A Demonstration of VisDPT: Visual Exploration of Differentially Private Trajectories

Xi He, Nisarg Raval, Ashwin Machanavajjhala
Dept. of Computer Science, Duke University, USA
{hexi88, nisarg, ashwin}@cs.duke.edu

ABSTRACT

The release of detailed taxi trips has motivated numerous useful studies, but has also triggered multiple privacy attacks on individuals' trips. Despite these attacks, no tools are available for systematically analyzing the privacy risk of released trajectory data. While, recent studies have proposed mechanisms to publish synthetic mobility data with provable privacy guarantees, the questions on – 1) how to explain the theoretical privacy guarantee to non-privacy experts; and 2) how well private data preserves the properties of ground truth, remain unclear. To address these issues, we propose a system – VisDPT that provides rich visualization of sensitive information in trajectory databases and helps data curators understand the impact on utility due to privacy preserving mechanisms. We believe VisDPT will enable data curators to take informed decisions while publishing sanitized data.

1. INTRODUCTION

Billions of detailed anonymized taxi trips for big cities such as New York and Beijing, have been made publicly available. These have fueled a number of studies on human mobility patterns for a variety of applications such as traffic planning and road constructions [6, 11, 14, 15, 16]. However, privacy concerns over these datasets [7, 10, 12, 13] have motivated multiple attempts to identify passengers, or to learn their sensitive information from these datasets. For example, publication of NYC Taxi data[1], leads to the release of sensitive information such as trips of celebrities and home addresses of frequent visitors of night clubs [10]. A study on GPS data of taxis in Beijing concluded that passengers' privacy can be exposed by re-identifying more than 55% of trajectories using origin and destination queries [13].

Despite increasing privacy awareness among data curators, the translation of the above attacks to a systematic and thorough privacy evaluation tool is a nontrivial task for data curators who are not experts in privacy. Geolocation analysis tools such as GEPETO [7, 8] have been designed to enhance visual exploration of data and help evaluate inference attacks on large data. However, this exploration process requires manual, incremental, and repeated efforts until suspicious trips are identified. Moreover, privacy claims that are not supported by standard privacy evaluation metrics can be inconsistent and biased, since they are sensitive to factors such as quality and quantity of data, and background knowledge of the adversary. Recent studies on publishing synthetic mobility data [9] with provable privacy guarantees namely *differential privacy* [5] – provides an orthogonal solution to the privacy attacks. However, the theoretical privacy guarantee and the effectiveness of preserving utility of the synthetic data are not trivial to understand.

To tackle these challenges, we present VisDPT – a web based visual exploration tool that helps data curators understand the privacy risks involved in publishing their data. VisDPT also enables data curators to understand the quality of data processed by sanitizing mechanisms. Through VisDPT, we make the following contributions:

1. An end-to-end system for generating a database of synthetic trajectories that incorporates our recent work *Differentially Private Trajectories* (DPT) [9], which was presented at VLDB 2015. DPT generates synthetic trajectories that have similar aggregate properties as the original trajectories, while ensuring differential privacy.

2. A class of privacy metrics to allow comprehensive examination of privacy risks of both raw and synthetic trajectories without writing repeated low-level queries.

3. An interactive visualization of a rich set of utility queries that are commonly used in mobility data analysis, such as origin-destination queries and frequent pattern queries.

People who are not privacy experts can use VisDPT to gain insights on how and to what extent the sensitive information is protected in a differentially private output. At the same time, privacy experts can explore the impact of noise on the utility of the perturbed dataset with VisDPT. To make the experience of attendees interesting and engaging, we design our demo in the form of games. In particular, the demo involves a privacy game and a utility game. The privacy game allows attendees to perceive the amount of sensitive information protected by differential privacy via visual comparisons between a non-private output and a private output over various queries. The utility game helps

[1] http://www.andresmh.com/nyctaxitrips/

Proceedings of the VLDB Endowment, Vol. 9, No. 13

attendees learn the impact of differentially private mechanisms on aggregate properties by visually inspecting utility query results under different levels of privacy guarantee. We believe VisDPT will enhance data curators' understanding of differentially private trajectories and encourage them to release properly sanitized trajectories for studies on human mobility in the future.

2. BACKGROUND

Given a spatial domain $\mathbf{\Sigma}$, a regular trajectory t is a sequence of locations (σ_i) observed and recorded at regular time intervals, where $\sigma_i \in \mathbf{\Sigma}$ for $i = 1, 2, \ldots$. The spatial domain is usually a set of latitude-longitude coordinates. We consider a trajectory database D consisting of $|D|$ individuals. We denote the input to our system by D_{raw} which is the ground truth. This ground truth usually inherits high uniqueness of individuals even at coarser temporal and spatial resolutions [4]. This means that an adversary can easily identify Alice from an anonymized database based on few locations she visited, as these locations do not appear in other individual's trajectories. We generalize this notion of uniqueness and define a class of privacy metrics such that data curator can visually explore the potential privacy risk on the raw trajectories D_{raw} with minimal effort. Furthermore, the curator can release a modified or synthetic version of D_{raw} which has no real individuals. However, the absence of real individuals in the synthetic data does not necessarily guarantee privacy. Hence, we extend the privacy metrics of the ground truth data D_{raw} to synthetic data D_{syn} which can be useful to evaluate various privacy mechanisms.

The privacy guarantee demonstrated in our system is *differential privacy* [5], which is a de-facto standard for protection of an individual's sensitive information. The compelling guarantee allows statistical indistinguishability (governed by a privacy parameter ϵ) of the output with or without any individual's presence in the database. Formally, we define ϵ-differential privacy in the context of trajectory databases as follows. Let D_1, D_2 be two neighboring trajectory databases, i.e., D_1 and D_2 differ in only one individual's trajectory t, written as $\|D_1 - D_2\| = 1$. A randomized mechanism M satisfies ϵ-differential privacy (DP), if for any possible output O of M, for $\epsilon > 0$, $\Pr[M(D_1) = O] \leq \exp(\epsilon) \Pr[M(D_2) = O]$ holds. In other words, the adversary cannot infer whether or not the individual's trajectory is present in the database from the output distribution generated by M (the released information) with high probability, where this probability depends on the privacy parameter ϵ. Usually, larger ϵ implies more disclosure about any single individual, but better accuracy on the released information.

Despite of many appealing properties of differential privacy, its theoretical guarantee is not trivial to understand. Hence, VisDPT incorporates our recent work on *Differentially Private Trajectories* (DPT) [9] to illustrate this privacy guarantee over trajectory databases. DPT is an end-to-end framework which takes D_{raw} as an input, and outputs a database of synthetic trajectories, with ϵ-differential privacy guarantee. This framework can be summarized as follows. First, DPT builds a probabilistic hierarchical reference system model from the raw database. This model is then perturbed into an ϵ-differentially private model with optimal amount of Laplace noise. Lastly, DPT samples private synthetic trajectories from the perturbed model. Due to space constraints, we refer readers to the paper [9] for details. For this demo, VisDPT will take the synthetic databases generated by DPT to show the difference between differentially private and non-differentially private trajectory databases, and the trade-off between privacy and utility using ϵ in the context of trajectories.

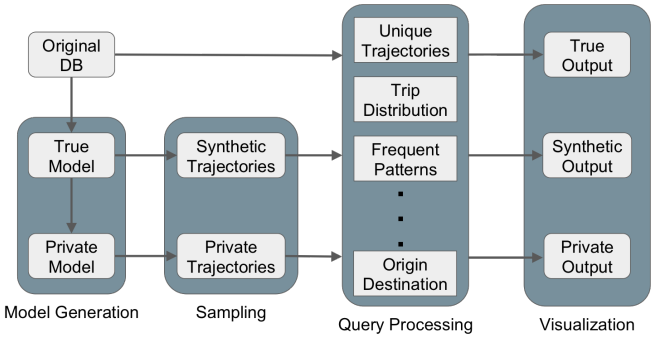

Figure 1: Outline of VisDPT framework

3. OVERVIEW OF VisDPT

Figure 1 shows the outline of VisDPT system which consists of four modules – 1) model generation, 2) sampling, 3) query processing, and 4) visualization. Given a database of taxi trips D_{raw}, VisDPT builds a probabilistic model $\mathcal{H}$ and sanitizes it to generate a private model $\tilde{\mathcal{H}}$ with a given privacy budget ϵ. The sampling phase uses $\mathcal{H}$ and $\tilde{\mathcal{H}}$ to generate synthetic trajectories D_{syn} and private trajectories $D^{\epsilon}_{\text{syn}}$ respectively. In this work, we use the mechanism proposed in our prior work DPT [9] for modeling and sampling trajectories. However, our framework can be easily extended to support multiple modeling and sampling techniques to compare their accuracy and privacy guarantee. The query processing module can execute various privacy and utility queries (e.g., finding unique trajectories, calculating trip distribution, etc.) on original as well as sampled databases. The results of these queries are presented by the visualization module which displays quantitative results (such as statistics and graphs) together with three different map views (one for each database). The map views help users visually inspect and compare the results by overlaying output trips or distribution. Next, we explain the privacy and utility queries supported in our first version of VisDPT.

3.1 Privacy Metrics

The notion of "uniqueness" has been developed as one of the privacy risk assessments on raw trajectory database [4]. The test based on this notion approximates the number of points required to uniquely identify the trajectory of an individual. However, synthetic data built from the raw database has no real individuals and hence the re-identifying test on the ground truth may not be directly applicable for the synthetic data. Therefore, we first develop a class of privacy metrics for raw trajectory database, and extend these metrics to synthetic database.

Metrics for Raw Database. The vulnerability of D_{raw} to re-identification attacks and the uniqueness of a trajectory can be well reflected by this class of privacy metrics. Each privacy metric is associated with a similarity function, denoted by $I_t(t')$ for $t, t' \in D_{\text{raw}}$ which outputs 0 if t' is

similar to t, otherwise it outputs 1. The *uniqueness of a trajectory* $t \in D_{\text{raw}}$ depends on the set of trajectories similar to t in D_{raw}, denoted by $S_t = \{t' \in D_{\text{raw}} | I_t(t') = 0\}$. We say t is unique if $|S_t| = 1$, i.e., S_t only contains t itself. We denote the set of unique trajectories in D_{raw} by $T_I(D_{\text{raw}})$. Then, the *vulnerability of a raw database* D_{raw} to re-identification attacks is defined as the fraction of unique trajectories in the raw database. Instances of this class of privacy metrics using two types of similarity functions are presented below.

Point based similarity. The point based similarity $I_t^{p,\theta}(t')$ computes the similarity between trajectories based on the similarity of locations at fixed p positions. Let $\{\sigma_1, \ldots, \sigma_p\} \subset t$ and $\{\sigma'_1, \ldots, \sigma'_p\} \subset t'$, are locations at p positions in respective trajectories. Then, $I_t^{p,\theta}(t') = 0$, if all these p pairs of points are within distance θ, i.e., $||\sigma_i - \sigma'_i||_2 < \theta$ $\forall i = 1, \ldots, p$, otherwise it is 1.

Distance based similarity. The distance based similarity $I_t^{d,\theta}(t')$ computes the similarity between trajectories using a given distance function $d(\cdot, \cdot)$. Examples of such distance functions are edit distance [3] and Fréchet distance [2]. We say $I_t^{d,\theta}(t') = 0$, if $d(t, t') < \theta$, otherwise it is 1.

Metrics for Synthetic Database. From the privacy perspective, the synthetic dataset should not have any trajectories that can be exploited for re-identification attacks. One class of such trajectories are the ones that are similar to the unique trajectories in the raw database, i.e., $T_I(D_{\text{raw}})$. The set of such trajectories in the synthetic dataset, $T_I(D_{\text{syn}})$, is computed as follows. For each $t \in T_I(D_{\text{raw}})$, we compute the set of trajectories in D_{syn} that are similar to t. We denote this set as $S_{I_t}(D_{\text{syn}}) = \{t' \in D_{\text{syn}} | I_t(t') = 0\}$. Then, $T_I(D_{\text{syn}}) = \cup_{t \in T_I(D_{\text{raw}})} S_{I_t}(D_{\text{syn}})$. It is important to know how similar these trajectories are to the unique trajectories. Hence, for all trajectories $T_I(D_{\text{syn}})$ we compute the distribution of similarity (with the corresponding unique trajectories in the raw database) using a distance function $d(\cdot, \cdot)$ mentioned above.

3.2 Utility Metrics

VisDPT presents a rich class of utility metrics for popular taxi data analysis, such as patterns [14], paths taken by taxi [15, 16] and trip distribution [6, 11].

Diameter. The diameter for a trajectory $t = (\sigma_1, \ldots, \sigma_n)$ is the maximum Euclidean distance between any pair of locations in t, i.e., $\max_{i,j} d(\sigma_i, \sigma_j)$ $\forall i, j = 1, ..., n$. $Q_d(D)$ denotes the empirical distribution of the diameter on the trajectory database D.

For rest of the queries, we map all trajectories onto a uniform grid Σ_v with resolution v and compute the distribution at a cell level, to handle the sparsity issue of databases. For every trajectory $t = (\sigma_1, \ldots, \sigma_n)$ in D, we find the mapped sequence of cells $(c_1, \ldots, c_n) \in \Sigma_v^n$, where $\sigma_i \in c_i$, i.e., location σ_i is in the cell c_i.

Origin/Destination. Given a uniform grid Σ_v, the origin distribution $Q_s(D, v)$ denotes the empirical distribution of the starting points of D over Σ_v. The destination distribution $Q_e(D, v)$ is defined similarly for the ending points of trajectories in D.

Trip/Path. Trip distribution measures the distribution of ending points of trajectories over Σ_v that are originating from a given cell $c_s \in \Sigma_v$, denoted by $Q_t(D, c_s)$. The path query $Q_p(D, c_s, c_e)$ denotes the set of trajectories starting at c_s and ending in c_e.

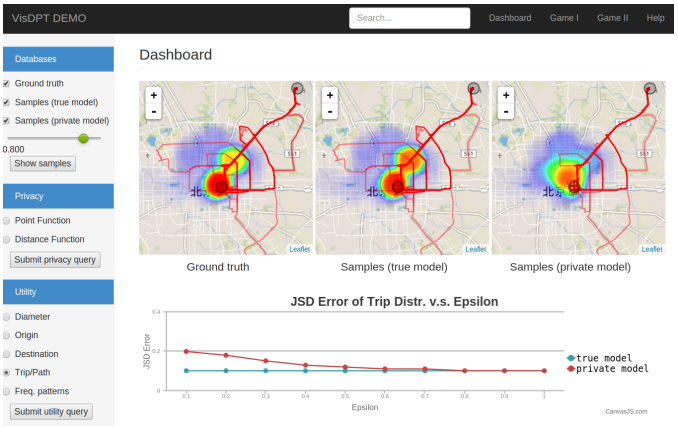

Figure 2: Snapshot of Dashboard

Frequent Patterns. Frequent patterns is a popular metric to analyze the location sequences that are visited most often. The top l k-patterns, denoted by $Q_l^k(\text{D})$ are the l most frequent k-grams (sequences of k cells) appearing across all mapped trajectories in D.

If query results are in the form of distribution, their empirical accuracy is evaluated as $JSD(Q(D_{\text{raw}}), Q(D_{\text{syn}}))$, where $JSD(\cdot, \cdot)$ is the Jensen Shannon divergence with range $[0, \ln 2]$. If query results are in the form of set such as Q_f, their accuracy is measured as $F_1(Q(D_{\text{raw}}), Q(D_{\text{syn}}))$, where $F_1(\cdot, \cdot)$ is the F_1 score with range $[0, 1]$ (i.e., harmonic mean of precision and recall) – a similarity measure between item sets.

4. DEMONSTRATION OVERVIEW

In this demo, we expect the attendees to gain insights on – 1) how and to what extent the sensitive information (such as unique trajectories) is protected in a differentially private output; and 2) the impact of noise on the utility of the perturbed dataset. The dashboard for the demo comprises of a control panel (left) and a viewing area (right) as shown in Figure 2. The control panel allows users to select samples from databases (D_{raw}, D_{syn} and $D_{\text{syn}}^{\epsilon}$) and to perform various analysis on these databases. The results are displayed in three different map windows (one for each database) in the viewing area. Each map window displays a visualization of the query result such as heat maps or trajectory paths and allows operations such as point selection. The overall quantitative results (e.g., statistics, graphs, etc.) will be displayed below for various ϵ values. The attendee can select ϵ value, to change the view for the third window with the corresponding database ($D_{\text{syn}}^{\epsilon}$).

For the purpose of this demo, we use a dataset of approximately 4.3 million taxi trips recorded by 8602 taxi cabs in Beijing, China, during May 2009 [1]. The trajectories cover the region of Beijing within the bounding box (39.788N, 116.148W) and (40.093N, 116.612W) – approximately 34 km × 40 km. The raw sampling rate of these trajectories ranges from 30 seconds to 5 minutes. In the interest of time, we precompute all synthetic databases and queries for the demo. The demo comprises of two tasks designed as games that can be played by the attendees. First, we explain them the aim of the game followed by a quick demonstration of how to play (and win) the game. Then,

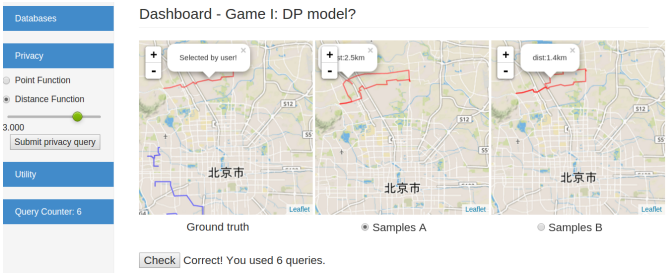

Figure 3: Privacy Game: guess DP model

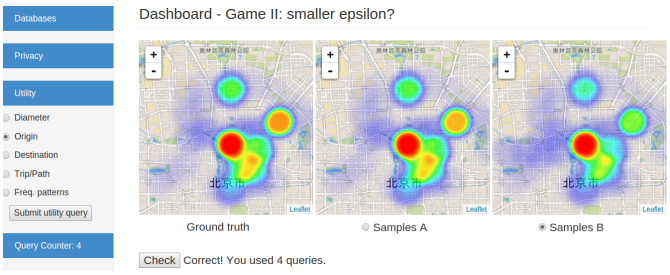

Figure 4: Utility Game: guess ϵ

the attendees are given a chance to explore the system via these games as described below.

Privacy Game. This game aims to differentiate private results from non-private results based on the existence of unique trajectories. The demonstrators will first show how to explore unique trajectories in the ground truth D_{raw} by selecting one of the privacy metrics and adjusting the corresponding threshold θ for its similarity function. Corresponding unique trajectories will be displayed on the first window. Next, demonstrator will select one of the displayed unique trajectories by clicking on it. The other two windows will show synthetic trajectories in D_{syn} and $D^{\epsilon}_{\text{syn}}$ (fixed ϵ) which are close to the selected trajectory (i.e., having distance $< \theta$) and their corresponding similarity (distance) values. The attendee will observe that the differentially private model is unlikely to output trajectories similar to any unique trajectories in the ground truth, while non-perturbed output preserves unique trajectories.

After this learning, the attendee will be shown three windows (Figure 3), where the first window displays trajectories in D_{raw}, but the positions of the other two windows for D_{syn} and $D^{\epsilon}_{\text{syn}}$ are at random and are unknown to the attendee. His goal is to guess which window (results) are from $D^{\epsilon}_{\text{syn}}$ by querying on the three databases. If he chooses the correct window then he wins, otherwise, the window positions are shuffled and he retries. We measure the number of attempts in terms of the number of queries made by an attendee on D_{raw} which includes adjustment of θ and selection of a trajectory. One can think of this information as the background information available to the attendee to achieve the goal.

Utility Game. The aim of this game is to provide deeper understanding of the trade-off between the aggregate utility of synthetic databases and privacy budget ϵ in differential privacy. The demonstrators will select the utility query in the control panel and the results are visualized on the corresponding three windows accompanied by an error plot of queries over $D^{\epsilon}_{\text{syn}}$ w.r.t to D_{raw} for different ϵ. Sliding bar for ϵ can be adjusted to explain the relationship between query outputs and ϵ – smaller ϵ yields poor utility.

The attendee will be shown three windows in this game (Figure 4), with the first window showing results on D_{raw}. The other two windows show results on $D^{\epsilon_1}_{\text{syn}}$ and $D^{\epsilon_2}_{\text{syn}}$, with a randomly chosen ϵ_1 and ϵ_2 in the range $[0, 10]$. The attendee has to identify the window with the smaller ϵ by visually comparing the perturbed outputs with the output of the first window (ground truth). The attendee wins if he correctly identifies the window, otherwise, the window positions are reshuffled and he retries. The attempt of the attendee in terms of the number of queries executed by him will be recorded.

In summary, the demo of VisDPT will allow attendees to see the effectiveness of differential private trajectories on their privacy and utility guarantees. We hope that VisDPT will be a useful tool for data curators to make informed decisions, and hence encourage the release of properly sanitized trajectories for studies on human mobility in the future.

5. REFERENCES

[1] Taxi trajectory open dataset, Tsinghua university, China. http://sensor.ee.tsinghua.edu.cn, 2009.

[2] B. Aronov, S. Har-Peled, C. Knauer, Y. Wang, and C. Wenk. Fréchet distance for curves, revisited. ESA, 2006.

[3] L. Chen, M. T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. SIGMOD, 2005.

[4] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.*, 2013.

[5] C. Dwork. Differential privacy. In *ICALP*, 2006.

[6] N. Ferreira, J. Poco, H. Vo, J. Freire, and C. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *VCG*, 2013.

[7] S. Gambs, M.-O. Killijian, and M. del Prado Cortez. Gepeto: A geoprivacy-enhancing toolkit. WAINA, 2010.

[8] S. Gambs, M. O. Killijian, I. Moise, and M. N. del Prado Cortez. Mapreducing gepeto or towards conducting a privacy analysis on millions of mobility traces. In *IPDPSW*, May 2013.

[9] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava. Dpt: Differentially private trajectory synthesis using hierarchical reference systems. *VLDB*, 2015.

[10] neustar Research. Riding with the stars: Passenger privacy in the nyc taxicab dataset, 2014.

[11] P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz, and C. Ratti. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, 2014.

[12] T. W. Schneider. Analyzing 1.1 billion nyc taxi and uber trips, with a vengeance, 2015.

[13] P. Sui, T. Wo, Z. Wen, and X. Li. Privacy risks in publication of taxi gps data. In *HPCC/CSS/ICESS*, 2014.

[14] M. Xu, J. Wu, Y. Du, H. Wang, G. Qi, K. Hu, and Y. Xiao. Discovery of important crossroads in road network using massive taxi trajectories. *CoRR*, 2014.

[15] J. Yuan, Y. Zheng, X. Xie, and G. Sun. Driving with knowledge from the physical world. In *SIGKDD*, 2011.

[16] J. Yuan, Y. Zheng, X. Xie, and G. Sun. T-drive: Enhancing driving directions with taxi drivers' intelligence. *TKDE*, 2012.