# Differentially Private Trajectory Data

Àlex Miranda-Pascual  Patricia Guerra-Balboa

Chair of IT-Security, KASTEL Security Research Labs, Karlsruhe Institute of Technology

Karlsruher Institut für Technologie

## Motivation

The value of an interest in trajectory data is becoming increasingly apparent. Traffic jam prediction, urban planning, route guidance, and smart cities are just a few of their many applications. However, it comes with a significant privacy risk, as trajectory data are extremely privacy-invasive.

Our research goal is to provide effective methods that allow for accurate trajectory data analysis for the mentioned applications without compromising individual privacy.

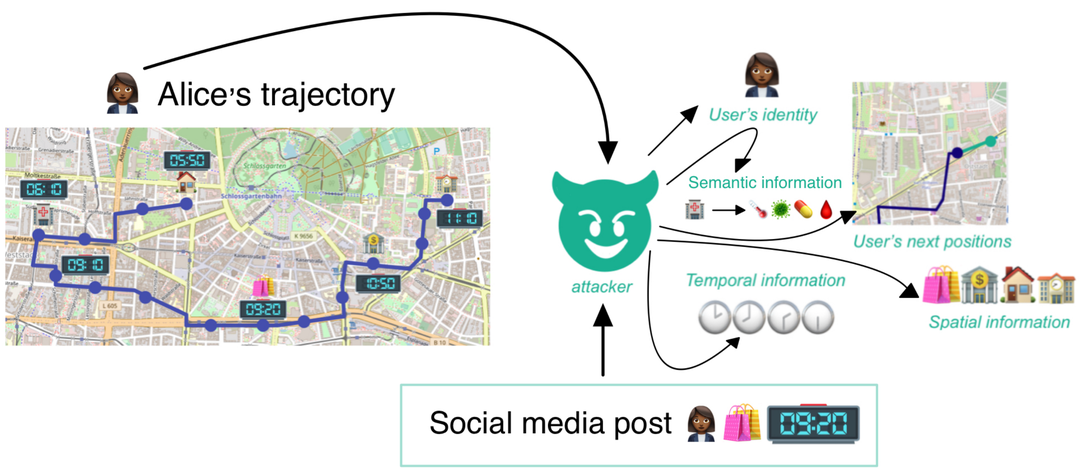



Figure 1. Trajectories can reveal precise patterns of behavior, allowing attackers to infer sensitive aspects of an individual's life, including health status, religious beliefs, social relationships, or sexual preferences.

## Trajectories Properties Affecting Privacy

High dimensional data

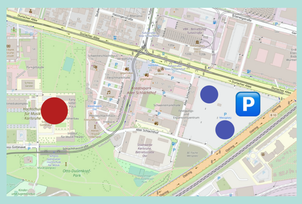
$T = p_1 \to p_2 \to p_3 \to \cdots \to p_m$

Points $p_i = (x_i, y_i, t_i, s_i)$ include:
Spatial information: $(x_i, y_i)$
Temporal information: $t_i$
Categorical information: $s_i$

Trajectory database

$$D = \begin{cases} T_1: & p_1^{(1)} & p_2^{(1)} & \cdots & p_{m_1}^{(1)} \\ T_2: & p_1^{(2)} & p_2^{(2)} & \cdots & p_{m_2}^{(2)} \\ \vdots & \vdots & \vdots & & \\ T_r: & p_1^{(r)} & p_2^{(r)} & \cdots & p_{m_r}^{(r)} \end{cases}$$

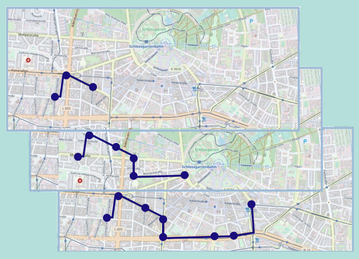
Geophysical restrictions



Semantic context of locations

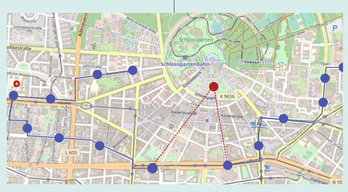
Diversity and uniqueness

Streaming scenario



Correlation

Auto-correlation | Correlation between trajectories

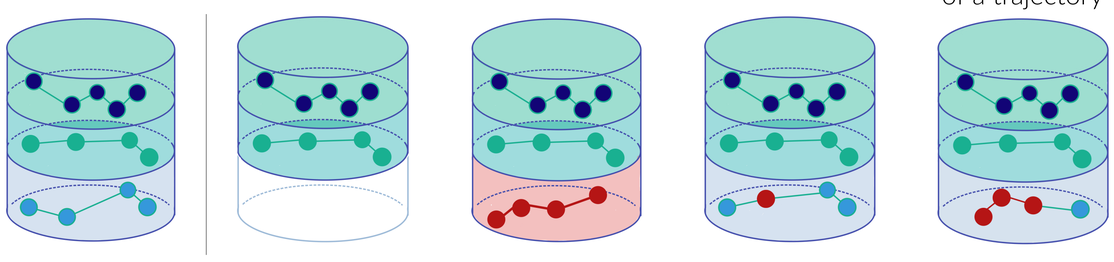


## Differential Privacy (DP)

Differential privacy is a privacy notion that bounds the effect of a single change in the database.

**Neighboring databases: How do we define a single change?**

| $\mathcal{G}$: | Unbounded | Bounded | Event-level | $\ell$-trajectory |
|---|---|---|---|---|
| *Original* | Adding/removing a trajectory | Changing a trajectory | Changing a point | Changing $\ell$ consecutive points of a trajectory |



DP aims to make $\mathcal{G}$-neighboring databases indistinguishable so that an analyst can extract statistics about the entire population, while an adversary cannot learn more than a limited amount about any user. Thus, in the case of $\mathcal{G}$ = unbounded, it aims to protect the presence of any user in the database.

### DP for any Neighborhood Definition

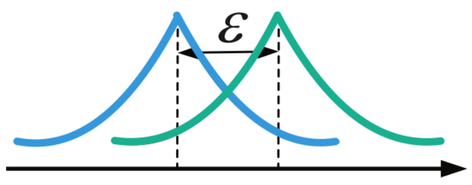
Let $\mathbb{D}$ be a database class and $\mathcal{G}$ a neighborhood definition. Then, a randomized algorithm $\mathcal{M}$ with domain $\mathbb{D}$ is $\varepsilon$-DP$_\mathcal{G}$ if for all $\mathcal{G}$-neighboring databases, $D, D' \in \mathbb{D}$, and all measurable $S \subseteq \mathrm{Range}(\mathcal{M})$,

$$\mathrm{P}\{\mathcal{M}(D) \in S\} \leq \mathrm{e}^{\varepsilon}\, \mathrm{P}\{\mathcal{M}(D') \in S\}.$$

### Sensitivity

Let $f\colon \mathbb{D} \longrightarrow \mathbb{D}'$ be a deterministic map. We define the sensitivity of $f$ with respect to $\mathcal{G}$ and $\mathcal{G}'$ as

$$\Delta f = \max_{\substack{D, D' \in \mathbb{D} \\ \mathcal{G}\text{-neighb.}}} \mathrm{dist}_{\mathcal{G}'}(f(D), f(D')).$$




Distance $\mathrm{dist}_\mathcal{G}(D, D')$ is the minimum number of $\mathcal{G}$-neighboring databases between $D$ and $D'$.

### Independent Composition Theorem

For all $i \in [k]$, let $\mathcal{M}_i$ with domain $\mathbb{D}_i$ be mutually independent $\varepsilon_i$-DP$_{\mathcal{G}_i}$ mechanisms, and let $f_i\colon \mathbb{D} \longrightarrow \mathbb{D}_i$ be arbitrary maps with finite sensitivity. Then, mechanism $\mathcal{M}$ with domain $\mathbb{D}$ defined such that $\mathcal{M}(D) = (\mathcal{M}_1(f_1(D)), \ldots, \mathcal{M}_k(f_k(D)))$ for all $D \in \mathbb{D}$ is $\varepsilon$-DP$_\mathcal{G}$ with

$$\varepsilon = \max_{\substack{D, D' \in \mathbb{D} \\ \mathcal{G}\text{-neighb.}}} \sum_{i=1}^{k} \varepsilon_i\, \mathrm{dist}_\mathcal{G}(f_i(D), f_i(D')) \leq \max_{\substack{D, D' \in \mathbb{D} \\ \mathcal{G}\text{-neighb.}}} \sum_{i\,:\, f_i(D) \neq f_i(D')} \varepsilon_i\, \Delta f_i.$$

## State of the Art: DP Mechanisms for Trajectory Data

| Masking | | | | Synthetic generation | |
|---|---|---|---|---|---|
| Local DP | Global DP | | | | |
| Perturbation of semantic trajectories | Noisy counts | Clustering | Interpolation and sampling | Traditional approaches | Machine learning approaches |
| ✓ Public knowledge<br>✓ Time & semantics<br>✗ Discrete domain<br>✗ Low/short resolution & length | ✓ Base mechanism<br>✗ Discrete domain<br>✗ Low/short resolution & length | ✓ Continuous domain<br>✗ Geospatial inconsistency<br>✗ Not DP | ✓ Continuous domain<br>✗ Approximate DP | ✓ Continuous domain<br>✗ Low resolution<br>✗ Aggregated statistics | ✓ Global distribution<br>✗ No specific DP method |

## Results and Discussion

### Main limitations

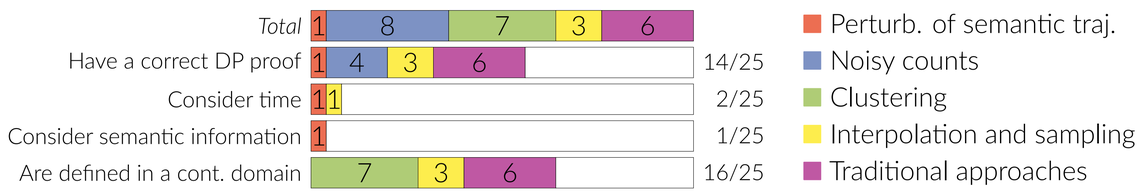
- Difficulty in defining protection mechanisms with acceptable utility guarantees
- Lack of consensus in the literature regarding evaluations and comparisons

### Other limitations

- Incorrect DP proofs.
- No masking mechanisms in the continuous domain satisfy DP. It is difficult to bound sensitivities in a continuous set of query responses.
- Most mechanisms ignore the temporal dimension, leaving temporal data vulnerable to attacks.
- Outputs may contain physically impossible trajectories.
- Difficult to deal with correlation, since DP is defined for independent data.

### Results in numbers

The number of mechanisms that:




## Conclusions and Current Work

- We analyzed both the theoretical and practical aspects of DP in trajectory data privacy, finding the gaps and limitations of privacy and utility of current proposals.
- We provided a systematization of knowledge of the privacy notions, utility metrics, and privacy-enhancing mechanisms for trajectory data.
- We designed and proved theoretical aspects of DP regarding composition that helps for streaming scenarios like route advice and traffic-jam prevention.

To address the limitations of the current mechanisms, we have started to explore the following ideas:

### Graph Data

Targets:

- Develop a discretization that avoids the continuous-domain problem and thus the sensitivity bounds. This makes the methods suitable for traffic-jam prediction and other use cases.
- Establish the geophysical framework (road networks) within the mechanism to avoid inconsistent/unrealistic data.
- Prevent filtering attacks by considering autocorrelation in the mechanisms.

### Suppression

Targets:

- Reduce the noise added by any DP mechanism by detecting and removing hard-to-protect locations and trajectories. It improves the overall utility with no penalty to the privacy level.
- Reduce the sensitivity bounds for the DP mechanisms.

### Composition

Targets:

- Estimate a tight privacy budget after sequential outputs of a mechanism running in streaming.
- Support high-dimension handling by slicing data and running mechanisms on parallel subsets.
- Estimate a tight privacy budget after different epochs in a machine learning DP training.

## Information on the Authors and References

Àlex Miranda-Pascual
alex.pascual@kit.edu
PhD researcher

Patricia Guerra-Balboa
patricia.balboa@kit.edu
PhD researcher

[1] À. Miranda-Pascual, P. Guerra-Balboa, J. Parra-Arnau, J. Forné, and T. Strufe, "SoK: Differentially private publication of trajectory data," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PoPETs)*, 2023.

[2] P. Guerra-Balboa, À. Miranda-Pascual, J. Parra-Arnau, and T. Strufe, "The composability properties of differential privacy for general granularity notions," *Under review in the 37th IEEE Comput. Secur. Found. Symp. (CSF)*, 2024.