Tuning the Utility-Privacy Trade-Off in Trajectory Data

Maja Schneider*
University of Leipzig & ScaDS.AI
Dresden/Leipzig, Germany
mschneider@informatik.uni-leipzig.de

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

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

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

Erhard Rahm
University of Leipzig & 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 improving the user experience in apps. But such data can also leak private information about those who produced it. It can reveal not only a person's whereabouts but also very personal information about them, such as their points of interest (POI), which in turn can reveal a person's age, gender, religion or home and work address. Location privacy preserving mechanisms (LPPM) can mitigate this issue by transforming the 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 LPPMs. Our tool makes it easy for data owners to investigate the value of their data, choose the right mechanism and tune its parameters to achieve a good utility-privacy trade-off.

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

location privacy, utility-privacy trade-off, trajectory data, visualization

ACM Reference Format:
Maja Schneider, Jonathan Schneider, Lea Löffelmann, Peter Christen, and Erhard Rahm. 2018. Tuning the Utility-Privacy Trade-Off in Trajectory Data. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Trajectory data collected through mobile sensors, for example on smartphones or in vehicles, are valuable not only in the context of building smart cities, but also for research and commercial enterprises aiming to improve their services [2]. The data enable useful applications such as urban planning, traffic forecasting or personalization. On the other hand, trajectory data are inherently privacy-sensitive [8]. Attacks have shown to reveal private attributes about the data producers, such as their identity, gender, age, or religious affiliation [4, 17]. In particular, a person's points of interest (POI), e.g., their home and work locations, can be exploited by an adversary to obtain private information.

Thus, trajectory data need to be protected before they are published or shared with non-trusted parties. Location privacy preserving mechanisms (LPPM) transform location data in such a way that private attributes can no longer be derived. This is usually achieved by perturbing or masking the data, which on the other hand decreases its utility [9]. Thus, an LPPM needs to be designed in a way that both privacy is protected and the usefulness of the data is preserved.

As a data owner it is therefore essential to not only find the right LPPM, suitable for the respective data and privacy requirements, but to tune its parameters so that a good utility-privacy trade-off is achieved, that sufficiently protects the private information of the data producer but at the same time does not impair the data utility too much for the desired application. Finding such a trade-off is not trivial and requires experimental investigations. Furthermore, an LPPM's privacy parameters and their actual effect on the respective data can be hard to understand.

To get a better understanding of the privacy risk that is present in the data, of how well private information are protected by a certain LPPM and to what extent data utility is affected, it is helpful to support data owners by summarizing and visualizing this information so that finding a suitable algorithm 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 particulate matter levels as well as geographic location. Once a sufficiently large data set has been created, the LSP wants to sell the data to interested parties to generate additional income, for example on a data trading platform. Now, if the LSP sells the raw data, it may happen that the data is

acquired by a competitor who is able to uncover the LSP's customers from the data and uses it to entice the customers away. Even if the data are bought by a non-competitor, such as a municipality, wishing to analyze traffic congestion in the 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 the delivery drivers, for example, where they live or which doctor they went to during lunch break. The LSP therefore needs to protect the private information in the data by privacy-preserving mechanisms. At the same time, the LSP can only sell the data if they are sufficiently accurate to enable certain applications, such as traffic modeling. Also, the LSP needs to estimate the value of the data to know what price can be asked for it.

Contribution. We present *Privacy Tuna*, a visualization framework that enables trajectory data owners to assess the privacy risks of their data, apply privacy-preserving mechanisms 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 privacy-preserving mechanisms can be applied to protect the data. *Privacy Tuna* can be easily extended by custom algorithms.
- *Analysis of privacy and utility*: In *Privacy Tuna*, privacy and utility of route data are intuitively contrasted 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 researching trajectory data [14]. These systems support pre-processing tasks, such as trajectory cleaning (e.g. segmentation, calibration, enrichment) and compression. Research 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, 10, 15, 16], but to the best of our knowledge there are none that offer interactive privacy preservation while monitoring utility and privacy with different metrics.

2 PRIVACY OF TRAJECTORY DATA

Mobile devices are increasingly collecting information about the users' location, which can violate their privacy. Points of interest (POI) are particularly likely to reveal private information about users, such as their home or work location, gender, age, education level, or marital status [17]. Moreover, human mobility behavior appears to be so unique that a few points on a trajectory are sufficient to identify a person with a high degree of certainty [4], so that ultimately 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 or investigative journalists.

To address these issues, it is essential to assess the actual privacy risk for a user. To this end different metrics and privacy notions were formulated. Formal models, such as Differential Privacy (DP), give a mathematical guarantee for the user's privacy. Privacy metrics on the other hand estimate the knowledge of an adversary who gains access to the data.

Differential Privacy. A widely accepted standard for guaranteeing privacy is *Differential Privacy* (DP) [5]. Originally introduced in the context of relational data, 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 a database, in location privacy it needs to hide the presence of 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. Better algorithms take into account the correlation of consecutive points but also suffer from higher utility degradation [7].

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*. Error metrics estimate the error in the adversary's reconstruction, measured for example with *expectation of distance error*, 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* [12], are often used for this task. Another approach called *D-Tour* [13] analyzes deviations from an ideal trajectory to identify POIs. To reduce the privacy risk from such attacks, LPPMs like *Promesse* [11] use smoothing techniques that resample the 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, (2) to select a suitable privacy algorithm that prevents such information leakage, (3) to measure the utility of their data to understand its value, and (4) to tune the parameters of a privacy algorithm 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. At the core of the framework sits a Flask backend containing the business logic, which is a collection of Python methods for protecting privacy in trajectory data and measuring privacy and utility. It features a selection of LPPMs and privacy and utility metrics, which can be extended with custom methods. Data is visualized in a web application, implemented with the Angular framework and using

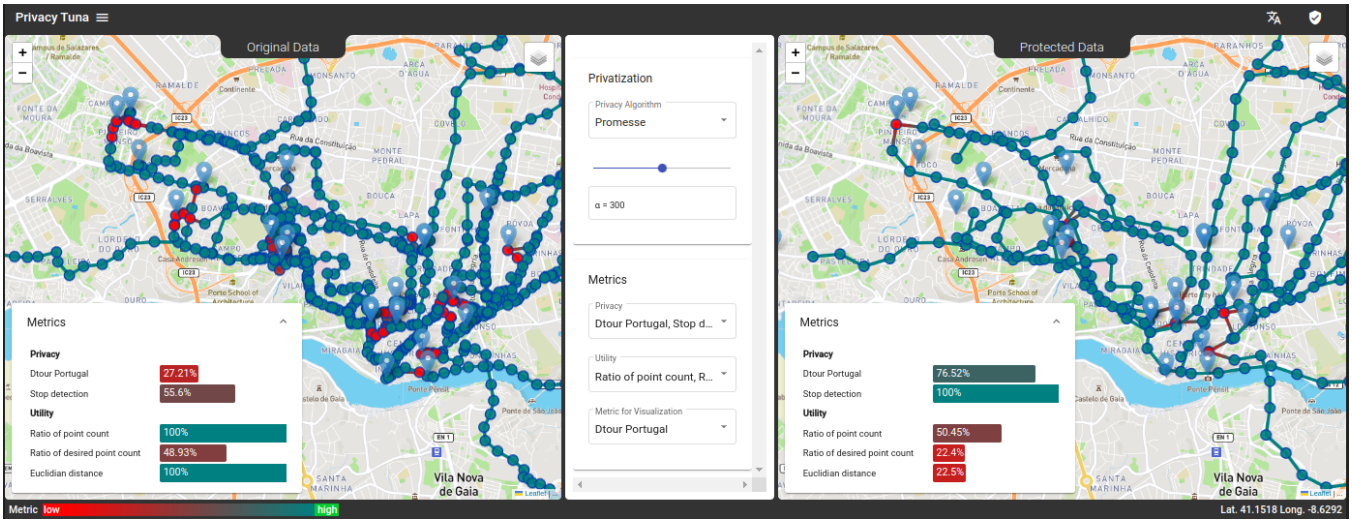

Figure 1: Tuning the utility-privacy trade-off in taxi routes from the *Porto Taxi* data set with the *Privacy Tuna* framework. Customer pick-up and drop-off points mark points of interest (POI) that are privacy-sensitive and require special protection. Points in the map are colored according to a chosen metric, in this case by the risk of POI detection via the *D-Tour* 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 privatization with the *Promesse* algorithm (right map) the trajectories are smoothed and the privacy is better protected, which shows by less red points and higher privacy metric values. At the same time though the utility metric values are degraded. By adjusting the slider we can try to find a good balance of both.

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. Users can import trajectory data in JSON format, where each data row contains a route id, a list of point coordinates belonging to that route, and optional measurements for each point (for example, particulate matter measurements).

Scalability. When plotting high-dimensional data, such as trajectory data comprising many data points, a typical challenge is the scalability of the 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, users can filter data by the time range, the route id or by statistical characteristics, such as the average point distance and the point count. Additionally, users have the possibility to make their data more sparse by dropping points with a certain frequency from a route or by deleting routes altogether. Also, to speed up plotting, only data points are rendered that lie within the current map section.

Investigating the privacy of the data. After users have uploaded the trajectory data to the *Privacy Tuna* framework and selected the desired routes from the database overview, these are visualized in a map and available for exploration and further processing. Data can be explored in more detail by zooming into the interesting areas. In the menu users can select a number of privacy and utility metrics that are calculated for all the selected routes and contrasted with bar plots in the 'Metrics overview' box. The metrics are normalized to a value between zero and one hundred percent, to be intuitively comparable. To explore different levels of detail for the 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. Also, 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. The actual POI locations are shown as markers in the map. By choosing a POI detection algorithm as basis for coloring the 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 the data. To estimate the utility of the data users can select multiple utility metrics which are visualized next to the privacy metrics in the 'Metrics overview' box, and can be chosen, too, as basis for coloring data points in the maps. Certain metrics compare the information loss of a privatization, such as the Euclidean distance between original and privatized 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. Therefore the utility value is calculated independently for the original and the privatized data.

Generic metrics evaluate the utility of data based on statistical analysis, such as comparing data distributions. In contrast, by selecting application-specific metrics of utility, the data owner can better assess what types of applications the data are suitable for, and thus promote the 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 distance of distributions between measurement values of the data 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 the data, users can select an LPPM from the menu to protect the data. Several algorithms are available 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. 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 with the slider and observing the 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 [1] we will use a selection of routes from the real world trajectory data set *Porto Taxi* [13]. 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 and apply all necessary steps to transform the raw data into a set of protected data. Attendees will 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.

Investigate privacy. 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 see that some POIs are successfully identified, which affects the privacy of individuals when data is shared. Additionally, we choose 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, the 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.

[1] A demonstration video is available at https://github.com/majaschneider/privacytuna.

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 the 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, that indicate whether enough data are available and whether the distortion is small enough so that the analysis is accurate. Finally, we use the export function to store the privatized data set based on the current settings.

ACKNOWLEDGMENTS

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by the Sächsische Staatsministerium für Wissenschaft Kultur und Tourismus in the program Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI

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] Tao Jiang, Helen J. Wang, and Yih Chun Hu. 2007. Preserving location privacy in wireless LANs. *Proc. MobiSys* (2007), 246–257.
[8] John Krumm. 2007. Inference attacks on location tracks. In *Proc. 5th Int. Conf. Pervasive Comput.* Springer, 127–143.
[9] 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.
[10] 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.
[11] 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.
[12] Vincent Primault, Sonia Ben Mokhtar, Cedric Lauradoux, and Lionel Brunie. 2014. Differentially Private Location Privacy in Practice. In *Proc. MoST*.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.