Disclosure Control by Computer Scientists: An Overview and an Application of Microaggregation to Mobility Data Anonymization

Josep Domingo-Ferrer and Michal Sramka
Universitat Rovira i Virgili, Department of Computer Engineering and Maths,
UNESCO Chair in Data Privacy,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
E-mail: josep.domingo@urv.cat, michal@sramka.net

1 Introduction

Privacy-preserving data mining (PPDM) is a subdiscipline of computer science which in many respects is parallel to statistical disclosure control (SDC) within statistics. See [12] for a survey of recent developments in PPDM.

We focus here on the connections between k-anonymity, a concept arisen in the PPDM community, and microaggregation, a family of methods developed within SDC. This is discussed at a conceptual level in Section 2. We then move to anonymization of mobility data, *i.e.* trajectories, a very dynamic area in PPDM and a completely neglected one in SDC. In Section 3 we apply the microaggregation approach to k-anonymize real-world trajectories. We present a new distance measure for spatio-temporal data that facilitates the microaggregation process. The measure naturally considers *both* spatial and temporal aspects and can be fine-tuned for specific applications and instantiated with existing measures for spatial data, sequences, or time series. Conclusions are summarized in Section 4.

2 K-Anonymity via microaggregation

One of the guiding principles for guaranteeing anonymity to an individual is to hide the individual's uniqueness and make him/her look indistinguishable within the group. This is often referred to as "hide-in-the-crowd" or "safety-in-number" principle. More formally, a set of similar objects forms an *anonymity set*. An object in an anonymity set $\mathcal{A}$ enjoys *(perfect) anonymity* if it cannot be recognized from the other objects in $\mathcal{A}$ with probability greater than $1/|\mathcal{A}|$.

K-Anonymity. The above idea gave birth to the k-anonymity notion [13, 14]. A *microdata* set consists of attributes that are either identifying, key, or sensitive. *Identifying attributes* are those attributes that unambiguously and uniquely identify an individual, *e.g.*, social security numbers. To achieve anonymity, they are encrypted/removed from the microdata before publishing. *Key attributes* (or *quasi-identifiers*) are attributes that, in combination, can be potentially used to identify an individual (by linking them with some external information). Examples are date of birth or age, job, address, or gender. Unlike identifiers, key attributes cannot be removed from the microdata, because any attribute is potentially a key attribute. Finally, *sensitive attributes* are attributes which contain information that is considered to be private/sensitive to the corresponding individual, *e.g.*, salary, religion, medical conditions.

A *protected dataset* is said to satisfy *k-anonymity* for $k > 1$ if, for each combination of key attributes, at least k records exist in the dataset sharing that combination. This means that

an individual can be confused with at least $k-1$ other individuals based on a combination of his/her key attributes. In other words, no individual can be identified with probability better than $1/k$.

Microaggregation. Microaggregation is a family of microdata anonymization methods which can be operationally defined in terms of the following two steps:

1. *Partition:* the set of original objects is partitioned into several clusters in such a way that objects in the same cluster are similar to each other and there are at least k objects per cluster.

2. *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 objects; each object is replaced by the cluster prototype.

The generic concept of microaggregation as described above has been proposed in [7]. Joint multivariate microaggregation of all key attributes with minimum group size k has been proposed in [8] as an alternative to achieve k-anonymity; besides being simpler, this alternative has the advantage of yielding complete data without any coarsening (nor categorization in the case of numerical data). Microdata anonymization through condensation [3] or hybrid-data generation [9] is a variation of microaggregation which masks the data in clusters by replacing them with synthetic data.

3 Anonymizing real-world trajectories

Huge amounts of movement data are automatically generated by technologies such as GPS, GSM, RFID, etc. The ever increasing capacity to store data allows such object movement data to be collected in large databases. Analyzing the movement data can lead to useful or previously unknown knowledge. In particular, publishing such data is essential to improve transportation (intelligent transportation, traffic monitoring and planing, congestion trends) or to understand the dynamics of the economy in a region (supply chain management, market trends, etc.). However, there are obvious threats to the privacy of individuals if their movement data (trajectories) are published in a way which allows re-identification of the individual behind a trajectory.

We propose to use microaggregation to anonymize trajectories. For the purpose of clustering trajectories, we present a new generic distance measure for trajectories that can be instantiated for specific cases and applications. We also briefly discuss the methods and techniques that are of use in masking the clustered trajectories.

Related work. There are a few methods for anonymizing object movement data in the literature about anonymization and location-based privacy [4]. Here we concentrate on the related work in trajectory anonymization that employs microaggregation. Abul et al. [1] proposed microaggregation-like approach that achieves so-called (k, δ)-anonymity for trajectories. They used Euclidean distance to compare trajectories that have the same time span; trajectories not having the same time span could not be compared and therefore were anonymized as different datasets. Recently, Abul et al. [2] extended their previous work by using the Edit Distance on Real Sequences (EDR) to compare trajectories. The EDR is unfortunately not suitable for microaggregation as it does not differentiate between close or distant trajectories (only about how many changes, no matter how close or far, are needed to move from one to to the other trajectory).

Trajectories. Movement data can be modeled in many ways. Various properties of the mobile object (*e.g.*, position, direction, speed) are sampled at particular times as the object moves. Without loss of generality, we restrict ourselves to positions in a plane. Then, a triple (t, x, y) denotes that at time t an object is in the position (x, y).

Definition 1. A *trajectory* is an ordered set of triples $T = \{(t_1, x_1, y_1), \ldots, (t_n, x_n, y_n)\}$, where $t_i < t_{i+1}$ for all $1 \leq i < n$.

Definition 2. Two trajectories $T_i = \{(t_1^i, x_1^i, y_1^i), \ldots, (t_n^i, x_n^i, y_n^i)\}$ and $T_j = \{(t_1^j, x_1^j, y_1^j), \ldots, (t_m^j, x_m^j, y_m^j)\}$ are said to be *p%-contemporary* if

$$p = 100 \cdot \min(\frac{I}{t_n^i - t_1^i}, \frac{I}{t_m^j - t_1^j})$$

with $I = \max(\min(t_n^i, t_m^j) - \max(t_1^i, t_1^j), 0)$.

Intuitively, two trajectories are 100%-contemporary if and only if they start at the same time and end at the same time; two trajectories are 0%-contemporary trajectories if and only if they occur during non-overlapping time intervals.

A new generic distance for trajectories. A distance measure for trajectories is needed to partition trajectories into clusters. An optimal distance needs to consider both spatial and temporal aspects of trajectories and needs to be able to compare any two trajectories. Here we propose such a distance measure based on the p-contemporary definition above. The measure can be instantiated using new or existing distance measures and fine-tuned for different applications. A specific instantiation of the proposed measure that used time-synchronization of trajectories and Euclidean distance has been successfully used in trajectory anonymization in [10, 11].

There are existing distance measures for time series or for comparison of two time-shifted trajectories. See [5] for a survey. Some of the distances, *e.g.*, the Euclidean distance, require that trajectories be defined over the same time span and that, at each time, both measured trajectories be defined. In these cases, it is often assumed that the objects move in constant speed and so the remaining points can be interpolated. As far as we know, nobody has considered extrapolation to an underlying map. Let Δ be a distance function for trajectories, *e.g.*, the Euclidean distance.

Definition 3. Consider a set of trajectories $\mathcal{T} = \{T_1, \ldots, T_N\}$ where each trajectory is written as $T_i = \{(t_1^i, x_1^i, y_1^i), \ldots, (t_{n^i}^i, x_{n^i}^i, y_{n^i}^i)\}$. The *distance between trajectories* is defined as follows. If $T_i, T_j \in \mathcal{T}$ are p%-contemporary with $p > 0$, then

$$d(T_i, T_j) = \frac{1}{p} \cdot \Delta(T_i, T_j) \ .$$

If $T_i, T_j \in \mathcal{T}$ are 0%-contemporary but there is at least one subset of $\mathcal{T}$

$$\mathcal{T}^k(ij) = \{T_1^{ijk}, T_2^{ijk}, \ldots, T_{n^{ijk}}^{ijk}\} \subseteq \mathcal{T}$$

such that $T_1^{ijk} = T_i$, $T_{n^{ijk}}^{ijk} = T_j$ and T_ℓ^{ijk} and $T_{\ell+1}^{ijk}$ are p_ℓ%-contemporary with $p_\ell > 0$ for $\ell = 1$ to $n^{ijk} - 1$, then

$$d(T_i, T_j) = \min_{\mathcal{T}^k(ij)} \left(\sum_{\ell=1}^{n^{ijk}-1} d(T_\ell^{ijk}, T_{\ell+1}^{ijk}) \right)$$

Otherwise $d(T_i, T_j)$ is not defined.

The computation of the distance between every pair of trajectories is not exponential as it could seem from the definition. A *distance graph* for $\mathcal{T} = \{T_1, \ldots, T_N\}$ is a weighted graph where nodes represent trajectories, two nodes are adjacent if the corresponding trajectories are p%-contemporary for some $p > 0$, and edge weights are the distances between the corresponding trajectories. The distance graph, and hence the distance $d(T_i, T_j)$ for *any two* trajectories, can be computed in polynomial time $O(N^3)$ using the Floyd-Warshall algorithm. In particular, missing distances are computed as the minimum cost path between the nodes of connected components of the graph.

Anonymizing clusters of trajectories. We can now use any microaggregation algorithm (for example, MDAV [8]) on trajectories to minimize the sum of intra-cluster distances measured with the above distance and ensuring that the cluster size is at least k. Once we have the clusters, we can swap triples between the trajectories inside the cluster provided that the times and the spatial coordinates of the swapped triples are below preselected temporal and spatial thresholds.

The resulting utility features are very interesting:

- Time information is taken into account and it is preserved;
- The original locations are preserved (no fake locations are introduced);
- The lengths of the microaggregated trajectories do not exactly match the lengths of the original trajectories, but they are strongly correlated,
- The shape of original trajectories is fairly well preserved;
- The number of discarded trajectories is much reduced in comparison to competing methods in the literature.

Regarding disclosure risk, experimental results reported in [10, 11] indicate that our method achieves a lower re-identification probability compared to other competitors like (k, δ)-anonymity when the distortion is the same.

4 Concluding remarks

We have discussed how the k-anonymity concepts (arisen the PPDM community) can be implemented using microaggregation (arisen in the SDC community) to anonymize mobility data (only considered in the PPDM community but which should also be considered in the SDC community). Specifically, a trajectory k-anonymization method based on microaggregation which has interesting utility and privacy-preserving properties has been presented.

Disclaimer and acknowledgments

This work was partly funded by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the Government of Catalonia through grant 2009 SGR 1135. The first author is partly supported as an ICREA-Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but the views expressed in this paper are their own and do not commit UNESCO.

References

[1] Abul, O., Bonchi, F., Nanni, M., (2008). Never walk alone: uncertainty for anonymity in moving objects databases. In *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008*, pp. 376-385, IEEE.

[2] Abul, O., Bonchi, F., Nanni, M., (2010). Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884–910.

[3] Aggarwal, C. C., Yu, P. S., (2004). A condensation approach to privacy preserving data mining. In *Advances in Database Technology - EDBT 2004*, LNCS 2992, pp. 183-199, Springer.

[4] Bonchi, F., (2009). Privacy preserving publication of moving object data. In *Privacy in Location-Based Applications, Research Issues and Emerging Trends*, LNCS 5599, pp. 190-215. Springer.

[5] Chen, L., Özsu, M. T., Oria, V., (2005). Robust and fast similarity search for moving object trajectories. In *2005 ACM SIGMOD Intl. Conference on Management of Data*, pp. 491-502, ACM Press.

[6] Dalenius, T., (1986). Finding a needle in a haystack - or identifying anonymous census records. *Journal of Official Statistics*, 2(3): 329-326.

[7] Domingo-Ferrer, J., Mateo-Sanz, J. M., (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201.

[8] Domingo-Ferrer, J., Torra, V., (2005). Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212.

[9] Domingo-Ferrer, J., González-Nicolás, U., (2010). Hybrid microdata using microaggregation. *Information Sciences*, 180(15):2834-2844.

[10] Domingo-Ferrer, J., Sramka, M., Trujillo-Rasua, R., (2010). Privacy-preserving publication of trajectories using microaggregation. In *3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS (SPRINGL 2010)*, San Jose CA, USA, ACM Press, pp. 26-33.

[11] Domingo-Ferrer, J., Trujillo-Rasua, R., Sramka, M., (2011). Microaggregation-based anonymization of mobility data. (Manuscript).

[12] Fung, B., Wang, K., Chen, R., Yu, P. S., (2010). Privacy-preserving data publishing: a survey on recent developments. *ACM Computing Surveys*, 42(4), Article no. 14.

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

[14] Samarati, P., (2001). Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

RÉSUMÉ

Nous revenons sur le concept de k-anonymat, surgi dans la communauté informatique, qui peut être réalisé par micro-agrégation, une famille de méthodes parue dans la communauté statistique. En particulier, on présente une méthode de micro-agrégation pour k-anonymiser des données de mobilité. Il s'agit d'un type de données activement traité par les informaticiens, mais assez négligé par les statisticiens officiels.