Differentially Private Real-time Data Release over Infinite Trajectory Streams

Kyoto University, Japan
Department of Social Informatics
Yang Cao, Masatoshi Yoshikawa
Outline

• Motivation: opportunity & privacy risk

• Problem definition and analysis

• Proposed solution

• Experiment results

• Conclusion & Future work
Motivation: Opportunity

- A great opportunity to utilize personal real-life data
  - Easy to collect Life log data, along with people's trajectory.
  - Trajectory streams are consisting of many people's trajectories

\(<uI\text{D},\text{time},\text{loc},\ldots>\)

A trajectory: time sequenced locations
Motivation: Opportunity

- A great opportunity to utilize personal real-life data
  - Easy to collect Life log data, along with people’s trajectory.
  - Trajectory streams are consisting of many people’s trajectories

- Statistics of trajectory streams are useful
  - e.g., Count: How many people at Pittsburgh station now?
  - e.g., Count: How many people at Pittsburgh station with Heart Rate>100 now?

Leverage statistics of trajectory streams to data-based innovations!

Health-aware Navigation System
Marketing Analysis
Intelligent Transportation System
Motivation: Privacy Risk

❖ Publish Statistics (of personal data) is risky

http://www.mathcs.emory.edu/~lxiong/cs573_s12/
Motivation: Privacy Risk

- Publish Statistics (of personal data) is risky

- E.g., Release the data of
  
  Q1: \( \text{COUNT(Sex=Female)} = A \).
  
  Q2: \( \text{COUNT(Sex=Female OR (Age=40 & Sex=Male & Employer="u1")}) = B \)

What if \( B = A + 1 \) ?

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Dis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>40</td>
<td>M</td>
<td>HIV</td>
</tr>
<tr>
<td>u2</td>
<td>30</td>
<td>M</td>
<td>-</td>
</tr>
</tbody>
</table>

http://www.mathcs.emory.edu/~lxiong/cs573_s12/
Motivation: Privacy Risk

❖ Publish Statistics (of personal data) is **risky**

❖ E.g., Release the data of

Q1: \( \text{COUNT(Sex=Female)} = A \).  
Q2: \( \text{COUNT(Sex=Female OR } \text{Age}=40 & \text{Sex=Male & Employer="u1") } = B \)

If \( B=A+1 \)

Q3: \( \text{COUNT(Sex=Female OR } \text{(Age}=42 & \text{Sex=Male & Employer="u1")} & \text{Diagnosis="HIV") = C} \)

\( C = 1 \) or \( 0 \)

http://www.mathcs.emory.edu/~lxiong/cs573_s12/
Motivation: Privacy Risk

- Publish Statistics (of personal data) is **risky**

- E.g., Release the data of
  - Q1: $\text{COUNT}(\text{Sex=Femal}) = A$
  - Q2: $\text{COUNT}(\text{Sex=Femal OR (Age=40 & Sex=Male & Employer=“u1”)}) = B$

If $B = A + 1$

- Q3: $\text{COUNT}(\text{Sex=Femal OR (Age=42 & Sex=Male & Employer=“u1” & Diagnosis=“HIV”)}) = C$

$C = 1$ or $0$

**Positively or negatively compromised!**

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Dis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>40</td>
<td>M</td>
<td>HIV</td>
</tr>
<tr>
<td>u2</td>
<td>30</td>
<td>M</td>
<td>-</td>
</tr>
</tbody>
</table>

http://www.mathcs.emory.edu/~lxiong/cs573_s12/
Motivation: Privacy Risk

- Publish Statistics (of personal data) is **risky**
- Linkage attack on anonymized data[1][2]

Motivation: Privacy Risk

❖ Publish Statistics (of personal data) is risky

❖ Linkage attack[1][2]

❖ Personal Trajectory data is highly sensitive!

❖ Four spatiotemporal data points can identify 95% of the individuals[3]

Goal: to publish statistics of trajectory streams by a Privacy Preserving Data Publishing (PPDP) method, for...

• open data
• utilizing untrusted cloud services
• data mining outsourcing

Our contributions

❖ A **rigorous** and **flexible** PPDP framework over infinite trajectory streams

❖ The first definition of **personalized privacy model** for spatio-temporal data

❖ Designed **algorithms** to publish *counts* in real-time.

---

**Privacy Model & PPDP algorithm**

- **real-time trajectory data**
- **users privacy preferences**
- **sensitive data**
- **noisy data**
- **published data utility**
- **Private Data**

E.g., Counts: How many people at Pittsburgh station now?

publishable!
Our contributions

❖ A rigorous and flexible PPDP framework over infinite trajectory streams

❖ The first definition of personalized privacy model for spatio-temporal data

❖ The protection is based on $\varepsilon$-Differential Privacy. — rigorousness

❖ Designed algorithms to publish counts in real-time.

❖ Personlized privacy & Published data utility is better. — flexibility

---

real-time trajectory data

users privacy preferences

Privacy Model & PPDP algorithm

Sensitive data → noisy data → Privacy Data

publishable!

E.g., Counts: How many people at Pittsburgh station now?
Outline

• Motivation: opportunity & privacy risk
• Problem definition and analysis
• Proposed solution
• Experiment result
• Conclusion & Future work
Problem Definition

- PPDP over infinite trajectory streams

Data collection process as follows:

(a) raw data

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
</tbody>
</table>

(b) trajectory representation of raw data

(c) raw statistics

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Problem Definition

- PPDP over infinite trajectory streams

- How to transform (c) to “safe” version (c’), while keeping it similar to (c) as much as possible?

(a) raw data

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

(b) trajectory representation of raw data

<table>
<thead>
<tr>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>park</td>
<td></td>
<td></td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td></td>
<td>bar</td>
<td></td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>park</td>
<td>office</td>
<td>gym</td>
<td>bar</td>
</tr>
</tbody>
</table>

(c) raw statistics

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Problem Definition

- PPDP over infinite trajectory streams

- How to transform \((\mathbf{c})\) to “safe” version \((\mathbf{c})'\), while keeping it similar to \((\mathbf{c})\) as much as possible?

- Ad-hoc methods CANNOT provide a reliable privacy guarantee
  - e.g., method of deleting the values of 1

It is **hard** to model the attacker’s **background knowledge** in this big data

\[
\begin{array}{c|ccccccc}
\text{locs} & \text{t1} & \text{t2} & \text{t3} & \text{t4} & \text{t5} & \ldots \\
\hline
\text{park} & 2 & 0 & 0 & 0 & 1 & \ldots \\
\text{office} & 0 & 1 & 0 & 0 & 0 & \ldots \\
\text{bar} & 0 & 1 & 0 & 0 & 2 & \ldots \\
\text{gym} & 0 & 0 & 1 & 0 & 0 & \ldots \\
\end{array}
\]

\((\mathbf{c})\) raw statistics

Ad-hoc PPDP algorithm

\[
\begin{array}{c|ccccccc}
\text{locs} & \text{t1} & \text{t2} & \text{t3} & \text{t4} & \text{t5} & \ldots \\
\hline
\text{park} & 2 & 0 & 0 & 0 & 0 & \ldots \\
\text{office} & 0 & 0 & 0 & 0 & 0 & \ldots \\
\text{bar} & 0 & 0 & 0 & 0 & 2 & \ldots \\
\text{gym} & 0 & 0 & 0 & 0 & 0 & \ldots \\
\end{array}
\]

\((\mathbf{c})'\) safe (?) statistics
Differential Privacy: A rigorous privacy definition


❖ As a *de facto* privacy standard for statistical data publishing.

❖ Randomized algorithm $A$ achieve ε-DP, if it satisfies

\[
\frac{\Pr[A(Q(D))]}{\Pr[A(Q(D^*))]} \leq e^{\epsilon}, \; \epsilon > 0
\]

$D^*$: database *except* any one individual’s data

Differential Privacy: A rigorous privacy definition

- **ε-Differential Privacy (ε-DP) [4]**
  - As a *de facto* privacy standard for statistical data publishing.
  - Randomized algorithm $A$ achieve ε-DP, if it satisfies
    \[
    \frac{\Pr[A(Q(D))]}{\Pr[A(Q(D^*))]} \leq e^\varepsilon, \quad \varepsilon > 0
    \]
    then $A$ achieves ε-DP; $\varepsilon$ is a given positive parameter.
  - $\varepsilon$: “Privacy budget” — unitary privacy level control.

---

Differential Privacy: A rigorous privacy definition

  - As a *de facto* privacy standard for statistical data publishing.
  - Randomized algorithm $A$ achieve ε-DP, if it satisfies
    \[ \frac{\Pr[A(Q(D))]}{\Pr[A(Q(D^*))]} \leq e^{\epsilon}, \; \epsilon > 0 \]
    then $A$ achieves ε-DP; $\epsilon$ is a given positive parameter.
  - $\epsilon$: “Privacy budget” — unitary privacy level control.
  - Robust under Linkage attack.
    - can be used as sub-procedure of $A$

---

Problem Analysis

- PPDP over infinite trajectory streams
  - How to apply $\varepsilon$-DP to PPDP of infinite trajectory streams ?
Problem Analysis

- PPDP over infinite trajectory streams

  How to apply $\varepsilon$-DP to PPDP of infinite trajectory streams?
  — depends on what we want to protect!
Problem Analysis

- PPDP over infinite trajectory streams

- How to apply $\varepsilon$-DP to PPDP of infinite trajectory streams?
  - depends on what we want to protect!
  - Two naive methods:
    1. to protect data of each one timestamp.
    2. to protect each user’s data of all timestamps.

(a) raw data

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
</tbody>
</table>

(b) trajectory representation of raw data

(1) park —> bar
(2) bar —> park
(3) park —> office —> gym —> bar

(c) raw statistics

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>
Problem Analysis

- PPDP over infinite trajectory streams

How to apply $\epsilon$-DP to PPDP of infinite trajectory streams?  
— depends on what we want to protect!

- Two naive methods:
  1. to protect data of each one timestamp.
  2. to protect each user’s data of all timestamps.

(a) raw data

(b) trajectory representation of raw data

(c) raw statistics

However, it is unrealistic for infinite trajectory streams

Problem Analysis
Problem Analysis

- PPDP over infinite trajectory streams

- How to apply $\varepsilon$-DP to PPDP of infinite trajectory streams?
  — depends on what we want to protect!
  - Two naive methods:
    (1) to protect data of each one timestamp.
    (2) to protect each user’s data of all timestamps.

Using Laplace Mechanism\textsuperscript{[4]} as sub-procedure:

(a) raw data

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
</tbody>
</table>

(b) trajectory representation of raw data

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

(c) raw statistics

Problem Analysis

- PPDP over infinite trajectory streams
  - How to apply $\varepsilon$-DP to PPDP of infinite trajectory streams? — depends on what we want to protect!
    - Two naive methods:
      1. to protect data of each one timestamp.
      2. to protect each user’s data of all timestamps.

Using Laplace Mechanism\textsuperscript{[4]} as sub-procedure:

- (c) raw statistics
- (c)' $\varepsilon$-DP statistics

Problem Analysis

- PPDP over infinite trajectory streams

- How to apply $\varepsilon$-DP to PPDP of infinite trajectory streams? — depends on what we want to protect!
  - Two naive methods:
    1. to protect data of each one timestamp.
    2. to protect each user’s data of all timestamps.

GPU Our observation:
In real-life, individuals may have different requirements on privacy.

(1) protecting any one patio-temporal data point

(2) protecting all length of trajectories
Problem Analysis

❖ PPDP over infinite trajectory streams

❖ How to apply $\varepsilon$-DP to PPDP of infinite trajectory streams?
— depends on what we want to protect!

▶ Two naive methods:
(1) to protect data of each one timestamp.
(2) to protect each user’s data of all timestamps.

⇒ Our observation:
In real-life, individuals may have different requirements on privacy.

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td></td>
<td>park</td>
<td></td>
<td></td>
<td>bar</td>
<td>...</td>
</tr>
<tr>
<td>u2</td>
<td></td>
<td>bar</td>
<td></td>
<td></td>
<td>park</td>
<td>...</td>
</tr>
<tr>
<td>u3</td>
<td></td>
<td>park</td>
<td>office</td>
<td>gym</td>
<td>bar</td>
<td>...</td>
</tr>
</tbody>
</table>

2-trajecory

3-trajecory

(our method) to protect any $\ell$-trajectory

privacy preference of user $i$ overall privacy : $\varepsilon$

$\ell$ successive spatio-temporal data points
Related work

❖ Differential Privacy on finite streams

❖ Differential Privacy on infinite streams

Related work

- **Differential Privacy on finite streams**
  1. to protect data of each one timestamp.
     - event-level privacy [8]
  2. to protect each user’s data of all timestamps.
     - user-level privacy [8]
     - FAST [7]: Laplace + Kalman filter (predict/correct the noisy data)

- **Differential Privacy on infinite streams**

![Diagram showing event-level and user-level privacy]

---


Related work

- **Differential Privacy on finite streams**
  1. To protect data of each one timestamp.
     - Event-level privacy [8]
  2. To protect each user’s data of all timestamps.
     - User-level privacy [8]
     - FAST [7]: Laplace + Kalman filter (predict/correct the noisy data)

- **Differential Privacy on infinite streams**
  - W-event privacy [6]

- ℓ-trajectory privacy (ℓ=3) (this study)

- W-event privacy (w=3) [6]

Outline

• Motivation: opportunity & privacy risk
• Problem definition and analysis
• **Proposed solution**
• Experiment result
• Conclusion & Future work
Overview of our solution

- A rigorous and flexible privacy model: $\ell$-trajectory privacy

- **PPDP algorithm** to publish *counts* in real-time.

(b) Infinite trajectories

(c) Raw statistics

(c)’ $l$-trajectory private data
Overview of our solution

- A rigorous and flexible privacy model: $\ell$-trajectory privacy
- **Challenge**: proof of how to achieve it by DP mechanism
- **PPDP algorithm** to publish *counts* in real-time.

**Overview of our solution**

- **$\ell$-trajectory privacy model**

![Diagram showing infinite trajectories and re-publish strategy]

- **(b) Infinite trajectories**
- **(c) raw statistics**
- **(c)' l-trajectory private data**
Overview of our solution

- A rigorous and flexible privacy model: \( \ell \)-trajectory privacy

  - **Challenge**: proof of how to achieve it by DP mechanism

- **PPDP algorithm** to publish *counts* in real-time.

  - **Challenge**: real-time & infinite streams — design a Greedy Algorithm (GA) to dynamically add noise at each timestamp;

---

(b) Infinite trajectories

(c) raw statistics

---

(c)' \( \ell \)-trajectory private data
Overview of our solution

- A rigorous and flexible privacy model: ℓ-trajectory privacy
  - **Challenge**: proof of how to achieve it by DP mechanism

- **PPDP algorithm** to publish *counts* in real-time.
  - **Challenge**: real-time & infinite streams — design a Greedy Algorithm (GA) to dynamically add noise at each timestamp;
  - **Challenge**: too much noise — re-publish the noisy data with Minimum Manhattan Distance (MMD) to the current data

![Diagram showing ℓ-trajectory privacy model, PPDP algorithm, and MMD re-publish strategy](image)

- (b) Infinite trajectories
- (c) raw statistics
Overview of our solution

- A rigorous and flexible privacy model: $\ell$-trajectory privacy

  - **Challenge**: proof of how to achieve it by DP mechanism

- **PPDP algorithm** to publish *counts* in real-time.

  - **Challenge**: real-time & infinite streams — design a Greedy Algorithm (GA) to dynamically add noise at each timestamp;

  - **Challenge**: to much noise — re-publish the noisy data with Minimum Manhattan Distance (MMD) to the current data.

---

(b) Infinite trajectories

(c) raw statistics

(c)' $l$-trajectory private data
How to achieve $\ell$-trajectory privacy

- We proved that how to achieve it by conventional DP mechanisms (e.g., Laplace/Exponential mechanism)

- $\varepsilon_i$ is privacy budget variables at each $t_i$

---

### Privacy budget variables:

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
</tbody>
</table>

- $\varepsilon_1$  
- $\varepsilon_2$  
- $\varepsilon_3$  
- $\varepsilon_4$  
- $\varepsilon_5$  

### Raw data:

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Trajectory representation of raw data

(a) raw data  
(b) trajectory representation of raw data  
(c) raw statistics
How to achieve $\ell$-trajectory privacy

- We proved that how to achieve it by conventional DP mechanisms (e.g., Laplace/Exponential mechanism)

- $\epsilon_i$ is privacy budget variables at each $t_i$

- the sum of $\epsilon_i$ at timestamp $i$ of any $\ell$-trajectory should be less than $\epsilon$

\begin{itemize}
  \item $\epsilon_1 \leq \epsilon_2 \leq \epsilon_3 \leq \epsilon_4 \leq \epsilon_5 \leq \cdots$
\end{itemize}

### (a) raw data

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### (b) trajectory representation of raw data

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

### (c) raw statistics

- Privacy budget variables: $\epsilon_i$ is privacy budget variables at each $t_i$. The sum of $\epsilon_i$ at timestamp $i$ of any $\ell$-trajectory should be less than $\epsilon$.

- Risk: $\epsilon_1 + \epsilon_5 \leq \epsilon$ $\epsilon_1 + \epsilon_2 + \epsilon_3 \leq \epsilon$ $\epsilon_2 + \epsilon_5 \leq \epsilon$ $\epsilon_2 + \epsilon_3 + \epsilon_5 \leq \epsilon$
How to achieve $\ell$-trajectory privacy

- We proved that how to achieve it by conventional DP mechanisms (e.g., Laplace/Exponential mechanism).

- $\varepsilon_i$ is privacy budget variable at each $t$.

- The sum of $\varepsilon_i$ at timestamp $i$ of any $\ell$-trajectory should be less than $\varepsilon$.

**Privacy budget variables**

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
</tbody>
</table>

**raw data**

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
<th>$\varepsilon 1$</th>
<th>$\varepsilon 2$</th>
<th>$\varepsilon 3$</th>
<th>$\varepsilon 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**trajectory representation**

(a) raw data

(b) trajectory representation

(c) raw statistics

If we know all data in advance:

Maximize: \[ \sum \varepsilon \]

- $\varepsilon 1 + \varepsilon 5 \leq \varepsilon$
- $\varepsilon 1 + \varepsilon 2 + \varepsilon 3 \leq \varepsilon$
- $\varepsilon 2 + \varepsilon 5 \leq \varepsilon$
- $\varepsilon 2 + \varepsilon 3 + \varepsilon 5 \leq \varepsilon$

approximate Maximum Utility
How to achieve $\ell$-trajectory privacy

- We proved that how to achieve it by conventional DP mechanisms (e.g., Laplace/Exponential mechanism).
- privacy budget variables at each $t_i$:
- the sum of $\epsilon_i$ at timestamp $t_i$ of any $\ell$-trajectory should be less than $\epsilon$

(a) raw data

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
</tbody>
</table>

(b) trajectory representation of raw data

(c) raw statistics

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>
How to achieve $\ell$-trajectory privacy

- We proved that how to achieve it by conventional DP mechanisms (e.g., Laplace/Exponential mechanism)
- Privacy budget variables $\epsilon_i$ at each $t_i$
- The sum of $\epsilon_i$ at timestamp $i$ of any $\ell$-trajectory should be less than $\epsilon$

(a) Raw data

<table>
<thead>
<tr>
<th>uID</th>
<th>tim</th>
<th>loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>u1</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t1</td>
<td>park</td>
</tr>
<tr>
<td>u2</td>
<td>t2</td>
<td>bar</td>
</tr>
<tr>
<td>u3</td>
<td>t2</td>
<td>office</td>
</tr>
<tr>
<td>u3</td>
<td>t3</td>
<td>gym</td>
</tr>
<tr>
<td>u1</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>u2</td>
<td>t5</td>
<td>park</td>
</tr>
<tr>
<td>u3</td>
<td>t5</td>
<td>bar</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

(b) Trajectory representation of raw data

(c) Raw statistics

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>
PPDP algorithm: GA

- **GA**: Greedy Algorithm to get approximately optimal $\varepsilon_i$ by incomplete information
PPDP algorithm: GA

- **GA**: Greedy Algorithm to get approximately optimal $\epsilon_i$ by incomplete information

- idea: exponential decay
  1. set $\epsilon_1=0$, then compute $\epsilon_3=w$
  2. use $w/2$ as the value of $\epsilon_3$ (reserve $w/2$ for $\epsilon_1$)

---

**Privacy budget variables:**

<table>
<thead>
<tr>
<th>locs</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>park</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>bar</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>gym</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

---

(a) raw data

(b) trajectory representation of raw data

(c) raw statistics
Republishing strategy to improve data utility of counts

- idea: in real-life data, the data has periodically repeated pattern

- re-publish **Adjacent** noisy (\(\text{Adj}\)) data (has been studied in [6])

- re-publish noisy data who is holding **Minimum Manhattan Distance** (\(\text{MMD}\)) to real data of current timestamp (using Exponential Mechanism[5])

---

PPDP algorithm: MMD

![Diagram showing published noisy data and current real data](image)

PPDP framework

- Personalized privacy control
- Proposed framework
  - dynamic budget allocation (GA)
  - Private Approximation strategy
  - Private publishing
- capable of publishing diverse statistical data over infinite trajectory streams
  - Adj / MMD is optimized for counts publishing
Outline

- Motivation: opportunity & privacy risk
- Problem definition and analysis
- Proposed solution
- Experiment result
- Future work
### Experiments

- **Four real-life trajectory datasets**

<table>
<thead>
<tr>
<th></th>
<th>PeopleFlow¹</th>
<th>Geolife²</th>
<th>T-Drive³</th>
<th>WorldCup98⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Desc.</strong></td>
<td>people moving; people moving; diverse type of mobility; sparse; Beijing</td>
<td>taxis moving; Beijing</td>
<td>webpage click streams</td>
<td></td>
</tr>
<tr>
<td><strong>Locs. Amt.</strong></td>
<td>18</td>
<td>56</td>
<td>21</td>
<td>1,000</td>
</tr>
<tr>
<td><strong>users Amt.</strong></td>
<td>11,406</td>
<td>170</td>
<td>2,698</td>
<td>550,762</td>
</tr>
<tr>
<td><strong>TimeStamps Amt.</strong></td>
<td>1,694</td>
<td>1,440</td>
<td>886</td>
<td>722</td>
</tr>
<tr>
<td><strong>Interval of TS</strong></td>
<td>5 mins</td>
<td>1 min</td>
<td>10 mins</td>
<td>1 hour</td>
</tr>
<tr>
<td><strong>Length of TS</strong></td>
<td>6 days</td>
<td>24 hours*</td>
<td>~7 days</td>
<td>~35 days</td>
</tr>
<tr>
<td><strong>Datapoints Amt.</strong></td>
<td>102,468</td>
<td>240,990</td>
<td>37,255</td>
<td>1,258,542</td>
</tr>
</tbody>
</table>

¹ [http://pflow.csis.u-tokyo.ac.jp/](http://pflow.csis.u-tokyo.ac.jp/)

*Aggregating 50,176 hours timestamps to 24 hours by omitting the date.*
## Experiments

<table>
<thead>
<tr>
<th>Budget Allocation</th>
<th>Approximation Strategy</th>
<th>Real-time</th>
<th>Utility Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>uniformly</td>
<td>X</td>
<td>O</td>
</tr>
<tr>
<td>LP</td>
<td>Approximately global optimised</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>GA+Adj</td>
<td>dynamically</td>
<td></td>
<td>O</td>
</tr>
<tr>
<td>GA+MMD</td>
<td>dynamically</td>
<td></td>
<td>O</td>
</tr>
<tr>
<td>FAST\textsubscript{fix} [7]</td>
<td>uniformly</td>
<td></td>
<td>O</td>
</tr>
</tbody>
</table>

Experiments

- Real data v.s. Noisy data ($\ell=20, \epsilon=1$)
Experiments

❖ Metrics

❖ Mean of Absolute Error (MAE)

\[
MAE(R,N) = \frac{1}{T |locs|} \sum_{i=1}^{T} \sum_{j=1}^{|locs|} |r_{i,j} - n_{i,j}|
\]

❖ Mean of Square Error (MSE)

❖ sensitivity to large error

\[
MSE(R,N) = \frac{1}{T |locs|} \sum_{i=1}^{T} \sum_{j=1}^{|locs|} (r_{i,j} - n_{i,j})^2
\]

❖ KL-divergence

❖ similarity of two distribution (the lower the more similar)

\[
D_{KL}(R \| N) = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{|locs|} \ln \left( \frac{r_{i,j}^*}{n_{i,j}^*} \right) r_{i,j}^*
\]

All of them are “the lower, the better”
Experiments

- MAE/MSE by varying $\ell$ ($\varepsilon=1$)
Experiments

- MAE/MSE by varying $\varepsilon$ ($\ell=20$)
Outline

• Motivation: opportunity & privacy risk
• Problem definition and analysis
• Proposed solution
• Experiment result

• Conclusion & Future work
Conclusion

- Proposed a rigorous and flexible **privacy model** for spatio-temporal data: l-trajectory privacy;
  - application: PPDM
  - business model: “privacy as services/money”

- **PPDP Framework** for spatio-temporal data;

- **Algorithms GA+MMD** for publishing private counts with high utility in real-time.
Future Work

- Improve algorithm GA+MMD for private counts
  - the counts should be non-negative & integer
- More flexible privacy models
  - $\ell$-trajectory privacy
  - Location-based privacy model
- PPDP / PPDM for mining personal spatio-temporal data
Thank you!
Any question?