Sharing and Generating Privacy-Preserving Spatio-Temporal Data Using Real-World Knowledge

Teddy Cunningham
University of Warwick
Coventry, United Kingdom
Teddy.Cunningham@warwick.ac.uk

***Abstract*—Privacy-preserving spatio-temporal data sharing is vital in many machine learning and analysis tasks, such as managing disease spread or tailoring public services to a population's travel patterns. Current methods for data release are insufficiently accurate to provide meaningful utility, and they carry a high risk of deanonymization or membership inference attacks. These limitations and public concern over privacy and data protection has limited the extent to which data is shared. This work presents approaches generating and publishing spatio-temporal data, such as geographic locations and trajectories, with differential privacy. In the first solution, differentially private spatial data is generated using kernel density estimation and a road network-aware approach. In the second solution, a local differentially private mechanism is developed by perturbing hierarchically-structured, overlapping *n*-grams of trajectory data. Both of the solutions incorporate publicly available information, such as the road network or categories of places of interests, to enhance the utility of the output data without negatively affecting privacy or efficiency. Experiments with real-world data demonstrate that the private data can perform as well as the non-private data in a range of practical data science tasks.**

***Index Terms*—Differential Privacy, Spatio-Temporal Data, Trajectories, Privacy**

I. Introduction

People's location is collected at large scale by a wide range of organizations (e.g. *Uber* and *Google Maps*), typically through mobile technologies. Being able to analyze and model location patterns is highly valuable to other businesses and researchers (and society as a whole) to enable a vast range of location-based applications, from tracking disease spread to reducing traffic congestion. However, such data is extremely private, for numerous personal, social, and financial reasons, and the risks concerning the violation of individuals' privacy presents a major impediment to the free sharing of such data. More recently, the coronavirus pandemic and the need for high quality contact tracing has emphasized the need for privacy-aware use of personal location data.

This research focuses on developing utility-focused algorithms for spatio-temporal data generation and publication with strong privacy guarantees, which are achieved through applying differential privacy (DP). A range of publicly available external knowledge is incorporated to boost utility of the output data with no cost to privacy. This contrasts with most existing DP mechanisms, which are typically very restrictive in their use of external knowledge. The publicly available external knowledge that can be utilized ranges from geographic knowledge (e.g., locations of rivers, seas, military compounds), to location- or domain-specific information (e.g., business opening hours, sports teams schedules), and even abstract commonsense knowledge (e.g., churches are likely to be busy on Sunday mornings, but not Tuesdays at 3am). Experiments with real-world data demonstrate that including this external knowledge can improve utility noticeably.

The work in Sections III and IV has been published in [1] and [2], respectively, and will be included in the thesis. An extension to the work of Section IV, which forms ongoing research, is also briefly discussed, and it will also be included in the thesis.

II. Differential Privacy

Definition 1 (ϵ-differential privacy [3], [4]): A randomized mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if, for any two datasets D, D' differing by one element and output $y \in \mathcal{Y}$:

$$\Pr[\mathcal{M}(D) = y] \leq e^{\epsilon} \Pr[\mathcal{M}(D') = y] \tag{1}$$

where ϵ is the privacy budget; higher privacy budgets generally mean less privacy but better utility.

The Laplace mechanism releases differentially private values of numerical functions of data [4]. For a function f acting on D, it adds random noise to the value of $f(D)$ such that:

$$\mathcal{M}_f = f(D) + \mathrm{Lap}(\tfrac{\Delta_f}{\epsilon}) \tag{2}$$

where, $\mathrm{Lap}(\cdot)$ denotes the Laplace distribution, and the scale of the noise is set by the sensitivity of f, $\Delta_f = \max_{D,D'} \ |f(D) - f(D')|$. The exponential mechanism [5] is an alternative method for releasing DP output. For any dataset D and output $y \in \mathcal{Y}$, the result of mechanism $\mathcal{M}$ is ϵ-differentially private if one randomly selects y such that:

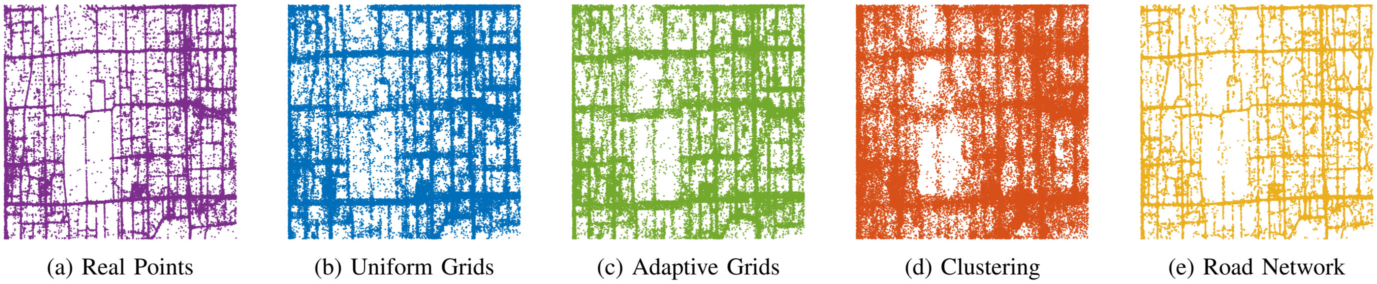
$$\Pr[\mathcal{M}(D) = y] = \frac{\exp(\epsilon q(D,y)/2\Delta_q)}{\sum_{y_i \in \mathcal{Y}} \exp(\epsilon q(D,y_i)/2\Delta_q)} \tag{3}$$

where, $q(D, y)$ is some quality function, and Δ_q is the sensitivity of the quality function (defined as for Δ_f).

Definition 2 (ϵ-local differential privacy [6]): A randomized mechanism $\mathcal{M}$ satisfies ϵ-local differential privacy if, for any two inputs x, x' and output y:

$$\Pr[\mathcal{M}(x) = y] \leq e^{\epsilon} \Pr[\mathcal{M}(x') = y] \tag{4}$$

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under Grant No. EP/L016400/1, European Research Council under Grant No. ERC-2014-CoG 647557, and by The Alan Turing Institute under EPSRC Grant No. EP/N510129/1.


(a) Real Points

(b) Uniform Grids

(c) Adaptive Grids

(d) Clustering

(e) Road Network

Fig. 1: Plots of real and synthetic data for data generation methods; picture reproduced from [1] with authors' permission

Whereas centralized DP allows the aggregator to add noise, LDP ensures that noise is added to data (typically by the user, or on their device) before it is shared with an aggregator.

Both DP and LDP have two important properties that are utilized in this work [7]. First, mechanism outputs can be manipulated without affecting the privacy guarantee. Second, an ϵ_i-(L)DP mechanism can be sequentially composed to give an overall privacy loss of $\epsilon = \sum_i \epsilon_i$. This allows the privacy budget, ϵ, to be split across the mechanism while still providing a strict upper bound on privacy leakage.

III. Private Location Data Generation

Most existing work on differentially private spatial data publication or generation (e.g., [8]–[11]) fails to output data in the same format as the input data, which limits its practical utility and the range of data analytics tasks for which it can be used. With this motivation, the section focuses on constructing two end-to-end pipelines for privately generating spatial point data. Both pipelines take an input point dataset $\mathcal{P}$ and seek to privately generate a synthetic point dataset $\mathcal{S}$ that preserves as much as of the underlying distribution of the real data as possible.

A. Partitioning-Based Approaches

This work presents three partitioning-based approaches, although other forms of partitioning can be used. Each approach utilizes a differentially private partitioning method from literature to divide the spatial domain into a set of finite regions (denoted as R_i). The first uses a uniform grid [12], the second accounts for uneven point distribution (as is common in spatial datasets) by using adaptive grids (also from [12]), and the third uses k-means-style clustering [13], [14]. Once partitioned, Laplace noise is added to the number of points in each region to get noisy counts: $\hat{n}_i = n_i + \text{Lap}(\frac{1}{\epsilon_1})$.

Synthetic data is generated using a kernel density estimation (KDE) based approach in which the kernel function is tuned to the size of the regions. Given a kernel function ϕ, the kernel density estimator, $\hat{f}(\mathbf{x})$, for a dataset of size N is:

$$\hat{f}(\mathbf{x}) = \tfrac{1}{N} \textstyle\sum_{j=1}^{N} \phi(\mathbf{x} - \mathbf{x}_j) \tag{5}$$

Given the widespread use of its one-dimensional counterpart in other DP work, this work uses the two-dimensional Laplace distribution (in its polar form) for the kernel function, defined as:

$$\phi(\mathbf{x} - \mathbf{x}_j) \equiv \phi(r_j, \theta_j) = \tfrac{\exp(-r_j/h_i)}{2\pi h_i} \tag{6}$$

where $r_j = \|\mathbf{x} - \mathbf{x}_j\|$, θ_j is the angle between $\mathbf{x}$ and $\mathbf{x}_j$, and h_i is a normalization (or smoothing) factor. To obtain a DP-compliant kernel for region R_i, one must tune ϕ for each region R_i such that the probability ratio between the two most distal points in R_i is no more than e^{ϵ}, as required by Definition 1. Hence, the smoothing parameter for R_i is set to: $h_i = \frac{\|R_i\|}{\epsilon_2}$, where $\|R_i\|$ is the maximum distance between any two locations (not necessarily in $\mathcal{P}$) in R_i. Once the KDE is constructed, $\hat{n}_i$ points are generated in each region. When doing so, public geographic knowledge from maps etc. is used to prevent points from being generated in geographic areas that would be nonsensical (e.g., seas, rivers, military compounds).

Using DP's composition theorem means that the overall privacy loss for each point is $\epsilon_1 + \epsilon_2 = \epsilon$.

B. Road Network-Aware Approach

Many datasets exhibit a degree of underlying structure and, in the case of spatial datasets, this underlying structure may be public knowledge (e.g., the road network). This underlying structure can be exploited when generating synthetic data, as will now be outlined in this three-step method, in which the road network is modeled as a graph $\mathcal{G}(\mathcal{E}, \mathcal{V})$.

First, the noisy number of points along each edge $e \in \mathcal{E}$ are obtained: $\hat{n}_e = n_e + \text{Lap}(\frac{1}{\epsilon_1})$. Micro-histograms are constructed to obtain the distribution of points along edges (both parallel and perpendicular to each edge). The number of bins for each micro-histogram is $\mathcal{O}\left(\sqrt{\epsilon \hat{n}_e}\right)$, which can be shown to minimize the total error. Laplace noise is added to each histogram bin count using ϵ_2 and ϵ_3 to control the noise, which means the overall privacy leakage for each point is $\epsilon_1 + \epsilon_2 + \epsilon_3 = \epsilon$. Finally, $\hat{n}_e$ synthetic points are generated along each edge by randomly sampling from the noisy micro-histograms.

Fig. 1 shows synthetic data samples based on Beijing taxi data. The road network approach generates synthetic data that is visually more faithful to the original data than the partitioning-based approaches, which demonstrates the benefit of incorporating publicly-available external knowledge.

IV. Trajectory Sharing with LDP

Despite generating strong techniques for doing so, generating or publishing private point data has a fundamental limitation: it fails to consider the spatio-temporal correlations that exist between consecutive points. Furthermore, the centralized setting of DP relies on a trusted aggregator, which is not always realistic or practical. Hence, this section focuses on publishing temporally-ordered sequences (i.e., trajectories) of places of interest (POIs) using LDP.

A. Challenges

The key challenge in this problem is trying to preserve existing spatio-temporal correlations between consecutive points, in addition to aiming to ensure that the error between any single real and perturbed trajectory point is minimized. This challenge is complemented by the ubiquitous aim of providing a strong privacy guarantee through a mechanism that is efficient and scalable for city-size applications.

A naïve solution to these challenges would be to perturb the entire trajectory as one entity, by modeling them as individual points in high-dimensional space. However, this approach quickly becomes computationally infeasible as the number of possible trajectories (all of which must be instantiated) is $\mathcal{O}(\frac{P^{\tau}T!}{\tau!(T-\tau)!})$, where P is the number of POIs, τ is the length of the trajectory, and T is the number of discrete timesteps at which events can occur.

B. n-Gram Solution

These challenges are addressed through a solution that uses the exponential mechanism to perturb hierarchically-structured overlapping n-grams (i.e., contiguous trajectory subsequences of length n) in accordance with LDP. Using n-grams means that the spatio-temporal relationship between adjacent points can be captured, without needing to instantiate an infeasibly large domain set. By *overlapping* the n-grams, it is possible to capture more information for each point (thereby providing utility benefits), without affecting the overall privacy guarantee. Publicly known semantic information regarding the POIs (e.g., category, opening hours) is used to construct a quality function that the exponential mechanism uses to ensure that semantically similar n-grams are more likely to be returned by the mechanism, which improves utility.

Space, time, and much of the public external knowledge that can be utilized (e.g., POI categories) have intrinsic hierarchical structures. For example, a POI is located on a street, in a suburb, and in a city—all levels of a spatial hierarchy. Similarly, 'Italian restaurant', 'restaurant', and 'food and drink' are all tags that can be associated with a pizza restaurant. This solution exploits these hierarchies by structuring the n-gram space into a hierarchy of combined space-time-category regions. This has three benefits: a) these regions preserve correlations between the three attributes, which helps to enhance utility, b) utility also benefits from a smaller domain size, and c) smaller domain sets ensure that the mechanism remains scalable for urban-size settings. The mechanism can operate at different granularities depending on other attributes, which may be public knowledge, such as the relative popularity of POIs.

The overall number of perturbations is $\tau + n - 1$, given by $\tau - n + 1$ perturbations of overlapping n-grams and $2(n-1)$ supplementary perturbations, which use shorter n-grams and are necessary to ensure points at the end of trajectories are covered n times. By assigning each perturbation a fraction of the overall privacy budget, $\epsilon' = \frac{\epsilon}{\tau+n-1}$, the overall privacy guarantee is at the user-level and is bounded by ϵ.

C. Extensions

This work can be extended to a more complex and compelling trajectory sharing problem in which multiple services wish share private data on the same set of users between each other. This problem is important given that it would allow services to learn richer patterns about their customers, while still giving users strong trajectory-level privacy guarantees. For example, if Anna purchases a phone from a shop and then buys headphones online, their bank will know the value of both transactions but not the items purchased, whereas the stores will know the items purchased at their outlets, but not items bought elsewhere. By allowing private trajectory sharing, all three services have the opportunity to learn more about events they capture (e.g., the bank can learn what item was purchases) as well as events they do not capture (e.g., each store can learn information about the other transaction). Many of the main challenges from Section IV-A apply, with the additional challenges of preserving correlations between attributes within the same event, and ensuring that events common to both services can be preserved in the shared perturbed data. Addressing this practical problem forms ongoing work.

Acknowledgments

The author thanks Hakan Ferhatosmanoglu for his supervision, as well as Graham Cormode and Divesh Srivastava, who have all contributed towards the research described in this paper.

References

[1] T. Cunningham, G. Cormode, and H. Ferhatosmanoglu, "Privacy-preserving synthetic location data in the real world," in *SSTD*, 2021.

[2] T. Cunningham, G. Cormode, H. Ferhatosmanoglu, and D. Srivastava, "Real-world trajectory sharing with local differential privacy," *PVLDB*, vol. 14, no. 11, 2021.

[3] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*. Springer, 2006.

[4] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*. Springer, 2006.

[5] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *IEEE FOCS*, 2007.

[6] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *IEEE FOCS*, 2013.

[7] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, 2014.

[8] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava, "Dpt: Differentially private trajectory synthesis using hierarchical reference systems," *PVLDB*, vol. 8, no. 11, 2015.

[9] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *ACM SIGSAC*, 2015.

[10] S. Ghane, L. Kulik, and K. Ramamohanarao, "Publishing spatial histograms under differential privacy," in *SSDBM*, 2018.

[11] M. E. Gursoy, V. Rajasekar, and L. Liu, "Utility-optimized synthesis of differentially private location traces," 2020.

[12] W. Qardaji, W. Yang, and N. Li, "Differentially private grids for geospatial data," in *IEEE ICDE*, April 2013.

[13] D. Su, J. Cao, N. Li, E. Bertino, and H. Jin, "Differentially private k-means clustering," in *ACM CODASPY*. ACM, 2016.

[14] D. Su, J. Cao, N. Li, E. Bertino, M. Lyu, and H. Jin, "Differentially private k-means clustering and a hybrid approach to private optimization," *ACM Trans. Priv. Secur.*, vol. 20, no. 4, 2017.