Mobile phone Origin-Destination Matrix Anonymization and Analysis

Benoît Matet, Étienne Côme, Angelo Furno, Latifa Oukhellou

Mobile phone Origin-Destination Matrix Anonymization and Analysis

B. Matet[1,2]*, E. Côme*[1]*, A. Furno*[2]*. (1) Univ. Gustave Eiffel, COSYS, GRETTIA F-77447 Marne-la-Vallée, France, benoit.matet@univ-eiffel.fr. (2 Univ. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT,F-69518, Lyon, France*

Personal trajectory data are a category of personal data consisting in sequences of spatiotemporal points that represent the trajectory of users over a given timespan. It is clear that a dataset of trajectories must be anonymized in order to protect user's privacy. Data protection approaches rely mostly on k-anonymization [1], which is achieved when all users in the dataset are indistinguishable from at least k−1 other users. k-anonymization is usually attained through generalization and suppression, i.e., replacing exact values with intervals and deleting outliers. However, processing of whole trajectories with k-anonymization proves to be difficult to achieve for $k > 5$ and does not offer a truly foolproof anonymization scheme. Moreover, we argue that according to European regulation, any link inferred between two spatiotemporal points constitutes a data breach. As such, k-anonymized trajectories do not offer enough protection to be safely released in open data. We propose simplifying the problem to its extreme by considering trajectories defined only by their first and last spatiotemporal points, i.e., an Origin/Destination (OD)-matrix. Although they leave out most of the trajectory information, OD-matrices are a key element in the transport analysis framework as they can be used to understand the dynamic of transport demand, make long-term prediction, and allow for transport simulations. We achieve k-anonymization of an OD-matrix with $6 \leq k \leq 16$ by setting a quadtree structure for the spatial generalization of origins and destinations separately. The aggregation of origins consists in finding a quadtree whose set of leaves represent a spatial partition of the area of study. We formalize the problem as finding the tree that minimizes the Mean Squared Error compared to a given target outgoing volume. Each origin leaf is then associated to a second quadtree-based spatial partition for destination areas. The problem of finding destinations is formalized as finding the tree that minimizes a generalization error [2]. Resulting volumes v_{od} from an origin *o* to a destination *d* are considered anonymized if $v_{od}>k$ and are suppressed if $v_{od}\leq k$. As the objective functions for both problems are modular, we can easily find the optimal solutions, as well as the best solution under a given suppression threshold, by reducing it to a Tree Knapsack Problem. These simplifications let us find the least destructive anonymization as measured by the generalization error for high values of k, ensuring complete anonymization of the data for a reasonable computational cost. We apply our approach at scale on massive mobile network data from French telecom operator Orange. The results show that we are able to reconstruct the hourly profiles for in- and outgoing mobility flows of areas in Lyon and retrieve the land use of the city via clustering of the profiles. We are also able to retrieve typical profiles of OD-matrices with respect to the hour of the day via Latent Dirichlet Allocation (LDA).

Acknowledgements

This work is supported by the French ANR research project PROMENADE (grant number ANR-18-CE22-0008).

References

[1] M. Fiore, P. Katsikouli, E. Zavou, M. Cunche, F. Fessant, D. Le Hello, U. Matchi Aivodji, B. Olivier, T. Quertier, and R. Stanica. Privacy in trajectory micro-data publishing : a survey, 2020.

[2] Y. Liang and R. Samavi. Optimization-based k-anonymity algorithms. Computers Security, 93:101753, 2020.