Demonstration Paper

Tuning the Utility-Privacy Trade-Off in Trajectory Data

Maja Schneider
Leipzig University & ScaDS.AI
Dresden/Leipzig, Germany
maja.schneider@informatik.uni-leipzig.de

Jonathan Schneider
Leipzig University & ScaDS.AI
Dresden/Leipzig, Germany
js18hoju@studserv.uni-leipzig.de

Lea Löffelmann
Leipzig University & ScaDS.AI
Dresden/Leipzig, Germany
ll69xupa@studserv.uni-leipzig.de

Peter Christen
School of Computing,
Australian National University
Canberra, Australia
peter.christen@anu.edu.au

Erhard Rahm
Leipzig University & ScaDS.AI
Dresden/Leipzig, Germany
rahm@informatik.uni-leipzig.de

ABSTRACT

Trajectory data, often collected on a large scale with mobile sensors in smartphones and vehicles, are a valuable source for realizing smart city applications, or for improving the user experience in mobile apps. But such data can also leak private information, such as a person's whereabouts and their points of interest (POI). These in turn can reveal sensitive information, for example a person's age, gender, religion, or home and work address. Location privacy preserving mechanisms (LPPM) can mitigate this issue by transforming data so that private details are protected. But privacy-preservation typically comes at the cost of a loss of utility. It can be challenging to find a suitable mechanism and the right settings to satisfy privacy as well as utility. In this work, we present Privacy Tuna, an interactive open-source framework to visualize trajectory data, and intuitively estimate data utility and privacy while applying various LPPMs. Our tool makes it easy for data owners to investigate the value of their data, choose a suitable privacy-preserving mechanism and tune its parameters to achieve a good utility-privacy trade-off.

1 INTRODUCTION

Trajectory data collected through mobile sensors are a valuable resource not only in the context of building smart cities, but also for research and commercial enterprises aiming to improve their services [2]. Such data enable useful applications, such as urban planning, traffic forecasting or personalization. On the other hand, trajectory data are inherently privacy-sensitive [12]. Attacks have shown to reveal private attributes about data producers, such as their identity, gender, age, or religious affiliation [4, 22]. In particular, a person's points of interest (POI), for example their home and work locations, can be exploited by an adversary to obtain such private information.

Trajectory data therefore need to be protected before they can be published or shared with non-trusted parties. Location privacy preserving mechanisms (LPPM) [13] transform location data in such a way that sensitive attributes can no longer be derived. This is usually achieved by perturbing or masking location data, which however can result in a decrease of the utility of such data. Thus, an LPPM needs to be designed in a way that both privacy is protected and the usefulness of the data is preserved. Since privacy and utility typically work against each other, this dilemma is also called the utility-privacy trade-off.

As a data owner it is therefore essential to not only find the right LPPM, suitable for the respective data and privacy requirements, but also to tune its parameters such that a good utility-privacy trade-off is achieved. The objective is to sufficiently protect any private information of the data producer but at the same time not to impair the data utility too much for the desired application. Finding such a trade-off is not trivial and requires experimental investigations because a suitable solution can depend on the actual data set. Furthermore, an LPPM's privacy parameters and their actual effect on the respective data can be difficult to understand for data owners who are not privacy experts.

To get a better understanding of the privacy risk that is present in their location data, of how well the private information is protected by a certain LPPM, and to what extent data utility is affected, it is helpful to support data owners by summarizing and visualizing such information so that finding a suitable LPPM and tuning its parameters becomes easier and more intuitive.

Example. A logistics service provider (LSP) equips its delivery vehicles with mobile sensors to regularly record temperature as well as geographic location. Once a sufficiently large data set has been created, the LSP wants to sell this data to interested parties to generate additional income, for example on a data trading platform.

Now, if the LSP sells their raw data, it may happen that this data is acquired by a competitor who is able to uncover the LSP's customers from the data and uses it to entice these customers away. Even if the data are bought by a non-competitor, such as a municipality wishing to analyze traffic congestion in a city, the LSP needs to trust the buyer to keep the data and therewith the customers confidential. In addition, a buyer might uncover private information about delivery drivers, for example where they live or which doctor they went to during their lunch break.

The LSP therefore needs to protect the private information in their data using a privacy-preserving mechanism. At the same time, the LSP can only sell their data if they are sufficiently accurate to enable certain applications, such as temperature or traffic modeling. Furthermore, the LSP needs to estimate the value of their data to know what price can be asked for it.

In order to achieve this goal, the LSP can use Privacy Tuna to (1) analyze the actual privacy risk for their drivers and customers who are represented in the data by choosing a POI detection attack as privacy metric, (2) select a suitable LPPM that prevents information leakage from POIs, and (3) tune the parameters of this LPPM so that POIs are reliably removed and also the utility of the data is sufficient for an application like traffic forecasting.

Contribution. We present PRIVACY TUNA[1], an interactive open-source visualization framework that enables trajectory data owners to assess the privacy risks of their data, apply LPPMs and intuitively tune the trade-off between utility and privacy by adjusting the parameters of the algorithm. It offers the following key features:

- *Data exploration*: PRIVACY TUNA offers data filters and visualizes trajectory data before and after the application of an LPPM on two adjacent maps.
- *Privacy-preservation*: Different LPPMs can be applied to protect the data. PRIVACY TUNA can be easily extended with custom algorithms.
- *Analysis of privacy and utility*: In PRIVACY TUNA, privacy and utility of trajectory data are intuitively compared and visualized in more detail on the maps. The framework can be extended with custom privacy and utility metrics.

Related Frameworks. There are several database systems capable of efficiently storing, managing, and analysing trajectory data [19]. These systems support pre-processing tasks, such as trajectory cleaning (e.g. segmentation, calibration, enrichment) and compression. Analysis tasks can be solved, such as calculating trajectory similarity, searching, joining or clustering trajectories, or classifying them. Several systems are able to visualize and explore trajectory data [3, 14, 20, 21] but do not offer features to evaluate and protect privacy.

There are two frameworks similar to ours, namely GEPETO [7] and VisDPT [8]. GEPETO is a tool for investigating geo-privacy, which offers visualization of data and application of sanitization measures and inference attacks, such as heuristics to uncover the beginning and end of a user's trip. VisDPT is a framework to generate synthetic trajectories from a probabilistic model built from ground truth and private data from a privatized model. Privacy and utility of both data sets can be evaluated with queries so that graphical and quantitative results can be compared for each data set. For both tools there is currently no maintained public version available. PRIVACY TUNA unites several features of both tools, like the side-by-side view of non-private and private data or the availability of privacy attacks. In addition, it offers a comparison of multiple privacy and utility metrics at the same time, which are normalized to be more comparable and intuitively understandable. Furthermore, additional attributes of the data, such as measurements of temperature at each data point, can be incorporated into utility evaluation and visually inspected to help understand the worth of the data.

2 PRIVACY OF TRAJECTORY DATA

Mobile devices are increasingly collecting information about their users' location, which can violate their privacy. Points of interest (POI) are particularly likely to reveal private information about a user, such as their home or work location, gender, age, education level, or marital status [22]. Moreover, human mobility behavior appears to be so unique that a few points on a trajectory are often sufficient to identify a person with a high degree of certainty [4], so that ultimately an identity can be linked to private POIs. In general, disclosing a person's whereabouts can be potentially dangerous for them, e.g., if they are celebrities, investigative journalists, or members of the military [11].

To address these issues, it is essential to assess the actual privacy risk for a user when revealing their location data. To this end different metrics and privacy notions were formulated.

Differential Privacy. A widely accepted standard for guaranteeing privacy is *Differential Privacy* (DP) [5]. Originally introduced in the context of relational databases, it provides a mathematical guarantee that the influence of a user's data onto the outcome of a query over a database is limited. The level of privacy is thereby controlled by a privacy budget ϵ, which is spent to a certain degree with each query. DP can be achieved by adding random noise to the data, for example drawn from a Laplace distribution [6].

While in relational databases DP hides to a certain degree the presence of a user in that database, in location privacy it needs to hide a user's location. *Geo-Indistinguishability* (Geo-I) [1] states that a user's perturbed location is equally likely as any other within a certain radius around the user's true location. A simple LPPM for obtaining point-wise Geo-I is *Noise 2D Point* [1], where noise is added to the longitude and latitude of each individual point in a trajectory. More advanced algorithms take into account the correlation of consecutive points in a trajectory but they can also suffer from higher utility degradation [10].

Metrics of uncertainty, error and attack success. Uncertainty metrics describe how confident an adversary can be about their estimated information about a user. In the context of location privacy this can reflect an adversary's uncertainty in assigning observed locations to a user or reconstructing the actual locations. Uncertainty can be measured, for example, with *entropy* [18]. Error based metrics estimate the error in the adversary's reconstruction, measured for example with the *expectation of distance error* [9] that considers the distance between the real trajectory of a user and the adversary's estimated reconstruction.

The privacy risk can also be measured by the success of an adversary's attack, for example, when trying to infer the POIs of a user. Stop detection methods based on temporal and spatial clustering, such as *DJ Cluster* [16], are often used for this task. Another approach called *D-Tour* [17] analyzes deviations from an ideal trajectory to identify POIs. To reduce the privacy risk from such attacks, LPPMs like *Promesse* [15] use smoothing techniques that resample trajectory points to eliminate point clusters.

3 BALANCING UTILITY AND PRIVACY

PRIVACY TUNA is a framework that enables data owners to (1) understand and estimate the risks from potential privacy leakage obtained from the location information of their data, to (2) select a suitable privacy algorithm that prevents such information leakage, to (3) measure the utility of their data to understand its value, and to (4) tune the parameters of an LPPM so that a good balance between utility and privacy is achieved.

System architecture. The PRIVACY TUNA framework consists of multiple components that communicate via a REST interface.[2] At the core of the framework sits a Flask backend holding a selection of methods for protecting privacy in trajectory data and for measuring privacy and utility, implemented in Python. It features a selection of LPPMs, such as *Noise 2D Point* and *Promesse*, and several privacy and utility metrics, such as *D-Tour* and Euclidean distance. This backend can be extended with custom methods. Data is visualized in a web application, implemented with the

[1]The source code, an interactive demonstrator, and a demonstration video are available at: https://github.com/majaschneider/privacytuna.

[2]Details of used software components and their sources are available at: https://github.com/majaschneider/privacytuna.

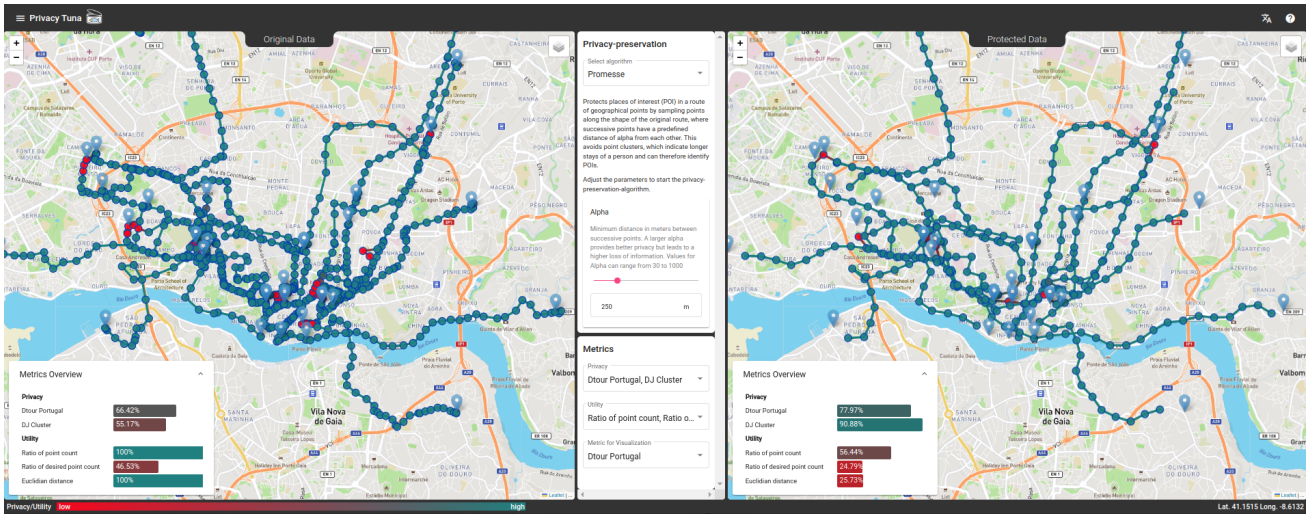

Figure 1: Tuning the utility-privacy trade-off in taxi routes from the *Porto Taxi* [17] data set with the Privacy Tuna framework. Customer pick-up and drop-off points mark points of interest (POI), shown as blue markers, which are sensitive and require special protection. Points in the map are colored according to a chosen metric, in this case by the risk of POI detection via the *D-Tour* [17] algorithm. The original data (left map) shows a higher privacy risk, indicated by red points in the map and low privacy values in the 'Metrics overview' box. After application of the *Promesse* [15] algorithm (right map) the trajectories are smoothed and their privacy is better protected, which is shown by less red points and higher privacy metric values. However, at the same time the utility metric values are degraded. By adjusting the slider we can find a good balance between privacy and utility.

Angular framework and using Leaflet for map plots. A second backend, build with SpringBoot, orchestrates the data flow between the different components. It connects to a PostgreSQL database with PostGIS extension that is used for temporal data storage. In addition, it features an optional identity and access management via KeyCloak. The user can import trajectory data in JSON format, where each data row contains a route identifier, a list of point coordinates belonging to that route, and optional measurement values for each point (for example, temperature measurements).

Scalability. When plotting large volumes of data, such as trajectory data comprising many data points, a typical challenge is the scalability of visualization. Therefore, once data is uploaded to the Privacy Tuna framework, the data is shown in the database overview table, where it can explored. To reduce the amount of data that is to be plotted, the user can filter data by time range, the route identifier or by statistical characteristics, such as the average point distance and the number of points in a trajectory. Additionally, the user has the possibility to make their data more sparse by dropping points with a certain frequency from a route or by deleting routes altogether.

Investigating the privacy of trajectory data. After the user has uploaded their trajectory data to the Privacy Tuna framework and selected the desired routes from the database overview, these routes are visualized in a map and available for exploration and further processing. Data can be explored in more detail by zooming into the areas of interest. In the menu the user can select a number of privacy and utility metrics, which are calculated for all selected routes and compared with bar plots in the 'Metrics overview' box. These metrics are normalized to a value between zero and one hundred percent, to be intuitively comparable. To explore different levels of detail for the selected metrics, data points in the maps can be colored according to their respective value of a certain metric, which can be chosen by the user. Furthermore, metric information is available on point and route level via pop-up windows.

Because the risk of disclosing POIs is particularly relevant when analyzing privacy, different POI detection attacks are available as privacy metrics, including the *D-Tour* and *DJ Cluster* algorithm. The actual POI locations are shown as blue markers on the map, as can be seen in Fig. 1. By choosing a POI detection algorithm as the basis for coloring points, the risk of detecting them can be easily matched with the true locations. This helps to get an intuitive understanding of the privacy requirements of the data. Utility and privacy metrics can easily be extended by custom metrics in the Flask backend, which is beneficial when data is required to satisfy a specific application, such as traffic forecasting.

Investigating the utility of trajectory data. To estimate the utility of the data the user can select multiple utility metrics, which are visualized next to the privacy metrics in the 'Metrics overview' box, and can be chosen as the basis for coloring data points in the maps. Certain metrics, such as Euclidean distance, compare the loss of information resulting from the application of an LPPM between original and protected trajectory points. Therefore, the utility of the original data is defined as 100% and the loss is depicted as a decrease in utility for the protected data. Other utility metrics objectively estimate the value of a data set for a certain application, such as the traffic density, for example. In such cases, the utility value is calculated independently for the original and the protected data.

While generic metrics evaluate the utility of data based on statistical analysis, for example by comparing data distributions, with application-specific metrics of utility the user can better assess what types of applications their data are suitable for. The user can thus promote their data for sale in a more targeted

manner. The utility value achieved in this process also gives a good indication of the monetary value of the data. In this context, metrics that indicate the differences of measured value distributions, such as that of temperature measurements, are particularly useful. Measurements themselves can be chosen for coloring the data in the map.

Protecting trajectory data and tuning the trade-off. After identifying the privacy risk in their data, the user can select an LPPM from the menu to protect their data. Several algorithms are available in Privacy Tuna, including *Noise 2D Point* and *Promesse*, that either protect POIs or perturb the data to achieve DP. For each algorithm a description of the mechanism and its parameters is displayed. The parameters can be set manually or with a slider. The accordingly protected data are then drawn in the second map on the right side, as can be seen in Fig. 1. The formerly selected utility and privacy metrics are applied to the protected data and shown in the 'Metrics overview' box in the right map. By adjusting the parameter values and observing the changing utility and privacy values, the algorithm can be tuned to achieve a good balance of both, if possible.

4 DEMONSTRATION

In the Privacy Tuna demonstration we will use a selection of routes from the real world trajectory data set *Porto Taxi* [17]. The data consists of cab rides, where each customer pick-up and drop-off point marks a privacy-sensitive POI. Conference attendees will take on the role of a trajectory data owner wanting to sell their data to the municipality of Porto. Attendees will be able to interact with the framework, as shown in Fig. 1, and apply all necessary steps to transform the raw data into a set of protected data. They will be able to explore the data and get to know its worth for specific applications, which they can use in their role for advertising the data and estimating a sensible selling price. We will demonstrate the following steps:

Upload and select data. We introduce the data upload functionality and demonstrate how the amount of data can be reduced, using the attribute filter and making routes more sparse by dropping certain points to increase the average distance between points. We select several routes and investigate their actual risk of leaking private POIs by choosing the *D-Tour* attack as privacy metric for coloring the data. We will demonstrate that some POIs are successfully identified, which affects the privacy of individuals when this data would be shared. Additionally, we select several utility metrics to analyze how useful the data is for traffic analyses.

Find a suitable privacy mechanism. We then select several LPPMs and investigate how suitable they are for protecting POIs, and how much they degrade utility. Attendees can observe that by using *Noise 2D Point*, which applies point-wise DP, privacy is actually getting worse and can only be mitigated with very high noise values that heavily decrease the utility. We will show that an algorithm like *Promesse*, which smoothes the route, is better suited to hide POIs but also retain a high utility.

Tune the parameters to find a good utility-privacy trade-off. We use the slider to adjust the LPPM's parameters and observe how the privacy and utility metrics change. We try to find a good balance so that both privacy and utility are high enough. Attendees, in their role as a data owner wanting to sell their data, will investigate how useful the protected data is for traffic analyses in order to estimate a reasonable selling price. This will be done by observing several utility metrics, which indicate whether enough data are available and whether the distortion is small enough for the analysis to be accurate. Finally, we use the export function to store the protected data set based on the current settings.

ACKNOWLEDGMENTS

This work is partially funded by the German Federal Ministry of Education and Research under grant DE4L 01MD19008D and ScaDS.AI Dresden/Leipzig 01IS18026B.

REFERENCES

[1] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. *Proc. ACM CCS*, 901–914.
[2] Abdelkarim Ben Ayed, Mohamed Ben Halima, and Adel M. Alimi. 2015. Big data analytics for logistics and transportation. *IEEE ICALT* (2015), 311–316.
[3] Siming Chen, Xiaoru Yuan, Zhenhuang Wang, Cong Guo, Jie Liang, Zuchao Wang, Xiaolong Luke Zhang, and Jiawan Zhang. 2016. Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data. *IEEE Transactions Vis. Comput. Graph.* 22, 1 (2016), 270–279.
[4] Yves Alexandre De Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Rep.* 3 (2013), 1–5.
[5] Cynthia Dwork. 2006. Differential privacy. In *Proc. Int. Colloq. Automata, Lang., Program.* Springer, 1–12.
[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC*. Springer, 265–284.
[7] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núnez del Prado Cortez. 2010. GEPETO : a GEoPrivacy-Enhancing TOolkit. In *AINA Workshops*. IEEE, 1071–1076.
[8] Xi He, Nisarg Raval, and Ashwin Machanavajjhala. 2016. A demonstration of VisDPT: Visual exploration of differentially private trajectories. *Proc. VLDB Endowment* 9, 13 (2016), 1489–1492.
[9] Baik Hoh and Marco Gruteser. 2005. Protecting location privacy through path confusion. In *Proceedings - First International Conference on Security and Privacy for Emerging Areas in Communications Networks, SecureComm 2005*. 194–205. https://doi.org/10.1109/ICUMT.2009.5345458
[10] Tao Jiang, Helen J. Wang, and Yih Chun Hu. 2007. Preserving location privacy in wireless LANs. *Proc. MobiSys* (2007), 246–257.
[11] Joshua Eaton. 2018. Jogging data reveals locations of sensitive military bases. https://archive.thinkprogress.org/fitness-tracker-data-military-secrets-9e71a355d418/
[12] John Krumm. 2007. Inference attacks on location tracks. In *Proc. 5th Int. Conf. Pervasive Comput.* Springer, 127–143.
[13] Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, and Yong Xiang. 2018. Location Privacy and Its Applications: A Systematic Study. *IEEE Access* 6 (2018), 17606–17624.
[14] Dongyu Liu, Di Weng, Yuhong Li, Jie Bao, Yu Zheng, Huamin Qu, and Yingcai Wu. 2017. SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations. *IEEE Transactions Vis. Comput. Graph.* 23, 1 (2017), 1–10.
[15] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. 2015. Time distortion anonymization for the publication of mobility data with high utility. *IEEE Trustcom/Big-DataSE/ISPA* 1 (2015), 539–546.
[16] Vincent Primault, Sonia Ben Mokhtar, Cedric Lauradoux, and Lionel Brunie. 2014. Differentially Private Location Privacy in Practice. In *Proc. MoST*.
[17] Maja Schneider, Lukas Gehrke, Peter Christen, and Erhard Rahm. 2022. D-TOUR: Detour-based point of interest detection in privacy-sensitive trajectories. *Proc. LNI* P-326 (2022), 219–230.
[18] C. E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 3 (1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
[19] Sheng Wang, Zhifeng Bao, J. Shane Culpepper, and Gao Cong. 2021. A Survey on Trajectory Data Management, Analytics, and Learning. *ACM Comput. Surveys* 54, 2 (2021), 1–36.
[20] Sheng Wang, Yunzhuang Shen, Zhifeng Bao, and Xiaolin Qin. 2019. Intelligent Traffic Analytics. *Proc. 12th ACM Int. Conf. Web Search and Data Mining* (2019), 778–781.
[21] Zuchao Wang, Min Lu, Xiaoru Yuan, Junping Zhang, and Huub Van De Wetering. 2013. Visual traffic jam analysis based on trajectory data. *IEEE Transactions Vis. Comput. Graph.* 19, 12 (2013), 2159–2168.
[22] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. You are where you go: Inferring demographic attributes from location check-ins. *Proc. WSDM* (2015), 295–304.