Subspace Differential Privacy

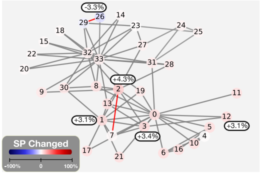
Jie Gao

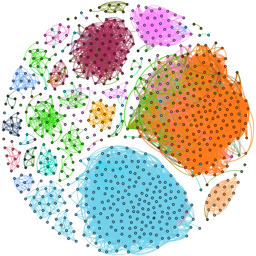
Rutgers University
http://sites.rutgers.edu/jie-gao

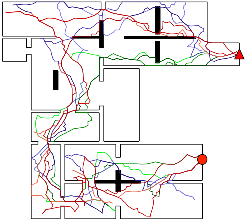
January 4th 2023

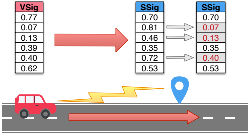
Data Challenges from Ubiquitous Sensing

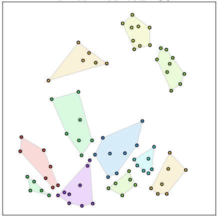
Enormous amount of inter-connected data collected from everyday living environment.

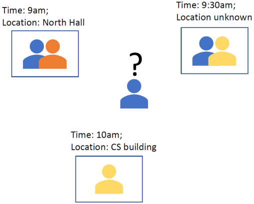




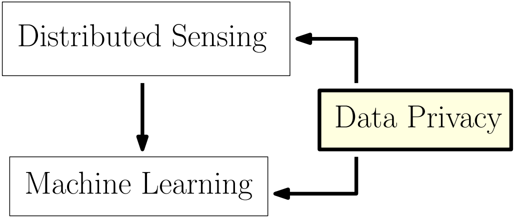








Motivation: Data Privacy in Distributed Sensing

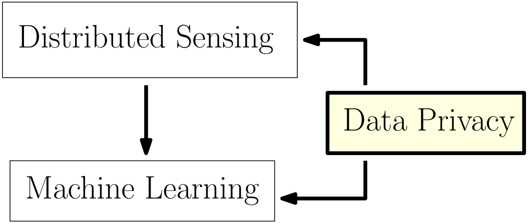


Motivation: Data Privacy in Distributed Sensing

Wearable devices

Cameras, Microphones, Gyro sensors

Envrionment sensors (for localization, tracking, activity recognition)



Identity

Location

Health condition

Social relationship

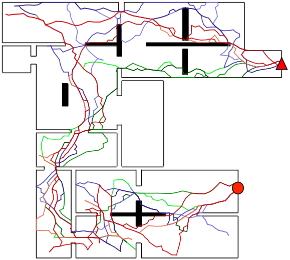
Personal health management, anomaly detection

Efficient energy management, improved comfort (smart building)

Civil engineering, traffic management, city planning

Application Scenario: Occupancy Sensing

(Utility) Goal: gather occupant counts.
Privacy Concern: Location + Identity.



Question: privacy model?

Outline

- **Review of differential privacy.**
- Challenges of applying DP in distributed sensing.
- Subspace DP: embracing invariants.
- Future directions.

Differential Privacy

Answer queries of a database of sensitive data entries.

- What is the average salary?
- What is the average salary of everyone except Bob?

Differential Privacy

Answer queries of a database of sensitive data entries.

- What is the average salary?
- What is the average salary of everyone except Bob?

DP: Add noises to the query output s.t. Bob's info is not revealed.

Differential Privacy

[Dwork, McSherry, Nissim and Smith, 2006]
A randomized mechanism M is ε-differentially private if for any two adjacent datasets D and D' (i.e., differ by one data entry), for a query f and any measurable subset $H \in \mathrm{Range}(f)$,

$$\Pr[f(D) \in H] \leq \exp(\varepsilon) \cdot \Pr[f(D') \in H].$$

Differential Privacy

[Dwork, McSherry, Nissim and Smith, 2006]
A randomized mechanism M is ε-differentially private if for any two adjacent datasets D and D' (i.e., differ by one data entry), for a query f and any measurable subset $H \in \mathrm{Range}(f)$,

$$\Pr[f(D) \in H] \leq \exp(\varepsilon) \cdot \Pr[f(D') \in H].$$

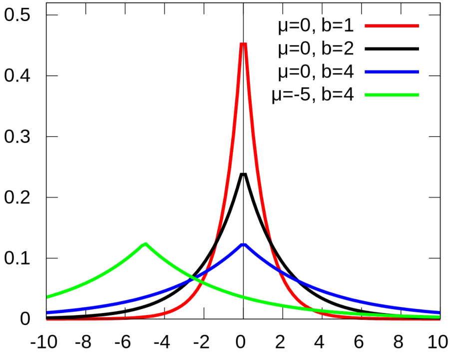
A randomized mechanism M is (ε, δ)-differentially private if for any two adjacent datasets D and D' (i.e., differ by one data entry), for a query f and any measurable subset $H \in \mathrm{Range}(f)$,

$$\Pr[f(D) \in H] \leq \exp(\varepsilon) \cdot \Pr[f(D') \in H] + \delta.$$

Laplace Mechanism

Laplace mechanism: add noise with distribution Lap(b):

$$P(x|b) = \frac{1}{2b}\exp(-\frac{|x|}{b})$$



Laplace Mechanism

The level of noise is usually determined in terms of *sensitivity*.

Laplace Mechanism

The level of noise is usually determined in terms of *sensitivity*.

The sensitivity of a function f is the largest possible difference in the output of f between any pair of adjacent databases:

$$\Delta f = \max_{(D,D')} |f(D) - f(D')|.$$

Laplace Mechanism

The level of noise is usually determined in terms of *sensitivity*.

The sensitivity of a function f is the largest possible difference in the output of f between any pair of adjacent databases:

$$\Delta f = \max_{(D,D')} |f(D) - f(D')|.$$

To achieve ε-differential privacy, a noise x of $\mathrm{Lap}(\Delta f/\varepsilon)$ suffices.

$$P\big(x = f(D) - f(D')\big) \sim \exp(\frac{|f(D) - f(D'|}{\Delta f/\varepsilon}) \leq \exp(\varepsilon)$$

Composition and Post-Processing

[Composition] Let M_1 and M_2 be randomized algorithm that are ε_1, ε_2-differentially private respectively. The output $(M_1(D), M_2(D))$ is $(\varepsilon_1 + \varepsilon_2)$-differentially private.

Composition and Post-Processing

[Composition] Let M_1 and M_2 be randomized algorithm that are ε_1, ε_2-differentially private respectively. The output $(M_1(D), M_2(D))$ is $(\varepsilon_1 + \varepsilon_2)$-differentially private.

[Post-Processing] For any deterministic or randomized function F defined over the image of the mechanism M, if M satisfies ε-differential privacy, so does $F(M)$.

Outline

- Review of differential privacy.
- **Challenges of applying DP in distributed sensing.**
- Subspace DP: embracing invariants.
- Future directions.

(New) Challenges for Data Privacy

- Accuracy: minimize error.

(New) Challenges for Data Privacy

- Accuracy: minimize error.
- Unbiasedness.

(New) Challenges for Data Privacy

- Accuracy: minimize error.
- Unbiasedness.
- Transparency: data-independent.

(New) Challenges for Data Privacy

- Accuracy: minimize error.
- Unbiasedness.
- Transparency: data-independent.

(New) Challenges for Data Privacy

- Accuracy: minimize error.
- Unbiasedness.
- Transparency: data-independent.

Structural or External Constraints:

- Structures from its physical nature.
  - Occupacy data: total headcount fixed;

(New) Challenges for Data Privacy

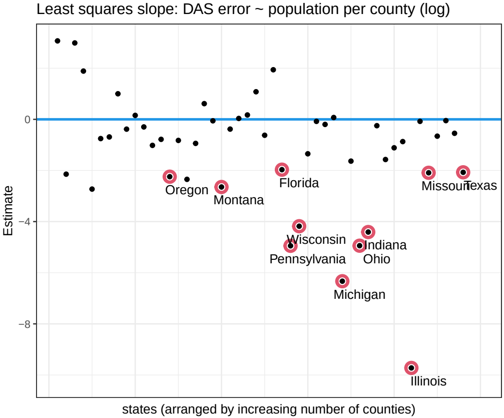
- Accuracy: minimize error.
- Unbiasedness.
- Transparency: data-independent.

Structural or External Constraints:

- Structures from its physical nature.
  - Occupacy data: total headcount fixed;
- External Invariant Constraints:
  - Disclosure Avoidance System (DAS) of 2020 Decennial Census.
  - Invariants: population total at state level; total housing units.

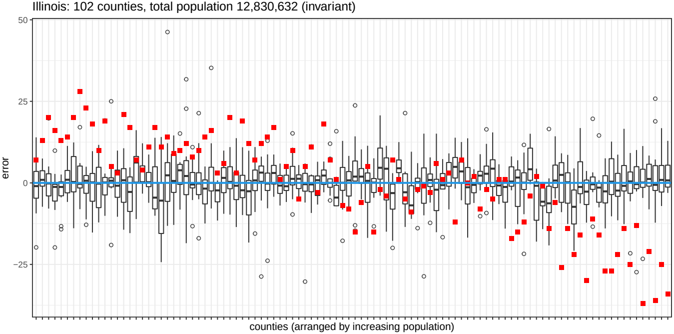
US CENSUS data

Add noise to county population count; while state population is accurate. TopDown Algorithm (Abowd et al., 2019): projection & rounding after DP)



US CENSUS data

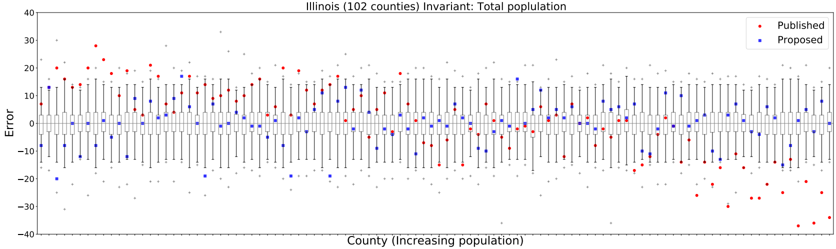
TopDown Algorithm vs Subspace DP (our method)



Counties of Illinois in increasing true population sizes, DAS errors (red squares) show a clear negative trend bias while our method (boxplots) shows no bias.

US CENSUS data

Integer subspace DP: population count should be integer values.



Outline

- Review of differential privacy.
- Challenges of applying DP in distributed sensing.
- **Subspace DP: embracing invariants.**
- Future directions.

Subspace Differential Privacy

- Database $x = (x_1, x_2, \cdots, x_N)^T$, with histogram $h(x)$.
- (Linear) counting queries: return the counts by a predicate p.

Example: Universe=$\{a, b, c, d\}$,

$$x = \begin{bmatrix} b \\ a \\ a \\ c \end{bmatrix}$$

Histogram

$$h(x) = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Subspace Differential Privacy

- Database $x = (x_1, x_2, \cdots, x_N)^T$, with histogram $h(x)$.
- (Linear) counting queries $A(x)$: return the counts by a predicate p.

Example: Query: how many entries of type $\{a, c\}$ are there?
Return: $A \cdot h(x)$, where

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}$$

and

$$h(x) = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Subspace Differential Privacy

- Database $x = (x_1, x_2, \cdots, x_N)^T$, with histogram $h(x)$.
- (Linear) counting queries $A(x)$: return the counts by a predicate p.
- ε-DP: $M(x)$
- Linear invariants: constraints C, $CM(x) = CA(x)$, $\forall x$.

Subspace Differential Privacy

- Database $x = (x_1, x_2, \cdots, x_N)^T$, with histogram $h(x)$.
- (Linear) counting queries $A(x)$: return the counts by a predicate p.
- ε-DP: $M(x)$
- Linear invariants: constraints C, $CM(x) = CA(x)$, $\forall x$.

Remark:

- Classical DP does not work – considering two databases differing by one entry but do not meet the same invariants.

Subspace Differential Privacy

- Database $x = (x_1, x_2, \cdots, x_N)^T$, with histogram $h(x)$.
- (Linear) counting queries $A(x)$: return the counts by a predicate p.
- ε-DP: $M(x)$
- Linear invariants: constraints C, $CM(x) = CA(x)$, $\forall x$.

Remark:

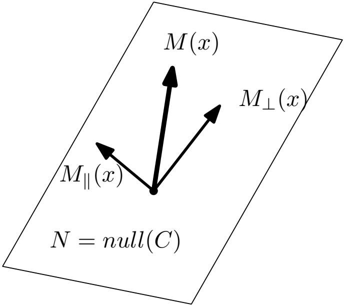
- Classical DP does not work – considering two databases differing by one entry but do not meet the same invariants.
- DP within a subspace N– defined by the contraints.

$$\Pr[P_N M(x) \in H] \leq \exp(\varepsilon) \cdot \Pr[P_N M(x') \in H].$$

Subspace Differential Privacy

$M(x) = M_{\|}(x) + M_{\perp}(x)$, where $M_{\|}(x) \in row(C)$ and $M_{\perp}(x) \in N = null(C)$.

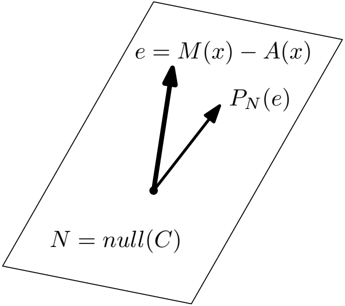
- ε-DP: $M_{\perp}(x) = P_N M(x)$ is differentially private.
- Linear invariants: constraints C, $CM_{\|}(x) = CM(x) = CA(x)$, $\forall x$.



Two Frameworks: Projection vs Extension

Projection: Convert a DP mechanism M to subspace DP.
$M'(x) = A(x) + P_N(M(x) - A(x)$, where $N = null(C)$.

- Linear invariants: $CM'(x) = CA(x)$, $\forall x$.
- Noise $e = M(x) - A(x)$, Project noise e into null space N.



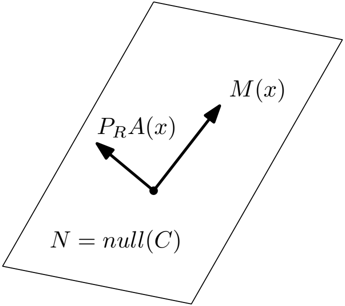
Need a full DP scheme M first.

Two Frameworks: Projection vs Extension

Extension: a DP mechanism M for query $P_N A(x)$ (within null space of C).

$M'(x) = P_R A(x) + M(x)$, where $R = row(C)$.

- Linear invariants: $CM'(x) = CP_R A(x) = CA(x)$, $\forall x$.
- DP follows from the DP of $M(x)$ in the null space.



Properties of Subspace DP

For both projection and extension mechanism:

- Unbiased if M is unbiased (true for Laplace and Gaussian mechanism).

Properties of Subspace DP

For both projection and extension mechanism:

- Unbiased if M is unbiased (true for Laplace and Gaussian mechanism).
- Extension mechanism – optimality of subspace scheme translates.

Properties of Subspace DP

For both projection and extension mechanism:

- Unbiased if M is unbiased (true for Laplace and Gaussian mechanism).
- Extension mechanism – optimality of subspace scheme translates.
- Transparency: our design mechanism depends only on invariants C (public knowledge) and not x (private data).

Properties of Subspace DP

For both projection and extension mechanism:

- Unbiased if M is unbiased (true for Laplace and Gaussian mechanism).
- Extension mechanism – optimality of subspace scheme translates.
- Transparency: our design mechanism depends only on invariants C (public knowledge) and not x (private data).
- Distributed implementation: pre-calculate the noise among distributed agents (with shared seeds).

Integer Subspace Differential Privacy

What if we want the output to be interger value?

- Grid points with additional linear constraints – Lattice space.

Integer Subspace Differential Privacy

What if we want the output to be interger value?

- Grid points with additional linear constraints – Lattice space.
- Solve the integer linear system $Ae = 0$ with interger coefficients – linear Diophantine equation.

Integer Subspace Differential Privacy

What if we want the output to be interger value?

- Grid points with additional linear constraints – Lattice space.
- Solve the integer linear system $Ae = 0$ with interger coefficients – linear Diophantine equation.
- Solving by the Smith normal form: $\exists U, V$ and diagonal matrix D such that $UAV = D$. Then, $e = Vw$.

$$D = \begin{pmatrix} \alpha_1 & 0 & 0 & & \cdots & & 0 \\ 0 & \alpha_2 & 0 & & \cdots & & 0 \\ 0 & 0 & \ddots & & & & 0 \\ \vdots & & & \alpha_r & & & \vdots \\ & & & & 0 & & \\ & & & & & \ddots & \\ 0 & & & \cdots & & & 0 \end{pmatrix} \cdot \quad w = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ e_1 \\ \vdots \\ e_k \end{pmatrix}$$

Integer Subspace Differential Privacy

What if we want the output to be interger value?

- Grid points with additional linear constraints – Lattice space.
- Solve the integer linear system $Ae = 0$ with interger coefficients – linear Diophantine equation.
- Solving by the Smith normal form: $\exists U, V$ and diagonal matrix D such that $UAV = D$. Then, $e = Vw$.

$$\mathrm{D} = \begin{pmatrix} \alpha_1 & 0 & 0 & & \cdots & & 0 \\ 0 & \alpha_2 & 0 & & \cdots & & 0 \\ 0 & 0 & \ddots & & & & 0 \\ \vdots & & & \alpha_r & & & \vdots \\ & & & & 0 & & \\ & & & & & \ddots & \\ 0 & & & \cdots & & & 0 \end{pmatrix} \cdot \ \mathrm{w} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ e_1 \\ \vdots \\ e_k \end{pmatrix}$$

- Highly non-trivial to sample discrete Laplace distribution on a lattice space.

Outline

- Review of differential privacy.
- Challenges of applying DP in distributed sensing.
- Subspace DP: embracing invariants.
- **Future directions.**

Big Picture

It is challenging to design privacy schemes when

Big Picture

It is challenging to design privacy schemes when

- Real world data has all kinds of special structures.

Big Picture

It is challenging to design privacy schemes when

- Real world data has all kinds of special structures.
- There is external information or prior knowledge on the data.

Big Picture

It is challenging to design privacy schemes when

- Real world data has all kinds of special structures.
- There is external information or prior knowledge on the data.

Next: two examples of why the problem is interesting/challenging.

- Networked data;
- Privacy in learning.

Run a Survey

Goal: what is the fraction of the population who smoke cigarette?

Run a Survey

Goal: what is the fraction of the population who smoke cigarette?

Individual: may not trust the data collector.

Run a Survey

Goal: what is the fraction of the population who smoke cigarette?
Individual: may not trust the data collector.

Random response: flip a coin

- If HEAD, tell the truth.
- If TAIL, report YES/NO uniformly randomly.

Run a Survey

Goal: what is the fraction of the population who smoke cigarette?
Individual: may not trust the data collector.

Random response: flip a coin

- If HEAD, tell the truth.
- If TAIL, report YES/NO uniformly randomly.

Analysis:

- If A is a smoker, report YES with probability 3/4.

Run a Survey

Goal: what is the fraction of the population who smoke cigarette?
Individual: may not trust the data collector.

Random response: flip a coin

- If HEAD, tell the truth.
- If TAIL, report YES/NO uniformly randomly.

Analysis:

- If A is a smoker, report YES with probability $3/4$.
- If A is not a smoker, report YES with probability $1/4$.

Run a Survey

Goal: what is the fraction of the population who smoke cigarette?
Individual: may not trust the data collector.

Random response: flip a coin

- If HEAD, tell the truth.
- If TAIL, report YES/NO uniformly randomly.

Analysis:

- If A is a smoker, report YES with probability $3/4$.
- If A is not a smoker, report YES with probability $1/4$.
- The total fraction of YES is $p/2 + 1/4$, where p is the true answer.

Run a Survey

Random response: flip a coin

- If HEAD, tell the truth.
- If TAIL, report YES/NO uniformly randomly.

(ε, δ)-differential privacy:

$$\text{Prob}\{YES|smoker\} \leq \text{Prob}\{YES|nonsmoker\} \cdot e^{\varepsilon} + \delta$$

Run a Survey

Random response: flip a coin

- If HEAD, tell the truth.
- If TAIL, report YES/NO uniformly randomly.

(ε, δ)-differential privacy:

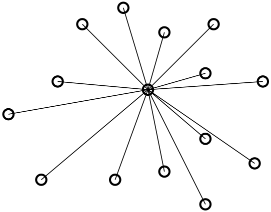
$$\text{Prob}\{YES|smoker\} \leq \text{Prob}\{YES|nonsmoker\} \cdot e^{\varepsilon} + \delta$$

$$\frac{3}{4} \leq \frac{1}{4} \cdot e^{\varepsilon} + \delta$$

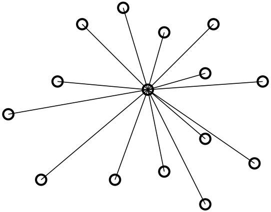
Random response is $(\ln 3, 0)$-differentially private.

Run a Survey

- What if the collector also knows the social network G?
- Smoking is a contagious behavior.



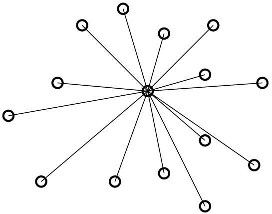
3/4 of all friends report YES



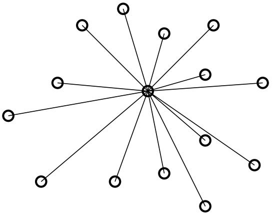
1/4 of all friends report YES

Run a Survey

- What if the collector also knows the social network G?
- Smoking is a contagious behavior.



3/4 of all friends report YES



1/4 of all friends report YES

Insight: Differential privacy is not broken but there is no ‘privacy’, as sensitive data can be inferred from friend circle.

Privacy in Learning

OccuTherm occupancy:

- Easy to train a neural network model to learn identity information.

Summary

With big data and ubiquitous sensing, there are major privacy concerns.

Privacy protection methods should respect the structure in data to defend against statistical inference attacks.

Acknowledgement

Subspace DP

- Jie Gao, Ruobin Gong, Fang-Yi Yu, Subspace Differential Privacy, AAAI-22.
- Prathamesh Dharangutte, Jie Gao, Ruobin Gong, Fang-Yi Yu, Integer Subspace Differential Privacy, AAAI-23.

Addressing structures in data.

- On Privacy of Socially Contagious Attributes, Aria Rezaei, Jie Gao, ICDM'19.
- Application-Driven Privacy-Preserving Data Publishing with Correlated Attributes, Aria Rezaei, Chaowei Xiao, Jie Gao, Bo Li, Sirajum Munir, EWSN 2021.