A lightweight approach for origin-destination matrix anonymization

Benoît Matet, Etienne Come, Furno Angelo, Loïc Bonnetain, Latifa Oukhellou, Nour Eddin El Faouzi

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

A Lightweight Approach for Origin-Destination Matrix Anonymization

Benoît Matet[1,2], Etienne Côme[1], Angelo Furno[2], Loïc Bonnetain[2], Latifa Oukhellou[1], Nour-Eddin El Faouzi[2]

1- Univ. Gustave Eiffel, COSYS, GRETTIA
F-77447 Marne-la-Vallée, France

2- Univ. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT,
F-69518, Lyon, France

Abstract. Personal trajectory data are becoming more and more accessible and have a high value in transport planning and mobility characterisation, at the cost of a risk for user's privacy. Addressing this risk is usually computationally expensive and can lead to losing most of the data utility. We explore a new, light-weight approach to Origin/Destination-matrix anonymization that is easily scalable. We apply it to trip records from New York City Taxi and Limousine Commission (TLC) to illustrate how it can combine foolproof anonymity with a good spatial precision for a reasonable computational cost.

1 Introduction

Personal trajectory data are a category of personal data consisting in sequences of spatiotemporal points that represent the trajectory of users over a given time span. It is clear that a dataset of trajectories must be completely anonymized in order to protect user's privacy. Data protection approaches rely mostly on k-anonymization, which is achieved when all users in the data are indistinguishable from at least $k-1$ other users. k-anonymization is usually attained through *generalization and suppression* [1, 2], *i.e.,* replacing values with less specific but semantically consistent values shared with other users, and deleting outlier users, respectively. However, k-anonymization of whole trajectories proves to be difficult to achieve for $k > 5$ and struggle to offer a truly foolproof anonymization [3]. In this paper, we consider 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 [4], make long-term prediction [5] and allow for transport simulations [6]. Anonymization of OD-matrices has been explored in [7] with a uniform aggregation to achieve k-anonymization, and in [8, 9] with a tree-based spatial aggregation to achieve *differential privacy*. Differential privacy is arguably a more elaborate and rigorous anonymization framework than k-anonymization, but we will not consider it as our goal for the following reasons : *i)* as of today, differential privacy is not recognized by European regulators, who promote k-anonymizaton instead, *ii)* In our case differential privacy introduces a noise in the data that is acceptable only for high enough volumes,

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

which means the data is already k-anonymized. In this paper, we use a quadtree to achieve k-anonymization of an OD-matrix with $6 \leq k \leq 16$. Further simplifying the problem, we treat the aggregations of origins and destinations as two separate problems. The aggregation of origins gives a first spatial partition for origin areas, each origin being associated to a second spatial partition for destination areas. For each origin, finding the spatial partition for destinations that minimizes a *generalisation error* [2] reduces the anonymization problem to a tree-knapsack optimisation problem (TKP), or ordered knapsack problem [10]. Resulting volumes v_{od} from an origin o to a destination d are considered anonymized if $v_{od} > k$ and are suppressed if $v_{od} \leq k$. These simplifications let us find the least destructive anonymization as measured by the generalisation error for high values of k. The rest of this paper is organized as follows: in Section 2, we formalize the problem and propose an algorithm to solve it. In Section 3, we apply our approach to an extract of New York City Taxi and Limousine Commission (TLC) trip records and evaluate the quality of the resulting data.

2 Anonymization methodology

2.1 Problem setting

We consider a region partitioned in a uniform grid U containing $|U| = N \times N$ initial tiles. Let $6 \leq k \leq 16$ be our anonymity threshold. We perform spatial aggregation based on a quadtree, *i.e.,* a tree-like data structure where each non-leaf node n represents an area and children of n represent the four quadrants in which the area can be divided. A leaf node has 0 children and represent a unit of spatial information. Thus n has either 0 or 4 children. We consider $\mathcal{Q}$ the set of quadtrees whose root represents the complete study area and whose vertices correspond to a non-empty set of tiles in U. Note that in order to have $\mathcal{Q} \neq \emptyset$, N must be a power of 2 without loss of generality. In the remainder of this paper, for all $q \in \mathcal{Q}$ we note $\mathcal{L}(q)$ the set of the leaves of q and for all node $n \in q$ we note $|n|$ the number of tiles in U represented by n. For each $q \in \mathcal{Q}$, $\mathcal{L}(q)$ is a spatial partition of the study area. We note q^U the quadtree $q \in \mathcal{Q}$ such that $\mathcal{L}(q) = U$. Each $q \in \mathcal{Q}$ is a subtree of q^U, and q^U has exactly $1+4+16+\ldots+|U| = \frac{1}{3}(4|U|-1)$ vertices. We apply generalization on origins and destinations separately. Origin aggregation aims at finding a spatial partition $\mathcal{L}(q^{ori}) = \{o_1, \ldots, o_i, \ldots\}$ for which the outgoing volumes v_o are the closest to a target volume v_{target}. Formally, q^{ori} represents the solution to the optimization problem, defined in Eq. 1:

$$q^{ori} = \arg\min_{q \in \mathcal{Q}} \sum_{o \in \mathcal{L}(q)} (v_o - v_{target})^2 \tag{1}$$

For each origin $o \in \mathcal{L}(q^{ori})$, destination aggregation aims at finding a spatial partition $\mathcal{L}(q_o^{dest}) = \{d_1, \ldots, d_i, \ldots\}$ that minimizes the generalization error defined in Eq. 2. This error, defined for a couple (o, d), corresponds to the *individual information loss* [2], which independently penalize generalization of each

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

attribute. For OD trips, we only have two attributes namely their origin and destination and we may measure by $|o|$ and $|d|$ their spatial generalizations in number of tiles. This leads to the following loss function:

$$G(o,d) = \begin{cases} v_{od}\frac{|o|+|d|}{|U|} \text{ if } v_{od} \geq k \\ v_{od}\frac{|U|+|U|}{|U|} \text{ if } v_{od} < k. \end{cases} \tag{2}$$

Volumes $v_{od} < k$ must be suppressed, and are therefore counted as if they were aggregated at the highest level, leading to the maximal cost of 1 per attribute. The associated optimization problem is defined by:

$$q_o^{dest} = \arg\min_{q \in \mathcal{Q}} \sum_{d \in \mathcal{L}(q)} G(o,d). \tag{3}$$

Treating suppressed volumes as aggregated to the highest level leads to a disproportionate penalty and to solutions that have close to no suppression at all, resulting in coarse generalization. Trajectory data are known to consistently contain some hard-to-generalize outliers [1]. In order to allow suppression of those outliers, we set a suppression threshold S interpreted as a maximal number of suppressed trips allowed, and we apply a coefficient $0 \leq \delta < 1$ to the cost of suppression. Generalization error for a trip becomes $G_\delta(o,d)$ defined as:

$$G_\delta(o,d) = \begin{cases} G(o,d) \text{ if } v_{od} \geq k \\ \delta G(o,d) \text{ if } v_{od} < k, \end{cases} \tag{4}$$

and the corresponding optimization problem becomes:

$$\begin{aligned} q_o^{dest} = \arg\min_{q \in \mathcal{Q}} & \sum_{d \in \mathcal{L}(q)} G_\delta(o,d), \\ s.t. & \sum_{\substack{d \in \mathcal{L}(q): \\ v_{od} \leq k}} v_{od} < S. \end{aligned} \tag{5}$$

2.2 Problem solution

All objective functions considered above are *modular*, meaning that for any partition $\mathcal{L}(q)$ of U, the objective function can be expressed as $\sum_{a \in \mathcal{L}(q)} g(a)$, where $g(a)_{a \in \mathcal{L}(q)}$ are *partial costs* that are independent of each other. In the absence of constraints as in Eq. 1 and Eq. 3, modularity makes it easy to recursively compute for any node $n \in q^U$ the smallest partial cost achievable $g^*(n)$: if n has no children, then $g^*(n) = g(n)$, else $g^*(n) = \min(g(n), \sum_{c \in children(n)} g(c))$. With this naive method, we solve Eq. 1 after visiting each node of q^U once, *i.e.,* in exactly $\frac{1}{3}(4|U|-1)$ steps. Applied to the problem defined in Eq. 5 for each $o \in \mathcal{L}(q^{ori})$, this approach returns a solution $\tilde{q}_o^{dest}$ which is not guaranteed to respect the constraint. However, Eq. 5 can be recast as a type of knapsack problem. In this case, each node n of $\tilde{q}_o^{dest}$ is considered as an item with weight w being the volume that will get suppressed if n is split, and benefit b being the

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

gain in generalization error induced by splitting n. The maximum capacity for the knapsack problem is then the suppression threshold S. Then, selecting which areas to split amounts to a knapsack problem with the following variants: *i)* the weight w may be zero if splitting n does not lead to suppressing additional volumes; *ii)* the benefit b may be negative if splitting causes too much suppression; *iii)* items follow a dependency tree: we can only split an area if its parent have been split. This problem has already been explored under the name of ordered knapsack problem in [10]. As our problem is small enough (an initial grid of $|U| = 128 \times 128$ gives $\frac{1}{3}(4|U| - 1) = 21\,845$ items for the knapsack problem), we can find the exact solution with a dynamic programming approach.

2.3 Reporting of small volumes

Suppressing trips alters the total volume of the data, which alters its representativity. To keep as much information as possible, we add a post processing step for each origin o to aggregate all destinations d s.t. $v_{od} < k$ as a single destination D_o. This step is not integrated into the optimization problem as the resulting cost function would not be modular (reported volumes impact several areas of the destinations partition). Reporting volumes allows us to reach nearly 100% k-anonymization, with edge cases when $v_{oD_o} < k$. Such flows that fail to be anonymized by this post-processing step are suppressed.

2.4 Algorithm Complexity

Dynamic tree knapsack solving involves recursively merging dynamic solutions of all children of each node. As each dynamic solution has a maximum size of S and merging two solutions s_1, s_2 requires going through $s_1 \times s_2$ iterations, merging all four children of each node has a complexity $T_{merge} = \mathcal{O}(S^4)$. As a quadtree contains $\frac{1}{3}(4|U| - 1) = \mathcal{O}(|U|)$ nodes, the overall complexity of knapsack solving is $T_{TKP} = \mathcal{O}(|U|S^4)$. However, dynamic knapsack solutions are in practice never of size S as only few configurations are relevant. The empirical complexity is in fact constant with respect to the suppression threshold S (Fig.1).

3 Results and discussion

We applied our approach to a subset of TLC Trip Records data[1] between 12^{th} January, 2009 and 31^{st} April, 2009 containing 41 712 990 OD rows for a total of 77 696 280 trips, separated into 2 616 OD-matrices each matrix corresponding to a one-hour time span. Our initial grid U was set with a mesh $m_U = 200$ m for a study area of 12800×12800 m^2, which corresponds to $|U| = 64^2 = 4096$ tiles. Data aggregated to U presents only 9% of 6-anonymity overall. We solved for $k \in \{6, 11, 16\}$ with $\delta = 0.01$ and $S = 10\%$ of total volume for each matrix. For each k, v_{target} is manually set based on performance obtained over one week of data. High values for v_{target} yield large origin aggregations and small

[1] publicly available at: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

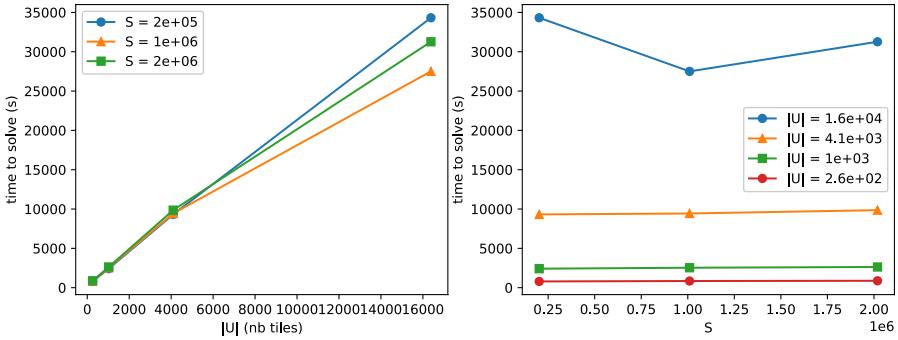

Fig. 1: Total computation time w.r.t. number of tiles in the initial uniform grid $|U|$ (left) and suppression threshold S (right) for one month of data

destinations, while low values yield small origins and large destinations. We choose v_{target} so as to minimize the difference between the mean size of origins and destinations. Solving for all matrices took approximately 16 hours on a personal computer with a 2,9 GHz Intel Core i5 and 8Go RAM. We measure the precision of the aggregated data with the Mean Area of Origins (MAO) and Mean Area of Destinations (MAD) : $MAO = \frac{\sum_{od} v_{od} \times |o| m_U^2}{\sum_{od} v_{od}}$ and $MAD = \frac{\sum_{od} v_{od} \times |d| m_U^2}{\sum_{od} v_{od}}$. MAO and MAD represent the spatial precision of the aggregated data in m^2 and are tied to the total generalization error $G_{\delta,tot}$ with :

$$G_{\delta,tot} = \sum_{o,d} G_\delta(o,d) = (\frac{\sum_{od} v_{od}}{|U|}) * (\text{MAO} + \text{MAD}) + \delta * \sum_{\substack{o,d \\ v_{od}<k}} v_{od}. \quad (6)$$

We compare our solution to a naive approach using a uniform spatial aggregation of 4×4, 8×8 and 16×16 initial tiles. MAD for those naive approaches are far greater than their spatial aggregation, as the post-processing step of reporting suppressed trips for each origin brings together several destinations as a single large destination area. Results of information loss are shown in Table 1.

4 Conclusion

We introduce a lightweight solution for general-purpose OD-matrix aggregation that loses minimal information. The proposed approach relies on assuming quadtree structure, summarizing whole trips to ODs and performing uniform temporal aggregation. We argue that such heavy hypotheses are relevant in many applications such as traffic simulation, mobility patterns analyses, or dynamic of transport demand. Solving the problem separately for spatial partitions of origins and destinations introduces a hyperparameter v_{target} that requires fine-tuning, but allows us to rapidly process huge volumes of data. This approach is especially suitable for anonymizing mobile phone operator data that are characterized by huge volumes and high sensitivity.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

k	approach	MAO (km^2)	MAD (km^2)	$G_{\delta,tot}$	% k-anon
6	our approach, $v_{target} = 400$	1.71	1.67	1.7e+06	99.97%
6	naive 4*4 agg, reported	0.64	3.60	2e+06	99.61%
6	naive 8*8 agg, reported	2.56	3.41	2.8e+06	99.90%
6	naive 16*16 agg, reported	10.24	10.35	9.8e+06	99.94%
11	our approach, $v_{target} = 700$	2.93	2.86	2.7e+06	99.93%
11	naive 4*4 agg, reported	0.64	7.83	4e+06	99.23%
11	naive 8*8 agg, reported	2.56	4.81	3.5e+06	99.78%
11	naive 16*16 agg, reported	10.24	10.60	9.9e+06	99.91%
16	our approach, $v_{target} = 1000$	3.77	3.56	3.5e+06	99.90%
16	naive 4*4 agg, reported	0.64	11.69	5.8e+06	98.86%
16	naive 8*8 agg, reported	2.56	6.22	4.2e+06	99.67%
16	naive 16*16 agg, reported	10.24	10.89	1e+07	99.88%

Table 1: Performances of k-anonymization with our approach compared to naive tile aggregation, for various k

Acknowledgement

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

References

[1] Marco Gramaglia and Marco Fiore. Hiding mobile traffic fingerprints with glove. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '15, New York, NY, USA, 2015. Association for Computing Machinery.

[2] Yuting Liang and Reza Samavi. Optimization-based k-anonymity algorithms. *Computers & Security*, 93:101753, 2020.

[3] Marco Fiore, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica. Privacy in trajectory micro-data publishing : a survey, 2020.

[4] Lijun Sun and Kay Axhausen. Understanding urban mobility patterns with a probabilistic tensor factorization framework. *Transportation Research Part B Methodological*, 91:511–524, 06 2016.

[5] Jintao Ke, Xiaoran Qin, Hai Yang, Zhengfei Zheng, Zheng Zhu, and Jieping Ye. Predicting origin-destination ride-sourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network, 10 2019.

[6] Cuauhtemoc Anda, Sergio Arturo Ordonez Medina, and Pieter Fourie. Multi-agent urban transport simulations using od matrices from mobile phone data. *Procedia Computer Science*, 130:803–809, 01 2018.

[7] Ling Yin, Qian Wang, Shih-Lung Shaw, Zhixiang Fang, Jinxing Hu, Ye Tao, and Wei Wang. Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. *PLOS ONE*, 10(10):1–23, 10 2015.

[8] Graham Cormode, Magda Procopiuc, Entong Shen, Divesh Srivastava, and Ting Yu. Differentially private spatial decompositions, 2012.

[9] Wahbeh Qardaji, Weining Yang, and Ninghui Li. Differentially private grids for geospatial data. *Proceedings - International Conference on Data Engineering*, 09 2012.

[10] David Johnson and K. Niemi. On knapsacks, partitions, and a new dynamic programming technique for trees. *Mathematics of Operations Research - MOR*, 8:1–14, 02 1983.