Movement Data Anonymity through Generalization

[Position Paper]

Gennady Andrienko
Fraunhofer IAIS Sankt Augustin, Germany

Natalia Andrienko
Fraunhofer IAIS Sankt Augustin, Germany

Fosca Giannotti
KddLab ISTI-CNR Pisa, Italy

Anna Monreale
Computer Science Dept. University of Pisa

Dino Pedreschi
Computer Science Dept. University of Pisa

ABSTRACT

In recent years, spatio-temporal and moving objects databases have gained considerable interest, due to the diffusion of mobile devices (e.g., mobile phones, RFID devices and GPS devices) and of new applications, where the discovery of consumable, concise, and applicable knowledge is the key step. Clearly, in these applications privacy is a concern, since models extracted from this kind of data can reveal the behavior of group of individuals, thus compromising their privacy. Movement data present a new challenge for the privacy-preserving data mining community because of their spatial and temporal characteristics.

In this position paper we briefly present an approach for the generalization of movement data that can be adopted for obtaining k-anonymity in spatio-temporal datasets; specifically, it can be used to realize a framework for publishing of spatio-temporal data while preserving privacy. We ran a preliminary set of experiments on a real-world trajectory dataset, demonstrating that this method of generalization of trajectories preserves the clustering analysis results.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial databases and GIS; K.4.1 [**Public Policy Issues**]: Privacy

Keywords

k-anonymity, Privacy, Spatio-temporal, Clustering

1. INTRODUCTION

Many Knowledge Discovery techniques have been developed that provide new means for improving personalized services through the discovery of patterns which represent typical or unexpected customer's and user's behavior. However, the collection and the disclosure of personal, often sensitive, information increase the risk of citizen's privacy violation. For this reason, many recent research works have focused on privacy-preserving data mining [2, 13, 6, 7]. In general, these approaches allow to extract knowledge while trying to protect the privacy of individuals represented in the dataset. Some of these techniques return anonymized data mining results, while others provide anonymized datasets to the companies/research institution in charge of their analysis. The pervasiveness of location-aware devices, e.g., PDAs and cell phones with GPS technology, RFID devices enables to collect a great amount of traces left by moving objects and to analyze their motion patterns. Clearly, in this context privacy is a concern: location data allows inferences which may help an attacker to discovery personal and sensitive information like habits and preferences of individuals. Hiding car identifiers for example replacing them with pseudonyms as shown in [13] is insufficient to guarantee anonymity, since location represents a property that could allow the identification of the individual. In particular, sensitive information about individuals can be uncovered with the use of visual analytics methods. Therefore, in all cases when privacy concerns are relevant, such methods must not be applied to original movement data. The data must be anonymized, that is, transformed in such a way that sensitive private information could no more be retrieved.

In this position paper we present a method for the generalization of movement data that can be adopted for obtaining a form of anonymity in spatio-temporal datasets. The main idea is to hide locations by means of generalization, specifically, replacing exact positions in the trajectories by approximate positions, i.e. points by areas. This method of generalization can be used in a privacy-preserving framework of spatio-temporal data in order to generate an anonymous dataset which satisfy the k-anonymity property. In the literature, most of anonymization approaches proposed in the spatio-temporal context are based on randomization techniques, space translations of points and suppression of some potions of a trajectory. To the best of our knowledge only the work in [15] uses spatial generalization to achieve anonymity for trajectory datasets; however, a fixed grid hierarchy is used in this work to discretize the spatial dimension. In contrast, the novelty of our approach lies in finding a suitable tessellation of the territory into areas depending of the input trajectory dataset. As a result of our approach, we obtain anonymous trajectories with high analytical utility, if compared with previous works (both randomization based and generalization based): in particular, we show how the results of clustering analysis are faithfully preserved. The con-

ACM SPRINGL '09 November 3, 2009. Seattle, WA, USA

cept of spatial generalization has been also used in the works on privacy in location-based services [5, 8, 9], where the goal is on-line anonymization of individual location-based queries, while our aim is privacy-preserving data publishing, which requires the anonymization of each entire trajectory. A detailed discussion appears in Section 2.

The rest of the paper is organized as follows. Section 2 discusses the relevant related works on privacy issue in spatio-temporal data. Section 3 introduces the problem definition. In Section 4 we describe the generalization approach for trajectories. Section 5 describes the experimental results on the clustering analysis. In Section 6 we discuss about the possible ideas to adapt the proposed generalization method to anonimyze movement data. Finally, Section 7 concludes.

2. RELATED WORK

Many research works have focused on techniques for privacy-preserving data mining [2] and for privacy-preserving data publishing. The first operation before data publishing is to replace personal identifier with pseudonyms. In [13] authors showed that this simple operation is insufficient to protect privacy. In this work, Samarati and Sweeney propose k-anonymity to make each record indistinguishable with at least $k-1$ other records. k-anonymity is the most popular method for the anonymization of spatio-temporal data. It is often used both in the works on privacy issues in location-based services (LBSs) and on anonymity of trajectories.

In LBSs context a trusted server usually has to handle the requests of users and to pass them on to the service providers. In general, it has to provide a on-line service without compromise the anonymity of the user. The different systems proposed in literature to make the requests indistinguishable from $k-1$ other requests use a space generalization, called *spatial-cloaking* [5, 8, 9]. In our context the anonymization process is off-line, as we want to anonymize a static database of trajectories. To the best of our knowledge only three works address the problem of k-anonymity of moving objects by a data publishing perspective [1, 10, 15]. In the work [1], the authors study the problem of privacy-preserving publishing of moving object database. They propose the notion of (k, δ)-anonymity for moving objects databases, where δ represents the possible location imprecision. In particular, this is a novel concept of k-anonymity based on co-localization that exploits the inherent uncertainty of the moving objects whereabouts. In this work authors also propose an approach, called *Never Walk Alone* based on trajectory clustering and spatial translation. In [10] Nergiz et al. address privacy issues regarding the identification of individuals in static trajectory datasets. They provide privacy protection by: (1) first enforcing k-anonymity, meaning every released information refers to at least k users/trajectories, (2) then reconstructing randomly a representation of the original dataset from the anonymization. Yarovoy et al. in [15] study problem of k-anonymization of moving object databases for the purpose of their publication. They observe the fact that different objects in this context may have different quasi-identifiers ans so, anonymization groups associated with different objects may not be disjoint. Therefore, a novel notion of k-anonymity based on spatial generalization is provided. In this work, authors propose two approaches in order to generate anonymity groups that satisfy the novel notion of k-anonymity. These approaches are called *Extreme Union* and *Symmetric Anonymization*.

Another approach based on the concept of k-anonymity is proposed in [11], where a framework for k-anonymization of sequences of regions/locations is presented. The authors also propose an approach that is an instance of the proposed framework and that allows to publish protected datasets while preserving the data utility for sequential pattern mining tasks. This approach, called *BF-P2kA*, uses a prefix tree to represent the dataset in a compact way. Given a threshold k generates a k-anonymous dataset while preserving the sequential pattern mining results.

Finally, in a very recent work [14], a suppression-based algorithm is suggested. Given the head of the trajectories, it reduces the probability of disclosing the tail of the trajectories. This work is based on the assumption that different attackers know different and disjoint portions of the trajectories and the data publisher knows the attacker knowledge. So, the solution is to suppress all the dangerous observations.

3. PROBLEM DEFINITION

A moving object dataset is a collection of trajectories $D = \{T_1, T_2, \ldots, T_m\}$ where each T_i is a trajectory represented by a sequence of spatio-temporal points:

$$T_i = (x_1, y_1, t_1), (x_2, y_2, t_2), \ldots, (x_n, y_n, t_n)$$

and $(t_1 < t_2 < \cdots < t_n)$.

Given a moving object dataset D our goal is to provide an anomymized version of D that guarantees the privacy of the individuals while preserving some interesting analysis results such as clustering analysis. For this aim we want to use a k-anonymity approach based on spatial generalization.

In this position paper we describe a method for generalization of movement data that can be adapted for obtaining anonymity in a moving object dataset. The idea is to hide personal information by means of generalization, specifically, replacing exact positions in the trajectories by approximate positions, i.e. points by areas.

4. TRAJECTORY GENERALIZATION

The approach to generalize movement data consists of two main steps: (1) generating a division of the territory into areas and (2) generalizing the original trajectories.

4.1 Division of the Territory into Areas

The generalization method generates an appropriate division of the territory into areas. The method is based on extracting characteristic points from the trajectories, which are the positions of start and end, significant turns (i.e. the change of the movement direction is above a given threshold), and significant stops (i.e. the time of staying in the same position is above a threshold). The extracted points are grouped into spatial clusters. The central points of the clusters are used as generating points for Voronoi tessellation of the territory, which produces suitable areas. Since the areas are built around clusters of characteristic points, the resulting abstraction conveys quite well the principal characteristics of the movement. The level of the abstraction can be controlled through the parameters of the clustering method.

We demonstrate the work of the method by example of a subset car trajectories got by the European project GeoPKDD. Thanks to this project we received a real-world and

large dataset of trajectories of cars equipped with GPS and moving in the city of Milan (Italy). For our preliminary experiments we considered 4287 trajectories. Figure 1(a) presents a map with the original trajectories. In Figure 1(b), there are the characteristic points extracted from the trajectories (54362 points in total). Both the trajectories and the characteristic points are shown with 10% opacity, to enable the estimation of the densities in different places. In Figure 1(c), the characteristic points have been clustered. The clusters are represented by colouring. We use a special spatial clustering algorithm with a parameter defining the maximum radius (spatial extent) of a cluster. The clusters in Figure 1(c) have been obtained for the value 500 metres of this parameter. Figure 1(d) presents the centroids of the point clusters and the Voronoi cells, which have been built using the centroids as generating points. Besides the cluster centroids, we add generating points around the boundaries of the territory and in the areas where there are no characteristic points from the trajectories. This is done for the cells to be more even in sizes and shapes.

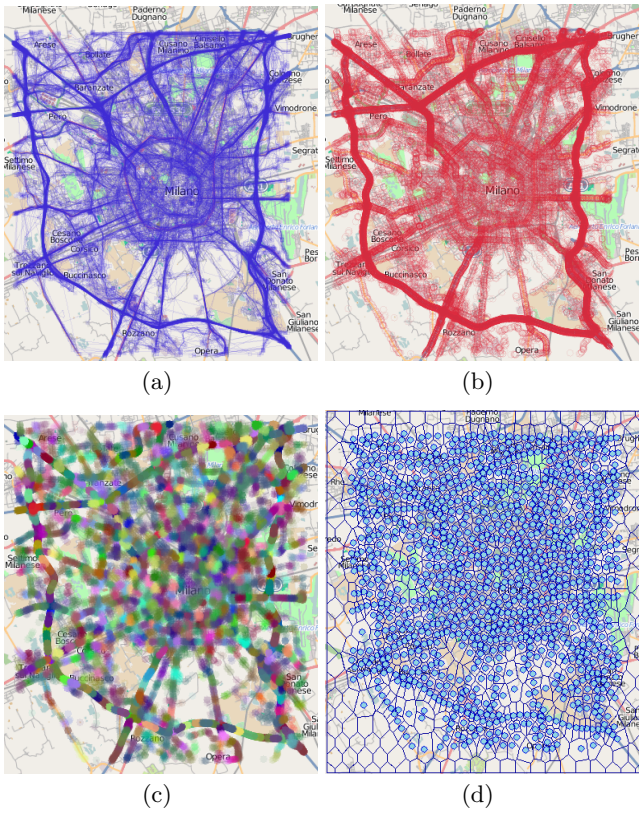

Figure 1: a) Original subset of 4287 trajectories (10% opacity). b) The characteristic points extracted from the trajectories (10% opacity). c) Clustered characteristic points. d) Centroids of the clusters and the Voronoi tessellation of the territory.

4.2 Generalization method

After obtaining the division of the territory, the trajectories are generalized in the following way. We apply place-based division of trajectories into segments. For each trajectory, the area a_1 containing its first point p_1 is found. Then, the second and following points of the trajectory are checked for being inside a_1 until finding a point p_i not contained in a_1. For this point p_i, the containing area a_2 is found.

The trajectory segment from the first point to the i-th point is represented by the vector (a_1, a_2). Then, the procedure is repeated: the points starting from p_{i+1} are checked for containment in a_2 until finding a point p_k outside a_2, the area a_3 containing p_k is found, and so forth up to the last point of the trajectory.

In the result, the trajectory is represented by the sequence of areas $\{a_1, a_2, \ldots, a_n\}$. There may be also a case when all points of a trajectory are contained in one and the same area a1. Then, the whole trajectory is represented by the sequence $\{a_1\}$. For each area a_i in the sequence, there is a corresponding time interval starting with the time moment of the first position in a_i and ending with the time moment of the last position in a_i.

As most of the methods for analysis of trajectories are suited to work with positions specified as points, the sequence of areas $\{a_1, a_2, \ldots, a_n\}$ is replaced, for practical purposes, by the sequence $c_1, c_2, \ldots, c_n$ consisting of the centroids of the areas $\{a_1, a_2, \ldots, a_n\}$. As a result, we obtain generalized trajectories. Figure 2 illustrates the generalized trajectories of the cars from Milan.

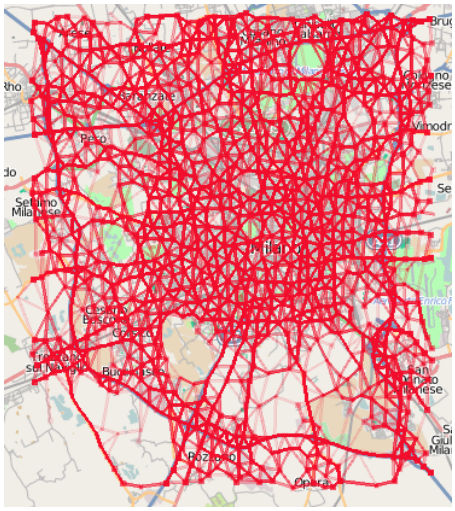

Figure 2: Generalized trajectories of cars from Milan (10% opacity); the line thickness is 2 pixels.

5. CLUSTERING ANALYSIS

An important property of this method for protecting personal data is that the resulting transformed data are suitable at least for some kinds of analysis. In particular, it is possible to analyze the flows between the areas and statistics of the visits of the areas. One may also analyze the statistics of the travel times between different pairs of areas, not only neighboring. Frequently occurring sequences of visited areas can be discovered by means of data mining techniques. It is also possible to apply cluster analysis to the modified trajectories. Thus, we have made several experiments with clustering of the original car trajectories from Milan and generalized versions of these trajectories using the generic density-based clustering algorithm *OPTICS* [4] with a suitable distance function.

We found that the results of clustering the original and the generalized trajectories are very similar when the distance threshold (the parameter of the clustering algorithm) for the generalized trajectories is about one half of the distance threshold for the original trajectories. Figure 3 shows the biggest clusters obtained from the original set of trajectories (the first group of 12 clusters in the figure) and the biggest clusters obtained from the set of generalized trajectories (the last group of 12 clusters). The clusters are represented in an aggregated form. The results have been obtained using the density-based clustering algorithm OPTICS with the distance function "*route similarity*" [3, 12] and the required minimum of 5 neighbors of a core object. The distance thresholds used is 500m for the first group and 250m for the second group. The labels A, B, etc. establish the correspondence between the clusters in two results. The clusters of the second group corresponding to the clusters G and L of the first group are not among the largest 12 clusters (they are on the 15th and 14th places, respectively). Analogously, the clusters of the first group corresponding to the 11th and 12th clusters of the second group (clusters with label N and Q in the figure) are on the 14th and 17th places, respectively.

The results of the experiments allow us to believe that the idea has a good potential.

6. GENERALIZATION VS K-ANONYMITY

The approach described in Section 4, given a dataset of trajectories allows us to generate a generalized version of it. In order to adapt this method to the anonymization of movement data, some extensions are required. In particular, it is necessary to ensure that:

1. each area contains positions from the trajectories of k different people, where k is a parameter.
2. the dispersion of the positions in each area is not less than a specified threshold (another parameter).
3. for each pair of areas a and b there are either none or at least k people who come from a to b (possibly, with visiting some other areas in between).

The satisfaction of these anonymity conditions is easy to check. Thus, the map in Figure 4 visualizes the numbers of different trajectories that visited the areas of the territory division (Voronoi cells) used for the generalization. The cells where the first two conditions are not satisfied must be enlarged to include more positions. This is done by producing a new Voronoi tessellation after excluding the generating points of the "problematic" cells. Similarly, when too few people come from a to b, either a or b is excluded. To choose between a and b, the total number of incoming and outgoing links with the magnitudes below k is counted for each of them. Excluded is the area where this number is bigger.

The generalization-based anonymization method is currently under development. We need to do further investigations for checking whether any risks to personal privacy are indeed precluded when trajectories are anonymized in this way.

7. CONCLUSION

In this position paper, we present an approach for generalization of movement data. We think that this method

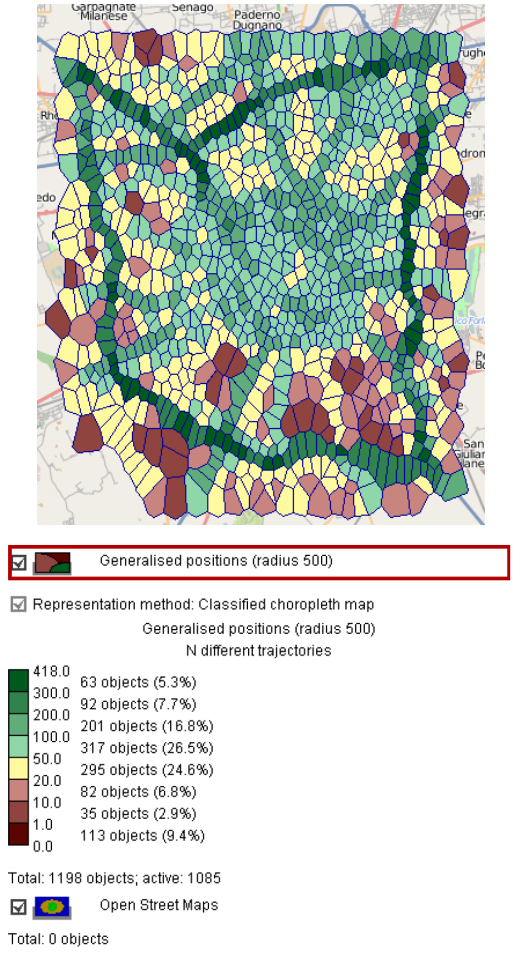

Figure 4: The map shows the numbers of different trajectories that visited the areas of the territory division by area coloring. The areas that do not contain any points from the trajectories are hidden.

can be adopted to realize a framework for anonymization of spatio-temporal data based on spatial generalization and the k-anonymity concept. Through a preliminary set of experiments on a real-life mobility dataset, we showed that the proposed technique preserves clustering results.

In future work, we intend to investigate further the protection model against the re-identification attack, that can be obtained using the method proposed in this paper.

8. REFERENCES

[1] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *ICDE*, pages 376–385, 2008.

[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD*, pages 439–450. ACM, 2000.

[3] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *VAST in press*, 2009.

[4] M. Ankerst, M. M. Breunig, H.P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *SIGMOD*, pages 49–60, 1999.

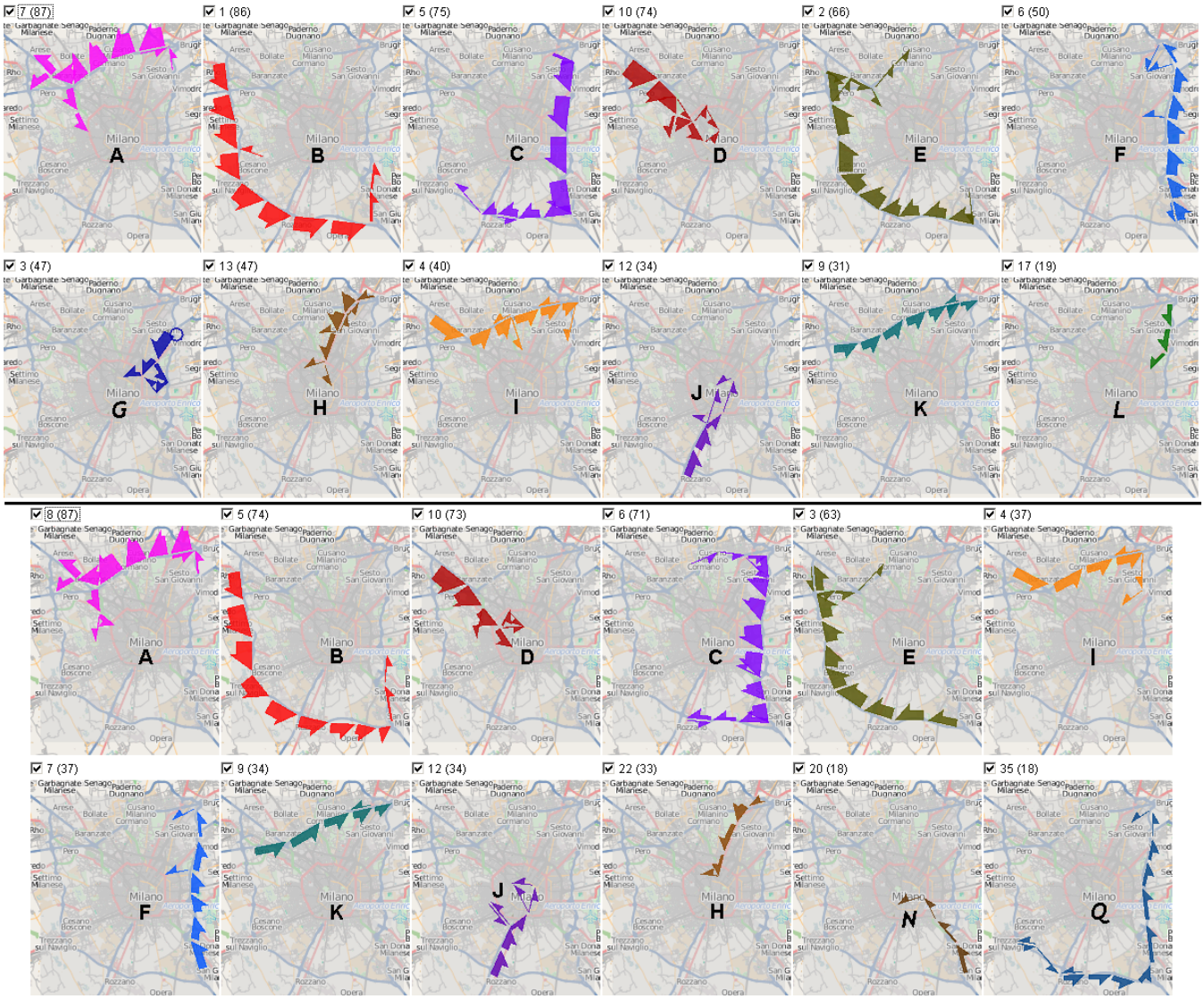

Figure 3: Comparison of clustering results for the original (top) and generalized (bottom) trajectories. 12 biggest clusters from each result are visible.

[5] M. Gruteser and D. Grunwald. A methodological assessment of location privacy risks in wireless hotspot networks. In *SPC*, pages 10–24, 2003.

[6] R. J. Bayardo Jr. and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, 2005.

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, page 25, 2006.

[8] M. F. Mokbel, C. Chow, and W. G. Aref. The new casper: Query processing for location services without compromising privacy. In *VLDB*, pages 763–774, 2006.

[9] M. F. Mokbel, C. Chow, and W. G. Aref. The new casper: A privacy-aware location-based database server. In *ICDE*, pages 1499–1500, 2007.

[10] M. E. Nergiz, M. Atzori, and Y. Saygin. Perturbation-driven anonymization of trajectories. Technical Report 2007-TR-017, ISTI-CNR, Pisa, 2007.

[11] R. G. Pensa, A. Monreale, F. Pinelli, and D. Pedreschi. Pattern-preserving k-anonymization of sequences and its application to mobility data mining. In *PiLBA*, 2008.

[12] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko. Visually-driven analysis of movement data by progressive clustering. *Information Visualization*, 7(3/4):225–239, 2007.

[13] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.

[14] M. Terrovitis and N. Mamoulis. Privacy preservation in the publication of trajectories. In *MDM*, pages 65–72, 2008.

[15] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang. Anonymizing moving objects: how to hide a mob in a crowd? In *EDBT*, pages 72–83, 2009.