Hiding Mobile Traffic Fingerprints with GLOVE *

Marco Gramaglia†
University Carlos III of Madrid
IMDEA Networks Institute
mgramagl@it.uc3m.es

Marco Fiore
CNR - IEIIT
Inria
marco.fiore@ieiit.cnr.it

ABSTRACT

Preservation of user privacy is paramount in the publication of datasets that contain fine-grained information about individuals. The problem is especially critical in the case of mobile traffic datasets collected by cellular operators, as they feature high subscriber trajectory uniqueness and they are resistant to anonymization through spatiotemporal generalization. In this work, we first unveil the reasons behind such undesirable features of mobile traffic datasets, by leveraging an original measure of the anonymizability of users' mobile fingerprints. Building on such findings, we propose GLOVE, an algorithm that grants k-anonymity of trajectories through specialized generalization. We evaluate our methodology on two nationwide mobile traffic datasets, and show that it achieves k-anonymity while preserving a substantial level of accuracy in the data.

CCS Concepts

•**Security and privacy** → **Pseudonymity, anonymity and untraceability;** *Data anonymization and sanitization;* •**Networks** → **Network privacy and anonymity;**

Keywords

Mobile traffic data; k-anonymity.

*This work was supported by the French National Research Agency under grant ANR-13-INFR-0005 ABCD and by the EU FP7 ERA-NET program under grant CHIST-ERA-2012 MACACO. It was performed using mobile communication data made available by SONATEL and Orange within the D4D Challenge.

†This work was carried out while Marco Gramaglia was at CNR-IEIIT.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CoNEXT '15 December 01-04, 2015, Heidelberg, Germany
© 2015 ACM. ISBN 978-1-4503-3412-9/15/12…$15.00
DOI: http://dx.doi.org/10.1145/2716281.2836111

1. INTRODUCTION

Public disclosure of datasets containing *micro-data*, i.e., information on individuals collected via surveys, transaction records, positioning and service logs, is an increasingly frequent practice. Indeed, these datasets yield fine-grained data about large populations that has proven central to seminal studies across research disciplines. Preserving user privacy in publicly accessible micro-data is a critical task, and naive anonymization techniques (e.g., hashing of users' identifiers) often offer inadequate protection. This has been repeatedly demonstrated by re-identification attacks on, e.g., medical records [1] or web service databases [2], which disclosed health conditions or political views of users.

Our work focuses on *movement micro-data* extracted from mobile traffic collected by cellular network probes. These data describe the movement of thousands to millions of subscribers over time periods of weeks to months. They have become an important instrument for large-scale analyses in sociology, demography, epidemiology, and computer science: recent surveys are available in [3, 4]. Like other types of micro-data, mobile traffic datasets are prone to attacks on individual privacy. Specifically, they suffer from (1) high uniqueness and (2) low anonymizability.

1. **High uniqueness.** Mobile subscribers have very distinctive patterns that often make them unique even within a very large population. Experiments showed that 50% of the mobile subscribers in a 25 million-strong dataset could be uniquely detected with minimal knowledge about their movement patterns, namely the three locations they visit the most frequently [5]. Similarly, an individual could be pinpointed among 1.5 million other mobile customers with a probability almost equal to one, by just knowing five random spatiotemporal points in his mobile traffic data [6].

We remark that uniqueness does not imply re-identifiability, since the sole knowledge of a specific subscriber's trajectory cannot disclose his identity: thus, in this work we do not re-identify any user present in the mobile traffic datasets we analyze. However, that link may become possible via cross-database linkage: in a recent

attempt, georeferenced check-in's of Flickr and Twitter users were leveraged to bring a de-anonymization attack on a mobile traffic dataset [7]. Several hundreds could be pinpointed with a 90% confidence level, and the authors argue that complete re-identification would have been possible with limited additional side information.

Uniqueness is thus a vulnerability that can be exploited for de-anonymization. Standard countermeasures rely on non-technical solutions, i.e., non-disclosure agreements that bound the scope of the activities carried out on the datasets, and prevent publication of data analysis 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.

Such a practice strongly limits the availability of mobile traffic datasets, as well as the reproducibility of related research. Mitigating the uniqueness of subscriber trajectories becomes then a very desirable facility that can entail more privacy-preserving datasets, and favor their open circulation. Here, the second problem of mobile traffic datasets comes into play.

2. **Low anonymizability.** The legacy solution to reduce uniqueness in micro-data datasets is generalization: data precision is reduced up to the point where no individual is uniquely distinguishable. However, previous studies showed that blurring users in the crowd, by reducing the spatiotemporal granularity of their movements, is hardly a solution in the case of mobile traffic datasets. Reliable anonymization is attained only under very coarse generalization, e.g., by disclosing users' locations at city-level precision [5]. In addition, a power-law relationship exists between uniqueness and spatiotemporal generalization of mobile traffic: additional privacy comes at an increasingly higher cost in terms of data resolution [6].

In conclusion, not only mobile traffic datasets yield highly unique trajectories, but the latter are also hard to anonymize. Ensuring individual privacy in these datasets easily compromises their utility. Our work tackles this precise problem, with a two-fold contribution.

First, we carry out a thorough investigation of the reasons behind the inconvenient properties of mobile traffic datasets outlined above. To that end, we define an original measure of the level of anonymizability of the mobile fingerprints left by subscribers as they interact with the cellular network. When applied to two nationwide datasets of mobile traffic, our measure offers novel insights on the causes behind the high uniqueness and poor anonymizability of this type of movement micro-data, which were not individuated in [5, 6].

Second, we propose a novel anonymization algorithm for mobile traffic datasets, which builds on the insights above. The algorithm, aptly named GLOVE, hides mobile fingerprints through so-called specialized generalization. GLOVE achieves indistinguishability of all users in our reference datasets while preserving substantial accuracy in the data. It yields a dramatic improvement over previous attempts at anonymization of mobile traffic, which could not attain a similar level of privacy without disrupting data utility.

Table 1: Movement micro-data from mobile traffic.

id	mobile fingerprint							
a	c_1,8	c_2,14	c_3,17					
b	c_4,8	c_5,15	c_6,15	...	c_{12},15	c_{13},15	c_{14},16	c_{15},17
c	c_{16},7	c_{17},20						

2. PROBLEM AND POSITIONING

In this section, we first formalize the general problem of user trajectory uniqueness in mobile traffic datasets, by introducing some fundamental definitions (Sec. 2.1). Then, we outline the scope of our work with respect to the general problem, so as to dispel any doubt on the applications and limitations of our methodology. Specifically, we first establish the precise objective we target (Sec. 2.2), and the attacker model we assume (Sec. 2.3). Then, we introduce a suitable privacy model (Sec. 2.4) under such objective and attacker model.

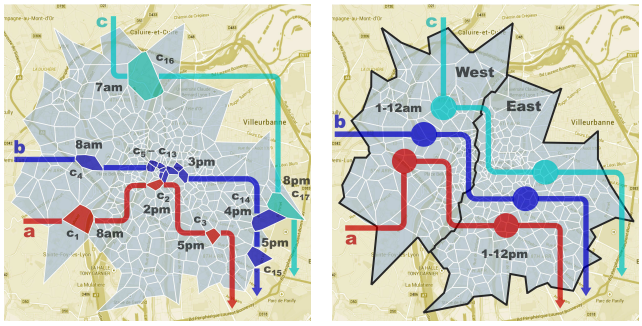
2.1 Definitions

Mobile traffic data is collected by mobile operators through probes deployed in their networks. Every mobile communication activity, either triggered by a user or autonomously initiated by his device, generates network events that are timestamped and associated to the current location of the device[1]. Mobile traffic data thus embeds information about the movement of individuals, since sequences of events can be regarded as a proxy of subscribers' trajectories. In this paper, we term the space and time information associated to each logged event a *spatiotemporal sample*. The complete set of samples associated to a specific user during the traffic recording period is the *mobile fingerprint* of that user.

An illustration is provided in Fig. 1a, which portrays the trajectories of three individuals, denoted as a, b, and c, respectively, across an urban area. User a interacts with the radio access infrastructure at 8 am, while he is in cell c_1 along his trajectory. Then, he triggers additional mobile traffic activities at 2 pm, while located in a cell c_2 in the city center, and at 5 pm, from a cell c_3 in the South-East outskirts. Thus, the mobile fingerprint of a is (c_1,8; c_2,14; c_3,17). The same for users b and c.

Mobile fingerprints are then collected into databases of movement micro-data. Tab. 1 provides a database example for the trajectories in Fig. 1a. The first column in Tab. 1 reports, for each mobile subscriber, an *identifier*,

[1]The actual precision (e.g., in space and time) of the information recorded varies significantly with the nature of the probes used for data collection [4]. Our discussion is independent of the mobile traffic data collection technique, and the solutions proposed in this work befit mobile traffic data featuring any level of precision.



(a) Original (b) Generalized

Figure 1: Example of mobile fingerprints of three subscribers. (a) Original dataset granularity: user locations are represented 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 precision.

e.g., his name, IMSI, or phone number. Since identifiers allow direct identification, they are never disclosed in movement micro-data. Instead, the common practice is to replace each identifier with a *pseudo-identifier*, i.e., a value that is unique for every individual in a same dataset, but changes across databases. Random strings or hashed identifiers are typical examples of pseudo-identifiers. Unfortunately, pseudo-identifiers do not guarantee indistinguishability: users' trajectories are often different from each other, making mobile fingerprints unique within the database [5, 6]. As discussed in Sec. 1, uniqueness is a vulnerability that can be exploited for cross-database correlation and user re-identification [7].

2.2 Objective

The ultimate objective our work contributes to is *Privacy-Preserving Data Publishing (PPDP)*, i.e., the provisioning of methods for the publication of information that is both privacy-preserving and useful. In our case, information maps to movement micro-data from mobile traffic, i.e., mobile fingerprints. Practical PPDP entails then the following requirements [8].

P1. **Publication of data, and not of data mining results**. We aim at producing privacy-preserving datasets of mobile fingerprints rather than anonymized datasets of classifiers, association rules, or aggregate statistics. This sets our goals apart from those of Privacy-Preserving Data Mining (PPDM), where the precise usage the data will be put to is known in advance.

P2. **Truthfulness at the record level**. Each published record, i.e., mobile fingerprint, must correspond to an existing individual in real life. In addition, samples in each mobile fingerprint must map to locations actually visited by the subscriber at that time. Randomized, perturbed, permuted, or synthetic data does not meet this requirement.

Our solution will obey the two principles above, which is why we do not target specific data usages and we discard some options for data anonymization. These concepts are unfolded in Sec. 2.4 below.

2.3 Attacker model

According to the classification of privacy preservation solutions proposed in [8], four different types of attacks can be envisioned against published micro-data: (i) *record linkage* aims at uniquely distinguishing an individual in the database; (ii) *attribute linkage* aims at correlating data pertaining to a same individual across different databases; (iii) *table linkage* aims at inferring whether an individual is present in a database; (iv) a *probabilistic attack* aims at improving some belief on an individual, by accessing the database. In this work we tackle the first category, i.e., record linkage attacks. We regard our approach as a sensible initial step towards a complete suite of solutions capable of guaranteeing PPDP against all types of attacks mentioned above.

Also, unlike previous works that have considered partial attacker's knowledge of the subscribers' mobile fingerprints (e.g., preferred locations [5] or random sample subsets [6]), we do not assume any specific adversary knowledge. This maps to ensuring so-called *quasi-identifier-blind anonymity* [9], i.e., accept that an attacker can be aware of any portion of the target user's trajectory, including the entirety of it. This choice is motivated by the fact that there is currently no reliable model of the attacker's knowledge [10], and making hypothesis in that sense may be dangerous. In this perspective, potential data providers will not accept to disclose datasets whose anonymization is only robust to one well-defined attacker model that surmises abilities or prior information of the opponent.

Summarizing the discussion above, we target releasing data that is robust to record linkage attacks, under the most general model of attacker's knowledge.

2.4 Privacy model

Our privacy model is consistent with the objective and attacker model presented in Sec. 2.2 and Sec. 2.3.

First, we adopt k-*anonymity* as a criterion of indistinguishability, among the many proposed for micro-data. k-anonymity commends that each individual in a dataset must be indistinguishable from at least k-1 other users in the same dataset. In our case, each mobile fingerprint needs to be hidden in a crowd of other k identical ones. This criterion is known to have limitations when confronted to attacks aiming at attribute linkage, at localizing users, or at disclosing their presence and meetings [11, 12]. However, k-anonymity is an effective countermeasure against the record linkage attacks we target (see Sec. 2.3), and thus perfectly fits our needs.

Second, we consider k-anonymization of the *full-length* mobile fingerprint of each user in the dataset. Indeed, this is the only way to ensure data robustness indepen-

dently of the attacker's knowledge, i.e., abiding by the quasi-identifier-blind anonymity principle (see Sec. 2.3). We stress that full-length fingerprint anonymization is a more demanding task than protecting the same data from attacks that narrow the adversary's capability (e.g., by assuming that the adversary only knows a limited set of popular locations [5] or random spatiotemporal samples [6] from the target user's trajectory).

Third, we adopt *spatiotemporal generalization* and *suppression* as techniques to achieve the k-anonymity criterion in the movement micro-data we target. As anticipated in Sec. 1, spatiotemporal generalization relies on reducing data precision in space and time so as to make samples of different mobile fingerprints identical. Suppression allows instead removing some data, either individual samples or whole fingerprints (i.e., users), from the dataset, because they do not fulfill the anonymity criterion.

Overall, we seek a solution for the full-length k-anonymization of mobile fingerprints through spatiotemporal generalization and possibly suppression. This privacy model fully conforms to the PPDP principles set forth in Sec. 2.2, as follows.

- P1 dictates that the anonymized data must be analysis-agnostic. In other words, (i) the anonymized data must retain the same format of the original data, and (ii) be as close as possible to the original data. Our privacy model fulfills the first point, since it returns trajectories of spatiotemporal samples, semantically identical to those in the original data. It also satisfies the second point, as the only way it acts on the trajectories is by reducing their granularity; and, it does so in a way that the accuracy loss is minimized.

- P2 limits the set of transformations that can be applied on the mobile fingerprints to those that do not inject new, fabricated spatiotemporal samples in the data. Both techniques considered in our work, i.e., spatiotemporal generalization and suppression conform to this principle.

We provide an example of the privacy model in Fig. 1b, for the mobile fingerprints of Fig. 1a. There, spatiotemporal generalization reduces the granularity of the data in space (cells are aggregated into two macroscopic East and West regions) and time (the temporal precision is reduced to 12-hour intervals). The three subscribers a, b and c are 3 now have identical full-length fingerprints (West,1-12; East,13-24), i.e., they are 3-anonymized. Clearly, the generalization induces a loss of accuracy in the data. E.g., in the example of Fig. 1b, the mobile fingerprint that allows hiding a, b and c is very coarse both in space and time. This is precisely the problem of the low anonymizability of mobile traffic datasets introduced in Sec. 1: in this type of micro-data, even guaranteeing 2-anonymity requires a reduction of granularity so severe to impair data utility [5, 6]. It is our goal to attain such anonymity while preserving substantial accuracy in the mobile traffic data.

As a final remark, we reckon that the privacy model we propose has limitations and may not suit all data analyses. Specifically, we believe that k-anonymized data better fits studies on, e.g., the routine behaviors of individual subscribers (e.g., home and work locations, next location predictions), or aggregate statistics on user populations (e.g., investigation of land uses, commuting flows, population distributions). Instead, analyses targeting outlying behaviors (e.g., visits to unusual locations by individuals, overnight mobility flows) may be distorted if run on k-anonymized data. Indeed, uncommon movement patterns introduce, by their own nature, some uniqueness that must be necessarily removed to attain indistinguishability.

3. MOBILE TRAFFIC DATASETS

For the purpose of our study, we extract movement micro-data, in the form of subscribers' mobile fingerprints, from two datasets of mobile traffic released by Orange within Data for Development Challenges [13].

- **Ivory Coast.** The first dataset describes five months of Call Detail Records (CDR) in the whole Ivory Coast. It contains the complete spatiotemporal trajectories for a subset of 50,000 randomly selected users, re-drawn every two weeks. Thus, the dataset contains information about ten 2-week periods. We ran a preliminary screening, filtering out users that have less than one sample per day in their fingerprint. Then, we merged all remaining fingerprints into a database of 82,000 records. This dataset is indicated as `d4d-civ` in the following.

- **Senegal.** The second dataset is derived from CDR collected over the whole Senegal for one year. It contains a randomly selected subset of 320,000 users over a rolling 2-week period, for a total of twenty-five periods. We did not filter out subscribers, since the dataset is already limited to users that are active for more than 75% of the 2-week time span. In our study, we consider one representative 2-week period among those available. This dataset is referred to as `d4d-sen` in the rest of the paper.

In both mobile traffic datasets, the information about user position is provided as a latitude and longitude pair that corresponds to an antenna location. We mapped the latter to a two-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[2]. As for the temporal dimension, the maximum precision granted by both datasets is one minute, and this is also our finest time granularity.

[2]At 100-m spatial granularity, each grid cell contains at most one antenna location from the original dataset: the process does not cause any loss in data accuracy.

4. MEASURING ANONYMIZABILITY

Our first objective is understanding the causes behind the high uniqueness and low anonymizability of movement micro-data extracted from mobile traffic. To that end, we propose and leverage a measure of the level of anonymizability of a mobile fingerprint, which estimates how easy (or difficult) it is to hide a given fingerprint in a dataset. Coherently with the scope of our work discussed in Sec. 2, the measure is based on the k-anonymity criterion and assumes spatiotemporal generalization to achieve it. Thus, the measure evaluates the spatiotemporal loss of accuracy required to make the fingerprint of a subscriber indistinguishable from those of k-1 other users in the same dataset. We name our measure the *k-gap* of a mobile fingerprint, and denote it as Δ_a^k in the case of a user a that needs to be k-anonymized.

4.1 Sample stretch effort

The k-gap of a fingerprint depends on the cost of k-anonymizing the spatiotemporal samples that compose it. We thus start by providing an expression of the *sample stretch effort*, i.e., the spatiotemporal loss of accuracy required to merge two samples through generalization. The sample stretch effort between the i-th sample of a's fingerprint and the j-th sample of b's fingerprint is denoted as $\delta_{ab}(i,j)$ in the following.

Let us consider the i-th sample of user a's fingerprint. We indicate the spatial information it conveys as a tuple $\sigma_i^a = (x_i^a, dx_i^a, y_i^a, dy_i^a)$, which outlines the boundaries of the geographical rectangle where a is located. Similarly, the temporal information in the sample is $\tau_i^a = (t_i^a, dt_i^a)$, meaning that a was within the area σ_i^a at some point in time during the interval between t_i^a and $t_i^a + dt_i^a$. As indicated in Sec. 3, $dx_i^a = dy_i^a = 100$ m and $dt_i^a = 1$ min, for all original fingerprints in our reference datasets.

A generic formulation of the sample stretch effort $\delta_{ab}(i,j)$ that accounts for generalization along both spatial and temporal dimensions is then

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

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 normalization factors $w_\sigma = w_\tau = 1/2$ ensure that $\delta_{ab}(i,j) \in [0,1]$.

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

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

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

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

Formally, the stretch in space ϕ_σ^* is computed as

$$\phi_\sigma^*\left(\sigma_i^a, \sigma_j^b\right) = \frac{\left[l_\sigma\left(\sigma_i^a, \sigma_j^b\right) + r_\sigma\left(\sigma_i^a, \sigma_j^b\right)\right] n_a}{n_a + n_b} + \\ + \frac{\left[l_\sigma\left(\sigma_j^b, \sigma_i^a\right) + r_\sigma\left(\sigma_j^b, \sigma_i^a\right)\right] n_b}{n_a + n_b}, \quad (4)$$

where

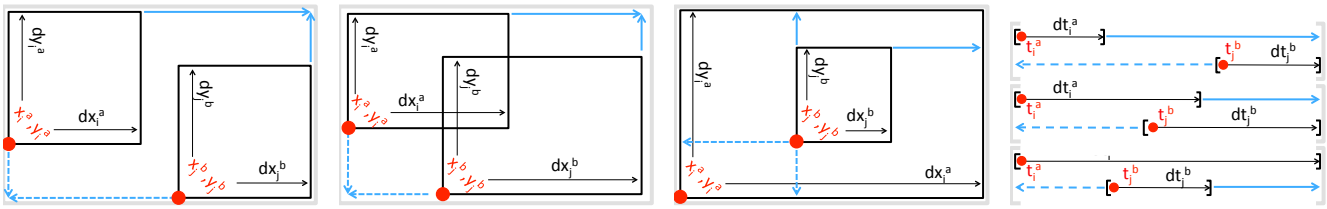
$$l_\sigma\left(\sigma_i^a, \sigma_j^b\right) = [x_i^a - \min(x_i^a, x_j^b)] + \\ + [y_i^a - \min(y_i^a, y_j^b)], \quad (5)$$

$$r_\sigma\left(\sigma_i^a, \sigma_j^b\right) = [\max(x_i^a + dx_i^a, x_j^b + dx_j^b) - x_i^a - dx_i^a] + \\ + [\max(y_i^a + dy_i^a, y_j^b + dy_j^b) - y_i^a - dy_i^a]. \quad (6)$$

The l_σ and r_σ functions quantify the *left stretch* and *right stretch* in space, i.e., they measure how much the boundaries of the first sample, σ_i^a, need to be extended along the longitudinal and latitudinal axes, in order to cover the bounding rectangle of the second sample, σ_j^b. Graphical examples are provided in Fig. 2a–2c. 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 fingerprints of two individual subscribers, then $n_a = n_b = 1$. However, our definitions above can accommodate the case where a and b are not two subscribers, but two groups of subscribers whose fingerprints have already been made indistinguishable. In that case, n_a and n_b represent the number of subscribers whose mobile fingerprints have already been generalized into fingerprints a and b, respectively. Then, the rationale for the weighted sum is that stretching a sample of fingerprint a reduces the accuracy in the data of n_a users, and the same is true for the n_b users in b:

[3]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 small interest to most studies. However, an important remark is that the ϕ_σ^{max} and ϕ_τ^{max} values also determine the derivative of the linear relationship between granularity loss and information loss. Since the two contributions in (2) and (3) are simply summed in (1), the ratio between ϕ_σ^{max} and ϕ_τ^{max} has a precise physical meaning: it indicates which loss of accuracy in space is equivalent to which loss of accuracy in time. Thus, the aforementioned values of ϕ_τ^{max} and ϕ_τ^{max} are also chosen in a way to assign the same weight to a spatial generalization of $\sim$0.5 km and a temporal generalization of $\sim$15 min. The rationale is that the vast majority of data mining processes start suffering some information loss only after the data accuracy falls below either of these values, which makes them equivalent from a data utility viewpoint.



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

Figure 2: Examples of the stretch needed to merge two spatiotemporal samples. 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. Different levels of overlap between the spatial and temporal components of samples are shown. (a,b,c) Spatial stretch. (d) Temporal stretch.

thus, weighting the stretches properly accounts for the number of users affected by the generalization.

Equations are similar in the case of time, where

$$\phi^*_\tau\left(\tau^a_i,\tau^b_j\right) = \frac{\left[l_\tau\left(\tau^a_i,\tau^b_j\right)+r_\tau\left(\tau^a_i,\tau^b_j\right)\right]n_a}{n_a+n_b} + \\ + \frac{\left[l_\tau\left(\tau^b_j,\tau^a_i\right)+r_\tau\left(\tau^b_j,\tau^a_i\right)\right]n_b}{n_a+n_b}, \quad (7)$$

$$l_\tau\left(\tau^a_i,\tau^b_j\right) = [t^a_i - \min(t^a_i,t^b_j),] \quad (8)$$

$$r_\tau\left(\tau^a_i,\tau^b_j\right) = [\max(t^a_i+dt^a_i,t^b_j+dt^b_j) - t^a_i - dt^a_i]. \quad (9)$$

Again, l_τ and r_τ mark the left stretch and right stretch in time; representative examples are provided in Fig. 2d. The contributions of a and b stretches in (7) are weighted by the number of subscribers involved, as in (4).

4.2 k-gap of mobile fingerprints

We can now define the *fingerprint stretch effort*, i.e., the spatiotemporal loss of accuracy required to merge two whole fingerprints via generalization. Considering the fingerprints of (groups of) users a and b, the effort, denoted as Δ_{ab}, is computed as

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

Here, m_a and m_b are the cardinalities of the fingerprints of a and b, respectively. The expression in (10) finds, for each sample in the longer fingerprint, the sample at minimum stretch effort in the shorter fingerprint. Δ_{ab} is the average of all such sample stretch efforts[4].

[4]We emphasize that all solutions adopted in the design of the fingerprint stretch effort in (10) are primarily driven by a scalability rationale. Given the extremely large size of mobile traffic datasets, we opted for very simple formulations: examples are the sum of space and time contributions in (1), the constant spatial and temporal thresholds in (2) and (3), or the rectangular stretches in (4) and (7). This approach limits the computational complexity of calculating (10), an operation that has to be repeated millions of times in order to characterize the anonymizability of a mobile traffic dataset. Although the deliberate simplicity of these formulations paves the road to more complex proposals, the results in Sec. 5 and Sec. 7 prove that the expression of Δ_{ab} in (10) already does an effective job of estimating the cost of merging mobile fingerprints.

The k-gap Δ^k_a of a generic mobile user a that is to be k-anonymized can then be computed as the average stretch effort of a's fingerprint from those of the nearest k-1 other users in the dataset. Formally

$$\Delta^k_a = \frac{1}{k-1}\sum_{b\in\mathbb{N}^{k-1}_a}\Delta_{ab}, \quad (11)$$

where $\mathbb{N}^{k-1}_a$ is the set of $k-1$ users b with the lowest fingerprint stretch effort to that of a.

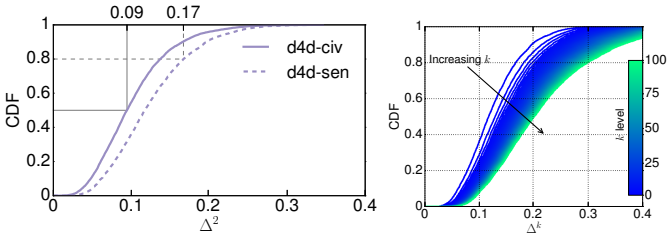
The expression in (11) returns a measure $\Delta^k_a \in [0,1]$ that indicates how hard it is to hide subscriber a in a crowd of k users in the same dataset. If $\Delta^k_a = 0$, user a is already k-anonymous. If $\Delta^k_a = 1$, the user is completely isolated, and k-anonymization makes all his samples so coarse in space and time that they are uninformative.

5. ANONYMIZABILITY ANALYSIS

The k-gap in (11) can be intended as a dissimilarity measure, and employed in legacy definitions used to assess micro-data sparsity, e.g., (ϵ,δ)-sparsity [2]. However, these definitions are less informative than complete distributions. Thus, in this section, we characterize the level of anonymizability of a dataset through the Cumulative Distribution Function (CDF) of the k-gap of all users in that dataset. We use the mobile traffic datasets in Sec. 3 as our reference case studies.

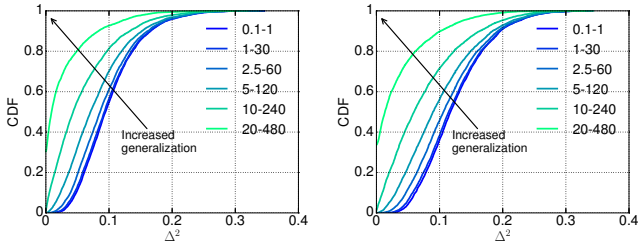
5.1 The good: anonymity is close to reach

Our baseline result is portrayed in Fig. 3a. The plot depicts the CDF of k-gap in the `d4d-civ` and `d4d-sen` mobile traffic datasets, when considering 2-anonymity as the privacy criterion. We observe that the two curves are quite similar, and both are at zero in the x-axis origin. This means that $\Delta^2_a > 0\ \forall a$, i.e., no mobile subscriber is 2-anonymous in either of the original datasets. The result is in line with previous analyses carried out on different mobile traffic datasets [5, 6], which confirms that the high uniqueness of subscribers' trajectories is an intrinsic property of any mobile traffic dataset, and not a specificity of those we consider in our study.



(a) $k = 2$ (b) $2 \leq k \leq 100$

Figure 3: CDF of k-gap. (a) $k = 2$, `d4d-civ` and `d4d-sen` datasets. (b) varying k, `d4d-sen` dataset.



(a) `d4d-civ` (b) `d4d-sen`

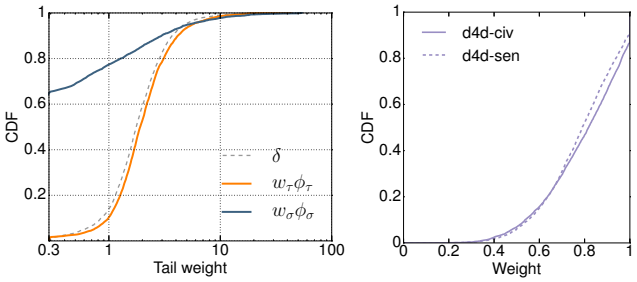
Figure 4: CDF of k-gap, for $k = 2$ and varied spatiotemporal generalizations (labeled in km–min), in the `d4d-civ` and `d4d-sen` mobile traffic datasets.

Interestingly, we remark the probability mass is, for both datasets, below 0.2, i.e., it is not far from the origin. This is good news, as it implies that the fingerprint stretch effort needed to make most users 2-anonymous is fairly low. As an example, 50% of the users in the `d4d-civ` dataset have a k-gap of 0.09 or less, which maps, on average, to a combined spatiotemporal generalization of less than one km and little more than 20 minutes. In other words, the 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. Similar considerations hold in the `d4d-sen` case: e.g., 80% of the dataset population has a k-gap of 0.17 or less, i.e., has an average spatial and temporal distances of 1.7 km and 41 minutes from 2-anonymity.

One may wonder how more stringent privacy requirements affect these results. Fig. 3b shows the evolution of the anonymizability of the `d4d-sen` datasets when k varies from 2 to 100. Identical results were obtained in the `d4d-civ` case and are omitted here. As expected, higher values of k require that a user is hidden in a larger crowd, and thus shift the distributions towards the right, i.e., need a coarser generalization. However, quite surprisingly, the shift is not dramatic, as the cost of k-anonymity grows sub-linearly with k.

5.2 The bad: generalization does not work

Unfortunately, the easy anonymizability suggested by the results above is only apparent. Fig. 4 shows the impact of spatiotemporal generalization on k-gap, $k = 2$.



(a) TWI (b) Ratio

Figure 5: (a) CDF of the Tail Weight Index computed on the distributions of sample stretch efforts (overall, and separated into spatial and temporal components) for all users in the `d4d-civ` datasets. (b) CDF of the temporal-to-spatial component ratios in the overall sample stretch effort, for all users in the `d4d-civ` and `d4d-sen` datasets. All results refer to $k = 2$.

Each curve corresponds to a different level of generalization of samples in mobile fingerprints, from the original dataset granularity of 100 meters and 1 minute, to an uninformative granularity of 20 km and 8 hours. As one could expect, increased generalization pushes the distribution towards the left, i.e., makes the dataset 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 cannot 2-anonymize but ∼35% of mobile users' trajectories. The result is again in agreement with previous studies [5, 6], and confirms the second property of mobile traffic datasets pointed out in Sec. 1, i.e., their low anonymizability.

5.3 The why: long-tailed time diversity

The results shown up to this point yield an apparent incongruity: spatiotemporal generalization performs poorly (Fig. 4), yet the fingerprint stretch effort needed to attain k-anonymity is in theory low (Fig. 3). In fact, the fingerprint stretch effort is an average of multiple sample stretch efforts, as per (10). We then hypothesize that the discrepancy above has roots in the diversity across the stretch efforts associated to different samples in a same fingerprint.

In order to test our proposition, we evaluate the statistical dispersion of the sample stretch efforts within each fingerprint. To that end, for each user a in the dataset, we retrieve the set $\mathbb{N}_a^{k-1}$ of k-1 other subscribers whose fingerprints are the closest to that of a, according to (11). Then, we disaggregate all the fingerprint stretch efforts Δ_{ab} between a and the users $b \in \mathbb{N}_a^{k-1}$ into sample stretch efforts δ_{ab}, as per (10). Finally, we separately collect the spatial and temporal components of all such sample distances, in (1), 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 sam-

ples of a's fingerprint, separately in the spatial and temporal dimensions. 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: the higher the TWI, the heavier the tail [14]. Specifically, we compute, for all fingerprints in a dataset, the TWI of three CDFs: the total stretch effort per sample (δ), as well as the associated spatial ($w_\sigma \phi_\sigma$) and temporal ($w_\tau \phi_\tau$) components.

Fig. 5a shows the CDFs of the TWI computed on all fingerprints in the `d4d-civ` dataset. Identical results were obtained in the `d4d-sen` case and are omitted here. The TWI in the spatial dimension is below 1.5 in around 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 $\geq$ 1.5 in around 70% of cases[5]. As a result, the TWI of the overall stretch effort (δ) distribution is shaped after that of temporal components.

Quantitative analyses confirm this last resolution. The plot in Fig. 5b 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 all users 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 fingerprints, 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 fingerprint stretch effort; in 15% of cases, the cost of anonymization is fully determined by the temporal stretch. We conclude that the temporal component of a mobile fingerprint is much harder to anonymize than the spatial one. In other words, *where* an individual generates mobile traffic activity is easily masked, but hiding *when* he carries out such activity is not.

5.4 Takeaways

The results presented in this section let us postulate that typical mobile fingerprints 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 fingerprints.

However, mobile fingerprints 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. Their impact is dramatic since, in order to hide a subscriber, one has to reduce granularity in space and time until *all* of his samples are merged within the fingerprints of k-1 more users in the same dataset. As a result, the single sample that is the hardest to anonymize in a fingerprint dooms all the others to undergo the same loss of accuracy it requires to be k-anonymized. Ultimately, this makes spatiotemporal generalization ineffective in attaining k-anonymity.

Overall, not only our analysis is consistent with previous results on the high uniqueness and low anonymizability of mobile traffic datasets [5, 6], but it provides, for the very first time, a rigorous explanation for such undesirable features. This new understanding of the characteristics of mobile fingerprints also represents the cornerstone for the development of anonymization techniques that better fit the specificity of mobile traffic datasets, as discussed next.

6. GLOVE

We leverage the insights in Sec. 5.4 to design a novel algorithm for the k-anonymization of movement microdata extracted from mobile traffic datasets. The algorithm, named GLOVE, builds on the fact that: (i) the vast majority of spatiotemporal samples in mobile fingerprints can be hidden with limited loss of accuracy; (ii) only a smaller portion of samples requires drastic generalization. At the light of these observations, GLOVE uses a *specialized generalization*, where each sample undergoes an independent, minimal reduction of granularity that hides it in a crowd of k.

6.1 Algorithm in a nutshell

The pseudocode of GLOVE is listed in Alg. 1. The inputs to the algorithm are the mobile fingerprint dataset and the value of k, i.e., the target k-anonymity level. In the initialization phase, the fingerprint stretch efforts between all fingerprint pairs are calculated according to (10), and stored in a matrix S (line 2). The algorithm then iterates until all fingerprints have been k-anonymized (line 3). At each iteration, the two fingerprints that have not yet been k-anonymized and are at minimum stretch effort in S are identified and removed from the database and from S (lines 5–6). They are then merged into a single fingerprint, which includes a number of subscribers equal to the sum of those already hidden into the two original fingerprints (lines 7–8). The resulting fingerprint is included in the database, computing its stretch effort, via (10) again[6], to all other yet-to-be-anonymized fingerprints (lines 9–12).

6.2 Fingerprint merge

The pseudocode in Alg. 1 involves a merging operation that returns one generalized fingerprint from two original fingerprints (line 7). We propose a two-stage process to perform the merging, in Fig. 6a. In the first stage, each sample in the longer fingerprint a is matched to that in the shorter fingerprint b at minimum sample stretch effort, computed as in (1). Then, all samples in

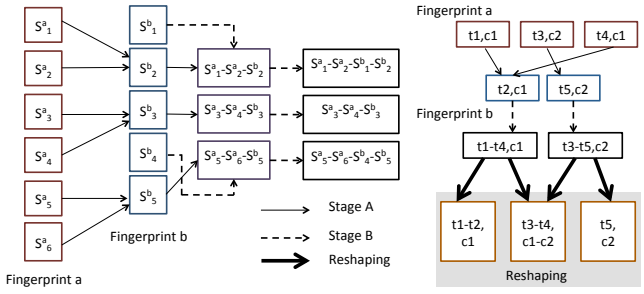
[5]An exponential distribution with parameter equal to one has TWI 1.6, whereas a fat-tailed Pareto distribution with shape equal to one has TWI 14.

[6]We recall that the expression in (10) can accommodate the case where the input fingerprints are already the result of a merge, see Sec. 4.2.

input : Anonymization level k
input : Mobile fingerprint dataset $\mathbb{M}$
output: Anonymized fingerprint dataset $\mathbb{M}$

foreach $a,b \in \mathbb{M}, a \neq b$ **do**
  S $[a,b]$=`calcStretch` (a,b) ;
end
while $\exists\, a,b \in \mathbb{M}$ s.t. $a.k < k,\ b.k < k$ **do**
  $a,b \leftarrow$ `leastStretch(S)` ;
  `remove(`$\mathbb{M}$,S,a,b`)` ;
  $m \leftarrow$ `merge(`a,b`)` ;
  $m.k = a.k + b.k$;
  `add(`$\mathbb{M}$,m`)` ;
  if $m.k < k$ **then**
    foreach $c \in \mathbb{M}$ s.t. $c.k < k$ **do**
      S $[c,m]$=`calcStretch` (c,m) ;
    end
  end
end

Algorithm 1: GLOVE algorithm pseudocode.



(a) Merging (b) Reshaping

Figure 6: (a) Example of fingerprint merging operation. (b) Example of reshaping operation.

fingerprint a pointing to a same sample in fingerprint b (e.g., s_1^a and s_2^a, pointing at s_2^b) are merged with the latter. In the second stage, the same procedure is run on samples of the shorter fingerprint that have not been merged during the first stage (e.g., s_1^b). These samples are matched with those resulting from the first stage (e.g., s_1^a–s_2^a–s_2^b in the case of s_1^b).

At both stages, samples are merged through spatiotemporal generalization. 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. Reusing the notation introduced in Sec. 4.1, the generalization is realized as follows:

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

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

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 Fig. 6a), the operations can be run iteratively, merging one sample at a time.

It is important to note that equations (12) and (13) realize our principle of specialized generalization, since: (i) the loss of granularity is the minimal required to hide each sample; (ii) the generalization is different among samples, breaking the dependency of all samples in a fingerprint from the hardest-to-anonymize one.

As a last remark, 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. 6b: there, locations $\sigma_2^a = c_2$ and $\sigma_1^b = c_1$ are farther in space than instants $\tau_2^a = t3$ and $\tau_2^b = t5$ are in time. Thus, the merging of fingerprints a and b leads to generalized samples of m that overlap in time, but refer to different geographical locations. The resulting fingerprint is formally correct, but it is difficult to read or analyze. We run a reshaping process that resolves all temporal overlappings, either partial or complete, by creating a new sample for each such case. The new sample covers the overlapping time intervals, and its spatial granularity is obtained by merging the geographical areas of the overlapping samples it replaces, as per (12) and (13). Reshaping has a cost in terms spatial granularity, but it improves the usability of anonymized data.

6.3 Complexity analysis

Attaining optimal k-anonymity is a NP-hard problem in movement micro-data databases [15]. GLOVE takes a greedy approach that requires: (i) computing the fingerprint stretch effort among all possible pairs of users in the original mobile traffic dataset; (ii) iteratively merging the two closest fingerprints and recomputing the stretch effort between the merged fingerprint and all those remaining in the dataset.

Let us denote as $|\mathbb{M}|$ the number of users in the dataset, and as $\bar{n}$ their average fingerprints length. The complexity of the first operation (i) above maps to the calculation of (10), whose cost is $\mathcal{O}\left(\bar{n}^2\right)$, for all $|\mathbb{M}|^2$ user pairs, and is thus $\mathcal{O}\left(|\mathbb{M}|^2 n^2\right)$. The complexity of the second operation (ii) is the sum of two contributions. On the one hand, the merge has a cost $\mathcal{O}\left(\bar{n}^2\right)$, and needs to be repeated $k = \mathcal{O}\left(|\mathbb{M}|\right)$ times (where k is the desired k-anonymity level, and cannot exceed the number of users in the dataset) for all users $|\mathbb{M}|$, leading to $\mathcal{O}\left(|\mathbb{M}|^2 \bar{n}^2\right)$. On the other hand, the recalculation of the stretch efforts for the new merged fingerprint requires computing (10) against all remaining users, which are $\mathcal{O}\left(|\mathbb{M}|\right)$, for an overall cost $\mathcal{O}\left(|\mathbb{M}|\bar{n}^2\right)$. Overall, GLOVE runs in polynomial time, and is quadratic in both the number of users and the fingerprint length.

Moreover, a strategic aspect in the design of GLOVE is that all of its key calculations are highly parallelizable. The implementation used in this paper relies on the Nvidia CUDA architecture for GPU computing. The calculations in (10), (12) and (13) were easily mapped to match the parallel computation capability of a GPU.

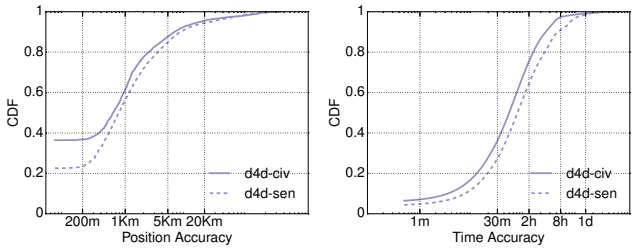


Figure 7: Spatiotemporal accuracy in the `d4d-civ` and `d4d-sen` datasets, 2-anonymized with GLOVE.

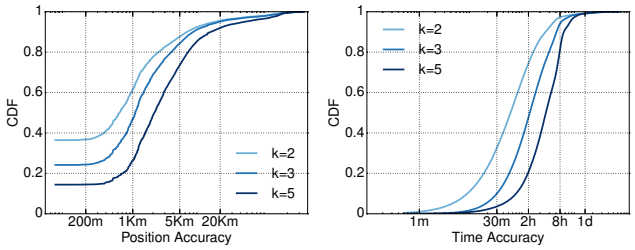


Figure 8: Spatiotemporal accuracy in the `d4d-civ` dataset, k-anonymized with GLOVE.

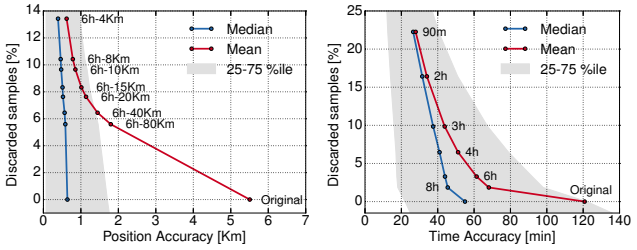


Figure 9: Spatiotemporal accuracy in the `d4d-civ` dataset, 2-anonymized with GLOVE and suppression.

A non-optimized proof-of-concept version of the software executes the calculations in (10) on 20–50,000 fingerprint pairs per second, using a single-GPU, low-end GeForce GT 740 card with 384 CUDA cores at 1 GHz. On this machine, the `d4d-civ` and `d4d-sen` datasets could be 2-anonymized with GLOVE in roughly 60 hours each. However, we believe that much better performance can be expected by running an optimized version of the software on dedicated, high-end hardware[7].

7. PERFORMANCE EVALUATION

GLOVE guarantees, by design, k-anonymity of all mobile fingerprints in a dataset. This is a result that legacy spatiotemporal generalization could not achieve, even under severe loss of granularity, as shown in Fig. 4. Clearly, the question is at which cost, in terms of precision loss, GLOVE attains k-anonymity.

Fig. 7 shows the accuracy of GLOVE-anonymized fingerprint samples in the `d4d-civ` and `d4d-sen` datasets, for the baseline case of 2-anonymity. The two plots outline the spatial and temporal accuracy of the anonymized data. We observe that 20% to 40% of the samples retain their original spatial accuracy, and have a temporal error of 30 minutes or less, largely sufficient to accurately characterize human mobility [16]. Even for larger fractions of samples, the loss of spatiotemporal granularity is tolerable: 70% to 80% of samples have a spatial error of less than 2 km and a temporal error of less than 2 hours – a level of accuracy that can support a large variety of studies in networking, sociology, or transportation research. Also, we recall that, under similar levels of generalization, no single subscriber could be 2-anonymized in Fig. 4.

Although 2-anonymity already satisfies the indistinguishability principle, higher privacy levels are possible, at a cost in terms of accuracy. Fig. 8 details the trade-off for the `d4d-civ` dataset. Identical trends were found for `d4d-sen`, and are omitted here. The percentage of samples with unvaried position accuracy drops to 25% for $k = 3$ 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 type of analysis to be carried out on the data, the fraction of exploitable samples may be significantly reduced for $2 < k \leq 5$.

For $k > 5$, the anonymized dataset becomes hardly exploitable: if such a level of protection is required, then one may try to simplify the problem, by, e.g., making assumption about the attacker's knowledge. This would allow modifying GLOVE operation so as to target, e.g., partial fingerprint anonymization, which is less expensive to achieve than the full-length version we are targeting in this work. Allowing suppression is another option, and we discuss it next.

7.1 Combining GLOVE with suppression

Suppression allows discarding hard-to-anonymize samples from fingerprints and is easily integrated in GLOVE. Indeed, specialized generalization can be combined with removal of samples whose temporal or spatial stretch efforts in (12) and (13) exceed some threshold. Fig. 9 shows the improvement of spatiotemporal accuracy (x axis) when imposing different thresholds to the spatial and temporal stretch (tags along curves), which results on discarding some percentage of samples (y axis). The plot refers to the the `d4d-civ` dataset. Trends are similar in the `d4d-sen` case. 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

[7]On a related point, dataset anonymization for PPDP is a one-time operation that is performed just once on the original data before it is released. Data releasing typically occurs with months of delay with respect to the data collection phase, due to data cleaning and legal clearance issues. Thus, the processing time is a much less relevant issue in PPDP than in other use cases.

		d4d-civ		d4d-sen		abidjan		dakar	
		W4M-LC	GLOVE	W4M-LC	GLOVE	W4M-LC	GLOVE	W4M-LC	GLOVE
$k=2$	Discarded fingerprints	1,104 (1.3%)	0	430 (0.1%)	0	3,387 (16.8%)	0	994 (1.4%)	0
	Created samples (×1000)	4,444 (24.9%)	0	5,302 (17.9%)	0	1,004 (20.8%)	0	1,949 (17.4%)	0
	Deleted samples (×1000)	1,325 (7.5%)	1,482 (8.3%)	1,577 (5.3%)	4,175 (14.1%)	997 (20.7%)	194496 (4.03%)	686 (6.1%)	535 (4.8%)
	Mean position error [m]	10,190.88	1,013.71	9,392.85	1,312.28	2,939.70	1,323.12	1,889.34	1,249.24
	Mean time error [min]	1,151.51	60.21	1,037.74	69.31	2,769.74	57.21	1,172.66	58.47
$k=5$	Discarded fingerprints	1,271 (1.5%)	0	3740 (1.2%)	0	204 (0.1%)	0	614 (0.8%)	0
	Created samples (×1000)	8,018 (44.9%)	0	8,863 (29.9%)	0	3,524 (73.8%)	0	3.310 (17.4%)	0
	Deleted samples (×1000)	1,713 (9.6%)	1,482 (8.3%)	2,179 (7.3%)	5,004 (16.9%)	652 (13.5%)	213 (4.41%)	877 (7.8%)	721 (6.4%)
	Mean position error [m]	23,534.062	5,129.9	19,881.9	5,694.2	4,033.4	1,870.10	3,365.22	1,596.77
	Mean time error [min]	3,455.94	171.01	2.600.64	408.30	3,334.4	146.90	2,030.13	147.53

Table 2: Comparative analysis of W4M-LC [17] and GLOVE. Results for anonymized datasets with $k = 2$ and $k = 5$.

stretch above 20 km, and whose temporal stretch[8] is above 6 h. Similarly, the average temporal accuracy is halved by suppressing just 4% of samples, i.e., thresholding at 6 h. Not only the mean, but also the median and 25^{th}–75^{th} percentile range are noticeably improved.

Interestingly, the accuracy gain is the most significant when only a small percentage of samples is removed form the dataset. We conclude that minimal suppression allows discarding a limited number of hard-to-anonymize outliers in the data, with a consequent large gain in accuracy.

7.2 Comparative analysis

Several solutions were proposed for the k-anonymization of trajectories, and are reviewed in Sec. 8. Among those, the only technique that can hide movement micro-data along both spatial and temporal dimensions is Wait for Me (W4M) [17], and we thus select it as the state-of-the-art benchmark for GLOVE.

The approach of W4M builds on the representation of an uncertain trajectory as a cylindrical volume that has a diameter δ in space and stretches through 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 cylindrical volume. An important remark is that W4M allows both suppression and creation of new synthetic samples. The latter operation is leveraged to improve the matching among trajectories in a cluster, and assumes that mobile objects (i.e., subscribers in our case) effectuate linear constant-speed movements between spatiotemporal samples. We use W4M with linear spatiotemporal distance and chunking (LC), i.e., the version intended for large databases, and indeed the only scaling to our mobile traffic data[9].

We ran W4M-LC on the `d4d-civ` and `d4d-sen` datasets, as well as on subsets of the same, geographically limited to major cities in Ivory Coast and Senegal, and named `abidjan` and `dakar`. We used the suggested settings for W4M-LC, i.e., $\delta = 2$ km and a 10% trashing (allowing the removal of trajectories that are difficult to cluster) [17]. 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.

[8]The left plot of Fig. 9 also considers a temporal stretch threshold, since spatial thresholding alone yielded marginal accuracy gain.

[9]Available at `http://kdd.isti.cnr.it/W4M/`.

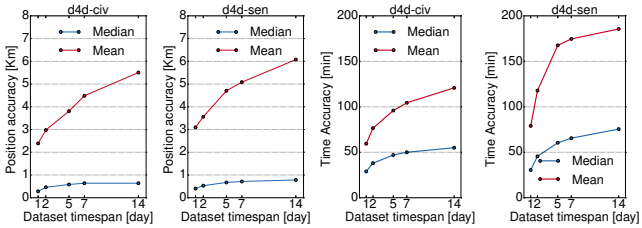


Figure 10: Spatiotemporal accuracy in time-subsets of `d4d-civ` and `d4d-sen`, 2-anonymized with GLOVE.

Differences are significant. In all scenarios, W4M-LC creates a substantial amount of synthetic samples, tallying 17% to 74% of the original data. Such samples do not correspond to actual user movements, and thus violate the PPDP truthfulness principle (P2 in Sec. 2.2). Even worse, fabricated samples do not help in attaining a sufficient level of accuracy in the anonymized data: the mean error introduced by the perturbations in W4M-LC is between 2-3 km (citywide datasets, $k = 2$) and 19-20 km (countrywide datasets, $k = 5$) in space, and between 16 hours ($k = 2$) and more that two days ($k = 5$) in time. The result is hardly exploitable for data analysis purposes. GLOVE yields a much higher average precision, around 1 km and 1 hour in all $k = 2$ cases, and around 1 km (citywide) to 5 km (countrywide) and 3 hours when $k = 5$. Moreover, these figures are obtained at an affordable cost (in the range 4%-17%) in terms of sample suppression.

We believe that the poor result of W4M-LC is due to the nature of the data. The technique was designed for the anonymization of trajectories sampled at frequencies that are high and similar for all moving objects. This is the case of, e.g., the GPS logs considered in the original performance evaluation of W4M-LC [17]. Instead, mobile traffic dataset contain trajectories whose sampling is very heterogeneous and typically sparse. In this context, the dedicated solution provided by GLOVE grants superior performance.

7.3 Generality analysis

Our evaluation is dependent on the data we use. We thus derive results that let us speculate on how the analysis would generalize to other datasets.

First, we investigate how the timespan of a mobile

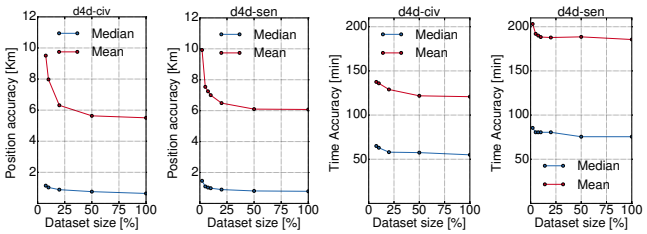


Figure 11: Spatiotemporal accuracy in user-subsets of `d4d-civ` and `d4d-sen`, 2-anonymized with GLOVE.

fingerprint dataset affects its k-anonymized version accuracy. To that end, we extract datasets of different duration, from one day to two weeks, from the original `d4d-civ` and `d4d-sen` datasets. Fig. 10 shows their spatiotemporal accuracy, upon 2-anonymization with GLOVE. We observe that shorter datasets yield a higher accuracy, both in space and time, once they have been anonymized. This is not surprising, since a lower dataset timespan reduces the length of mobile fingerprints, which then become easier to match to 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, all curves are not linear, and the loss of accuracy seems to decrease as datasets become longer; this effect is more evident for the median (typical users) than for the average (hard-to-anonymize individuals). We hypothesize that the result is due to the weekly periodicity known to drive human activities: long datasets spanning over multiple weeks may not be much harder to anonymize than a one-week dataset, as most of the diversity among mobile fingerprints is already present in the latter.

As a second test, we evaluated the impact of a diverse spatial extension of the datasets, using the `abidjan` and `dakar` subsets introduced before. Anonymizing such datasets with GLOVE yields a spatiotemporal accuracy similar to that obtained in the nationwide datasets, as exemplified by the values in Tab. 2. The fact that datasets featuring very different geographical coverage are similarly anonymized is explained by the locality of human activities: the median and average radius of gyration of users are 1.8 km and 12 km in `d4d-civ`, and 2 km and 10 km in `d4d-sen`. Thus, the mobile fingerprints of most individuals are confined to a limited geographical region the size of a city, and they are typically hidden among those of other users in the same area.

Finally, an interesting question is to which extent reducing the number of users in a dataset makes them more distinguishable. Fig. 11 shows how the spatiotemporal accuracy varies when considering datasets that comprise from 5% to 100% of the subscribers in `d4d-civ` and `d4d-sen`. Clearly, datasets with a lower number of users tend to be harder to anonymize, since the crowd that one can leverage to hide himself becomes thinner. However, the effect is only remarkable when retaining a rather low user fraction. We conclude that the anonymizability of our datasets is impaired only when the number of users falls below a few tens of thousands.

8. 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 [1, 2]. It also differs from confidentiality problems in other types of databases extracted from mobile traffic data, e.g., networks of subscriber relationships [18] or aggregate mobile demands at the access network [19]. The related techniques are designed for database formats and semantics that are completely different from ours.

Within the domain of movement micro-data, we focus on the anonymization of spatiotemporal trajectories, which are are not to be confounded with other kinds of movement micro-data: (i) location-based services (LBS), and (ii) spatial trajectories.

In LBS, the goal is anonymizing georeferenced queries, i.e., individual spatiotemporal points. This is done via temporal [20], spatial [21], or personalized spatiotemporal generalization [22]; encryption is also an option [23]. In all cases, hiding individual samples is a subset of the problem we tackle, which concerns instead complete spatiotemporal trajectories. Even in presence of linked queries (i.e., sequences of points that are subject to tracking), solutions rely on pseudo-identifier replacement [24]. While changing the user pseudo-identifier at will does not impair LBS operation, it is not an option in our case. First, it would disrupt the integrity of trajectories and thus the utility of the dataset. Second, it would not solve to the uniqueness problem we tackle.

Spatial trajectories are instead sequences of geographical locations without any temporal reference. They represent routes traveled by users, regardless of when they do so. Solutions for the anonymization of spatial trajectories tend to rely on spatial generalization [25]; however, they do not preserve the temporal dimension of the data, which is paramount in our context.

Among the works that addressed the anonymization of spatiotemporal trajectories, ours is the first to tackle the problem of Privacy-Preserving Data Publishing of mobile traffic datasets. Previous research has relied on techniques that do not meet the requirements we established in Sec. 2.2, namely the replacement of pseudo-identifiers mentioned above [15, 26], or perturbations and permutations that displace users at locations they may have never actually visited [17, 27, 28].

Even when they employ PPDP-compliant approaches, e.g., suppression [29] and generalization [30], past solutions do not fit our needs. Some only work on very short spatiotemporal trajectories of three samples each [29], whereas mobile fingerprints typically include hundreds of samples per week. Others are intended for datasets where the positions of all users are sampled with identical periodicity, e.g., via GPS logging [27, 28, 30]: in that case, the anonymization process only concerns the spatial dimension, since temporal hiding is implicit.

The only approach proposed to date that is capable of handling the anonymization of full-length finger-

prints along both space and time dimensions is Wait for Me (W4M) [17]. Although it creates fabricated samples that do not match the real-world user mobility and thus it does not fulfill the truthfulness principle of PPDP (P2 in Sec. 2.2), we picked W4M as a benchmark for our comparative analysis in Sec. 7.2.

9. CONCLUSIONS

We presented GLOVE, an algorithm for the k-anonymization of movement micro-data extracted from mobile traffic. Its design builds on novel insights into the nature of mobile fingerprint anonymizability. GLOVE attains complete k-anonymization of two reference datasets while preserving a level accuracy that would not even grant partial 2-anonymization under legacy spatiotemporal generalization. Also, it outperforms existing solutions for k-anonymization of movement micro-data, intended for dense spatiotemporal trajectories.

Overall, the research presented in this paper advances the current state of the art in best practices for the anonymization of mobile traffic datasets. However, our work also has limitations that open the way to future improvements. First, the privacy model behind the design of GLOVE is effective in countering a subset of the possible attacks on mobile traffic data: it thus represents a first step towards complete PPDP. Also, it may not fit some data analyses, e.g., targeting outlier behavior detection. Second, GLOVE was shown to grant at most 5-anonymity while maintaining a significant level of data granularity in the reference datasets. Although additional tests with other datasets may affect this bound, the result suggests that higher levels of privacy may require the use of suppression or the introduction of assumptions on adversary knowledge.

As a concluding remark, we recall that in this work we address the problem of uniqueness in mobile users' trajectories. Uniqueness does not implies direct re-identifiability of mobile users, and thus we do not de-anonymize any subscriber in the datasets we study.

10. REFERENCES

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

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

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

[4] D. Naboulsi, M. Fiore, R. Stanica, S. Ribot, *"Large-scale Mobile Traffic Analysis: a Survey,"* IEEE Communications Surveys and Tutorials, to appear.

[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] 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.

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

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

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

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

[13] 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].

[14] D. Hoaglin, F. Mosteller, J.W. Tukey *"Understanding robust and exploratory data analysis,"* Wiley, 1983.

[15] C. Bettini, X.S. Wang, S. Jajodia, *"Protecting Privacy Against Location-Based Personal Identification,"* SDM, 2005.

[16] C. Iovan, A.-M. Olteanu-Raimond, T. Couronne, Z. Smoreda, *"Moving and Calling,"* Geographic Information Science at the Heart of Europe, D. Vandenbroucke, B. Bucher, J. Crompvoets (editors), Springer, 2013.

[17] O. Abul, F. Bonchi, M. Nanni, *"Anonymization of moving objects databases by clustering and perturbation,"* Information Systems, 35(8):884–910, 2010.

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

[19] G. Acs, C. Castelluccia, *"A case study: privacy preserving release of spatio-temporal density in Paris,"* ACM KDD, 2014.

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

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

[22] B. Gedik, L. Liu, *"Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms,"* IEEE Transactions on Mobile Computing 7(1):1–18, 2008.

[23] M. Herrmann, A. Rial, C. Diaz, B. Preneel, *"Practical privacy-preserving location-sharing based services with aggregate statistics,"* ACM WiSec, 2014.

[24] J. Meyerowitz, R.R. Choudhury, *"Hiding stars with fireworks: location privacy through camouflage,"* ACM MobiCom, 2009.

[25] A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, S. Wrobel *"Movement Data Anonymity through Generalization,"* Transactions on Data Privacy 3(2):91–121, 2010.

[26] Y. Song, D. Dahlmeier, S. Bressan, *"Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data,"* PIR, 2014.

[27] O. Abul, F. Bonchi, M. Nanni, *"Never walk alone: Uncertainty for anonymity in moving objects databases,"* IEEE ICDE, 2008.

[28] J. Domingo-Ferrer, R. Trujillo-Rasúa *"Microaggregation- and permutation-based anonymization of movement data,"* Information Science, 208:55–80, 2012.

[29] B.C.M. Fung, M. Cao, B.C. Desai, H. Xu, *"Privacy protection for RFID data,"* ACM SAC, 2009.

[30] 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.