dataset, which we make available within the context of the NetMob 2023 Data Challenge [18]. The challenge is organized with the NetMob 2023 conference [19] with the aim to derive and exploit new insights from service-level mobile network data.

References

[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[3] B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel, "Exploring the mobility of mobile phone users," *Physica A: statistical mechanics and its applications*, vol. 392, no. 6, pp. 1459–1473, 2013.

[4] K. S. Kung, K. Greco, S. Sobolevsky, and C. Ratti, "Exploring universal patterns in human home-work commuting from mobile phone data," *PloS one*, vol. 9, no. 6, p. e96180, 2014.

[5] T. Louail, M. Lenormand, M. Picornell, O. Garcia Cantu, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, "Uncovering the spatial structure of mobility networks," *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.

[6] G. Miritello, R. Lara, M. Cebrian, and E. Moro, "Limited communication capacity unveils strategies for human interaction," *Scientific reports*, vol. 3, no. 1, pp. 1–7, 2013.

[7] M. Seppecher, L. Leclercq, A. Furno, D. Lejri, and T. Vieira da Rocha, "Estimation of urban zonal speed dynamics from user-activity-dependent positioning data and regional paths," *Transportation Research Part C: Emerging Technologies*, vol. 129, p. 103183, 2021.

[8] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, "Dynamic population mapping using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 15888–15893, 2014.

[9] M. Lenormand, M. Picornell, O. G. Cantú-Ros, T. Louail, R. Herranz, M. Barthelemy, E. Frías-Martínez, M. San Miguel, and J. J. Ramasco, "Comparing and modelling land use organization in cities," *Royal Society open science*, vol. 2, no. 12, p. 150449, 2015.

[10] G. Khodabandelou, V. Gauthier, M. Fiore, and M. A. El-Yacoubi, "Estimation of static and dynamic urban populations with mobile network metadata," *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 2034–2047, 2018.

[11] F. Batista e Silva, S. Freire, M. Schiavina, K. Rosina, M. A. Marín-Herrera, L. Ziemba, M. Craglia, E. Koomen, and C. Lavalle, "Uncovering temporal changes in europe's population density patterns using a data fusion approach," *Nature communications*, vol. 11, no. 1, 2020.

[12] J. E. Steele, P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. de Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, X. Lu, E. Wetter, A. J. Tatem, and L. Bengtsson, "Mapping poverty using mobile phone and satellite data," *Journal of The Royal Society Interface*, vol. 14, no. 127, p. 20160690, 2017.

[13] N. Pokhriyal and D. C. Jacques, "Combining disparate data sources for improved poverty prediction and mapping," *Proceedings of the National Academy of Sciences*, vol. 114, no. 46, pp. E9783–E9792, 2017.

[14] E. Moro, D. Calacci, X. Dong, and A. Pentland, "Mobility patterns are associated with experienced income segregation in large us cities," *Nat Commun*, vol. 12, no. 4633, 2021.

[15] I. Ucar, M. Gramaglia, M. Fiore, Z. Smoreda, and E. Moro, "News or social media? socio-economic divide of mobile service consumption," *Journal of The Royal Society Interface*, vol. 18, no. 185, p. 20210350, 2021.

[16] S. Mishra, Z. Smoreda, and M. Fiore, "Second-level digital divide: A longitudinal study of mobile traffic consumption imbalance in france," in *Proceedings of the ACM Web Conference 2022*, pp. 2532–2540, 2022.

[17] E. Union, "Eu general data protection regulation (gdpr): Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," 2016.

[18] "The NetMob 2023 Data Challenge." https://netmob2023challenge.networks.imdea.org/. Accessed: 2023-05-30.

[19] "NetMob 2023." https://netmob.org/. Accessed: 2023-05-30.