Anonymizing Trajectory Data: Limitations and Opportunities

Patricia Guerra-Balboa[1], **Àlex Miranda Pascual**[1,2], **Javier Parra-Arnau**[1,2], **Jordi Forné**[2], **Thorsten Strufe**[1]

[1]KASTEL Security Research Labs, Karlsruhe Institute of Technology
[2]Dept. of Network Engineering, Universitat Politècnica de Catalunya
patricia.balboa@kit.edu, alex.miranda.pascual@upc.edu, javier.parra-arnau@kit.edu, jordi.forne@upc.edu, strufe@kit.edu

Abstract

A variety of conditions and limiting properties complicate the anonymization of trajectory data, since they are sequential, high-dimensional, bound to geophysical restrictions and easily mapped to semantic points of interest and regions with known properties like suburban neighborhoods, industrial areas or city-centers. Learning the places where one has been is extremely privacy-invasive. However, analyzing real trajectories holds numerous promises, ranging from better informed traffic management, to location recommendations or computational social science, infrastructure and even urban development planning.

The aim of this paper is to establish various challenges, stemming from ideas and also limitations of existing proposals for the anonymization of trajectories, and subsequently identify research opportunities. Keeping both utility and privacy challenges prominent, we sketch the way towards establishing a useful research framework and propose possible research venues towards privacy-preserving trajectory publication.

Introduction

Everyday, the value and interest of location and trajectory data are becoming more and more noticeable not only in our lives but also among data-analytics companies. At the same time, the ability of personal devices (e.g., wearables, smartphones) and navigation systems to accurately collect, process, and analyze these data is growing at an unprecedented rate thanks to recent technological advances. Traffic management, urban planning, transportation systems design, routing advice, or homeland security are just a few of the many applications relying nowadays on trajectory analyses (Gangadharan 2013).

Despite the economic and societal good that comes from data analytics in general, raising tensions exist with the perceived risks to individuals' privacy (Tarnoff 2018; Ovide 2020). To deal with these tensions, current legal frameworks in Europe and other regions limit the collection, processing and sharing of personal data. The European General Data Protection Regulation (GDPR) indeed requires the development of methods to anonymize those personal data as one way to circumvent processing restrictions. Facilitating privacy-preserving analysis of location trajectories therefore is, not only a scientific challenge, but also a legal requirement.

A trajectory database is one where each record is a trajectory, that is to say, a sequence of timestamped locations (such as GPS coordinates). However, as we shall describe in the following sections, anonymizing this type of databases is no easy task. Well-known metrics and techniques in the field of data privacy, such as k-anonymity (Samarati and Sweeney 1998) or ϵ-differential privacy (ϵ-DP) (Dwork 2006), either are not immediately applicable to sequential and high-dimensional databases, or the guarantees of privacy they could provide are not clear. Besides, each data point in a trajectory strongly depends on its predecessor by the natural properties of motion; and this correlation poses important challenges to the implementation of privacy mechanisms.

Likewise, a profound understanding of the particularities of trajectory data is vital for their protection. Although at first sight it may appear they are innocuous to user privacy, trajectories may reveal accurate behavioral patterns, in terms of when and for how long a particular individual does what, allowing an attacker to infer circumstances and trends affecting sensitive aspects of an individual's life, including health status, religious beliefs, social relationships, or sexual preferences.

To make matters worse, trajectories always have a spatio-temporal context. With publicly available information such as street maps and the maximum velocity of a transport means, or with background knowledge about data subjects (such as their place of residence or work), adversaries can improve their attacks against privacy algorithms (Dai et al. 2020; Yang et al. 2020). In this sense, research shows that knowing only four spatio-temporal points is enough to uniquely identify 95% of the individuals (De Montjoye et al. 2013).

Background knowledge can also identify false or impossible trajectories. Adding noise naïvely to protect a trajectory may thus fail to protect the privacy of the respective individual. Furthermore, certain noisy coordinates may be *unreachable*[1], or *geo-spatially incoherent*, if they include impossible locations (e.g., driving through a building or lake). Also, new

[1]*Reachability* refers to the property of real trajectories that movement between two consecutive locations is attainable in the time given (Domingo-Ferrer and Trujillo-Rasua 2012). It depends on the movement speed of the users.

noisy trajectories could be "semantically" identical: for example, regardless of your position on a road or the specific parking spot you occupy in a parking lot, the sensitive information is that you are driving on that road or parked in that parking lot. Therefore, any coordinate disturbance that does not change the semantic property of your position does not provide any kind of privacy.

To top it off, numerous applications of trajectories-data analyses involve repeated computations, since the goal is typically one of monitoring, e.g., of traffic conditions. However, regularly publishing updated versions of an underlying database that are useful is a highly challenging task. The main reason is that one needs to ensure that the combination of information from any already anonymized data does not compromise individuals' privacy. On the one hand, syntactic anonymization methods cannot ensure this in a practical, real-world scenario, where multiple data controllers are very likely to anonymize data independently. On the other hand, although DP can take advantage of a composition property to preserve (to a limited extent) the privacy guarantee after repeated data releases, it sadly is at the cost of a significant degradation in data utility (Bambauer, Muralidhar, and Sarathy 2013; Fredrikson et al. 2014). This important limitation is typically addressed with unreasonably large values of ϵ, which unfortunately may vanish any expectation of privacy for individuals (Domingo-Ferrer, Sánchez, and Blanco-Justicia 2021; Ruggles et al. 2019).

All issues above raise serious concerns about the current state of the art for trajectory anonymization, particularly about whether existing technology can effectively guarantee individuals' privacy and strike an acceptable balance between disclosure risk and utility.

In this paper, we examine the state of the art on privacy-preserving trajectory publication, where the goal is to publish a database with personal trajectories or statistics thereof, while ensuring certain privacy and utility guarantees. Our analysis of current anonymization technology covers syntactic and semantic notions of privacy and is organized into metrics of privacy and utility and anonymization mechanisms. Based on this analysis, we establish various challenges, stemming from ideas and also limitations of existing proposals, for the anonymization of trajectories, and identify opportunities for future research.

The rest of the paper is organized as follows. First, we overview the state of the art on trajectory data anonymization. After that, we elaborate on the limitations and problems identified in our analysis of the literature. Finally, we envisage opportunities and solutions, and draw some concluding remarks.

Trajectories and Data Sets

There are a few types of trajectories that are used in trajectory privacy. The simplest of them, containing the basic structure present in the rest, are called *raw trajectories* and consist of an ordered sequence of spatio-temporal points $T = (x_1, y_1, t_1) \rightarrow \cdots \rightarrow (x_n, y_n, t_n)$. More complex representations called *semantic trajectories* exists, where the location evolves from a simple coordinate point to a point-of-interest (POI), which are provided with semantic meaning, such as a name and description, and possibly other information such as number of visitors or opening hours. Even more complex trajectories exist, called *multiple aspect trajectories* (Mello et al. 2019), which additionally consider any possible type of recordable information, like weather variations, transportation mode, or the heart rate or emotions of individuals.

Trajectory databases consist of multiple trajectories from different individuals (or moving objects) over a common region. Notable differences between them exist. Some data models consist of trajectories of equal length, including some of which are additionally uniformly distributed in time (i.e., every trajectory has a spatio-temporal point for every x minutes) (Hua, Gao, and Zhong 2015); while others are less regular, with spatio-temporal points only appearing when the user arrives (or stays) at a notable location (Cao and Yoshikawa 2015). We note that each of them comes with their own limitations and opportunities.

Measuring Privacy and Utility

Privacy metrics. There exist two well-known families of privacy notions in the field of statistical disclosure control (Hundepool et al. 2012), namely, syntactic and semantic notions (Clifton and Tassa 2013).

In the syntactic case, k-anonymity (Samarati and Sweeney 1998) and its extensions (such as l-diversity (Machanavajjhala et al. 2007) and t-closeness (Li, Li, and Venkatasubramanian 2007)) are classical representatives in the field. Several attempts have been made to translate or adapt these notions for trajectory data. In (Abul, Bonchi, and Nanni 2008), for example, a data set is said to satisfy (k, δ)-anonymity if, for any trajectory, there exists $k-1$ other trajectories such that at every timestep the corresponding locations are no more than $\delta/2$ away, allowing us to place these k trajectories in a "cylinder" of radius δ. Likewise, Poulis et al. (2014) proposes adapting k^m-anonymity as follows: a data set is k^m-anonymous if every continuous subtrajectory of length at most m is contained in at least k trajectories. Gramaglia et al. (2017) also extend this notion to introduce $k^{\tau,\epsilon}$-anonymity, where the authors consider the maximum additional knowledge that the attacker is allowed to learn. Similarly, $(K, C)_L$-privacy (Chen et al. 2013) implies that the adversary cannot distinguish the victim's trajectory, of which they know at most L locations, from $K-1$ other records, with a confidence in the inference better than C.

In the semantic case, DP is probably the best-known notion. Although it was originally proposed to protect the outcomes of queries to a static database (i.e., interactive setting)[2], mechanisms to publish or generate DP static data sets (i.e., non-interactive setting) appeared soon after. Few recent works, however, have tackled the problem of publishing protected versions of a *dynamic* database with DP guarantees, when the publication strategy is to release all available data or a synopsis thereof (e.g., histograms) at regular time in-

[2]The assumption is that an anonymization mechanism sits between the user submitting queries and the database answering them.

stants (Dwork et al. 2010; Li et al. 2015; Chen, Shen, and Jin 2015), or to protect only new or updated data at a given release time (Fan and Xiong 2014; Chen et al. 2017; Leal et al. 2018; Fioretto and Hentenryck 2019). The main obstacle one encounters when protecting dynamic data of this kind, especially when the goal is to publish the database itself (Leal et al. 2018) (rather than statistics derived from it), is that the privacy budget will be consumed completely at some time instant by the sequential composition property of DP. This means that the level of protection is obviously finite in time, and that making the data useful is a challenging task: the larger the number of releases, the smaller the ϵ assigned to each of them and therefore the more noise needs to be added.

To partly mitigate this problem, some of the works mentioned above and others have relied on alternative definitions of DP like *event-level* privacy (Dwork 2008, 2010), which differs from *user-level* privacy. In the field of trajectory data, the latter means to protect the whole trajectory history of any user, whilst the former protects a single location point (i.e., the event). Kellaris et al. (2014) introduces a balance of both notions called w*-event privacy*, which instead protects windows of w sequential events (see also Figure 1, left).

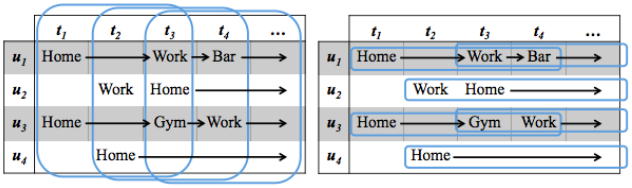

Figure 1: Left table illustrates a privacy model of w-event, while right shows that of ℓ-trajectory privacy ($w = \ell = 3$). In both tables, each column represents the trajectory of a user u_j, where locations are non-uniformly distributed through time.

As an improvement over w-event privacy, Cao and Yoshikawa (2015) reformulate the notion of ϵ-DP for trajectory data, obtaining ℓ*-trajectory privacy*. This model is slightly based on user-level privacy (see Figure 1, right). The authors define a data stream as $S_t = \{D_1, \ldots, D_t\}$, where D_i corresponds to the data set at the timestamp i and t denotes the current time. Two streams are said to be ℓ-neighbouring if they are completely equal except in a single continuous subtrajectory of length ℓ, where each location must differ. A mechanism Λ that takes as input prefixes of data streams S_t is said to be ℓ*-trajectory private* if for any ℓ-neighbouring data streams S_t and S'_t,

$$\Pr[\Lambda(S_t) = N_t] \leq \mathrm{e}^{\epsilon} \cdot \Pr[\Lambda(S'_t) = N_t]$$

for any possible output data stream N_t.

Utility metrics. A variety of metrics have been proposed to quantify the utility of anonymized trajectories.

In the special case of raw trajectories, we find some common general metrics. For example, numerous authors rely on statistical measures, like the similarity of distributions regarding *trajectory length* and *most visited places* (Luca et al. 2020). *Clustering-based methods*, such as (Hua, Gao, and Zhong 2015; Chen et al. 2020; Li et al. 2017), and others often build upon the *Hausdorff distance*, which basically provides a (pessimistic) notion of a distance between trajectories, based on the Euclidean distance between the corresponding locations.

Examples of more specific metrics, closer to actual applications of trajectory analytics, comprise *frequent sequential patterns mining* (Chen, Acs, and Castelluccia 2012), which looks at the difference of the most frequent sequential patterns between raw and sanitized data, and *count query* (Abul, Bonchi, and Nanni 2008; Chen, Acs, and Castelluccia 2012; Chen, Fung, and Desai 2011; Wang et al. 2021c). The latter assesses the utility of a count query Q through the error measure shown in Eq. (1), where $Q(D)$ denotes the number of occurrences of the sequence Q in the database D.

$$\mathrm{error}(Q(\hat{D})) = \frac{|Q(\hat{D}) - Q(D)|}{\max\{Q(D), Q(\hat{D})\}}. \quad (1)$$

In an entirely analogous manner, *spatio-temporal range queries* (Hua, Gao, and Zhong 2015) compute an error measure similar to Eq. (1). However, in this case, the query counts all points at a specific space region over all trajectories in a given time interval.

A utility metric that tackles semantic trajectories is proposed in (Cunningham et al. 2021). Here, each location is represented in three dimensions, namely, spatial, temporal and categorical. Accordingly, each dimension is assigned an associated distance function: time and physical distance for the spatial and temporal dimensions, and the authors define a categorical distance function as the difference in the semantic meaning between two locations using a hierarchy (in a logical way: e.g., $d_c(\text{Bar}, \text{Restaurant}) < d_c(\text{Bar}, \text{Church})$). The authors then use this distance to define *preservation range queries*, which output the percentage of locations in a data set at a distance no greater than δ from the corresponding permuted location. The authors also investigate another utility metric that quantifies *hotspot*[3] *preservation*, by computing the spatio-temporal distance between the hotspots' location in the original and the sanitized data sets.

Mechanisms: Achieving Privacy

In this section, we examine the most relevant mechanisms that aim to enforce the aforementioned privacy notions in the context of trajectory anonymization. We first describe those mechanisms providing syntactic guarantees, to afterwards cover those ensuring semantic notions.

Syntactic privacy. There exist three main general anonymization techniques to enforce syntactic privacy (Tortelli Portela, Vicenzi, and Bogorny 2019): *suppression*, the removal of those location samples or entire trajectories that cause privacy issues; *generalization*, making records indistinguishable from others by reducing

[3] *Hotspots* are defined as location and time range during which the location is visited by a large number of users, e.g., a stadium during a match, or a train station at rush hour.

the trajectories' precision or by grouping samples into larger ranges; and *(perturbative) masking*, which comprises a multitude of techniques including data *perturbation*, based on noise addition, location *merging* or *clustering*, or the creation of new entries either by *dummy generation* or probabilistic *condensation*, to name a few. Suppression and generalization techniques are also grouped into *non-perturbative masking* (Hundepool et al. 2012), since they preserve the truthfulness of data without distorting it, albeit losing information.

The vast majority of anonymization technology combines several of the general techniques mentioned above. Next, we succinctly describe the most relevant works.

Abul, Bonchi, and Nanni (2008) implement the method of Never Walk Alone (NWA), which uses data clustering and spatial translation to ensure the output database is (k, δ)-anonymous. Other methods that also use some type of masking include a location permutation method described (Domingo-Ferrer and Trujillo-Rasua 2012). This method clusters trajectories using microaggregation[4] and then permutes the locations by sensitive-attribute generalization and local suppression. Dai et al. (2018) extends the idea to additionally consider the semantic dimension. Poulis et al. (2014) proposes methods where locations are merged into pairs until k^m-anonymity is satisfied, one by merging the nearest locations first, and another where the semantic similarity of locations is taken into account.

Grouping similar trajectories and removing some of them to ensure k-anonymity are also frequent methods used in the literature, such as in (Pensa et al. 2008; Dong and Pi 2018). In the former, authors construct a prefix trees and prune subsequences which fail to achieve the k threshold in counts; whilst in the latter, authors introduce a method that studies the frequency of subtrajectories and removes the infrequent ones, grouping the rest into representatives to ensure k-anonymity.

In (Nergiz et al. 2009), the authors anonymize trajectories, first by ensuring k-anonymity via suppression and generalization techniques. Here specific points are replaced by cell grids. They also introduce a way to return to the original domain by randomly reconstructing representations from the original data set. Another method based on generalization and suppression appears in (Monreale et al. 2010), classifying locations into areas, with (Monreale et al. 2011) serving as an extension of it, where they generalize locations respecting their semantic meaning using a notion similar to l-diversity.

Finally, Chen et al. (2013) define a method based on local suppression, which removes only some instances from the data set so to ensure $(K, C)_L$-privacy and preserve instances of spatio-temporal points and frequent sequences in the trajectory data.

Semantic privacy. Next, we examine anonymization algorithms that, either publish trajectory databases satisfying DP, or provide statistics (e.g., counts of similar trajectories) about the underlying database with DP guarantees.

[4]*Microaggregation* is a type of clustering with a bound to the number of elements in each cell.

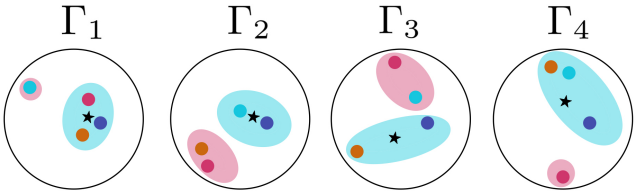

Figure 2: Example of how trajectory data are anonymized through clustering techniques. Different trajectories are represented in different colours, with points corresponding to the physical location over each time step. The colored areas represent the groups defined by the selected partition, and stars denote the centroids of each subset. In this example, trajectories are of length $|T| = 4$ and the selected partition contains $m = 2$ subsets.

Noisy counts is a common approach based on the Laplacian mechanism. An archetypal example is Chen, Fung, and Desai (2011); Chen, Acs, and Castelluccia (2012), which relies on variable n-gram models of trajectories, to add Laplacian noise to the counts of the most representative n-grams before release. Additionally, in (Chen, Acs, and Castelluccia 2012) a method to generate synthetic data from released n-grams is proposed.

Clustering-based mechanisms[5] constitute another approach used in trajectory privacy (Hua, Gao, and Zhong 2015; Li et al. 2017; Chen et al. 2013). The idea is to merge concurrent locations from different trajectories following a probabilistic partitioning based on the exponential mechanism. More specifically, the authors suggest a score function to measure distances between trajectories crossing locations at each time step. Using the exponential mechanism and this score function, they choose one of the candidate partitions (into m groups) of Γ_i, the universe of locations in the trajectory database at time t_i. Finally, all the locations of each subset are clustered together and replaced by their corresponding centroid (see Figure 2).

After selecting a partition and replacing the actual locations by centroids, the original location universe in each time step Γ_i is replaced by a smaller one, $\tilde{\Gamma}_i$, which contains perturbed information. The new trajectories are constructed from this reduced universe $\tilde{\Gamma}_i$. In addition to this process, the authors propose several DP release methods, which usually add trajectories drawn from $\tilde{\Gamma}_i$ at random, with its count attributed following the Laplacian mechanism, until obtaining a sanitized database of the same size as the original. The resulting data set is claimed to meet the privacy guarantees.

On the other hand, Cunningham et al. (2021) introduce a mechanism for *perturbing semantic trajectories* that satisfies ϵ-local differential privacy (ϵ-LDP) (Kasiviswanathan et al. 2011). The authors also find a way of implementing public knowledge into the privacy mechanism to improve its utility without affecting the privacy budget ϵ. They use

[5]Some authors also refer to this mechanism as *generalization*, but it is not used in this paper to avoid confusion with the previously-defined generalization technique.

this public knowledge to partition the set of all POIs into spatio-temporal-categorical regions, such that each one contains some number of POIs, whilst still preserving hotspots. Essentially, the mechanism can be divided into four parts: first, it generalizes every location into the corresponding region; it partitions these new trajectories into n-grams which are then perturbed, following the exponential mechanism, to ensure ϵ-LDP; then a reconstruction of the trajectory is done by minimizing a distance function defined over the three dimensions; and finally the mechanism returns to the initial domain by randomly picking a location in each section, making sure that consecutive locations in a trajectory are reachable in the corresponding time.

Limitations in Privacy Guarantees

Although syntactic notions can in general provide high utility data, they are susceptible to various well-known attacks (e.g., *background knowledge* or *attribute-linkage attacks*). This, together with the fact that they are not composable (Soria-Comas and Domingo-Ferrer 2016), limit the application of syntactic technology to continuously protect trajectory data. In the remainder of the section, we first present some works emphasizing this limitation, and then review proposals relying on semantic protection.

Trujillo-Rasua and Domingo-Ferrer (2013) analyze NWA (Abul, Bonchi, and Nanni 2008) and show that is not able to protect a new trajectory added to a set of already k-indistinguishable trajectories. Likewise, Chen et al. (2013) state that a data set with high sequential correlation cannot be properly protected with simple k-anonymity methods.

In an experiment comparing three ways of noise addition for trajectory data, Jiang et al. (2013) conclude that adding noise to each position is better than adding it to each spatial coordinate or to the whole trajectory. Also, models that generalize only one of the trajectories' dimensions, such as (Abul, Bonchi, and Nanni 2008; Monreale et al. 2011; Dong and Pi 2018), are susceptible to attacks on the other dimensions, as these still hold sensitive information, and should be avoided.

Even if DP was presented as a strong privacy guarantee, the exact notion of privacy it provides is sometimes not clear, as explained in (Lee and Clifton 2011). Moreover, several papers (Cao et al. 2017; Wang et al. 2021a; Yang, Sato, and Nakagawa 2015; Abowd et al. 2021; Xiao and Xiong 2015) have set the weakness of this notion when correlations among attributes in the database are notorious. Unfortunately, in trajectory data, locations visited by individuals are correlated through its movement's nature (physical laws of motions, speed restrictions of roads, usual human behaviour, etc.) (De Montjoye et al. 2013). As a consequence, a privacy violation could occur even if DP mechanisms are applied.

An analysis (Cao and Yoshikawa 2015) of DP notions tailored for trajectory data suggests that event-level privacy is not safe because of the vulnerable nature of this type of data with respect to background knowledge attacks. The authors of this work also claim that w-event privacy fails in trajectory protection since user's trajectories are sparse and are not uniformly distributed over the timeline, and might therefore not fit in the window.

While previous limitations focus on rather general aspects, the remainder of this subsection will analyze deficiencies in specific proposals for DP protection.

In *clustering-based methods*, a first problem arises when we consider speed limitations in cities, where the probabilistic merging of locations can likely produce impossible trajectories in real-world situations, i.e., trajectories such that two consecutive locations are unreachable in the given time; naturally, this may provide adversaries with new venues for attack. Besides, all proposals (Hua, Gao, and Zhong 2015; Li et al. 2017; Chen et al. 2013) building on this clustering method seem to be flawed. The apparent defect to our eyes is the following: the release of the centroids alone (i.e., without connecting them to form a trajectory) at each time step is a process that the authors ensure to guarantee $\epsilon|T|$-DP through the application of exponential mechanism at each time step and the sequential composition property. Here, $|T|$ denotes the length of a trajectory T (which is supposed to be constant in the database). However, the authors aim to publish trajectories, naturally by linking the centroids. But, to do so, they use the unprotected, original data to find the proper sequence. Unfortunately, a data-dependent operation of this kind violates DP.

Chen, Fung, and Desai (2011) and Chen, Acs, and Castelluccia (2012) define *noisy counts performances* and investigate whether adding noise to the frequency counts of n-grams really provides privacy. Since the released sequences are the original ones and there is no sampling or other process in-between, a certain specific sequence could identify an individual, even if the frequency of these sequences are modified. We notice that, even though the mechanism might be effective against *re-identification attacks*, an attacker could indeed discover other locations visited by the individual. For instance, if any appearance of the sequence $p_1 \rightarrow p_2$ in the data set is always followed (or preceded) by the spatio-temporal point p_3, then an attacker could learn that any user that visits p_1 and p_2, must surely also visit p_3, although they would not be able to identify the user's entire trajectory. Similarly, an analogous probabilistic attack can be created if a high percentage of sequences $p_1 \rightarrow p_2$ are followed by p_3.

Limitations in Utility Guarantees

This section describes several limitations affecting data utility. We proceed next with general limitations, and then with more specific ones related to metrics and evaluation methodologies.

General problems. We first highlight some general issues that are inherent to the nature of trajectory data. Chen et al. (2013) point out some of them that imply a significant utility loss when applying k-anonymity-based methods. Data sets with sparse or short trajectories pose a great challenge for these anonymization methods, since trajectories can have little overlap, which leads to an unavoidable data and utility loss. Similarly, for semantic notions, sparseness produces high sensitivities and also triggers the need

of adding more noise to achieve the same privacy budget. An example of these problems are seen later when we approach *noisy counts*. Also, generalization methods could be inefficient for high-dimensional databases, due to the *curse of dimensionality* (Aggarwal 2005).

Another general problem appears in anonymization methods that are based on perturbative masking. Gramaglia et al. (2017) states that to preserve truthfulness of data, it cannot rely on randomized, perturbed, permuted or synthetic data, since the addition of fictitious data introduces unpredictable biases in the final sanitized data sets. Furthermore, this type of mechanisms can also lead to the creation of impossible trajectories, with unreachable locations or geo-spatial inconsistencies. For example, the *clustering-based methods* used in (Hua, Gao, and Zhong 2015; Li et al. 2017; Chen et al. 2013) can yield new locations which might be illogical, such as coordinates on top of buildings or rivers.

Metrics. In the first instance, the *Hausdorff distance* might seem very appropriate to quantify the utility of an anonymized trajectory database, as it is a well-established metric for distance measurements. However, its main limitation is that it does not take into consideration the temporal dimension. This way, two trajectories over the same physical route, but with time variations, are considered to be the same trajectory under this distance, which clearly may lead to a huge utility loss. If we take a look at the possible applications that motivate the need for private release of trajectory data, as for instance traffic jams prediction, we can anticipate that this metric is likely to hide major problems, especially in terms of flow[6].

The shortcomings of *distribution of length* and *distribution of most visited places*, are even more apparent. We could have very similar length and most-visited-places distributions in two databases that are really different from all other aspects of the trajectories, such as shape or time. It is clear that there are too many possible trajectories of the same length with nothing else in common. Likewise, there are actually different paths that cross through common popular locations, so *distributions of most visited locations* is probably not the most representative metric to make sure algorithms are working correctly.

Lastly, we would like to stress the adequacy of the current usage of *spatio-temporal range queries*, *count query* and *frequent sequential pattern*. Even though these utility metrics are more close to reality applications of trajectory data, privacy designers should be careful when choosing their parameters. For example, if one chooses long radii or time intervals, or if one takes big ranges of K in the top K frequent sequences, the evaluation of how useful is an anonymized database is not going to be representative.

Methodologies. Next, we report deficiencies in methodologies of the state of the art to assess utility.

In general, privacy designers should be careful about the database employed to evaluate utility. One potential issue might result from biased databases with unrealistic data distributions, giving us an erroneous perception that the developed algorithms may provide high-utility data. One example that illustrates this issue is Microsoft's T-driver database[7] (Yuan et al. 2010, 2011), which consists in a database of taxis from Beijing. The problem with this type of database is that the particular and specialized behaviour of taxi drivers may not be representative of the whole population, which probably is going to make long stops at the specific location (e.g., at work or at the grocery store). Another example are public transport databases, such as the Montreal Transit Corporation[8] (STM) database. These types of databases have an abnormally small universe of locations (e.g., bus stops, train stations), and short and frequent stop times. Furthermore, public transport vehicles follow prefixed routes and stops, which are shared between different lines.

Another limitation we have identified in existing methodologies for clustering-based anonymization (Hua, Gao, and Zhong 2015; Li et al. 2017; Chen et al. 2013) is that the score function (of the exponential mechanisms of DP), only depends on distance, and therefore does not take into consideration any measurement regarding time. Sadly, doing so may spoil data accuracy for a wide variety of interesting usage purposes that are heavily dependent on timing and flow.

Similarly, a problem that appears when we put time aside is related with stationary sequences, which can be observed in trajectories with uniform time-stamps, when, for example, a car is stopped at a location. After a driver makes a stop (at work, for shopping, or any other usual task), the spatial location is going to remain the same at each time step until the car starts to move again (e.g., see Figure 2, where the dark blue point is exactly in the same location in each time step, because it represents a stop position in the trajectory). For example, if the car remains stationary from t_i to t_j, we have the following trajectory subsequence:

$$\begin{aligned}(x_0, y_0, t_0) \to \cdots \to (x_i, y_i, t_i) \to (x_i, y_i, t_{i+1}) \to \\ \to (x_i, y_i, t_{i+2}) \to \cdots \to (x_i, y_i, t_j) \to \\ \to (x_{j+1}, y_{j+1}, t_{j+1}) \to \cdots .\end{aligned}$$

Since merging locations probabilistically is only based on distances, the sanitized data will likely not reflect this stop, because the constant spatial point along time (same coordinates along various time steps) is going to be substituted by the corresponding centroids at each time step. In Figure 2 we can see that the locations of the dark blue stationary trajectory are going to be changed into pairwise different locations at each time step. This produces an apparent random movement which removes the stop.

Another issue we have observed is the poor utility that *noisy counts* approaches (Chen, Fung, and Desai 2011; Chen, Acs, and Castelluccia 2012) might offer, which results from implicitly assuming that raw trajectories contain a large number of common prefixes and n-grams. Since

[6] *Traffic flow* is the number of vehicles that cross a certain section of the road per unit of time.

[7] https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/

[8] https://www.stm.info/en/about/developers/available-data-description

the anonymization process adds noises to real counts of n-grams, if the counts of these ones are small, then the noise added to each one will become higher with fatal consequences in utility terms. Unfortunately, real world databases do not follow this assumption very often (i.e., we cannot assume there will be many common n-grams).

Opportunities

In this section, we sketch out possible strands of future research that may overcome some of the deficiencies identified in the previous section. Given the technical limitations of current syntactic notions to protect dynamic data, we focus on anonymization technology providing semantic guarantees.

Usually in data processing, clustering methods are a good option. In our review of the state of the art, we found various problems regarding where those methods were applied, but obviously this does not mean that we have to ruled them out completely. Instead of getting rid of clustering, what we seek is to find a methodology that overcomes the reported limitations. For this purpose, we are interested in an algorithm that takes time into account, and we believe that clustering of entire trajectories, instead at each time instant, could be more useful in terms of utility. Merging entire trajectories would reduce problems such as the temporal-related inconsistencies of the resulting trajectories. It would also solve the problem presented, in terms of privacy, which involves drawing the new trajectories from the clustered locations. Clustering methods that not only depend on physical but also on temporal dimension could provide better utility in flow terms, and help address certain inconveniences like the removal of stops.

In this sense, topological clustering based on persistent homology in a high-dimensional way (including time), could be an interesting approach for two main reasons: its qualitative predominance and its low computational cost (Pokorny, Goldberg, and Kragic 2016; Pokorny, Hawasly, and Ramamoorthy 2016). The main advantage of utilizing this type of clustering would be being able to merge *entire* trajectories according to topological properties, rather than just grouping a set of locations at every time step. However, how to enforce DP and deal with high dimensional data in this context is by no means obvious.

On the other hand, the main problem with enforcing DP in trajectories is that data correlation can violate its privacy guarantee. One possibility to deal with this is to adapt alternative notions of privacy (based on the original idea of DP) like *APL-free ϵ-DP* (Wang et al. 2021b), to a trajectory context. Similarly, some topological models have been proposed to identify privacy issues using *relations*, *Dowker simplicial complexes* and *lattices* (Erdmann 2017). The adaptation of these models and methodologies to our trajectory backdrop, where attributes will be regarded as locations visited by users, seems a great opportunity to explore.

On the other hand, as mentioned in the previous section, (Cunningham et al. 2021) prevents the publication of impossible trajectories in perturbation-based mechanisms, by detecting reachability violations and re-anonymizing any impossible trajectory. This way, the authors ensure what the sanitized database consists of well-defined trajectories. Because the algorithm in question only uses public knowledge (e.g., maximum movement velocity), it does not rely on any data-dependent operation to perform all this processing and hence it does not consume any privacy budget. Entirely analogous algorithms could of course be incorporated into any perturbative mechanism like this to ensure that all trajectories are well-defined. This could allow us to address geospatial inconsistencies, which would also translate into data sets of higher utility.

Just to finish it off, we would like to mention certain metrics of utility that, depending on the privacy designer's needs, could be more effective to evaluate future algorithms and results. This includes a good use of the state of the art, incorporating metrics of different context as, for instance, flow measurement, where we find some accuracy metrics extracted from non-privacy specific literature (Luca et al. 2020), such as *RMSE*, *MAE*, *MAPE*, *MSE*, and *CPC*. Also, it could be useful to explore temporal metrics based on waiting times and circadian rhythms, and temporal location patterns.

Summary

The first part of this paper has analyzed the state of the art of anonymization technology for trajectory data. We have examined how these data are commonly represented and which aspects they may capture; and reviewed the most relevant metrics and anonymization mechanisms enforcing syntactic and semantic protection. This dissection of current technology has allowed us to delve into the limitations of the current solutions, in terms of the promised privacy guarantees, and the utility left after anonymization. This second part of our work has more specifically identified technical impediments, whereby an important part of the examined technology may not effectively protect individuals' privacy and/or preserve most of the utility of trajectory data, in a continuous data publication scheme.

Acknowledgments

Javier Parra Arnau is the recipient of an Alexander von Humboldt research fellowship. This work also received the support from "la Caixa" Foundation (fellowship code LCF/BQ/PR20/11770009), the European Union's H2020 programme (Marie Skłodowska-Curie grant agreement No 847648), from the Spanish Government under the project "COMPROMISE" (PID2020-113795RB-C31/AEI/10.13039/501100011033), and from the BMBF project "PROPOLIS" (16KIS1393K). The authors at KIT are supported by KASTEL Security Research Labs (Topic 46.23 of the Helmholtz Association) and Germany's Excellence Strategy (EXC 2050/1 'CeTI').

References

Abowd, J.; Ashmead, R.; Cumings-Menon, R.; Garfinkel, S.; Kifer, D.; Leclerc, P.; Sexton, W.; Simpson, A.; Task, C.; and Zhuravlev, P. 2021. An Uncertainty Principle is a Price of Privacy-Preserving Microdata. *arXiv preprint arXiv:2110.13239*.

Abul, O.; Bonchi, F.; and Nanni, M. 2008. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. *2008 IEEE 24th International Conference on Data Engineering*, 376–385.

Aggarwal, C. C. 2005. On k-Anonymity and the Curse of Dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, 901–909. VLDB Endowment. ISBN 1595931546.

Bambauer, J.; Muralidhar, K.; and Sarathy, R. 2013. Fool's Gold: An Illustrated Critique of Differential Privacy. Arizona Legal Studies Discussion Paper No. 13-47, James E. Rogers College of Law, The University of Arizona.

Cao, Y.; and Yoshikawa, M. 2015. Differentially private real-time data release over infinite trajectory streams. In *2015 16th IEEE International Conference on Mobile Data Management*, volume 2, 68–73. IEEE.

Cao, Y.; Yoshikawa, M.; Xiao, Y.; and Xiong, L. 2017. Quantifying differential privacy under temporal correlations. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, 821–832. IEEE.

Chen, R.; Acs, G.; and Castelluccia, C. 2012. Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM conference on Computer and communications security*, 638–649.

Chen, R.; Fung, B.; and Desai, B. C. 2011. Differentially private trajectory data publication. *arXiv preprint arXiv:1112.2020*.

Chen, R.; Fung, B. C.; Mohammed, N.; Desai, B. C.; and Wang, K. 2013. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, 231: 83–97.

Chen, R.; Shen, Y.; and Jin, H. 2015. Private Analysis of Infinite Data Streams via Retroactive Grouping. In *Proc. Int. Conf. Inform., Knowl. Manage. (CIKM)*, 1061–1070. ACM.

Chen, S.; Fu, A.; Shen, J.; Yu, S.; Wang, H.; and Sun, H. 2020. RNN-DP: A new differential privacy scheme base on Recurrent Neural Network for Dynamic trajectory privacy protection. *Journal of Network and Computer Applications*, 168: 102736.

Chen, Y.; Machanavajjhala, A.; Hay, M.; and Miklau, G. 2017. PeGaSus: Data-Adaptive Differentially Private Stream Processing. In *Proc. ACM Conf. Comput., Commun. Secur. (CCS)*, 1375–1388. ACM.

Clifton, C.; and Tassa, T. 2013. On Syntactic Anonymity and Differential Privacy. *Trans. Data Priv.*, 6(2): 161–183.

Cunningham, T.; Cormode, G.; Ferhatosmanoglu, H.; and Srivastava, D. 2021. Real-world trajectory sharing with local differential privacy. *arXiv preprint arXiv:2108.02084*.

Dai, C.; Pi, D.; Becker, S. I.; Wu, J.; Cui, L.; and Johnson, B. 2020. CenEEGs: Valid EEG selection for classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(2): 1–25.

Dai, Y.; Shao, J.; Wei, C.; Zhang, D.; and Shen, H. T. 2018. Personalized Semantic Trajectory Privacy Preservation through Trajectory Reconstruction. *World Wide Web*, 21(4): 875–914.

De Montjoye, Y.-A.; Hidalgo, C. A.; Verleysen, M.; and Blondel, V. D. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1): 1–5.

Domingo-Ferrer, J.; Sánchez, D.; and Blanco-Justicia, A. 2021. The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning). *Commun. ACM*, 64(7): 33—-35.

Domingo-Ferrer, J.; and Trujillo-Rasua, R. 2012. Microaggregation- and permutation-based anonymization of movement data. *Information Sciences*, 208: 55–80.

Dong, Y.; and Pi, D. 2018. Novel Privacy-preserving algorithm based on frequent path for trajectory data publishing. *Knowledge-Based Systems*, 148: 55–65.

Dwork, C. 2006. Differential privacy. In *Proc. Int. Colloq. Automata, Lang., Program.*, 1–12. Springer-Verlag.

Dwork, C. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, 1–19. Springer.

Dwork, C. 2010. Differential privacy in new settings. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, 174–183. SIAM.

Dwork, C.; Naor, M.; Pitassi, T.; and Rothblum, G. N. 2010. Differential privacy under continual observation. In *Proc. ACM Int. Symp. Theory Comput. (STOC)*, 715–724. ACM.

Erdmann, M. 2017. Topology of privacy: Lattice structures and information bubbles for inference and obfuscation. *arXiv preprint arXiv:1712.04130*.

Fan, L.; and Xiong, L. 2014. An Adaptive Approach to Real-Time Aggregate Monitoring With Differential Privacy. *IEEE Trans. Knowl. Data Eng.*, 26(9): 2094–2106.

Fioretto, F.; and Hentenryck, P. V. 2019. OptStream: Releasing Time Series Privately. *J. Artif. Intell. Res.*, 65(1): 423–456.

Fredrikson, M.; Lantz, E.; Jha, S.; Lin, S.; Page, D.; and Ristenpart, T. 2014. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, 17–32. USA: USENIX Association.

Gangadharan, S. P. 2013. Big data for the people: it's time to take it back from our tech overlords. Accessed on 2021-01-18.

Gramaglia, M.; Fiore, M.; Tarable, A.; and Banchs, A. 2017. $k^{\tau,\epsilon}$-anonymity: Towards Privacy-Preserving Publishing of Spatiotemporal Trajectory Data. *ArXiv*, abs/1701.02243.

Hua, J.; Gao, Y.; and Zhong, S. 2015. Differentially private publication of general time-serial trajectory data. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, 549–557. IEEE.

Hundepool, A.; Domingo-Ferrer, J.; Franconi, L.; Giessing, S.; Nordholt, E. S.; Spicer, K.; and de Wolf, P.-P. 2012. *Statistical Disclosure Control*. Wiley.

Jiang, K.; Shao, D.; Bressan, S.; Kister, T.; and Tan, K.-L. 2013. Publishing Trajectories with Differential Privacy Guarantees. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*,

SSDBM. New York, NY, USA: Association for Computing Machinery. ISBN 9781450319218.

Kasiviswanathan, S. P.; Lee, H. K.; Nissim, K.; Raskhodnikova, S.; and Smith, A. 2011. What Can We Learn Privately? *(SIAM) J. Comput.*, 40(3): 793–826.

Kellaris, G.; Papadopoulos, S.; Xiao, X.; and Papadias, D. 2014. Differentially Private Event Sequences over Infinite Streams. *Proc. VLDB Endow.*, 7(12): 1155–1166.

Leal, B. C.; Vidal, I. C.; Brito, F. T.; Nobre, J. S.; and Machado, J. C. 2018. δ-DOCA: Achieving privacy in data streams. In *Proc. Int. Workshop Data Priv. Manage. (DPM)*, volume 11025 of *Lecture Notes Comput. Sci. (LNCS)*, 279–295. Barcelona, Spain.

Lee, J.; and Clifton, C. 2011. How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, 325–340. Springer.

Li, H.; Xiong, L.; Jiang, X.; and Liu, J. 2015. Differentially Private Histogram Publication for Dynamic Datasets: An Adaptive Sampling Approach. In *Proc. Int. Conf. Inform., Knowl. Manage. (CIKM)*, 1001–1010. ACM.

Li, M.; Zhu, L.; Zhang, Z.; and Xu, R. 2017. Achieving differential privacy of trajectory data publishing in participatory sensing. *Information Sciences*, 400: 1–13.

Li, N.; Li, T.; and Venkatasubramanian, S. 2007. t-Closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, 106–115. Istanbul, Turkey.

Luca, M.; Barlacchi, G.; Lepri, B.; and Pappalardo, L. 2020. A Survey on Deep Learning for Human Mobility. *arXiv preprint arXiv:2012.02825*.

Machanavajjhala, A.; Kifer, D.; Gehrke, J.; and Venkitasubramaniam, M. 2007. l-Diversity: Privacy beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1): 3–es.

Mello, R.; Bogorny, V.; Alvares, L.; Santana, L.; Ferrero, C.; Frozza, A. A.; Schreiner, G.; and Renso, C. 2019. MASTER: A multiple aspect view on trajectories. *Transactions in GIS*.

Monreale, A.; Andrienko, G.; Andrienko, N.; Giannotti, F.; Pedreschi, D.; Rinzivillo, S.; and Wrobel, S. 2010. Movement Data Anonymity through Generalization. *Transactions on Data Privacy*, 3: 91–121.

Monreale, A.; Trasarti, R.; Pedreschi, D.; Renso, C.; and Bogorny, V. 2011. C-safety: A framework for the anonymization of semantic trajectories. *Transactions on Data Privacy*, 4: 73–101.

Nergiz, M.; Atzori, M.; Saygin, Y.; and Güç, B. 2009. Towards Trajectory Anonymization: A Generalization-Based Approach. *Transactions on Data Privacy*, 2: 47–75.

Ovide, S. 2020. Just Collect Less Data, Period. Accessed on 2021-01-18.

Pensa, R.; Monreale, A.; Pinelli, F.; and Pedreschi, D. 2008. Pattern-Preserving k-Anonymization of Sequences and its Application to Mobility Data Mining. volume 397.

Pokorny, F. T.; Goldberg, K.; and Kragic, D. 2016. Topological trajectory clustering with relative persistent homology. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 16–23. IEEE.

Pokorny, F. T.; Hawasly, M.; and Ramamoorthy, S. 2016. Topological trajectory classification with filtrations of simplicial complexes and persistent homology. *The International Journal of Robotics Research*, 35(1-3): 204–223.

Poulis, G.; Skiadopoulos, S.; Loukides, G.; and Gkoulalas-Divanis, A. 2014. Apriori-based algorithms for k^m-anonymizing trajectory data. *Transactions of data privacy*, 7: 165–194.

Ruggles, S.; Fitch, C.; Magnuson, D.; and Schroeder, J. 2019. Differential Privacy and Census Data: Implications for Social and Economic Research. *AEA Papers and Proceedings*, 109: 403–08.

Samarati, P.; and Sweeney, L. 1998. Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression. Tech. rep., SRI Int.

Soria-Comas, J.; and Domingo-Ferrer, J. 2016. Big Data Privacy: Challenges to Privacy Principles and Models. *Data Sci. Eng.*, 1(1): 21–28.

Tarnoff, B. 2018. Big data for the people: it's time to take it back from our tech overlords. Accessed on 2021-01-18.

Tortelli Portela, T.; Vicenzi, F.; and Bogorny, V. 2019. Trajectory Data Privacy: Research Challenges and Opportunities. In *GEOINFO 2019 conference*.

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

Wang, H.; Xu, Z.; Jia, S.; Xia, Y.; and Zhang, X. 2021a. Why current differential privacy schemes are inapplicable for correlated data publishing? *World Wide Web*, 24: 1–23.

Wang, J.; Li, Z.; Lui, J. C.; and Sun, M. 2021b. Topology-Theoretic Approach To Address Attribute Linkage Attacks In Differential Privacy. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 1–6. IEEE.

Wang, W.; Yang, G.; Bao, L.; Ma, K.; Zhou, H.; and Bai, Y. 2021c. Travel Trajectory Frequent Pattern Mining Based on Differential Privacy Protection. *Wireless Communications and Mobile Computing*, 2021.

Xiao, Y.; and Xiong, L. 2015. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1298–1309.

Yang, B.; Sato, I.; and Nakagawa, H. 2015. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD international conference on Management of Data*, 747–762.

Yang, Y.; Cai, J.; Yang, H.; Zhang, J.; and Zhao, X. 2020. TAD: A trajectory clustering algorithm based on spatial-temporal density analysis. *Expert Systems with Applications*, 139: 112846.

Yuan, J.; Zheng, Y.; Xie, X.; and Sun, G. 2011. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 316–324.

Yuan, J.; Zheng, Y.; Zhang, C.; Xie, W.; Xie, X.; Sun, G.; and Huang, Y. 2010. T-drive: driving directions based on

taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, 99–108.