A Generalisation-based Approach to Anonymising Movement Data

Gennady Andrienko[1], Natalia Andrienko[1], Fosca Giannotti[2,4], Anna Monreale[3], Dino Pedreschi[3,4], Salvatore Rinzivillo[2]
(1) Fraunhofer IAIS (Intelligent Analysis and Information Systems), Sankt Augustin, Germany
(2) KDDLab, ISTI, CNR, Pisa, Italy
(3) KDDLab, Computer Science Department, University of Pisa, Italy
(4) Center for Complex Network Research, Northeastern University, Boston, MA

ABSTRACT

The possibility to collect, store, disseminate, and analyze data about movements of people raises very serious privacy concerns, given the sensitivity of the information about personal positions. In particular, sensitive information about individuals can be uncovered with the use of data mining and visual analytics methods. In this paper we present a method for the generalization of trajectory data that can be adopted as the first step of a process to obtain *k*-anonymity in spatio-temporal datasets. 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.

1. INTRODUCTION

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 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. In particular, sensitive information about individuals can be uncovered with the use of data mining (Agrawal and Srikant, 2000) and visual analytics methods (Andrienko et al, 2007). 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 (Giannotti and Pedreschi, 2007).

We base our work on a method for generalization of movement data (Andrienko and Andrienko, 2010) that can be adapted for anonymization of movement data. 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. This method of generalization can be used as a starting point for an anonymization process in order to generate a dataset that satisfies the *k*-anonymity property, i.e. a dataset where each entry is indistinguishable with at least *k-1* other entries (Samarati and Sweeney, 1998). In the literature, the proposed methods for anonymization of movement data are based on randomization, point suppression and space translation (Abul et al., 2008, Nergiz et al. , 2007, Yarovoy et al. , 2009). The concept of spatial generalization as been widely used in the works on privacy for location based services (Gruteser et al., 2003, Mokbel et al., 2006, Mokbel et al., 2007). In (Yarovoy et al. , 2009) the authors use a spatial generalization technique based on a fixed spatial grid to discretize the movement data.

The novelty of our approach is the use of a spatial tessellation derived from the input movement data (see Section 2). The generalized trajectories maintain a high analytical utility as proved by several experiments on a real dataset of GPS-tracked data using a density based clustering method, namely OPTICS (Ankerst et al., 1999), with a suitable set of distance functions (Andrienko et al.,

2007) (see Section 3). We also discuss a general idea to assure the privacy protection into the generalized dataset (see Section 4).

2. GENERALIZATION METHOD

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.

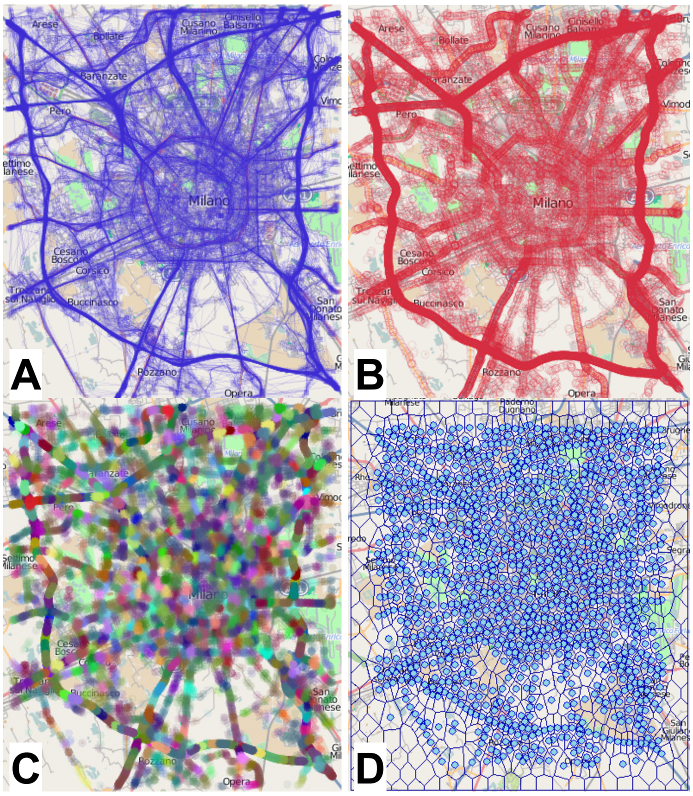

Figure 1: A) The original subset of 4,287 trajectories, shown with 10% opacity. B) The characteristic points extracted from the trajectories, shown with 10% opacity (54,362 points in total). C) The characteristic points have been clustered with the parameter maximum cluster radius = 500m. D) The centroids of the clusters and the Voronoi tessellation of the territory.

We demonstrate the work of the method by example of a subset consisting of 4,287 real GPS car trajectories from the city of Milan. Figure 1A presents a map with the original trajectories. In 1B, there are the characteristic points extracted from the trajectories. Both the trajectories and the characteristic points are shown with 10% opacity, to enable the estimation of the densities in different places. In Figure 1C, the characteristic points have been clustered. The clusters are represented by coloring. We use a special spatial clustering algorithm with a parameter defining the maximum radius (spatial extent) of a cluster. The clusters in Figure 1C have been obtained for the value 500 meters of this parameter. Figure 1D 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.

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 pi 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 $\{a_1, a_2, ..., a_n\}$. There may be also a case when all points of a trajectory are contained in one and the same area a_1. Then, the whole trajectory is represented by the sequence $\{a_i\}$. 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.

When a trajectory is represented by a sequence of at least two areas, two possibilities exist for two consecutive areas a_i and a_{i+1},: (1) a_i and a_{i+1} are adjoining areas, i.e. having a common edge; (2) a_i and a_{i+1} are not adjoining. In the second case, there is an optional possibility to insert intermediate areas between a_i and a_{i+1} so that each two consecutive areas will be adjoining. We do this by means of linear interpolation. Let p_m be the last point of the trajectory contained in a_i. We build a straight line between p_m and p_{m+1}, which is contained in a_{i+1}, and find all areas intersected by this line. These areas are inserted in the sequence between a_i and a_{i+1}. The corresponding intermediate points of the trajectory are computed as the points of the crossing line having the minimum distances to the generating points of the cells.

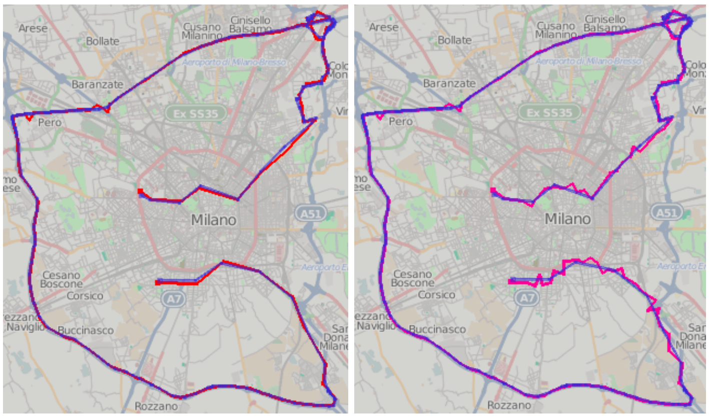

Figure 2: Both screenshots of a map display represent one and the same original trajectory (in blue) and two variants of corresponding generalized trajectories: without interpolation (red, left) and with interpolation (purple, right).

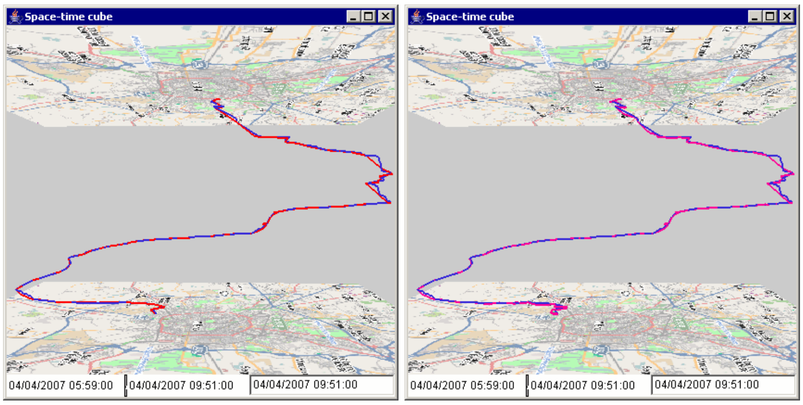

Figure 3: The trajectories demonstrated in Figure 2 are displayed in a space-time cube (where the vertical dimension represent time) for enabling comparison of the temporal characteristics. Left: without interpolation, right: with interpolation.

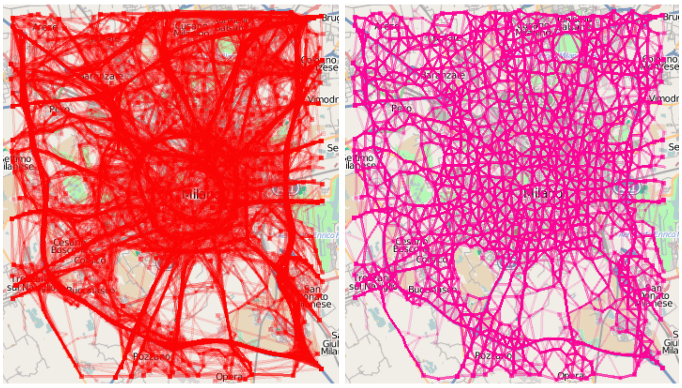

Figure 4: The generalised trajectories of the cars from Milan are shown with 10% opacity; the line thickness is 2 pixels. Left: the variant without interpolation, right: the variant with interpolation.

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, …, a_n}* is replaced, for practical purposes, by the sequence *{c_1, c_2, …, c_n}* consisting of the centroids of the areas *{a_1, a_2, …, a_n}*. As a result, we obtain generalized trajectories. Figures 2-4 illustrate the results of the generalization of the cars trajectories from Milan.

As can be seen from the illustrations, the variant of the generalization involving interpolation may introduce additional distortions in the original trajectories in cases when the distances between recorded positions are larger than the sizes of the areas. On the other hand, this may increase the degree of anonymity.

3. PRESERVATION OF PATTERNS IN THE GENERALIZED TRAJECTORIES

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 (with interpolation) of these trajectories using the generic density-based clustering algorithm OPTICS with a suitable distance function (Rinzivillo et al., 2008, Andrienko et al., 2009).

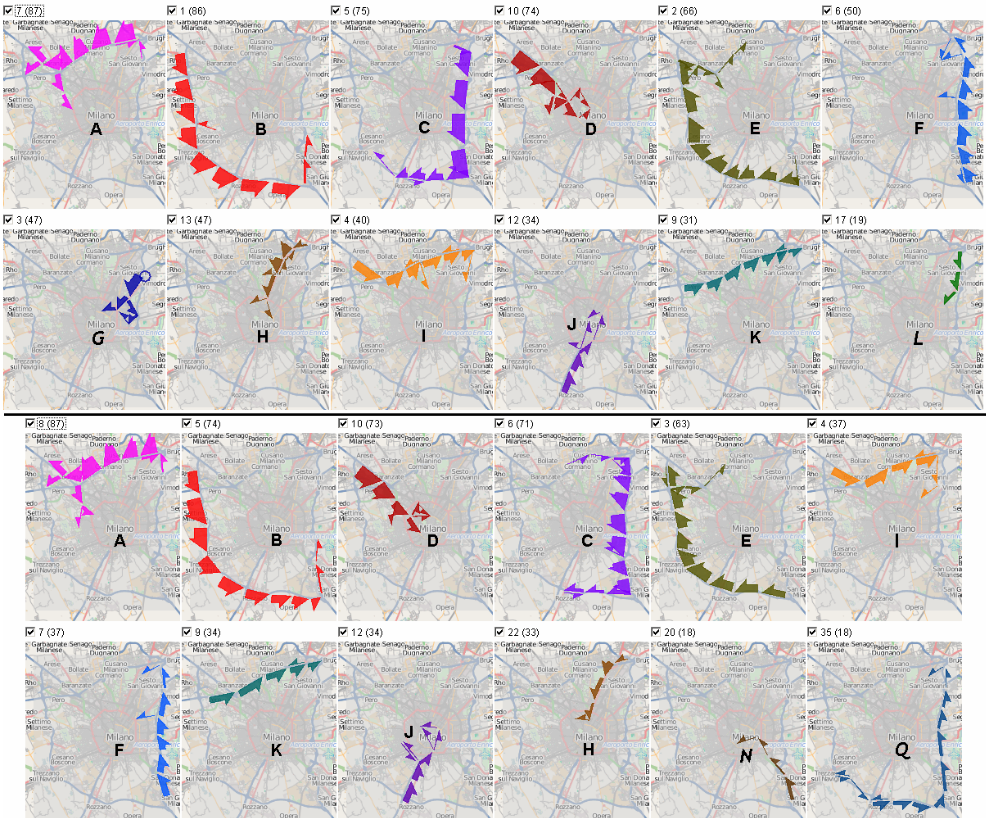

Figure 5: Comparison of clustering results for the original (top) and generalized (bottom) trajectories. 12 biggest clusters from each result are visible.

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 5 demonstrates the clusters obtained from the set of original trajectories and from the set of generalized trajectories, respectively. In both cases, we have used the same distance function "route similarity". The distance threshold is 500m in the first case and 250m in the second case. 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 demonstrate that the general structure and shapes of the clusters are preserved.

4. GENERALIZATION VS K-ANONYMITY.

The approach described in Section 2, 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).

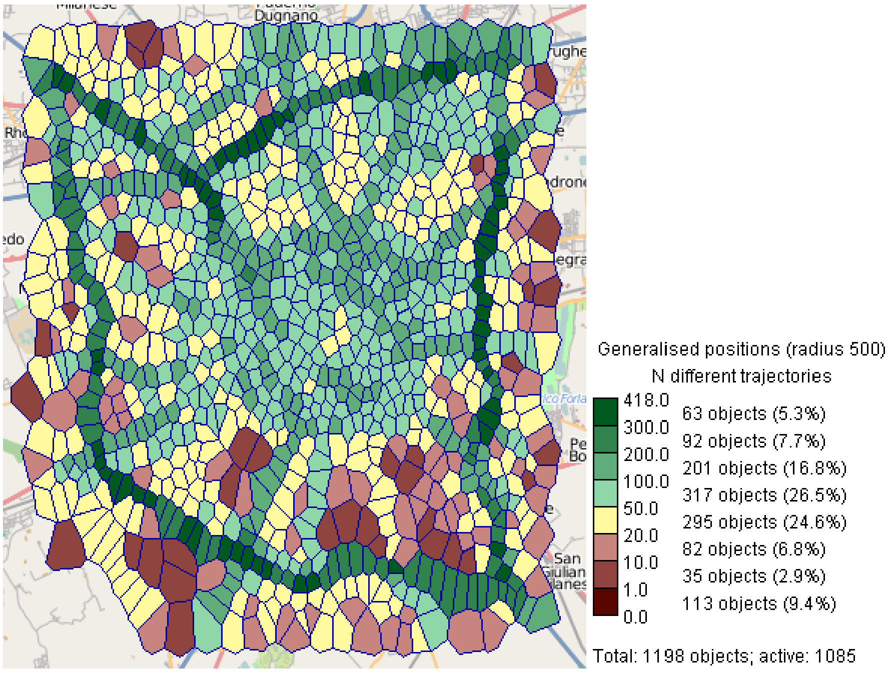

Figure 6: The map shows the numbers of different trajectories that visited the areas of the territory division by area coloring. The areas that do not contain any points from the trajectories are hidden

The satisfaction of these anonymity conditions is easy to check. Thus, the map in Figure 6 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 a neighboring place *b*, we unite the cells by computing a new centroid for all points belonging to these cells and then run the tessellation method again. Figures 7 and 8 demonstrate the flows between cells for the original tessellation and after performing several iterations of uniting neighboring cells with small number of moves between them.

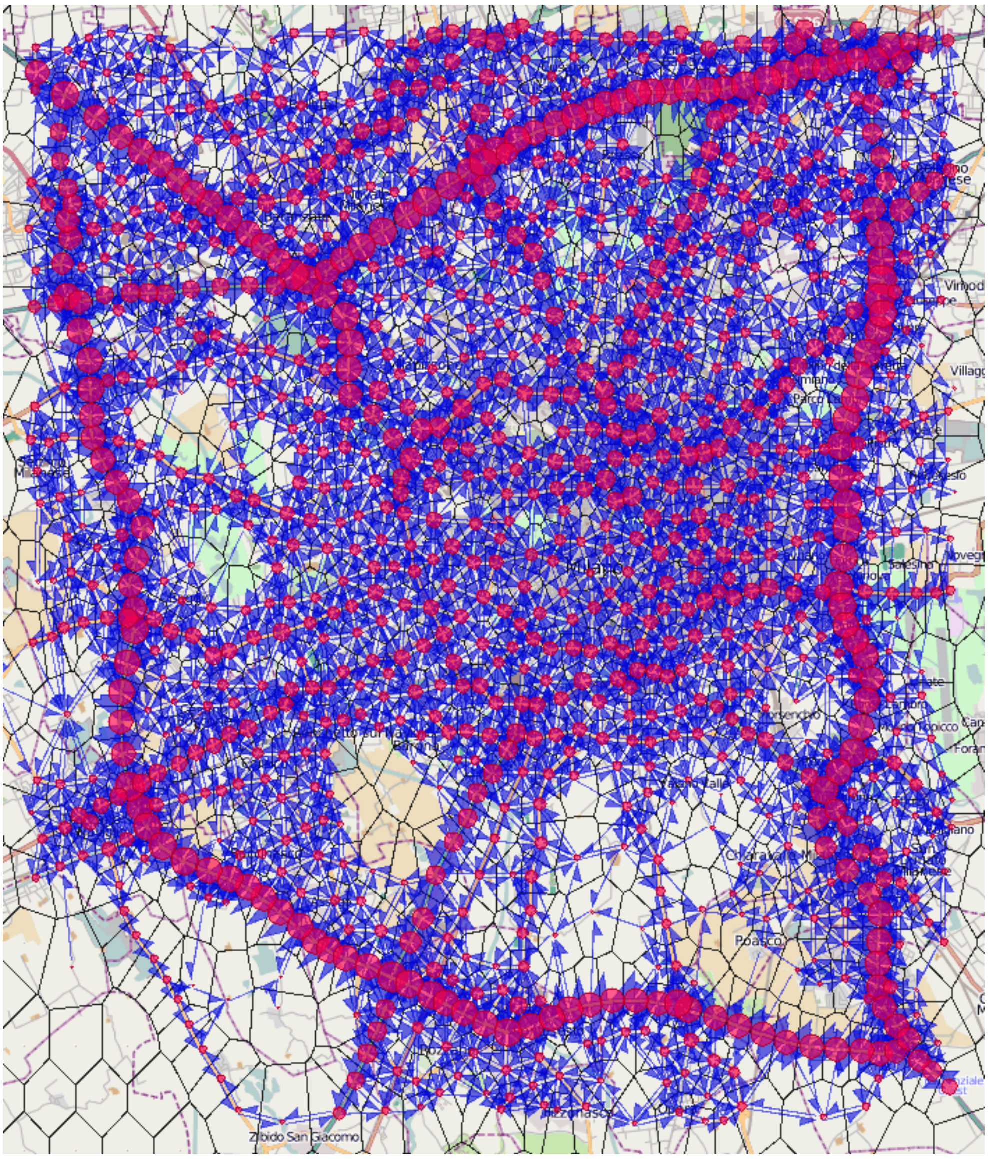

Figure 7: Presence in cells (N of different trajectories shown by graduated circles) and flows between the cells (N of moves in two directions shown by arrows with variable thickness) displayed for the original tessellation with 500 meters bounds: 1,212 polygons connected by 5,646 aggregate moves.

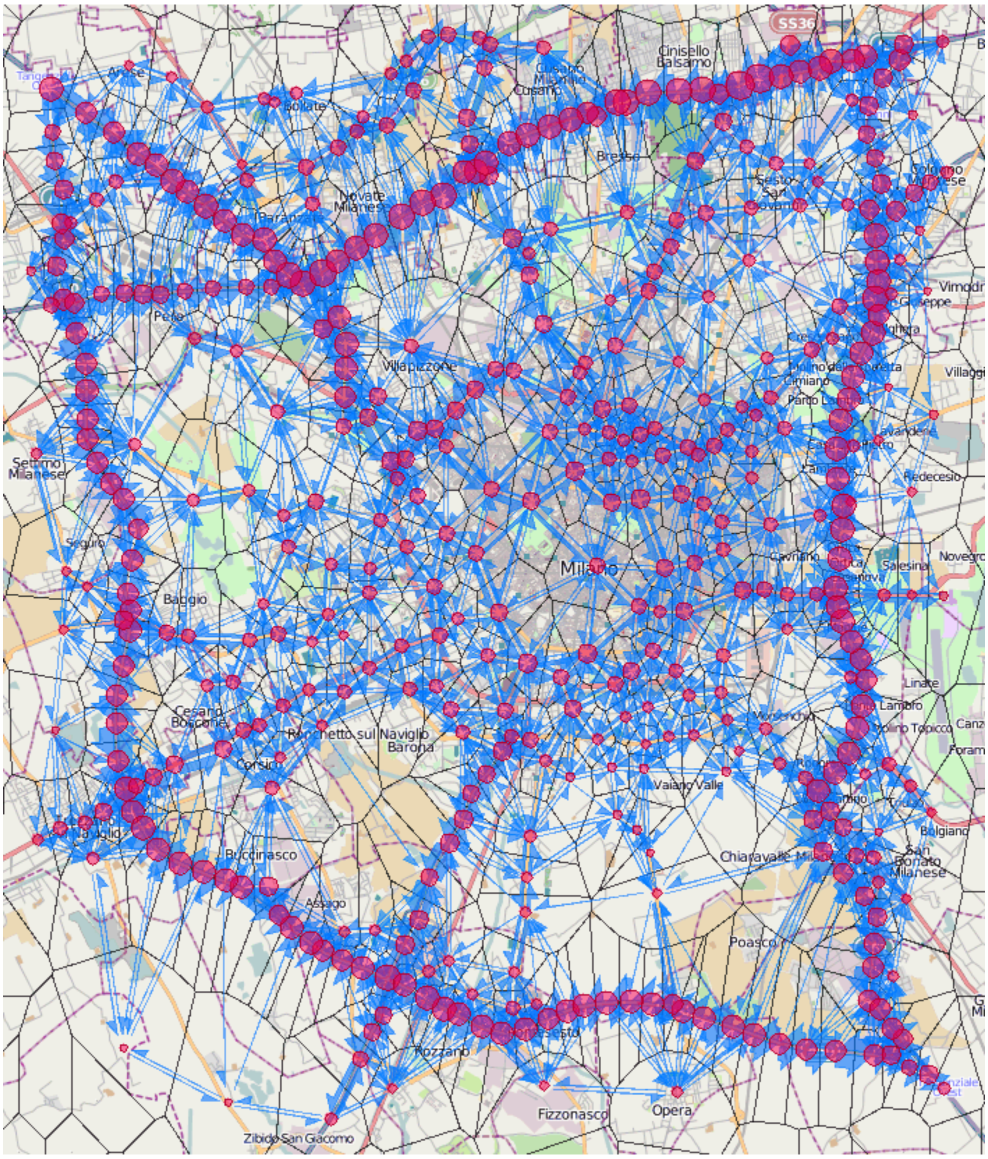

Figure 8: Presence in cells (N of different trajectories shown by graduated circles) and flows between the cells (N of moves in two directions shown by arrows with variable thickness) displayed for the tessellation after varying the aggregation level: 535 polygons connected by 2,379 active moves

Besides increasing the degree of generalization in data-sparse regions, trajectories can be further anonymised by removing the segments that occur in less than *k* trajectories.

The generalization-based trajectory 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.

5. CONCLUSION.

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 appearance 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 research community because of their spatial and temporal characteristics.

In this 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. In future work, we intend to perform a formal evaluation of the method and investigate further the protection model against the re-identification attack.

BIBLIOGRAPHY

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

Agrawal, R. and Srikant, R.. Privacy-preserving data mining. In SIGMOD,pp. 439-450. ACM, 2000.

Andrienko, N., Andrienko, G.: Spatial Generalization and Aggregation of Massive Movement Data, IEEE Transactions on Visualization and Computer Graphics (TVCG), in press (2010).

Andrienko, G., Andrienko, N., and Wrobel, S. Visual Analytics Tools for Analysis of Movement Data, ACM SIGKDD Explorations, v.9 (2), pp.38-46, 2007.

Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., and Giannotti, F. 2009. Interactive Visual Clustering of Large Collections of Trajectories. In VAST 2009.

Ankerst, M., Breunig, M., Kriegel, H.-P., and Sander, J. Optics: Ordering points to identify the clustering structure. In Proc. ACM SIGMOD, 49–60. 1999.

Giannotti, F., and Pedreschi, D., eds. 2007. Mobility, Data Mining and Privacy - Geographic Knowledge discovery. Springer, Berlin.

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

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

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

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

Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., and Andrienko, G. 2008. Visually–driven analysis of movement data by progressive clustering, Information Visualization, 7(3/4), 225-239.

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

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