The Fréchet/Manhattan distance and the trajectory anonymisation problem

Christof Ferreira Torres[1] and Rolando Trujillo-Rasua[1,2]

[1] University of Luxembourg, CSC
[2] Interdisciplinary Centre for Security, Reliability and Trust
rolando.trujillo@uni.lu

Abstract. Mobile communication has grown quickly in the last two decades. Connections can be wirelessly established from almost any habitable place in the earth, leading to a plethora of connection-based tracking mechanisms, such as GPS, GSM, RFID, etc. Trajectories representing the movement of people are consequently being gathered and analysed in a daily basis. However, a trajectory may contain sensitive and private information, which raises the problem of whether spatio-temporal data can be published in a private manner.
In this article, we introduce a novel distance measure for trajectories that captures both aspect of the microaggregation process, namely clustering and obfuscation. Based on this distance measure we propose a trajectory anonymisation heuristic method ensuring that each trajectory is indistinguishable from $k-1$ other trajectories. The proposed distance measure is loosely based on the Fréchet distance, yet it can be computed efficiently in quadratic time complexity. Empirical studies on synthetic trajectories show that our anonymisation approach improves previous work in terms of utility without sacrificing privacy.

1 Introduction

Not long ago, visual identification was the only mean to collect spatio-temporal data from people. Nowadays this task is far easier since there is no need of direct human intervention for monitoring and tracking. Instead, surveillance cameras, social networks, credit card transactions, and many other worldwide adopted technologies and services, automatically collect this type of data. Today's pervasiveness of location-aware devices like mobile phones and GPS receivers helps even further companies and governments to easily collect huge amount of information about people's movement.

Analysing and mining this type of information, also known as trajectories, might reveal new trends and previously unknown knowledge to be used in traffic, sustainable mobility management, urban planning and supply chain management. By doing so, resources might be optimised and business and government decisions can be solid and well-founded. In this sense, both companies and citizens profit directly from the publication and analysis of databases of trajectories.

Despite of all these benefits, there are obvious threats to people's privacy if their movement data are published in a way which allows re-identification of the

person behind a trajectory. Just considering the locations visited by a trajectory, it may reveal sensitive information about users like religious, political, or sexual preferences. The privacy threat grows when the time information exposes user's habits that may be used for unauthorized advertisement and user profiling.

A tentative solution to achieve anonymity is de-identification by means of removing identifying attributes of individuals. However, this is often insufficient to preserve privacy due to other type of attributes called *quasi-identifiers*, which are non-identifying attributes that together with external information might uniquely identify the individual behind a record. Unfortunately, in the case of spatio-temporal data, every location can be regarded as a quasi-identifier [26]. Therefore, just knowing some locations visited by an individual could be enough to identify his trajectory in a database. As an example, let's consider a GPS application recording trajectories of citizens. Daily routine indicates that an early morning trajectory is likely to begin at the user's home and end at the user's workplace. This simple assumption might be enough to accurately re-identify a user's trajectory.

The above problem has been addressed relying on *k-anonymity* [19, 18, 20], a widely used privacy notion. A set S is said to satisfy k-anonymity if each combination of quasi-identifier attribute values is shared by at least k records in S. Therefore, considering that all identifying attributes have been removed, k-anonymity ensures that no anonymised record can be correctly linked to an individual with probability higher than $1/k$. In microdata, the set of quasi-identifiers is typically considered small and known in advance. In spatio-temporal data, however, a similar assumption can hardly hold; any location can be regarded as a quasi-identifier. As a result, anonymisation methods aimed at achieving k-anonymity on microdata cannot be directly applied on spatio-temporal data and vice versa.

Contributions. In this article we propose a distance measure for trajectories specially suited for clustering and obfuscation. The distance is loosely based on the Fréchet distance [3], yet it is efficiently computable. The novel construction has significant advantages: (i) it can deal with non-overlapping trajectories, (ii) it outputs, in addition to a distance value, a set of matching points that are exploited later in the obfuscation process, and (iii) it considers the shape of the trajectories due to the very nature of the Fréchet distance. We use the proposed distance measure as the basis of a trajectory anonymisation technique that releases datasets satisfying k-anonymity, regardless of the adversary knowledge. We show, through experiments on synthetic spatio-temporal data, that our approach outperforms previous comparable work in terms of utility.

Outline of the paper. This paper is structured as follows. Section 2 next provides related work. Section 3 introduces a novel distance measure based on the Fréchet distance and the Manhattan norm. A microaggregation-based method for trajectory anonymisation is proposed in Section 4, which is empirically evaluated in Section 5. Finally, Section 6 draws conclusions and future work.

2 Related work

Trajectory k-anonymity is aimed at hiding a single trajectory into a crowd of at least $k-1$ other trajectories. The idea is that every trajectory in the published dataset be indistinguishable from $k-1$ other trajectories and, as a consequence, an adversary cannot identify the individual behind a trajectory with probability higher than $1/k$.

An approach to achieve k-anonymity is by means of suppression of attribute values, which is generally used on discrete and/or semantic data where perturbation methods are not well suited. One of the first suppression-based methods for trajectory anonymization is due to Terrovitis and Mamoulis [21]. They consider trajectories to be sequences of addresses taken from an address domain $\mathcal{P}$. The adversary controls subsets of addresses of $\mathcal{P}$, and thus his knowledge is represented as projections of original trajectories over the addresses in $\mathcal{P}$ that are in the adversary's knowledge. A greedy algorithm aimed at guaranteeing that no address unknown by the adversary can be linked with an user with probability higher than a given threshold is proposed in [21]. The main problem with this approach is that dealing with all possible adversary's knowledge becomes harder than the original k-anonymity problem, which is already known to be NP-Hard [13]. There exist other suppression-based methods in the literature, e.g., [6]. However, they target privacy notions different to k-anonymity.

Like Terrovitis and Mamoulis in [21], Yarovoy et al. also consider an adversary controlling a subset of user's locations or quasi-identifiers [26], with the distinction that such a subset may differ for different users. Trajectory k-anonymity is defined in terms of a bipartite attack graph relating original trajectories with the anonymised trajectories. The authors propose to create anonymised groups through generalisation with respect to the joint set of quasi-identifiers from the users within the group. K-anonymity is thus achieved by creating anonymised groups such that the bipartite attack graph is symmetric and the degree of each vertex representing an anonymised trajectory is at least k. It is worth remarking that the privacy model considered in this article is different, as any user's location is regarded as a quasi-identifier.

Another generalisation-based approach was proposed by Monreale et. al [14]. As in [21], they ignore the time information. Therefore, k-anonymity is achieved if the generalisation of every original trajectory is a sub-trajectory of the generalisation of $k-1$ other trajectories. In order to preserve the utility of the original dataset, a Voronoi tessellation of the geographical area is created so that each location is transformed into the Voronoi cell that contains it. Utility is measured by simply comparing clustering results.

In [8, 9], Domingo-Ferrer et al. propose a different approach based on microagrregation and permutation rather than on generalisation. First, they introduce a novel distance measure that consider both spatial and temporal aspects of trajectories. The distance measure is flexible enough to be used either for spatio-temporal data or time series. Based on this distance measure, the authors propose to create clusters of trajectories so as to minimise the intra-cluster distance. Within a given cluster, locations are randomly swapped with other $k-1$

unswapped close locations. Locations that cannot be swapped are removed and so are the trajectories without swapped locations.

Abul, Bonchi, and Nanni [1, 2] proposed two trajectory anonymisation methods: Never Walk Alone (NWA) and Wait For Me (W4M). Both are partially based on microggregation [7]. The microggregation technique works as follows. The dataset of trajectories is partitioned into several clusters of size at least k and at most $2k - 1$. To do so, NWA relies on the Euclidean distance while W4M uses on the edit distance on real sequences (EDR) [5]. Trajectories within a cluster are perturbed by using space translation. The claimed privacy of these proposals has proven to be flawed [23], though.

In [15, 16], Nergiz et al. consider a trajectory to be a sequence of square geographical areas where a user moves randomly within a given time frame. For clustering, the authors use the log cost metric that balances the spatial and temporal distortion with user-provided weights. Since the log cost metric is based on point matching, the anonymisation process is directly inferred from the clustering process, which improves efficiency.

Recently, Gao et al. proposed a privacy-preserving technique that does not target trajectory k-anonymity directly, as most previous work do, but a trade-off between privacy and utility [12]. Privacy is measured in terms of anonymity sets that are created based on a similarity measure that takes the angles and directions of the trajectories into account. Utility relies on the classical Euclidean distance.

In the literature we can find a variety of distance measures for trajectories and time-series. Vlachos et al. proposed two distance measures based on the Longest Common Subsequence problem (LCSS) [25]. The first one matches only points that are within a given spatio-temporal region. Unmatched points are discarded and taken as outliers. This criterion for outliers detection is smoothed in their second distance measure by using a weighted matching function that considers the distance between points. Another distance measure that has been designed to cope with noise is the Edit Distance on Real sequences (EDR) [5]. The problem is that it requires a fixed and global distance threshold that defines whether a location is too far from another location. A survey on distance measures for trajectory clustering can be found in [28].

3 A distance measure for trajectory microaggregation

We consider trajectories describing the movements of objects on the surface of the earth. Even though a movement is assumed to be continuous, it is typically described by a finite polyline. Formally, a trajectory is defined as a sequence of time-stamped locations $\tau = \ell_1 \cdots \ell_n$ such that $\ell_i.t < \ell_{i+1}.t \ \forall i \in \{1, \cdots, n-1\}$ where $\ell.t$, $\ell.x$, and $\ell.y$, denote the time, latitude, and longitude of the location ℓ, respectively. In general, trajectories can be recorded at different and irregular sampling rates, are not noise-free, and the velocity between two consecutive locations is assumed to be constant. A collection of trajectories is called a spatio-

temporal database. For large databases, the size of a trajectory is considered to be significantly smaller than the size of the database.

The choice of the distance measure is critical in microaggregation. It influences the way trajectories are clustered and usually it also impacts on the anonymisation process. There exist different factors that characterise a trajectory distance measure. For example, a distance measure may consider only trajectories within a given timespan, or look for spatial similarity regardless of direction and sampling rate, or take into account trajectory's features such as speed and angle.

The distance measure we propose in this article is loosely based on the Fréchet distance [11]. The Fréchet distance, also known as the *dog-leash* distance, assumes that a person walks over one trajectory and his dog over the other trajectory. Both may travel at independent but positive speed. The Fréchet distance outputs the minimum-length leash required for that person to walk his dog. Intuitively, the shorter the leash the closer the two curves.

Alt and Godau proposed in 1995 an algorithm to compute the Fréchet distance for two polylines [3] with computational complexity $\mathcal{O}(pq\log(p+q))$ where p and q are the size of the polylines. To the best of our knowledge, this computational complexity has not been improved significantly without making assumptions on the curves. We thus consider variations of the Fréchet distance such as the Coupling distance [10] and the Dynamic Time Warping (DTW) distance [27], which are significantly simpler and runs in $O(pq)$ time complexity.

We say that a sequence $L = (u_{a_1}, v_{b_1}) \cdots (u_{a_n}, v_{b_n})$ is a coupling between two trajectories $U = u_1 \cdots u_p$ and $V = v_1 \cdots v_q$ if the following conditions are satisfied:

- $a_1 = 1$ and $b_1 = 1$
- $a_n = p$ and $b_n = q$
- For every $i \in \{1, \cdots, n-1\}$ it holds that $a_{i+1} = a_i$ or $a_{i+1} = a_i + 1$, and $b_{i+1} = b_i$ or $b_{i+1} = b_i + 1$

A coupling can be seen as sequence of matching points, as defined in the Edit Distance on Real sequences (EDR) [5]. The difference, however, is that a coupling respects the order of the locations and also ensures that all points are considered.

Definition 1 (Coupling distance). *Let $U = u_1 \cdots u_p$ and $V = v_1 \cdots v_q$ be two trajectories and let $\mathcal{L}$ be the set of all couplings between U and V. Let $\|.\|$ denote a norm on $\mathcal{L}$. The coupling distance is defined as follows:*

$$coupling\_dist(U, V) = \min\{\|L\| | L \in \mathcal{L}\}$$

The coupling distance can be computed by a simple dynamic algorithm. The norm that directly relates to the original discrete Fréchet distance is the Infinite norm. Given $L = (u_{a_1}, v_{b_1}) \cdots (u_{a_n}, v_{b_n})$, the Infinite norm $\|L\|_\infty$ is the longest distance between a pair of linked locations in L, i.e., $\|L\|_\infty =$

$\max_{i\in\{1,\cdots,n\}} d(u_{a_i}, v_{b_i})$. Another relevant norm, which we use in this article, is the Manhattan norm, defined as $\|L\|_1 = \sum_{i\in\{1,\cdots,n\}} d(u_{a_i}, v_{b_i})$.

Using the Infinite norm in the coupling distance has a clear interpretation in microaggregation of trajectories, that is, the longest distance that ought to be covered in order to spatially translate a trajectory into another one. However, accounting for the longest distance may lead to non-robust behaviors, because small variations in the trajectories can cause large variations in the distance function. For this reason, we propose to use the infinity norm to compute the optimal coupling between trajectories, yet we consider the average Manhattan norm to represent the actual distance between them. We claim that the average Manhattan norm approximates better the required distortion to microaggregate trajectories. Formally, the distance measure used in the present article is defined as follows.

Definition 2 (Fréchet/Manhattan coupling distance). *Let $U = u_1 \cdots u_p$ and $V = v_1 \cdots v_q$ be two trajectories and let $\mathcal{L}$ be the set of all couplings between U and V. Let $L \subseteq \mathcal{L}$ such that for every $l \in L$ it holds that $\|l\|_\infty$ is minimum amongst the couplings in $\mathcal{L}$. The average coupling distance is defined as:*

$$\min_{l\in L} \frac{1}{|l|}\|l\|_1$$

Computing the Fréchet/Manhattan distance is a bit more elaborated than computing the coupling distance. Nevertheless, it can still be computed in $\mathcal{O}(pq)$ time complexity as shown by Algorithm 1. Given two trajectories $U = u_1 \cdots u_p$ and $V = v_1 \cdots v_q$, we create a matrix I of size $p \times q$ where we store the optimal coupling with respect to the Infinite norm. Such computation is performed by the standard dynamic approach proposed in [10]. In order to determine the Fréchet/Manhattan distance, we consider another matrix M where we store the optimal coupling distance with respect to the Manhattan norm among those optimal couplings with respect to the Infinite norm. To do so, we need to find where those optimal couplings with respect to the Infinite norm come from. Let us analyse what is the impact of having the pair (u_x, v_y) in an optimal coupling l. First, we should notice that if $(u_x, v_y) \in l$ then $\|l\|_\infty \geq d(u_x, v_y)$. Indeed, if $d(u_x, v_y) < \min\{[x-1, y], [x-1, y-1], [x, y-1]\}$ then $\|l\|_\infty = \min\{[x-1, y], [x-1, y-1], [x, y-1]\}$, otherwise $\|l\|_\infty = d(u_x, v_y)$. We thus store in a set C all pairs that lead to an optimal coupling with respect to the Infinite norm amongst the pairs $\{[x-1, y], [x-1, y-1], [x, y-1]\}$. Finally, $M[x, y]$ is computed as $\min\{M[x, y]/L[x, y] | [x, y] \in C\}$ where $L[x, y]$ is the size of the optimal coupling with respect the Manhattan norm for the subtrajectories $u_1 \cdots u_x$ and $v_1 \cdots v_y$.

4 A microaggregation-based approach

The anonymisation method proposed in this article is based on k-microaggregation, that is, a process whereby clusters of at least k *homogeneous* trajectories are anonymised independently. A usual homogeneity criterion is the sum of squared

Algorithm 1 Average coupling distance

Require: Two trajectories $U = u_1 \cdots u_p$ and $V = v_1 \cdots v_q$

Let I, M, L be three matrices of size $p \times q$. Intuitively, I and M represent the Infinity and Manhattan norms, respectively, while L is the length of the optimal coupling
Let d represent the Euclidean distance.
$I[1,1] = M[1,1] = d(u_1, v_1)$
$L[1,1] = 1$
for $i = 2$ to p **do**
    $I[i,1] = \max\{I[i-1,1], d(u_i, v_1)\}$
    $M[i,1] = M[i-1,1] + d(u_i, v_1)$
    $L[i,1] = i$
end for
for $j = 2$ to q **do**
    $I[1,j] = \max\{I[1,j-1], d(u_1, v_j)\}$
    $M[1,j] = M[1,j-1] + d(u_1, v_j)$
    $L[1,j] = j$
end for
for $i = 2$ to p **do**
    for $j = 2$ to q **do**
        Let R be the set $\{[i-1,j], [i-1,j-1], [i,j-1]\}$
        Let $C \subseteq R$ such that for every $[x,y] \in C$ it holds that $I[x,y] \leq d(u_i, v_j)$
        if C is empty **then**
            Let $[a,b] \in R$ such that for every $[x,y] \in R$ it holds that $I[a,b] \leq I[x,y]$
            $I[i,j] = I[a,b]$
            Add to C every element $[x,y]$ in R such that $I[x,y] = I[a,b]$
        else
            $I[i,j] = d(u_i, v_j)$
        end if
        Let $[x,y] \in C$ such that $M[x,y]/L[x,y] \leq M[a,b]/L[a,b]$ for every $[a,b] \in C$
        $M[i,j] = M[x,y] + d(u_i, v_j)$
        $L[i,j] = L[x,y] + 1$
    end for
end for
return $M[p,q]/L[p,q]$

pairwise distances between trajectories within a cluster (intra-cluster distance). Hence, an optimal microaggregation can be intuitively defined as the one maximising the within-groups homogeneity.

The optimal microaggregation problem for multivariate points, like trajectories, has proven to be NP-hard [17]. That justifies the use of heuristics in microaggregation-based approaches for trajectory anonymisation [24]. An additional challenge to be addressed is that distance measures between trajectories tend to be computationally expensive. This implies that computing all pairwise distances between trajectories in a large database is may not be feasible.

Below, we detail the two main components of our microaggregation-based approach, namely the proposed heuristic for trajectory clustering and the obfuscation technique.

4.1 Clustering

We use a greedy approach to address the k-microaggregation problem explained above. Each cluster is represented by a *pivot* trajectory, and contains $k-1$ other trajectories that are close to the pivot trajectory. In other words, we consider as homogeneity criterion the sum of squared distances between the pivot trajectory and the other trajectories in the cluster. Once a cluster C is created from a pool D of trajectories, all trajectories in C are removed from D. As shown by Algorithm 2, this process is repeated until D contains less than k trajectories.

Algorithm 2 Trajectory clustering

Require: $D = \{\tau_1, \ldots, \tau_N\}$ a set of trajectories; a distance measure $d : D \times D \to \mathbb{R}$; a natural number δ representing the number of clusters generated at each iteration; and an anonymization parameter k

Let $\mathcal{C}$ be an empty set of clusters of trajectories
while $|D| \geq k$ **do**
  Let τ'_1 be a random trajectory in D
  Let τ'_δ be the farthest trajectory to τ'_1 with respect to d
  Let $\tau'_1, \tau'_2, \cdots, \tau'_\delta$ be δ trajectories in D that minimizes the sum of squares $\sum_{i \in [1..\delta-1]} d(\tau'_{i+1}, \tau'_i)^2$
  Let C_0 be an empty set of trajectories and $d_0 = \infty$
  for all $i = 1$ to δ **do**
    Create the cluster of trajectories C_i containing τ'_i and the closest $k-1$ trajectories to τ'_i
    Compute $d_i = \sum_{\tau \in C_i} d(\tau'_i, \tau)^2$
    if $d_i < d_0$ **then**
      $C_0 = C_i$ and $d_0 = d_i$
    end if
  end for
  $\mathcal{C} = \mathcal{C} \cup \{C_0\}$
  Remove all trajectories in C_0 from D
end while
return $\mathcal{C}$

In our approach, depicted in Algorithm 2, finding the optimal set of clusters is equivalent to finding the optimal sequence of pivot trajectories. The most effective greedy solution to this problem is to choose the best cluster amongst the $|D|$ clusters that can be created considering each trajectory in D a pivot trajectory. However, that requires the computation of all pairwise distances between the trajectories in D. As a trade-off, given a natural number $\delta \ll |D|$, we choose a random trajectory τ'_1 in D and find the sequence $\tau'_1, \tau'_2, \cdots \tau'_\delta$ such that: i) τ'_δ

is the farthest trajectory to τ'_1 and ii) the sum of squares $\sum_{i\in[1..\delta-1]} d(\tau'_{i+1}, \tau'_i)^2$ is minimum. We thus choose the best cluster amongst the δ clusters that can be built considering either τ'_1, or τ'_2, $\cdots$, or τ'_δ, as pivot. Note that, if $\delta = |D|$ then we actually find the optimal set of clusters.

4.2 Obfuscation technique

Our privacy-preserving method for the publication of trajectories is based on the clustering technique and the Fréchet/Manhattan coupling distance described above. Even though the coupling distance deals well with trajectories recorded at different sampling rates, the lower is the sampling rate the better it approximates the classical Fréchet distance. We thus use linear interpolation to decrease and homogenise the sampling rate of two trajectories as follows. Let $U = u_1 \cdots u_p$ and $V = v_1 \cdots v_q$ be two trajectories. For every $i \in \{1, \cdots, p\}$, we insert in V by using linear interpolation a new point at time $v_1.t + \frac{(v_q.t - v_1.t)(u_i.t - u_1.t)}{u_p.t - u_1.t}$. An analogous procedure is used to increase the sampling rate of U with respect to V. Note that, the trajectories U' and V' resulting from re-sampling U and V, respectively, have equal size.

We use the Fréchet/Manhattan distance in Algorithm 2 to partition a collection of trajectories $\{\tau_1, \cdots, \tau_N\}$ into a set of homogeneous clusters $\{C_1, \cdots, C_m\}$. For every $i \in \{1, \cdots, m\}$, let X be the pivot trajectory in the cluster C_i as considered in Algorithm 2. For each $Y \in C_i$ $(X \neq Y)$, let $(u_{a_1}, u_{b_1}) \cdots (u_{a_n}, v_{b_n})$ be the optimal coupling between X and Y with respect to the Fréchet/Manhattan distance, where $U = u_1 \cdots u_p$ and $V = v_1 \cdots v_q$ are the re-sampling of X and Y, respectively. For each $j \in \{1, \cdots, n\}$ and if $u_{a_j} \in X$, i.e., if u_{a_j} is an original location of X rather than an interpolated location added during the re-sampling procedure, we add to the set $S(u_{a_i})$ the location v_{b_i}. Once this process is finished for all trajectories in C_i, we consider, for every location $x \in X$, the set $S(x)$ containing those locations from other trajectories in C_i that formed a pair with x in an optimal coupling. We always include x into $S(x)$ whenever $S(x)$ is not empty. The anonymised trajectory for the cluster C_i will be that formed by the average locations obtained from the sets $\{S(x) | x \in X, S(x) \neq \emptyset\}$. A pseudo-code description of this procedure is given in Algorithm 3.

4.3 Privacy analysis

Several notions of trajectory k-anonymity exist. For example, in [14, 21], the adversary ignores the time dimension. In [1, 2], an adversary is considered unable to distinguish two locations if their distance is below a predefined threshold. In [9], the model is defined considering that original locations must be preserved, which means that random spatial distortion is disallowed.

In this article we consider trajectory k-anonymity as a property of the anonymised dataset regardless the adversary capabilities. Our notion of k-anonymity is indeed similar to that presented in [15, 16] for generalised trajectories.

Algorithm 3 Trajectory anonymization algorithm

Require: $\{\tau_1, \ldots, \tau_N\}$ a collection of original trajectories; a number δ to be used in the clustering process; the Fréchet/Manhattan distance d; an anonymisation parameter k

1: Use the clustering technique defined by Algorithm 2 on input $\{\tau_1, \ldots, \tau_N\}$, the distance measure d, δ, and k, to obtain a set of clusters $\{C_1, \cdots, C_m\}$
2: Let $D^\star$ be an empty set of trajectories
3: **for** $i = 1$ to m **do**
4: Let X be the pivot trajectory in C_i as defined in Algorithm 2
5: Let $S(x)$ be an empty set for every $x \in X$
6: **for** $Y \in C_i$ and $X \neq Y$ **do**
7: Let $(u_{a_1}, v_{b_1}) \cdots (u_{a_n}, v_{b_n})$ be the optimal coupling between X and Y with respect to the Fréchet/Manhattan distance d, where $U = u_1 \cdots u_p$ and $V = v_1 \cdots v_q$ are the re-sampling of X and Y, respectively
8: **for** $j = 1$ to n **do**
9: **if** $u_{a_j} \in X$ **then**
10: $S(u_{a_j}) = S(u_{a_j}) \cup \{v_{b_j}\}$
11: **end if**
12: **end for**
13: **end for**
14: Let τ be an empty trajectory
15: **for** $x \in X$ and $S(x) \neq \emptyset$ **do**
16: $S(x) = S(x) \cup \{x\}$
17: Add to τ the average location formed by the locations in $S(x)$
18: **end for**
19: $D^\star = D^\star \cup \{\underbrace{\tau, \ldots, \tau}_{k}\}$
20: **end for**
21: **return** $D^\star$

Definition 3 (Trajectory k-anonymity). *Let $D^\star$ be a collection of trajectories. $D^\star$ meets trajectory k-anonymity if every trajectory in $D^\star$ is equal to other $k-1$ trajectories in $D^\star$.*

Theorem 1. *Let D be a collection of original trajectories and $D^\star$ the output of Algorithm 3 on input D. $D^\star$ satisfies trajectory k-anonymity.*

Proof. The proof trivially follows from the fact that Algorithm 3 produces k equal trajectories for each cluster (see Step 19 in Algorithm 3). □

5 Empirical evaluation

As the privacy-preserving anonimisation technique introduced in this article replaces a cluster of k close, but potentially different, trajectories by k identical trajectories, it is of paramount importance to evaluate utility loss in this method. Next, we introduce spatial-range queries as a measure of utility loss. We finally compare our anonymisation approach with other state-of-the-art privacy preserving techniques.

5.1 Trajectory analysis and utility measures

There exist a plethora of trajectory analysis techniques developed within the Geographic Information Science and Data Mining fields. These techniques may look for movement patterns such as flocking, leadership, commuting, and encounter, or may be aimed at answering basic queries such as nearest neighbor or range queries.

In this article we mainly focus on queries that are used for aggregate statistics. This queries are typically measurable, and thus they can be defined as functions on the domain of all spatio-temporal databases ranging over a metric space. Let $\mathcal{D}$ be the universe of all possible collections of trajectories and let (M, d) be a metric space. A spatio-temporal query Q is formally defined as a function $Q : \mathcal{D} \to M$. Examples of measurable queries are traffic density, travel time, peak hours, amongst many others.

Measurable queries can be naturally used to define utility measures for anonymization techniques as follows. Let $D \in \mathcal{D}$ be an original spatio-temporal database and $D^\star \in \mathcal{D}$ its anonymized version. Given a measurable query $Q : \mathcal{D} \to M$, we measure utility loss by the formula $d(Q(D), Q(D^\star))$. The closer this measure to zero the better $D^\star$ approximates D with respect to Q.

A well-known type of measurable query in trajectory analysis is *spatio-temporal range queries*, which were introduced by Trajcevski et al. in [22] in 2004. In particular, we consider the two following queries.

- *Sometime_Definitely_Inside(T, R, t_b, t_e)* is *true* if and only if there exists a time $t \in [t_b, t_e]$ at which trajectory T is inside region R.
- *Always_Definitely_Inside(T, R, t_b, t_e)* is *true* if and only if at every time $t \in [t_b, t_e]$, trajectory T is inside region R.

At a first sight, it may seem that the query *Always_Definitely_Inside*(AI) is stronger than *Sometime_Definitely_Inside*(SI). However, with the later we can formulate questions at a local level like: how many users pass through the Grand Place in Belgium?, whilst with AI the shape of trajectories becomes more relevant and might be useful for questions like: how many users take the toll highway placed between Barcelona and Tarragona cities?

Other important points to be remarked are the area of R and the time interval $[t_b, t_e]$. Both provide flexibility when dealing with uncertain or perturbed trajectories. Asking for trajectories passing through a single location at a given time-stamp is meaningless in this type of imprecise data. The size of the area and the time interval should not be too large either, though.

Similarly to [1, 2, 9], we used both queries to define a distortion metric of the anonymised dataset $\mathcal{T}^\star$ with respect to original dataset $\mathcal{T}$. The idea is to define a large set of queries according to some distribution of regions and time intervals. The same set of queries is applied to both datasets $\mathcal{T}^\star$ and $\mathcal{T}$ and the number of trajectories satisfying SI and AI are counted as shows the following SQL style code.

- Query $\mathcal{Q}_1(\mathcal{T}, R, t_b, t_e)$:

SELECT COUNT (*) FROM $\mathcal{T}$ WHERE SI($\mathcal{T}$.traj, R, t_b, t_e)

– Query $\mathcal{Q}_2(\mathcal{T}, R, t_b, t_e)$:
SELECT COUNT (*) FROM $\mathcal{T}$ WHERE AI($\mathcal{T}$.traj, R, t_b, t_e)

Two different *range query distortions* $\text{SID}(\mathcal{T}, \mathcal{T}^\star)$ and $\text{AID}(\mathcal{T}, \mathcal{T}^\star)$ are defined by using the accumulative queries $\mathcal{Q}_1$ and $\mathcal{Q}_2$, respectively.

– $\text{SID}(\mathcal{T}, \mathcal{T}^\star) = \frac{1}{|\xi|} \sum_{\forall < R, t_b, t_e > \in \xi} \frac{|\mathcal{Q}_1(\mathcal{T}, R, t_b, t_e) - \mathcal{Q}_1(\mathcal{T}^\star, R, t_b, t_e)|}{\max(\mathcal{Q}_1(\mathcal{T}, R, t_b, t_e), \mathcal{Q}_1(\mathcal{T}^\star, R, t_b, t_e))}$ where ξ is a large set of SI queries.
– $\text{AID}(\mathcal{T}, \mathcal{T}^\star) = \frac{1}{|\xi|} \sum_{\forall < R, t_b, t_e > \in \xi} \frac{|\mathcal{Q}_2(\mathcal{T}, R, t_b, t_e) - \mathcal{Q}_2(\mathcal{T}^\star, R, t_b, t_e)|}{\max(\mathcal{Q}_2(\mathcal{T}, R, t_b, t_e), \mathcal{Q}_2(\mathcal{T}^\star, R, t_b, t_e))}$ where ξ is a large set of AI queries.

Both metrics SID and AID are bounded by 0 and 1. The minimum is achieved when $\mathcal{Q}_i(\mathcal{T}, R, t_b, t_e) = \mathcal{Q}_i(\mathcal{T}^\star, R, t_b, t_e)$, and the maximum if $\mathcal{Q}_i(\mathcal{T}^\star, R, t_b, t_e) = 0$, where $i \in \{1, 2\}$. Therefore, the lower the range query distortion the lower the utility loss of the anonymised dataset.

5.2 Implementation details of the considered methods

We compare our method with the generalisation-based and permutation-based approach proposed in [16] and [9], respectively. The generalisation-based method relies on a distance threshold, which allows the Log-cost distance measure to discard outlier locations. Because in this section we only consider noiseless synthetic data, we have set up such a distance threshold to its maximum value. The permutation-based method, instead, discard outlier locations during the obfuscation process by considering both a distance and a time threshold. Again, we set up both thresholds to their maximum values so as to avoid outlier removal in a noiseless dataset. The permutation-based method considered in this article is the one named *SwapLocations* in [9].

5.3 Results on synthetic trajectories

We compare our anonymisation method with other approaches by using a synthetic dataset generated with Brinkhoff's framework [4], which is used often to evaluate privacy-preserving approaches. Synthetic data generated with Brinkhoff's generator have the advantage of being easily transferable and reproducible. We thus provide next the parameters used to generate the dataset of trajectories considered in our experiments.

The generation parameters over the map of Oldenburg were: 6 moving object classes and 3 external object classes; 5 moving objects and 3 external object generated per time-stamp; the maximum lifespan of a trajectory was set up to $1,000$ time-stamps; speed 10; and report probability $1,000$. This resulted in $5,000$ synthetic trajectories provided by Brinkhoff's generator [4], which contain a total of $492,105$ locations in the German city of Oldenburg and 98.421 locations per trajectory in average.

In order to generate spatial-range queries, we considered regions whose radius randomly distributes over the interval of natural numbers $[0, 500]$. The maximum of this interval is a small fraction of the average length of each trajectory, which is 7284. Remark that the smaller the spatial interval the tighter is the spatial-range query and the harder become for an anonymisation technique to apply spatial distortion without bringing down utility. We respect to the time dimension we considered different time intervals $[0, 0], [0, 300], [0, 600], [0, 1800], [0, 3600]$. For a given time interval $[0, t]$, we generate a spatial-range query by choosing a random interval $[t_b, t_e]$ such that $0 \leq t_b \leq t_e \leq t$.

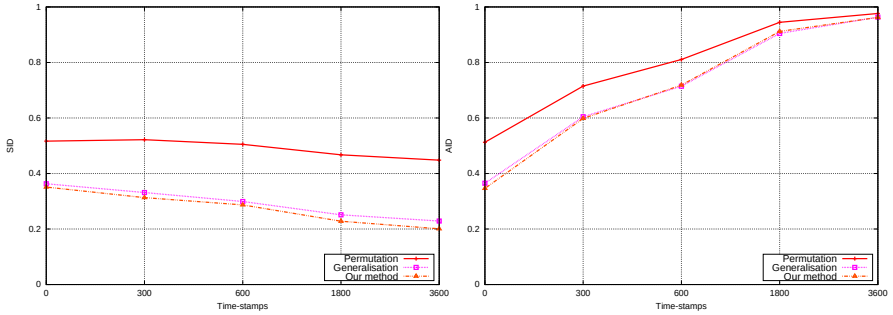
We generated for each time interval $100,000$ spatial-range queries of both types: $\mathcal{Q}_1$ and $\mathcal{Q}_2$. Armed with these set of queries, we computed the range query distortions SID and AID of the anonymised data sets provided by three different anonymisation methods: the Generalisation-based approach [16], the Permutation-based approach [9], and our method. Each anonymisation method provided three different datasets satisfying k-anonymity with $k \in \{2, 4, 8\}$. The results are depicted in Figure 1.

It can be seen from Figure 1 that our method performs better than the approaches proposed in [16, 9] for every cluster size and every time interval. The improvement in terms of utility increases as the offered privacy increases. For $k = 2$, our method is just slightly better than the generalisation-based approach, while for $k \in \{4, 8\}$ our method performs significantly better. This means that our technique clusters and anonymises trajectories more efficiently.

Figure 1 also shows that more research on trajectory anonymisation techniques ought to be conducted. The ideal range query distortion is zero, and none of the three considered techniques gets close to this optimal value. This issue can be overcome by considering larger datasets of original trajectories. Intuitively, the larger the dataset the easier is to find clusters with low intra-cluster distance. Other solution approach consists in removing outlier trajectories, that is, trajectories that cannot be clustered with other $k - 1$ trajectories without dramatically increasing the intra-cluster distance. The study and evaluation of these solution approaches, as well as reporting on results over real-life datasets, are left as future work.

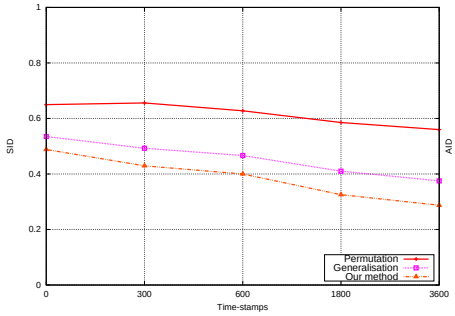
6 Conclusions

In this article we have introduced a novel distance measure for trajectories, which is well suited for both clustering and anoymisation. The proposed distance measure resembles to other types of coupling distance measures, such as the Fréchet distance, with the particularity that the Infinite norm and the Manhattan norm are considered together. To demonstrate the suitability of our distance measure, we presented a trajectory-anonymisation heuristic method that creates cluster with low intra-cluster distance and satisfies trajectory k-anonymity. Empirical results show that our method offers better utility than other state-of-the-art methods, such as the generalisation-based and permutation-based approaches.

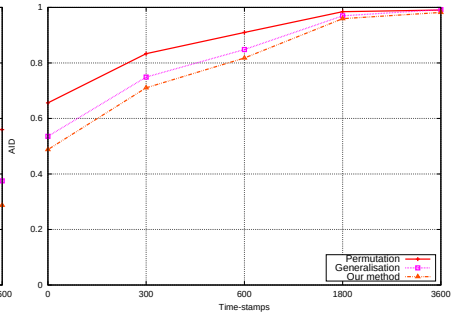


(a) $k = 2$ (PSI)

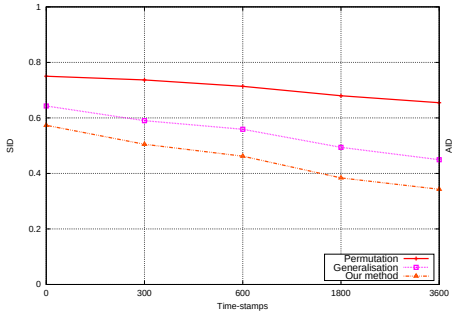
(b) $k = 2$ (DAI)



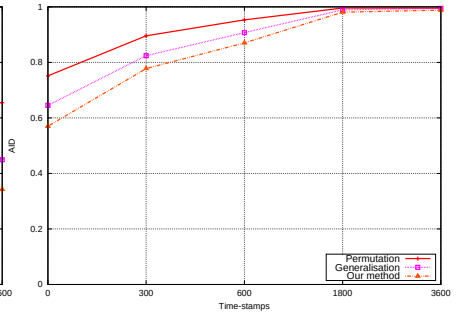
(c) $k = 4$ (PSI)



(d) $k = 4$ (DAI)



(e) $k = 8$ (PSI)



(f) $k = 8$ (DAI)

Fig. 1. Six charts showing the range query distortions of three different anonymisation methods. Charts on the left depict the SI query distortion (SID), while charts on the right show the AI query distortion (AID).

Future work will be directed towards reaching optimal range-query distortion values.

References

1. Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: uncertainty for anonymity in moving objects databases. In *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008,Cancun, Mexico, 7-12 April 2008*, pages 376–385. IEEE, 2008.
2. Osman Abul, Francesco Bonchi, and Mirco Nanni. Anonymization of moving objects databases by clustering and perturbation. *Inf. Syst.*, 35(8):884–910, 2010.
3. Helmut Alt and Michael Godau. Computing the Fréchet distance between two polygonal curves. *Int. J. Comput. Geometry Appl.*, 5:75–91, 1995.
4. Thomas Brinkhoff. A framework for generating network-based moving objects. *Geoinformatica*, 6(2):153–180, 2002.
5. Lei Chen, M. Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, 14-16 June 2005*, pages 491–502. ACM, 2005.
6. Rui Chen, Benjamin C. M. Fung, Noman Mohammed, Bipin C. Desai, and Ke Wang. Privacy-preserving trajectory data publishing by local suppression. *Inf. Sci.*, 231:83–97, 2013.
7. Josep Domingo-Ferrer and Josep Maria Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.*, 14(1):189–201, 2002.
8. Josep Domingo-Ferrer, Michal Sramka, and Rolando Trujillo-Rasúa. Privacy-preserving publication of trajectories using microaggregation. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, SPRINGL '10, pages 26–33, New York, NY, USA, 2010. ACM.
9. Josep Domingo-Ferrer and Rolando Trujillo-Rasua. Microaggregation- and permutation-based anonymization of movement data. *Inf. Sci.*, 208:55–80, 2012.
10. Thomas Eiter and Heikki Mannila. Computing Discrete Fréchet Distance. Technical report, Christian Doppler Laboratory for Expert Systems,, TU Vienna, Austria, 1994. Technical Report CD-TR 94/64.
11. Maurice Fréchet. Sur quelques points du calcul functionnel [On some points of functional calculus]. *Rendiconti del Circolo Matematico di Palermo*, 22:1–74, 1906.
12. Sheng Gao, Jianfeng Ma, Cong Sun, and Xinghua Li. Balancing trajectory privacy and data utility using a personalized anonymization model. *J. Netw. Comput. Appl.*, 38:125–134, February 2014.
13. Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '04, pages 223–228, New York, NY, USA, 2004. ACM.
14. Anna Monreale, Gennady Andrienko, Natalia Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. Movement data anonymity through generalization. *Trans. Data Privacy*, 3(2):91–121, 2010.
15. Mehmet Ercan Nergiz, Maurizio Atzori, and Yücel Saygin. Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS*

and LBS, SPRINGL 2008, Irvine, California, USA, 4 November 2008, pages 52–61. ACM, 2008.

16. Mehmet Ercan Nergiz, Maurizio Atzori, Yücel Saygin, and Baris Guc. Towards trajectory anonymization: a generalization-based approach. *Trans. Data Privacy*, 2(1):47–75, 2009.
17. Anna Oganian and Josep Domingo-ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18:345–354, 2001.
18. Pierangela Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. on Knowl. and Data Eng.*, 13(6):1010–1027, November 2001.
19. Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.
20. Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzz.*, 10(5):557–570, 2002.
21. Manolis Terrovitis and Nikos Mamoulis. Privacy preservation in the publication of trajectories. In *IEEE International Conference on Mobile Data Management*, pages 65–72, Los Alamitos, CA, USA, 2008. IEEE Computer Society.
22. Goce Trajcevski, Ouri Wolfson, Klaus Hinrichs, and Sam Chamberlain. Managing uncertainty in moving objects databases. *ACM Trans. Database Syst.*, 29:463–507, 2004.
23. Rolando Trujillo-Rasua and Josep Domingo-Ferrer. On the privacy offered by (k, δ)-anonymity. *Information Systems*, 38(4):491 – 494, 2013.
24. Rolando Trujillo-Rasua and Josep Domingo-Ferrer. Privacy in spatio-temporal databases: A microaggregation-based approach. In *Advanced Research in Data Privacy*, pages 197–214. 2015.
25. M. Vlachos, D. Gunopulos, and G. Kollios. Robust similarity measures for mobile object trajectories. In *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*, pages 721–726, Sept 2002.
26. Roman Yarovoy, Francesco Bonchi, Laks V. S. Lakshmanan, and Wendy Hui Wang. Anonymizing moving objects: how to hide a mob in a crowd? In *Proceedings of the 12th International Conference on Extending Database Technology, EDBT 2009, Saint Petersburg, Russia, 24-26 March 2009*, volume 360 of *ACM International Conference Proceeding Series*, pages 72–83. ACM, 2009.
27. Byoung-Kee Yi, H. V. Jagadish, and Christos Faloutsos. Efficient retrieval of similar time sequences under time warping. In *Proceedings of the Fourteenth International Conference on Data Engineering*, ICDE '98, pages 201–208, Washington, DC, USA, 1998. IEEE Computer Society.
28. Zhang Zhang, Kaiqi Huang, and Tieniu Tan. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 03*, ICPR '06, pages 1135–1138, Washington, DC, USA, 2006. IEEE Computer Society.