A4-3

Differentially Private Data Publishing over Infinite Trajectory Streams

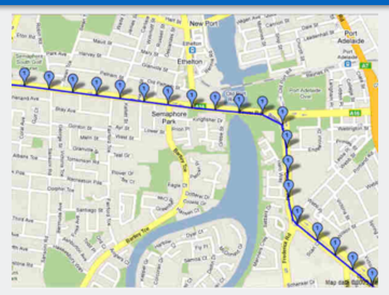
Yang CAO, Masatoshi YOSHIKAWA
Kyoto University, Department of Social Informatics

Abstract

Recently, it becomes easy to collect our own data in real life such as GPS locations.
Can we utilize personal data without privacy violation ?
In this work, we proposed

- ℓ-trajectory privacy model providing personalized privacy control, and
- an algorithm framework to achieve this privacy goal, meanwhile keeping high data utility.

Introduction



A ℓ-trajectory: ℓ successive accessed locations by a user.

	t1	t2	t3	t4	t5	…
u1	park				bar	…
u2		bar			park	…
u3	park	office	gym		bar	…

a 3-trajectory

(a) Raw Data

privacy risk

locs	t1	t2	t3	t4	t5	…
park	2	0	0	0	1	…
office	0	1	0	0	0	…
bar	0	1	0	0	2	…
gym	0	0	1	0	0	…

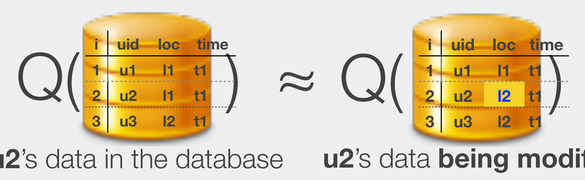
(b) Publish Real Statistics

- **Opportunity**: Data from our real-life is extremely useful for data-based innovations.
- **Risk**: **Linkage attack**[1] on anonymized dataset.
  Trajectory is sensitive: 4 data points tell us who you are[2].

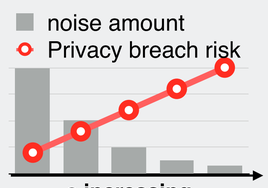
Model

ε-Differential Privacy (ε-DP) is a de facto model for Privacy Persevering Data Publishing (PPDP).
Its idea is to cover sensitive data by adding random noises to satisfy

$$\Pr[Q(D)] \le e^{\varepsilon} \cdot \Pr[Q(D^*)], \quad \varepsilon > 0$$

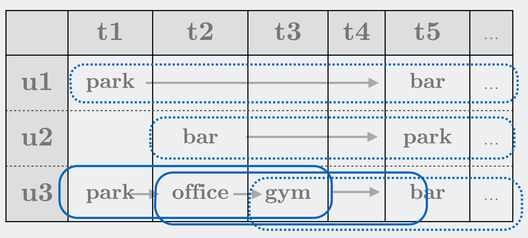


ε is Privacy Budget given by user.

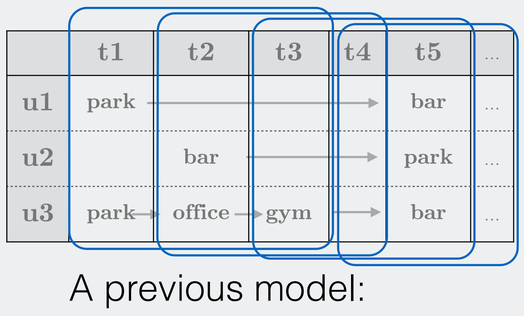


ℓ-trajectory privacy is based on ε-DP

- to limit the impact of any single ℓ-trajectory to the query result
- to make sure any ℓ-trajectory under ε-DP
- **User-level** Protection (v.s. Event-level [3])



ℓ-trajectory privacy (ℓ=3)



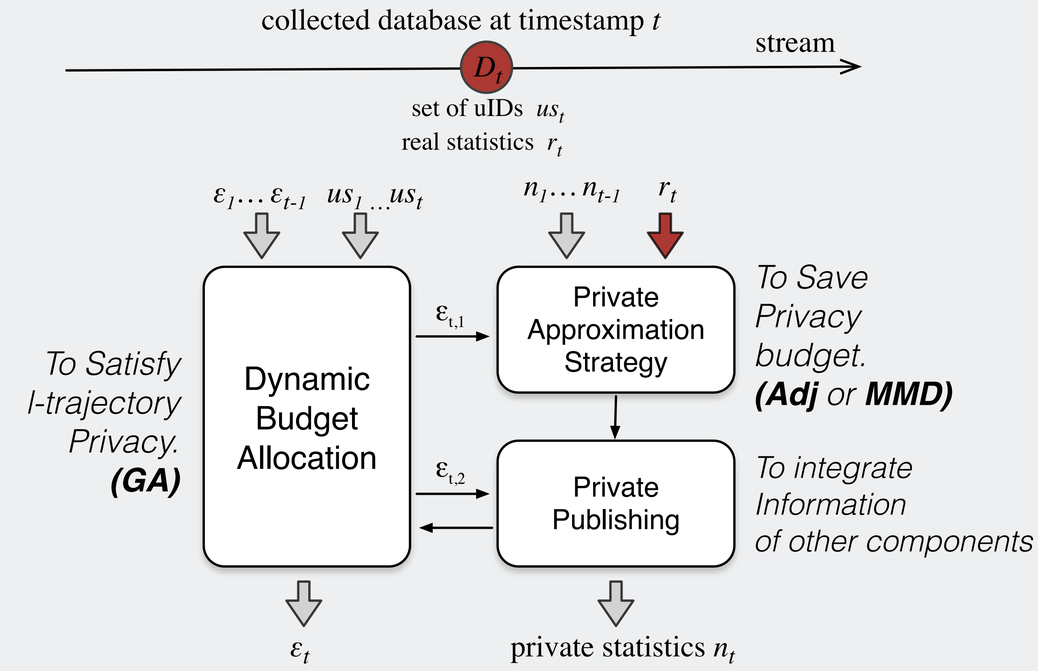
A previous model:
w-event privacy[3] (w=3)

Theorem 2 : (feasibility of ℓ-trajectory privacy) If the sum of privacy budgets on any ℓ-trajectory is less than or equal to ε, we can obtain ℓ-trajectory privacy.

Algorithms

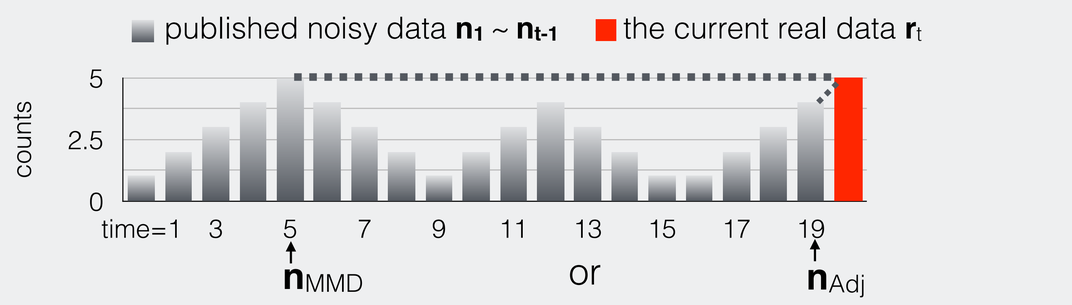
Proposed Framework of PPDP over Infinite Trajectory Streams

- For **Infinite** Streams (v.s. Finite streams)
- **Real-time** (v.s. offline publishing)



GA: A greedy algorithm to dynamically allocate privacy budget according to the data distribution.

Two different strategies based on republishing:
If it is benefit, republish —
Adj : the last noisy data, or
MMD: most-similar noisy data among $\mathbf{n_1} \sim \mathbf{n_{t-1}}$



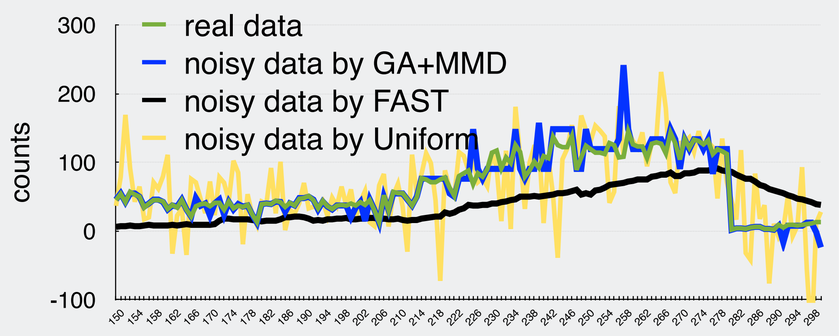
Experiments

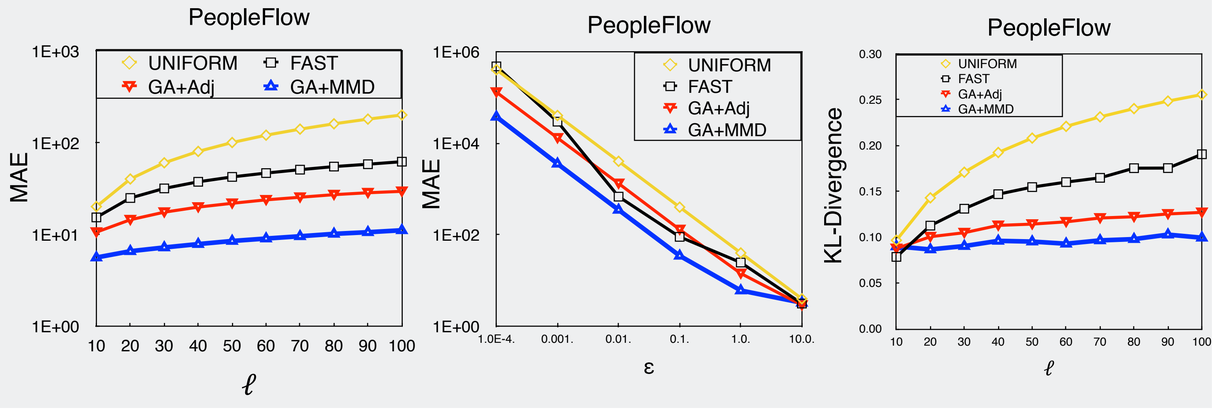
Experiments on four real-life datasets show that **GA+MMD** has relatively high data utility.
FAST[4] is a competitor PPDP framework for time-series data.
Uniform is a baseline method of uniform budget allocation to satisfy ℓ-trajectory privacy.

GA+MMD 's noisy data is closer to the real data. →

MAE:
(Mean of Absolute Error)
the lower the less noise.

KL-Divergence:
the lower the similar to original data's distribution





[1] C. Dwork, "A Firm Foundation for Private Data Analysis," Commun. ACM, Jan. 2011.
[2] Y.-A. de Montjoye et al, "Unique in the Crowd: The privacy bounds of human mobility," Sci. Rep.,Mar. 2013.
[3] G. Kellaris et al, "Differentially Private Event Sequences over Infinite Streams," VLDB'14.
[4] L. Fan et al,"FAST: Differentially Private Real-time Aggregate Monitor with Filtering and Adaptive Sampling," SIGMOD'13.