Differentially private count queries over personalized-location trajectory databases

Fatemeh Deldar, Mahdi Abadi *

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, P.O. Box 14115-194, Tehran, Iran

Abstract
Differential privacy is a technique for releasing statistical information about a database without revealing information about its individual data records. Also, a personalized-location trajectory database is a trajectory database where locations have different privacy protection requirements and, thus, are privacy conscious. This data article is related to the research article entitled “PLDP-TD: Personalized-location differentially private data analysis on trajectory databases” (Deldar and Abadi, 2018 [1]), in which we introduced a new differential privacy notion for personalized-location trajectory databases, and devised a novel differentially private algorithm, called PLDP-TD, to implement this new privacy notion. Here, we describe how the datasets in the research article were obtained and measure the relative error of PLDP-TD for different non-zero count query sets.

© 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Specifications table

| Subject area | Computer science |
|--------------|------------------|
| More specific subject area | Information security; Data privacy; Differential privacy |
| Type of data | Table, figure, text file |
| How data was acquired | Processing of synthetic and real trajectory data |

DOI of original article: http://dx.doi.org/10.1016/j.pmcj.2018.06.005

*Corresponding author. Fax: +98 21 82884325.
E-mail addresses: f.deldar@modares.ac.ir (F. Deldar), abadi@modares.ac.ir (M. Abadi).
Data format: Raw, analyzed

Experimental factors: The continuous spatial domain was discretized by partitioning it into a uniform grid and then each grid cell was considered as a location. Then, the latitude-longitude coordinates in each input trajectory dataset were replaced by their corresponding locations. Afterwards, a privacy descriptor was assigned to each location, resulting in a personalized-location trajectory dataset.

Experimental features: A differentially private algorithm was applied to each personalized-location trajectory dataset and the average relative error of noisy answers to non-zero count queries was measured.

Data source location: N/A

Data accessibility: Data are with this article.

Value of the data

- The obtained personalized-location trajectory datasets can be used as benchmark datasets to examine the effectiveness of various differentially private algorithms developed for trajectory databases.
- The method used for obtaining personalized-location trajectory datasets can be used to obtain personalized datasets in other contexts.
- The personal privacy budget allocation strategy can be used as a reference strategy for effective non-uniform budget allocation in any tree-based structure.
- The results showed that by providing non-uniform privacy guarantees, the quality of noisy answers to count queries significantly improves.

1. Data

Our personalized-location trajectory datasets are obtained based on three synthetic and real trajectory datasets, namely, City80K [2], Geolife [3], and Taxi [4]. The City80K dataset simulates the routes of 80,000 citizens in a metropolitan area with 26 city blocks in 24 hours. The Geolife dataset collects the GPS trajectories of 182 users in Beijing, China, during a period of over five years (from April 2007 to August 2012), including not only common activities like go home and go to work, but also some entertainments and sports activities, such as shopping, hiking, and cycling. We choose the trajectories whose latitude-longitude coordinates (moving points) are between $\frac{1}{2}39^\circ40\ ^{\prime}$ latitude and $\frac{1}{2}115^\circ8^\prime117^\circ4^\prime$ longitude, and break a trajectory if the interval time between its subsequent and current latitude-longitude coordinates is larger than one minute. Since each recorded trajectory consists of a sequence of latitude-longitude coordinates in a continuous spatial domain, we discretize the spatial domain by partitioning it into a $64 \times 64$ grid and then consider each grid cell as a location. We further replace the latitude-longitude coordinates by the obtained locations. Finally, the Taxi dataset contains approximately 5.2 million GPS trajectories of 8602 taxi cabs in Beijing, China, recorded during a 1-month period in May 2009. The trajectory data cover a region of Beijing

| Dataset  | No. of trajectories | Type of spatial domain | No. of locations | Average trajectory length | Maximum trajectory length |
|----------|---------------------|------------------------|-----------------|--------------------------|-------------------------|
| City80K  | 80,000              | Discrete               | 26              | 4.52                     | 24                      |
| Geolife  | 1,028,434           | Continuous             | 4096            | 19.52                    | 601                     |
| Taxi     | 5,251,604           | Continuous             | 256             | 11.32                    | 53,095                  |

Table 1
Summary of trajectory datasets.
restricted between [39.8, 40.1] latitude and [116.1, 116.6] longitude. Similar to the Geolife dataset, we discretize its spatial domain into a finite number of grid cells and then consider each grid cell as a location. However, due to smaller region size, we consider a $16 \times 16$ grid. Table 1 shows a summary of these trajectory datasets, providing information on the number of trajectories, the type of spatial domain, the number of locations, and the average and maximum trajectory lengths.

We randomly assign one of the five privacy descriptors, namely, Very Low, Low, Medium, High, and Critical, to each location of the trajectory datasets. We assume that the distribution of privacy descriptors of different locations in a trajectory dataset follows a Zipf distribution with shape parameter 0.50. We construct five non-zero count query sets on each trajectory dataset, each of which has different maximum query size (namely, 1, 2, 3, 4, and 5 for the City80K dataset; as well as, 2, 4, 8, 14, and 20 for the Geolife and Taxi datasets) and contains at most 10,000 randomly-generated count queries. It is important to note that by a non-zero count query set, we mean a set of count queries that inevitably exist in the original trajectory database. Figs. 1a to c show the distribution of privacy descriptors for different locations of the trajectory datasets.

2. Experimental design, materials and methods

Personalized-location differential privacy (PLDP), proposed in our research paper [1], is a new differential privacy notion for trajectory databases that relaxes privacy guarantee for trajectory data records with low privacy levels in order to improve the utility of data analysis. In fact, by introducing this concept, data owners are allowed to take the privacy protection requirements of different locations into consideration. Based on this new privacy notion, we have devised a differentially private algorithm, called PLDP-TD [1], which allocates a personal privacy budget to each trajectory data record based on the privacy protection requirements of locations it contains. PLDP-TD employs two different strategies for personal privacy budget allocation: non-adaptive and adaptive. In this data
article, we first compare the impact of these strategies on the average relative error of noisy answers to count queries.

Table 2 reports the average relative errors (in percent) of PLDP-TD with non-adaptive and adaptive strategies for different non-zero count query sets on City80K, Geolife, and Taxi under different total privacy budgets. From the table, we can observe that the average relative error of the adaptive strategy is lower than that of the non-adaptive one. This is due to the fact that the adaptive strategy tries to make effective use of privacy budgets. Therefore, we continue to experiment with this strategy. From now on, for convenience, we refer to PLDP-TD with adaptive strategy simply as PLDP-TD.

In continue, we compare the average relative error of PLDP-TD in the normal case with the case when all locations of the underlying geographical map have the same privacy descriptor “Critical”. When applying PLDP-TD to the second case, we refer to it simply as DP-TD. Figs. 2a to i compare the average relative error (in percent) of PLDP-TD with that of DP-TD for different non-zero count query sets on City80K, Geolife, and Taxi under different total privacy budgets. Obviously, we can observe
that the average relative error of PLDP-TD is significantly smaller than that of DP-TD, which is the result of relaxing privacy guarantee for trajectory data records that have low privacy levels.

Transparency document. Supplementary material

Transparency document associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2018.08.104.

References

[1] F. Deldar, M. Abadi, PLDP-TD: personalized-location differentially private data analysis on trajectory databases, Pervasive Mob. Comput. 49 (2018) 1–22. http://dx.doi.org/10.1016/j.pmcj.2018.06.005.
[2] R. Chen, B.C.M. Fung, N. Mohammed, B.C. Desai, K. Wang, Privacy-preserving trajectory data publishing by local suppression, Inf. Sci. 231 (2013) 83–97. http://dx.doi.org/10.1016/j.ins.2011.07.035.
[3] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.-Y. Ma, Understanding mobility based on GPS data, in: Proceedings of the 10th International Conference on Ubiquitous Computing, ACM, New York, NY, USA, 2008, pp. 312–321. http://dx.doi.org/10.1145/1409635.1409677.
[4] Taxi trajectory open dataset, Tsinghua University, 2009, (http://sensor.ee.tsinghua.edu.cn).