Chen et al. *EPJ Data Science* (2019) 8:30
https://doi.org/10.1140/epjds/s13688-019-0206-8

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Complete trajectory reconstruction from sparse mobile phone data

Guangshuo Chen[1,2*], Aline Carneiro Viana[1], Marco Fiore[3] and Carlos Sarraute[4]

*Correspondence:
guangshuo.chen@inria.fr
[1]INRIA, Université Paris-Saclay, Palaiseau, France
[2]École Polytechnique, Université Paris-Saclay, Palaiseau, France
Full list of author information is available at the end of the article

Abstract

Mobile phone data are a popular source of positioning information in many recent studies that have largely improved our understanding of human mobility. These data consist of time-stamped and geo-referenced communication events recorded by network operators, on a per-subscriber basis. They allow for unprecedented tracking of populations of millions of individuals over long periods that span months. Nevertheless, due to the uneven processes that govern mobile communications, the sampling of user locations provided by mobile phone data tends to be sparse and irregular in time, leading to substantial gaps in the resulting trajectory information. In this paper, we illustrate the severity of the problem through an empirical study of a large-scale Call Detail Records (CDR) dataset. We then propose Context-enhanced Trajectory Reconstruction, a new technique that hinges on tensor factorization as a core method to complete individual CDR-based trajectories. The proposed solution infers missing locations with a median displacement within two network cells from the actual position of the user, on an hourly basis and even when as little as 1% of her original mobility is known. Our approach lets us revisit seminal works in the light of complete mobility data, unveiling potential biases that incomplete trajectories obtained from legacy CDR induce on key results about human mobility laws, trajectory uniqueness, and movement predictability.

Keywords: Human mobility; Mobile phone dataset; Data enrichment; Seamless trajectory reconstruction; Location predictability; Trajectory unicity

1 Introduction

Over the past ten years, the proliferation of mobile devices indirectly fuelled many original studies on human mobility, as mobile phone data collected by network operators have become an invaluable source of information about individual movement patterns at scale. Data extracted from mobile communications have supported many breakthroughs [1, 2], including seminal works demonstrating that individual movements are regular [3] and predictable [4], as well as unique [5]. Mobile phone data have allowed unveiling major patterns in human movements, such as home-work commuting [6], or frequent and recurrent motifs [7]. The novel understanding of individual movements enabled by network operator data has facilitated innovation in, *e.g.*, transportation systems [8], urban planning [9], communication infrastructure deployment [10, 11], and epidemics control [12].

Call Detail Records (CDR) are the current de-facto standard for mobile phone data used in human mobility studies. As illustrated in Table 1, CDR are logs of events generated dur-

Table 1 The sample of CDR dataset. In this case, each CDR entry describes a voice call. All user identifiers and cell tower ids are coded to hide personal information. The GPS location of the cell towers is provided in the data source

UserID	Event Time	Cell Tower	Caller	Callee	In/Out	Duration (s)
38DA6	2015-05-01 18:26:50	1921	38DA6	163B7	Out	52
78EC3	2015-05-01 14:16:09	2189	53808	78EC3	In	600
9FAFE	2015-05-01 23:20:09	2189	9FAFE	7BBF1	Out	41
708A2	2015-05-01 08:21:10	1988	96EC4	708A2	In	37
A27AD	2015-05-01 21:51:09	2189	EA33F	A27AD	In	108
9F3C7	2015-05-01 13:21:25	20102	C5691	9F3C7	In	134
D4578	2015-05-01 17:03:46	20103	D4578	5B9A3	Out	30
F904A	2015-05-01 23:24:03	1998	F904A	F3C88	Out	10
4CCEA	2015-05-01 20:11:38	21104	4CCEA	5EF18	Out	438
A77B8	2015-05-01 09:40:26	21104	A77B8	BD3E5	Out	33
D5761	2015-05-01 20:34:40	1999	DAA24	D5761	In	600

ing mobile communications, such as voice calls, text messages, and possibly mobile data traffic sessions, which are collected for billing purposes [2]. Those events are associated with individual mobile devices and are time-stamped and geo-referenced, which makes CDR an obvious proxy for the trajectories of the mobile network subscribers. Interestingly, CDR encompass huge populations (*e.g.*, millions of users) over large (*e.g.*, citywide or nationwide) geographic regions, and cover long periods (*e.g.*, months or years): they thus enable human mobility investigations at unprecedented scales.

A drawback of CDR-based trajectories (or simply *trajectories* hereafter) is that they are often largely incomplete. As a matter of fact, telecommunication events are punctual and provide information at specific time instants, which are also sparse and irregularly distributed in time. Thus, CDR offer a very partial view of the overall mobility of each user, with substantial continuous periods where location information is entirely absent [13]. From a research standpoint, the reduced completeness of CDR yields severe consequences: (i) as many studies require exhaustive knowledge of personal trajectories, they focus on few highly active individuals that generate a large number of CDR events, and discard the rest—in several important works, this amounts to ignoring over 98% of the subscriber population [3, 4]; (ii) when the whole user base is considered, the dependability of the results is questioned by the fact that they are derived from partial movement information, and potential biases due to missing locations are difficult to assess [14].

A solution to the problem of mobile phone data sparsity is *trajectory reconstruction*, which aims at completing individual movement information by recovering the unspecified positions of each user. However, as detailed in Sect. 2, the literature on trajectory reconstruction for mobile phone data is relatively thin, and mostly focuses on a *partial* reconstruction that still leaves (possibly large) gaps in our knowledge of personal mobility. In this paper, we present a novel and more effective approach to trajectory reconstruction, named *Context-enhanced Trajectory Reconstruction* (CTR), which allows completing *all* missing trajectory data with reasonable accuracy. CTR enables human mobility analyses over much larger populations, and without biases due to incomplete information. We apply CTR to the problem of reconstructing trajectories from CDR, since they represent the primary type of movement data employed for large-scale human mobility analysis [1, 2]. However, our method is general and can be applied to other classes of mobile phone data; specifically, by running CTR on data originally collected at a frequency of a few hundreds samples per day, one could aim at completing individual trajectories at high temporal res-

olutions, *e.g.*, of minutes. The design, evaluation, and application of our solution yield the following contributions.

- First, we provide evidence of the severe sparsity that affects mobile phone data, by analyzing the CDR of 1.8 million users collected during three consecutive months. We quantify the phenomenon by means of relevant metrics, including the duration, sampling frequency, and *completeness* of individual trajectories in the data. In particular, our results show that legacy preprocessing techniques that discard users based on arbitrary completeness thresholds, in fact, ignore a vast user population with substantial and potentially serviceable mobility information. Details are in Sect. 3.
- Second, we introduce our novel approach, CTR, which leverages well-known features of human mobility to customize *tensor factorization* in a way that befits our problem. This original methodology lets CTR transform sparse mobile phone data into seamless individual trajectories that span the full dataset duration, for all users. Details are in Sects. 4 and 5.
- Third, we validate the proposed strategy with ground truth data. Comprehensive performance evaluation shows that CTR achieves full reconstruction of individual trajectories on an hourly basis, with a median displacement between 1 and 2 network cells that depends on the sparsity of the original CDR data. This effectively means that, in the reconstructed data, a user is typically placed in the correct cell or in one that is very close to it. Such a level of accuracy is acceptable for metropolitan-scale analyses (where the urban surface is typically covered by hundreds of cells) and is excellent for national-scale studies (as inter-city mobility is perfectly captured). Details are in Sect. 6.
- Fourth, we demonstrate the importance of trajectory reconstruction in CDR-driven human mobility analysis. Specifically, we revisit three seminal studies [3–5] by using the complete mobility of 1.7 million users, instead of the incomplete trajectories of a small fraction of especially active users as in the original works. Our results show that key results in these studies may change, even quite dramatically, in presence of complete mobility, proving that trajectory reconstruction is an indispensable first step for analyses of human mobility that rely on mobile phone data. Details are in Sect. 7.

We then draw conclusions and comment on the perspectives of our work in Sect. 8.

2 Related work

A number of strategies adopted in mobile phone data analysis aim at mitigating the sparsity and irregularity of the individual trajectories extracted from CDR. We classify current solutions depending on whether they are targetting (i) time discretization, (ii) user filtering, or (iii) trajectory reconstruction, and discuss them below.

2.1 Time discretization

A very common practice in mobile phone data analyses, time discretization reduces the temporal resolution of the data so as to make trajectories appear complete. Indeed, discretizing time allows associating a single location to each time step, implying that the location is valid across the whole interval covered by the step. We remark that time discretization is basically unavoidable in CDR, due to the limited granularity of mobile communication events. Interestingly, however, it is performed at quite different temporal resolutions by works in the literature, spanning from 15 minutes [14] to 1 hour [3, 4], or even

to 2 hours [5]. Also related to data alteration in time, some studies focus on specific periods (*e.g.*, work hours, commuting hours, nighttime hours), which allows them to treat the (time-discretized) data as complete even if just such target periods are covered [13]. Ultimately, time discretization can be regarded as a data preprocessing step, which does not explicitly address the problem of trajectory incompleteness in CDR: indeed, it does not offer any guarantee in terms of the fraction of time intervals during which location information is available for a given user.

2.2 User filtering

It is customary for works on mobile phone data to filter the available user base before further processing. User filtering can be adopted to eliminate outliers (*e.g.*, call centers [15]), but it is more often considered to retain trajectories that provide sufficient mobility information. Indeed, when extensive knowledge of the user mobility is required by the nature of the investigation, the straightforward way to deal with incomplete trajectory data is that of filtering users so as to exclude from the analysis those with insufficient location information. This is typically performed by imposing thresholds on the amount of mobility data available in CDR, though there is hardly a standard to define them. As a reference, several important works opt for a choice of users who have a minimum average sampling frequency of 0.5 events per hour, and who are observed at two unique locations at least [3, 4, 14].

User filtering comes with a cost in terms of below-threshold users that are excluded from the analysis. Unfortunately, in practical cases involving CDR datasets, even the loose requirements set by the thresholds mentioned above risk to lead to severe reductions in the examined user population. For instance, two very well-known studies analyze just 1.67% [3] and 0.45% [4] of the total subscribers available in their datasets upon filtering, leaving substantial richness in the original data untapped.

2.3 Trajectory reconstruction

The objective of trajectory reconstruction is to infer the positions of the users at times during which the original mobile phone data does not provide such information. By recovering the missing location data, a successful reconstruction makes user movements more complete, hence effectively obviates the problem of discarding vast portions of users caused by filters on mobility data.

The literature on trajectory reconstruction from mobile phone data, including CDR, is relatively thin, and mostly focuses on *partial* reconstruction. A fairly standard solution is extending the validity of instantaneous locations over longer time intervals, by assuming that users do not move for some periods (*e.g.*, in the order of hours) before and after every recorded event in the CDR dataset [16]. Enhancements to this approach have also been proposed, by dynamically adapting the duration of location validity periods via supervised learning [13, 17]. A different strategy hinges on probabilistic reconstruction, implemented by modeling positions between two CDR events as random variables [18]. However, all these techniques incur into non-negligible spatial errors, and yet, as they target partial reconstruction, cannot return full trajectories that cover the entirety of time [13].

Spatiotemporal interpolation [19] is the only method proposed to date for complete trajectory reconstruction, *i.e.*, a reconstruction that returns seamless location information for each user at the target time discretization. Multiple interpolation methods can

be adopted to this end, and the approach has been proven effective with dense CDR entries featuring more than 1000 events per day. However, users with such a high level of communication activity are infrequent, and indeed only 0.07% of the total subscribers in the original CDR dataset are retained for analysis in [19]. When applied to typical CDR data, interpolation methods are ineffective and provide no clear improvement over a naive error-prone method where users are associated with the last location they are seen at [17].

The solution we propose in this paper also falls in the trajectory reconstruction category, and, like interpolation, aims at inferring complete individual trajectories. However, we seek an approach that effectively also operates on very sparse positioning data, and thus can be applied to entire CDR datasets.

3 Sparsity in CDR-based trajectory data

As a preliminary step to our study, we carry out a data-driven investigation of sparsity in individual trajectories extracted from mobile phone data. Our goal is quantifying the phenomenon, so as to demonstrate the significance of the problem and to motivate the need for a dedicated solution. To this end, we employ a CDR dataset collected from 1.8 million prepaid mobile network subscribers of a major Latin America mobile network operator. The data consists of 778 million voice call entries generated by customers during a consecutive 90-day period. The customers appear within a large geographic area of several thousands of km^2 covered by 4200 cell towers. As illustrated in Table 1, each CDR entry contains the caller's and the callee's hashed identifiers, the call duration, the event time, and the identifier of the cell tower which the caller's device was connected to when the call originated. The mapping of each cell tower identifier and its location in latitude/longitude is also provided by the data source. The results presented below refer to the CDR-based trajectories of all users in the dataset.

3.1 Trajectory completeness

We first quantify the amount of missing information in the individual trajectories. To this end, we assume a total observation time $\mathcal{T}$ and discretize it into time intervals of duration τ, which we will refer to as the *temporal resolution*. We define the *completeness* of a specific trajectory as the fraction of time intervals for which we have at least one location sample. As an example, let us assume that $\mathcal{T} = 7$ days and the temporal resolution is $\tau = 1$ hour: then, a trajectory with locations in 80 different hours has a completeness of $80/(7 \times 24) = 0.476$.

In practice, $\mathcal{T}$ corresponds to the duration of the dataset from which the trajectories are extracted, and τ to the temporal resolution required for the mobility analysis. As both values may vary based on the nature of the mobility study, we investigate their impact on completeness through comprehensive parametric analysis. Instead, our notion of completeness is independent of how positions are determined in the presence of multiple samples in the same time interval. Hence we do not need to account for that aspect at this stage.

Figure 1 portrays the cumulative distribution function (CDF) of the completeness of 1.8 million trajectories from our reference dataset when different combinations of $\mathcal{T}$ and τ are considered. Each plot refers to a different temporal resolution τ, while curves identify different observation periods $\mathcal{T}$. We observe the following.

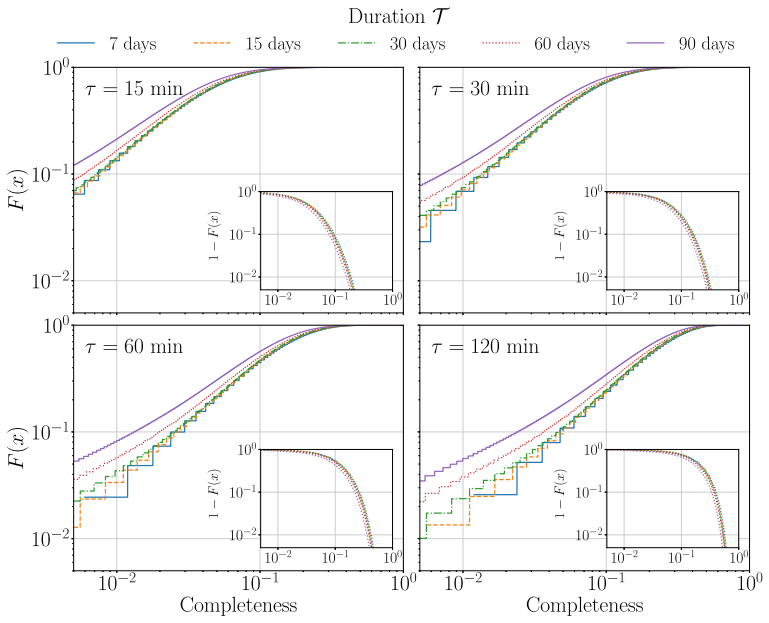

Figure 1 CDF of completeness. Cumulative distributions of the completeness of trajectories over 1.8 million users for different combinations of the observing duration $\mathcal{T}$ and the temporal resolution τ

- Seamless trajectories with completeness equal to 1 are extremely rare in mobile phone data, under any combination of $\mathcal{T}$ and τ. Even with short observation periods ($\mathcal{T} = 7$ days) and low resolution ($\tau = 2$ hours), not a single complete trajectory is extracted from our reference dataset, despite the large user population. In fact, less than 1% of users would satisfy a minimum completeness requirement of 0.5, which allows half of the locations to be unknown; and, more than 20% of trajectories do not even meet the degree of completeness 0.1, *i.e.*, miss 9 locations out of 10. The fact that such poor figures refer to the best case, *i.e.*, the less demanding combination of $\mathcal{T}$ and τ, illustrates well the severity of the sparsity problem in CDR datasets.
- The duration of the observation period has a marginal impact on trajectory completeness. More precisely, and as one would expect, slightly higher completeness is recorded for trajectories that span a shorter observation period when $\mathcal{T} \in [30, 90]$ days; however, and quite interestingly, completeness hardly varies when $\mathcal{T} \leq 30$ days. We argue that the result can be linked to the weekly periodicity of human activities (which entail a reduced difference for $\mathcal{T} \in [7, 30]$ days), plus seasonal effects such as summer holidays that only intervene over long timescales ($\mathcal{T} \geq 30$ days in our dataset).
- The temporal resolution has a stronger effect on completeness. For instance, when $\mathcal{T} = 30$ days, only 8% of the trajectories have completeness above 0.1 when $\tau = 15$ minutes, whereas the same percentage grows to 26%, 53% and 75% when $\tau = 30$, 60 and 120 minutes, respectively. To better understand the scaling of completeness with the temporal resolution, we show in Fig. 2 the mean and median values of completeness, computed across all users, versus τ. We conclude that completeness

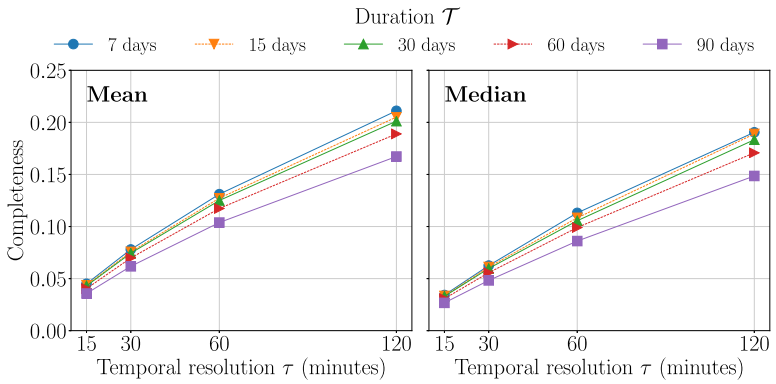

Figure 2 Completeness versus temporal resolution. The left and right plots show the mean and median of the completeness of trajectories, respectively, as the function of temporal resolution τ

grows almost linearly (Pearson correlation coefficients are at 0.981 and 0.983, for the mean and median, respectively) with τ in the range 15 to 120 minutes, under all observation periods $\mathcal{T}$. The takeaway message is that one cannot easily escape the tradeoff that higher temporal resolution incurs into curbed completeness in CDR data.

Overall, our results highlight that the vast majority of trajectories from a typical CDR dataset only capture a rather small portion of the complete user mobility. Based on Fig. 2, we quantify the fraction of known positions between 5% and 20% on average, depending on the temporal resolution needed for the analysis. These results are well aligned with the outcome of user filtering in previous studies that leverage mobile phone data, where only a small subset of users displays sufficient completeness and is retained for analysis [3, 4, 19].

3.2 A statistical model of completeness

As a further contribution to the characterization of the sparsity problem, we derive a data-driven model of CDR-based trajectory completeness. The model can be leveraged to obtain a quick, approximate estimate of the completeness of trajectories extracted from a CDR dataset, by merely providing the timespan of the CDR data and the desired temporal resolution.

In order to derive the model, we fit theoretical distributions to the completeness probability density function (PDF) and link the distribution variables to the system parameters $\mathcal{T}$ and τ. Earlier studies have shown that mobile communication activity tends to follow long-tailed distributions across users [20]. Since the usage patterns of mobile devices directly determine completeness, it makes sense to look at heavy-tailed theoretical distributions. We run a maximum likelihood estimation of the variables of six standard long-tailed distributions (Generalized Pareto, Lévy, Power-law, Lognormal, Gamma, and Weibull) on the empirical PDF of completeness. We then evaluate the quality of fitting via the coefficient of determination R^2, and by running the Kolmogorov-Smirnov test to obtain the D_{KS} statistic.

Table 2 summarizes the results. Goodness-of-fit tests consistently point at either of two distributions, *i.e.*, Weibull and lognormal, depending on the combination of system parameters. An illustrative example, for $\mathcal{T}$ = 60 days and τ = 60 minutes, is provided in

Table 2 Empirical fittings of the completeness of trajectories. Fitting quality of six standard long-tailed distributions to the empirical PDF of trajectory completeness, in terms of R^2 and D_{KS}. Rows refer to different combinations of observation period $\mathcal{T}$ and time resolution τ. Best-fit R^2 and D_{KS} are highlighted in bold

Duration	Resolution	**Weibull**		Lognormal		Gamma		Pareto		Levy		Power law	
$\mathcal{T}$	τ	D_{KS}	R^2	D_{KS}	R^2	D_{KS}	R^2	D_{KS}	R^2	D_{KS}	R^2	D_{KS}	R^2
7 d	15 min	0.0334	**0.9990**	**0.0318**	0.9973	0.3710	0.4679	0.0495	0.9969	0.2509	0.8046	0.3538	0.5031
	30 min	**0.0302**	**0.9993**	0.0345	0.9967	0.0548	0.9962	0.1716	0.8973	0.2649	0.7848	0.3587	0.4894
	60 min	**0.0278**	**0.9990**	0.0372	0.9958	0.0475	0.9977	0.1198	0.9516	0.2888	0.7506	0.4162	0.2041
	120 min	**0.0375**	**0.9972**	0.0443	0.9938	0.0629	0.9853	0.1043	0.9768	0.3238	0.6977	0.2456	0.7450
15 d	15 min	0.0264	**0.9986**	**0.0233**	0.9984	0.3594	0.5007	0.0627	0.9893	0.2620	0.7853	0.3949	0.4120
	30 min	**0.0208**	**0.9995**	0.0264	0.9980	0.0271	0.9991	0.0724	0.9851	0.2765	0.7647	0.3576	0.4975
	60 min	**0.0188**	**0.9997**	0.0279	0.9974	0.0257	0.9987	0.0935	0.9719	0.2997	0.7315	0.3176	0.6076
	120 min	**0.0254**	**0.9987**	0.0332	0.9961	0.0913	0.9572	0.1880	0.8780	0.3349	0.6792	0.2721	0.7116
30 d	15 min	0.0239	**0.9985**	**0.0207**	**0.9985**	0.3514	0.5261	0.0700	0.9835	0.2619	0.7872	0.3829	0.4365
	30 min	0.0205	**0.9992**	0.0216	0.9983	**0.0203**	0.9991	0.1528	0.8976	0.2763	0.7661	0.3912	0.4149
	60 min	**0.0149**	**0.9996**	0.0263	0.9975	0.0289	0.9982	0.0895	0.9702	0.3003	0.7346	0.3131	0.6212
	120 min	**0.0239**	**0.9984**	0.0315	0.9962	0.0995	0.9479	0.1069	0.9538	0.3337	0.6893	0.3154	0.6176
60 d	15 min	0.0266	**0.9980**	**0.0239**	0.9977	0.1990	0.8284	0.0458	0.9934	0.2527	0.8076	0.3850	0.4156
	30 min	**0.0233**	**0.9985**	0.0234	0.9976	0.0286	0.9974	0.1944	0.8669	0.2649	0.7903	0.3897	0.4212
	60 min	**0.0207**	**0.9985**	0.0264	0.9970	0.0310	0.9980	0.0772	0.9769	0.2883	0.7622	0.3451	0.5635
	120 min	**0.0245**	**0.9975**	0.0316	0.9958	0.0754	0.9750	0.0946	0.9617	0.3209	0.7196	0.3491	0.5070
90 d	15 min	**0.0298**	**0.9954**	0.0329	**0.9954**	0.3619	0.5248	0.0322	0.9954	0.2371	0.8390	0.3866	0.4528
	30 min	0.0336	**0.9958**	**0.0335**	0.9950	0.0583	0.9864	0.0398	0.9942	0.2487	0.8264	0.3819	0.4774
	60 min	**0.0305**	**0.9960**	0.0355	0.9944	0.0454	0.9913	0.0611	0.9843	0.2705	0.8020	0.3231	0.6123
	120 min	**0.0369**	**0.9940**	0.0379	0.9932	0.0486	0.9927	0.0760	0.9743	0.2998	0.7687	0.3359	0.5885

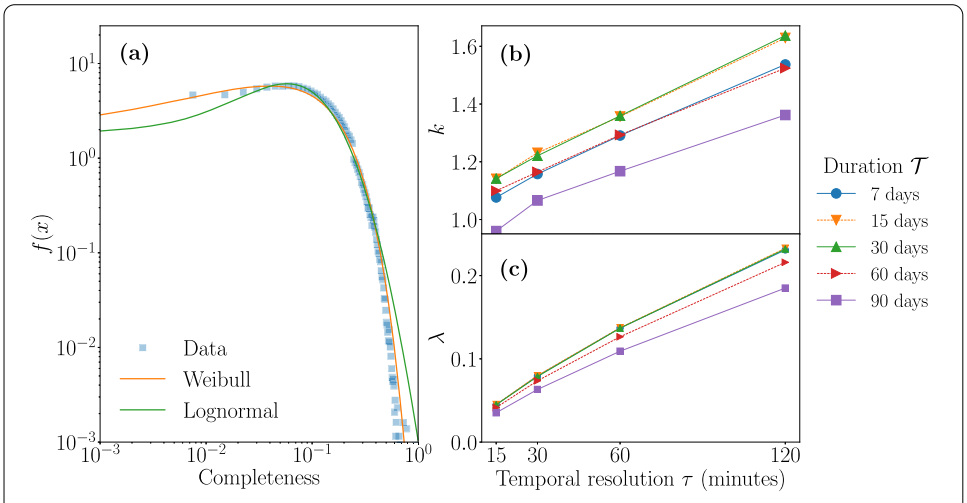

Figure 3 Fitting completeness into long-tailed distributions. (**a**) Empirical PDF of the completeness of 1.8 million user trajectories, with Weibull and lognormal theoretical fittings, for $\mathcal{T} = 60$ days and $\tau = 60$ minutes. (**b**)(**c**) Weibull variables k and λ versus the temporal resolution τ and observation period $\mathcal{T}$

Fig. 3(a). The Weibull distribution emerges as a clear winner, as it prevails in 32 out of 40 cases in Table 2.

The PDF of the Weibull distribution is expressed as $f_{\text{Weibull}}(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}e^{-(x/\lambda)^k}$. Plots (b) and (c) of Fig. 3 show that the estimated Weibull parameters k and λ scale linearly (Pearson correlation coefficients of 0.89 and 0.98, respectively) with the most important system

parameter, τ. Also, the observation time $\mathcal{T}$ introduces an offset to the linear relationship. We thus model $\hat{k} = [\tau \times (4.569 - 0.01037\mathcal{T}) + 1086.5 - 1.14515\mathcal{T}] \times 10^{-3}$ and $\hat{\lambda} = [\tau \times (1.814 - 0.00379\mathcal{T}) + 26.23 - 0.07659\mathcal{T}] \times 10^{-3}$, and the CDF of CDR-based trajectory completeness as $f_{\text{Weibull}}(x; \hat{k}, \hat{\lambda})$.

4 Missing location inference

In this section, we formalize the problem of trajectory reconstruction from sparse and irregular sampling and discuss how popular features of human mobility can be leveraged to design a sensible solution to the problem.

4.1 Trajectory reconstruction problem

Let us consider mobile phone data spanning an observation period $\mathcal{T}$. User trajectories are inferred by discretizing $\mathcal{T}$ into intervals of duration τ, hence $\mathcal{T} = \{1, \ldots, N\}$ can be regarded as the set of all time intervals. The actual (*i.e.*, ground-truth) trajectory of a generic individual in the dataset is denoted by $L_{\mathcal{T}} = \{\mathbf{l}_i \mid i \in \mathcal{T}\}$, where $\mathbf{l}_i$ is the *representative location* of the user during the ith time slot. When a single mobile communication event is associated to a time interval i, the representative location $\mathbf{l}_i$ maps to the geographic location associated with the event. This is the geographical position of the antenna that is the most representative of the user location. Although our discussion is general and can accommodate any definition of antenna representativeness, we opt for a strategy of selecting the antenna that handles the most communications during the target time slot i. We remark that this approach allows mitigating the so-called "ping-pong" (also referred to as "flickering" or "oscillation") effect that affects mobile phone data. This phenomenon consists in the mobile device association bouncing across nearby antennas, so that the location proxy for a same user changes even in absence of physical movement. The effect occurs at fast timescales of seconds [21, 22], hence oscillations are highly likely to occur within a single time slot i, which spans tens of minutes or more in our case. In each slot, oscillations are then canceled by the antenna selection process above.

As discussed in Sect. 3, trajectories extracted from mobile phone data are usually incomplete, *i.e.*, no event is recorded during many time intervals, and $\mathbf{l}_i = \emptyset$ for a large fraction of i's. Let $\Omega = \{i \in \mathcal{T} \mid \mathbf{l}_i \neq \emptyset\}$ be the set of time instant for which we have location information for the considered user. We define the *observed trajectory* of non-null locations as $L_\Omega = \{\mathbf{l}_i \mid i \in \Omega\}$; this is the user mobility information that can be derived from the mobile phone data. Let us also denote by $\Omega^C = \mathcal{T} - \Omega$ the time slots set during which the user location is unknown.

The trajectory reconstruction problem is then that of inferring, based on the sole knowledge of L_Ω, an estimated complete trajectory $\hat{L}_{\mathcal{T}} = \{\hat{\mathbf{l}}_i \neq \emptyset \mid i \in \Omega \cup \Omega^C\}$ such that

$$\arg\min_{\hat{L}_{\mathcal{T}}} \frac{1}{|\mathcal{T}|} \sum_{i=1}^{N} \|\mathbf{l}_i - \hat{\mathbf{l}}_i\|, \tag{1}$$

where $\|\cdot\|$ denotes the geographic distance between the two argument locations. The expression in (1) implies that the desired solution minimizes the mean absolute error (MAE) between the positions of the user in the complete ground-truth and reconstructed trajectories, at the same time instant.

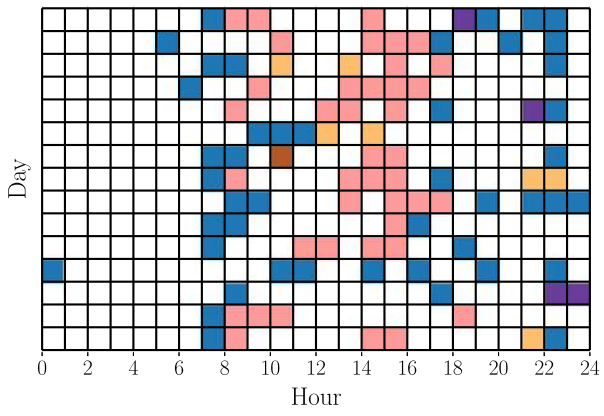

Figure 4 Example of the observed trajectory of a user. The trajectory is recored during $\mathcal{T} = 15$ days and has a degree of completeness in 0.27. Each grid element maps to the representative location associated to each time interval of $\tau = 1$ hour. Colors map to different places: the most active daytime and nighttime locations are marked in pink and blue respectively; white blocks are missing locations. Figure best viewed in colors

4.2 Design rationale

In order to solve the problem in (1), we have to rely on the known locations of a user, L_Ω, to infer those missing. Several well-known properties of human mobility play in our favor and can be leveraged to solve the problem effectively, as follows.

- *Prevalence of static phases.* People tend to spend a substantial amount of their time at a few fixed locations (*e.g.*, workplace, school, shopping mall, sports center), where they linger for long continuous periods in the order of hours. Transitions among these locations are instead fairly rapid: in fact, people typically perceive movement phases between points of actual interest as a waste of time and strive to reduce them to a minimum. This results in a pattern of long static phases with fast movements in between [18]. The prevalence of static behaviors allows adopting temporal resolutions that are granular enough to capture the important stay points of each individual, yet are still tractable for reconstruction. For instance, in the example of Fig. 4, the user tends to stay for long times at the same location during working hours, and considering $\tau = 1$ hour as done in the plot does not lose substantial positioning information during such hours.
- *Overnight invariance.* A user is typically at the same location, *i.e.*, at home, during nighttime, as also demonstrated by the fact that even sparse CDR data can be effectively employed to identify dwelling units [14, 23]. The observation that most nighttime locations match is important for trajectory reconstruction since mobile phone events tend to be especially scarce overnight, yet knowledge gaps can be filled with a limited amount of observed data during night hours. As an illustrative example, the user in Fig. 4 is always found at the same place early in the morning, late in the evening, and once overnight: it is apparent that such a location can be sensibly extended to all night hours, which are otherwise very poorly sampled.
- *Regularity of movement.* Human mobility is strongly regular, from multiple perspectives. People frequently return to a same, limited set of locations [3, 24], which results in repetitive sequences of visits to a few places [7, 24], and infrequent trips to other locations in a (typically large) geographical region [24]. Regularity also occurs in time, as the locations above are visited in highly periodic patterns [4]. When

considered jointly, these results have critical implications for trajectory reconstruction: they suggest that sparse location data may still be highly informative of the mobility of one individual if such data are sufficiently distributed in time to capture diverse moments of the day and week. This is typically the case for mobile phone data, which feature a combination of temporal irregularity of the sampling and long observation periods $\mathcal{T}$. As an example, in the case of Fig. 4, the fact that the user generates CDR events at different moments of the morning, working hours, and evening allows depicting a fairly complete picture of her mobility pattern during a typical day, by observing mobile communication samples for a sufficiently long amount of time (15 days in the considered sample).

5 Context-enhanced trajectory reconstruction

Context-enhanced Trajectory Reconstruction (CTR) is a novel trajectory reconstruction approach that solves the problem in (1), by receiving the observed trajectory L_Ω as the input, and generating an estimated complete trajectory $\hat{L}_\mathcal{T}$. To this end, CTR builds upon the observations in Sect. 4.2, via the three steps detailed next. The notation employed throughout the discussion is summarized in the Abbreviations section.

5.1 Nighttime trajectory reconstruction

The first step in CTR aims at reconstructing missing portions of a trajectory during the night hours. The rationale is that, as seen in Sect. 4.2, these are easily reconstructed from the extractable knowledge about users' home locations, which is also straightforward to derive from mobile phone data. Specifically, we assume that nighttime spans between 10 PM and 9 AM of the subsequent day: this period is adapted to the typical schedule in the Latin America country where our data is collected, but are easily adjusted to any other settings with statistical data on the local population habits. For each user in the dataset, we proceed as follows.

We first identify the most frequent location visited by the user during nighttime, which we name $\hat{\mathbf{l}}_H$. Then, we verify that $\hat{\mathbf{l}}_H$ accounts for a conservative 80% (or more) of the positions of the user during the night period. If this is not the case, we deem that the result is not consistent enough to establish with high confidence the actual home location of the user, and we skip the nighttime trajectory reconstruction step. Otherwise, we identify $\hat{\mathbf{l}}_H$ as the home location of the user. When applied to our reference dataset of 1.8 million subscribers, this strategy allowed assigning a home location to 0.72 million users, equivalent to 42% of the whole population.

If a home location $\hat{\mathbf{l}}_H$ is identified, each 10 PM–9 AM period is processed separately for every day. The user is considered to be at home only in periods where no location other than $\hat{\mathbf{l}}_H$ is observed in the data: in this case, all time intervals i in the period are added to a set Ω^H. Otherwise, if at least one location different from $\hat{\mathbf{l}}_H$ is present during a 10 PM–9 AM period, we consider that the user may not have spent that specific night at home, and no modification is made to Ω^H. Once all days have been processed, we assign the home location to all slots in Ω^H, *i.e.*, $\hat{\mathbf{l}}_i = \hat{\mathbf{l}}_H$, $\forall i \in \Omega^H$. We then update the known trajectory as $\hat{L}_\Omega = \{\mathbf{l}_i \mid i \in \Omega\} \cup \{\hat{\mathbf{l}}_i \mid i \in \Omega^H\}$, and the set of missing locations by $\Omega^C = \mathcal{T} - \Omega - \Omega^H$. In our reference dataset, an average of 82% of the nighttime slots of the aforementioned 0.72 million users could be reconstructed this way.

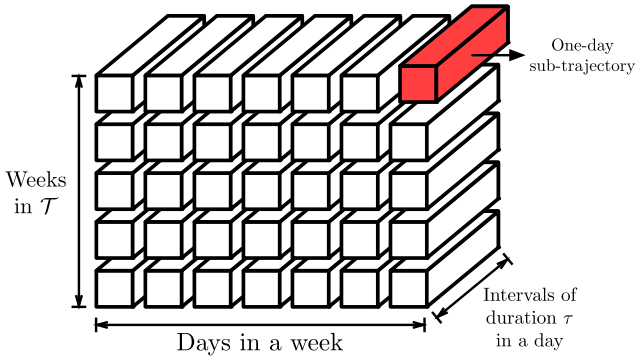

Figure 5 Example of the location tensor of a user $\mathcal{X}$. A trajectory $L_{\mathcal{T}}$ is represented as daily and weekly combinations of one-day sub-trajectories

5.2 Seamless trajectory reconstruction with tensor factorization

The second step of CTR aims at reconstructing the trajectory during the remaining time intervals in Ω^C, by tailoring state-of-the-art *tensor factorization* techniques to our problem. Tensor factorization exploits redundancy to recover missing data; in our context, redundancy is created by the regularity of human movement, which creates repeated patterns of visited locations over the many days and weeks covered by a CDR dataset, as discussed in Sect. 4.2.

We start by organizing the trajectory data of a user into a three-dimensional tensor format, as illustrated in Fig. 5. The *location tensor* $\mathcal{X}$ has dimensions that reflect the weekly, daily, and instantaneous (with granularity τ) mobility of the user. Therefore $\mathcal{X} \in \mathbb{R}^{N_w \times N_d \times 2N_\tau}$, where N_w is the number of weeks in the observation period, N_d is the number of days in a week, and N_τ is the number of time slots in a day.

A trajectory $L_{\mathcal{T}}$ is transformed into $\mathcal{X}$ via the following two steps:

- First, the trajectory is split into multiple one-day sub-trajectories. Each one-day sub-trajectory is then converted into a one-dimensional vector: for instance, the sub-trajectory of the j^{th} day of the ith week is denoted by $\mathbf{x}_{i,j}$ and satisfies $\mathbf{x}_{i,j} = (\mathbf{l}^{(x)}_{n_{i,j,1}}, \mathbf{l}^{(y)}_{n_{i,j,1}}, \ldots, \mathbf{l}^{(x)}_{n_{i,j,N_\tau}}, \mathbf{l}^{(y)}_{n_{i,j,N_\tau}}) \in \mathbb{R}^{2N_\tau}$, where $\mathbf{l}^{(x)}_k$ and $\mathbf{l}^{(y)}_k$ are the two coordinates that identify the location at time step k, for $\mathbf{l}_k = (\mathbf{l}^{(x)}_k, \mathbf{l}^{(y)}_k)$.
- Second, we enter all of the one-day sub-trajectories of a user trajectory into the location tensor by organizing them into a matrix of one-day sub-trajectories for all N_w weeks in the observation period $\mathcal{T}$, *i.e.*, $\mathcal{X} = [\mathbf{x}_{i,j}]_{N_w \times N_d} = \{\mathcal{X}_{i,j,k}\}_{N_w \times N_d \times 2N_\tau}$.

The location tensor $\mathcal{X}$ is partially filled with the original incomplete trajectory from mobile phone data, as well as with the inferred overnight locations as per Sect. 5.1. At this stage, the location tensor includes information for time slots in $\Omega \cup \Omega^H$. However, the tensor is still relatively sparse, as the set Ω^C of time intervals where locations are missing is still large. In our reference dataset, the location tensors have an average density of 0.63 (where 1 denotes a complete tensor) when first generated.

In order to complete the location tensor of an individual, we employ tensor factorization, which decomposes the tensor into hyper-parameters and then uses them to infer missing values. Factorization has proven very effective in recovering highly incomplete structural data in other contexts [25, 26]. Formally, we formulate an equivalent tensor factorization (TF) problem as follows. First, we perform a canonical polyadic decompo-

sition (CPD) [27] of the tensor $\mathcal{X} \in \mathbb{R}^{N_w \times N_d \times 2N_\tau}$ into three R-rank matrices $\mathbf{A} \in \mathbb{R}^{N_w \times R}$, $\mathbf{B} \in \mathbb{R}^{N_d \times R}$, and $\mathbf{C} \in \mathbb{R}^{2N_\tau \times R}$, where each value $\mathcal{X}_{i,j,k}$ in the tensor is approximated as $\mathcal{X}_{i,j,k} = \sum_{\delta=1}^{R} \mathbf{A}_{i,\delta}\mathbf{B}_{j,\delta}\mathbf{C}_{k,\delta}$. In our experiments, we set the rank R to the minimum cardinality of the tensor dimensions, *i.e.*, $R = \min(N_w, N_d, 2N_\tau)$. This value entails a small number of hyperparameters in tensor decomposition, and is well within the upper bound of the tensor rank $\min(N_w N_d, 2N_w N_\tau, 2N_d N_\tau)$, which cannot be otherwise computed with a finite algorithm [27]. For simplicity, we employ the concise notation $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ for the CPD using matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. The TF problem is then

$$\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}} = \underset{\mathbf{A},\mathbf{B},\mathbf{C}}{\arg\min} \sum_{(i,j,k)\in\Omega\cup\Omega^H} \left(\mathcal{X}_{i,j,k} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_{i,j,k}\right)^2 + \lambda\left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2\right), \tag{2}$$

where λ is a penalty parameter to avoid overfitting [27] and $\|\cdot\|_F$ is the Frobenius norm. The solution to the problem in (2) is an estimation of the complete tensor $\hat{\mathcal{X}} = [\![\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}]\!]$, which thus recovers the missing locations in Ω^C. In other words, we are seeking a tensor that has R-rank CPD decomposition and that fits the observed locations as closely as possible.

In fact, (2) is a standard TF problem, which does not account for contextual specificities that can improve the solution. Specifically, the formulation in (2) treats each dimension and every single value of the tensor equally, while human mobility patterns exhibit non-uniform importance of dimensions and values. Indeed, (i) daily periodicities tend to be stronger than weekly ones [3], and (ii) consecutive time slots, hence nearby values in a same one-day sub-trajectory vector, show a strong correlation in terms of locations of the user [18]. In the light of these considerations, we customize the TF problem for location inference, by introducing two additional elements to the optimization problem in (2).

The first element emphasizes daily repetitive mobility patterns by means of Toeplitz matrices. Toeplitz matrices allow modelling relationships between specific elements in the location tensor. For instance, given a matrix $\mathbf{P} = [p_{i,j}]_{N\times N}$ and a Toeplitz matrix $\mathbf{Q} = \text{Toeplitz}(1, -1, 0)_{N\times N}$, then the product $\|\mathbf{PQ}\|_F^2$ represents the sum of the differences between consecutive values in the original matrix $\mathbf{P}$, *i.e.*, expands to $\sum_{i=1}^{N}\sum_{j=0}^{N}\|(p_{i-1,j} - p_{i,j})\|_F^2$. Therefore, we construct a matrix $\mathbf{D} = \text{Toeplitz}(1, -1, 0)_{N_d\times N_d}$ as illustrated in Fig. 6; then, the expression $\|\mathcal{X} \times_d \mathbf{D}\|_F^2$, where $\times_d$ is the *tensor-matrix* product [27] such that $(\mathcal{X} \times_d \mathbf{D})_{i,m,k} = \sum_{n=1}^{N_d} \mathcal{X}_{i,n,k}\mathbf{D}_{m,n}$, represents the sum of squared differences of location coordinates at the same hour of consecutive days. In our case, $\mathcal{X}$ is decomposed via CPD

$$\text{Toeplitz}(\boxed{1, -1, 0})_{N_d\times N_d} = \begin{bmatrix} \boxed{1} & \boxed{-1} & \boxed{0} & 0 & \cdots & 0 \\ 0 & \boxed{1} & \boxed{-1} & \boxed{0} & \ddots & \vdots \\ 0 & 0 & 1 & -1 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{N_d\times N_d}$$

Figure 6 Example of matrix $\textbf{Toeplitz}(1, -1, 0) \in \mathbb{R}^{N_d\times N_d}$. The diagonal is filled by 1's, the first upper diagonal is filled by −1's, and all other values are 0's

into $[\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]$, hence we have

$$\begin{aligned}\left([\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]\times_d \mathbf{D}\right)_{i,m,k} &= \sum_{n=1}^{N_d}\sum_{\delta=1}^{R}\mathbf{A}_{i,\delta}\mathbf{B}_{n,\delta}\mathbf{C}_{k,\delta}\mathbf{D}_{m,n}\\ &= \sum_{\delta=1}^{R}\mathbf{A}_{i,\delta}\left(\sum_{n=1}^{N_d}\mathbf{D}_{m,n}\mathbf{B}_{n,\delta}\right)\mathbf{C}_{k,\delta}\\ &= \sum_{\delta=1}^{R}\mathbf{A}_{i,\delta}(\mathbf{DB})_{m,\delta}\mathbf{C}_{k,\delta} \end{aligned} \tag{3}$$

and equivalently, $[\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]\times_d \mathbf{D} = [\![\mathbf{A},\mathbf{DB},\mathbf{C}]\!]$. Minimizing $\|[\![\mathbf{A},\mathbf{DB},\mathbf{C}]\!]\|_F^2$ implies reducing the diversity of locations at a same hour across individual days.

A similar approach is then adopted to ensure that the solution to the TF problem favors the similarity of locations in consecutive time intervals. Here, we construct a Toeplitz matrix $\mathbf{T} = \text{Toeplitz}(1,0,-1,0)_{2N_\tau\times 2N_\tau}$, and denote by $\times_\tau$ the tensor-matrix product such that $(\mathcal{X}\times_\tau \mathbf{D})_{i,m,k} = \sum_{n=1}^{2N_\tau}\mathcal{X}_{i,m,n}\mathbf{T}_{k,n}$. Recall that each daily sub-trajectory is a one-dimensional vector $(\mathbf{l}_1^{(x)},\mathbf{l}_1^{(y)},\mathbf{l}_2^{(x)},\mathbf{l}_2^{(y)},\dots)\in\mathbb{R}^{2N_\tau}$, hence the x-axis (respectively, y-axis) of two consecutive locations are separated by a coordinate of the other type. Then, including in the minimization problem the term $\|\mathcal{X}\times_\tau \mathbf{T}\|_F^2 = \|[\![\mathbf{A},\mathbf{B},\mathbf{TC}]\!]\|_F^2$ allows constraining the diversity of locations between consecutive hours in a same day.

Embedding both elements above in (2) leads to an enhanced TF problem

$$\begin{aligned}&\underset{\mathbf{A},\mathbf{B},\mathbf{C}}{\arg\min}\sum_{(i,j,k)\in\Omega\cup\Omega^H}\left(\mathcal{X}_{i,j,k}-[\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]_{i,j,k}\right)^2+\lambda\left(\|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2+\|\mathbf{C}\|_F^2\right)\\ &\quad+\lambda_d\left\|[\![\mathbf{A},\mathbf{DB},\mathbf{C}]\!]\right\|_F^2+\lambda_\tau\left\|[\![\mathbf{A},\mathbf{B},\mathbf{TC}]\!]\right\|_F^2\end{aligned} \tag{4}$$

where λ_d and λ_τ weight the importance of similarity across time slots that are 24-hour apart and adjacent, respectively. The problem in (4) is a combination of multiple least square problems, and is efficiently solved by means of an alternating least square (ALS) technique [27].

The complete tensor estimation $\hat{\mathcal{X}}$ contains the original locations of the user extracted from the mobile phone data during time instants in Ω, the locations inferred during night hours in Ω^H as per Sect. 5.1, and those estimated by solving (4) for all time slots in Ω^C. Therefore, $\hat{\mathcal{X}}$ can be used to fill an updated trajectory $\hat{L}_\mathcal{T} = \{\mathbf{l}_i \mid i\in\Omega\}\cup\{\hat{\mathbf{l}}_i \mid i\in\Omega^H\cup\Omega^C\}$, which is *seamless*, *i.e.*, covers the whole observation period $\mathcal{T}$.

5.3 Homogeneous quantization of locations

The locations in the original mobile phone data, as well as those inferred for night hours, match the position of cellular network towers. Instead, the solution of the enhanced TF problem in (4) returns locations whose coordinates are real values, which may not match the position of actual cellular towers. Since it is desirable that the reconstructed trajectory has consistent properties (including a homogeneous quantization of the locations that is based on the cellular infrastructure deployment), the third step in `CTR` aims at associating each location estimated as per Sect. 5.2 to the position of a real-world cellular tower.

Let $\mathcal{C} = \{\mathbf{c}_j \mid 1\le j\le N_c\}$ be the set of locations of the N_c cell towers deployed in the target geographic region. Our goal is that of replacing the estimated locations $\hat{\mathbf{l}}_i$ such that

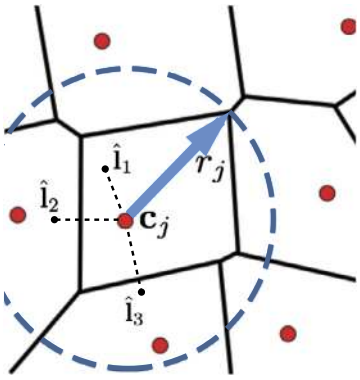

Figure 7 Example of cellular radius r_j for cellular tower $\mathbf{c}_j$. Solid black lines outline the Voronoi tessellation of space, while the solid blue arrow denotes the cell radius r_j. The dashed blue circle outlines the region whose radius is the cell radius r_j. Because the region covers the estimated locations $\hat{\mathbf{l}}_1, \ldots, \hat{\mathbf{l}}_3$, the cell tower $\mathbf{c}_j$ is a possible candidate for associating these estimated locations to actual cell towers

$i \in \Omega^C$ by one of the cell tower positions $\mathbf{c}_j \in \mathcal{C}$. To this end, we first design a sensible cost function for determining the suitability of a match $(\hat{\mathbf{l}}_i, \mathbf{c}_j)$ between a generic location and a cell tower position. The cost function, denoted by $f(\hat{\mathbf{l}}_i, \mathbf{c}_j)$, accounts for (i) the geographic distance between $\hat{\mathbf{l}}_i$ and $\mathbf{c}_j$, and (ii) the popularity of the tower $\mathbf{c}_j$, as:

$$f(\hat{\mathbf{l}}_i, \mathbf{c}_j) = \begin{cases} 1/e_j & \text{if } \|\mathbf{c}_j, \hat{\mathbf{l}}_i\| \leq r_j, \\ +\infty & \text{otherwise.} \end{cases} \tag{5}$$

In expression (6) above: e_j is the number of communication events that are generated by all the users in the dataset and associated to the cellular tower $\mathbf{c}_j$ in the mobile phone data; $\|\cdot\|$ denotes the geographic distance between the two location parameters; and, r_j is the *cellular radius* of tower $\mathbf{c}_j$, *i.e.*, the maximum distance between the actual tower position $\mathbf{c}_j$ and the farthest corner of its Voronoi tessellation cell [28] as illustrated in Fig. 7.

The cell tower replacing an estimated location $\hat{\mathbf{l}}_i$ is that satisfying

$$\arg\min_{\mathbf{c}_j \in \mathcal{C}} f(\hat{\mathbf{l}}_i, \mathbf{c}_j). \tag{6}$$

In other words, minimizing the cost function limits the selection of matching cell towers for an estimated location $\hat{\mathbf{l}}_i$ to those within distance r_j of $\hat{\mathbf{l}}_i$; then, it picks the candidate having the highest probability of occurrence in the data, hence adopting a wisdom-of-the-crowd approach that has already proven effective in personal mobility analysis [29].

6 Validation

In order to validate the proposed CTR solution, we employ a set of ground-truth trajectories with completeness equal to 1, presented in Sect. 6.1. We generate incomplete trajectories from the ground-truth data, following the same sampling process observed in our reference CDR dataset, as described in Sect. 6.2. Then, complete trajectories are reconstructed with CTR from the downsampled ones, and compared against the ground-truth, in Sect. 6.3: the level of agreement lets us comment on the quality of the reconstruction.

6.1 Fine-grained ground-truth trajectories

We have access to data about the mobile Internet sessions of a small subset of the 1.8 million users in the CDR dataset introduced in Sect. 3. As more events are generated

from Internet access rather than voice calls, they entail a sensibly higher frequency in location sampling. When considered jointly with CDR, rich mobile Internet data allows obtaining the complete trajectories of 1450 users over an observation period $\mathcal{T}$ of two consecutive weeks, and with a temporal resolution of $\tau = 1$ hour. If multiple samples are recorded in the same time slot, the unique location is determined as that of the cell tower which most events are associated with. In other words, we have 1450 complete trajectories having observation sets $\Omega = \mathcal{T}$. In each of them, all the locations are known, *i.e.*, $L_\Omega = \{\mathbf{l}_i \mid i \in \Omega\} = L_\mathcal{T}$. We remark that the available fine-grained ground-truth trajectories bound our validation to a two-week period. We leave to future analyses, based on mobility data covering longer periods, the assessment of the impact of the trajectory duration on the accuracy of the completion process.

6.2 Incomplete trajectories

Incomplete trajectories are generated by downsampling the ground-truth ones. For instance, a trajectory with completeness 0.3 is derived from a two-week ground-truth trajectory having $14 \times 24 = 336$ time slots by keeping the locations in 101 time slots and discarding the rest. Here, the sampling process must be chosen appropriately: for instance, a uniform random selection would result in a temporal distribution of retained locations that is inconsistent with that produced by mobile communication events in CDR. We ensure that the incomplete trajectories mimic the temporal sampling properties of those extracted from CDR as follows: (i) we construct an empirical distribution of the probability that events are recorded at each time slots, based on CDR data of all 1.8 million users in our dataset; (ii) we set the desired completeness value k that shall characterize the output incomplete trajectory; (iii) we retain a fraction k of the overall location samples in one complete trajectory according to a random selection of time slots that follows the empirical distribution above. Performing this approach once leads to an artificial missing set Ω^C, and results in a down-sampled incomplete trajectory $L_\Omega = \{\mathbf{l}_i \mid i \in \Omega\}$ (recall that $\Omega = \mathcal{T} - \Omega^C$).

This approach lets us produce a large amount of downsampled trajectories out of the 1450 ground-truth ones, and gives us enough flexibility to carry out a comprehensive validation. Specifically, we generate 3×10^5 incomplete trajectories with completeness ranging from 0.01 to 0.5, which allows testing CTR trajectory reconstructions in a wide diversity of settings.

6.3 Quality of CTR trajectory reconstruction

We run CTR on the incomplete CDR-like data and update the incomplete CDR-like trajectories as $\hat{L}_\mathcal{T} = \{\mathbf{l}_i \mid i \in \Omega\} \cup \{\hat{\mathbf{l}}_i \mid i \in \Omega^C\}$. We then assess the quality of the reconstructed trajectories $\hat{L}_\mathcal{T}$ against the complete ground-truth known locations $L_\mathcal{T}$.

The performance metric we adopt is *cell displacement*, which expresses the MAE used in (1) in terms of mobile network cells, as

$$\Delta(\hat{L}_\mathcal{T}, L_\mathcal{T}) = \frac{1}{|L_\mathcal{T}| - |L_\Omega|} \sum_{i \in \mathcal{T} \setminus \Omega} \frac{\|\mathbf{l}_i, \hat{\mathbf{l}}_i\|}{r_j}. \tag{7}$$

We recall that $\|\cdot\|$ denotes the geographic distance between the two location parameters, and r_j is the cellular radius of the tower j to which the user is associated at time step i, *i.e.*,

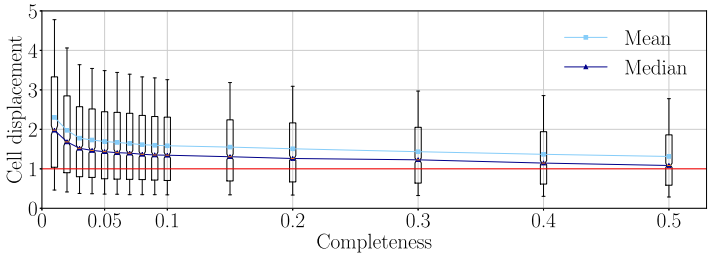

Figure 8 Performance of CTR in cell displacement. Cell displacement of locations inferred via CTR with respect to the actual ones, versus the completeness of the original trajectory. Candlesticks highlight the mean (light blue), median (dark blue), 25th and 75th percentiles (box), and 10th and 90th percentiles (errorbars). The horizontal line (red) highlights a one-cell displacement

for which $\mathbf{l}_i = \mathbf{c}_j$. Expression (7) implies that the cell displacement is computed only based on locations that are missing in the original data (*i.e.*, for time instants in $\mathcal{T} \setminus \Omega$), and known user positions do not affect it. Unlike MAE, cell displacement compensates for the very diverse density (hence radius) that cellular towers tend to have in different regions of a large geographic area and provides a relative measure that is comparable across all locations.

The validation results are summarized in Fig. 8, where the cell displacement measured in the trajectories reconstructed by CTR is shown against the completeness of the input CDR-like data. The candlesticks report the mean and median cell displacement, as well as the 10th, 25th, 75th, and 90th percentiles. The horizontal red line highlights a cell displacement of 1: below this value, the estimation error is lower than spatial precision of the original mobile phone data, *i.e.*, the geographical coverage radius of the cell tower the user is presently associated to. Therefore, the error of the completion process is smaller than that inherent to the original data.

The most important remark is that the median cell displacement is always between 1 and 2, which denotes a substantial accuracy in the reconstructed trajectories: even under completeness levels as low as 1%, estimated locations are typically placed in a cell that is immediately adjacent to the correct one, or one cell further. More complete original trajectories provide CTR with added information to fill gaps in the data, leading to an improved performance where the typical estimation error is pushed closer to the correct cell. Interestingly, the gain of accuracy is higher at low completeness levels, and a trajectory with just 5% completeness already reduces the median cell displacement to less than 1.5. Some variability is observed around the median, which is however natural, given the heterogeneity that characterizes the mobility patterns and communication activities of different users.

The level of reconstruction accuracy achieved by CTR is acceptable for metropolitan-scale analyses: in urban areas, network cells usually span a few hundreds of square meters and cover, *e.g.*, (portions of) individual neighborhoods, hence the displacement in Fig. 8 would still allow locating users fairly precisely and investigating mobility flows at an inter-neighborhood level. The reconstructed trajectory precision is instead excellent for regional- or national-scale studies, since cell displacements of 1 or 2—over large surfaces covered by tens of thousands of cell towers—allow to capture human mobility at, *e.g.*, inter-city level, perfectly.

7 Revisiting key results in the literature

As mentioned in Sect. 1, mobile phone data are the cornerstone of many seminal works on large-scale human mobility analysis. All such works (i) employ raw CDR datasets that are highly incomplete, and (ii) possibly adopt substantial user filtering. However, if and how trajectory incompleteness and user filtering effect—and possibly bias—the outcome of these studies is unclear. Our aim in this section is shedding light on this matter, which also lets us demonstrate the importance of trajectory reconstruction in mobile phone data analysis.

To this end, we run CTR on the reference dataset presented in Sect. 3 and reconstruct all trajectories which have completeness higher than 1% in the original data. Selecting such a lower threshold on completeness bounds the typical cell displacement to a small value according to the validation in Sect. 6, hence ensure high confidence in the quality of the estimated locations; moreover, a 1% completeness threshold allows retaining 95% of the user population. Overall, we reconstruct the complete trajectories of around 1.7 million individuals in a large geographic region, with a temporal resolution of $\tau = 1$ hour during three months.

We then reproduce three key studies on human mobility, using both the original CDR dataset and the complete trajectory data returned by CTR, and discuss the eventual differences that we observe in the results. The three analyses are separately presented next.

7.1 Laws of individual mobility

In an important study appeared in Nature in 2008, Gonzalez *et al.* [3] derive a seminal model of individual human mobility based on CDR data. The data-driven analysis reveals that the travel distances of single users between two consecutive locations follow a truncated power-law. Formally, for any given individual, her travel distance Δr between two consecutive events follows a truncated power-law $P(\Delta r|r_g) \sim r_g^{-\alpha} F(\Delta r/r_g)$, where r_g is the radius of gyration,[a] and $F(x)$ is a function such that $F(x) \sim x^{-\alpha}$ for $x < 1$ and $F(x)$ rapidly decreases for $x > 1$. This expression indicates that human movements resemble a *Lévy flight* that follows a power-law $P(\Delta r) \sim \Delta r^{-\alpha}$ within a geographical region bounded by r_g; long displacements beyond such region are instead increasingly rare.

These key insights are obtained from a random sample of 1.67% of the total 6 million users, and the trajectories of the selected users are largely incomplete, with less than one location per user and per day recorded on average. We run the same analysis on the seamless trajectories of 1.7 million users, and show in Fig. 9 how considering complete movement data of a full user population affects the results. Clearly, we look now at a complete set of events—one per time slot—and the travel distance Δr is computed between the locations in any two subsequent time slots. This ultimately results in a more comprehensive view of individual travel distances. We make the following considerations.

- As a preliminary result in their analysis, Gonzalez *et al.* find that the travel distances and radiuses of gyration aggregated over the whole user population follow truncated power laws $(x + x_0)^{-\beta} e^{-x/k}$. We confirm that this is also the case under complete trajectory data, as shown in the top plots of Fig. 9. The exponent values are also consistent, as Gonzalez *et al.* find $\beta_{\Delta_r} = 1.77$ and $\beta_{r_g} = 1.65$, whereas $\beta_{\Delta_r} = 1.78$ and $\beta_{r_g} = 1.68$ from our complete trajectory data. However, we remark a sensible difference in the cutoff values. Travel distances can be sensibly higher in complete data; also, k_{r_g} is at 400 km in [3], which is not far from the 340 km from our original

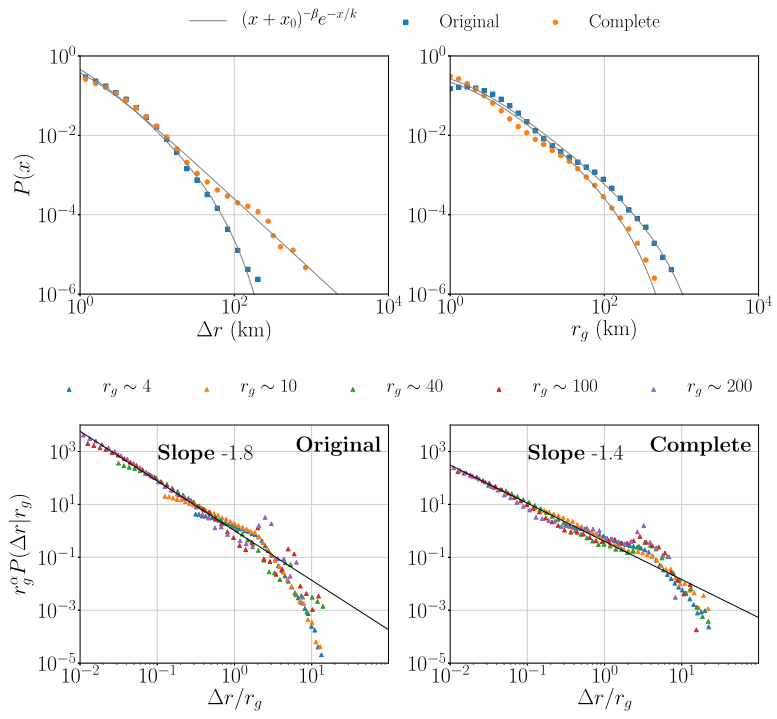

Figure 9 Revisited results from [3]. Distributions of travel distance Δr, radius of gyration r_g, and conditional travel distance Δr (with respect to $\Delta r/r_g$) for original and complete trajectories

(*i.e.*, incomplete) CDR-based trajectories but is reduced in the complete data, where the cutoff occurs instead at 175 km. We ascribe the difference to the fact that completion often leads to add missing locations that are far from the linear interpolation between CDR positions, such as those generated by infrequent but recurrent long-haul trips. Our conclusion is that, CDR data sparsity risks to both underestimate long trips, and overestimate the region within which the Lévy flight behavior of human mobility occurs.

- Concerning individual movements, we can reproduce the truncated power-law behavior found by Gonzalez *et al.* with both our sparse CDR-based trajectories and the complete ones reconstructed via CTR. This is illustrated in the bottom plots of Fig. 9. In the original work, the authors find an exponent $\alpha \approx 1.2$, and we estimate the α parameter at 1.8 and 1.4 for original and complete trajectories. These exponents are qualitatively consistent since they all are in the $(1, 3)$ range that characterizes Lévy walks [30].
- We ascribe the quantitative differences with respect to [3] to the inherent specificity of each dataset. Although we also use locations from voice calls as in [3], the data refer to very different cultural and economic settings (*i.e.*, a European country and a South American one) and geographical span (our country covers a territory that is four times wider than the largest country in Europe).

Overall, *our results corroborate the findings in* [3], *and show that the resulting conclusions hold also with seamless trajectory data from larger populations.*

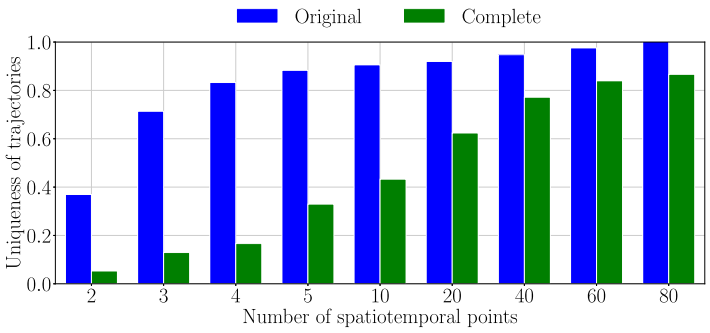

Figure 10 Revisited results from [5]. Ratio of trajectories with uniqueness $|U| = 1$ for a number of spatiotemporal points in the abscissa, for original (left) and complete (right) trajectories

7.2 Uniqueness of individual trajectories

The work by De Montjoye *et al.* [5], published in Scientific Reports in 2013, is the first to raise significant concerns on the privacy of users whose communication activities are recorded in CDR datasets. The study unveils the very high *uniqueness* of individual trajectories: knowing a few random time-stamped locations of an individual allows pinpointing her trajectory within the mobile phone data of a large user population. Notably, the authors show that 4 random $\mathbf{l}_i$'s identify one specific user among 1.5 million individuals 95% of times. In practice, the result implies that an adversary would need minimal knowledge about the mobility of a victim in order to perform a re-identification attack on a CDR dataset. The tests leading to the conclusion above consider a large user population but rely on raw CDR data that is highly incomplete: on average, a user only has approximately 19 unique locations observed in a month from 4200 cell towers.

Figure 10 depicts how complete trajectories affect the results. The plots show how many trajectories in the dataset are uniquely identified by knowing a random number of time-stamped locations (in the abscissa), with incomplete CDR data, and with seamless reconstructed trajectories. The result with incomplete data is very well aligned with that of De Montjoye *et al.* However, the difference with complete trajectories is striking: for the low number of locations considered in [5], the uniqueness is reduced by around one order of magnitude when the location information is seamless. The contrast lets us argue that *the extremely high uniqueness identified by De Montjoye et al. is mainly caused by the diverse temporal patterns of the mobile communications of each user, rather than by his/her distinctive mobility.* In other words, many users tend to have similar complete trajectories, and sampling them randomly yields limited uniqueness; it is instead the sparse sampling of the trajectories inherent to CDR that introduces an artificial temporal diversity and leads to dramatically increased uniqueness.

We thus conclude that actual human mobility is sensibly less unique than what suggested by incomplete trajectories inferred from mobile phone data. Note, however, that this does not mean that user privacy is preserved in reconstructed trajectory data, nor it negates the privacy warning by De Montjoye *et al.* As shown in Fig. 10, re-identification is still very possible and yet becomes more difficult. For instance, pinpointing a user in 90% of cases requires knowledge of 80 random spatiotemporal samples, instead of the 5 suggested by the original study.

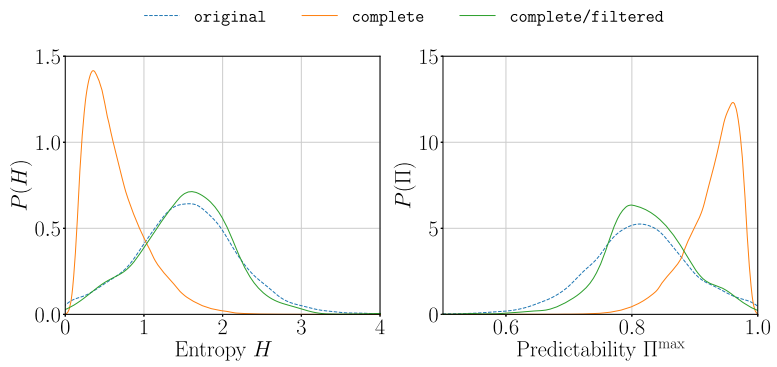

Figure 11 Revisited results from [4]. Distributions of the entropy and maximum theoretical predictability derived with (i) the incomplete original CDR-based trajectories of 8000 users filtered with the method in [4], (ii) the complete trajectories of all 1.7 million users, and (iii) the complete trajectories of the same 8000 users considered in the first case

7.3 Next-location predictability

The high theoretical predictability of human mobility was first exposed by Song *et al.* [4], in a famous paper that appeared in Science in 2010. In that work, the authors model the movement of each user as a time series of locations extracted from mobile phone data. They then measure the entropy rate of the probability of finding a particular time-ordered subsequence in the trajectory of the user. Finally, they compute the maximum predictability of each outcome from the entropy rate, by using Fano's inequality. The analysis considers 45,000 users (out of a total population of 10 million) with trajectory completeness higher than 0.2, at least 0.5 calls/hour and more than 2 unique recorded locations during an observation period of 14 weeks. Based on these data, the theoretical predictability is found to have a very high upper bound at around 93%.

We repeat the experiment with our reference dataset and trajectory reconstruction method. We run the filtering adopted by Song *et al.* on our original CDR data and extract 8000 users: this maps to around 0.44% of the available population, which is very close to the 0.45% considered in [4]. The probability distribution of the maximum predictability across the retained users is labeled as `original` in Fig. 11. The mean predictability is at 81%, which is high yet quite far from the 93% found by Song *et al.*: As users are selected in the exact same way in the two cases, we ascribe the difference to the diverse mobility habits of the user populations in the two datasets, which are collected in different countries, at different scales, and during different time periods.

More interesting to our study is thus the comparison of the predictability obtained with the incomplete trajectories of 0.44% of the users, and that resulting from the seamless trajectories of the full available population (i.e., 1.7 million users), reconstructed through `CTR`. The curve labeled as `complete` in Fig. 11 is considerably shifted to the right, with a much-reduced variance around the peak, now at 94%. Our hypothesis for this outcome is that the filtering introduced in [4] dramatically reduces the set of users, favoring individuals who are very active from a mobile communication viewpoint. It has been repeatedly demonstrated that interactions with the cellular network are more frequent for users with higher mobility [31–33]. As a result, *the sparse nature of CDR data lets Song et al. introduce an unwanted bias in their study, which is ultimately focused on very mobile individuals whose displacements are more difficult to anticipate.* Instead, considering a much

larger population lets us account for the vast majority of fairly static users, and reveals that people's movements are on average even easier to predict than estimated in [4].

In order to verify our hypothesis, we show in Fig. 11 the predictability distribution for the complete, reconstructed trajectories of the same 8000 users that are retained from the original data by the filtering proposed in [4]. The result, tagged as `complete/filtered` is very consistent with that derived with the sparse CDR-based trajectories of those subscribers, *i.e.*, the `original` curve. The result supports our intuition that it is the population sampling and not the reconstruction process that determines the striking difference in the figure.

8 Conclusions

We have investigated the problem of sparsity in trajectories inferred from mobile phone data. We have quantified the issue in a representative real-world dataset, proposed a novel solution, named `CTR`, to reconstruct a seamless trajectory from sparse CDR data, and validated our methodology using ground-truth movement patterns. We have also demonstrated the importance of trajectory reconstruction by revisiting well-known results on human mobility based on raw CDR, and showing that complete trajectories can in some cases affect the outcome of those analyses substantially.

Funding
This work was supported by the ERANET CHIST-ERA MACACO and is performed in the context of the EMBRACE Associated Team of Inria and STIC AmSud MOTIF.

Abbreviations
$\mathcal{T}$,observation period of a trajectory; $L_{\mathcal{T}} = \{\mathbf{l}_i \mid i \in \mathcal{T}\}$,ground-truth trajectory; $\Omega = \{i \in \mathcal{T} \mid \mathbf{l}_i \neq \emptyset\}$,observed period of a trajectory; $L_{\Omega} = \{\mathbf{l}_i \mid i \in \Omega\}$,observed trajectory; $\Omega^C = \mathcal{T} - \Omega$,unknown period of a trajectory; $\hat{L}_{\mathcal{T}} = \{\hat{\mathbf{l}}_i \neq \emptyset \mid i \in \Omega \cup \Omega^C\}$,estimated trajectory; $\emptyset$,empty set; $\mathbf{l}_i$,location of the ith time slot.

Availability of data and materials
The datasets we use in this study contain sensitive personal information and are protected by non-disclosure agreements with the data owners, which prevents public disclosure. We understand and appreciate the need for transparency in research, however we cannot share access to confidential datasets used.

Ethics approval and consent to participate
Data collection was approved by the data owners, and written informed consent has been obtained for all study participants.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Designed the study: GC, AV, and MF. Collected and analyzed the data: GC and CS. All authors wrote and approved the final manuscript.

Author details
[1]INRIA, Université Paris-Saclay, Palaiseau, France. [2]École Polytechnique, Université Paris-Saclay, Palaiseau, France. [3]CNR-IEIIT, Torino, Italy. [4]Grandata Labs, San Francisco, USA.

Endnote
[a] The radius of gyration is a scalar metric assessing the overall mobility of a user. For a trajectory $L_{\mathcal{T}}$, it is computed as $r_g = \sqrt{\frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \|\mathbf{l}_i, \mathbf{l}_C\|^2}$, where $\mathbf{l}_C = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \mathbf{l}_i$ is the center of mass of all the locations in $L_{\mathcal{T}}$.

Publisher's Note

Received: 21 January 2019 Accepted: 3 September 2019 Published online: 12 October 2019

References

1. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4(1):10. https://doi.org/10.1140/epjds/s13688-015-0046-0
2. Naboulsi D, Fiore M, Ribot S, Stanica R (2016) Large-scale mobile traffic analysis: a survey. IEEE Commun Surv Tutor 18(1):124–161. https://doi.org/10.1109/comst.2015.2491361
3. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958
4. Song C, Qu Z, Blumm N, Barabasi A-L (2010) Limits of predictability in human mobility. Science 327(5968):1018–1021. https://doi.org/10.1126/science.1177170
5. de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. Sci Rep 3(1):1376. https://doi.org/10.1038/srep01376
6. Ahas R, Silm S, Saluveer E, Järv O (2009) Modelling home and work locations of populations using passive mobile positioning data. In: Gartner G, Rehrl K (eds) Location based services and TeleCartography II: from sensor fusion to context models. Springer, Berlin, pp 301–315. https://doi.org/10.1007/978-3-540-87393-8_18
7. Schneider CM, Belik V, Couronne T, Smoreda Z, Gonzalez MC (2013) Unravelling daily human mobility motifs. J R Soc Interface 10(84):20130246. https://doi.org/10.1098/rsif.2013.0246
8. Ferreira N, Poco J, Vo HT, Freire J, Silva CT (2013) Visual exploration of big spatio-temporal urban data: a study of New York city taxi trips. IEEE Trans Vis Comput Graph 19(12):2149–2158. https://doi.org/10.1109/TVCG.2013.226
9. Zhang D, Zhao J, Zhang F, He T (2015) coMobile: real-time human mobility modeling at urban scale using multi-view learning. In: Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems. SIGSPATIAL '15. ACM, New York, pp 40:1–40:10. https://doi.org/10.1145/2820783.2820821
10. Zang H, Bolot JC (2007) Mining call and mobility data to improve paging efficiency in cellular networks. In: Proceedings of the 13th annual ACM international conference on mobile computing and networking. MobiCom '07. ACM, New York, pp 123–134. https://doi.org/10.1145/1287853.1287868
11. Oliveira EMR, Viana AC (2014) From routine to network deployment for data offloading in metropolitan areas. In: 2014 eleventh annual IEEE international conference on sensing, communication, and networking (SECON), pp 126–134. https://doi.org/10.1109/SAHCN.2014.6990335
12. Frias-Martinez E, Williamson G, Frias-Martinez V (2011) An agent-based model of epidemic spread using human mobility and social network information. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, pp 57–64. https://doi.org/10.1109/PASSAT/SocialCom.2011.142
13. Chen G, Hoteit S, Viana AC, Fiore M, Sarraute C (2018) Enriching sparse mobility information in call detail records. Comput Commun 122:44–58. https://doi.org/10.1016/j.comcom.2018.03.012
14. Ranjan G, Zang H, Zhang Z-L, Bolot J (2012) Are call detail records biased for sampling human mobility? Mob Comput Commun Rev 16(3):33. https://doi.org/10.1145/2412096.2412101
15. Sarraute C, Blanc P, Burroni J (2014) A study of age and gender seen through mobile phone usage patterns in Mexico. In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014), pp 836–843. https://doi.org/10.1109/ASONAM.2014.6921683
16. Jo H-H, Karsai M, Karikoski J, Kaski K (2012) Spatiotemporal correlations of handset-based service usages. EPJ Data Sci 1(1):1. https://doi.org/10.1140/epjds10
17. Hoteit S, Chen G, Viana A, Fiore M (2016) Filling the gaps: on the completion of sparse call detail records for mobility analysis. In: Proceedings of the eleventh ACM workshop on challenged networks. CHANTS '16. ACM, New York, pp 45–50. https://doi.org/10.1145/2979683.2979685
18. Ficek M, Kencl L (2012) Inter-call mobility model: a spatio-temporal refinement of call data records using a Gaussian mixture model. In: 2012 proceedings IEEE INFOCOM, pp 469–477. https://doi.org/10.1109/INFCOM.2012.6195786
19. Hoteit S, Secci S, Sobolevsky S, Ratti C, Pujolle G (2014) Estimating human trajectories and hotspots through mobile phone data. Comput Netw 64:296–307. https://doi.org/10.1016/j.comnet.2014.02.011
20. Seshadri M, Machiraju S, Sridharan A, Bolot J, Faloutsos C, Leskove J (2008) Mobile call graphs: beyond power-law and lognormal distributions. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '08. ACM, New York, pp 596–604. https://doi.org/10.1145/1401890.1401963
21. Iovan C, Olteanu-Raimond A-M, Couronné T, Smoreda Z (2013) Moving and calling: mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In: Vandenbroucke D, Bucher B, Crompvoets J (eds) Geographic information science at the heart of Europe. Springer, Cham, pp 247–265. https://doi.org/10.1007/978-3-319-00615-4-14
22. Katsikouli P, Fiore M, Furno A, Stanica R (2019) Characterizing and removing oscillations in mobile phone location data. In: IEEE WoWMoM 2019—20th IEEE international symposium on a world of wireless, mobile and multimedia networks, Washington DC, United States. https://hal.inria.fr/hal-02110719
23. Douglass RW, Meyer DA, Ram M, Rideout D, Song D (2015) High resolution population estimates from telecommunications data. EPJ Data Sci 4(1):4. https://doi.org/10.1140/epjds/s13688-015-0040-6
24. Oliveira EMR, Viana AC, Sarraute C, Brea J, Alvarez-Hamelin I (2016) On the regularity of human mobility. Pervasive Mob Comput 33:73–90. https://doi.org/10.1016/j.pmcj.2016.04.005
25. Kong L, Xia M, Liu X-Y, Chen G, Gu Y, Wu M-Y, Liu X (2014) Data loss and reconstruction in wireless sensor networks. IEEE Trans Parallel Distrib Syst 25(11):2818–2828. https://doi.org/10.1109/tpds.2013.269
26. Karatzoglou A, Amatriain X, Baltrunas L, Oliver N (2010) Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In: Proceedings of the fourth ACM conference on recommender systems. RecSys '10. ACM, New York, pp 79–86. https://doi.org/10.1145/1864708.1864727
27. Kolda TG, Bader BW (2009) Tensor decompositions and applications. SIAM Rev 51(3):455–500. https://doi.org/10.1137/07070111x
28. Portela JN, Alencar MS (2006) Cellular network as a multiplicatively weighted Voronoi diagram. In: CCNC 2006. 2006 3rd IEEE consumer communications and networking conference, 2006, vol 2, pp 913–917. https://doi.org/10.1109/CCNC.2006.1593171

29. Jeong J, Leconte M, Proutiere A (2016) Cluster-aided mobility predictions. In: Proceedings of the 35th annual IEEE international conference on computer communications. INFOCOM'16. https://doi.org/10.1109/INFOCOM.2016.7524491
30. Shlesinger MF, Zaslavsky GM, Frisch U (1995) Lévy flights and related topics in physics. Lecture notes in physics, vol 450. Springer, Berlin. https://doi.org/10.1007/3-540-59222-9
31. Paul U, Subramanian AP, Buddhikot MM, Das SR (2011) Understanding traffic dynamics in cellular data networks. In: 2011 proceedings IEEE INFOCOM, pp 882–890. https://doi.org/10.1109/INFCOM.2011.5935313
32. Couronné T, Smoreda Z, Raimond AO (2013) Chatty Mobiles: individual mobility and communication patterns. CoRR abs/1301.6553. http://arxiv.org/abs/1301.6553
33. Hess A, Marsh I, Gillblad D (2015) Exploring communication and mobility behavior of 3G network users and its temporal consistency. In: 2015 IEEE international conference on communications (ICC), pp 5916–5921. https://doi.org/10.1109/ICC.2015.7249265