WP. 32
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (v): Privacy for new types of microdata: sequence data and mobility data

Anonymization of trajectory data

Prepared by Josep Domingo-Ferrer and Rolando Trujillo Universitat Rovira i Virgili, Spain

Anonymization of Trajectory Data

Josep Domingo-Ferrer and Rolando Trujillo-Rasua

Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, UNESCO Chair in Data Privacy, Av. Països Catalans 26, E-43007 Tarragona, Catalonia.
{josep.domingo,rolando.trujillo}@urv.cat

Abstract. Trajectories of mobile objects, are automatically collected in huge quantities. Publishing and exploiting such data is essential to improve planning, but it threatens the privacy of individuals: re-identification of the individual behind a trajectory is easy unless precautions are taken. We present two heuristics for privacy-preserving publication of trajectories. Both of them publish only true locations. The first heuristic is based on trajectory microaggregation and on location permutation; it effectively achieves trajectory k-anonymity. The second heuristic is based only on location permutation; it gives up trajectory k-anonymity and aims at a different property named location k-diversity. The advantage of the second heuristic is that it takes into account reachability constraints when computing anonymized trajectories.

1 Introduction

Trajectories of mobile objects (individuals, cars, etc.) are routinely collected or at least collectible by such technologies as GPS, RFID, GSM, etc. The availabily of trajectories, that is, mobility data, is extremely useful for public and corporate planning purposes. However, publication of original collected mobility data would result in obvious privacy disclosure: even if de-identified, trajectories are easily linkable [22] to the individuals they correspond to and they tell a lot about that individual's lifestyle and habits. Furthermore, sensitive locations (hospitals, etc.) visited by individuals may be disclosed.

Unfortunately, the traditional anonymization and sanitization methods for microdata [12, 10, 15] cannot be directly applied to spatio-temporal data without considerable expense in computation time and information loss. Hence, there is a need for specific anonymization methods to thwart privacy attacks and therefore reduce privacy risks associated with publishing trajectories.

In this paper, we sketch two methods for trajectory anonymization which yield anonymized trajectories formed by fully accurate true original locations and whose distinctive features are: (i) The first method aims at trajectory k-anonymity. (ii) The second method takes reachability constraints into account, and it tries to reduce the fraction of discarded locations by replacing trajectory k-anonymity with location

k-diversity. Full details on the proposed methods, plus formal statements and proof of privacy guarantees and experimental results can be found in [9].

2 Related work

Trajectories can be modeled and represented in many ways [11]. Without loss of generality, we consider a trajectory to be a time-stamped path in a plane. More formally, let *time-stamped location* be a triple (t, x, y) with t being a time-stamp and (x, y) a *location* in $\mathbb{R}^2$. Intuitively, the time-stamped location denotes that at time t an object is at location (x, y).

Definition 1 (Trajectory) *A* trajectory *is an ordered set of time-stamped locations*

$$T = \{(t_1, x_1, y_1), \ldots, (t_n, x_n, y_n)\} \ , \tag{1}$$

where $t_i < t_{i+1}$ *for all* $1 \leq i < n$.

Definition 2 (Sub-trajectory) *A trajectory* $S = \{(t'_1, x'_1, y'_1), \ldots, (t'_m, x'_m, y'_m)\}$ *is a* sub-trajectory *of* T *in Expression* (1)*, denoted* $S \preceq T$*, if there exist integers* $1 \leq i_1 < \ldots < i_m \leq n$ *such that* $(t'_j, x'_j, y'_j) = (t_{i_j}, x_{i_j}, y_{i_j})$ *for all* $1 \leq j \leq m$.

Hereinafter, we will use *triple* as a synonym for time-stamped location. When there is no risk of ambiguity, we also say just "location" to denote a time-stamped location.

Several anonymity notions and methods for trajectories have been proposed. Among those works, the closest to our approach is the notion of (k, δ)-anonymity [1, 2]. In the original method –Never Walk Alone (NWA) [1]–, the set of trajectories is partitioned into disjoint subsets in which trajectories begin and end at roughly the same time; then trajectories within each set are clustered using the Euclidean distance. In the follow-up method –Wait For Me (W4M) [2]–, the original trajectories are clustered using the edit distance on real sequences (EDR) [5]. Both approaches proceed by anonymizing each cluster separately. Two trajectories T_1 and T_2 are said to be co-localized with respect to δ in a certain time interval $[t_1, t_n]$ if for each triple (t, x_1, y_1) in T_1 and each triple (t, x_2, y_2) in T_2 with $t \in [t_1, t_n]$, it holds that the spatial Euclidean distance between both triples is not greater than δ. Anonymity in this context means that each trajectory is co-localized with at least $k - 1$ other trajectories. Anonymization is achieved by spatial translation of trajectories inside a cluster of at least k trajectories having the same time span. In the special case when $\delta = 0$, the method produces one centroid/average trajectory that represents each and all trajectories in the cluster. *Ad hoc* preprocessing and outlier removal facilitate the process. The problem with the NWA method is that partitioning the set of all trajectories into subsets sharing the same time span may produce too many subsets with too few trajectories inside each of them; clearly, a subset with less than

k trajectories cannot be k-anonymized. Also, setting a value for δ may be awkward in many applications, *e.g.* trajectories recorded using RFID technology.

Another k-anonymity based notion for trajectories consisting of ranges of points and ranges of times has been proposed in [18] and [19]. It uses clustering to minimize the "log cost metric"; this balances the spatial and temporal translations with user-provided weights. Minimizing the log cost therefore maximizes utility. The clusters are anonymized by matching points of the trajectories and generalizing them into minimum bounding boxes. Unmatched points are suppressed and so are some trajectories. The anonymized data are not released; instead, synthetic "atomic" trajectories (having unit x-range, y-range and time range) are generated by sampling the bounding boxes. This approach does not release standard trajectories but only trajectories with unit ranges.

In [17], k-anonymity means that an original trajectory T is generalized into a trajectory $g(T)$ (without the time information) in such a way that $g(T)$ is a subtrajectory of the generalizations of at least $k-1$ other original trajectories. Ignoring the time information during anonymization and complex plane tessellations used to achieve the k-anonymity are the main drawbacks of this method. Utility is measured by comparing clustering results.

Some approaches assume that the data owner anonymizing the database knows exactly what the adversary's knowledge is. If the adversary is assumed to know different parts of trajectories, then those are removed from the published data [21]. However, this work only considers sequential place visitation without real timestamps. If the adversary is assumed to use some prediction of continuation of a trajectory based on previous path and speed, then uncertainty-aware path cloaking [13, 14] can suppress these trajectories; this however results in high information loss.

In contrast to these methods, we perform traditional microaggregation over all original trajectories —we do not specially and separately consider trajectories having the same time span and we consider trajectories over locations, not ranges, without stripping the time information. We publish synthetic trajectories which are analogous to condensed or hybrid microdata [3, 7]. However, our synthetic trajectories are formed by locations covered by the original trajectories. This means that the location points of our anonymized trajectories remain on the underlying network map.

3 Anonymization methods

We present two anonymization methods, called SwapLocations and ReachLocations, respectively, both of which yield anonymized trajectories formed by original locations. The first of them is partially based on microaggregation [6] of trajectories and partially based on permutation of locations. The second method is based on

permutation of locations. The main difference between the SwapTriples method [8] and the two new methods we propose here is that the latter effectively guarantee trajectory k-anonymity (SwapLocations) or location k-diversity (ReachLocations). To that end, an original triple is discarded if it cannot be swapped randomly with another triple drawn from a set of $k-1$ other original triples.

3.1 The SwapLocations method

This method first partitions the set of trajectories into several clusters, using microaggregation. Then, each cluster is anonymized using the SwapLocations function (see below).

We limit ourselves to clustering algorithms which try to minimize the sum of the intra-cluster distances. To do so, several distance measures can be used [20, 4, 16, 5]. However, we suggest the use of the distance measure proposed in [8] since it naturally considers both spatial and temporal aspects of trajectories. The cardinality of each cluster must be approximately k, with k an input parameter; if the number of trajectories is not a multiple of k, one or more clusters must absorb the up to $k-1$ remaining trajectories, hence those clusters will have cardinalities between $k+1$ and $2k-1$. The purpose of setting k as the cluster size is to fulfill trajectory k-anonymity.

The SwapLocations function begins with a random trajectory T in C. The function attempts to cluster each unswapped triple λ in T with another $k-1$ unswapped triples belonging to different trajectories such that: i) the time-stamps of these triples differ by no more than a time threshold R^t from the time-stamp of λ; ii) the spatial coordinates differ by no more than a space threshold R^s. If no $k-1$ suitable triples can be found that can be clustered with λ, then λ is removed; otherwise, random swaps of triples are performed within the formed cluster. As a result, at least one of the trajectories returned by SwapLocations has all its triples swapped.

3.2 The ReachLocations method

The ReachLocations method takes reachability constraints into account: from a given location, only those locations at a distance below a threshold *following a path in an underlying graph* (*e.g.*, urban pattern or road network) are considered to be directly reachable. Enforcing such reachability constraints while requiring full trajectory k-anonymity would result in a lot of original locations being discarded. To avoid this, trajectory k-anonymity is changed by another useful privacy definition: location k-diversity.

Computationally, this means that trajectories are *not* microaggregated into clusters of size k. Instead, each location is k-anonymized independently using the entire set of locations of all trajectories. To do so, a cluster C_λ of "unswapped" locations is created around a given location λ, *i.e.* $\lambda \in C_\lambda$. The cluster C_λ is constrained as follows: i) it must have the lowest intra-cluster distance among those clusters of k "unswapped" locations that contain the location λ; ii) it must have locations

belonging to k different trajectories; and iii) it must contain only locations at a path from λ at most R^s long and with time-stamps differing from t_λ at most R^t. Then, the spatial coordinates (x_λ, y_λ) are swapped with the spatial coordinates of some random location in C_λ and both locations are marked as "swapped". If no cluster C_λ can be found, the location λ is removed from the data set and will not be considered anymore in the subsequent anonymization. This process continues until no more "unswapped" locations appear in the data set.

4 Privacy guarantees

The main difference between the SwapTriples method in [8] and the SwapLocations method here is that, in the latter, no original location survives unswapped in an anonymized trajectory. Therefore, any sub-trajectory $S' \preceq S \preceq T_1$ has the same probability of being a sub-trajectory of the anonymized version of T_1 ($T_1^\star$) than of being a sub-trajectory of any of the $k-1$ anonymized trajectories $T_2^\star, \cdots, T_k^\star$. Thus, given S, an adversary is not able to link T_1 with $T_1^\star$ with probability higher than $\frac{1}{k}$. In this sense, the SwapLocations method achieves *trajectory k-anonymity*.

On the other hand, the ReachLocations method guarantees that any triple λ in an original trajectory T appears in the anonymized trajectory $T^\star$ corresponding to T if and only if λ was not removed and was swapped with itself, which happens with probability at most $\frac{1}{k}$. This is what we call *location k-diversity*.

It should be remarked that, in ReachLocations two successive locations λ_j^i and λ_{j+1}^i of an original trajectory T_i may be cloaked with respective sets of $k-1$ locations belonging to different sets of $k-1$ original trajectories; this is why we cannot speak of trajectory k-anonymity for ReachLocations.

5 Conclusions

We have presented two permutation-based heuristic methods to anonymize trajectories with the common features that: i) places and times in the anonymized trajectories are true original places and times with full accuracy; ii) both methods can deal with trajectories with partial or no time overlap. Thanks to microaggregation, the first method achieves trajectory k-anonymity. The second method has the feature of taking reachability constraints into account, that is, it assumes a territory constrained by a network of streets or roads; to avoid removing too many locations, it aims at location k-diversity rather than trajectory k-anonymity.

References

[1] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: uncertainty for anonymity in moving objects databases. In *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008*, Cancun, Mexico, 7-12 April 2008, pages

376–385. IEEE, 2008.

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

[3] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. In *Proceedings of the 9th International Conference on Extending Database Technology, EDBT 2004*, Heraklion, Crete, Greece, 14-18 March 2004, volume 2992 of *Lecture Notes in Computer Science*, pages 183–199. Springer, 2004.

[4] E. M. Arkin, L. P. Chew, D. P. Huttenlocher, K. Kedem, and J. S. B. Mitchell. An efficiently computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):209–216, 1991.

[5] L. Chen, M. T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of 2005 ACM SIGMOD International Conference on Management of Data*, Baltimore, Maryland, USA, 14-16 June 2005, pages 491–502. ACM, 2005.

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

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

[8] J. Domingo-Ferrer, M. Sramka, and R. Trujillo-Rasúa. Privacy-preserving publication of trajectories using microaggregation. In *Proceedings of the SIGSPATIAL ACM GIS 2010 International Workshop on Security and Privacy in GIS and LBS, SPRINGL 2010*, San Jose, California, USA, 2 November 2010. ACM, 2010.

[9] J. Domingo-Ferrer and R. Trujillo-Rasua. Microaggregation- and permutation-based anonymization of mobility data (submitted manuscript, 2011).

[10] J. Domingo-Ferrer and E. Magkos (eds.), *Privacy in Statistical Databases*, LNCS 6344, Springer, 2011.

[11] L. Forlizzi, R. H. Güting, E. Nardelli, and M. Schneider. A data model and data structures for moving objects databases. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD 2000*, Dallas, Texas, USA, 16-18 May 2000, pages 319–330. ACM, 2000.

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

[13] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Preserving privacy in GPS traces via uncertainty-aware path cloaking. In *Proceedings of the 2007 ACM Conference on Computer and Communications Security, CCS 2007*, Alexandria, Virginia, USA, 28-31 October 2007, pages 161–171. ACM, 2007.

[14] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Achieving guaranteed anonymity in gps traces via uncertainty-aware path cloaking. *IEEE Transactions on Mobile Computing*, 9(8):1089–1107, 2010.

[15] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, and P.-P. DeWolf, *Handbook on Statistical Disclosure Control*, CENEX SDC, March 2007. `http://neon.vb.cbs.nl/casc/handbook.htm`

[16] T. W. Liao. Clustering of time series data - a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[17] A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel. Movement data anonymity through generalization. *Transactions on Data Privacy*, 3(2):91–121, 2010.

[18] M. E. Nergiz, M. Atzori, and Y. Saygin. Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS, SPRINGL 2008*, Irvine, California, USA, 4 November 2008, pages 52–61. ACM, 2008.

[19] M. E. Nergiz, M. Atzori, Y. Saygin, and B. Guc. Towards trajectory anonymization: a generalization-based approach. *Transactions on Data Privacy*, 2(1):47–75, 2009.

[20] R. Shonkwiler. Computing the Hausdorff set distance in linear time for any l(p) point distance. *Information Processing Letters*, 38(4):201–207, 1991.

[21] M. Terrovitis and N. Mamoulis. Privacy preservation in the publication of trajectories. In *Proceedings of the 9th International Conference on Mobile Data Management, MDM 2008*, Beijing, China, 27-30 April 2008, pages 65–72. IEEE, 2008.

[22] V. Torra, J. Abowd, and J. Domingo-Ferrer. Using Mahalanobis distance-based record linkage for disclosure risk assessment. In *Privacy in Statistical Databases-PSD 2006*, LNCS 4302, pp. 233-242, 2006.