*N*ever *W*alk *A*lone: Uncertainty for Anonymity in Moving Objects Databases

Osman Abul [#1], Francesco Bonchi [*2], Mirco Nanni [*3]

[#]*Computer Engineering Dept., TOBB University*
Ankara, Turkey
[1]osmanabul@etu.edu.tr

[*]*Pisa KDD Laboratory, ISTI - CNR*
Pisa, Italy
[2]francesco.bonchi@isti.cnr.it
[3]mirco.nanni@isti.cnr.it

***Abstract*— Preserving individual privacy when publishing data is a problem that is receiving increasing attention. According to the k-anonymity principle, each release of data must be such that each individual is indistinguishable from at least $k-1$ other individuals. In this paper we study the problem of anonymity preserving data publishing in moving objects databases. We propose a novel concept of k-anonymity based on co-localization that exploits the inherent uncertainty of the moving object's whereabouts. Due to sampling and positioning systems (e.g., GPS) imprecision, the trajectory of a moving object is no longer a polyline in a three-dimensional space, instead it is a cylindrical volume, where its radius δ represents the possible location imprecision: we know that the trajectory of the moving object is within this cylinder, but we do not know exactly where. If another object moves within the same cylinder they are indistinguishable from each other. This leads to the definition of (k, δ)-anonymity for moving objects databases.**

We first characterize the (k, δ)-anonymity problem and discuss techniques to solve it. Then we focus on the most promising technique by the point of view of information preservation, namely space translation. We develop a suitable measure of the information distortion introduced by space translation, and we prove that the problem of achieving (k, δ)-anonymity by space translation with minimum distortion is NP-hard. Faced with the hardness of our problem we propose a greedy algorithm based on clustering and enhanced with ad hoc pre-processing and outlier removal techniques. The resulting method, named $\mathcal{NWA}$ ($\mathcal{N}$ever $\mathcal{W}$alk $\mathcal{A}$lone), is empirically evaluated in terms of data quality and efficiency.

Data quality is assessed both by means of objective measures of information distortion, and by comparing the results of the same spatio-temporal range queries executed on the original database and on the (k, δ)-anonymized one. Experimental results show that for a wide range of values of δ and k, the relative error introduced is kept low, confirming that $\mathcal{NWA}$ produces high quality (k, δ)-anonymized data.

I. Introduction

With today's pervasiveness of mobile phones and other location-aware devices, the amount of traces left by moving objects and daily collected by service providers is continuously increasing. The wealth of *space-time trajectories* left by these personal devices and their human companions is expected to enable novel classes of applications, for instance in traffic and sustainable mobility management, where the discovery of behavioral patterns is the key step. Clearly, in these applications privacy is a concern, since location data enables intrusive inferences, which may reveal habits, social customs, religious and sexual preferences of individuals, and can be used for unauthorized advertisement and user profiling.

As an example, consider a traffic control application that collects vehicle movements. In a naïve tentative of preserving anonymity, the car identifiers are not disclosed but instead replaced with pseudonyms. However, as shown in [1] such operation is insufficient to guarantee anonymity, since location represents a property that in some cases can lead to the identification of the individual. For example, if one is known to follow almost every morning the same route, it is very likely that the starting point is the home of the individual and the ending point is the working place. Joining this information with some telephone directories we can easily link the trajectory to its owner.

In this paper we study the problem of *anonymity preserving data publishing in moving objects databases*. In particular, we extend the classical concept of k-anonymity [2] to deal with this particular form of data, and to exploit its inherent uncertainty [3], [4], [5]. In fact the energy in a mobile device is very limited, so it is impossible for a mobile object to continuously send out its location information. To reduce the energy consumption, many methods [6] are developed for predicting an expected location of a mobile object at a given time t, using some predictive model, e.g., Kalman Filter, linear model, etc. If the actual location of the mobile object differs more than a uncertainty threshold δ from the predicted location, then the mobile object reports the new location, otherwise it does not. The uncertainty threshold δ, that is a parameter in our framework, has a real-world technological counterpart defined by an agreement between the server and the mobile device. For sake of presentation, in the following we assume a common δ, although our framework can easily handle different δs for different users, as discussed in Section VII.

II. Related Work and Our Contribution

The traditional k-anonymity framework [2] focuses on relational tables: the basic assumptions are that each tuple in the table corresponds uniquely to an individual, and that attributes are divided in *quasi-identifiers* (i.e., attributes that can be linked to external information to reidentify the individual to whom the information refers), and *sensitive attributes*. Although it has been shown that it presents some flaws and limitations [7], and that finding an optimal k-anonymization is NP-hard [8], the k-anonymity model is still practically relevant and in recent years a large research effort has been devoted to develop algorithms for k-anonymity (see [9], [10]).

Moving objects databases (MOD) [11] is another rather young research area that has received a lot of interest in recent years. Several different MOD problems have been tackled, ranging from indexing [12], [13], [14], representing and querying [15], [16], [17], updating and modelling imprecision and communication costs [3], [18]. Existing work about anonymity of spatio-temporal moving points has been mainly developed in the context of *location based services* (LBS). In this context a trusted server is usually in charge of handling users' requests and passing them to the service providers, and the general goal is to provide the service on-the-fly without threatening the anonymity of the user that is requiring the service.

The concept of *location k-anonymity* for LBS was first introduced in [19] and later extended in [20] to deal with different values of k for different requests. The underlying idea is that a message sent from a user is k-anonymous when it is indistinguishable from the spatial and temporal information of at least $k-1$ other messages sent from different users. The proposed solution is based on a spatial subdivision in areas, and on *delaying the request* as long as the number of users in the specified area does not reach k. The work in [20] instead of using the same k for all messages, allows each message to specify an independent anonymity value and the *maximum spatial* and *temporal tolerance resolutions* it can tolerate based on its privacy requirements. The work described in [21] proposes a privacy system that takes into account only the spatial dimension: the area in which location anonymity is evaluated is divided into several regions and position data is delimited by the region. Anonymity is required in two different ways: the first, called *ubiquity*, requires that a user visits at least k regions; the second, called *congestion*, requires the number of users in a region to be at least k. High ubiquity guarantees the location anonymity of every user, while high congestion guarantees location anonymity of local users in a specified region. In [22] the concept of *mix zones* is introduced. A mix zone is an area where the location based service providers can not trace users' movements. When a user enters a mix zone, the service provider does not receive the real identity of the user but a pseudonym that changes whenever the user enters a new mix zone. In this way, the identities of users are kept confused. A similar classification of areas, named *sensitivity map* is introduced in [23]: it classifies locations as either *sensitive* or *insensitive*, and describes three algorithms that hide users' positions in sensitive areas.

Contrary to the notions of mixed zones and sensitivity maps, the approach introduced in [1] is geared on the concept of *location based quasi-identifier*, i.e., a spatio-temporal pattern that can uniquely identify one individual. How to exploit this interesting concept in the case of data publishing is a serious, challenging, open problem not addressed in [1] nor in other work. In our framework we do not take in consideration the possibility of having spatio-temporal quasi-identifiers, instead we simply develop a technique to make k trajectories be indistinguishable in their whole. Once understood how to introduce quasi-identifiers in the context of publishing a database of moving objects, it will be interesting to adapt our techniques to deal with this case.

All the work described above is developed for location based services. In this paper, instead, we face the problem by the perspective of *privacy aware data publishing*, i.e., the same context of classical k-anonymity. In our setting, we have a static database of moving objects and we want to publish it (for instance for analysis purpose) is such a way that the anonymity of the individuals is preserved, but also the *quality* of the data is kept high. On the contrary, in the LBS context the aim is to provide the service without learning user's exact position, and the data can also be forgotten once that the service has been provided. In other terms, in our context anonymity is *off-line* and *data-centric*, while in the LBS context is a sort of *on-line service-centric* anonymity. A solution to the first problem is not, in general, a solution to the second (and viceversa), and both problems are important. Consider, for instance, the concept of *mix zones* previously described: it is a solution for LBS since it still allows to provide the service, but it is not for data publishing, since the quality of the data is completely destroyed. While several works exist about anonymity in LBS, to the best of our knowledge this is the first paper addressing the problem of k-anonymity of trajectories by a data publishing perspective. In this context, our main contribution is the introduction of concept of (k, δ)-anonymity, i.e., anonymity exploiting the inherent uncertainty of the moving object's whereabouts. In the next section we provide the problem definition, while in Section IV we discuss possible techniques to enforce (k, δ)-anonymity, arguing that in our context the most challenging and the most promising technique by the point of view of information preservation is *space translation*. We develop a suitable measure of the information distortion introduced by space translation, and we prove that the problem of achieving (k, δ)-anonymity by space translation with minimum distortion is NP-hard. Thus we propose a two-steps greedy method: in the first step we group trajectories in clusters having at least k elements; in the second step we perform the minimum space translation needed to achieve (k, δ)-anonymity.

Several previous works used *k-member clustering* for *k-anonymity* [24], [25], [26]. Our proposal differs from these results, not only because we work with trajectories of moving objects instead of the usual relations, but also because we introduce uncertainty in the model. While in previous proposals

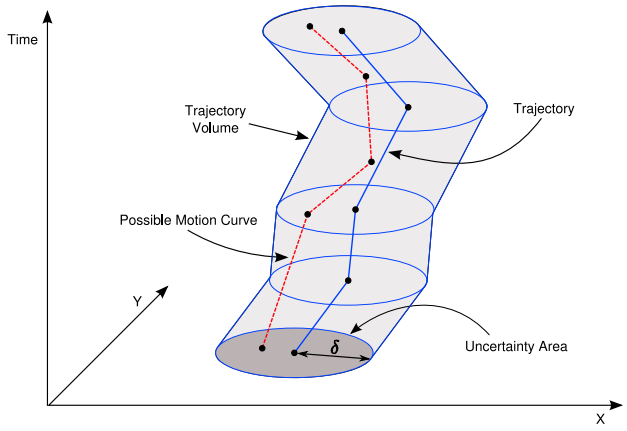


Fig. 1. Uncertain trajectory: uncertainty area, trajectory volume and possible motion curve.

what is released is the centroid of a cluster together with the cardinality of the cluster, in our approach each trajectory is released distinctly while the anonymity is guaranteed by the position uncertainty. This leads to an important benefit of our approach: *the fact that the released data has been previously anonymized is much less evident in our approach*. It is worth noting that publishing one unique representative per cluster, for a number of times equal to the population of the cluster, can be seen as a special instance ($\delta = 0$, i.e., no uncertainty) of our framework.

III. Problem Definition

Following [4] an *uncertain trajectory* is defined as a cylindrical volume of radius δ.

Definition 1 (Uncertain Trajectory[4]): A trajectory of a moving object is a polyline in three-dimensional space represented as a sequence of spatio-temporal points: $(x_1, y_1, t_1), (x_2, y_2, t_2) \ldots (x_n, y_n, t_n)(t_1 < t_2 < \cdots < t_n)$. During the time segment $[t_i, t_{i+1}]$ the object is assumed to move along a straight line from (x_i, y_i) to (x_{i+1}, y_{i+1}) at a constant speed. Given a trajectory τ between times t_1 and t_n, and an uncertainty threshold δ, the pair $\langle \tau, \delta \rangle$ defines an uncertain trajectory. For each point (x, y, t) along τ, its uncertainty area is the horizontal disk (i.e., circle and its interior) with radius δ and centered at (x, y, t), where (x, y) is the expected location at time $t \in [t_1, t_n]$. The trajectory volume of $\langle \tau, \delta \rangle$, denoted $Vol(\tau, \delta)$ is the union of all such disks for all $t \in [t_1, t_n]$. A possible motion curve of τ is any continuous function $f_{PMC^\tau} : Time \rightarrow \mathbb{R}^2$ defined on the interval $[t_1, t_n]$ such that for any $t \in [t_1, t_n]$, the spatio-temporal point $(f_{PMC^\tau}(t), t)$ is inside the uncertainty area at time t: we also adopt the notation $f_{PMC^\tau} \subset Vol(\tau, \delta)$. □

Definition 1 is graphically represented in Figure 1. Intuitively, two trajectories are indistinguishable if they are defined in the same time interval and they follow *almost* the same route w.r.t. the uncertainty threshold.

Definition 2 (Co-localization): Two trajectories τ_1, τ_2 defined in $[t_1, t_n]$ are said to be co-localized w.r.t. δ, iff for each point (x_1, y_1, t) in τ_1 and (x_2, y_2, t) in τ_2 with $t \in [t_1, t_n]$, it holds that $Dist((x_1, y_1), (x_2, y_2)) \leq \delta$, where $Dist$ is the Euclidean distance: $Dist((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. We write $Coloc_\delta(\tau_1, \tau_2)$ omitting the time interval $[t_1, t_n]$. □

Another way to express the co-localization of trajectories is to say that each one is a possible motion curve of the other:

$$Coloc(\tau_1, \tau_2) \iff \tau_1 \subset Vol(\tau_2, \delta) \iff \tau_2 \subset Vol(\tau_1, \delta)$$

Given an anonymity threshold k, we can define an *anonymity set* as a set of at least k trajectories that are co-localized.

Definition 3 (Anonymity Set of Trajectories): Given a position uncertainty threshold δ and an anonymity threshold k, a set S of trajectories is a (k, δ)-anonymity set iff $|S| \geq k$ and $\forall \tau_i, \tau_j \in S \,.\, Coloc_\delta(\tau_i, \tau_j)$. □

The following properties further characterize an anonymity set of trajectories.

Proposition 1: A set of trajectories S, with $|S| \geq k$, is a (k, δ)-anonymity set iff it exists a trajectory τ_c s. t. all the trajectories in S are possible motion curves of τ_c within an uncertainty radius of $\delta/2$: i.e., $\forall \tau \in S \,.\, \tau \subset Vol(\tau_c, \delta/2)$.

Given a (k, δ)-anonymity set S, the trajectory τ_c is obtained by taking, for each $t \in [t_1, t_n]$, the point (x, y) that is the center of the minimum bounding circle of all the points at time t of all trajectories in S. □

Therefore, an anonymity set of trajectories can be bounded by a cylindrical volume of radius $\delta/2$. In Figure 2, we graphically represent this property.

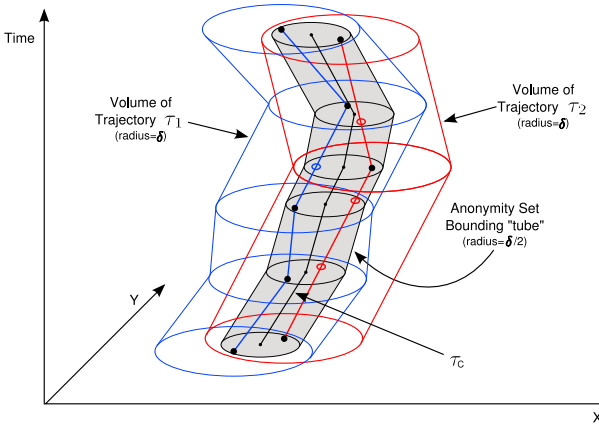


Fig. 2. A $(2, \delta)$-anonymity set formed by two co-localized trajectories, their respective uncertainty volumes, and the central cylindrical volume of radius $\delta/2$ that contains both trajectories.

The problem introduced in this paper is that of (k, δ)-anonymizing a database of trajectories of moving objects.

Problem 1 ((k, δ)-anonymity): Given a dataset of trajectories $\mathcal{D}$, an uncertainty threshold δ and an anonymity threshold k, the problem of (k, δ)-anonymity requires to transform $\mathcal{D}$ in a dataset $\mathcal{D}'$, such that for each trajectory $\tau \in \mathcal{D}'$ it exists a (k, δ)-anonymity set $S \subseteq \mathcal{D}'$, $\tau \in S$; and the distortion between $\mathcal{D}$ and $\mathcal{D}'$ is minimized. □

In the following we assume that all the trajectories in the dataset $\mathcal{D}$ have the *same sampling time*. Since in Definition 1 trajectories are defined on the continuous time (i.e., they are assumed to move along a straight line and at a constant speed, between two consecutive observations), this assumption does not produce any loss of generality.

IV. Techniques for Trajectory Anonymity

In the following we discuss various techniques that could be used to enforce trajectory anonymity. We start discussing the basic techniques used in the classical k-anonymity setting, *generalization* and *suppression* [27], then we discuss the *condensation* approach [28], and finally we introduce the technique adopted in this paper, namely *space translation*.

According to Definition 2, two trajectories to be co-localized must be defined over the same time interval. Although in real-world data it is quite unusual to have two trajectories starting and ending at the exact same time instants, in practice this problem can be tackled by allowing small time gaps, or by selecting coarser time samplings, or more in general, by introducing small information loss that is however necessary to achieve (k, δ)-anonymity. This issue will be addressed in practice in Section V-A. In the rest of this section we study techniques for enforcing (k, δ)-anonymity, always focusing on a maximal subset of trajectories that have the same time span, or, in other words, elaborating separately each single equivalence class induced by the *same time span* relation. Given a dataset of trajectories $\mathcal{D}$ and a time interval T, we denote:

$$\mathcal{D}_T = \{\tau \in \mathcal{D} | \tau \textit{ is defined exactly in } T\}.$$

A. Trajectory Suppression

Given a set of trajectories $\mathcal{D}_T$, enforcing (k, δ)-anonymity only by means of *trajectory suppression* is feasible and easy: it requires to remove all trajectories in $\mathcal{D}_T$ that do not belong to any anonymity set.

The main drawbacks of trajectory suppression is that it changes the size of the database, and that if used alone this technique usually introduces a too strong information loss. However, when combined with some other techniques, it can be very effective because, by removing few outliers, it can often enhance the overall data quality. In Section V we will embed outliers detection and suppression within our method for (k, δ)-anonymization.

B. Trajectory Generalization

In the classical k-anonymity setting, generalization of attributes that are quasi-identifiers, i.e., replacing real values with less specific but consistent values [27], is the main technique. Given the domain of an attribute there are various ways to generalize its values. For instance, ZIP code "44705" can be generalized to "$4470*$" (i.e., [44700-44709]), or to "$447**$" (i.e., [44700-44799]), or even to "$*$" corresponding to *maximum domain generalization*.

In the case of trajectories, generalizing can be achieved by coarsening space granularity, i.e., by substituting a point (x, y) with an area (e.g., a rectangle), containing (x, y). However, generalization-based anonymization techniques usually rely on predefined generalization hierarchies, whose choice heavily influences the quality and usefulness of anonymized data. In the context of moving objects, many different generalization hierarchies can be chosen, based on the underlying geography: for instance on the road network, or on the city's districts, or city's regions of interest. Which one is the best choice is difficult to know, and a poor choice could result in a poor quality of the anonymized data. Moreover, as discussed in [26], generalization-based anonymization techniques can produce high information loss due to unnecessary generalization. This problem is partially solved by the *hierarchy-free multidimensional* approach proposed in [10], that however has the main limitation of introducing inconsistency in the domain.

But the main reason for not adopting generalization techniques is that in our context we miss a concept of quasi-identifiers and thus we face the *curse of dimensionality* [29] as explained in the following. Under the same sampling time assumption, a dataset $\mathcal{D}_T$ can be easily seen as a table having one attribute (column) x_t and one attribute y_t for each $t \in T$, and one row for each trajectory $\tau \in \mathcal{D}_T$. Since in our context we must assume every column of $\mathcal{D}_T$ to be a quasi-identifier, we can easily end up with an extremely large number of quasi-identifiers, thus making very difficult to anonymize the data without an unacceptably high amount of information loss. This is confirmed by an experiment that we conducted using the *Mondrian* algorithm [10] over a real-world small dataset of trajectories (the Trucks dataset described in Section VI-A). After a pre-processing to create equivalence classes $\mathcal{D}_T$ of reasonable size, and their flattening in relational tables, we applied *Mondrian*: in the resulting anonymized dataset, almost all (99.9%) points were maximally generalized, thus yielding complete information loss. Another important advantage of our approach over generalization-based techniques is that the final anonymized dataset has the same spatial granularity for all the trajectories as the original dataset.

C. Condensation

The condensation approach introduced in [28] is a perturbation-like approach which aims at preserving the inter-attribute correlations of data. It starts by partitioning the original data into clusters of exactly k elements, then it regenerates, for each group, a set of k fake elements that approximately preserve the distribution and covariance of the original group. The record regeneration algorithm tries to preserve the eigenvector and eigenvalues of each group. The general idea is that valid data mining models (in particular, classification models) can be built from the reconstructed data without significant loss of accuracy. Condensation has been applied by the same authors also to sequences [30]. Developing condensation-like anonymization of trajectories could be an interesting approach. However, our objective is keep the data as close to the original as possible, exploiting the inherent uncertainty of position data to reduce the amount of distortion needed to anonymize data. Moreover condensation

is data mining (classification) oriented and its quality strongly depends on the subsequent data mining analysis performed on the perturbed data: i.e., the same anonymization does not work well for all possible subsequent analyses, and it is not easy to assume that the analysis to be performed is always known in advance. When the objective is to enforce (k, δ)-anonymity with minimum distortion (Problem 1) both generalization and condensation seem not to be good options.

D. Space Translation for (k, δ)-anonymity

Enforcing (k, δ)-anonymity by means of space translation involves moving some trajectory points from the original location to another location. The objective is to obtain (k, δ)-anonymity while keeping original and translated routes as similar as possible. Since this is the objective, a metric measuring the distortion is needed. The natural choice is the sum of point-wise distances between the original and translated trajectories as defined next. In the following, given $(x, y, t) \in \tau$, we denote the (x, y) position of τ at time t as $\tau[t]$, the x position of τ at time t as $\tau[t](x)$, and similarly for y.

Definition 4 (Translation distortion): Let $\tau' \in \mathcal{D}'_T$ be the translated version of $\tau \in \mathcal{D}_T$. The translation distortion cost of τ' w.r.t. τ is $TD(\tau, \tau') = \sum_{t \in T} Dist(\tau[t], \tau'[t])$. The total distortion of anonymized dataset $\mathcal{D}'_T$ w.r.t. $\mathcal{D}_T$ is defined as $TTD(\mathcal{D}_T, \mathcal{D}'_T) = \sum_{\tau \in \mathcal{D}_T} TD(\tau, \tau')$. □

Problem 2 ((k, δ)-anonymity by space translation): Given a dataset of trajectories $\mathcal{D}_T$ all defined over the same time interval T, an uncertainty threshold δ and an anonymity threshold k, transform $\mathcal{D}_T$ into $\mathcal{D}'_T$ such that:

- $\mathcal{D}'_T$ is the same set of trajectories $\mathcal{D}_T$, possibly containing space translated points,
- $\mathcal{D}'_T$ is (k, δ)-anonymous, and
- total translation distortion $TTD(\mathcal{D}, \mathcal{D}')$ is minimized.

□

Theorem 1: (k, δ)-anonymity by space translation problem is NP-hard.

PROOF. can be found in our technical report [31]. □

Faced with the hardness of our problem, we propose a two-stage greedy method for providing (k, δ)-anonymity: in the first stage we find clusters of trajectories containing at least k members, and in the second stage we apply space translation to move all the trajectories of a cluster within an uncertainty tube of radius $\delta/2$, making the cluster become an anonymity set according to Definition 3.

More in details, in the first step we produce *Candidate optimal clustering*, i.e., clustering such that each cluster has a population size between k and $2k - 1$. In fact if the cluster size is at least $2k$, then it can be further divided into at least two sub-clusters satisfying the k-member constraint, and yielding less distortion as proven in [24]. Moreover, since under the same time sampling and same time span assumptions trajectories can be seen as vectors, the distance functions we adopt is the usual *Euclidean distance*.

Once we have extracted a candidate optimal clustering, the next issue is how to minimize the cost of space translation needed to transform each cluster in an anonymity set of trajectories. The next Lemma provides the minimum distortion space translation for the case $\delta = 0$.

Lemma 1 (Minimum Distortion Space Translation): For any cluster p_i of a given candidate optimal clustering $\mathcal{P} = \{p_1, \ldots, p_n\}$ of a dataset $\mathcal{D}_T$, and given $\delta = 0$, the minimum distortion is obtained when all points are moved to the arithmetic mean of the cluster, denoted $\tau_c[t]$, for each $t \in T$, i.e.:
$\tau_c[t](x) = \frac{\sum_{\tau \in p_i} \tau[t](x)}{|p_i|}$ and $\tau_c[t](y) = \frac{\sum_{\tau \in p_i} \tau[t](y)}{|p_i|}$.

PROOF. can be found in our technical report [31]. □

From the lemma above it follows that for any given candidate optimal clustering it is easy to find minimum cost space translation for (k, δ)-anonymity, when $\delta = 0$. However, when $\delta > 0$ there is no analytical expression minimizing the distortion. In fact, according to Proposition 1, all points must be moved within an uncertainty disk of radius $\delta/2$, and points that are already inside such disk have null translation cost regardless of the distance to the disk center. So, even slight changes in the position of the disk, can significatively change the surface of distortion function. Thus, for the cases $\delta > 0$, our strategy for $SpaceTranslation$ is as follows:

1) first obtain cluster center as for $\delta = 0$ (Lemma 1) ;
2) then move points lying outside the disk onto the disk perimeter, along the direction from the original location to the disk center.

Algorithm 1 summarizes the generic two-stage method that we propose. Given a dataset of trajectories $\mathcal{D}$, the algorithm applies to each equivalence class of same time span $\mathcal{D}_T$ existing in $\mathcal{D}$.

Algorithm 1 Two-stage Method for (k, δ)-anonymity

Input: $\mathcal{D}_T$, k, δ
Output: $\mathcal{D}'_T$
1: $\gamma \leftarrow CandidateOptimalClustering(\mathcal{D}_T, k)$;
2: $\mathcal{D}'_T \leftarrow SpaceTranslation(\gamma, \delta)$;
3: **return** $\mathcal{D}'_T$;

The method is generic since it allows different approaches and heuristics in the first step. Thus, the most important issue we face now, is how to find the best candidate optimal clustering. Unfortunately, the number of all such clusterings is exponential in the size of $\mathcal{D}_T$. So, we need to resort to sub-optimal clustering schemes.

V. THE $\mathcal{NWA}$ ALGORITHM

In order to assess which clustering approach is most suitable for our purposes, we have extended a large variety of well known clustering schemes to make them handle trajectories and the constraint that each cluster must have a population of at least k and at most $2k - 1$ elements. We have prototyped and experimentally compared them. For lack of space we must omit this part of our investigation. However the interested reader can refer to our technical report [31]. The result of such

preliminary experimentation was that a *Greedy Clustering (*GC*)* scheme represents the best trade-off between effectiveness and efficiency, and thus it is chosen as the building block for our method.

GC iteratively selects a pivot trajectory and makes a cluster out of it and of its $k-1$ unvisited nearest neighbors, starting from a random pivot and choosing next ones as the farthest unvisited trajectories w.r.t. previous pivots. Being simple and extremely efficient, GC allows us to iteratively repeat it until clusters satisfying some criteria of compactness are built. Starting from this clustering scheme, and further enhancing it with techniques aimed at improving its effectiveness and at making it a working anonymization tool for real-world data and applications, we obtain our $\mathcal{NWA}$ method summarized in Algorithm 2.

$\mathcal{NWA}$ is developed along three main phases:

- **Pre-processing:** aimed at enforcing larger equivalence classes of trajectories w.r.t. *same time span*;
- **Clustering:** based on GC method and enhanced with techniques to keep low the radius of produced clusters, at the price of suppressing some outlier trajectories;
- **Space Translation:** transforming each cluster found into a (k, δ)-anonymity set.

Algorithm 2 $\mathcal{NWA}$

Input: $\mathcal{D}, k, \delta, \pi$
Output: $\mathcal{D}'$

1: $initialize(Max\_Trash)$;
2: $\mathcal{D}' \leftarrow \emptyset$;
3: $\mathcal{D}^{ec} \leftarrow \mathcal{NWA}_{preproc}(\mathcal{D}, \pi)$;
4: **for all** $\mathcal{D}_T \in \mathcal{D}^{ec}$ **do**
5: **if** $|\mathcal{D}_T| \geq k$ **then**
6: $Trash\_quota(T) \leftarrow \left\lfloor \frac{|\mathcal{D}_T|}{|\mathcal{D}|} * Max\_Trash \right\rfloor$;
7: $\gamma \leftarrow \mathcal{NWA}_{clust}(\mathcal{D}_T, k, Trash\_quota(T))$;
8: $\mathcal{D}' \leftarrow \mathcal{D}' \cup SpaceTranslation(\gamma, \delta)$;
9: **return** $\mathcal{D}'$;

The input of the algorithm are a database of trajectories $\mathcal{D}$, an anonymity threshold k, an uncertainty threshold δ, and the time granularity π used in the pre-processing step to create equivalence classes of trajectories, as explained in the next section. The output of the algorithm is a (k, δ)-anonymized database $\mathcal{D}'$. Moreover, $\mathcal{NWA}$ makes use of an additional parameter, Max_Trash, that is hidden to the user and automatically estimated by the algorithm in line 1 (in our experiments it was set to 10% of the dataset size): Max_Trash bounds the maximum allowed *trash*, i.e., outlier trajectories to be suppressed.

A. Pre-processing

The fist task of $\mathcal{NWA}$ is the partitioning of the input database into equivalence classes w.r.t. time span, i.e. groups containing all the trajectories that have the same starting time and the same ending time. As mentioned in Section IV, if performed on the raw input data this often produces a large number of very small equivalence classes, possibly leading to very low quality (k, δ)-anonymization. In order to overcome this problem, we developed a simple pre-processing procedure, summarized in Algorithm 3, able to enforce larger equivalence classes at the price of a small information loss. The preprocessing is driven by an integer parameter π: only one timestamp every π can be the starting or ending point of a trajectory. For instance, if the original data was sampled at a frequency of one minute, and $\pi = 60$, all trajectories are pre-processed in such a way that they all start and end at full hours. To do that, the first and the last *suitable* timestamps occurring in each trajectory are detected (lines 3 and 4), and then all the points of the trajectory that do not lay between them are removed (line 6). Finally, after the transformation trajectories are partitioned into equivalence classes w.r.t. their new starting and ending points (line 8).

Algorithm 3 $\mathcal{NWA}_{preproc}$

Input: $\mathcal{D}, \pi$
Output: $\mathcal{D}^{ec}$

1: **for all** $\tau \in \mathcal{D}$ **do**
2: Let $[t_b, t_e]$ be the time span of τ;
3: $i \leftarrow \min\{\ t \mid t \geq t_b \wedge t\ mod\ \pi = 0\}$;
4: $j \leftarrow \max\{\ t \mid t \leq t_e \wedge t\ mod\ \pi = 0\}$;
5: **if** $i \leq j$ **then**
6: $\tau' \leftarrow \tau[i, j]$; // Project τ over $[i, j]$
7: **insert** τ' **in** $\mathcal{D}_{[i,j]}$;
8: $\mathcal{D}^{ec} \leftarrow \bigcup \left\{\ \mathcal{D}_{[i,j]} \mid i\ mod\ \pi = 0 \wedge j\ mod\ \pi = 0\right\}$;
9: **return** $\mathcal{D}^{ec}$;

B. Clustering

At a very general level, the clustering procedure, named $\mathcal{NWA}\_clust$ (Algorithm 4), follows the same structure of GC, by selecting a sequence of *pivot* trajectories that play the role of cluster centers, each one chosen as the farthest trajectory from the previous pivot (excepted the first one, chosen as the farthest trajectory from the dataset center, see lines 4 and 6); forming a cluster of exactly k trajectories around each pivot with its $(k-1)$-nearest neighbors (line 7); and, finally, assigning each remaining object to its closest pivot (line 14). The difference introduced in $\mathcal{NWA}\_clust$ is a constraint added to the clusters to be formed, i.e., they must have radius not larger than a threshold max_radius, both when clusters are created (lines 8–12) and when they are enlarged with the remaining objects (lines 15–17). max_radius is automatically initialized by the algorithm in step 1 (in our experiments it was simply set to 0.5% of the semi-diagonal of the spatial minimum bounding box of the dataset). When a cluster cannot be created around a new pivot, the latter is simply *deactivated* (line 12: $Active$ is the set of actual acceptable pivots) — i.e., it will not be used as pivot but, in case, it can be used in the future as member of some other cluster — and the process goes on with the next pivot. When a remaining object cannot be added to any cluster without violating max_radius, it is simply trashed (line 17). We notice that this process can lead to solutions with a too

large trash, in which case the whole procedure is restarted from scratch relaxing the max_radius constraint (line 18, in our experiments implemented by multiplying max_radius by a factor 1.5), reiterating the operation till a clustering with sufficiently small trash is obtained (line 19). At the end, the set of clusters obtained is returned as output, thus implicitly discarding the trashed trajectories.

Algorithm 4 $\mathcal{NWA}_{clust}$

Input: $\mathcal{D}, k, Trash_{max}$
Output: γ

1: $initialize(max\_radius)$;
2: **repeat**
3: $Active \leftarrow \mathcal{D}$; $Clustered \leftarrow \emptyset$; $Pivots \leftarrow \emptyset$; $Trash \leftarrow \emptyset$;
4: $\tau_p \leftarrow$ average trajectory of $\mathcal{D}$;
5: **while** $Active \neq \emptyset$ **do**
6: $\tau_p \leftarrow \arg\max_{\tau \in Active} Dist(\tau_p, \tau)$;
7: $c_{\tau_p} \leftarrow \{\tau_p\} \cup \{k-1$ nearest neighbors of τ_p in $\mathcal{D} \setminus Clustered\}$;
8: **if** $\max_{\tau \in c_{\tau_p}} Dist(\tau_p, \tau) \leq max\_radius$ **then**
9: $Active \leftarrow Active \setminus c_{\tau_p}$;
10: $Clustered \leftarrow Clustered \cup c_{\tau_p}$;
11: $Pivots \leftarrow Pivots \cup \{\tau_p\}$;
12: **else** $Active \leftarrow Active \setminus \{\tau_p\}$;
13: **for all** $\tau \in \mathcal{D} \setminus Clustered$ **do**
14: $\tau_p \leftarrow \arg\min_{\tau' \in Pivots} Dist(\tau', \tau)$;
15: **if** $Dist(\tau_p, \tau) \leq max\_radius$ **then**
16: $c_{\tau_p} \leftarrow c_{\tau_p} \cup \{\tau\}$;
17: **else** $Trash \leftarrow Trash \cup \{\tau\}$;
18: $increase(max\_radius)$;
19: **until** $|Trash| \leq Trash_{max}$;
20: **return** $\{c_{\tau_p} | \tau_p \in Pivots\}$;

From another viewpoint, we can see the whole process as a constrained clustering task, where the maximum trash size expresses a hard constraint, while the maximum cluster radius is a soft constraint that can be relaxed as much as needed. Indeed max_radius handles trade-off between quality and running time, since small values might lead to more compact clusters, but possibly requiring more iterations to reach a feasible value for max_radius. If running time is not a crucial issue, max_radius becomes a non-critical parameter, and can be initialized to a very small value.

VI. Experimental Evaluation

In this section we report the empirical evaluation we have conducted in order to assess the performance of our method, in terms of the quality of the data maintained in the (k, δ)-anonymization process, and in terms of efficiency. In particular, when measuring data quality, we are interested in the differences holding between the original data $\mathcal{D}$ and its (k, δ)-anonymized version, $\mathcal{D}'$. For this purpose, we adopt both objective measures of information distortion (Section VI-B), and more usability-oriented measures, such as the difference in the output of the same spatio-temporal range queries executed on $\mathcal{D}$ and $\mathcal{D}'$ (Section VI-C).

A. Experimental Data

We experimented on both a real-world trajectory dataset, and a synthetic one. The first one, contains 273 trajectories of real trucks movement data [32]. The second one has been generated using Brinkhoff's network-based synthetic generator of moving objects [33]: it contains 100,000 trajectories and it represents one day movement over the road-network of the city of Oldenburg (Germany). The former dataset is referred as TRUCKS, and the latter as OLDENBURG henceforth. In Table I we report the characteristics of the two datasets and the pre-processing performed on them. For each dataset $\mathcal{D}$ we report: the half-diagonal of the minimum bounding box ($MBB\_radius(\mathcal{D})$) of the spae in $\mathcal{D}$, the number of trajectories ($|\mathcal{D}|$), the pre-processing step used (π), the resulting number of equivalence classes w.r.t. same time span relation ($|\mathcal{D}^{ec}|$), the maximum time span and maximum population of an equivalence class. It should be noted that the TRUCKS dataset is much sparser than OLDENBURG: it contains a much smaller number of trajectories, moving in a larger space.

TABLE I
DATASETS USED IN THE EXPERIMENTS.

$\mathcal{D}$	MBB $radius(\mathcal{D})$	$\lvert\mathcal{D}\rvert$	π	$\lvert\mathcal{D}^{ec}\rvert$	Max time $\mathcal{D}_T \in \mathcal{D}^{ec}$	Max pop. $\mathcal{D}_T \in \mathcal{D}^{ec}$
TRUCKS	65969.6	273	415	12	2085	103
OLDEN.	35779.3	100K	5	435	141	3499

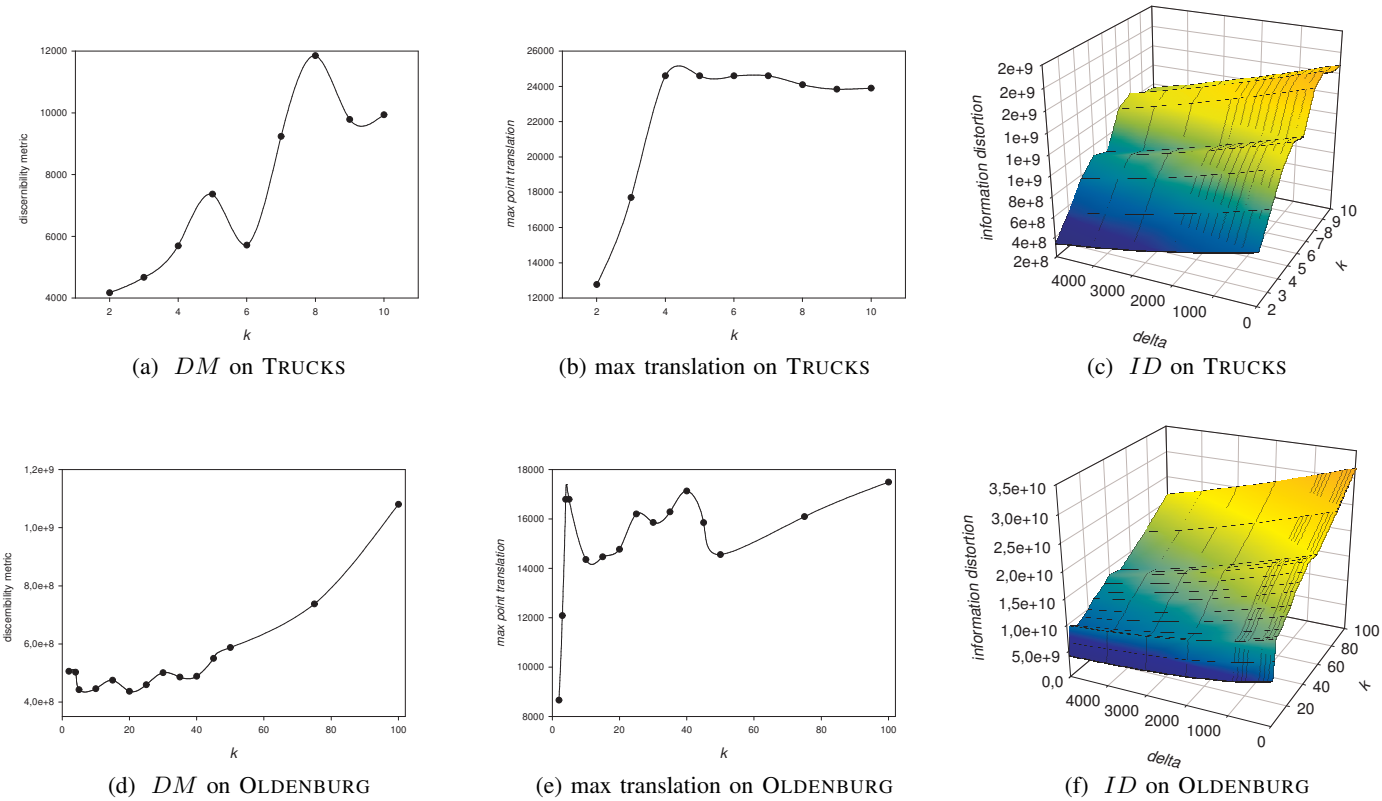
B. Discernibility and Distortion

An interesting metric introduced in [9] is *discernibility*, that measures the data quality of the anonymized dataset based on the size of each anonymity set. Given a clustering $\mathcal{P} = \{p_1, \ldots, p_n\}$ of $\mathcal{D}$, where p_n represents the trash bin, the discernibility metric is defined as:

$$DM(\mathcal{D}) = \sum_{i=1}^{n-1} |p_i|^2 + |p_n||\mathcal{D}|$$

Intuitively, discernibility represents the fact that data quality shrinks as more data elements become indistinguishable. Discernibility measures are reported in Figure 3(a) and (d). It is worth noting that $\mathcal{NWA}$ does not produce solutions exhibiting monotone discernibility w.r.t. k: this is due to the strong influence of trash on discernibility.

Since discernibility only captures the clustering step of our method, and not the space translation (hence it is independent from δ), in the following we develop an ad hoc measure of information distortion. Recall that information distortion occurs in three different ways in $\mathcal{NWA}$. First, in the pre-processing step, some initial and final points of a trajectory are possibly cut with the aim of building larger equivalence classes of trajectories having the same time span. Second, trajectories ending in the trash bin are completely removed and



(a) DM on TRUCKS (b) max translation on TRUCKS (c) ID on TRUCKS

(d) DM on OLDENBURG (e) max translation on OLDENBURG (f) ID on OLDENBURG

Fig. 3. Discernibility and distortion empirical evaluation.

will not appear in the released dataset $\mathcal{D}'$. Third, trajectories not ending in the trash bin are space-translated to achieve (k, δ)-anonymity. Our aim is to develop a unique measure able to capture these three different kinds of information distortion. We note that: (i) the finer-grained data element at which information distortion occurs is the point, and (ii) a point can be either translated or suppressed. If we consider trajectories as sets of points, and suppression as *maximal translation*, we obtain a uniform measure capturing the three different information distortions. For each trajectory $\tau \in \mathcal{D}$, let τ' be its correspondent in the (k, δ)-anonymized dataset $\mathcal{D}'$. Note that τ' can possibly be an empty set of points, in the case the trajectory τ ended in the trash bin. For each time t in which τ is defined, we measure:

$$ID(\tau[t], \tau'[t]) = \begin{cases} Dist(\tau[t], \tau'[t]) & if\ \tau'[t]\ is\ defined; \\ \Omega & otherwise. \end{cases}$$

where Ω is a constant value used to penalize removed points and corresponding to the *maximal point translation* recorded in the experiment. Given a trajectory τ, let T be the time interval in which τ is defined. The total distortion produced by τ in the (k, δ)-anonymization process is:

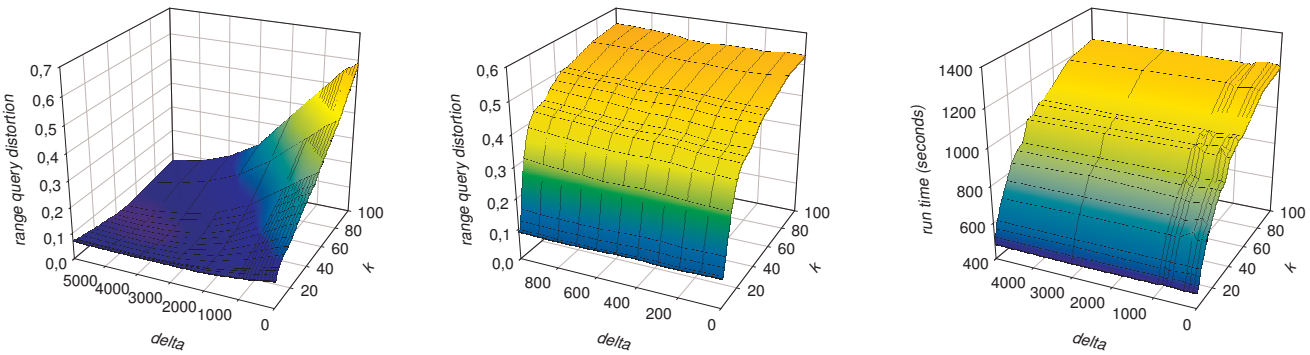
$$ID(\tau, \tau') = \sum_{t \in T} ID(\tau[t], \tau'[t])$$

For instance, given a τ defined in $[t_1, t_n]$, if τ ends in the trash, it produces an information distortion of Ω for each timestamp: i.e., $ID(\tau, \tau') = n\Omega$. Finally, the total information distortion introduced by (k, δ)-anonymizing $\mathcal{D}$ is: $ID(\mathcal{D}, \mathcal{D}') = \sum_{\tau \in \mathcal{D}} ID(\tau, \tau')$.

The results for the total information distortion are shown in Figure 3(c) and (f), where we can see that, as expected, small values of k and large values of δ yield a low distortion, that increases quasi-monotonically as k grows and δ decreases. Since our information distortion measure is based on the Ω parameter, we plot the values obtained for *maximal point translation* (used to set Ω) for $\delta = 200$, in Figure 3(b) and (e). Though there is not a strictly monotonic growth, the maximal point translation quickly reaches high values and remains relatively stable, meaning that removed points are almost always paid a high cost.

C. Spatio-temporal Range Queries

Since the purpose of releasing data is usually to query or to analyze it, the best way of measuring utility is to compare the results between queries evaluated on the original dataset $\mathcal{D}$ and its (k, δ)-anonymized version $\mathcal{D}'$. In the following we report such comparison adopting *spatio-temporal range queries with uncertainty*, according to the model of [4], that is also at the basis of our problem definition in Section III. Therefore, the same uncertainty level δ at the basis of our framework, is both used in the (k, δ)-anonymization process and in the querying process for evaluation purposes.



(a) OLDENBURG (b) OLDENBURG (c) OLDENBURG

Fig. 4. (a) range query distortion for $\mathcal{Q}_1$, (b) range query distortion for $\mathcal{Q}_2$, (c) run times.

In [4] it is introduced a set of six (Boolean) predicates that give a qualitative description of the relative position of a moving object τ with respect to a region R, within a given time interval $[t_b, t_e]$. In particular we are interested in the condition $inside(R, \tau)$. Since the location of the object changes continuously, we may ask if such condition is satisfied *sometime* or *always* within $[t_b, t_e]$; moreover, due to the uncertainty, the object may *possibly* satisfy the condition or it may *definitely* do so, at a particular timestamp $t \in [t_b, t_e]$. If there exists some possible motion curve f_{PMC^τ} (recall Definition 1) which at the time t is inside the region R, there is a possibility that the moving object will take f_{PMC^τ} as its actual motion and will be inside R at t. However, the moving object may have chosen another possible motion curve as its actual motion. Similarly, if every possible motion curve f_{PMC^τ} is inside the region R at the time t, then regardless of which one describes the actual objects motion, the object is guaranteed to be inside the region R at time t. Thus, we have two domains of quantification, with two quantifiers in each.

In the following, we focus only on the two extreme cases, namely $Possibly\_Sometime\_Inside$ (P_S_I), corresponding to a double $\exists$, and $Definitely\_Always\_Inside$ (D_A_I), corresponding to a double $\forall$. More formally:

$$Possibly\_Sometime\_Inside(\tau, R, t_b, t_e) \equiv$$
$$\equiv (\exists f_{PMC^\tau})(\exists t \in [t_b, t_e]) inside(R, f_{PMC^\tau}(t), t)$$
$$Definitely\_Always\_Inside(\tau, R, t_b, t_e) \equiv$$
$$\equiv (\forall f_{PMC^\tau})(\forall t \in [t_b, t_e]) inside(R, f_{PMC^\tau}(t), t)$$

In the experiments, we compare the results of the following queries over $\mathcal{D}$ and its (k, δ)-anonymized version $\mathcal{D}'$.

Query $\mathcal{Q}_1$:

```
SELECT COUNT(*)
FROM MOD
WHERE P_S_I(MOD.traj,R,t_b,t_e)
```

Query $\mathcal{Q}_2$:

```
SELECT COUNT(*)
FROM MOD
WHERE D_A_I(MOD.traj,R,t_b,t_e)
```

In Figure 4(a) we report the measure *range query distortion*, i.e., $|\mathcal{Q}_1(\mathcal{D}) - \mathcal{Q}_1(\mathcal{D}')|/max(\mathcal{Q}_1(\mathcal{D}), \mathcal{Q}_1(\mathcal{D}'))$ for various values of k and δ, and averaged over 1000 different runs with randomly chosen circular regions R having radius between 500 and 5000, and randomly chosen time interval $[t_b, t_e]$ with duration ranging from 2 to 8 hours. The experimental results show a very low distortion for a wide range of values of δ and k (in most cases below 10%), raising only for high values of k in combination with small values of δ. Similarly, in Figure 4(b) the distortion for $\mathcal{Q}_2$ is reported. In this case we obtain higher error (even if always below 60%), as the high selectivity of $\mathcal{Q}_2$ makes it very sensitive to any data distortion, and thus it is intrinsically harder for our method geared on space translation.

D. Efficiency

The algorithm was implemented in C, and all experiments were performed on a Intel Xeon 2Ghz processor with 1Gb of RAM over a Linux 2.6.14 platform. Run time measurements are reported in Figure 4(c). The results confirm that our prototype is very robust and efficient, as it is able to (k, δ)-anonymize in few minutes a dataset containing 100k trajectories (approx. 350Mb). Performances decrease as k grows, while they are unaffected by δ.

Our software (source code and executables) can be downloaded from: `www-kdd.isti.cnr.it/NWA/`.

VII. CONCLUSIONS

In this paper we introduced the novel concept of (k, δ)-anonymity for privacy preserving data publication from moving objects databases, that exploits the inherent uncertainty of location in order to reduce the amount of distortion needed to anonymize data. We deeply characterized the problem and we developed a simple, yet effective and efficient, method. Although experimental results show that our method introduces reasonable distortion, there is still room for improvements. The rigid pre-processing described in Section V-A could be avoided by adopting a time-tolerant distance function, such as EDR [34], for the clustering step. We are developing this new method. Also more sophisticated techniques to handle the trade-off between cluster radius and trash rate are under investigations. In this paper we assumed a uniform uncertainty

level δ for all the moving points. In some applications this could not be the case, and different moving objects could have different uncertainty level δ. Our method can be straightforwardly extended to deal with this situation by taking, for each cluster, the minimum δ appearing in the cluster (but more sophisticated techniques could be devised).

It has been recently recognized in [7] and in many other works, that k-anonymity alone does not put us on the safe side, because although one individual is hidden in a group (thanks to equal values of the *quasi-identifiers*), if the group has not enough diversity of the *sensitive attributes* then an attacker can still associate one individual to sensitive information. However, in the context of moving object data the problem is very challenging, because position is a peculiar kind of information that could be considered to be sensitive and quasi-identifier at the same time. Therefore, since anonymity requires similarity on the quasi identifiers, and diversity requires dissimilarity on the sensitive information, it seems that in the case of moving object data they are two conflicting goals. In this paper we did not consider any concept of quasi-identifier for trajectories, and thus we did not tackle the diversity problem: these are interesting and open research problems that deserve deep investigation.

ACKNOWLEDGMENT

This work was carried out during the tenure of Osman Abul's ERCIM fellowship at ISTI - CNR. Francesco Bonchi and Mirco Nanni are supported by the EU project GeoPKDD (IST-6FP-014915). The authors wish to thank Kristen LeFevre for providing the implementation of the *Mondrian* algorithm.

REFERENCES

[1] C. Bettini, X. S. Wang, and S. Jajodia, "Protecting Privacy Against Location-Based Personal Identification." in *Proc. of the Second VLDB Workshop on Secure Data Management (SDM'05)*.

[2] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information (abstract)," in *Proc. of the 17th ACM Symp. on Principles of Database Systems (PODS'98)*.

[3] O. Wolfson, S. Chamberlain, S. Dao, L. Jiang, and G. Mendez, "Cost and imprecision in modeling the position of moving objects." in *Proc. of the 14th IEEE Int. Conf. on Data Engineering (ICDE'98)*.

[4] G. Trajcevski, O. Wolfson, K. Hinrichs, and S. Chamberlain, "Managing uncertainty in moving objects databases." *ACM Trans. Database Syst.*, vol. 29, no. 3, pp. 463–507, 2004.

[5] D. Pfoser and C. S. Jensen, "Capturing the uncertainty of moving-object representations." in *Proc. of the 6th International Symp. on Advances in Spatial Databases (SSD'99)*.

[6] A. Jain, E. Y. Chang, and Y.-F. Wang, "Adaptive stream resource management using kalman filters." in *Proc. of the 2004 ACM SIGMOD Int. Conf. on Management of Data*.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity," in *Proc. of the 22nd IEEE Int. Conf. on Data Engineering (ICDE'06)*.

[8] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Anonymizing tables." in *Proc. of the 10th Int. Conf. on Database Theory (ICDT'05)*.

[9] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymity," in *Proc. of the 21st IEEE Int. Conf. on Data Engineering (ICDE'05)*.

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity." in *Proc. of the 22nd IEEE Int. Conf. on Data Engineering (ICDE'06)*.

[11] R. H. Güting and M. Schneider, *Moving Objects Databases*. Morgan Kaufmann, 2005.

[12] G. Kollios, D. Gunopulos, and V. J. Tsotras, "On indexing mobile objects." in *Proc. of the 18th ACM Symp. on Principles of Database Systems (PODS'99)*.

[13] P. K. Agarwal, L. Arge, and J. Erickson, "Indexing moving points." in *Proc. of the 19th ACM Symp. on Principles of Database Systems (PODS'00)*.

[14] M. Hadjieleftheriou, G. Kollios, V. J. Tsotras, and D. Gunopulos, "Indexing spatiotemporal archives." *VLDB Journal*, vol. 15, no. 2, pp. 143–164, 2006.

[15] A. P. Sistla, O. Wolfson, S. Chamberlain, and S. Dao, "Modeling and querying moving objects." in *Proc. of the 13th IEEE Int. Conf. on Data Engineering (ICDE'97)*.

[16] D. Pfoser, C. S. Jensen, and Y. Theodoridis, "Novel approaches in query processing for moving object trajectories." in *Proc. of the 26th Int. Conf. on Very Large Databases (VLDB'00)*.

[17] R. H. Güting, M. H. Böhlen, M. Erwig, C. S. Jensen, N. A. Lorentzos, M. Schneider, and M. Vazirgiannis, "A foundation for representing and quering moving objects." *ACM Trans. Database Syst.*, vol. 25, no. 1, pp. 1–42, 2000.

[18] O. Wolfson, A. P. Sistla, S. Chamberlain, and Y. Yesha, "Updating and querying databases that track mobile units." *Distributed and Parallel Databases*, vol. 7, no. 3, pp. 257–387, 1999.

[19] M. Gruteser and D. Grunwald, "Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking." in *Proc. of the First Int. Conf. on Mobile Systems, Applications, and Services (MobiSys 2003)*.

[20] B. Gedik and L. Liu, "Location Privacy in Mobile Systems: A Personalized Anonymization Model." in *Proc. of the 25th Int. Conf. on Distributed Computing Systems (ICDCS'05)*.

[21] H. Kido, Y. Yanagisawa, and T. Satoh, "Protection of Location Privacy using Dummies for Location-based Services." in *Proc. of the 21st IEEE Int. Conf. on Data Engineering (ICDE'05)*.

[22] A. R. Beresford and F. Stajano, "Mix Zones: User Privacy in Location-aware Services." in *Proc. of the Second IEEE Conf. on Pervasive Computing and Communications Workshops (PERCOM'04)*.

[23] M. Gruteser and X. Liu, "Protecting Privacy in Continuous Location-Tracking Applications." *IEEE Security & Privacy Magazine*, vol. 2, no. 2, pp. 28–34, 2004.

[24] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control." *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 189–201, 2002.

[25] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving anonymity via clustering." in *Proc. of the 25th ACM Symp. on Principles of Database Systems (PODS'06)*.

[26] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques." in *Proc. of the 12th Int. Conf. Database Systems for Advanced Applications, (DASFAA'07)*.

[27] L. Sweeney, "k-anonymity privacy protection using generalization and suppression," *International Journal on Uncertainty Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, 2002.

[28] C. C. Aggarwal and P. S. Yu, "A condensation approach to privacy preserving data mining." in *Proc. of the 9th Int. Conf. on Extending Database Technology, (EDBT'04)*.

[29] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality." in *Proc. of the 31st Int. Conf. on Very Large Databases (VLDB'05)*.

[30] C. C. Aggarwal and P. S. Yu, "On anonymization of string data." in *Proc. of the 2007 SIAM Int. Conf. on Data Mining*.

[31] O. Abul, F. Bonchi, and M. Nanni, "$\mathcal{N}$ever $\mathcal{W}$alk $\mathcal{A}$lone: Trajectory anonymity via clustering," ISTI-C.N.R., Tech. Rep. ISTI 2007-TR-010, March 2007, http://puma.isti.cnr.it/download.php?doc=cnr.isti/cnr.isti/2007-TR-010/%2007-TR-010.pdf.

[32] E. Frentzos, K. Gratsias, N. Pelekis, and Y. Theodoridis, "Nearest neighbor search on moving object trajectories." in *Proceedings of the 9th Int. Symp. on Advances in Spatial and Temporal Databases (SSTD'05)*.

[33] T. Brinkhoff, "Generating traffic data." *IEEE Data Eng. Bull.*, vol. 26, no. 2, pp. 19–25, 2003.

[34] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories." in *Proc. of the 2005 ACM SIGMOD Int. Conf. on Management of Data*.