Privacy Preserving Publication of Moving Object Data

Francesco Bonchi

Yahoo! Research
Avinguda Diagonal 177, Barcelona, Spain
bonchi@yahoo-inc.com

Abstract. The increasing availability of space-time trajectories left by location-aware devices is expected to enable novel classes of applications where the discovery of consumable, concise, and actionable knowledge is the key step. However, the analysis of mobility data is a critic task by the privacy point of view: in fact, the peculiar nature of location data might enable intrusive inferences in the life of the individuals whose data is analyzed. It is thus important to develop privacy-preserving techniques for the publication and the analysis of mobility data.

This chapter provides a brief survey of the research on *anonymity preserving data publishing of moving objects databases.*

While only few papers so far have tackled the problem of anonymity in the off-line case of publication of a moving objects database, rather large body of work has been developed for anonymity on relational data on one side, and for location privacy in the on-line, dynamic context of *location based services* (LBS), on the other side. In this chapter we first briefly review the basic concepts of k-anonymity on relational data. Then we focus on the body of research about privacy in LBS: we try to identify some useful concepts for our static context, while highlighting the differences, and discussing the inapplicability of some of the LBS solutions to the static case. Next we present in details some of the papers that recently have attacked the problem of moving objects anonymization in the static context. We discuss in details the problems addressed and the solutions proposed, highlighting merits and limits of each work, as well as the various problems still open.

1 Introduction

Recent years have witnessed the pervasiveness of location-aware devices, e.g., GSM mobile phones, GPS-enabled PDAs, location sensors, and active RFID tags. This new capability of localizing moving objects and persons enables a wide spectrum of possible novel applications that were simply infeasible only few years ago. Those applications can be roughly divided in two large groups:

on-line: such as monitoring the moving objects, real-time analysis of their motion patterns, and development of location-based services;

C. Bettini et al. (Eds.): Privacy in Location-Based Applications, LNCS 5599, pp. 190–215, 2009.
© Springer-Verlag Berlin Heidelberg 2009

off-line: such as the collection of the traces left by these moving objects, the off-line analysis of these traces with the aim of extracting behavioral knowledge in support of, e.g., mobility-related decision making processes, sustainable mobility, and intelligent transportation systems [1].

The latter scenario, which is the focus of this paper, is rapidly gaining a great deal of attention as witnessed by the amount spatio-temporal data mining techniques that have been developed in the last years [2,3,4,5,6,7]. Such techniques may be used, for instance, by governments to measure day-by-day variability in mobility behavior[1], by researchers to study masses mobility behavior[2] or by companies to track employees and maximize their efficiency[3]. 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.

More concretely, consider a *traffic management application* on a city road network, where the trajectories of vehicles equipped with GPS devices are recorded and analyzed by the city municipality traffic management office. This is not unrealistic: in the context of the GeoPKDD[4] project, we received a dataset of this kind from the city of Milan (Italy). Indeed, many citizens accept to equip their car with GPS devices, because this way they obtain a substantial discount on the mandatory car insurance. Suppose now that the office owning the data is going to outsource the data mining analysis to an external laboratory. In a naïve tentative of preserving anonymity, the car identifiers are not disclosed but instead replaced with pseudonyms. However, as shown in [8], such operation is insufficient to guarantee anonymity, since location represents a property that in some circumstances can lead to the identification of the individual. For example, if Bob is known to follow almost every working morning the same route, it is very likely that the starting point is Bob's home and the ending point is his working place. Joining this information with some telephone directories we can easily link Bob's trajectory to Bob's identity.

As another example, a recent paper by Terrovitis and Mamoulis analyzes the case of a company in Hong Kong, called Octopus[5], that collects daily trajectory data of Hong Kong residents who use Octopus smart RFID card [9]. The data could be disclosed to a third external party, e.g., for *reach of poster analysis* in Hong Kong. The reach of a poster defines the percentage of people who have at least one contact with a given poster (or a posters network) within a specified period of time. The reach allows to determine the optimal duration of some advertisement and to tune the formation of poster networks[6].

[1] See `http://www.fhwa.dot.gov/ohim/gps/`

[2] See the projects of the sens*able* City Lab `http://senseable3.mit.edu/`

[3] `http://www.denverpost.com/headlines/ci_4800440`

[4] GeoPKDD: Geographic Privacy-aware Knowledge Discovery and Delivery, EU project IST-6FP-014915, webpage: `http://www.geopkdd.eu/`

[5] `http://www.octopuscards.com`

[6] See for instance the Swiss Poster Research `http://www.spr-plus.ch/`

As Terrovitis and Mamoulis [9] pointed out, when a person, say Alice, uses her Octopus card to pay at different convenience stores that belong to the same chain (e.g., 7-Eleven), by collecting her transaction history in all these stores, the company can construct a subset of her complete trajectory. If this constructed trajectory uniquely identifies Alice, then by matching it with the published trajectory database (this is usually called *"linkage attack"*), even though the users identifiers may be removed, Alice still can be re-identified, as can the other locations that she visited.

The two abovementioned examples regard cases of data released to a third party, but privacy issues do not arise only when the data must be published. As it happens for any other kind of data, collecting, storing and analyzing person-specific mobility data is subject to privacy regulations[7],[8]. Therefore it is important to develop concepts and methods for collecting, storing and publishing spatio-temporal data in a way that preserves privacy of the individuals.

The research problem surveyed in this paper can be stated as *anonymity preserving data publishing of moving objects databases*. More formally, we consider a static *moving object database* (MOD) $D = \{O_1, ..., O_n\}$ that correspond to n individuals, a set of m discrete time points $T = \{t_1, ..., t_m\}$, and a function $\mathcal{T} : D \times T \rightarrow \mathcal{R}^2$, that specifies, for each object O and a time t, its position at time t. The function $\mathcal{T}$ is called the trajectory. Indeed, $\mathcal{T}(O_i)$ denotes the trajectory of object O_i, i.e., $\mathcal{T}(O_i) = \{(x_i^1, y_i^1, t_1), ..., (x_i^m, y_i^m, t_m)\}$ is O_i's trajectory, with (x_i^j, y_i^j) representing the position of O_i at time t_j. The problem is how to transform D in such a way that it satisfy some form of anonymity, while most of its original utility is maintained in the transformed database D^*.

Introduced by Samarati and Sweeney [10,11,12], the concept of k-anonymity has established, also thanks to its simplicity, as the *de facto* standard solution to prevent linkage attacks in de-identified relational databases. The idea behind k-anonymity can be described as "hiding in the crowd", as it requires that each release of data must be such that each individual is indistinguishable from at least $k - 1$ other individuals. In the classical k-anonymity framework the attributes are partitioned into *quasi-identifiers* (i.e., a set of attributes whose values can be linked to external information to reidentify the individual), and *sensitive attributes* (publicly unknown, which we want to keep private). In order to provide k-anonymity, the values of the quasi-identifiers are generalized to be less specific so that there are at least k individuals in the same group, who have the same (generalized) quasi-identifer values. Although it has been shown that the k-anonymity model presents some flaws and limitations [13], and that finding an optimal k-anonymization is NP-hard [14,15], it remains a fundamental model of privacy with practical relevance.

Unfortunately (and quite obviously), the concept of k-anonymity can not be borrowed from relational databases as it is, because in the case of moving objects much more complexity is brought in by the peculiar nature of spatio-temporal

[7] `http://www.cdt.org/privacy/eudirective/EUDirective.html`

[8] `http://www.dataprotection.ie/documents/legal/6aiii.htm`

data. In the rest of this paper we will discuss this complexity, and we will analyze in details the few papers that have attacked this problem in the last two years.

Paper Content and Structure

So far only few papers have tackled the problem of anonymity in the off-line case of publication of a moving objects database. Instead, a lot of work has been done for anonymity on relational data, and for location privacy in the on-line, dynamic context of *location based services* (LBS). In Section 2 we review the basic concepts of k-anonymity on relational data, while in Section 3 we review the body of research on privacy in LBS. We try to identify some useful concepts for our static context, while highlighting the differences, and discussing the inapplicability of some of the LBS solutions to the static case. In particular, we recall the interesting concepts of *location based quasi-identifier* and *historical k-anonymity* introduced by Bettini *et al.* in [8].

Then we focus on four papers (all very recent) that, to the best of our knowledge, are the unique that have attacked the problem of MOD anonymization so far. Since one key concept is that of quasi-identifier, we use this concept to partition the methods in two groups: methods assuming no quasi-identifier and thus anonymizing the trajectories in their whole [16,17] (Section 4) and methods assuming some form of quasi-identifer [9,18] (Section 5).

Finally, in Section 6 we draw some conclusions and discuss the open research problems.

2 Relational Data Anonymity

Samarati and Sweeney showed that the simple de-anonymization of individual sources does not guarantee protection when sources are cross-examined: a sensitive medical record, for instance, can be uniquely linked to a *named* voter record in a publicly available voter list through some shared attributes. The objective of k-anonymity is to eliminate such opportunities of inferring private information through cross linkage.

The traditional k-anonymity framework [10,11,12,19] focuses on relational tables: the basic assumptions are that the table to be anonymized contains entity-specific information, that each tuple in the table corresponds uniquely to an individual, and that attributes are divided in *quasi-identifier* (i.e., a set of attributes whose values in combination can be linked to external information to reidentify the individual to whom the information refers); and *sensitive attributes* (publicly unknown and that we want to keep secret).

According to this approach, the data holder has the duty of identifying all possible attributes in the private information that can be found in other public databases, i.e., the attributes that could be exploited by a malicious adversary by means of cross linkage (the quasi-identifier).

Once the quasi-identifier is known, the "anonymization" of the database takes place: the data is transformed in such a way that, for every combination of values of the quasi-identifier in the sanitized data, there are at least k records that share

Table 1. (a) example medical table, (b) a 2-anonymous version of table (a), (c) an alternative 2-anonymous version of table (a), and (d) a 2-anonymous version of table (a) by full-domain generalization. These example tables are borrowed from [20].

Job	Birth	Postcode	Illness
Cat1	1975	4350	HIV
Cat1	1955	4350	HIV
Cat1	1955	5432	flu
Cat1	1955	5432	fever
Cat2	1975	4350	flu
Cat2	1975	4350	fever

(a)

Job	Birth	Postcode	Illness
Cat1	*	4350	HIV
Cat1	*	4350	HIV
Cat1	1955	5432	flu
Cat1	1955	5432	fever
Cat2	1975	4350	flu
Cat2	1975	4350	fever

(b)

Job	Birth	Postcode	Illness
*	1975	4350	HIV
*	*	4350	HIV
Cat1	1955	5432	flu
Cat1	1955	5432	fever
*	*	4350	flu
*	1975	4350	fever

(c)

Job	Birth	Postcode	Illness
*	*	4350	HIV
*	*	4350	HIV
*	*	5432	flu
*	*	5432	fever
*	*	*	flu
*	*	4350	fever

(d)

those values. One equivalence class of records sharing the same quasi-identifier values is usually called *anonymity set* or *anonymization group*.

The anonymization is usually obtained by (*i*) generalization of attributes (the ones forming the quasi-identifier), and (*ii*), when not avoidable, suppression of tuples [12].

An example medical data table is given in Table 1(a). Attributes `job`, `birth` and `postcode` form the quasi-identifier. Two unique patient records (corresponding to the first two rows) may be re-identified easily since their combinations of the attributes forming the quasi-identifier are unique. The table is generalized as a 2-anonymous table in Table 1(b): here the two patient records are indistinguishable w.r.t. the quasi-identifier and thus are less likely to be re-identified by means of cross linkage.

In the literature of k-anonymity, there are two main models. One model is *global recoding* [12,19,20,21,22] while the other is *local recoding* [12,14,20]. A common assumption is that each attribute has a corresponding conceptual hierarchy or taxonomy. Generalization replaces lower level domain values with higher level domain values. A lower level domain in the hierarchy provides more details and maintains more of the original information than a higher level domain. In global recoding, all values of an attribute come from the same domain level in the hierarchy. For example, all values in `birth` date are in years, or all are in both months and years. One advantage is that an anonymous view has uniform domains, but the price to pay is higher information loss. For example, a global recoding of Table 1(a) is in Table 1(d), but it clearly suffers from overkilling generalization.

With local recoding, values may be generalized to different levels in the domain. For example, Table 2 is a 2-anonymous table by local recoding. In fact one can say that local recoding is a more general model and global recoding is a special case of local recoding. Note that, in the example, known values are replaced by unknown values (*), indicating maximum generalization, or total loss of information.

As discussed by Domingo-Ferrer and Torra in [23] the methods based on generalization and suppression suffer of various drawbacks. To overcome some of these limitations the use of *microaggregation* for k-anonymity has also been proposed [23]. Microaggregation is a concept originating from the statistical disclosure control (SDC) research community. In particular, under the name microaggregation goes a family of perturbative SDC methods that have been developed both for continuous and categorical data, and that do not require a hierarchy [23,24,25,26]. Whatever the data type, microaggregation can be operationally defined in terms of the following two steps:

- *Partition:* the set of original records is partitioned into several clusters in such a way that records in the same cluster are similar to each other and so that the number of records in each cluster is at least k.
- *Aggregation:* An aggregation operator (for example, the mean for continuous data or the median for categorical data) is computed for each cluster and is used to replace the original records. In other words, each record in a cluster is replaced by the clusters prototype.

This approach, even if under different names, e.g. *k-member clustering for k-anonymity*, has been investigated in [27], and then extended in [28,29] to deal with attributes that have a hierarchical structure. Usually, after a clustering step what is released is the centroid of each cluster together with the cardinality of the cluster.

Another similar approach is introduced by Aggarwal and Yu in [30]. *Condensation* 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 preserves 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 [31].

Some limitations of the k-anonymity model, with consequent proposals for improvement, have emerged in the literature. One first drawback is that the difference between quasi-identifiers and sensitive attributes may be sometimes vague, leading to a large number of quasi-identifiers. This problem has been studied in [32], where Aggarwal analyzes the scalability of distortion w.r.t. the number of dimensions used in the k-anonymization process, showing that for sparse data the usability of the anonymized data could sensibly decrease. A possible solution based on k-anonymity parameterized w.r.t. a given public dataset has been proposed by Atzori in [33]. Xiao and Tao in [34] propose a decomposition of quasi-identifiers and sensitive data into independently shared databases. Kifer and Gehrke [35] suggest to share anonymized marginals (statistical models such as density estimates of the original table) instead of the private table.

Another drawback of simple k-anonymity is that it may not protect sensitive values when their entropy (diversity) is low: this is the case in which a sensitive value is too frequent in a set of tuples with same quasi-identifier values after k-anonymization [13,20,36,37]. Consider Table 1(b): although it satisfies 2-anonymity property, it does not protect two patients sensitive information, HIV infection. We may not be able to distinguish the two individuals for the first two tuples, but we can derive the fact that both of them are HIV infectious. Suppose one of them is the mayor, we can then confirm that the mayor has contracted HIV. Surely, this is an undesirable outcome. Note that this is a problem because the other individual whose generalized identifying attributes are the same as the mayor also has HIV. Table 3 is an appropriate solution. Since (*,1975,4350) is linked to multiple diseases (i.e. HIV and fever) and (*,*,4350) is also linked to multiple diseases (i.e. HIV and flu), it protects individual identifications and hides the implication.

Regardless of these limitations, k-anonymity remains a widely accepted model both in scientific literature and in privacy related legislation, and in recent years a large research effort has been devoted to develop algorithms for k-anonymity (see for instance [22,38,39] just to cite some of the most relevant ones).

3 Anonymity and Location Based Services

As witnessed by the other chapters in this volume, most of the existing work about anonymity of spatio-temporal moving points has been 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.

This is the main difference with our setting where we have a static database of moving objects and we want to publish it in 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 ideally the data might 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* and *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. However, although in our context the focus is on the quality of the data, while in LBS is on the quality of the service, it should be noted that both concepts of quality are intrinsically geared on the level of precision with which positions are represented.

In the following we review some of the proposals for privacy in LBS trying to identify useful concepts for our static context, while discussing why some of the LBS solutions are not suitable for our purposes.

The concept of *location k-anonymity* for location based services was first introduced in [40] Gruteser and Grunwald, and later extended by Gedik and Liu in [41] 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 [41] 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 proposed algorithm tries to identify the smallest spatial area and time interval for each message, such that there exist at least $k-1$ other messages from different users, with the same spatial and temporal dimensions. Domingo-Ferrer applied the idea of microaggregation for k-anonymity in location-based services [42].

Kido *et al.* [43] propose 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 information is delimited by the region it belongs to. 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 [44] Beresford and Stajano introduce the concept of *mix zones*. 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 entering a mix zone in the same time period are mixed. A similar classification of areas, named *sensitivity map* is introduced by Gruteser and Liu in [45]: locations are classified as either *sensitive* or *insensitive*, and three algorithms that hide users' positions in sensitive areas are proposed.

Contrary to the notions of mixed zones and sensitivity maps, the approach introduced by Bettini *et al.* in [8] is geared on the concept of *location based quasi-identifier*, i.e., a spatio-temporal pattern that can uniquely identify one individual. More specifically, a location based quasi-identifier (LBQID) is a sequence of spatio-temporal constraints (i.e., a spatial area and a time interval) plus a recurrence formula.

Example 1 (Borrowed from [8]). A user may consider the trip from the condominium where he lives to the building where he works every morning and the trip back in the afternoon as an LBQID if observed by the same service provider for at least 3 weekdays in the same week, and for at least 2 weeks. This LBQID may be represented by the spatio-temporal pattern:

$$\langle AreaCondominium\,[7am, 8am], AreaOfficeBldg\,[8am, 9am],$$
$$AreaOfficeBldg\,[4pm, 6pm], AreaCondominium\,[5pm, 7pm]\rangle$$
$$Recurrence : 3.Weekdays * 2.Weeks$$

where the various areas such as *AreaCondominium* identify sets of points in bidirectional space possibly by a pair of intervals $[x_1, x_2][y_1, y_2]$.

Where and how LBQID are defined for each single user, is an interesting open problem not addressed in [8], and we will discuss it again later in this chapter. In [8] Bettini *et al.* simply state that the derivation process of those spatio-temporal patterns to be used as LBQID will have to be based on statistical analysis of the data about users movement history: if a certain pattern turns out to be very common for many users, it is unlikely to be useful for identifying any one of them. Since in the LBS framework there is the trusted server which stores, or at least has access to, historical trajectory data, Bettini *et al.* argue that the trusted server is probably a good candidate to offer tools for LBQID identification. However, the selection of candidate patterns may also possibly be guided directly by the users.

The goal of introducing the concept of LBQID is that of ensuring that no sensitive data is released from the trusted server receiving the requests, to a service provider, when the data can be personally identified through a LBQID. On the technical level, considering the language proposed in Example 1, a timed state automata [46] may be used for each LBQID and each user, advancing the state of the automata when the actual location of the user at the request time is within the area specified by one of the current states, and the temporal constraints are satisfied.

In [8] the concept of *historical k-anonymity* is also introduced. Given the set of requests issued by a certain user (that corresponds to a trajectory of a moving object in our terminology), it satisfies historical k-anonymity if there exist $k - 1$ *personal histories of locations* (i.e., trajectories in our terminology) belonging to $k - 1$ different users such that they are *location-time consistent* (i.e., undistinguishable). What the framework in [8] aims at, is to make sure that if a set of requests from a user matches a location based quasi-identifier then it satisfies historical k-anonymity. The idea is that if a service provider can successfully track the requests of a user through all the elements of a LBQID, then there would be at least $k - 1$ other users whose personal history of locations is consistent with these requests, and thus may have issued those requests. This is achieved by generalization in space and time, essentially by increasing the uncertainty about the real user location and time of request. Generalization is performed by an algorithm that tries to preserve historical k-anonymity of the set of requests that have matched the current partial LBQID. If for a particular request, generalization fails, (i.e., historical k-anonymity is violated), the system will try to unlink future requests from the previous ones by means of the mixed zones technique [44] that we discussed earlier.

Although defined in the LBS context this work is very relevant to the problem surveyed in this chapter, i.e., anonymity in a static database of moving objects.

In other terms, a set of trajectories that satisfies historical k-anonymity may be considered safe also in our data-publishing context. The main difference, however, is the fact that they consider data point (requests) continuously arriving, and thus they provide *on-line anonymity*. More concretely, the anonymization group of an individual is chosen once and for all, at the time this is needed, i.e., when his trajectory matches a LBQID. This means that the $k-1$ moving objects passing closer to the point of the request are selected to form the anonymization group regardless of what they will do later, as this information is not yet available. Instead in our context the information about the whole history of trajectories is available, thus we must select anonymization groups considering the trajectories in their whole. This is the main reason why the solution proposed in [8] does not seem suitable for our static data-publishing context.

As discussed at the beginning of this section the difference is that in the LBS context the emphasis is on providing the service, while in our context the emphasis is on the quality maintained in the anonymized database. As another example consider again the concept of *mix zones* previously described: it is a solution for LBS, since it provides some sort of anonymity for the users while still allowing the service to be provided correctly; but it is not a solution for data publishing, since the quality of the data is completely destroyed. Just think about a data mining analysis (e.g., finding hot routes [6], frequent mobility patterns [47] clustering trajectories [4], etc.) on a dataset “anonymized” by means of the mixed-zone technique: the results would simply be unreliable.

Summarizing, in this section we have reviewed techniques for anonymity in the LBS context, and we have discussed the differences between this context and our context: why the anonymity solutions proposed in the former are not suitable in the latter. However, many techniques developed in the LBS research community, such as location perturbation, spatio-temporal cloaking (i.e., generalization), and the personalized privacy-profile of users, should be taken in consideration while devising solutions for anonymity preserving data publishing of MODs. Also the definitions of historical k-anonymity and LBQID might be borrowed in the context of data publishing: how to do it is a challenging open problem not addressed in [8] nor in other work.

4 Methods Based on Clustering and Perturbation

The problem of defining a concept of quasi-identifier for a database of moving objects, that is both realistic and actionable, is not an easy task. On the one hand, it is the nature itself of spatio-temporal information to make the use of quasi-identifiers unlikely, or at least, challenging. The natural spatial and temporal dependence of consecutive points in a trajectory, avoids from identifying a precise set of locations, or a particular set of timestamps to be the quasi-identifier. Moreover, as argued in [18], unlike in relational microdata, where every tuple has the same set of quasi-identifier attributes, in mobility data we can not assume to have the same quasi-identifier for all the individuals. It is very likely that various moving objects have different quasi-identifiers and this should be taken into account in modeling adversarial knowledge. But as shown in [18] allowing different quasi-identifiers for different objects creates many challenges: the main one being that the anonymization groups may not be disjoint.

A different order of challenges regarding the definition of a concept of quasi-identifier for MODs, arises at the practical level. It is not clear from where and how the quasi-identifier for each single user should be defined. Both [8] and [18] argue that the quasi-identifiers may be provided directly by the users when they subscribe to the service, or be part of the users personalized settings, or they may be found by means of statistical data analysis or data mining. However, the problem of spatio-temporal quasi-identifier definition in the real-world is an open issue.

Given the aforementioned challenges, Abul *et al.* [16], and Nergiz *et al.* [17], have tackled the problem of anonymity of MODs without considering any concept of quasi-identifier, thus anonymizing trajectories as a whole. This may be considered as implicitly and conservatively assuming that the adversary may identify each user in any location at any time. Since the proposed techniques provide protection under this very conservative setting, they also provide protection under less powerful adversary.

In this setting the output of anonymization can only be a set of anonymization groups each one containing identical, or at least very similar, sets of trajectories and having size at least k. From the above consideration, it follows that a natural approach is to tackle the anonymity problem by means of clustering, more precisely *microaggregation* or *k-member clustering*, i.e., clustering with the constraint on the minimum population of each cluster.

The first work tackling the anonymization of trajectories as a constrained clustering problem is [16]. In that paper Abul *et al.* 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 not simply 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. A graphical representation of an uncertain trajectory is reported in Figure 1(a).

If another trajectory moves within the cylinder (or uncertainty area) of the given trajectory, then the two trajectory are indistinguishable from each other (or in other terms, they're a *possible motion curve* of each other). This leads to the definition of (k, δ)-anonymity for moving objects databases. More formally, given an anonymity threshold k, Abul *et al.* define a (k, δ)-*anonymity set* as a set of at least k trajectories that are co-localized w.r.t. δ. Then they show that a set of trajectories S, with $|S| \geq k$, is a (k, δ)-anonymity set if and only if there exists a trajectory τ_c such that all the trajectories in S are possible motion curves of τ_c within an uncertainty radius of $\delta/2$. 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

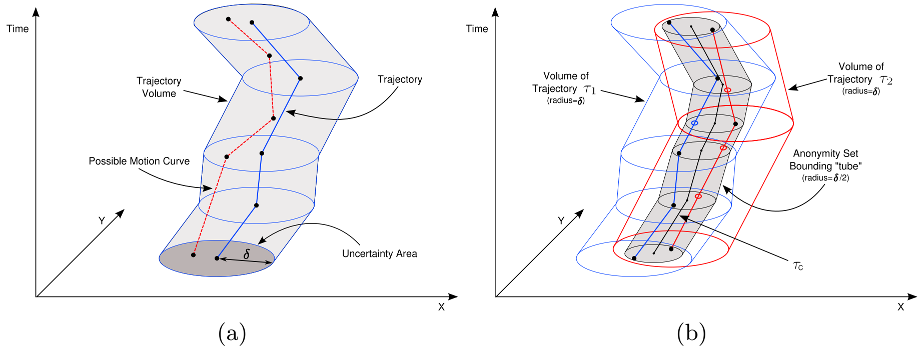

Time
Trajectory Volume
Trajectory
Possible Motion Curve
Y
δ
Uncertainty Area
X
(a)
Time
Volume of Trajectory τ1 (radius=δ)
Volume of Trajectory τ2 (radius=δ)
Anonymity Set Bounding "tube" (radius=δ/2)
Y
τC
X
(b)

Fig. 1. (a) an uncertain trajectory: uncertainty area, trajectory volume and possible motion curve. (b) an 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.

by a cylindrical volume of radius $\delta/2$. In Figure 1(b), we graphically represent this property.

The problem of (k,δ)-anonymizing a database of trajectories of moving objects requires to transform a MOD D in D^* such such that for each trajectory $\tau \in D^*$ it exists a (k,δ)-anonymity set $S \subseteq D^*$, $\tau \in S$, and the distortion between D and D^* is minimized.

In [16] a two-step method, based on clustering and perturbation, is devised to achieve (k,δ)-anonymity. In particular, as perturbation method is chosen *space translation*: i.e., slightly moving some observations in space. A suitable measure of the information distortion introduced by space translation is defined, and the problem of achieving (k,δ)-anonymity by space translation with minimum distortion is proven to be NP-hard.

In the first clustering step, the MOD D is partitioned in groups of trajectories, each group having size in the interval $[k, 2k-1]$. After having tried a large variety of clustering methods for trajectories under the k-member constraint, Abul *et al.* chose a simple greedy method as the best trade-off between efficiency and quality of the results. The resulting method, named $\mathcal{NWA}$ ($\mathcal{N}$ever $\mathcal{W}$alk $\mathcal{A}$lone), is further enhanced with ad-hoc preprocessing and outlier removal. In fact it is claimed by the authors (but also by other previous work, e.g., [29]), that outlier detection and removal might be a very important technique in clustering-based anonymization schemes: the overall quality of the anonymized database can benefit by the removal of few outlying trajectories.

The pre-processing step aims at partitioning the input database into larger 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. This is needed because $\mathcal{NWA}$ adopts Euclidean distance that can only be defined among trajectories having the same time span: if performed directly on the raw input data this often produces a large number of very small equivalence classes, possibly leading to very low quality anonymization. To overcome this problem, a simple pre-processing method is developed. The method enforces larger equivalence classes at the price of a small information loss. The pre-processing 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, and then all the points of the trajectory that do not lay between them are removed.

The greedy clustering method 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, the greedy algorithm allows to iteratively repeat it until clusters satisfying some criteria of compactness are built.

More in details, a compactness constraint is added to the greedy clustering method briefly described above: clusters to be formed must have a radius not larger than a given threshold. When a cluster cannot be created around a new pivot without violating the compactness constraint, the latter is simply *deactivated* — 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 the compactness constraint, it is considered an outlier and it is trashed. This process might lead to solutions with a too large trash, in which case the whole procedure is restarted from scratch relaxing the compactness constraint, reiterating the operation till a clustering with sufficiently small trash is obtained. At the end, the set of clusters obtained is returned as output, thus implicitly discarding the trashed trajectories.

In the second step, each cluster of trajectories is perturbed by means of the minimum spatial translation needed to push all the trajectories within a common uncertainty cylinder, i.e., transforming them in an anonymity set.

Data quality of the anonymized database D^* 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 D and D^*. In particular, as of objective measures Abul *et al.* adopt the total information distortion introduced by the spatial translation of points, and *discernibility*. Introduced in [38], discernibility is a simple measure of 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: $DM(D^*) = \sum_{i=1}^{n-1} |p_i|^2 + |p_n||D^*|$. Intuitively, discernibility represents the fact that data quality shrinks as more data elements become indistinguishable. The experiments reported in [16] show that discernibility is strongly influenced by the number of removed trajectories, and it does not provide any information about the amount of distortion introduced, thus resulting not much suitable for the cases of trajectory anonymization.

Abul *et al.* also report experiments on range query distortion, adopting the model of *spatio-temporal range queries with uncertainty* of [48]. In that work it is defined 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 the condition of interest is $inside(R, \tau)$. 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 (here the uncertainty is expressed by the same δ of the anonymization problem). If there exists some possible motion curve which at the time t is inside the region R, there is a possibility that the moving object will be inside R at t. Similarly, if every possible motion curve of the moving object 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, there are two domains of quantification, with two quantifiers in each. In [16], only the two extreme cases are used in the experimentation: namely $Possibly\_Sometime\_Inside$, corresponding to a double $\exists$, and $Definitely\_Always\_Inside$, corresponding to a double $\forall$. The query used is the count of the number of objects in D, and for comparison in D^*, satisfying $Possibly\_Sometime\_Inside$ (or $Definitely\_Always\_Inside$) for some randomly chosen region R and time interval $[t_b, t_e]$ (averaging on a large number of runs). Experimental results show that for a wide range of values of δ and k, the relative error introduced by the method of Abul *et al.* is kept reasonably low. Also the running time is shown to be reasonable even on large MODs.

Inspired by the *condensation* approach [30,31], Nergiz *et al.* [17] tackles the trajectory anonymization problem by means of grouping and reconstruction. In their framework no uncertainty is part of the input of the problem: an anonymity set is defined as a set of size $\geq k$ of *identical* trajectories (this correspond to the case of $\delta = 0$ in the setting of [16]). To make this feasible it is necessary to generalize points in space and time. Since they consider generalization, trajectory is defined as a sequence of 3D spatio-temporal volumes. In other terms each observation, each point in a trajectory is represented by intervals on the three dimensions: $[x_1, x_2]$, $[y_1, y_2]$, and $[t_1, t_2]$.

Therefore, a k-anonymization of a MOD D is definesd as a another MOD D^* such that:

- for every trajectory in D^*, there are at least $k-1$ other trajectories with exactly the same set of points;
- there is a one to one relation between the trajectories $tr \in D$ and trajectories $tr^* \in D^*$ such that for each point $p_i \in tr^*$ there is a unique $p_j \in tr$ such that $t_i^1 \leq t_j^1$, $t_i^2 \geq t_j^2$, $x_i^1 \leq x_j^1$, $x_i^2 \geq x_j^2$, $y_i^1 \leq y_j^1$ and $y_i^2 \geq y_j^2$.

Given a set of trajectories that are going to be anonymized together, the anonymity set is created by *matching* points, and then by taking the 3D minimum bounding box that contains the matched points. A depiction of the proposed anonymization process is provided in Figure 2.

As clustering strategy Nergiz *et al.* adapt the condensation based grouping algorithm given in [30]. The cost of the *optimal* anonymization is adopted as

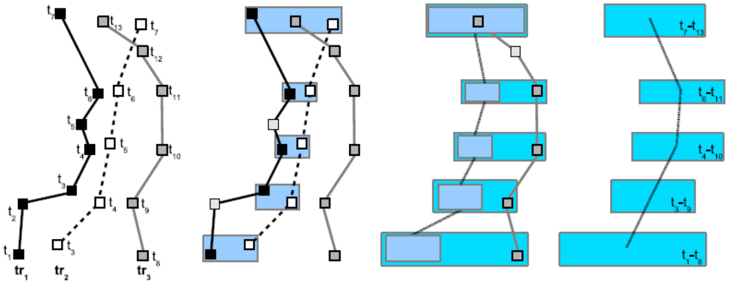


Fig. 2. Anonymization of three trajectories tr_1, tr_2 and tr_3, based on point matching and removal, and spatio-temporal generalization

distance metric between two trajectories. Finding the optimal anonymization of two trajectories is the same as finding the point matching between the two trajectories such that anonymizing the trajectories through the matching minimizes the generalization cost. A similar alignment problem is well studied for strings (where the goal is to find an alignment of strings such that total pairwise edit distance between the strings is minimized) in the context of DNA comparisons. Alignment problem for two trajectories is polynomial and can be solved by using a dynamic programming approach.

The resulting greedy algorithm, named *multi TGA*, at each iteration creates an empty group G, randomly samples one trajectory $tr \in D$, puts tr into G, and initialize the group representative $rep_G = tr$. Next, the closest trajectory $tr' \in TR \setminus G$ to rep_G is selected and added to G, and then rep_G is updated as the bounding box anonymizing rep_G and tr'.

The main drawback of this algorithm is the costly operation of finding the closest trajectory to the group representative. In order to decrease the number of times that such costly operation must be performed, a new algorithm (named *fast TGA*) is introduced: in *fast TGA* all the $k-1$ closest trajectories to the group representative are chosen in one pass. However, another drawback arises, as the challenge now becomes the computation of the optimal anonymization. In fact, while optimal matching between two trajectories is easy, computing the optimal point matching for $n > 2$ trajectories in NP-hard. For tackling this problem Nergiz *et al.* rely on heuristics that have proven to be effective in the string alignment problem.

After providing their generalization-based approach to k-anonymity of trajectories, Nergiz *et al.* discuss some drawbacks of such approach, and suggest that in many cases it might be more suitable to publish a reconstructed MOD, instead of a generalized one. In particular, they claim that generalization suffers from two main shortcomings. Firstly, the use of minimum bounding boxes in anonymization discloses uncontrolled information about exact locations of the points: e.g., in the case of two trajectories, two non-adjacent corners give out the exact locations. Secondly, it is challenging to take full advantage of information

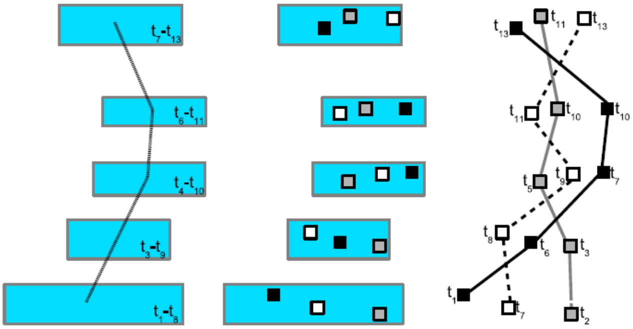


Fig. 3. Example of reconstruction starting from the anonymization of Figure 2

contained in generalized MODs as most data mining and statistical applications work on atomic trajectories.

Therefore Nergiz *et al.* adapt the reconstruction approach [30] and publish reconstructed data rather than data anonymized by means of generalization. An example reconstruction is shown in Figure 3. The output after reconstruction is atomic and suitable for trajectory data mining applications.

For assessing the quality of the resulting anonymization Nergiz *et al.* focus on the utility of the data for mining purposes. In particular, they chose a standard clustering method and compare the results obtained by clustering the original MOD D and its anonymized version D^*. In order to asses the result of clustering, they consider every pair of trajectories and verify whether both are in the same cluster, in the clustering given by D, and whether they are in the same cluster, in the clustering given by D^*. Then they measure accuracy, precision and recall.

5 Methods Based on Quasi-identifier

In this section we review two recent approaches to anonymization of MODs that adopt some concept of quasi-identifier.

The basic assumption of work by Terrovitis and Mamoulis [9] is that the adversaries own portions of the moving objects, and different adversaries owns different parts. The portion of a trajectory known by an adversary may be used to perform a linkage attack if the MOD is published without paying attention to anonymity. The privacy that is required is that, from the data publication, an adversary can not learn anything more than what he already knows.

As motivating example, they analyze the case of a company in Hong Kong called Octopus that collects daily trajectory data of Hong Kong residents who use Octopus smart RFID card. As we discussed in the motivating example in Section 1, when Alice uses her Octopus card to pay at different convenience stores that belong to the same chain (e.g., 7-Eleven), she left a sequence of traces, in some sense, giving away to the company a portion of her own trajectory. If this

Table 2. (a) an example MOD D, and (B) a local MOD D^A (A's knowledge)

t_{id}	trajectory
t_1	$a_1 \rightarrow b_1 \rightarrow a_2$
t_2	$a_1 \rightarrow b_1 \rightarrow a_2 \rightarrow b_3$
t_3	$a_1 \rightarrow b_2 \rightarrow a_2$
t_4	$a_1 \rightarrow a_2 \rightarrow b_2$
t_5	$a_1 \rightarrow a_3 \rightarrow b_1$
t_6	$a_3 \rightarrow b_1$
t_7	$a_3 \rightarrow b_2$
t_8	$a_3 \rightarrow b_2 \rightarrow b_3$

(a)

t_{id}	trajectory
t_1^A	$a_1 \rightarrow a_2$
t_2^A	$a_1 \rightarrow a_2$
t_3^A	$a_1 \rightarrow a_2$
t_4^A	$a_1 \rightarrow a_2$
t_5^A	$a_1 \rightarrow a_3$
t_6^A	a_3
t_7^A	a_3
t_8^A	a_3

(b)

projection of her trajectory uniquely identifies Alice, then by matching it with the published trajectory database, even though the IDs of users may be removed, Alice still can be re-identified, as can the other locations outside the portion of Alice's trajectory that 7-Eleven already knows.

More formally, Terrovitis and Mamoulis consider trajectories being simple sequences of addresses, corresponding to the places in which the Octopus card is used. Let $\mathcal{P}$ be the domain of all addresses where the Octopus card is a accepted. Since commercial companies might have multiple branches, $\mathcal{P}$ can be partitioned in m disjoint non-empty sets of addresses $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m$ such that each set contains all and only the addresses of the different branches of a company. Or in other terms, each adversary i controls a portion of addresses $\mathcal{P}_i$. For each trajectory t in the input MOD D, each adversary i holds a portion (or a projection) t^i. In general, each adversary i holds a local database D^i containing the projections of all $t \in D$ with respect to $\mathcal{P}_i$. The adversary has no knowledge about trajectories having empty projection; therefore, $\mathcal{P}_i$ can be smaller than the database of the publisher. A trajectory may appear multiple times in D and more than one trajectories may have the same projection with respect to $\mathcal{P}_i$. The most important property of a t^i is that adversary i can directly link it to the identities of all persons that pass through it, in its local database (e.g., loyalty program). Consider the example MOD D given in Table 2(a). Each sequence element is a shop address, where the corresponding user did his/her card transactions. Locations are classified according to the possible adversaries. For example, all places denoted by a_i are assumed to also be tracked by company A (e.g., 7-Eleven). Table 2(b) shows the knowledge of A. This knowledge D^A can be combined with D, if D is published, to infer private information. For instance, if D is published, A will know that t_5^A actually corresponds to t_5, since t_5 is the only trajectory that goes through a_1 and a_3, and no other location of company A. Therefore A is 100% sure that the user whose trajectory is t_A^5, visited b_1.

Therefore, the problem tackled by Terrovitis and Mamoulis in [9] can be formulated as follows. Given a MOD D, where each trajectory $t \in D$ is a sequence of values from domain $\mathcal{P}$, construct a transformed database D^*, such that if D^* is public, for all $t \in D$, every adversary i cannot correctly infer any location $\{p_j | p_j \in t \wedge p_j \notin t^i\}$ with probability larger than a given threshold P_{br}.

This problem is similar to the l-diversity problem defined in [13]. The main differences with the problem of privacy preserving publication in the classic relational context, are that in this context quasi-identifiers are variable-length sequences of locations, and that there can be multiple sensitive values (i.e., locations) per trajectory and these values are different from the perspectives of different adversaries. The second difference is that the algorithm which transforms D in D^* must consider linkage attacks to different sensitive values from different adversaries at the same time. One important point is that anonymization is based on the assumption that the data owner is aware of the adversarial knowledge, i.e., which adversary holds which portion of data, or in other terms, the data owner is exactly aware of the partition $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m$ of $\mathcal{P}$.

As objective function of utility Terrovitis and Mamoulis adopt the average difference between the original trajectories in D and the published ones in D^*. The method used to sanitize D from the possible linkage attacks is based on the identification, and the consequent suppression, of certain points that causes potential threats. This is done taking under consideration the benefit in terms of privacy and the deviation from the main direction of the trajectory. Since the problem of finding the optimal set of points to delete from D in order to derive a secure D^* and achieve the minimum possible information loss is harder NP-hard, they proposes a greedy algorithm that iteratively suppresses locations, until the privacy constraint is met. The algorithm simulates the attack from any possible adversary, and then solves the identified privacy breaches. The algorithm is empirically evaluated by measuring the effective cost and the number of points suppressed.

While this problem statement fits perfectly the scenario described above, it is not easily adaptable for other cases, e.g., the scenario of mobility data, where a set of cars equipped with GPS move on a road-network, or users with mobile phones move in a city. In this cases it seems less reasonable to assume that the data owners knows which are the spatio-temporal points known by an adversary.

Yarovoy *et al.* [18] in another recent work addresses the problem of privacy-preserving publication of MODs, considering quasi-identifiers, and using spatial generalization as anonymization technique. The authors argue that unlike in relational microdata, where every tuple has the same set of quasi-identifier attributes, in mobility data we can not assume a set of particular locations, or a set of particular timestamps, to be a quasi-identifier for all the individuals. It is very likely that various moving objects have different quasi-identifiers and this should be taken into account in modeling adversarial knowledge.

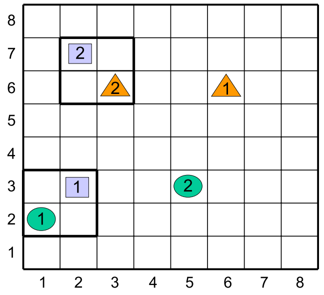
More precisely, given a MOD $D = \{O_1, ..., O_n\}$ that correspond to n individuals, a set of m discrete time points $T = \{t_1, ..., t_m\}$, and a function $\mathcal{T} : D \times T \rightarrow \mathbb{R}^2$, that specifies, for each object O and a time t, its position at time t, they consider timestamps as attributes with objects' positions forming their values, and they assume quasi identifiers to be sets of timestamps. As said above, a fixed set of timestamps can not be the quasi-identifier for all the moving objects. To capture this, they define the quasi-identifier as a function:

MOB	t_1	t_2
O_1	$(1,2)$	$(5,3)$
O_2	$(2,3)$	$(2,7)$
O_3	$(6,6)$	$(3,6)$

(a)

MOB	t_1	t_2
O_1	$[(1,2),(2,3)]$	$(5,3)$
O_2	$[(1,2),(2,3)]$	$[(2,6),(3,7)]$
O_3	$(6,6)$	$[(2,6),(3,7)]$

(b)



(c)

Fig. 4. Assuming $QID(O_1) = \{t_1\}$, $QID(O_2) = QID(O_3) = \{t_2\}$: (a) original database; (b) a 2-anonymity scheme that is not safe, and (c) its graphical representation

$QID : \{O_1, ..., O_n\} \rightarrow 2^{\{t_1,...,t_n\}}$. That is, every moving object may potentially have a distinct quasi-identifier.

The main issue in anonymizing MOD is that, due to the fact that different objects may have different QID, anonymization groups associated with different objects may not be disjoint, as illustrated below.

Example 2. Consider the trajectories in Figure 4(a) and illustrated in Figure 4(c). Let $k = 2$ and $QID(O_1) = \{t_1\}$, $QID(O_2) = QID(O_3) = \{t_2\}$. Intuitively the best (w.r.t. information loss) anonymization group for O_1 w.r.t. its QID $\{t_1\}$ is $AS(O_1) = \{O_1, O_2\}$. This is illustrated in Figure 4(c) with a dark rectangle. This means in the anonymized database we assign the region $[(1,2),(2,3)]$ to O_1 and O_2 at time t_1. The best anonymization group for O_2 as well as for O_3 w.r.t. their QID $\{t_2\}$ is $\{O_2, O_3\}$. Thus, in the anonymized database, O_2 and O_3 will both be assigned to the common region $[(2,6),(3,7)]$ (the second dark rectangle) at time t_2. Clearly, the anonymization groups of O_1 and O_2 overlap.

Due to this fact providing a robust and sound definition of k-anonymity in the case of MOD is challenging, as it will be clarified below. In order to explain why, we first need to introduce some basic definitions.

Given a MOD D, a distorted version of D is any database D^* over the same time points $\{t_1, ..., t_n\}$, where D^* contains one row for every moving object O in D, and either $D^*(O,t) = D(O,t)$ or $D(O,t) \sqsubseteq D^*(O,t)$, where with $\sqsubseteq$ we denote spatial containment among regions. The goal, as usual, is to find a distorted version of the MOD D, denoted by D^*, such that on the one hand, when published, D^* is still useful for analysis purposes, and on the other, a suitable version of k-anonymity is satisfied. The anonymization technique considered is *space generalization*. In the input MOD D, each position is an exact point, but with the application of a grid, each point may be regarded as a cell (as in Figure 4(c)), and generalized points are rectangles made of these cells.

Generalization obviously results in information loss. Yarovoy *et al.* measure information loss as the reduction in the probability with which the position of an object at a given time can be accurately determined. More formally, given a distorted version D^* of a MOD D, the information loss is defined as: $\mathtt{IL}(D, D^*) = \Sigma_{i=1}^{n} \Sigma_{j=1}^{m} (1 - 1/area(D^*(O_i, t_j))$; where $area(D^*(O_i, t_j))$ denotes the area of the region $D^*(O_i, t_j)$. As an example, consider the generalized MOD D^* as in Figure 4(b). The information loss associated with D^* is $2 \times (1 - 1/4) + 2 \times (1 - 1/4) = 3$.

A basic building block for devising any notion of anonymity is a notion of indistinguishability. Let D^* be a distorted version of a MOD D, two moving objects O, O' are indistinguishable in D^* at time t provided that $D^*(O,t) = D^*(O',t)$, i.e., both are assigned to the same region in D^*. The most obvious way of defining k-anonymity is the following: a distorted version D^* of a MOD D satisfies k-anonymity provided that for every moving object O in D, $\exists k-1$ other distinct moving objects $O_1, ..., O_{k-1}$ in D^*: $\forall t \in QID(O)$, O is indistinguishable from each of $O_1, ..., O_{k-1}$ at time t.

According to this definition the database in Figure 4(b) is 2-anonymous and thus "safe". This obvious definition of k-anonymity still suffers privacy breaches. Indeed, due the fact that anonymization groups may not be disjoint, it is possible that by combining overlapping anonymization groups, some moving objects may be uniquely identified, as explained next. Recall the previous example. There, I_1 and I_2 are in the same anonymization group (i.e., have the same generalized location) at time point t_1 (i.e., the QID of I_1), while I_2 and I_3 are in the same anonymization group at time point t_2 (i.e., the QID of I_2 and I_3). However, when the adversary tries to map the three moving objects O_1, O_2, O_3 to the three individuals I_1, I_2, I_3, with the adversary knowledge of QID values of these three moving objects, he can infer that I_1 must be mapped to either O_1 or O_2, while I_2 (and I_3) should be mapped to either O_2 or O_3. If I_1 is mapped to O_2, we cannot find a consistent assignment for I_2, I_3. As a result, the adversary can conclude that O_1 must map to I_1. Thus, a more sophisticated definition of k-anonymity is needed in order to avoid privacy breaches in the case of moving object databases.

Given the considerations above, Yarovoy *et al.* [18] define k-anonymity by formalizing the attack described above. In particular they define an attack graph associated with a MOD D and its distorted version D^*, as the bipartite graph G consisting of nodes for every individual I in D (called *I-nodes*) and nodes for every moving object id O (called *O-nodes*) in the published database D^*. G contains an edge (I, O) iff $D(O,t) \sqsubseteq D^*(O,t), \forall t \in QID(I)$.

An assignment of individuals to moving objects is consistent provided there exists a perfect matching in the bipartite graph G. Consider the distorted database shown in Figure 4(b): the corresponding attack graph is shown in Figure 5(a). It is obvious that the edge (I_1, O_1) must be a part of every perfect matching. Thus, by constructing the attack graph an attacker may easily conclude that MOB O_1 can be re-identified as I_1.

One of the key shortcomings in the straightforward definition of k-anonymity given above is that while it ensures every I-node corresponding to an individual has at least k neighbors, it does not have any restriction on the degree of the O-nodes. *What if we required that in addition, every O-node must have degree at*

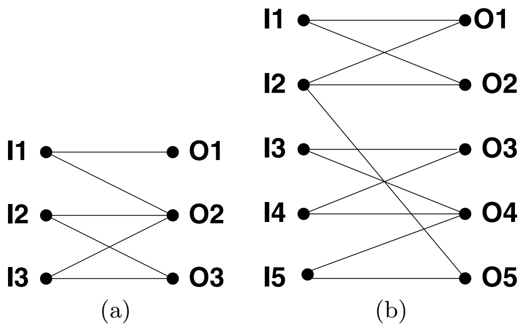


Fig. 5. Attack graphs for different anonymization schemes: (a) for D^* in Figure 4(b); (b) for a hypothetical database D^* satisfying modified definition of k-anonymity

least k? Suppose we say that a distorted database is k-anonymous provided in the corresponding attack graph, every I-node as well as every O-node has degree $\geq k$. Figure 5(b) shows a possible attack graph that satisfies this condition. In this graph, every I-node and every O-node has degree 2 or more. Yet, O_5 can be successfully re-identified as I_5 as follows. Suppose O_5 is instead assigned to I_2, to which it is adjacent as well. Then it is easy to see that no I-node can be assigned to one of O_1, O_2. Thus, *the edge* (I_2, O_5) *cannot be a part of any perfect matching*. Thus, this edge can be pruned, leaving I_5 as the only I-node to which O_5 can be assigned.

This example is subtler than the previous example and clearly shows the challenges involved in devising a notion of k-anonymity that does not admit privacy breaches.

The attack model is formalized as following. The attacker first constructs an attack graph associated with the published distorted version of D and the known QIDs as described above. Then, he repeats the following operation until there is no change to the graph:

1. Identify an edge e that cannot be part of any perfect matching.
2. Prune the edge e.

Next, he identifies every node O with degree 1. He concludes the (only) edge incident on every such node must be part of every perfect matching. There is a privacy breach if the attacker succeeds in identifying at least one edge that must be part of every perfect matching.

Finally k-anonymity is defined. Let D be a MOD and D^* its distorted version. Let G be the attack graph w.r.t. D, D^*. Then D^* is k-anonymous provided that (i) every I-node in G has degree k or more; and (ii) G is symmetric, i.e., whenever G contains an edge (I_i, O_j), it also contains the edge (I_j, O_i). An immediate observation is that in an attack graph that satisfies the above conditions, every O-node will have degree k or more as well.

Yarovoy *et al.* [18] develop two different algorithms and show that both of them satisfy the above definition of k-anonymity. One main challenge in devising these algorithms arises again from the fact that anonymization groups may not be disjoint: in particular, is overlapping anonymization groups can force the algorithm to revisit earlier generalizations, and possibly re-generalize them with other objects. For computing the anonymity group of a given moving object, both algorithms use a method based on Hilbert index of spatial objects for efficient indexing of trajectories. In the empirical comparison of the two algorithms, Yarovoy *et al.* report statistics on the size of the equivalence classes created in the anonymization, as well as the average information loss introduced. They also report range query distortion similarly to [16].

6 Conclusions and Open Research Issues

We provided an overview of a rather young research effort concerning how to anonymize a moving objects database. While only few papers have been published so far on this problem, much large body of work has been developed for location privacy in the on-line, dynamic context of location based services. We briefly reviewed this body of research trying to clarify why, even if apparently similar, the problem of anonymization becomes deeply different when tackled in a static instead of a dynamic context. However, many techniques developed in the LBS research community, such as location perturbation, spatio-temporal cloaking (i.e., generalization), and the personalized privacy-profile of users, should be taken in consideration while devising solutions for anonymity preserving data publishing of MODs. Also the definitions of *historical k-anonymity* and *location based quasi-identifier* introduced in [8] might be borrowed in the context of data publishing.

We discussed the challenge of deriving quasi-identifiers in the context of mobility data: as argued by some authors, they might be defined by the users themselves, or they might be “learnt” by mining a MOD. Finding a realistic and actionable definition of quasi-identifiers, as well as devising methodology to derive them, are important open problems.

Yarovoy *et al.* [18] argue that, contrarily to the classic relational setting, in MODs quasi-identifiers can only be defined on the individual basis, i.e., each moving object must have his own quasi-identifier. They also show how many computational challenges arise from this assumption. The main one is that the anonymization problem is no longer about finding a partition of objects in disjoint anonymity sets, because due to the different quasi-identifiers, anonymity sets may overlap.

An interesting approach is the one of Terrovitis and Mamoulis [9], that instead of defining quasi-identifiers by the user perspective, consider the linkage attacks that are possible given that different adversaries have knowledge of different parts of the data.

Other authors [16,17] instead do not consider quasi-identifiers and focus on anonymizing trajectories in their whole, by grouping together similar trajectories, and slightly perturbing them, to make them undistinguishable. Even if these approaches avoid the challenges connected to quasi-identifiers, they still face some open problems. One of these is the so-called diversity problem [13] introduced for relational data. In that context it is shown 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-identifier attributes), if the group has not enough diversity of the sensitive attributes then an attacker can still associate one individual to sensitive information. Also in the MOD context, if we are able to know that one individual belong to a group, even if we are not able to identify exactly his trajectory, we can still discover some sensitive information.

Another line of research, not yet started, is about developing ad-hoc anonymization techniques for the intended use of the data: for instance, with respect to a specific spatio-temporal data mining analysis.

A close and interesting research area is the so called *privacy-preserving data mining*, i.e., instead of anonymizing the data for a privacy-aware data publication, the focus of privacy is shifted directly to the analysis methods. Privacy preserving data mining, is an hot and lively research area which has seen the proliferation of many completely different approaches having different objectives, application contexts and using different techniques [49,50,51,52,53]. However, very little work has been done about developing privacy-preserving mining techniques for spatio-temporal and mobility data [54,55,56,57]: as said for the anonymization of MOD, this research topic is rather young and we expect to see many new proposals in the next future.

References

1. Giannotti, F., Pedreschi, D. (eds.): Mobility, Data Mining and Privacy - Geographic Knowledge Discovery. Springer, Heidelberg (2008)
2. Lee, J.G., Han, J., Li, X.: Trajectory outlier detection: A partition-and-detect framework. In: Proc. of the 24th IEEE Int. Conf. on Data Engineering (ICDE 2008) (2008)
3. Lee, J.G., Han, J., Li, X., Gonzalez, H.: raClass: trajectory classification using hierarchical region-based and trajectory-based clustering. In: Proc. of the 34th Int. Conf. on Very Large Databases (VLDB 2008) (2008)
4. Lee, J.G., Han, J., Whang, K.Y.: Trajectory clustering: a partition-and-group framework. In: Proc. of the 2007 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2007) (2007)
5. Li, X., Han, J., Kim, S., Gonzalez, H.: Anomaly detection in moving object. In: Intelligence and Security Informatics, Techniques and Applications. Studies in Computational Intelligence, vol. 135. Springer, Heidelberg (2008)
6. Li, X., Han, J., Lee, J.G., Gonzalez, H.: Traffic density-based discovery of hot routes in road networks. In: Papadias, D., Zhang, D., Kollios, G. (eds.) SSTD 2007. LNCS, vol. 4605, pp. 441–459. Springer, Heidelberg (2007)
7. Nanni, M., Pedreschi, D.: Time-focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems 27(3), 267–289 (2006)
8. Bettini, C., Wang, X.S., Jajodia, S.: Protecting Privacy Against Location-Based Personal Identification. In: Jonker, W., Petković, M. (eds.) SDM 2005. LNCS, vol. 3674, pp. 185–199. Springer, Heidelberg (2005)
9. Terrovitis, M., Mamoulis, N.: Privacy preservation in the publication of trajectories. In: Proc. of the 9th Int. Conf. on Mobile Data Management (MDM 2008) (2008)

10. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: Proc. of the 17th ACM Symp. on Principles of Database Systems (PODS 1998) (1998)
11. Samarati, P., Sweeney, L.: Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion. In: Proc. of the IEEE Symp. on Research in Security and Privacy, pp. 384–393 (1998)
12. Sweeney, L.: k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty Fuzziness and Knowledge-based Systems 10(5) (2002)
13. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. In: Proc. of the 22nd IEEE Int. Conf. on Data Engineering (ICDE 2006) (2006)
14. Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., Zhu, A.: Anonymizing tables. In: Eiter, T., Libkin, L. (eds.) ICDT 2005, vol. 3363, pp. 246–258. Springer, Heidelberg (2005)
15. Meyerson, A., Willliams, R.: On the complexity of optimal k-anonymity. In: Proc. of the 23rd ACM Symp. on Principles of Database Systems (PODS 2004) (2004)
16. Abul, O., Bonchi, F., Nanni, M.: $\mathcal{N}$ever $\mathcal{W}$alk $\mathcal{A}$lone: Uncertainty for anonymity in moving objects databases. In: Proc. of the 24nd IEEE Int. Conf. on Data Engineering (ICDE 2008) (2008)
17. Nergiz, E., Atzori, M., Saygin, Y.: Towards trajectory anonymization: a generalization-based approach. In: Proc. of ACM GIS Workshop on Security and Privacy in GIS and LBS (2008)
18. Yarovoy, R., Bonchi, F., Lakshmanan, L.V.S., Wang, W.H.: Anonymizing moving objects: How to hide a MOB in a crowd? In: Proc. of the 12th Int. Conf. on Extending Database Technology (EDBT 2009) (2009)
19. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. 13(6), 1010–1027 (2001)
20. Wong, R.C.W., Li, J., Fu, A.W.C., Wang, K. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2006) (2006)
21. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: Proc. of the 21st IEEE Int. Conf. on Data Engineering (ICDE 2005) (2005)
22. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity. In: Proc. of the 2005 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2005) (2005)
23. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous -anonymity through microaggregation. Data Min. Knowl. Discov. 11(2), 195–212 (2005)
24. Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregates method. In: Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa, Statistics Canada, pp. 195–204 (1993)
25. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. 14(1), 189–201 (2002)
26. Torra, V.: Microaggregation for categorical variables: A median based approach. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 162–174. Springer, Heidelberg (2004)
27. Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., Zhu, A.: Achieving anonymity via clustering. In: Proc. of the 25th ACM Symp. on Principles of Database Systems (PODS 2006) (2006)

28. Li, J., Wong, R.C.W., Fu, A.W.C., Pei, J.: Achieving *k*-anonymity by clustering in attribute hierarchical structures. In: Tjoa, A.M., Trujillo, J. (eds.) DaWaK 2006. LNCS, vol. 4081, pp. 405–416. Springer, Heidelberg (2006)
29. Byun, J.W., Kamra, A., Bertino, E., Li, N.: Efficient k-anonymization using clustering techniques. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 188–200. Springer, Heidelberg (2007)
30. Aggarwal, C.C., Yu, P.S.: A condensation approach to privacy preserving data mining. In: Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K., Ferrari, E. (eds.) EDBT 2004. LNCS, vol. 2992, pp. 183–199. Springer, Heidelberg (2004)
31. Aggarwal, C.C., Yu, P.S.: On anonymization of string data. In: Proc. of the 2007 SIAM Int. Conf. on Data Mining (2007)
32. Aggarwal, C.C.: On k-anonymity and the curse of dimensionality. In: Proc. of the 31st Int. Conf. on Very Large Databases (VLDB 2005) (2005)
33. Atzori, M.: Weak *k*-anonymity: A low-distortion model for protecting privacy. In: Katsikas, S.K., López, J., Backes, M., Gritzalis, S., Preneel, B. (eds.) ISC 2006. LNCS, vol. 4176, pp. 60–71. Springer, Heidelberg (2006)
34. Xiao, X., Tao, Y.: Anatomy: Simple and effective privacy preservation. In: Proc. of the 32nd Int. Conf. on Very Large Databases (VLDB 2006) (2006)
35. Kifer, D., Gehrke, J.: Injecting utility into anonymized datasets. In: Proc. of the 2006 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2006) (2006)
36. Øhrn, A., Ohno-Machado, L.: Using boolean reasoning to anonymize databases. Artificial Intelligence in Medicine 15(3), 235–254 (1999)
37. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Proc. of the 23rd IEEE Int. Conf. on Data Engineering (ICDE 2007) (2007)
38. Bayardo, R., Agrawal, R.: Data privacy through optimal k-anonymity. In: Proc. of the 21st IEEE Int. Conf. on Data Engineering (ICDE 2005) (2005)
39. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: Proc. of the 22nd IEEE Int. Conf. on Data Engineering (ICDE 2006) (2006)
40. Gruteser, M., Grunwald, D.: 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) (2003)
41. Gedik, B., Liu, L.: Location Privacy in Mobile Systems: A Personalized Anonymization Model. In: Proc. of the 25th Int. Conf. on Distributed Computing Systems (ICDCS 2005) (2005)
42. Domingo-Ferrer, J.: Microaggregation for database and location privacy. In: Etzion, O., Kuflik, T., Motro, A. (eds.) NGITS 2006. LNCS, vol. 4032, pp. 106–116. Springer, Heidelberg (2006)
43. Kido, H., Yanagisawa, Y., Satoh, T.: Protection of Location Privacy using Dummies for Location-based Services. In: Proc. of the 21st IEEE Int. Conf. on Data Engineering (ICDE 2005) (2005)
44. Beresford, A.R., Stajano, F.: Mix Zones: User Privacy in Location-aware Services. In: Proc. of the Second IEEE Conf. on Pervasive Computing and Communications Workshops (PERCOM 2004) (2004)
45. Gruteser, M., Liu, X.: Protecting Privacy in Continuous Location-Tracking Applications. IEEE Security & Privacy Magazine 2(2), 28–34 (2004)

46. Bettini, C., Wang, X.S., Jajodia, S.: Testing complex temporal relationships involving multiple granularities and its application to data mining. In: Proc. of the 15th ACM Symp. on Principles of Database Systems (PODS 1996) (1996)
47. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory pattern mining. In: Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2007) (2007)
48. Trajcevski, G., Wolfson, O., Hinrichs, K., Chamberlain, S.: Managing uncertainty in moving objects databases. ACM Trans. Database Syst. 29(3), 463–507 (2004)
49. Clifton, C., Marks, D.: Security and privacy implications of data mining. In: Proc. of the 1996 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 1996), February 1996, pp. 15–19 (1996)
50. O'Leary, D.E.: Knowledge discovery as a threat to database security. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 507–516. AAAI/MIT Press (1991)
51. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2000), pp. 439–450 (2000)
52. Clifton, C., Kantarcioglu, M., Vaidya, J.: Defining privacy for data mining. In: Natural Science Foundation Workshop on Next Generation Data Mining, November 2002, pp. 126–133 (2002)
53. Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y.: State-of-the-art in privacy preserving data mining. ACM SIGMOD Record 33(1), 50–57 (2004)
54. Inan, A., Saygin, Y.: Privacy-preserving spatio-temporal clustering on horizontally partitioned data. In: Tjoa, A.M., Trujillo, J. (eds.) DAWAK 2006. LNCS, vol. 4081, pp. 459–468. Springer, Heidelberg (2006)
55. Abul, O., Atzori, M., Bonchi, F., Giannotti, F.: Hiding sequences. In: Proceedings of the Third ICDE International Workshop on Privacy Data Management (PDM 2007) (2007)
56. Abul, O., Atzori, M., Bonchi, F., Giannotti, F.: Hiding sensitive trajectory patterns. In: ICDM 2007, pp. 693–698 (2007)
57. Bonchi, F., Saygin, Y., Verykios, V.S., Atzori, M., Gkoulalas-Divanis, A., Kaya, S.V., Savas, E.: Privacy in spatiotemporal data mining. In: [1], pp. 297–333