On the intricacies of per individual cellular network datasets generation

Anne Josiane Kouam*, Aline Carneiro Viana*, Alain Tchana†,
* *Inria*, France † *Grenoble INP*, France

***Index Terms*—Cellular networks, CDRs, generative models**

I. Introduction

With mobile devices becoming proxies for human presence and activity, datasets collected by mobile operators – i.e., Call Detail Records (CDRs) – are nowadays acknowledged as a common tool to study human behavior in multiple research domains and industries, such as sociology [1], epidemiology [2], transportation [3], and networking [4] (cf. Fig. 1a). CDRs describe time-stamped and geo-referenced event types (e.g., calls, SMS, data) generated by each mobile device interacting with operator networks (cf. Table I). They comprise city-, region-, or country-wide areas and usually cover long time periods (months or years); no other technology provides an equivalent per-device precise scope today.

Yet, the exploitation of real-world CDRs for research faces many limitations (cf. §II). First, *accessibility*: CDRs datasets are not publicly available, imposing strict mobile operators' agreements. Second, *usability*: CDRs are usually available in an aggregated form (i.e., grouped mobility flows and coarse spatiotemporal information), limiting related analyses' preciseness. Third, *privacy*: even anonymized, CDRs describe sensitive information of users' habits, which hardens their shareability [5]. Fourth, *flexibility*: Restricted access to CDRs limits advanced research requiring data richness in terms of population size, duration, or geographical coverage.

This paper introduces the autonomous generation of realistic CDRs to solve the above challenges. In particular, (1) we detail the motivations of such a solution by establishing the scope of such generated traces and describing how it provides new avenues for research advances, and (2) we share our feasibility study of realistic CDRs generation by presenting the related requirements and challenges.

II. Motivation

We first discuss the striking CDRs' research dependencies and the relevance of enabling realistic CDRs' generation.

*a) **CDRs value recognition**:* Generated by the continuous interaction of a urban-wide population with cellular networks, CDRs represent a rich source of knowledge, valuable to many research communities. For a quantitative appreciation, Fig. 1a identifies as many as 14 different research domains leveraging CDRs, among 100 items selected from a 5-year sample set of 1022 publications (gotten from Google Scholar). This clearly shows a great diversity of domains on this sample only ($\sim 10\%$) and considering the 5-year period.

*b) **Limitations in CDRs exploitation**:* Unlike WiFi networks, cellular networks are mobile operators' exclusive property, hardening outside access to collected CDRs. CDRs access is usually granted through NDAs and is often hardly available for most researchers, time demanding, or imprecise due to privacy laws, bringing *accessibility issues*.

Though strongly necessary, privacy compliance asks for CDRs information aggregation, which hardens their usability and limits the exactness of related investigation. Aggregation usually concerns flows, space, time, and event information in CDRs. E.g., the CDRs available at [6] describe aggregated flows of individuals and their number of generated events per intervals of 10 min and square grids of size 235 meters. This points a lack of *information precision* in available CDRs.

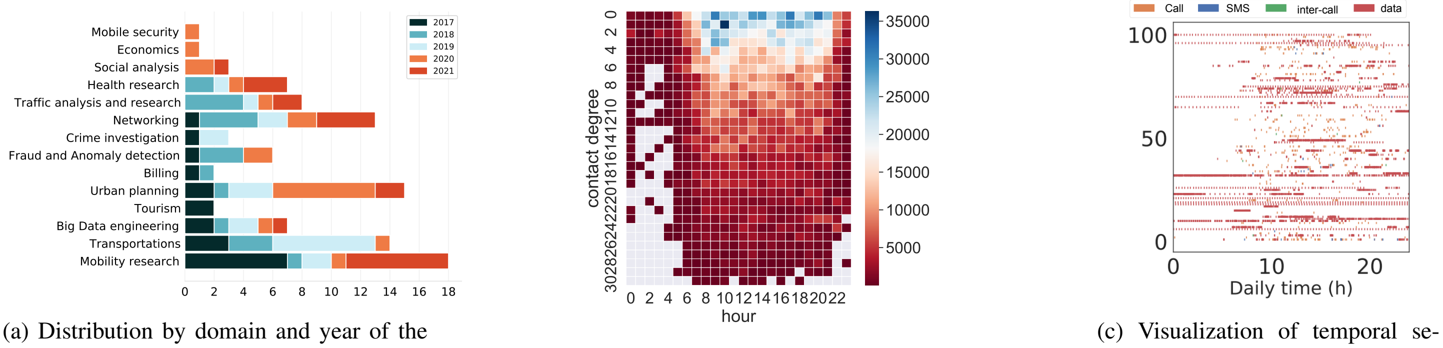
Not surprisingly and justifying the regular CDRs' imprecision issue, personal details of individuals' life habits, inferred from CDRs, calls for privacy-strict exploitation rules and impairs data shareability: e.g., when reconstructed [7] or not [5], majority of individuals' trajectories in CDRs (i.e., higher than 80%) can still be precisely identified, even if anonymized and being sparse in space and coarse in time.

Restricted access to CDRs impacts the *flexibility* of scaling up or adapting CDRs' research results in terms of the population size, the duration, or the covered geographical area, thus limiting advanced research requiring such data richness.

*c) **Conclusion**:* We claim that *an undeniable solution to these limitations is to empower the research community with the flexibility for realistic CDRs' generation that is both adapted to the corresponding research needs and free from the following restrictions: (1) The accessibility to real-world CDRs, (2) The impreciseness of exploiting aggregated CDRs, (3) The impracticality of doing individuals analyses without impeding privacy, and (4) The barrier of not being able to scale up or adapt provided CDRs datasets.* It is worth mentioning the large extent of resulting benefits is likely to profit mobile operators as well as emerging technologies (e.g., 5G/6G), services and applications (e.g., Tactile Internet), or emergencies (e.g., COVID epidemic understanding).

III. CDRs generation requirements

We elaborate on the requirements generated CDRs should meet to ensure broad applicability and reliably reproduce real-world CDRs. First, we present an overview of CDRs gener-



(a) Distribution by domain and year of the most relevant (sorted by Google Scholar) publications using CDRs from 2017 to 2021

(b) Total call duration over a week as a function of contact degree and time

(c) Visualization of temporal sequences of users' events for a real-world CDRs dataset.

Fig. 1

ation; then, we identify five attributes making the generation of realistic CDRs surprisingly challenging and complex.

*a) **Generation overview:*** Mobile traffic generation involves synthesizing timestamped datasets (from x_0 to x_T) only from a given context taken as a parameter, which is much complex traffic prediction, i.e., estimating the next data value at time T based on records from $t = 0$ to $T - 1$. Hence, a generative framework should be expressive enough to fully learn from provided datasets as a training phase, infer the inherent data distributions, and produce new datasets with identical distributions, i.e., realistic. Such frameworks are commonly named *generative models* [8], and while there have been some advances in the literature to efficiently build generative models (e.g., GAN for fake images synthesis), we show in the following the challenges still present for the case of CDRs dataset generation. In particular, we leverage a real-world, fully anonymized CDRs dataset provided by a major network operator in Africa to unveil the intricacies inherent to CDRs, making them difficult to model and generate.

*b) **Modeling inter-features correlations:*** A CDRs dataset comprises both mobility and traffic fields. Mobility fields are essentially user positions (i.e., network cell Ids), while traffic fields are related to network event types, i.e., call, SMS, and data, as described in Table I. CDRs therefore describes three-fold timely behaviors of network users: mobility (*where*), traffic (*what, how*), and social (*whom*) ones. An ideal CDRs generation model should be able to capture the implicit correlations between these features. For instance, Fig. 1b illustrates how social closeness to users' contacts (*social feature*) steadily impacts the hourly duration of calls made (*traffic feature*). Modeling such correlations is unfortunately not a straightforward task, and has never been addressed in the literature. Indeed to the best of our knowledge, state-of-the-art CDRs generation contributions provide the modeling of individual CDRs feature, e.g., [9] for mobility, [10] for data traffic, [11] for social properties.

TABLE I: CDRs fields classified into described users' features

	CDR field
General	Phone number IMEI Timestamp
Traffic	Event-type (call/SMS/data) Call duration Data session size
Social	Phone number of the correspondent
Mobility	Cell Id

.

*c) **Controllability:*** Generated CDRs should be used for a variety of case scenarios (cf. Fig. 1a). The designed generative model should thus allow users to modify the output CDRs by specifying parameters such as the duration, population size, or the mobility area related to a city's urbanization level, layout and infrastructure. Such a controllable generation calls for conditional generative models, rather than just the more common generation approaches based on classical Generative Adversarial Networks (GANs) for instance.

*d) **Modeling arbitrary network topologies:*** CDRs directly reflect the network topology (i.e., cell towers distribution) of its considered mobility area. Hence, there is a strong dependency between operators' CDRs and their network topology, which has to be captured to produce realistic generated CDRs. Such topology however varies with the considered city, requiring the generative model to be able to condition generation on context with arbitrary spatial size. This is a known non-trivial task in machine learning, as popular multilayer perceptron (MLP) or convolutional neural network (CNN) architectures only operate on input with fixed dimensions. More significantly, cellular network topologies are not regular but consist of heterogeneous cells whose shapes vary with the population density of the corresponding covered zones. This makes it impossible to leverage grid tessellations as commonly done in the literature for spatial coverage modeling [12].

*e) **Modeling temporal dynamics:*** Mobile network traffic demonstrates consistent long-term dynamics correlated to regular human activities (peak and off-peak daily hours, weekly working days and weed-ends, yearly vacation and working months). To some extent, the generated CDRs should faithfully reproduce such dynamics. Unfortunately, this requires first access to long-period-covering real-world data, along with tackling the challenge of learning long-term correlations from acquired datasets. While long short-term memory (LSTM) [13] neural networks are acknowledged as a suitable tool for this latter aspect, the complexity related to multi-various timely dynamics applied to multi-featured CDRs makes the training of such models incredibly thorny.

*f) **Modeling spatiotemporal correlations**:* Mobile traffic datasets include not only spatial or temporal dynamics but also spatiotemporal ones. Specifically, such correlations are induced by human activities in space fluctuations as a function of the time of the day. For instance, in urban life, the office period (9h-17h) presents more traffic events than the after-work one (18h-2h). Such events are concentrated in specific zones corresponding to the working zone of the city. In contrast, the after-work period includes displacement times and night activities, which are not made at specific spots (e.g., people can walk down the streets for their night activity), explaining why events are spread over a broader zone. A good generative model should, therefore, be able to reproduce such spatiotemporal dynamics realistically and regardless of the number of generated users, which is a varying parameter.

*g) **Modeling individuality**:* Last but not the least, reproducing CDRs description per individual demands being able to realistically capture, beyond the global aggregated behavior of the population, the individual behaviors of subscribers in terms of mobility and traffic. While mobility modeling and reproduction is well covered in literature, individuals' cellular traffic reproduction still lacks detailed investigations. In particular, cellular traffic presents a notable heterogeneity that challenges preciseness. As an illustration, Fig. 1c plots the traffic generated during a day by 100 randomly selected users from a real-world CDRs. In the Figure, each line plots a user's sequence of events. We can see a great diversity of users regarding events generation. For example, while some users make predominantly local calls, others make only data; some do not make international calls, and others make it frequently. Similarly, in terms of inter-event time, each user has a singular behavior with events either very timely close together, very far apart, or both. Statistical approaches [10], [14], [15], are limited in reproducing such traffic dynamics as they do not allow per-user modeling but per-user profile (i.e., group of users with similar behavior). This additional challenge refers to multivariate generative modeling, for which current literature is limited to *low-dimensional datasets handling in which each univariate component is independent of the others* [16]. Such techniques are, unfortunately, not adequate for CDRs typically encompassing thousands of network users whose individual timely traffic generations are related by social interactions and to their daily spatiotemporal habits.

IV. Conclusion

Despite the significant value of CDRs datasets, their limited accessibility and usability affect the reproducibility and effectiveness of research in many domains. This paper argues the generation of realistic CDRs is the solution to these limitations. We thus discuss issues to be considered in such solution, which we believe, shed light on the requirements to meet for realistic CDRs generation

References

[1] D. Rhoads, I. Serrano, J. Borge-Holthoefer, and A. Solé-Ribalta, "Measuring and mitigating behavioural segregation using call detail records," *EPJ Data Science*, 2020.

[2] H.-H. Chang, M.-C. Chang, M. Kiang, A. Mahmud, N. Ekapirat, K. Engø-Monsen, P. Sudathip, C. Buckee, and R. Maude, "Low parasite connectivity among three malaria hotspots in thailand," *Scientific Reports*, 2021.

[3] S. Qin, Y. Zuo, Y. Wang, X. Sun, and H. Dong, "Travel trajectories analysis based on call detail record data," in *Chinese Control And Decision Conference*, 2017.

[4] M. Ozturk, A. I. Abubakar, J. P. B. Nadas, R. N. B. Rais, S. Hussain, and M. A. Imran, "Energy optimization in ultra-dense radio access networks via traffic-aware cell switching," *IEEE Transactions on Green Communications and Networking*, 2021.

[5] Y.-A. Montjoye, C. Hidalgo, M. Verleysen, and V. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, 2013.

[6] T. Italia, "Telecommunications - SMS, Call, Internet - MI," 2015. [Online]. Available: https://doi.org/10.7910/DVN/EGZHFV

[7] G. Chen, A. C. Viana, M. Fiore, and C. Sarraute, "Complete Trajectory Reconstruction from Sparse Mobile Phone Data," *EPJ Data Science*, 2019.

[8] H. GM, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Computer Science Review*, 2020.

[9] M. Zilske and K. Nagel, "Studying the accuracy of demand generation from mobile phone trajectories with synthetic data," *Procedia Computer Science*, 2014.

[10] E. M. R. Oliveira, A. C. Viana, K. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets," *Computer Networks*, 2017.

[11] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi, "On the structural properties of massive telecom call graphs: Findings and implications," in *ACM CIKM*, 2006.

[12] K. Xu, R. Singh, M. Fiore, M. K. Marina, H. Bilen, M. Usama, H. Benn, and C. Ziemlicki, "Spectragan: Spectrum based generation of city scale spatiotemporal mobile network traffic data," in *IEEE CoNEXT*, 2021.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[14] A. Murtić, M. Maljić, S. L. Gruičić, D. Pintar, and M. Vranić, "Sna-based artificial call detail records generator," in *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018.

[15] M. Songailaitė and T. Krilavičius, "Synthetic call detail records generator," *CEUR Workshop proceedings*, 2021.

[16] M. Hofert, A. Prasad, and M. Zhu, "Multivariate time-series modeling with generative neural networks," *Econometrics and Statistics*, 2022.