Gramaglia, M., Fiore, M., Furno, A. & Stanica, R. (2021). GLOVE: Towards Privacy-Preserving Publishing of Record-Level-Truthful Mobile Phone Trajectories. *ACM/ IMS Transactions on Data Science, 2*(3), article nº 21, pp. 1–36.

DOI: 10.1145/3451178

GLOVE: towards privacy-preserving publishing of record-level-truthful mobile phone trajectories

MARCO GRAMAGLIA, University Carlos III of Madrid, Spain
MARCO FIORE, IMDEA Networks Institute, Spain
ANGELO FURNO, IFSTTAR-ENTPE, Univ. Lyon, France
RAZVAN STANICA, Univ. Lyon, INSA Lyon, Inria, CITI, France

Datasets of mobile phone trajectories collected by network operators offer an unprecedented opportunity to discover new knowledge from the activity of large populations of millions. However, publishing such trajectories also raises significant privacy concerns, as they contain personal data in the form of individual movement patterns. Privacy risks induce network operators to enforce restrictive confidential agreements in the rare occasions when they grant access to collected trajectories, whereas a less involved circulation of these data would fuel research and enable reproducibility in many disciplines. In this work, we contribute a building block towards the design of privacy-preserving datasets of mobile phone trajectories that are truthful at the record level. We present GLOVE, an algorithm that implements k-anonymity, hence solving the crucial *unicity* problem that affects this type of data while ensuring that the anonymized trajectories correspond to real-life users. GLOVE builds on original insights about the root causes behind the undesirable unicity of mobile phone trajectories, and leverages generalization and suppression to remove them. Proof-of-concept validations with large-scale real-world datasets demonstrate that the approach adopted by GLOVE allows preserving a substantial level of accuracy in the data, higher than that granted by previous methodologies.

CCS Concepts: • **Security and privacy** → **Data anonymization and sanitization**; *Pseudonymity, anonymity and untraceability*; Usability in security and privacy.

Additional Key Words and Phrases: mobile phone trajectories, data publishing, anonymization, truthfulness

ACM Reference Format:
Marco Gramaglia, Marco Fiore, Angelo Furno, and Razvan Stanica. 2021. GLOVE: towards privacy-preserving publishing of record-level-truthful mobile phone trajectories . 1, 1 (February 2021), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Mobile network operators can easily record time-stamped and geo-referenced network events generated by mobile devices (*e.g.*, upon attach and detach, voice call set-up and termination, text message transmission and reception, or data session establishment). Sequencing all events related to a same mobile phone returns irregularly sampled spatiotemporal trajectories that classify as *movement micro-data*, *i.e.*, data yielding mobility information about single individuals.

Mobile phone tracking can be concurrently performed on millions of users, hence the resulting trajectories provide an unprecedented portrait of the movements and undertakings of very large

Authors' addresses: Marco Gramaglia, mgramagl@it.uc3m.es, University Carlos III of Madrid, Madrid, Spain; Marco Fiore, marco.fiore@imdea.org, IMDEA Networks Institute, Madrid, Spain; Angelo Furno, angelo.furno@ifsttar.fr, IFSTTAR-ENTPE, Univ. Lyon, Lyon, France; Razvan Stanica, razvan.stanica@insa-lyon.fr, Univ. Lyon, INSA Lyon, Inria, CITI, Lyon, France.

XXXX-XXXX/2021/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

populations. This original micro-data has rapidly established as a fundamental instrument to scale up studies and infer novel knowledge across many disciplines, including physics, sociology, epidemiology, urban planning, network engineering, and transportation research [1, 2]. Unfortunately, operators are extremely reticent to publish mobile phone trajectory datasets. Even when granted, access is regulated by non-disclosure agreements that bound the scope of the data analysis, and prevent publication of results without prior verification by the relevant authorities. This is, *e.g.*, the solution adopted in the case of the mobile traffic data used in our study. Clearly, this practice hinders the development of novel mining techniques, the discovery of new knowledge, and the reproducibility and verifiability of research using mobile phone trajectory data.

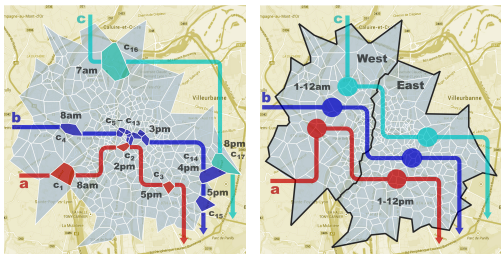
Privacy risks are a major impediment in this sense: individual movements can reveal a great deal about the habits of a person, including, *e.g.*, home and work locations, commuting and personal time schedules, or visits to sensitive places such as clinics and nightclubs. Thus, the disclosure of substantial information about the displacements of subscribers may violate basic privacy requisites.

The common countermeasure adopted by network operators is *pseudonymisation*, *i.e.*, replacing all personal *identifiers* (*e.g.*, name, phone number, IMSI, TMSI) with *pseudo-identifiers* (*e.g.*, random or hashed values). However, as for other types of micro-data –see, *e.g.*, the eminent cases of medical records [3] and web service databases [4]– such naive anonymization is not robust against re-identification. The foremost reason why pseudonymisation does not work is the high *unicity* that affects mobile phone trajectories. Unicity stems from the fact that mobile subscribers have very distinctive movement patterns, which make them univocally recognizable even in very large populations. Seminal experiments showed that 50% of the mobile phone trajectories in a 25 million-strong dataset turned out to be unique when considering the three most frequent locations they contain [5]. Similarly, any mobile phone trajectory could be pinpointed with near certainty among 1.5 millions entries by just knowing five of its spatiotemporal points picked at random [6].

Unicity is not a privacy threat per-se, but it is a vulnerability that creates substantial opportunities for re-identification, via attacks that cross-correlate the target dataset with other databases. Recent studies have cross-correlated mobile phone trajectories with publicly available social networks metadata from Flickr and Twitter [7] credit card records [8], or public transit logs [9]. In all these cases, some mobile phone trajectories could be associated to user identities with high probability, even if the trajectory dataset had been transformed by means of pseudo-identifiers.

Mitigating unicity becomes then a very desirable facility that makes datasets more privacy-preserving, favoring their disclosure. However, achieving this objective is especially challenging when one wants to preserve *truthfulness at the record level* in mobile phone trajectory data, *i.e.*, publish trajectories that correspond to actual individuals in real life. Such a level of dependability is important not to curb the extent of subsequent mining tasks: for instance, it ensures that no systematic bias is introduced by the data transformation when studying specific geographical regions or time intervals; or, it allows focusing on subsets of data subjects that reveal to be particularly interesting upon a statistical analysis on the whole user population [10].

The legacy solution to remove unicity while preserving truthfulness at the record level in other micro-data datasets (*e.g.*, relational databases) is generalization, which homogeneously reduces data precision to the point where no individual record is uniquely distinguishable. Previous studies showed that such an approach does not suit well mobile phone trajectories, whose unicity is reduced only when lowering their spatiotemporal granularity at levels that compromise data utility, such as day-long and citywide generalization [5]. In fact, a power-law relationship exists between unicity and spatiotemporal generalization, implying that stronger privacy guarantees come at a rapidly increasing cost in terms of reduced data resolution [6]. Overall, not only mobile phone trajectories have high unicity, but the latter is also hard to eliminate: ensuring privacy in these datasets easily compromises their utility. Our work tackles this precise problem, with a two-fold contribution.



(a) Original (b) Generalized

Fig. 1. Example of mobile phone trajectories. (a) Original data: mobile device locations are at cell level, and the temporal information has a hourly precision. (b) Spatiotemporal generalization: location is limited to the Eastern or Western half of the city, and time has 12-hour accuracy.

First, we carry out a thorough investigation of the reasons behind the inconvenient properties of mobile phone trajectories outlined above. We define an original measure of the *anonymizability* of spatiotemporal trajectories. When applied to real-world datasets, our measure offers new insights on the causes of the high unicity and poor anonymizability of this type of movement micro-data, which was not individuated in previous works.

Second, we propose an original algorithm that removes such unicity, by implementing the k-anonymity privacy criterion in datasets of mobile phone trajectories. The algorithm is named GLOVE, as its purpose is hiding the "digital fingerprints" left by mobile subscribers when they interact with the cellular network. By building on the insights above and a greedy use of the anonymizability measure, GLOVE preserves substantially higher data accuracy than previous attempts at k-anonymizing mobile phone trajectories.

2 PROBLEM AND POSITIONING

In this section, we first formalize the general problem of unicity in mobile phone trajectory datasets, by introducing some basic definitions (Sec. 2.1). Then, we outline the precise scope of our work with respect to the general problem above by establishing the objective we target (Sec. 2.2), and the attacker model we assume (Sec. 2.3). Based on these considerations, we introduce a suitable privacy model (Sec. 2.4). Finally, we clarify the applications and limitations of our methodology (Sec. 2.5).

2.1 Definitions

Mobile phone trajectory data is collected by operators through probes deployed in their networks. Every mobile communication activity, either in the user or in the control plane, generates network events that are time-stamped and mapped to the geographical location of the antenna the device is currently associated to. We term the space and time information associated to each logged event a *spatiotemporal sample* (also referred to as *sample* in the following). The complete sequence of spatiotemporal samples related to a same mobile device is a *mobile phone trajectory*, and it is a proxy for the movements of the subscriber owning the device. Mobile phone trajectories have important distinguishing features: (*i*) they are irregularly sampled, since the communication events are not deterministic; and, (*ii*) they are sparse in time, typically including a few tens of samples per day each. An illustrative example is provided in Fig. 1a, which portrays three mobile phone trajectories, denoted as a, b, and c, respectively, across an urban area. For instance, device a interacts with the network at 8 am, while in cell c_1. Then, it triggers additional network events at 2 pm, while in cell c_2 in the city center, and at 5 pm, from a cell c_3 in the South-East outskirts. Thus, the trajectory of a is (c_1,8; c_2,14; c_3,17).

id	mobile phone trajectory					
a	c_1,8	c_2,14	c_3,17			
b	c_4,8	c_5,15	...	c_{13},15	c_{14},16	c_{15},17
c	c_{16},7	c_{17},20				

name	observations	
Alice	c_6,15	c_{15},20

location	item	cost
c_{15},17	Imipramine	$50
c_4,11	Shoes	$120
c_8,13	Videogame	$35

Fig. 2. Movement micro-data from the trajectories in Fig. 1a (left), side information database owned by the adversary (middle), and location database with sensitive information (right).

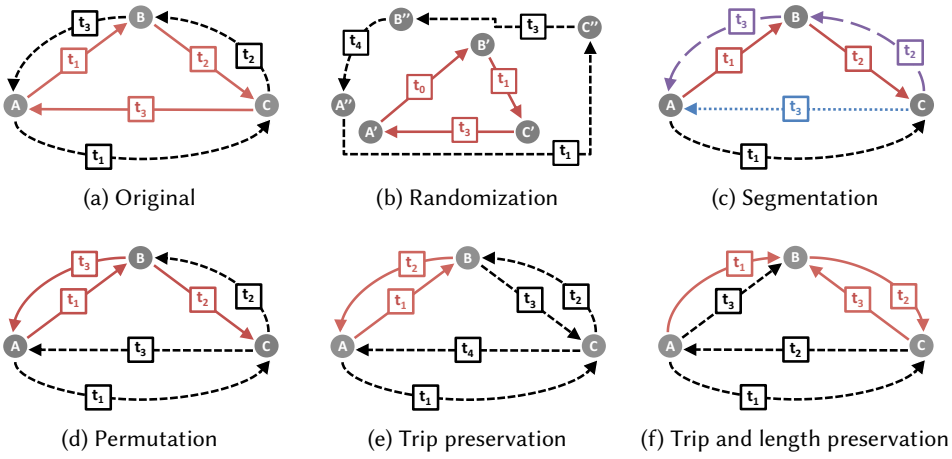


Fig. 3. Examples of mobile phone trajectories that are not truthful at the record level. (a) Original trajectories of two users, visiting three locations A, B and C. The first user (solid red arrows) has a trajectory $(A, t_1) \rightarrow (B, t_2) \rightarrow (C, t_3)$, whereas the second user (dashed black arrows) follows a pattern $(A, t_1) \rightarrow (C, t_2) \rightarrow (B, t_3)$. (b) Anonymized trajectories produced by randomizing locations and timestamps, as proposed in [11]. (c) Anonymized trajectories produced by changing the pseudo-identifier of each user at regular time intervals, as proposed in [12]. (d) Anonymized trajectories produced by iteratively swapping pseudo-identifiers among users, as proposed in [13]. (e) Anonymized synthetic trajectories produced by a model that preserves (*e.g.*, differentially private) statistics of flow volume between any two locations, such as that proposed in [14]. (f) Anonymized synthetic trajectories produced by a model that also preserves (*e.g.*, differentially private) statistics of the length of original trajectories, such as that proposed in [15]. The individual paths in (b–f) are not representative of actual users in (a).

Mobile phone trajectories are gathered into databases of movement micro-data. Fig. 2 (left) provides a toy example of such database for the trajectories in Fig. 1a. The first column of the table reports the character denoting the mobile device. In practical cases, the database is pseudonymised, and a, b, and c are pseudo-identifiers. Typical examples of pseudo-identifiers are random strings or irreversible hashes of the subscriber's name, phone number, IMSI or TMSI. Pseudo-identifiers remove the direct association of trajectories to persons, hence avoiding an explicit privacy breach in the database.

2.2 Objective

The ultimate objective to which our work contributes is *Privacy-Preserving Data Publishing (PPDP)*, *i.e.*, the provisioning of methods for the publication of information that is both privacy-preserving and useful for subsequent analysis. In our case, information maps to movement micro-data from mobile traffic, *i.e.*, mobile phone trajectories. Practical PPDP entails four principles, which we briefly review next.

P1. *The non-expert data publisher.* Mining of the data is performed by the data recipient, and not by the data publisher. The only task of the data publisher is to anonymize the data for publication. This principle stems from the consideration that, in a desirable open data circulation model, the publisher typically (*i*) is not aware of the identity of all possible recipients of the data, and (*ii*) does not have the knowledge to perform the analyses that such recipients will carry out [10].

P2. *Publication of data, and not of data mining results.* PPDP aims at producing privacy-preserving datasets of mobile phone trajectories rather than anonymized datasets of classifications, association rules, or aggregate statistics. The rationale is that micro-data grants infinitely higher flexibility in subsequent explorations, as it provides access to individual-level information that is lost in data mining outputs [10]. We remark that, when considered jointly with P1 above, this principle sets PPDP apart from Privacy-Preserving Data Mining (PPDM), where future usages of the data are known in advance. PPDP does not limit the scope of data analyses as PPDM does, and it is also a harder problem in general.

P3. *Truthfulness at the record level.* Each published record must correspond to an existing user in real life. In the case of mobile phone trajectories, this principle entails that each database record must describe the movement of an actual subscriber, with each spatiotemporal sample mapping to a location truly visited by the target individual at the associated time. We already mentioned in Sec. 1 the importance of this principle: if it is not met, each trajectory in the dataset describes a personal mobility pattern that does not correspond to any real user, which jeopardizes the credibility of follow-up individual-level studies based on the data [10].
Fig. 3 provides intuitive toy examples of this situation. The different diagrams evidence how randomization (b), segmentation (c), permutation (d), or models generating synthetic trajectories based on, *e.g.*, differentially private statistics (e, f), ultimately result in fictitious individual trajectories that do not correspond to the mobility of any real user. While they preserve specific aggregate statistics, these approaches create anonymized data with an inherent and unpredictable bias at record level, which makes all subsequent analysis of isolated trajectories unreliable, and severely limits the utility of the published data.

P4. *The data recipient could be an attacker.* Published data is openly available, thus data recipients could also be attackers. This makes PPDP different from database encryption, which assumes that only authorized and trustworthy recipients in possession of the required cryptographic material can access the clear text.

Our privacy model, laid in Sec. 2.4, obeys all these principles.

2.3 Attacker model

According to a popular classification [10], attacks on published micro-data can be categorized into four classes: *(i) record linkage* aims at linking the side information owned by the adversary with a single record in the target database; *(ii) attribute linkage* aims at linking the side information owned by the adversary with a sufficiently accurate value of a specific (and typically sensitive) attribute in the target database; *(iii) table linkage* aims at inferring whether an individual is present in the target database; *(iv)* a *probabilistic attack* aims at improving the overall knowledge or some precise belief of the adversary by accessing the target database.

In this work we tackle the first type of attack above, *i.e.*, record linkage, a toy example of which is provided in Fig.2. There, an adversary owns side information (two spatiotemporal samples, in the middle) about his victim, Alice, which he collected by, *e.g.*, seeing her there (middle table). The adversary has also access to a database of credit card transactions that contains a timestamped location of the retailer and (possibly sensitive) information on the item purchased (table on the right). A simple cross-database correlation lets the adversary link Alice's identity to trajectory *b*,

since the observed sample (c_6,15) only appears there, *i.e.*, is unique to that record; similarly, he can also link the first entry in the credit card database to trajectory b, as the (c_{15},17) location and time pair is only found in that record. The successful record linkages allow the adversary to learn that Alice bought a quite strong antidepressant, which indicates a medical condition that Alice may not want to disclose, hence represents a privacy breach. The example makes it clear that *(i)* a successful record linkage exploits the unicity of patterns of spatiotemporal samples in the trajectories, and *(ii)* pseudonymisation does not offer any protection in this case. Unfortunately, as discussed in Sec. 1, unicity is an intrinsic characteristic of datasets of mobile phone trajectories [5, 6], which paves the way for high probability of positive record linkage in real-world case studies [7–9].

The example also underscores how the chances of success of a record linkage attack depend on the adversary's side information: additional observations could have led to the re-identification of Bob's and Charlie's trajectories as well. Since no reliable model of the attacker's knowledge exists in the context of spatiotemporal data [17], we follow the common practice of assuming that any subset of a user's trajectory can potentially be part of the side information [18]. We stress that this is a worst-case scenario when considering that the target dataset and the adversary's side information are constituted of spatiotemporal samples only. By opting for this approach, our attacker model is more general and harder to address than models bounding the knowledge of the adversary to, *e.g.*, the set of locations most frequently visited by the target individual [5], or a fixed subset of his trajectory samples [6].

2.4 Privacy model

Our privacy model is designed abiding by the PPDP objective and record linkage attacker model outlined in Sec. 2.2 and Sec. 2.3, respectively.

First, we want to remove unicity in databases of mobile phone trajectories, so as to inhibit any possibility of record linkage against the data: formally, this maps to pursuing the privacy principle of *indistinguishability*. The suitable criterion to this end is *k-anonymity*, which commends that each record in a dataset must be indistinguishable from at least k-1 other records in the same dataset [3]. In our context, each mobile phone trajectory needs to be hidden in a crowd of k identical ones in the database.

Second, we consider k-anonymity of *full-length* mobile phone trajectories. This abides by the notion of *quasi-identifier-blind anonymity* in the milieu of trajectory data [18], and is the only way to ensure indistinguishability against an adversary that possesses a generic subset of the target user's trajectory. Note that full-length trajectory anonymization implicitly protects the data from attacks that narrow the side information available to the adversary, as assumed in several prominent previous studies [5, 6].

Third, we adopt *spatiotemporal generalization* and *suppression* as the techniques to achieve k-anonymity. As anticipated in Sec. 1, spatiotemporal generalization relies on reducing data precision in space and time so as to make samples of different mobile phone trajectories identical. Suppression allows instead removing some data, from individual samples to whole phone trajectories (*i.e.*, users), if their presence in the database prevents fulfilling the anonymity criterion.

Summarizing the points above, our target is the k-anonymization of full-length mobile phone trajectories through spatiotemporal generalization and suppression. This privacy model fully conforms to the PPDP principles set forth in Sec. 2.2, as follows.

- The model does not entail any requirement on the expertise of the publisher about data recipients or their purposes, since the adopted privacy criterion and data transformations are independent of those aspects. It thus satisfies principle P1.
- The model preserves the format of the original data, generating anonymized individual trajectories each composed by a sequence of spatiotemporal samples. This allows performing

with the published dataset any data mining task that could be run on the original data, hence meeting principle P2.
- The model employs generalization and suppression of samples, which exclusively reduce the granularity of the mobile phone trajectories, and do not inject new, fabricated information in the database. A proper implementation of these techniques thus guarantees that each anonymized record corresponds to the actual mobility of a real user, although with a lower precision in space and time. This satisfies principle P3.
- The model does not resort to encryption at any stage, hence does not make any assumption on the integrity of recipients. It thus fulfills principle P4.

A toy example of our privacy model is in Fig. 1b, for the mobile phone trajectories of Fig. 1a. The data is generalized in space, as cells are aggregated into macroscopic East and West regions; generalization also occurs in time, as the temporal accuracy is reduced to 12 hours. The three full-length trajectories a, b and c are now indistinguishable, and equal to (West,1-12; East,13-24): they are thus 3-anonymized. Note that the process preserves truthfulness at the record level, since each generalized spatiotemporal sample corresponds to a region actually visited by each user during the recorded time interval.

Clearly, generalization induces a loss of accuracy in the data. In the example of Fig. 1b, the 3-anonymized mobile phone trajectory a, b and c are very coarse in both space and time. This is precisely the problem outlined in Sec. 1: with trajectory micro-data, even guaranteeing a baseline 2-anonymity requires a reduction of granularity so severe to impair data utility [5, 6]. Our goal is then enforcing k-anonymity in datasets of mobile phone trajectories while preserving substantial accuracy, which represents a key contribution towards privacy-preserving publishing of record-level-truthful data.

2.5 Relevance and limitations

Achieving k-anonymity of mobile phone trajectories is an open problem. Several solutions proposed for other kinds of spatiotemporal micro-data can be applied to this context, as thoroughly discussed in Sec. 3. However, even the most suitable approaches yield unsatisfactory performance and leave substantial space for improvement, as also later demonstrated by our comparative evaluation in Sec. 7.3. In this work, we make a step ahead to close the methodological gap above, and develop a k-anonymization solution that preserves substantially increased utility in the trajectory micro-data.

We remark, however, that ours does not pretend to be a comprehensive solution to the problem of privacy-preserving publishing of mobile phone trajectories. The k-anonymity of full-length trajectories is a robust countermeasure against a well determined type of threat, *i.e.*, the record linkage attack presented in Sec. 2.3. In practice, our solution provides protection in cases where the unicity of information contained in the records is the issue, and just pinpointing a single user in the trajectory database leads to a privacy breach. This kind of risk is removed by k-anonymity.

Although record linkage is by far the most common attack against spatiotemporal trajectories considered in the literature to date, other threats can be envisioned against which k-anonymity has known limitations. For instance, it has been shown that this privacy criterion does not offer protection against adversaries aiming at attribute linkage (*e.g.*, by exploiting the homogeneity of sensitive attributes among k-anonymous records), at localizing users, or at disclosing presence, meetings and sensitive places [19, 20]. Addressing attacks other than record linkage is beyond the scope of this paper, and there exist other privacy criteria that are designed to counter such more complex threats. Nevertheless, as comprehensively discussed in [10], countermeasures to complex attacks like l-diversity [19], t-closeness [21], or $k^{\tau,\epsilon}$-anonymity [22] build on k-anonymity; if no dependable method implementing k-anonymity exists for mobile phone trajectory datasets, all these approaches are also unfeasible. And, strategies that build on entirely different methods

like differential privacy [23] have substantial limitations in the context we target, as thoroughly discussed in the next Section.

In light of these considerations, the work presented in this paper shall be regarded as a fundamental building block towards a complete suite of tools that enable robust PPDP of mobile phone trajectories. It is not as a definitive self-contained solution, rather a possibly important step towards that goal.

3 RELATED WORK

Our work deals with privacy preservation in movement micro-data. This is a very different problem from ensuring anonymity in relational micro-data [3, 4]. It also differs from confidentiality problems in other types of databases that can be extracted from mobile network operator data, *e.g.*, networks of subscriber relationships [24], or origin-destination matrices [25]. Anonymization techniques designed for such databases are not applicable to our context, due to the different data format and semantics.

Within the domain of movement micro-data, we focus on spatiotemporal trajectories, not to be confounded with other kinds of movement micro-data, *i.e.*,*(i)* positioning data collected in location-based services (LBS), and *(ii)* spatial trajectories. Below, we present privacy solutions for LBS and spatial trajectories, and detail how they are not applicable to our context. We then thoroughly discuss previous methods for PPDP of spatiotemporal trajectories, based on k-anonymity as well as on other privacy criteria.

Privacy in location-based services. In LBS, users query service providers for location-dependent information; as queries necessarily embed the position of the requester, providers may accumulate substantial data about the whereabouts of individual users over time, and mine such data to extract sensitive knowledge about the private habits of users. The relevant privacy problem is then transforming user positions before they are included in the queries, in a way that they do not reveal the actual location of the requester and yet allow the provider to return a meaningful reply. The transformation can be carried out by *(i)* an intermediate trusted server, or *(ii)* the user's device generating the query itself.

Server-based solutions are mostly based on k-anonymity, *i.e.*, aim at guaranteeing that the positioning information in each query is indistinguishable from that in other k-1 queries generated at approximately the same time, so that no single query makes a user uniquely identifiable. Different approaches can be used to this end, including temporal [26], spatial [27], and spatiotemporal generalization [28], or encryption [29]. Other criteria than k-anonymity have also been considered, and notably *geo-indisitinguishability* [30] and its extensions [31–33], which adapt *differential privacy* to the specific case of location data. Differential privacy is a protection strategy that offers provable privacy guarantees by commending that the presence of each user's data in a published dataset does not change substantially (*i.e.*, beyond a predetermined pivacy budget) the output of the analysis, thus formally bounding the privacy risk [23]. Also personal device-based methods mainly rely on the same approach, and more specifically on *local differential privacy* (LDP), a distributed variant on differential privacy that applies transformations at the data source directly [34].

Independently of the entity in charge of anonymization, LBS scenarios are different from ours. Privacy-preserving transformations in LBS aim at protecting individual queries, *i.e.*, single spatiotemporal points in isolation, which is a subset of the problem of anonymizing complete spatiotemporal trajectories we address. Moreover, the transformation of positioning data occurs in real-time, whereas we target PPDP, which is generally concerned with offline anonymization of large historical databases.

Privacy in LBS becomes a problem closer to the one we consider when untrusted servers can track users by linking LBS queries in time: in this case, privacy must be enforced on sequences

of points, *i.e.*, trajectories. Popular countermeasures involve pseudo-identifier replacement in *mix zones* where different user paths cross each other [35]. Enhancements include *confusion*, which uses temporal generalization to make it easier for paths to traverse mixing zones at the same time [36], or *camouflage*, which lets users generate queries that refer to their whole predicted trajectory between two mixing zones [37]. However, while appropriate for on-line querying, pseudo-identifier replacement is not a solution in the case of PPDP; indeed, it disrupts the integrity of trajectories, and severely compromises data utility. Moreover, it was recently shown that pseudo-identifier replacement does not work in mobile phone trajectory datasets: even by replacing pseudo-identifiers at every 6 hours, 80% of the trajectories remain uniquely identifiable within a million-record database by an adversary knowing just four random points visited by the target individual [12].

PPDP of spatial trajectories. Spatial trajectories are sequences of geographical locations without a temporal reference: they describe individual routes, regardless of when they are generated. Solutions for the anonymization of spatial trajectories typically rely on spatial generalization [38]. Yet, their design does not allow for straightforward extensions to the time dimension of the data. In fact, accounting for the temporal information makes the anonymization problem significantly more complex, and ultimately makes solutions for spatial trajectories unsuitable to our context.

PPDP of spatiotemporal trajectories with k-anonymity. A number of techniques have been proposed for the k-anonymization of spatiotemporal trajectories. However, none is fully compliant with the principles of PPDP, and none is specifically designed for mobile phone trajectory data. In [39], spatiotemporal trajectories are perturbed, by shifting the position of mobile users across space, so that at least k users are at the same location at each time instant. However, perturbation violates principle P3 of PPDP, as it displaces users to locations they may have never actually visited and creates fictitious mobility. The same holds for the permutation of samples among similar trajectories employed in [40]. In [41], suppression is used to remove spatiotemporal samples that hinder k-anonymity: the approach works on short (*e.g.*, three-sample long) subsets of trajectories, but does not scale to the full-length trajectories we consider.

A solution based on generalization was presented in [42]. However, this proposal, as well as those in [39, 40], target datasets where the positions of all users are sampled with identical periodicity. The anonymization algorithms only act on the spatial dimension, since temporal merging is implicit. The datasets evaluated in [39, 42] are either composed of GPS logs or synthetic trajectories, which allows accommodating the fixed sampling assumption. However, the same premise does not hold in the case of mobile traffic, where the sampling depends on mobile user activity and is extremely heterogeneous across the user population.

The only approaches proposed to date that are capable of handling the k-anonymization of full-length phone trajectories along both space and time dimensions are TGA [43] and W4M [44]. TGA relies on generalization and suppression, however it builds on computationally expensive logarithmic operations to determine which trajectories should be generalized and how samples should be merged. The method was shown to scale to datasets of thousands of trajectories, which are still orders of magnitude smaller than those typically collected by mobile network operators. For instance, the datasets we employ in our analysis include up to hundreds of thousands of users, and production-level datasets easily encompass millions of subscribers [2]. W4M is similarly based on generalization and suppression, but in addition it creates fabricated samples that are included in the individual trajectories to ease their k-anonymization. Such new samples are generated exclusively for data protection and without any consideration for the mobility of the real-world user, hence typically violate again principle P3 of PPDP. Notwithstanding their limitations, TGA and W4M represent the only competitor solutions for k-anonymization, and we thus employ them as benchmarks for comparative analysis, ultimately showing that GLOVE offers better performance in terms of anonymized data accuracy.

PPDP of spatiotemporal trajectories with differential privacy. A quite large body of works have investigated applications of data privacy to the PPDP of mobile phone data. However, the vast majority of the proposed solutions concerns the publication of aggregate statistics derived from the original spatiotemporal trajectories, and not of the trajectories themselves. Representative examples include the release of differentially private *quadtrees* [45], spatiotemporal density [46, 47], transit graphs [48], or histograms [49]. Nevertheless, none of these solution produces spatiotemporal trajectories, which is instead our objective.

Unfortunately, generating differentially private trajectory data is extremely challenging. The aggregate statistics above boil down to counting data, for which effective differential privacy mechanisms based on Laplacian or exponential noise are well understood. This is not the case for very high dimensional data like spatiotemporal trajectories, where each database record is constituted by a large number of time-space tuples. In absence of sound theoretical results, two approaches have been explored. The first is weakening the requirements of actual differential privacy, for instance by allowing that a portion of the spatiotemporal points in a trajectory can be disclosed, as in (ϵ, δ)-differential privacy; however, implementing even these diluted models requires questionable compromises, such as assuming that the origin and destination locations of each trip are not sensitive and can be published [50]. Another attempt extends k-anonymity to provide equivalent guarantees than differential privacy in the context of trajectory data, with so-called $k^{\tau,\epsilon}$-anonymity; yet, the practical algorithm assumes that the adversary's knowledge is limited to a single, short sequence of spatiotemporal points that are collected during a continuous time interval [22], and cannot cope with a more general and realistic attacker model like the one we consider.

The second, more popular, strategy is generating synthetic trajectories based on differentially private aggregate statistics computed from the original data. The process involves *(i)* deriving one or more summaries of the original data, *(ii)* adding noise to the summaries so as to meet differential privacy requirements, and *(iii)* using the noisy summaries to produce new individual trajectories. Summaries considered in the literature encompass *prefix-trees* [14] and derived representations [51, 52], *n-grams* [53], spatial distributions of origin-destination locations and travel distances [15], or transition probabilities among locations and trip duration [54]. No matter the input statistics, and in addition to severe scalability problems emerging in some cases [52], all solutions above fail to meet principle P3 of PPDP. Indeed, they create synthetic trajectories that are guaranteed to preserve the veracity of the (noisy) summaries at the base of the generation process. But, when descending to the individual record level, they do not provide any guarantee that a single trajectory is actually representative of the movement of a real-world user. Fig. 3e and Fig. 3f provide examples of that situation in a very simple case.

Overall, as of today, differential privacy and its derived definitions still appear impractical for PPDP of spatiotemporal trajectories that are truthful at the record level, which is the problem we tackle in this work.

4 MEASURING ANONYMIZABILITY

Our first objective is understanding the causes of the *(i)* high unicity and *(ii)* resistance to generalization that characterize movement micro-data from mobile phone trajectories. To that end, we propose and leverage a measure of the level of *anonymizability* of a mobile phone trajectory, which estimates how easy (or difficult) it is to hide a given spatiotemporal trajectory in a dataset.

Coherently with the scope of our work outlined in Sec. 2, the measure is based on the k-anonymity criterion and assumes that generalization is employed to achieve it; indeed, no suppression is considered at this stage. Therefore, the measure evaluates the loss of spatial and temporal accuracy required to make a mobile phone trajectory indistinguishable from k-1 others in the same dataset.

We name our measure the *k-gap* of a mobile phone trajectory, and denote it as Δ_a^k for a trajectory a under k-anonymity. In the remainder of this section, we formally introduce the *k-gap* measure.

4.1 Sample stretch effort

The k-gap of a mobile phone trajectory depends on the cost of k-anonymizing the spatiotemporal samples that compose it. We thus start by introducing the *sample stretch effort*, *i.e.*, the loss of spatiotemporal accuracy required to merge two samples through generalization.

Let us first define the generic i-th sample of mobile phone trajectory a, which, as mentioned in Sec. 2.1, consists of the space and time information associated to the i-th logged event in the considered trajectory. Formally, we denote the sample as (σ_i^a, τ_i^a), and represent it as the composition of a spatial region σ_i^a and a temporal interval τ_i^a: more precisely, the tuple $\sigma_i^a = (x_i^a, dx_i^a, y_i^a, dy_i^a)$ outlines the vertices x_i^a, $x_i^a + dx_i^a$, y_i^a, $y_i^a + dy_i^a$ of a geographical rectangle, while $\tau_i^a = (t_i^a, dt_i^a)$ corresponds to the time interval between t_i^a and $t_i^a + dt_i^a$. Then, a sample (σ_i^a, τ_i^a) conveys that trajectory a crosses the area σ_i^a at some point in time within τ_i^a.

The rationale for chosing the representation above is twofold. First, it can accommodate cases where the precision of the positioning information is not absolute, as it allows capturing geographical (via dx_i^a and dy_i^a) and temporal (via dt_i^a) uncertainty in the data; for instance, as per Sec. 5.1, $dx_i^a = dy_i^a = 100$ m and $dt_i^a = 1$ min for all original trajectories in our reference datasets. Second, the representation is instrumental to our analysis, since it can inherently model generalized samples that are already the result of a loss of spatial (via dx_i^a and dy_i^a) and temporal (via dt_i^a) accuracy.

A generic formulation of the sample stretch effort $\delta_{ab}(i, j)$ between the i-th sample of trajectory a and the j-th sample of trajectory b is then

$$\delta_{ab}(i,j) = w_\sigma \phi_\sigma \left(\sigma_i^a, \sigma_j^b\right) + w_\tau \phi_\tau \left(\tau_i^a, \tau_j^b\right). \tag{1}$$

Here, $\phi_\sigma, \phi_\tau \in [0, 1]$ are functions that respectively determine the loss of accuracy in space and time induced by the merging of the two samples. The factors w_σ and w_τ weight the spatial and temporal contributions in (1). In the following, we will assume an unbiased model where the two dimensions have the same importance, thus $w_\sigma = w_\tau = 1/2$, which also ensures that $\delta_{ab}(i, j) \in [0, 1]$.

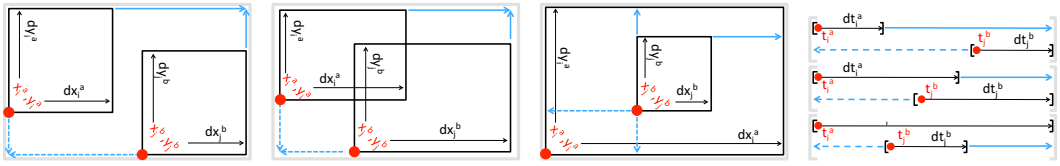
The functions ϕ_σ and ϕ_τ are designed by considering that both space and time generalizations induce a loss of information that is linear in the decrease of granularity, *i.e.*,

$$\phi_\sigma \left(\sigma_i^a, \sigma_j^b\right) = \begin{cases} \dfrac{\phi_\sigma^* \left(\sigma_i^a, \sigma_j^b\right)}{\phi_\sigma^{max}} & \text{if } \phi_\sigma^* \left(\sigma_i^a, \sigma_j^b\right) \leq \phi_\sigma^{max} \\ 1 & \text{otherwise,} \end{cases} \tag{2}$$

$$\phi_\tau \left(\tau_i^a, \tau_j^b\right) = \begin{cases} \dfrac{\phi_\tau^* \left(\tau_i^a, \tau_j^b\right)}{\phi_\tau^{max}} & \text{if } \phi_\tau^* \left(\tau_i^a, \tau_j^b\right) \leq \phi_\tau^{max} \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

In (2) and (3), the functions ϕ_σ^* and ϕ_τ^* model the stretch needed to match two samples in space and time, respectively. The constants ϕ_σ^{max} and ϕ_τ^{max} are the spatial and temporal thresholds above which the information loss is so severe that the data is not usable anymore[1].

[1]In our study, we set $\phi_\sigma^{max} = 20$ km and $\phi_\tau^{max} = 8$ hours, as we consider that a spatiotemporal granularity losing all intra-urban and morning-afternoon variability is of little interest to most data analyses.



(a) Space, disjoint (b) Space, partial overlap (c) Space, total overlap (d) Time

Fig. 4. Examples of spatiotemporal sample stretch. Light blue arrows indicate the left (dashed) and right (solid) stretch on the first (a's i-th) and/or second (b's j-th) samples. Plots refer to different levels of overlap between the spatial (a,b,c) and temporal (d) components.

Formally, the stretch in space ϕ^*_σ is computed as

$$\phi^*_\sigma\left(\sigma^a_i, \sigma^b_j\right) = \frac{\left[l_\sigma\left(\sigma^a_i, \sigma^b_j\right) + r_\sigma\left(\sigma^a_i, \sigma^b_j\right)\right] n_a}{n_a + n_b} + \\ + \frac{\left[l_\sigma\left(\sigma^b_j, \sigma^a_i\right) + r_\sigma\left(\sigma^b_j, \sigma^a_i\right)\right] n_b}{n_a + n_b}, \tag{4}$$

where

$$l_\sigma\left(\sigma^a_i, \sigma^b_j\right) = [x^a_i - \min(x^a_i, x^b_j)] + \\ + [y^a_i - \min(y^a_i, y^b_j)], \tag{5}$$

$$r_\sigma\left(\sigma^a_i, \sigma^b_j\right) = [\max(x^a_i + dx^a_i, x^b_j + dx^b_j) - x^a_i - dx^a_i] + \\ + [\max(y^a_i + dy^a_i, y^b_j + dy^b_j) - y^a_i - dy^a_i]. \tag{6}$$

The l_σ and r_σ functions quantify the *left stretch* and *right stretch* in space, *i.e.*, they measure the extent to which the boundaries of the first sample, σ^a_i, need to be stretched along the longitudinal and latitudinal axes, in order to fully cover the bounding rectangle of the second sample, σ^b_j. Graphical examples are in Fig. 4a–4c. In (4), the left and right stretches required for a's sample to geographically cover b's sample are summed with those required for b's sample to cover a's.

The sum in (4) is weighted by n_a and n_b. When a and b are the mobile phone trajectories of two individual subscribers, $n_a = n_b = 1$. However, the definitions above can accommodate the case where a and b do not refer to two subscribers, but to two groups of subscribers whose mobile phone trajectories have already been made indistinguishable. In that case, n_a and n_b represent the number of subscribers whose mobile phone trajectories have already been generalized into trajectories a and b, respectively. Hence, stretching a sample of trajectory a reduces the accuracy in the data of n_a users, and equivalently for b: therefore, the factors n_a and n_b ensure that the contributions to the global loss of data accuracy are weighted by the number of users affected by each generalization.

Equations are similar in the case of time, where

$$\phi_\tau^* \left(\tau_i^a, \tau_j^b\right) = \frac{\left[l_\tau \left(\tau_i^a, \tau_j^b\right) + r_\tau \left(\tau_i^a, \tau_j^b\right)\right] n_a}{n_a + n_b} + \\ + \frac{\left[l_\tau \left(\tau_j^b, \tau_i^a\right) + r_\tau \left(\tau_j^b, \tau_i^a\right)\right] n_b}{n_a + n_b}, \tag{7}$$

$$l_\tau \left(\tau_i^a, \tau_j^b\right) = [t_i^a - \min(t_i^a, t_j^b)], \tag{8}$$

$$r_\tau \left(\tau_i^a, \tau_j^b\right) = [\max(t_i^a + dt_i^a, t_j^b + dt_j^b) - t_i^a - dt_i^a]. \tag{9}$$

Again, l_τ and r_τ mark the left stretch and right stretch in time; illustrative examples are provided in Fig. 4d. As done in (4), the contributions of a's and b's stretches in (7) are weighted by the number of subscribers affected by the generalization.

4.2 k-gap of mobile phone trajectories

We can now define the *trajectory stretch effort*, *i.e.*, the spatiotemporal loss of accuracy required to merge two whole mobile phone trajectories via generalization. For two trajectories a and b, the associated effort Δ_{ab} is computed as

$$\Delta_{ab} = \begin{cases} \dfrac{1}{m_a} \displaystyle\sum_{i=1}^{m_a} \min_{j=1,\ldots,m_b} \delta_{ab}(i,j) & \text{if } m_a \geq m_b \\ \dfrac{1}{m_b} \displaystyle\sum_{j=1}^{m_b} \min_{i=1,\ldots,m_a} \delta_{ab}(i,j) & \text{otherwise.} \end{cases} \tag{10}$$

Here, m_a and m_b are the cardinality, *i.e.*, the number of samples, of trajectories a and b, respectively. The expression in (10) finds, for each sample in the longer phone trajectory, the sample at minimum stretch effort in the shorter trajectory. Δ_{ab} is the average of all such sample stretch efforts. The definition guarantees that $\Delta_{ab} = \Delta_{ba}, \forall a, b$.

The k-gap Δ_a^k of a trajectory a under k-anonymity is then computed as the average stretch effort of a from the nearest k-1 other trajectories in the dataset. Formally,

$$\Delta_a^k = \frac{1}{k-1} \sum_{b \in \mathbb{N}_a^{k-1}} \Delta_{ab}, \tag{11}$$

where $\mathbb{N}_a^{k-1}$ is the set of $k-1$ trajectories with the lowest trajectory stretch effort to a.

The expression in (11) returns a measure $\Delta_a^k \in [0,1]$ that indicates how costly it is to hide a trajectory a into a crowd of k other trajectories in the same dataset. If $\Delta_a^k = 0$, then a is already k-anonymous in the original data. If $\Delta_a^k = 1$, then a is so unique that k-anonymizing it makes all of its samples coarse beyond ϕ_σ^{max} and ϕ_τ^{max}, hence completely uninformative.

4.3 Design choices and implementation

We emphasize that all solutions adopted in the design of the sample and trajectory stretch efforts, and of the k-gap measure are primarily driven by a scalability rationale. Indeed, other formulations are possible for the expressions in (1)–(10), for instance non-linear relationships that include multiplicative, exponential or power-law terms in all these equations. However, mobile phone trajectory datasets are typically very large, *i.e.*, comprise from hundreds of thousands to tens of millions of records, each with hundreds of samples at least. Assessing the anonymizability of such large-sized datasets requires repeating the operations in (1)–(10) billions of times. In the light of this observation, we opted for elementary formulations that reduce the computational complexity of

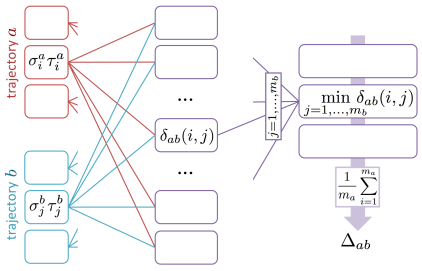


Fig. 5. Parallelization of calculations for the trajectory stretch effort Δ_{ab}.

calculations to a minimum. This led to considering a linear sum of space and time contributions in (1), constant spatial and temporal thresholds in (2) and (3), rectangular –rather than more complex polygonal or parametric geometrical forms– stretches in (4) and (7), and average –rather than higher-lever statistical measures– binding in (10) and (11).

In addition, the formulations in (1)–(10) also allow for a very efficient parallelization of the computation of trajectory stretch efforts. This is illustrated in Fig. 5. First, all possible sample stretch efforts $\delta_{ab}(i, j)$ need to be computed for all pairs i, j of samples in a and b: this is embarrassingly parallelizable, as the $m_a \times m_b$ calculations are independent of each other. Second, all m_b samples (assuming that $m_a > m_b$) that refer to the same sample i of a have to be compared, so that the minimum can be picked: the decision is independent for each sample of a, allowing again high parallelization. The final step involves averaging, an operation that is once more easily parallelized. These features make it possible to benefit from, *e.g.*, MapReduce paradigms or GPU architectures when implementing the code.

We also remark that, although we did not resort to such an approach in our implementation, the computational time of k-gap in (11) could be reduced by means of effective indexing techniques [55]. As a matter of fact, indexing allows pruning away irrelevant trajectories, and acccelerate the selection of the k-1 nearest trajectories required by the expression.

Finally, despite their deliberate simplicity, the formulations in (1)–(10) already do an effective job of estimating the cost of merging mobile phone trajectories. This is proven in Sec. 7, where the trajectory stretch effort is used by our proposed anonymization algorithm as a metric to select akin trajectories to be hidden into each other. There, they contribute to the solid performance of the solution in preserving the accuracy of the anonymized data.

5 ANONYMIZABILITY ANALYSIS

Thanks to the k-gap measure, we can investigate the poor anonymizability of mobile phone trajectories. To this end, k-gap could be used as a dissimilarity measure in legacy definitions used to assess micro-data sparsity like (ϵ, δ)-sparsity [4]. However, these definitions are more concise and less informative than complete distributions. We thus prefer characterizing the level of anonymizability of real-world datasets by analyzing the complete Cumulative Distribution Function (CDF) of the k-gap of all trajectories in each dataset.

5.1 Mobile phone trajectory datasets

For the purpose of our study, we extract movement micro-data, in the form of subscribers' mobile phone trajectories, from two datasets of mobile traffic released by Orange within their Data for Development Challenges [56]:

- **Ivory Coast.** The first dataset describes five months of Call Detail Records (CDR) in Ivory Coast. It contains the spatiotemporal trajectories of a subset of 50,000 randomly selected users, re-drawn every two weeks. Thus, the dataset contains information about a different

Table 1. Main features of the reference mobile phone trajectory datasets.

Dataset	Surface [Km^2]	BS	BS/Km^2	Users	Density [user/Km^2]	Samples [per user/h]	Timespan [days]
d4d-civ	322,463	1238	0.0038	82,728	0.26	0.75	14
d4d-sen	196,712	1666	0.0085	286,926	1.45	0.45	14

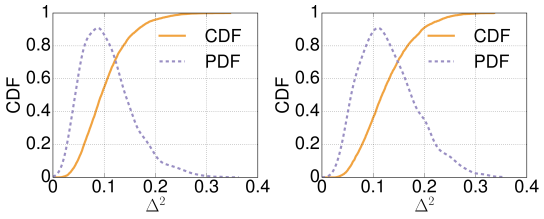


Fig. 6. CDF/PDF of k-gap, $k = 2$, in d4d-civ (left) and d4d-sen (right).

group of users during ten 2-week periods. Since many trajectories in the original dataset are exceedingly sparse, and have no samples at all in most of the days, we ran a preliminary screening, filtering out users that have less than one sample per day in their trajectory. Then, we merged the remaining trajectories from all periods into a single 2-week database of around 82,000 records. This dataset is indicated as d4d-civ.

- **Senegal.** The second dataset is derived from CDR collected in Senegal for one year. It contains a randomly selected subset of 320,000 users over a rolling 2-week period. In this case, users are already guaranteed to be active for more than 75% of the 14-day time span, therefore we consider a single representative period among those available, and filter out users who do not appear in each and every day. This yields a 2-week dataset with around 287,000 users, which is referred to as d4d-sen in the rest of the paper.

In both mobile traffic datasets, the positioning information is provided as a latitude and longitude pair, corresponding to the antenna location in d4d-civ and to a random point within the Voronoi cell associated to the antenna in d4d-sen. We mapped latitude and longitude to a bi-dimensional coordinate system using the Lambert azimuthal equal-area projection. We then discretized the resulting positions on a 100 m regular grid, which represents the maximum spatial granularity we consider. The maximum temporal precision granted by both datasets is one minute, and this is also our finest time granularity. The major features of the two datasets are summarized in Tab. 1.

5.2 The good: anonymity is close to reach

Our baseline result is portrayed in Fig. 6. The plot depicts the CDF of k-gap when considering 2-anonymity as the privacy criterion. Differences between d4d-civ and d4d-sen are minimal. Both CDFs are null at the x-axis origin, *i.e.*, no single mobile phone trajectory is 2-anonymous in either of the original datasets. The result is in line with previous analyses carried out on other equivalent datasets [5, 6]. This confirms that the high unicity of mobile phone trajectories is an intrinsic property of this type of datasets.

Observing the complete distribution provides however additional information that was not reported in previous studies. Interestingly, for both datasets the probability mass is below 0.2, *i.e.*, it is not far from the origin. This is good news, as it implies that the trajectory stretch effort needed to make most trajectories 2-anonymous is fairly low. As an example, 50% of the trajectories in d4d-civ have a k-gap of 0.09 or less, which maps, on average, to a combined spatiotemporal generalization of less than 1 km and little more than 20 minutes. This result seems to suggest that half of the individuals in the dataset can be 2-anonymized if the spatial granularity is decreased to 1 km, and the temporal precision is reduced to around 20 minutes. This is consistent across datasets, since in

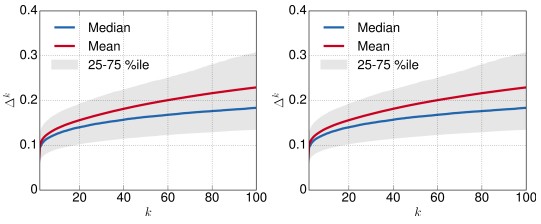


Fig. 7. k-gap statistics versus k, in d4d-civ (left) and d4d-sen (right).

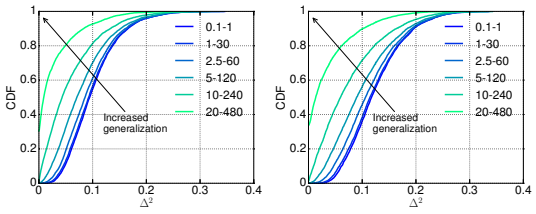


Fig. 8. CDF of k-gap, for $k = 2$ and various spatiotemporal generalizations (labeled in km–min), in d4d-civ (left) d4d-sen (right).

the d4d-sen case 80% of the trajectories have a k-gap of 0.17 or less, *i.e.*, an average distance from 2-anonymity of 1.7 km in space and 41 minutes in time.

One may wonder how more stringent privacy requirements affect these results. Fig. 7 shows the evolution of the k-gap distribution when k varies from 2 to 100. The curves represent the mean, median and first/third quartiles of the k-gap CDF, and the two plots refer to the d4d-civ and d4d-sen datasets. As expected, higher values of k require that a user is hidden in a larger crowd, and thus increase the effort in the generalization. However, quite surprisingly, the cost of k-anonymity appears to grow slowly and sub-linearly with k: ultimately, 100-anonymity does not appear much more difficult to achieve than 2-anonymity.

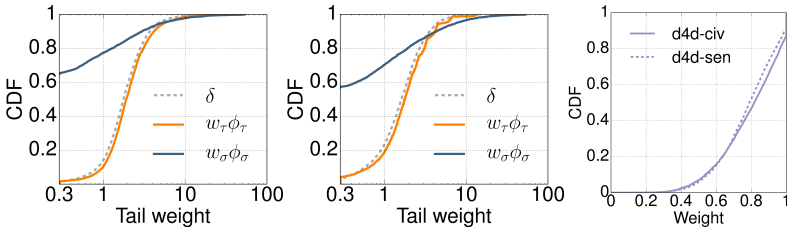
5.3 The bad: generalization does not work

Unfortunately, the easy anonymizability suggested by the results above is only apparent. Fig. 8 shows the impact of spatiotemporal generalization on k-gap, $k = 2$. Each curve corresponds to a different level of generalization of samples in mobile phone trajectories, from the original dataset granularity of 100 meters and 1 minute, to an uninformative granularity of 20 km and 8 hours, *i.e.*, 480 minutes. As one could expect, increased generalization pushes the distributions towards the left, *i.e.*, makes the datasets more privacy-preserving. However, the effect is mild: even a very coarse-grained generalized dataset where the spatiotemporal granularity is reduced to 20 km (the size of a large city) and 8 hours can 2-anonymize just ~35% of mobile phone trajectories. This result is again in agreement with previous studies [5, 6], and confirms that mobile phone trajectory datasets are resistant to anonymization via classical generalization.

5.4 The why: long-tailed time diversity

The results in Sec. 5.2 and Sec. 5.3 may seem incongruous: spatiotemporal generalization performs poorly (Fig. 8), yet the trajectory stretch effort needed to attain k-anonymity is in theory low (Fig. 6). In fact, the discrepancy is illusory, and due to the fact that the trajectory stretch effort in (10) is an average: evaluating the dispersion of the sample stretch efforts within each trajectory reconciles the views.

We retrieve, for each trajectory a in the dataset, the set $\mathbb{N}_a^{k-1}$ of the k-1 other trajectories that are the closest to a, according to (11). Then, we disaggregate the trajectory stretch efforts Δ_{ab} between



(a) TWI, d4d-civ (b) TWI, d4d-sen (c) Ratio

Fig. 9. (a,b) CDF of the Tail Weight Index computed on the distributions of sample stretch efforts (overall, and separated into spatial and temporal components) for all trajectories in d4d-civ and d4d-sen. (c) CDF of the temporal-to-spatial component ratios in the overall sample stretch effort, for all trajectories in d4d-civ and d4d-sen. All results for $k = 2$.

a and all $b \in \mathbb{N}_a^{k-1}$ into their individual sample stretch efforts, *i.e.*, $\min_{j=1,\ldots,m_b} \delta_{ab}(i,j)$ for each sample i of a as per (10). Finally, we collect the spatial and temporal components in (1) of all such sample stretch efforts into sets $\mathbb{S}_a^k = \{w_\sigma\phi_\sigma\}$ and $\mathbb{T}_a^k = \{w_\tau\phi_\tau\}$. The distributions of values in $\mathbb{S}_a^k$ and $\mathbb{T}_a^k$ unveil the stretch effort required to k-anonymize individual samples of trajectory a, in the spatial and temporal dimensions respectively. We are especially interested in studying the tails of such distributions, since they contain hard-to-anonymize samples that demand a high stretch effort (*i.e.*, a significant loss of accuracy) in space or time, in order to be hidden via generalization. We employ the Tail Weight Index (TWI) as a measure of the weight of the distribution tail [57]. The higher the TWI, the heavier the tail: to provide a reference, a negative exponential distribution with parameter equal to one has TWI 1.6, whereas a fat-tailed Pareto distribution with shape equal to one has TWI 14. We compute, for all trajectories in a dataset, the TWI of three CDFs: the total stretch effort per sample (labeled as δ in the plots), and the stretch efforts due to the spatial ($w_\sigma\phi_\sigma$) and temporal ($w_\tau\phi_\tau$) components.

Fig. 9a and Fig. 9b show the CDFs of the TWI computed on all trajectories in the d4d-civ and d4d-sen dataset. The TWI in the spatial dimension is below 1.6 in 75%–85% of cases: this implies that tail of spatial stretch distributions decays exponentially, if not faster, in the vast majority of cases. Instead, temporal stretch distributions are typically heavy tailed, with a TWI $\geq$ 1.6 in 70%–75% of cases. As a result, the TWI of the overall stretch effort distribution, denoted by δ, is shaped after that of temporal components.

Quantitative analyses confirm this last resolution. The plot in Fig. 9c shows the CDF of the temporal-to-spatial component ratios, *i.e.*, $\sum_{\mathbb{T}_a^k} w_\tau\phi_\tau / \sum_{\mathbb{S}_a^k} w_\sigma\phi_\sigma$, for every trajectory a in each of the two reference datasets. The CDF is skewed towards high values in the d4d-civ and d4d-sen datasets: in 95% of trajectories, the temporal stretch is larger than the spatial one; in half of the cases, the temporal stretch contributes to 80% or more of the total trajectory stretch effort; in 15% of cases, the cost of anonymization is fully determined by the temporal stretch. Overall, temporal components of mobile phone trajectories are much harder to anonymize than spatial ones: *where* an individual performs activities is easily masked; hiding *when* those occur is not.

5.5 Takeaways

The results presented in this section let us postulate that typical mobile phone trajectories are composed by a vast majority of spatiotemporal samples that are easily hidden among those of other users in the same dataset. This leads to a low k-gap of mobile phone trajectories.

However, mobile phone trajectories also feature a small but not negligible number of samples that create long tails in the sample stretch effort distributions. These samples are extremely difficult

to anonymize, mainly along their temporal dimension. Unfortunately, full-length k-anonymity requires that *all* samples of a phone trajectory are merged within those of k-1 other trajectories. This does not suit well the legacy *uniform generalization* adopted in the literature [5, 6], exemplified in Fig. 1b, and assumed in Sec. 5.3. Uniform generalization enforces the same reduction of granularity on all samples of all trajectories; as a result, the single sample that is the hardest to anonymize within a trajectory sets the (typically very high) bar for the loss of accuracy required to k-anonymize that trajectory.

In the remainder of the paper, we leverage this new understanding to design a more appropriate solution that effectively achieves k-anonymity in mobile phone trajectory datasets.

6 GLOVE

GLOVE is a novel algorithm for the k-anonymization of movement micro-data from mobile phone trajectories. It builds on the following insights from Sec. 5.5: *(i)* the vast majority of spatiotemporal samples in mobile phone trajectories can be hidden with limited loss of accuracy; *(ii)* only a reduced portion of samples require drastic generalization. Hence, GLOVE employs a *non-uniform generalization*, where the granularity of each generalized sample is not necessarily the same; instead, each sample undergoes an independent, minimal reduction of accuracy necessary to attain k-anonymity.

6.1 Algorithm in a nutshell

The pseudocode of GLOVE is listed in Alg. 1. The inputs to the algorithm are the mobile phone trajectory database $\mathbb{M}$ and the value of k, *i.e.*, the target k-anonymity level. In the initialization phase, all trajectories are tagged as 1-anonymous, *i.e.*, unique (line 2), and the trajectory stretch effort is calculated according to (10) for all trajectory pairs in $\mathbb{M}$, and stored in a matrix S (line 4).

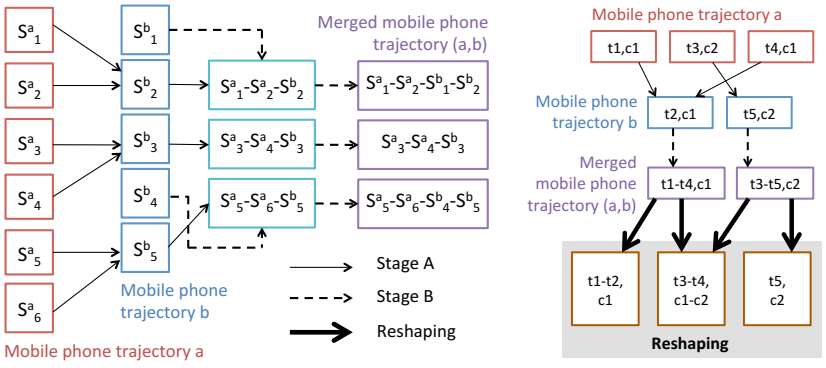
The algorithm then enters the main loop, which iterates until all trajectories have been k-anonymized (lines 7–19). At each iteration, function `minimumStretchEffort` identifies the pair of mobile phone trajectories a, b at minimum trajectory stretch effort in S, such that at least one of the two trajectories has not yet been k-anonymized (line 8). Function `removeRecords` removes the pair from $\mathbb{M}$ and S (line 9). The two trajectories are then generalized into a single trajectory by function `mergeTrajectories`; the resulting trajectory m is representative of a number of original trajectories equal to the sum of those hidden into a and b (lines 10–12). As the merging operation may result into a counter-intuitive representation of individual samples that are not disjoint in time, a function `reshapeTrajectory` is used fix the issue (line 11). As a last step, function `addRecord` enters m into the database, and, if the new generalized trajectory does not yet ensure an anonymity level k (line 14), S is updated by recomputing the trajectory stretch efforts between the new generalized trajectory and all other trajectories[2] in $\mathbb{M}$ via (10) again (lines 13–16). Once the main loop ends, a negligible number of less than k spurious trajectories may have to be removed as it does not meet the k-anonymity requirement.

We remark that the one presented above, and evaluated in the rest of the paper, is the *strict* version of GLOVE: in this case, the trajectories are freezed as soon as they reach the desired level k of anonymity, so that they are not used for further generalization. This has the advantage of preserving the level of detail of trajectories that are already k-anonymous, and ensures that all trajectories in the output database have a level of anonymity exactly equal to k. A different, *loose* variant does not enforce trajectory freezing, hence allows reusing trajectories that are already k-anonymous to hide others that are not yet so. The implementation is straightforward, as it is sufficient to remove the conditioning on $m.k < k$ (line 14) in Alg. 1. This has the advantage of not

[2] Recall that the expression in (10) can accommodate input trajectories that already generalize multiple original trajectories, see Sec. 4.

Algorithm 1: GLOVE algorithm pseudocode.

```
   input  : Anonymization level k
   input  : Mobile phone trajectory dataset 𝕄
   output : Anonymized phone trajectory dataset 𝕄
1  foreach a ∈ 𝕄 do
2      a.k = 1 ;
3      foreach b ∈ 𝕄, a ≠ b do
4          S [a,b] = trajectoryStretchEffort (a,b) ;
5      end
6  end
7  while ∃ a ∈ 𝕄 s.t. a.k < k and S ≠ ∅ do
8      a,b ← minimumStretchEffort(S) ;
9      removeRecords(𝕄,S,a,b) ;
10     m ← mergeTrajectories(a,b) ;
11     m ← reshapeTrajectory(m) ;
12     m.k = a.k + b.k ;
13     addRecord(𝕄,m) ;
14     if m.k < k then                                   // Comment to implement the loose variant
15         foreach c ∈ 𝕄 do
16             S [c,m]=trajectoryStretchEffort (c,m) ;
17         end
18     end
19 end
20 𝕄 ← 𝕄 \ {a ∈ 𝕄 s.t. a.k < k};
```



(a) Function `mergeTrajectories` (b) Function `reshapeTrajectory`

Fig. 10. (a) Example of operation of `mergeTrajectories`. (b) Example of operation of `reshapeTrajectory`.

leaving spurious trajectories out of the anonymized dataset, which may however contain records with uncontrolled different privacy levels k.

All functions in Alg. 1 are naive, except for `mergeTrajectories` and `reshapeTrajectory`, which we detail next.

6.2 Functions `mergeTrajectories` and `reshapeTrajectory`

Function `mergeTrajectories` returns one generalized mobile phone trajectory from two input trajectories (line 10). This operation is not immediate, and we propose a two-stage process to perform the non-uniform generalization.

The two stages are illustrated in Fig. 10a. In stage A, each sample in the longer mobile phone trajectory a is matched to that in the shorter trajectory b at minimum sample stretch effort, computed as in (1). Then, all samples in a pointing to a same sample in b (*e.g.*, s_1^a and s_2^a, pointing at s_2^b in the example) are merged with the latter into a single generalized sample. In stage B, the same procedure is run on samples of the shorter mobile phone trajectory that have not been merged during the stage A (*e.g.*, s_1^b in the example). These samples are matched with those resulting from

the first stage using (1) in a way that the sample stretch effort of each matching is minimum (*e.g.*, s_1^b is merged with the generalized sample s_1^a–s_2^a–s_2^b in the example).

At both stages, samples are merged through spatiotemporal generalization. Such an operation follows again the intuition in Fig. 4, *i.e.*, it is performed by stretching the spatial and temporal dimension of all the samples to be merged up to the point where they become indistinguishable. Formally, let us consider two generic samples, a's i-th and b's j-th, to be merged into a new sample, m's k-th. The generalization is realized as

$$\star_k^m = \min(\star_i^a, \star_j^b), \tag{12}$$

$$d\star_k^m = \max(\star_i^a + d\star_i^a, \star_j^b + d\star_j^b) - \star_k^m, \tag{13}$$

where $\star$ is to be replaced by x and y, or by t, in order to obtain the equations for spatial or temporal generalization, respectively. These operations simply stretch the new sample of m so that it covers the geographical areas and temporal intervals of both a's i-th sample and b's j-th sample. In case multiple samples must be merged together (*e.g.*, s_1^a, s_2^a, and s_2^b in the first stage of Fig. 10a) the operations can be run iteratively, merging one sample at a time: the result is invariant to the ordering of samples.

It is important to note that equations (12) and (13) realize our principle of non-uniform generalization, since: *(i)* the loss of granularity is the minimum required to hide each sample; *(ii)* the generalization is different among samples, breaking the dependency of all samples in a phone trajectory from the single sample that is the hardest to anonymize.

As anticipated, the merging operation may result into counter-intuitive representations of time, in cases where the minimum sample stretch effort is dominated by the spatial component. An example is provided in Fig. 10b: there, locations $\sigma_2^a = c_2$ and $\sigma_1^b = c_1$ are farther in space than instants $\tau_2^a = t_3$ and $\tau_2^b = t_5$ are in time. Thus, the merging of trajectories a and b leads to generalized samples of m that overlap in time, but refer to different geographical locations. The resulting generalized trajectory is formally correct, but it is difficult to read or analyze. Function `reshapeTrajectory` addresses this problem by means of a reshaping process that resolves temporal overlaps among samples, either partial or complete. The function first orders all samples in the trajectory according to their start time. It then scans the ordered list by looking for following samples with a start time lower than the end time of the current sample. For each identified overlap case, it creates a new sample that covers the overlapping time interval; the spatial granularity of the new sample spans the geographical areas of the samples it replaces, as per (12) and (13). The time span of the original samples is trimmed accordingly. Reshaping has a cost in terms of spatial granularity, since the new samples it generates span a larger geographical region that is the union of those in the overlapping original samples; however, it largely improves the usability of the anonymized data.

6.3 Computational complexity

Attaining optimal k-anonymity in movement micro-data databases is a NP-hard problem [58]. GLOVE takes a greedy approach that requires: *(i)* computing the trajectory stretch effort among all possible pairs of users in the original mobile traffic dataset; *(ii)* iteratively merging the two closest trajectories and recomputing the stretch effort between the merged trajectory and all those remaining in the dataset. By doing so, GLOVE inherently trades optimality for computational efficiency: the greedy selection of the two trajectories at minimum stretch effort may force at each iteration a generalization that is not part of the best solution possible. For instance, imagine the toy example of four trajectories, such that the stretch efforts are x for the pair (b, c), and $x + \epsilon$ (where ϵ is a very small value) for both (a, b) and (c, d); all other stretch efforts are much higher than x. Then, the optimal solution for $k = 2$ would be generalizing (a, b) and (c, d), for a total cost of $2(x + \epsilon)$. A greedy approach instead merges (b, c) first, barring the best solution, and determining

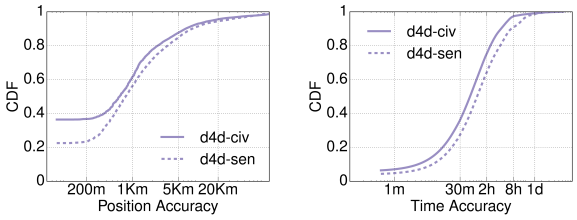


Fig. 11. Spatiotemporal accuracy in the d4d-civ and d4d-sen datasets, 2-anonymized with GLOVE.

an increased cost of the overall generalization. However, this is an unavoidable penality when scalability to very large trajectory dataset is a requirement.

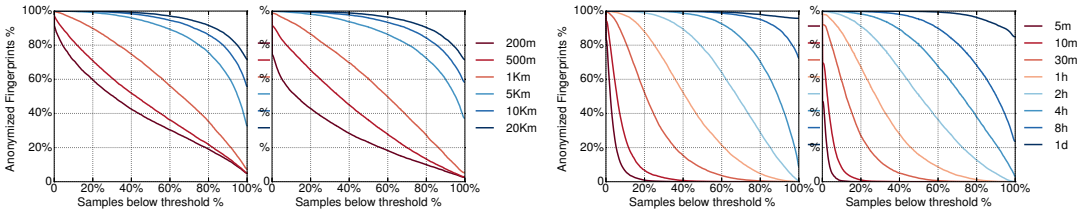
In order to determine the exact computational complexity of GLOVE, let us denote as $|\mathbb{M}|$ the number of mobile phone trajectories in the dataset, and as $\bar{n}$ their average length in samples. The complexity of operation *(i)* above maps to calculating (10), whose cost is $O\left(\bar{n}^2\right)$, for all $|\mathbb{M}|^2$ user pairs, and is thus $O\left(|\mathbb{M}|^2\bar{n}^2\right)$. The complexity of operation *(ii)* is the sum of two contributions. On the one hand, the merge has a cost $O\left(\bar{n}^2\right)$, and needs to be repeated k times (the desired k-anonymity level), for all users $|\mathbb{M}|$, leading to $O\left(k|\mathbb{M}|\bar{n}^2\right)$; note that this includes the $O\left(\bar{n}log(\bar{n})\right)$ cost of reshaping the $\bar{n}$ ordered samples, independently in each merged trajectory. On the other hand, the recalculation of the stretch efforts for the new generalized trajectory requires computing (10) against all other trajectories, which are $O\left(|\mathbb{M}|\right)$, with cost $O\left(|\mathbb{M}|\bar{n}^2\right)$. Overall, GLOVE runs in polynomial time: specifically, it is quadratic in both the number of trajectories and their length, since $k \ll |\mathbb{M}|$. We remark that $|\mathbb{M}|$ actually decreases as Algorithm 1 progress towards higher values of k, since trajectories are merged more and more, and their number is reduced: as a result, GLOVE run-times grow sublinearly with k.

A strategic aspect in the design of GLOVE is that it builds on expressions introduced in Sec. 4 that are highly parallelizable, as detailed in Sec. 4.3. This is also true for function `mergeTrajectories`, which repeats twice the operations needed for (10). The implementation used in this paper relies on the Nvidia CUDA architecture for GPU computing, with adequate mappings of the calculations in (10), (12) and (13). A non-optimized proof-of-concept version of the software executes the calculations in (10) on 20–50,000 phone trajectory pairs per second, using a single-GPU, low-end GeForce GT 740 card with 384 CUDA cores at 1 GHz. Again, we remark that indexing techniques may be a viable approach to improve such performance substantially, by, *e.g.*, cutting down the demanding calculation of all pairwise stretch efforts during the initialization of Alg. 1 (lines 3–5).

On this machine, the full d4d-civ and d4d-sen datasets were 2-anonymized with GLOVE in 60 hours each. We remark that dataset anonymization for PPDP is a one-time operation performed before data release. The latter typically occurs within months from the actual data collection, to allow for data cleansing and legal clearance. With weeks to complete the task, processing time is a much less relevant issue in PPDP than in other use cases. We also expect that performance can be sensibly improved by running an optimized version of the software on dedicated parallel architectures and high-end hardware.

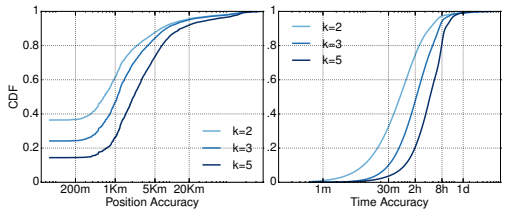
7 PERFORMANCE EVALUATION

GLOVE guarantees, by design, the k-anonymity of all mobile phone trajectories in a dataset. This is a result that uniform spatiotemporal generalization never achieves, as per Fig. 8. The capital question is, however, what is the cost paid in terms of data granularity reduction in order to attain k-anonymity. In the following, we investigate this aspect.



(a) d4d-civ, space (b) d4d-sen, space (c) d4d-civ, time (d) d4d-sen, time

Fig. 12. Percentage of mobile phone trajectories (y axis) with a given percentage of samples (x axis) below some spatial (a,b) and temporal (c,d) thresholds. Each curve refers to a threshold value, on the right of each plot. Plots are for the d4d-civ (a,c) and d4d-sen (b,d) datasets, and $k = 2$.



(a) d4d-civ, space (b) d4d-civ, time

Fig. 13. Accuracy of samples in k-anonymized d4d-civ dataset in space (a) and time (b), for k=2, 3, 5.

7.1 Accuracy analysis

Fig. 11 shows the accuracy of GLOVE-anonymized phone trajectory samples in the d4d-civ and d4d-sen datasets, for the baseline case of 2-anonymity. The two plots show CDFs of the spatial and temporal accuracy of the anonymized data, defined as the granularity of generalized samples in space and time, respectively. We observe that 20% to 40% of the samples retain a 100-m spatial accuracy, and have a temporal accuracy of 30 minutes or less. Even for larger fractions of samples, the loss of spatiotemporal granularity is tolerable: 70% to 80% of samples have an accuracy of 2 km or less in space, and of 2 hours or less in time. Recall that, at this granularity, no single trajectory was 2-anonymized with uniform generalization in Fig. 8.

Additional analyses are presented in Fig. 12. The four plots summarize the spatiotemporal accuracy of samples on a per-trajectory basis. Each curve refers to a threshold in space or time, and outlines how many trajectories in the dataset have a given percentage of samples with granularity finer than that threshold. These plots reveal that there exists heterogeneity in the data quality across the 2-anonymized trajectories: *e.g.*, some have all of their samples retaining an accuracy of 1 km and 2 hours or better, whereas others do not have a single sample with such a level of detail. Interestingly, this heterogeneity is higher in space than time: the curves in Fig. 12a and Fig. 12b cover more uniformly the abscissa than those in Fig. 12c and Fig. 12d. Hence, we can expect the anonymized trajectories to have especially diverse spatial accuracies. We impute this to the fact that our datasets capture environments that range from densely populated cities where many trajectories develop within the same neighborhood, to remote rural villages where the closest trajectories unfold at tens of km from each other.

Although 2-anonymity already satisfies the indistinguishability principle, higher privacy levels are possible, at a cost in terms of accuracy. The last plots in Fig. 13 detail the trade-off for the d4d-civ dataset. The percentage of samples with unvaried position accuracy drops to 25% for $k = 3$

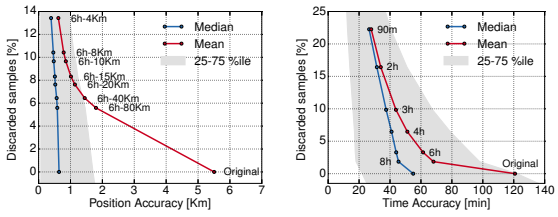


Fig. 14. Spatiotemporal accuracy in the d4d-civ dataset, 2-anonymized with GLOVE with suppression.

and 15% for $k = 5$; the percentage of samples with accuracy better than 2 km is 70% for $k = 3$ and 50% for $k = 5$. In time, 50% and 20% of samples feature a temporal accuracy better than 2 hours under $k = 3$ and $k = 5$, respectively. These figures point out that, depending on the usage of the anonymized data, the fraction of exploitable samples may be significantly reduced for $2 < k \leq 5$.

These values of k are already reasonable for data protection from a legal standpoint: for instance, the Italian code of conduct indicates $k = 3$ as the requisite to process personal data for statistical and scientific research purposes [59]. If a higher level of protection is required, then one may try to simplify the problem, by, *e.g.*, making assumptions about the attacker's knowledge. This would allow modifying GLOVE operation so as to target, *e.g.*, partial trajectory anonymization, which is less expensive than the full-length version we target in this work. Another option is allowing suppression, which we discuss next.

7.2 Augmenting GLOVE with suppression

Suppression allows discarding hard-to-anonymize samples from phone trajectories and is easily integrated in GLOVE. Indeed, non-uniform generalization can be combined with removal of samples whose temporal or spatial stretch efforts in (12) and (13) exceed some threshold. Fig. 14 shows the improvement of spatiotemporal accuracy (x axis) in the d4d-civ dataset when imposing different thresholds to the spatial and temporal stretch (tags along curves), which results in discarding some percentage of samples (y axis). Suppression can significantly improve the quality of the anonymized dataset. For instance, the average spatial accuracy shifts from more than 5 km to around 1 km when discarding less than 8% of samples, *i.e.*, by removing samples with a spatial stretch effort above 20 km, and whose temporal stretch effort[3] is above 6 h. Similarly, the average temporal accuracy is halved by suppressing just 4% of samples, *i.e.*, thresholding at 6 h. The median and 25^{th}–75^{th} percentile range are similarly improved by suppression.

Interestingly, as the percentage of removed samples increases, there is a diminishing returns effect on data accuracy. We conclude that minimal suppression of around 5% of samples represents a sweet spot in our reference datasets, as it significantly improves data quality by discarding a limited number of hard-to-anonymize outliers.

7.3 Comparative analysis

Several solutions were proposed for the k-anonymization of movement micro-data, as reviewed in Sec. 3. Among those, the only techniques that operate on both spatial and temporal dimensions are the Trajectory Generalization-based Approach (TGA) [43] and Wait for Me (W4M) [44]. We thus consider them as the state-of-the-art benchmarks for GLOVE.

TGA is close in spirit to GLOVE, as it relies on per-spatiotemporal point generalization. However, TGA yields several major differences when compared to our solution. First, TGA employs a different pairwise trajectory similarity metric, named log cost metric, which scales logarithmically the loss

[3]The results in Fig. 14 also consider a temporal stretch effort threshold, since spatial thresholding alone yielded very marginal gains in accuracy.

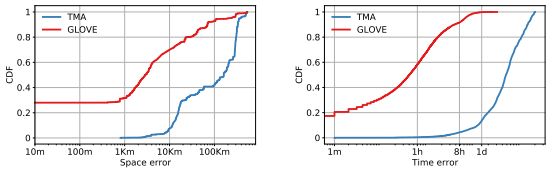


Fig. 15. Spatiotemporal accuracy in a random 1,000-trajectory sample of the d4d-civ dataset, 2-anonymized with GLOVE anf TGA.

of spatial and temporal accuracy, and also natively includes a penalization term for suppressing samples; when compared to our trajectory stretch effort, the metric is computationally expensive while its advantage in terms of anonymized data quality needs to be explored. Second, GLOVE greedily selects the pair of trajectories to be anonymized at each iteration according to the strategy described in Sec. 6.1; instead, TGA performs a random pick of the first trajectory, and then identifies the second one so as to minimize the log cost metric, which can clearly lead to a less optimal choice of trajectories from a global viewpoint. Third, TGA formulates the merge operation as a point matching problem, which entails that no two points of one trajectory can be generalized with a single point of the other trajectory; this leads to substantial suppression when the trajectories have a very different number of points, which is not the case with the more flexible approach of GLOVE.

We compare GLOVE and TGA on a reduced dataset of 1,000 trajectories randomly sampled from the original d4d-civ dataset. The data reduction is unavoidable and linked with the first insight of our comparative evaluation: the time required to compute a single log cost metric used by TGA is around 30 times (9.29 versus 0.32 seconds on average in our setup) higher than that needed for the trajectory stretch effort defined in (10). Such an expensive metric severely limits the scalability of TGA, preventing us from running larger or additional experiments.

Fig. 15 summarizes the results in terms of spatial and temporal resolution of the small-scale dataset upon 2-anonymization with GLOVE and TGA. Data accuracy is clearly better with our solution, with gains of one order of magnitude or more in both space and time. We ascribe the advantage of GLOVE to (*i*) the greedy choice of trajectories, as the random selection of TGA is especially detrimental in small datasets, and (*ii*) the higher freedom in selecting samples to be merged granted by the trajectory stretch effort used in the `mergeTrajectories` function in Alg. 1.

The other relevant term of comparison is W4M, which builds on the concept of uncertain trajectory, *i.e.*, a cylindrical volume that has a diameter δ in space and stretches along time. W4M groups similar trajectories into clusters of at least k elements each, and then performs the minimum spatiotemporal translation needed to push all the trajectories of a cluster within the same uncertain trajectory. An important remark is that W4M allows both the suppression and the creation of new synthetic samples. The latter operation improves the matching among trajectories in a cluster, and assumes that mobile objects (*i.e.*, phones in our case) perform linear constant-speed movements between samples. We use W4M with linear spatiotemporal distance and chunking (LC), *i.e.*, the version intended for large databases such as ours.

We run W4M-LC on the d4d-civ and d4d-sen datasets, using the suggested settings of $\delta = 2$ km and a 10% trashing, which enables entirely removing trajectories that are especially difficult to cluster [44]. Tab. 2 presents the results for $k = 2$ and $k = 5$, confronted to those achieved by GLOVE with suppression via thresholds set at 6 hours and 15 km.

Differences are significant. In all scenarios, W4M-LC creates a substantial amount of synthetic samples, tallying 17% to 45% of the original data. Such samples do not correspond to actual user movements, and thus violate the PPDP principle P3 of truthfulness at the record level. Even worse, fabricated samples do not help in attaining a sufficient level of accuracy in the anonymized data:

Table 2. Features of d4d-civ and d4d-sen k-anonymized with W4M-LC [44] and GLOVE, for k=2, 5.

		d4d-civ		d4d-sen	
		W4M-LC	GLOVE	W4M-LC	GLOVE
$k=2$	Discarded phone trajectories	1,104 (1.3%)	0	430 (0.1%)	0
	Created samples (×1000)	4,444 (24.9%)	0	5,302 (17.9%)	0
	Deleted samples (×1000)	1,325 (7.5%)	1,482 (8.3%)	1,577 (5.3%)	4,175 (14.1%)
	Mean position error [m]	10,190.88	1,013.71	9,392.85	1,312.28
	Mean time error [min]	1,151.51	60.21	1,037.74	69.31
$k=5$	Discarded phone trajectories	1,271 (1.5%)	0	3740 (1.2%)	0
	Created samples (×1000)	8,018 (44.9%)	0	8,863 (29.9%)	0
	Deleted samples (×1000)	1,713 (9.6%)	1,482 (8.3%)	2,179 (7.3%)	5,004 (16.9%)
	Mean position error [m]	23,534.062	5,129.9	19,881.9	5,694.2
	Mean time error [min]	3,455.94	171.01	2.600.64	408.30

the mean error introduced by the perturbations in W4M-LC roughly ranges between 10 km ($k = 2$) and 20 km ($k = 5$) in space, and between 16 hours ($k = 2$) and more than two days ($k = 5$) in time. The result is hardly exploitable for data analysis purposes. GLOVE yields a much higher average precision, at around 1 km and 1 hour in all $k = 2$ cases, and around 5 km and 3 hours when $k = 5$. These figures are obtained at an affordable cost (8% to 17%) in terms of sample suppression.

We believe that the poor result of W4M-LC is due to the nature of the target data: W4M-LC was originally designed for the anonymization of trajectories sampled at a frequency that is high, regular, and similar for all moving objects. This is the case of, *e.g.*, the GPS logs considered in the initial performance evaluation of W4M-LC [44]. Instead, as discussed in Sec. 2.1, the mobile phone trajectories our work focuses on are characterized by irregular and heterogeneous sampling, and are typically sparse in time. In this context, a dedicated solution such as GLOVE grants superior performance.

8 DATA FEATURES AND k-ANONYMITY

Our evaluation results are clearly specific to the reference data we use. However, in this section we carry out additional analyses that let us speculate about how GLOVE performance would generalize to other mobile phone trajectory datasets. Again, our main performance metric is the spatial and temporal accuracy of the anonymized data, *i.e.*, the granularity of generalized samples in space and time, respectively.

8.1 Dataset features

We first investigate how the macroscopic features that characterize the trajectory dataset as a whole affect the accuracy of its k-anonymized version.

Temporal span. We extract from the original d4d-civ and d4d-sen shorter datasets spanning one day to one week. Fig. 16 shows how the dataset timespan affects the spatiotemporal accuracy after 2-anonymization with GLOVE. We observe that shorter datasets yield a higher accuracy,

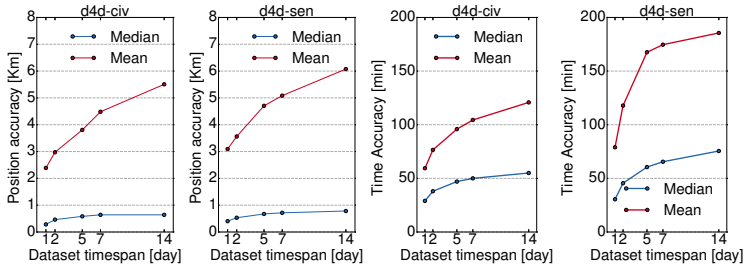


Fig. 16. Spatiotemporal accuracy versus the temporal span of the d4d-civ and d4d-sen (sub-)datasets, 2-anonymized with GLOVE.

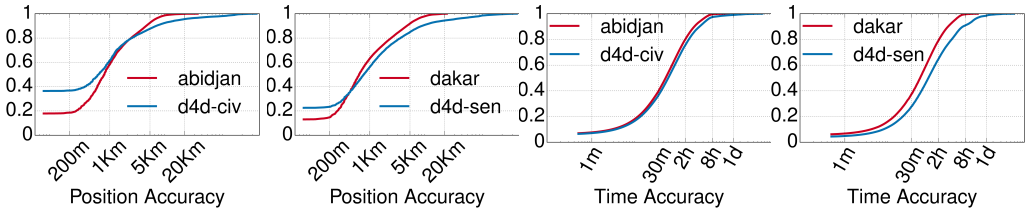


Fig. 17. Spatiotemporal accuracy in citywide abidjan and dakar datasets, compared to nationwide ones, after 2-anonymization with GLOVE.

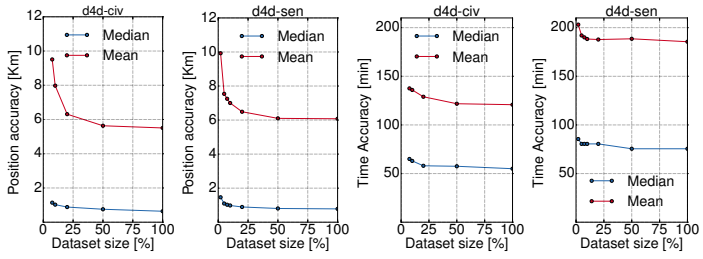


Fig. 18. Spatiotemporal accuracy versus the cardinality of the d4d-civ and d4d-sen (sub-)datasets, 2-anonymized with GLOVE.

both in space and time, once they have been anonymized. This is not surprising, since a lower dataset timespan reduces the length of mobile phone trajectories, which then become easier to hide into each other. The gain in accuracy can be very high, as 2-anonymized 1-day datasets are twice as precise than 2-week ones. Interestingly, the loss of accuracy grows sub-linearly with the dataset duration – especially in the median case, *i.e.*, for typical trajectories. We hypothesize that the weekly periodicity of human activities [60] makes datasets spanning multiple weeks not much harder to anonymize than one-week ones.

Geographical coverage. To assess the impact of the spatial extension of datasets, we extract data subsets that are geographically limited to the major cities in Ivory Coast and Senegal, *i.e.*, abidjan and dakar. The distributions in Fig. 17 show that their anonymization with GLOVE yields a spatiotemporal accuracy similar to that obtained in the nationwide datasets. This affinity is explained by the locality of human activities: the median and average radius of gyration of the trajectories are 1.8 km and 12 km in d4d-civ, and 2 km and 10 km in d4d-sen. Thus, most mobile phone trajectories are confined to a limited geographical region whose size is approximately that of a city: it is then very likely that trajectories are hidden with others unfolding in the same urban area. This makes the citywide or nationwide dataset anonymization costs comparable.

Cardinality. An interesting question is to what extent the number of trajectories in a dataset affects its anonymization. Fig. 18 shows how the spatiotemporal accuracy varies when considering datasets that comprise from 5% to 100% of the original records, *i.e.*, mobile phone trajectories, in d4d-civ and d4d-sen. We observe that datasets with a lower number of trajectories tend to be harder to anonymize, since the crowd that one can leverage to hide a trajectory becomes thinner. However, the effect is only remarkable when retaining a rather low fraction of the original trajectories. Specifically, the k-anonymized data precision seems impaired only when the number of users falls below a few tens of thousands.

8.2 Trajectory features

The features that characterize individual trajectories offer a different, complementary viewpoint on the accuracy of k-anonymized data, and we consider them next.

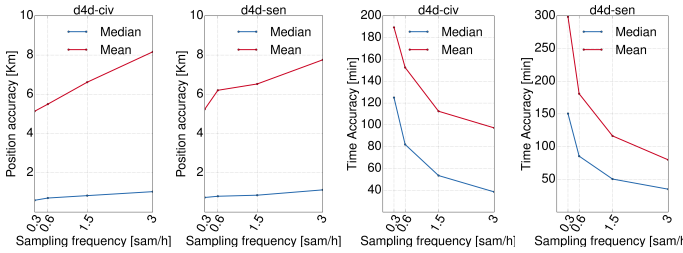
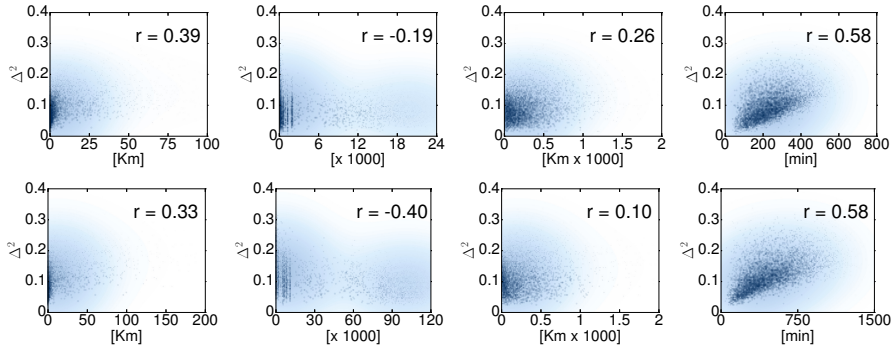


Fig. 19. Spatiotemporal accuracy versus the trajectory sampling frequency in d4d-civ and d4d-sen, 2-anonymized with GLOVE.



(a) RadGyr (b) LocDens (c) TravDist (d) StdDev

Fig. 20. Correlation between properties of individual mobile phone trajectories and the stretch effort needed to merge them with their GLOVE-selected pair. Plots refer to the d4d-civ (top) and d4d-sen (bottom) datasets, and $k = 2$.

Sampling frequency. The most prominent trait characterizing a mobile phone trajectory is its sampling frequency. Fig. 19 shows how this feature affects the spatiotemporal accuracy of 2-anonymized trajectories. Trends are very different in the spatial and temporal dimensions. In space, the sampling frequency has a marginal impact on typical cases, as shown by the quasi-constant median. In time, denser trajectories become much easier to hide: we observe a threefold improvement in resolution when trajectories are sampled from 0.3 to 3 times per hour, with the median accuracy growing from 120 to 40 minutes.

Overall, an increased sampling frequency renders mobile phone trajectories easier to k-anonymize. Although it introduces more samples, and thus increases the sheer number of merge operations, a higher frequency reduces the time difference between samples, which, as discussed in Sec. 5.5, is the main obstacle to trajectory anonymization.

Other properties. We also explore whether other properties of individual trajectories have a significant impact on the anonymized data resolution. We test the following properties: *(i)* the radius of gyration (RadGyr), which measures the level of mobility of a trajectory[4]; *(ii)* the local density (LocDens), *i.e.*, the number of trajectories with a radius of gyration overlapping with that of the target trajectory; *(iii)* the total travel distance of the trajectory (TravDist); *(iv)* the standard deviation of inter-sample time (StdDev), *i.e.*, the uniformity of trajectory samples in time.

[4]The radius of gyration r_g is a unidimensional measure of the distance covered by a trajectory that captures movement directionality. It is computed as $r_g = \sqrt{1/n \sum_i (r_i - 1/n \sum_i r_i)^2}$, where $r_i, i \in [1, n]$ is a bi-dimensional vector describing the location in the i-th sample of the trajectory.

Fig. 20 summarizes our results. The scatterplots depict the correlation between each property above and the trajectory stretch effort (used here as a scalar proxy of the combination of spatial and temporal accuracy) needed to 2-anonymize the trajectory with GLOVE. The exact correlation coefficient is reported as a number in each plot.

Overall, correlations with spatial properties have expected trends (*e.g.*, positive for RadGyr or negative for LocDens), but are weak in all cases: once more, the spatial dimension of a trajectory is not prevalent in determining its anonymizability. Instead, StdDev shows a fairly strong positive correlation, which means that trajectories whose samples are irregularly distributed in time are more difficult to hide. This result complements that on the sampling frequency above: ultimately, a mobile phone trajectory is easier to anonymize if it has a large number of samples that are uniformly spaced in time.

8.3 Summary

The analyses above let us identify which data characteristics affect the k-anonymization process to a larger extent than previously done in the literature. We conclude that datasets containing a large number of high-frequency and regularly sampled trajectories are the most suitable candidates for low-cost anonymization, from a data accuracy viewpoint. The spatial and temporal span of the dataset are less decisive, as long as the data covers at least a city for one week – a very common setting in the literature [2]. We thus speculate that k values higher than the limit we identified in `d4d-civ` and `d4d-sen`, *i.e.*, $k = 5$, may be sustainable when using GLOVE in datasets with a larger cardinality or denser trajectories.

9 UTILITY OF k-ANONYMIZED DATASETS

To conclude our performance evaluation, we assess whether the level of accuracy of datasets that are generated by GLOVE is sufficient for a dependable analysis of mobile phone trajectories. In other words, we investigate if the published trajectory data remains useful once k-anonymized with our proposed solution.

9.1 Results in the literature

A first set of encouraging observations comes from the literature. Concerning the data utility for spatial analyses, a study dedicated to trajectory data indicates that an accuracy of 2-7 km in the published samples allows to correctly model the majority of people movements [61]. The spatial granularity of samples in our k-anonymized datasets is well within such thresholds for k up to 5, as demonstrated for instance by Fig. 12 and Fig. 14.

As far as the temporal resolution is concerned, a sampling interval of 30 minutes is deemed sufficient for a proper characterization of human mobility [62]. Unfortunately, our reference datasets often feature sampling intervals longer than that threshold even in their original form, as shown in Tab. 1. This prevents a conclusive analysis, however a promising remark is that the k-anonymized trajectory samples have temporal stretches that are typically close to the 30-minute threshold identified above, especially when combining GLOVE and suppression as shown in Fig. 14. Moreover, the results in Sec. 8.2 suggest that such a temporal resolution of the anonymized data would improve in the case of input datasets with a more regular and frequent sampling. Based on these figures, it seems a reasonable speculation that the loss of granularity induced by GLOVE on k-anonymized high-utility mobile phone trajectory data would not represent an impairment to temporal analyses.

9.2 Experiments with basic mobility analysis

To further assess the utility of GLOVE-anonymized data, we run a set of classical data mining tasks on the original `d4d-civ` and `d4d-sen` datasets, and on their k-anonymized versions, and quantify the differences. In the case of the original data, we consider that spatiotemporal samples have 100-m

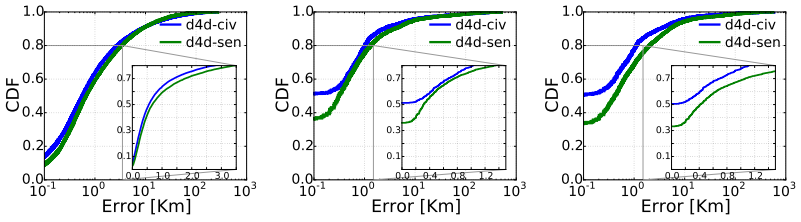


Fig. 21. Estimation error of center of mass (left), home (center) and work (right) locations with 2-anonymized d4d-civ and d4d-sen datasets.

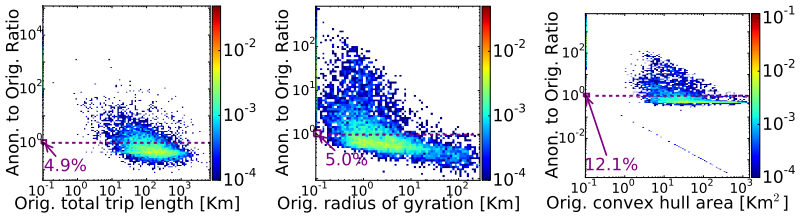


Fig. 22. Estimation error of total travel distance (left), radius of gyration (center), and convex hull (right) with 2-anonymized d4d-civ dataset.

and 1-minute spatial and temporal accuracy, respectively[5]. We remark that several tasks require that each spatiotemporal sample denotes a single location and a precise time instant: in these cases, when computing mobility metrics, we approximate the geographical location of a generalized sample $\sigma_i^a = (x_i^a, dx_i^a, y_i^a, dy_i^a)$ as the center of the area it spans, *i.e.*, $(x_i^a + dx_i^a/2, y_i^a + dy_i^a/2)$; similarly, the occurrence time of σ_i^a is mapped to the intermediate instant $t_i^a + dt_i^a/2$.

Center of mass. The center of mass denotes the pivotal location of a trajectory in space[6], and is often regarded as a simple but significant feature to summarize the region of movement of a user. The left plot in Fig. 21 shows the CDF of the error incurred when the center of mass is computed from 2-anonymized data with respect to the case where the same metric is derived from the original data. The error is below 0.5 km in around 50% of cases, and under 3 km in 80% of trajectories. The vast majority of centers of mass is thus located correctly, when considering the spatial resolution thresholds indicated in [61].

Home and work locations. Important places that are frequently visited by each individual were one of the very first targets for mobile phone data mining, and the inferred home and work locations represent an input for many subsequent analyses. Following a classical approach, we compute the home and work locations of each individual in a dataset as the most popular locations within the user's trajectory overnight (10 pm to 6 am) and during working hours (9 am to 5 pm), respectively. The center and right plots in Fig. 21 portray the CDF of the error induced by 2-anonymized data when inferring such important locations. Depending on the dataset, 35% to 50% of the home and work locations are unaffected by the anonymization. At least 70% of these locations are placed within 1 km of their original position, and the 7 km threshold of [61] is met by over 90% of the trajectories.

[5]We stress that 100 m is an artificial, higher spatial resolution than that of the original data, which is introduced during a preprocessing phase as described in Sec. 5.1. Therefore, the spatial errors computed in the remainder of this section shall be regarded as upper bounds.

[6]The center of mass c of a trajectory is computed $c = \frac{1}{n}\sum_i r_i$, where $r_i, i \in [1, n]$ is a bi-dimensional vector describing the location in the i-th sample of the trajectory.

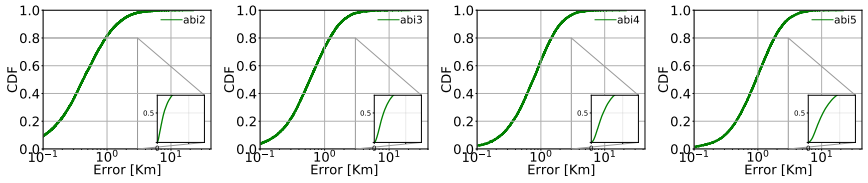


Fig. 23. Estimation error of center of mass with k-anonymized abidjan dataset, and $2 \leq k \leq 5$.

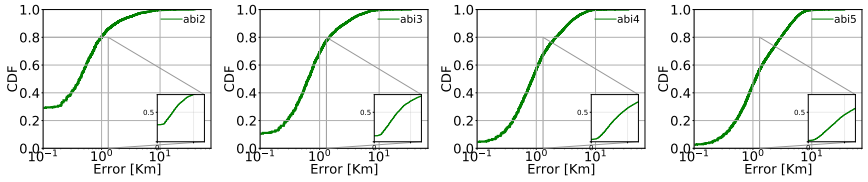


Fig. 24. Estimation error of home locations with k-anonymized abidjan dataset, and $2 \leq k \leq 5$.

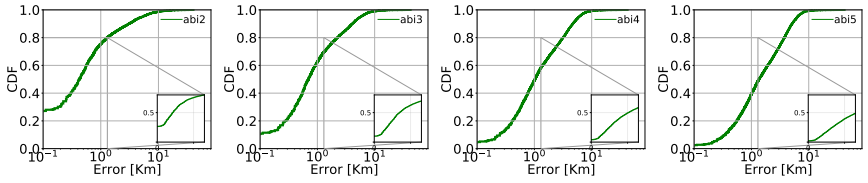


Fig. 25. Estimation error of work locations with k-anonymized abidjan dataset, and $2 \leq k \leq 5$.

Total travel distance. The total distance travelled by a user is another figure of merit often employed to characterize human mobility. The error in estimating the total length of each trajectory from GLOVE-processed data is shown in the left plot in Fig. 22, as the ratio between the values obtained from the anonymized and original data. The abscissa represents the original total travel distance. Around 5% of the trajectories are completely static, *i.e.*, are generated by subscribers who never move to a different cell, which can happen in rural regions where a single base station covers several villages; in these cases, the anonymized data retains the original precision, and the ratio is one. However, the vast majority of trajectories are not static, and we observe a clear trend of increasing accuracy (*i.e.*, the ratio getting closer to one on average) as the original travel distance increases. The heatmap is the densest for original travel distances in the range 30–800 km, where the original and the anonymized datasets match well.

Radius of gyration. As explained before, this metric quantifies the total bidimensional mobility of a trajectory. Results are in the center plot in Fig. 22, which is semantically identical to the total travel distance one. The trends are equivalent to those discussed for the previous measure, with 5% static users with perfect estimation and a decreasing error for higher-mobility trajectories. In this case, most trajectories feature a radius of gyration between 500 m and 20 km, where the error ratio is again close to one.

Convex hull. The convex hull is the smallest convex envelope that includes all spatial samples of a trajectory, and conveys information about the overall surface covered by an individual during the mobile phone dataset timespan. The right plot in Fig. 22 portrays the error ratio in this case. The percentage of correct, static trajectories jumps to 12%, since trajectory including two locations still have a null convex hull. The vast majority of non-static trajectories has a ratio close to the desired value of one.

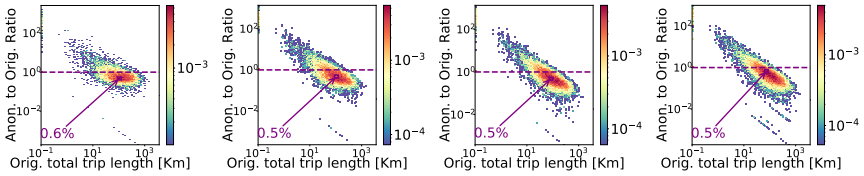


Fig. 26. Estimation error of total travel distance with k-anonymized abidjan dataset, and $2 \leq k \leq 5$.

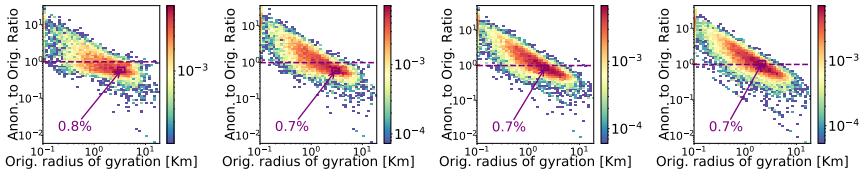


Fig. 27. Estimation error of radius of gyration with k-anonymized abidjan dataset, and $2 \leq k \leq 5$.

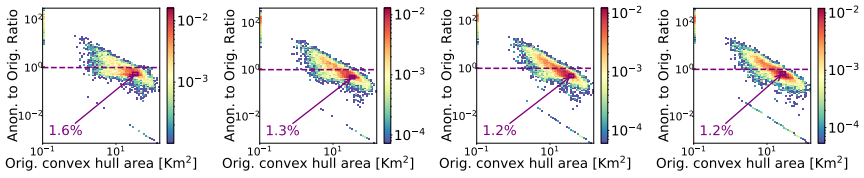


Fig. 28. Estimation error of convex hull with k-anonymized abidjan dataset, and $2 \leq k \leq 5$.

9.3 Impact of k

We discussed in Sec. 7 that increasing the value of k causes a sensible reduction in the accuracy of the data anonymized with GLOVE. A relevant question is whether such a loss of granularity affects in a significant manner the utility of the anonymized data for subsequent analyses. In order to address this aspect, we repeat the previous approach of evaluating classical geospatial data mining tasks in presence of k-anonymized datasets, with k ranging between 2 and 5. For these tests, we selected the citywide abidjan dataset, which we proved to yield equivalent properties to the national datasets in Sec. 8.1.

Fig. 23, Fig. 24, and Fig. 25 portray the results for the center of mass, home locations, and work locations, respectively. In all cases, the plots show the CDF of the error incurred when computing each metric on the k-anonymized individual trajectories, with respect to the same metric inferred from the original data.

In the case of the center of mass, we observe that increasing k only slightly worsen the error statistics. For instance, 80% of the trajectories have an error below 1 Km with $k = 2$, and that percentage is reduced to 60% when $k = 5$. the inset plots allow appreciating the limited changes in the CDF determined by k. The calculation of the center of mass of each individual appears thus very robust to the anonymization process.

Similar considerations hold for both home and work locations, where, however, the loss of data utility is more apparent. In particular, increasing k affects the fraction of individuals whose home and work locations are perfectly identified with the anonymized data: the value drops from 30% to 5% when k grows from 2 to 5. Also the percentage of home and work locations identified with an error below 1 Km falls from 80% to 45%.

Fig. 26, Fig. 27, and Fig. 28 illustrate instead the impact of higher privacy levels on the precision of estimates of the total travel distance, the radius of gyration, and the convex hull of individual trajectories, respectively. The plots adopt the same format used before for these measures, with the

value computed from the original data on the abscissa, and the error due to the anonymization process on the ordinate. In all figures, the plots refer to a value of k growing from 2 to 5 when moving from left to right.

The observed behavior is comparable across all measures: a higher k increases the estimation error. In particular an increase in k tends to determine a more evident overestimation of short travel distances, or small radiuses of gyration and convex hulls, as well as a more pronounced underestimation of long travel distances, or large radiuses and hulls. Still, the difference is not dramatic, and the overall shape of error clouds stays farily comparable across k values in all cases.

9.4 Summary

Overall, we conclude that the GLOVE-anonymized data allows running legacy mobile phone data mining tasks with results that retain substantial accuracy. We remark that these tasks analyse the mobility of users individually, hence only make sense if each trajectory actually corresponds to a real-world subscriber (*i.e.*, each value or point in Fig. 21–22 maps to one actual person). The fact that GLOVE builds on a PPDP model that is truthful at the record level lets it meet such a critical principle. In addition, the results for different values of k show that, while there exists an inherent trade-off between the privacy level and data utility, the trajectory data anonymized with GLOVE allows running basic geospatial analyses with an accuracy that is still acceptable when $k = 5$.

We also stress that the figures reported in our data utility analysis must be taken with a pinch of salt, as follows. First, the highest errors are only measured in presence of low-mobility trajectories, which are often useless in studies of human mobility and can be pruned from the original dataset before anonymization. Second, errors refer to datasets collected in developing countries, where users have limited mobile phone activity. As discussed in Sec. 7.1, subscribers generating sparse events contribute larger errors, and the accuracy of the anonymized data would likely improve in developed countries that are often the focus of fine-grained mobility analyses.

10 CONCLUSIONS

We presented GLOVE, an algorithm to k-anonymize movement micro-data from mobile phone trajectories. Its design builds on novel insights into the nature of mobile phone trajectory anonymizability. Specifically, we revealed that these trajectories are composed of a vast majority of spatiotemporal samples that are similar to those of other users, and are thus easily hidden via k-anonymity. However, mobile phone trajectories also typically feature a non-negligible number of highly unique spatiotemporal samples that can only be anonymized with severe generalization.

GLOVE takes advantage of this situation by adopting a non-uniform generalization where the loss of resolution is adapted on a per-sample basis. Thanks to this approach, GLOVE attains complete k-anonymization of real-world datasets while preserving a level of accuracy that would not grant partial 2-anonymization under legacy uniform generalization. Also, GLOVE outperforms existing solutions for k-anonymization of spatiotemporal trajectories, setting a new reference for the current state-of-the-art in anonymization of spatiotemporal trajectories from mobile phone data that are truthful at the record level.

This notwithstanding, GLOVE has limitations that pave the way for future research. First, the privacy model behind the design of GLOVE is effective in countering a subset of the possible attacks on mobile phone trajectories: it thus represents a first step towards complete PPDP, and not a definitive solution. Second, GLOVE should be regarded as a proof-of-concept, demonstrating the feasibility of the approach on fairly small-sized datasets, for low k values. Optimizations in the algorithm design and implementation are needed to improve its scalability, in terms of computational efficiency and anonymized data quality as k grows. Third, tests with additional mobile phone data are needed to generalize the results. Notably, the reference dataset used in

this study were collected in developing countries, whereas part of our analysis suggests that the performance of GLOVE may improve significantly in, *e.g.*, larger datasets collected in urban areas of developed countries.

As a concluding disclaimer, we would like to recall that this work addressed the problem of unicity in mobile users' trajectories. We note that unicity does not imply the actual re-identification of mobile users, and we do not try to de-anonymize any subscriber in the datasets we study.

REFERENCES

[1] V. Blondel, A. Decuyper, G. Krings, *"A survey of results on mobile phone datasets analysis,"* EPJ Data Science, 4(1), 2015.

[2] D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, *"Large-scale Mobile Traffic Analysis: a Survey,"* IEEE Communications Surveys and Tutorials, 18(1), 2016.

[3] L. Sweeney, *"k-anonymity: A model for protecting privacy,"* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002.

[4] A. Narayanan, V. Shmatikov, *"Robust de-anonymization of large sparse datasets,"* IEEE SP, 2008.

[5] H. Zang, J. Bolot, *"Anonymization of location data does not work: A large-scale measurement study,"* ACM MobiCom, 2011.

[6] Y. de Montjoye, C.A. Hidalgo, M. Verleysen, V. Blondel, *"Unique in the Crowd: The privacy bounds of human mobility,"* Nature Scientific Reports, 3(1376), 2013.

[7] A. Cecaj, M. Mamei, N. Bicocchi, *"Re-identification of Anonymized CDR datasets Using Social Network Data,"* IEEE PerCom Workshops, 2014.

[8] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, S. Lattanzi, *"Linking Users Across Domains with Location Data: Theory and Validation,"* ACM WWW, 2016.

[9] D. Kondor, B. Hashemian, Y.-A. de Montjoye, C. Ratti, *"Towards matching user mobility traces in large-scale datasets,"* arXiv:1709.05772 [cs.SI], 2017.

[10] B.C.M. Fung, K. Wang, R. Chen, P.S. Yu, *"Privacy-preserving data publishing: A survey of recent developments,"* ACM Computing Surveys, 42(4):14, 2010.

[11] R. Agrawal, R. Srikant, *"Privacy-preserving data mining,"* SIGMOD Record, 29(2):439–450, 2000.

[12] Y. Song, D. Dahlmeier, S. Bressan, *"Not so unique in the crowd: A simple and effective algorithm for anonymizing location data,"* PIR, 2014.

[13] J. Salas, D. Megías, V. Torra, *"Swapmob: Swapping trajectories for mobility anonymization,"* J. Domingo-Ferrer, F. Montes (editors), Privacy in Statistical Databases, 331–346, Springer International Publishing, 2018.

[14] R. Chen, B.C.M. Fung, B.C. Desai, N.M. Sossou, *"Differentially private transit data publication: A case study on the Montreal transportation system,"* ACM KDD, 2012.

[15] D.J. Mir, S. Isaacman, R. Cáceres, M. Martonosi, R.N. Wright, *"Dp-where: Differentially private modeling of human mobility,"* IEEE Big Data, 2013.

[16] A. Tockar, *"Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset,"* Technical Report, Neustar Research, Sep. 2014.

[17] R. Trujillo-Rasua, J. Domingo-Ferrer, *"On the privacy offered by (k,δ)-anonymity,"* Information Systems, 38:491–494, 2013.

[18] F. Bonchi, L.V.S. Lakshmanan, H. Wang, *"Trajectory anonymity in publishing personal mobility data,"* SIGKDD Explorations Newsletter, 13(1):30–42, 2011.

[19] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, *"l-diversity: Privacy beyond k-anonymity,"* ACM Trans. Knowledge Discovery from Data, 1(1):3, 2007.

[20] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, J.-P. Hubaux, *"Quantifying Location Privacy,"* IEEE SP, 2011.

[21] N. Li, T. Li, S. Venkatasubramanian, *"t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,"* IEEE ICDE, 2007.

[22] M. Gramaglia, M. Fiore, A. Tarable, A. Banchs, *"Preserving Mobile Subscriber Privacy in Open Datasets of Spatiotemporal Trajectories,"* IEEE INFOCOM, 2017

[23] C. Dwork, *"Differential privacy,"* ICALP, 2006.

[24] M. Hay, G. Miklau, D. Jensen, D. Towsley, P. Weis, *"Resisting structural re-identification in anonymized social networks,"* VLDB Endownment, 1(1), 2008.

[25] L. Sweeney, *"Practical Differentially Private Modeling of Human Movement Data,"* IFIP DBSec, 2016.

[26] M. Gruteser, D. Grunwald, *"Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking,"* ACM MobiSys, 2003.

[27] H. Kido, Y. Yanagisawa, T. Satoh, *"Protection of Location Privacy using Dummies for Location-based Services,"* IEEE ICDE, 2005.

[28] B. Gedik, L. Liu, *"Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms,"* IEEE Trans. Mobile Computing 7(1):1–18, 2008.
[29] M. Herrmann, A. Rial, C. Diaz, B. Preneel, *"Practical privacy-preserving location-sharing based services with aggregate statistics,"* ACM WiSec, 2014.
[30] M.E. Andrés, N.E. Bordenabe, K. Chatzikokolakis, C. Palamidessi, *"Geo-indistinguishability: differential privacy for location-based systems,"* ACM CCS, 2013.
[31] R. Assam, M. Hassani, T. Seidl, *"Differential private trajectory protection of moving objects,"* ACM IWGS, 2012.
[32] N.E. Bordenabe, K. Chatzikokolakis, C. Palamidessi, *"Optimal geo-indistinguishable mechanisms for location privacy,"* ACM CCS, 2014.
[33] Y. Xiao, L. Xiong, *"Protecting locations with differential privacy under temporal correlations,"* ACM CCS, 2015.
[34] J.C. Duchi, M.I. Jordan, M.J. Wainwright, "Local Privacy and Statistical Minimax Rates," IEEE FOCS, 2013.
[35] A.R. Beresford, F. Stajano, *"Mix Zones: User Privacy in Location-aware Services,"* IEEE PerCom, 2004.
[36] B. Hoh, M. Gruteser, H. Xiong, A. Alrabady, *Preserving privacy in GPS traces via uncertainty-aware path cloaking.* ACM CSS, 2007.
[37] J. Meyerowitz, R.R. Choudhury, *"Hiding stars with fireworks: location privacy through camouflage,"* ACM MobiCom, 2009.
[38] A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, S. Wrobel *"Movement Data Anonymity through Generalization,"* Trans. Data Privacy 3(2):91–121, 2010.
[39] O. Abul, F. Bonchi, M. Nanni, *"Never walk alone: Uncertainty for anonymity in moving objects databases,"* IEEE ICDE, 2008.
[40] J. Domingo-Ferrer, R. Trujillo-Rasúa *"Microaggregation- and permutation-based anonymization of movement data,"* Information Science, 208:55–80, 2012.
[41] B.C.M. Fung, M. Cao, B.C. Desai, H. Xu, *"Privacy protection for RFID data,"* ACM SAC, 2009.
[42] R. Yarovoy, F. Bonchi, L.V.S. Lakshmanan, W.H. Wang, *"Anonymizing moving objects: how to hide a mob in a crowd?,"* ACM EDBT, 2009.
[43] M.E. Nergiz, M. Atzori, Y. Saygin, B. Güç *"Towards Trajectory Anonymization: a Generalization-Based Approach,"* Trans. Data Privacy 2(1):47–75, 2009.
[44] O. Abul, F. Bonchi, M. Nanni, *"Anonymization of moving objects databases by clustering and perturbation,"* Information Systems, 35(8):884–910, 2010.
[45] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, T. Yu, *"Differentially private spatial decompositions,"* IEEE ICDE, 2012.
[46] G. Acs, C. Castelluccia, *"A case study: privacy preserving release of spatio-temporal density in Paris,"* ACM KDD, 2014.
[47] M. Alaggan, S. Gambs, S. Matwin, M. Tuhin, *"Sanitization of call detail records via differentially-private bloom filters,"* IFIP DBSec, 2015.
[48] S. Brunet, S. Canard, S. Gambs, B. Olivier, *"Novel differentially private mechanisms for graphs,"* IACR Cryptology, 2016:745, 2016.
[49] M. Hay, A. Machanavajjhala, G. Miklau, Y. Chen, D. Zhang, *"Principled evaluation of differentially private algorithms using dpbench,"* ACM SIGMOD, 2016.
[50] D. Shao, K. Jiang, T. Kister, S. Bressan, K.-L. Tan, *"Publishing trajectory with differential privacy: A priori vs. a posteriori sampling mechanisms,"* DEXA, 2013.
[51] J. Zhang, X. Xiao, X. Xie, *"Privtree: A differentially private algorithm for hierarchical decompositions,"* ACM SIGMOD, 2016.
[52] X. He, G. Cormode, A. Machanavajjhala, C.M. Procopiuc, D. Srivastava, *"Dpt: Differentially private trajectory synthesis using hierarchical reference systems,"* Proc. VLDB Endow. 8(11), 2015.
[53] R. Chen, G. Acs, C. Castelluccia, *"Differentially private sequential data publication via variable-length n-grams,"* ACM CCS, 2012.
[54] M.E. Gursoy, L. Liu, S. Truex, L. Yu, *"Differentially private and utility preserving publication of trajectory data,"* IEEE Transactions on Mobile Computing 18(10), 2018.
[55] V.T. de Almeida, R.H. Güting, *"Indexing the Trajectories of Moving Objects in Networks,"* Geoinformatica 9, 2005.
[56] V.D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, C. Ziemlicki, *"Data for Development: the D4D Challenge on Mobile Phone Data,"* arXiv:1210.0137 [cs.CY].
[57] D. Hoaglin, F. Mosteller, J.W. Tukey *"Understanding robust and exploratory data analysis,"* Wiley, 1983.
[58] C. Bettini, X.S. Wang, S. Jajodia, *"Protecting Privacy Against Location-Based Personal Identification,"* SDM, 2005.
[59] Code of conduct applying to the processing of personal data for statistical and scientific research purposes within the framework of the national statistical system. Article 5 – Criteria to Assess the Identification Risk. `http://www.garanteprivacy.it/garante/doc.jsp?ID=1115480`.
[60] H. Zang, J. Bolot, *"Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks,"* ACM MobiCom, 2007.

[61] M. Coscia, S. Rinzivillo, F. Giannotti, D. Pedreschi, *"Optimal Spatial Resolution for the Analysis of Human Mobility,"* IEEE/ACM ASONAM, 2012.

[62] C. Iovan, A.-M. Olteanu-Raimond, T. Couronne, Z. Smoreda, *"Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies,"* Geographic Information Science at the Heart of Europe, D. Vandenbroucke, B. Bucher, J. Crompvoets, Springer, 2013.