Differential Privacy and Its Applications

Gang Li, Tianqing Zhu, Wanlei Zhou
School of Information Technology
Deakin University, Australia

Road Map
• Background and Preliminary
  • Correlated Differential Privacy
  • Differentially Private Data Release
  • Differentially Private Data Analysis
  • Applications
  • Future Directions and Conclusion

Special Issues on Privacy
• IEEE Spectrum, Aug. 2014
  – On the Internet, nobody knows you are a dog (1993).
  – Interested parties not only know you are a dog, but also know the colour of your fur (2014)

Special Issues on Privacy
• Communication of ACM, Sept. 2014
  – Federal law governing student privacy and the release of student records suggests that anonymizing student data can hardly protect student privacy.

Special Issues on Privacy
• Science, Jan. 2015
  – Data pour out of us and our devices every second of every day, and people no longer control their personal privacy.

One of the 1st Concerns in Statistics
AOL Dataset Debacle

- AOL search data leak (2006):
  - 36 million search terms of 650,000 users
    - http://search-id.com
    - http://www.aolstalker.com
  - I Know who you are
    - Click history can uniquely identify a person
    - Find all log entries for AOL user 4417749
    - Multiple queries for businesses and services in Lilburn, GA (population 11K)
      - queries for Jarrett Arnold
      - Lilburn has 14 people with the last name Arnold
      - NYT contact them, finds out User 4417749 is Thelma Arnold

Breach of medical record

- 87% of Americans can be uniquely identified by
  - {zip code, gender, date of birth}
- Latanya Sweeney re-identify the medical record of an ex-governor of Massachusetts.
  
  [International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002]

Latanya Sweeney’s method:
- In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees.
- For twenty dollars she purchased the voter registration list for Cambridge Massachusetts
- This information can be linked

Linkage Attacks
- Using “innocuous” data in one dataset to identify a record in a different dataset containing both innocuous and sensitive data
- At the heart of the voluminous research on hiding small cell counts in tabular data
Breach of RecSys Ratings

Netflix Competition

- Simple Anonymization
  - Delete identity information such as User name, ID No.
  - The date of each rating and the title and year of release for each movie are provided
  - Some ratings are not sensitive, but some may be sensitive, for Users:
    - fine for Netflix to know,
    - not fine for public to know

IMDB Data

- Individuals may register for an account and rate movies
  - Need not be anonymous
  - Visible material includes ratings, dates, comments
  - By definition, these ratings not sensitive

Breach of RecSys Ratings

Linkage Attack

- We can re-identify a Netflix rater if we know just a little bit about her
  - 8 movie ratings (≤ 2 wrong, dates ± 2 weeks) → re-identify 99% of raters
  - 2 ratings, 3 days → re-identify 68% of raters
    - Relatively low confidence for the other 32% (especially with movies outside the top 10)
  - Even a handful of IMDB comments allows Netflix re-identification, in many cases
    - 50 IMDB users → re-identify 2 with very high probability, one from ratings, one from dates

Linkage Attack

- The Netflix attack works because the data are sparse and dissimilar, with a long tail.

Considering just movies rated, for 90% of records there isn’t a single other record that is more than 30% similar

I Know Where Your Cat Lives

- This project randomly selected one million images that include the word “cat” across public photo sites, which plots the location coordinates from each photo against a map to show where each cat lives.

--- July 22, 2014 Time
Background

Data are collected in:
• Medical, Bank, Social Network, …

Social benefits:
• better services, Finding correlations, Publishing official statistics, Data Mining, …

Something needs to be done for Privacy Breach:

Privacy Breach in Data Release/Sharing:
– Adversary with background information, may re-identify the user from the aggregated information of the dataset.
– Adversary observes the aggregated information of the dataset, removing a particular user will change the statistical information.

Who is an adversary?

Every user is potentially an adversary
– After data is released, we cannot prevent any user from performing any type of analysis on the released data
– Worst case scenario
  • Must account for disclosure risk from any and all types of analyses

Privacy Model

– a set of rules/assumptions to describe/measure the privacy of a data set.
– Privacy: (informal definition)
  • ? Adversary cannot learn anything new for personal information after he access the released dataset

Privacy Model

Traditional Privacy Model

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cryptography</td>
<td>• Not decreasing accurate. Suitable for multiple parties computation</td>
<td>• Computation complexity; people may not want to participate</td>
</tr>
<tr>
<td>Anonymization</td>
<td>• Easy to understand. Easy to implement for low dimensional dataset</td>
<td>• No guarantees on the quality of the dataset; NP-hard for high dimensional dataset; Weak privacy guarantee</td>
</tr>
<tr>
<td>Perturbation</td>
<td>Provide high privacy level</td>
<td>• Noise size is subjective</td>
</tr>
</tbody>
</table>
Weakness of Traditional PM

• Typical weakness
  – Privacy level is difficult to be measured and compared
  – The privacy guarantee is hardly to be proved theoretically
  – Susceptible to background attack/linkage attack

• Need a more semantic approach to privacy

What is Privacy?

• According to the Privacy Act, **Privacy** refers to the protection of users’ personal information
  – Personal Information
    – “… information or an opinion (including information or an opinion forming part of a database) about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion”

• Location information can be automatically added when an image is taken.

What is Privacy?

• According to Agrawal, **Privacy** refers to
  – “the right of individuals to determine for themselves when, how and to what extent information about them is communicated to others”

  [Agrawal’03]

• More complex than confidentiality, secret

What to promise?

• Respondent will fell safe submitting his data if
  “I know that my answer has no impact on the aggregated information.”

  \[
  M(\text{Data}) = M(\text{Data} - \text{MyRecord}) = \ldots = M(\Phi)
  \]

 “I know that any adversary looking at the aggregated information (constructed model or query results) \( R \) could not learn any new information with high probability about me”

  \[
  P(\text{MySecret} | R) = P(\text{MySecret})
  \]

What to promise?

• Respondent will fell safe submitting his data if
  “I know that my answer has no impact on the aggregated information.”

  \[
  M(\text{Data}) = M(\text{Data} - \text{MyRecord}) = \ldots = M(\Phi)
  \]

 “I know that any adversary looking at the aggregated information (constructed model or query results) \( R \) could not learn any new information with high probability about me”

  \[
  P(\text{MySecret} | R) = P(\text{MySecret})
  \]

  If \( R \) shows a trend in his population, the trend is true of the person with high probability, no matter data submitted or not
What to promise?

- Respondent will feel safe submitting his data if
  “If I knew the chance that the privatized aggregated information (constructed model or query results) R was nearly the same, whether or not I submitted my information”

Analogy with Cryptography

<table>
<thead>
<tr>
<th>Cryptography</th>
<th>Privacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plaintext</td>
<td>Individual information</td>
</tr>
<tr>
<td>Crypto Algorithm</td>
<td>Privacy Algorithm</td>
</tr>
<tr>
<td>Cipher text</td>
<td>Aggregate Information</td>
</tr>
<tr>
<td>From the Cipher text, adversary cannot obtain anything about Plaintext</td>
<td>From the aggregate information, adversary cannot obtain anything about individual information</td>
</tr>
</tbody>
</table>

- Anything about the plaintext that can be learned from the cipher text can be learned without the cipher text
- Intuition: prior and posterior views about an individual should not change too much

Differential Privacy

- An individual is in or out of the database should make little difference of the analytical output


Why does it work?

- To obtain aggregated information while preserve privacy of individuals

What does D.P. promise?

- DP promises the following:
  - The aggregated information (constructed model or query results) R gives minimal evidence about whether or not any given individual contributed to the data set
  - It protects all your personal information in the data, if only you provide information about yourself.
    - In its original form, it does not prevent adversary from inferring information about correlated groups
    - It does not prevent adversary from knowing individual’s information from aggregated information
What does D.P. promise?

It does NOT promise the following:

• If an adversary can not tell whether or not you submitted the data, he can not learn anything about you from the aggregated information.
  • with the right background information, an adversary can learn about you from aggregated information about the population, even if you didn’t submit the data.

What does D.P. promise?

It does NOT promise the following:

An adversary can not guess with high probability whether you submitted your data.

• In many applications, the adversary can guess with high probability that a person’s data is in
  • Correlated Data Set

2. DP Mechanism

• Differential Privacy
• Laplacian Mechanism
• Exponential Mechanism
• Privacy-Utility Trade off

Differential Privacy

Definition:

a mechanism $M$ is $\epsilon$-differential privacy if for all pairs of neighboring datasets $D$ and $D'$, and for all possible output $S$, satisfy with:

$$\Pr[M(D) \in S] \leq e^\epsilon \Pr[M(D') \in S]$$

$\epsilon$ is Privacy Budget

Privacy Budget

$\epsilon$ controls the privacy guarantee level of mechanism.

• A smaller $\epsilon$ represents a stronger privacy.
• Normally, it is less than 1.

Sensitivity

Sensitivity is a parameter determining how much perturbation is required in mechanisms.

• Global Sensitivity
• Local Sensitivity
Sensitivity: Global Sensitivity

- The **global sensitivity** considers the maximal difference between query results on neighboring datasets
  - indicates how much the difference should be hidden in mechanisms
  - Only related to query

Sensitivity: Example

Suppose we have a dataset D and two queries: \( f_1 = \text{Count}(\text{HIV}) \), \( f_2 = \text{Average}(\text{Age}) \). Let \( r \) represent the record.

<table>
<thead>
<tr>
<th>Job</th>
<th>Sex</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineer</td>
<td>Male</td>
<td>35</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>Engineer</td>
<td>Male</td>
<td>50</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>Lawyer</td>
<td>Male</td>
<td>35</td>
<td>HIV</td>
</tr>
<tr>
<td>Writer</td>
<td>Female</td>
<td>30</td>
<td>Flu</td>
</tr>
<tr>
<td>Writer</td>
<td>Female</td>
<td>30</td>
<td>HIV</td>
</tr>
<tr>
<td>Dancer</td>
<td>Female</td>
<td>30</td>
<td>HIV</td>
</tr>
<tr>
<td>Dance</td>
<td>Female</td>
<td>30</td>
<td>HIV</td>
</tr>
</tbody>
</table>

\[
\text{Count}(\text{HIV}) \quad f_1(D) = 4 \\
\text{Count}(\text{HIV}) \quad f_1(D) - r_1 = 4 \\
\text{Count}(\text{HIV}) \quad f_1(D) - r_2 = 4 \\
\text{Count}(\text{HIV}) \quad f_1(D) - r_3 = 4 \\
\text{Count}(\text{HIV}) \quad f_1(D) - r_4 = 4 \\
\text{Count}(\text{HIV}) \quad f_1(D) - r_5 = 3 \\
\text{Count}(\text{HIV}) \quad f_1(D) - r_6 = 3 \\
\text{Count}(\text{HIV}) \quad f_1(D) - r_7 = 3 \\
\]

\[
\Delta f_{GSC} = |4 - 3| = 1 \\
\]

\[
\text{Average}(\text{Age}) \quad f_2(D) = 34.3 \\
\text{Average}(\text{Age}) \quad f_2(D) - r_1 = 34.1 \\
\text{Average}(\text{Age}) \quad f_2(D) - r_2 = 31.6 \\
\text{Average}(\text{Age}) \quad f_2(D) - r_3 = 34.1 \\
\text{Average}(\text{Age}) \quad f_2(D) - r_4 = 35 \\
\text{Average}(\text{Age}) \quad f_2(D) - r_5 = 35 \\
\text{Average}(\text{Age}) \quad f_2(D) - r_6 = 35 \\
\text{Average}(\text{Age}) \quad f_2(D) - r_7 = 35 \\
\]

\[
\Delta f_{GSC} = |34.3 - 31.6| = 2.7 \\
\]

Differential Privacy Mechanism

- **Laplace Mechanism:**
  - suitable for numeric output
  - How many people in this room have blue eyes?

- **Exponential Mechanism:**
  - suitable for non-numeric output
  - What is the most common eye color in this rooms?

Laplace Mechanism

- In order for two worst-case neighboring data sets to produce a similar distribution of aggregated information, calibrated noise is added to span the sensitivity gap

- **Laplacian noise** is the easiest way

\[
\text{Laplace Mechanism} \quad f(D) \quad \text{true answer} \\
\]

\[
\text{Laplace Mechanism} \quad f(D) + \text{noise} \quad \text{noisy output} \\
\]

\[
\text{Laplace Mechanism} \quad M(D) = f(D) + \text{Laplace} \left( \frac{\Delta f}{\varepsilon} \right) \\
\]

Laplace Mechanism

- Let \( f(D) \) be a numeric query on dataset \( D \)
  - How many people in this room have blue eyes?
  - The sensitivity of \( f \Delta f = \max \| f(D) - f(D') \|_1 \)

- A Laplace Mechanism \( M \) is \( \varepsilon \)-differential privacy:

\[
M(D) = f(D) + \text{Laplace} \left( \frac{\Delta f}{\varepsilon} \right) \\
\]


Laplace Mechanism

Why it works?
- A Laplace Mechanism $M$ is $\epsilon$-differential privacy:
  - What we want is $\Pr[M(D) \in S] \leq e^{\epsilon}\Pr[M(D') \in S]$
  - Substituting the equation from Laplacian noise to it:

$$\frac{\Pr[M(D) \in S]}{\Pr[M(D') \in S]} = \frac{\frac{e^{\epsilon} \exp(-\frac{|f(D) - f(D')|}{\Delta f})}{2\Delta f}}{\frac{e^{\epsilon} \exp(-\frac{|f(D) - f(D')|}{\Delta f})}{2\Delta f}} = \exp(\frac{|f(D) - f(D')|}{\Delta f}) \leq e^{\epsilon}$$

Laplace Example

- Query: How many people has HIV?
  - DP answer = True answer + Noise
  - Sensitivity is 1, because the answer is changed most at 1 if one user is deleted.
  - If we define $\Delta f = 1$, the noise is sample from:
    - $DP$ answer $M(D)$:
      - $4 + 1 = 5$ (higher probability)
      - $4 - 1 = 3$ (lower probability)

Exponential Mechanism

- Exponential Mechanism is suitable for non-numeric output $R$
  - What is the most common eye color in this rooms?
    - i.e. $R = \{Brown, Blue, Black, Green\}$
  - Paired with a quality score $q$:
    - $q(D, r)$ represents how good an output $r$ is for dataset $D$
  - An exponential mechanism $A$ is $\epsilon$-differential privacy if:
    $$A(D, q) = \{\text{return } r \text{ with probability } \alpha \cdot \exp(\frac{\epsilon \cdot q(D, r)}{2\Delta q})\}$$
    Sensitivity of $q$: $\Delta q = \max ||q(D) - q(D')||_1$

Exponential Example

- What is the most common eye color in this rooms?
  - i.e. $R = \{Brown, Blue, Black, Green\}$
  - Pr($r$) $\propto \exp(\frac{\epsilon \cdot q(D, r)}{2\Delta q})$

Privacy vs Utility

Noise depends on $\Delta f$ and $\epsilon$, not on the database:
- Increasing $\epsilon$ flattens curve (smaller $b$);
  - more utility + less privacy
- Decreasing $\epsilon$ sharpens curve;
  - less utility + more privacy

$$M(D) = f(D) + Lap(\frac{2\epsilon}{\Delta f})$$

Advantage of DP

<table>
<thead>
<tr>
<th></th>
<th>Traditional PM</th>
<th>Differential Privacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy level can be measured and compared</td>
<td>No</td>
<td>Can be measured by the privacy budget</td>
</tr>
<tr>
<td>The privacy guarantee can be proved theoretically</td>
<td>No</td>
<td>DP definition</td>
</tr>
<tr>
<td>Resist background attack</td>
<td>No</td>
<td>DP assumes that attackers get to know everyone’s information except the one we will protect.</td>
</tr>
</tbody>
</table>

Impact of changing a single record

Option | Score | Sampling Probability |
-------|-------|----------------------|
Brown   | 23    | 0.25                 |
Blue    | 9     | 0.25                 |
Black   | 27    | 0.25                 |
Green   | 0     | 0.25                 |

Sensitivity of $q$: $\Delta q = \max ||q(D) - q(D')||_1$
Sensitivity: Local Sensitivity

Local sensitivity calibrates the record-based differences between query results on neighboring datasets and also satisfies the differential privacy definition.

- Related to record

$$\Delta f_{LS} = \max_{D'} \| f(D) - f(D') \|_2$$

Sensitivity: LS and GS

- For query such as median, max, the local sensitivity is smaller than global one.
- For query such as count and range, the local sensitivity may larger than global one.

Correlated Dataset

- Correlated Dataset
- Challenges for DP

DP Assumptions

- I.I.D:
  - Dataset consist of one certain distributions with generating records independently.
- Non-I.I.D:
  - dataset consists of all possible probability distributions over database instances, and parts of records may correlated to each other.
  - We call the dataset as coupled dataset.

DP in Correlated Dataset

Lots of the existing samples are correlated with each other.

- Traditional DP may underestimate the privacy risk on releasing the dataset [1].

[1] No Free Lunch in Data Privacy
DP in Correlated Dataset

- Alice and her 9 family members are living together. When Alice contracts the Flu, the entire family will also be infected.
- Deleting the record of Alice will impact 9 other records.
- And the count of Flu will change to either:
  - 90 (Alice got the Flu)
  - 100 (Alice is healthy).

90 Flus

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>A</td>
<td>1/0</td>
</tr>
<tr>
<td>Bob</td>
<td>A</td>
<td>1/0</td>
</tr>
<tr>
<td>Cathy</td>
<td>A</td>
<td>1/0</td>
</tr>
<tr>
<td>Douglas</td>
<td>A</td>
<td>1/0</td>
</tr>
<tr>
<td>Yang</td>
<td>B</td>
<td>1/0</td>
</tr>
<tr>
<td>Zoe</td>
<td>B</td>
<td>1/0</td>
</tr>
</tbody>
</table>

Name | Address | A | B | C |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bob</td>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cathy</td>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Douglas</td>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yang</td>
<td>B</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zoe</td>
<td>B</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

10 records

DP in Correlated Dataset

- If we have any correlations between records, then some DP algorithms leak more information than allowable.
- An attacker's posterior beliefs may differ significantly from the prior beliefs depending on the strength of the correlation.

Traditional Solution
- Add more noise than independent samples.
  - If there are k records are correlated to each other, then the sensitivity is at least k multiple original sensitivity s.
  - This method will introduce huge noise in the result.

Sensitivity: maximal difference between the query on the neighbouring dataset.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>100</td>
</tr>
<tr>
<td>C</td>
<td>200</td>
</tr>
</tbody>
</table>

A, B, C are independent. Deleting a row will change the count of 1 at the most. Sensitivity=1

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>200</td>
</tr>
<tr>
<td>C</td>
<td>100</td>
</tr>
</tbody>
</table>

A, B, C are linear dependent. Deleting a row in A will change 1 count in C at the most. Sensitivity=2

A*B=C

Deleting a row will change at most k items at the most. Sensitivity=k

Task for Correlated DP

Preserve Privacy in correlated datasets:
- Privacy: guarantee ε-DP
- Utility: decrease the noise

Challenges
- How to find out the correlated records
- How to measure the Sensitivity

Sensitivity of Correlated DP:
- Global Sensitivity: determined by the maximal difference in the whole dataset.
- Local Sensitivity: determined by each record.
  - Still doesn't consider the relationship between records.

• Traditional DP will introduce large noise if the sample is correlated to others.
Challenge 1: the correlation

With Background: curator or attacker get to know the background in advance.

With Background:
- curator or attacker get to know the background in advance.
- Without Background: curator knows nothing about the background.
  - 1. find out the relationship according to their attributes.
  - 2. Same values on a particular attributes.
  - 3. Time correlation
  - 4. Background: relatives, linear combination
  - 5. Pearson Correlation Coefficient (linear)

Challenge 2: the Sensitivity

According to the definition of DP, the sensitivity is measured by the $\max f(D)-f(D')$
- It may be very high when facing the correlated dataset.

How to accurately estimate the sensitivity?
- Related to specific records
- Related to specific queries

Sensitivity

Correlated parameter $\delta \in \mathbb{R}$ represent the correlation between each record. We maintain a correlated matrix, which consists all the relationship between records:

\[
\begin{bmatrix}
1 & \cdots & 0.2 \\
0.2 & \cdots & 1
\end{bmatrix}
\]

Property of correlated matrix:
- Asymmetric: $\theta_{ij} \neq \theta_{ji}$
- Diagonal: $\theta_{ii}(X) = X$
- Sparse: only part of the records are correlated to each other

Correlated Sensitivity

Correlated Sensitivity is measured by record sensitivity and query $Q$
- For each record, measure the record sensitivity
- For each query, measure the correlated sensitivity
- The calibrated noise is then derived from the correlated sensitivity

$Q(D) = Q(D) + \text{Lap}(\frac{CS}{\epsilon})$

Correlated Sensitivity

Record Sensitivity
- $CS_i = \Sigma_{j=0}^n \theta_{ij} (\|Q(D) - Q(D')\|_1)$
- $CS_i$ is the sensitivity of record $i$

For each query, correlated sensitivity is

$CS_q = \max_{i \in q} CS_i$

Where $q$ is a set of records related to query $Q$
Utility Analysis

- Let $D$ be any dataset, for query class $Q$, and any parameter $\alpha, \beta > 0$, CIM is $(\alpha, \beta)$-accurate for $Q$.

$$Pr(|Q(D) - \hat{Q}(D)| < \alpha) > 1 - \beta$$

3. Differentially Private Data Release

- Interactive Setting
- Non-interactive Setting

Overview

- Release Setting: Interactive, Non-interactive

Interactive Setting

- Input: Data set $D$ and set of query $F = \{f_1, \ldots, f_k\}$
- Output: Set of query answers $\hat{F} = \{\hat{f}_1, \ldots, \hat{f}_k\}$

Research Issues

- Query type and numbers
- Accuracy and Efficiency

Interactive: Static

$D$: Micro-Transaction Dataset

<table>
<thead>
<tr>
<th>Rich</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>30</td>
<td>Male</td>
</tr>
<tr>
<td>Male</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>20</td>
<td>Nu</td>
</tr>
<tr>
<td>Female</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Challenge:

1. Answer query accurately in limited privacy budgets
2. Efficiency
3. The number of queries it can answer

Interactive: Static

- Laplace: Add noise to true query answer directly [Dwork 06].

$$\hat{f}_i = f_i(D) + \text{Lap}(b)$$

- Can answer all types of queries

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In TCC'06
Disadvantages:
- Has poor performance when there are lots of queries or query sensitivity is high
- It can only answer sub-linear of $n$ queries accurately [Dinur03].

Why will the performance of Laplace mechanism get worse when there are lots of queries?

$$f(D) = f(D) + \text{Lap} \left( \frac{\Delta f}{\varepsilon} \right)$$

- If there are $k$ queries, to preserve $\varepsilon$-differential privacy, each query can only be allocated to $\frac{\varepsilon}{k}$ privacy budget, the noise for each query will be $\text{Lap} \left( \frac{\Delta f}{\varepsilon} \right)$.

Interactive: Static

- **Query Separation**: just answer part of query in $F$, for others, use median answers [Roth 10].

  $$F = \{f_1, f_2, f_3, f_4, \ldots, f_k\}$$

  $$A = \{f_1, f_2, \text{Med}_3, f_4, \ldots, \text{Med}_{k-1}, f_k\}$$

  - Answer limited type of queries, but find Median is a problem
  - A. Roth and T. Roughgarden. Interactive privacy via the median mechanism. STOC 2010.

- **Iteration**: Approximate the true dataset. set an initial dataset, answer query iteratively, and in each run, update the dataset until it approximate the original dataset [PMW Hardt 10, IDC Gupta 12]

  - Anupam Gupta, Aaron Roth, and Jonathan Ullman. Iterative constructions and private data release. TCC'12

  1. Initial the dataset $D_0$ as uniform distribution
  2. For $i = 2$ to $k$
     1. Select one query from $F$ and compare $f_i$ with $f_i(D_j)$
        - if the difference satisfied with some criteria, update $D_j$ to $D_{j+1}$
        - Otherwise, release $f_i(D_j)$
  2.2 Release $f_k(D_j)$

Because only **update step** will consume privacy budgets, it can answer more queries than Laplace mechanism.

- However, it is inefficient, which is exponential to the dimension

Solutions Comparison

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Typical Algorithm</th>
<th>Efficiency</th>
<th>Accuracy</th>
<th>Number of Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace/Exp</td>
<td>Laplace</td>
<td>Poly(N)</td>
<td>Not Accurate when $k$ is large</td>
<td>Linear of $N$ (Dinur03)</td>
</tr>
<tr>
<td>Query Separation</td>
<td>Median</td>
<td>Exp(k)</td>
<td>Accurate</td>
<td>Exp(N)</td>
</tr>
<tr>
<td>Iteration</td>
<td>PMIR, IDC</td>
<td>Exp(k)</td>
<td>Accurate</td>
<td>Exp(N)</td>
</tr>
</tbody>
</table>

Recall Challenges:
1. Efficiency
2. Answer query accurately in limited privacy budgets
3. The number of query it can answer
**Interactive: continual release**

### Dataset

- **D**: Set of datasets released by time

<table>
<thead>
<tr>
<th>Data Stream</th>
<th>( T_i )</th>
<th>( T_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3</td>
<td>0 3 1</td>
</tr>
</tbody>
</table>

- **Challenges:**
  1. How to hide the difference between each time step
  2. Decrease noise
  3. The number of time step \( T_n \) is pre-defined

**Laplace Mechanism [Dwork 06]**

Divided \( \epsilon \) in \( T_n \) parts, and add noise to the output in each step

<table>
<thead>
<tr>
<th>Data Stream</th>
<th>( T_i )</th>
<th>( T_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 1 1 0</td>
<td>0 1 0 1</td>
</tr>
</tbody>
</table>

We have to add noise of \( T_n \cdot Lap\left(\frac{\epsilon}{T}ight) \) (huge) in total

---

**Non-Interactive Setting**

- **Non-Interactive Setting**
  - Output: Set of query answers \( \hat{F} = \{\hat{f}_1, \ldots, \hat{f}_k\} \) in batch
  - Can we release a dataset?

- **Research Issues**
  - Accuracy and Efficiency

---

**Non-Interactive: Synthetic Data**

- **f(D): Synthetic Dataset**
  - Give a set of query \( F \), curator will generate a synthetic dataset to answer \( F \).

\[ F \longrightarrow \text{Original Dataset} \longrightarrow \text{Synthetic Dataset} \longrightarrow \hat{F} \]

- **Challenge:**
  1. Accuracy
  2. Efficiency

---

**Non-Interactive: Synthetic Data**

Dataset \( D \) can be considered as a set of records sampled from a universe \( U \):
Synthetic Data: Efficiency

Search a synthetic dataset in a universe is time consuming, which is exponential to the dimension d.

Synthetic Data: Accuracy

Synthetic dataset measurement [Kasiviswanathan08]:

\[ (\alpha, \delta) - \text{usefulness}. \] A dataset mechanism \( M \) is \((\alpha, \delta)\)-usefulness for queries in class \( F \) if with probability \( 1 - \delta \), for every \( f \in F \), and every dataset \( D \), when \( \hat{D} = M(D) \), we have

\[ \max |f(\hat{D}) - f(D)| \leq \alpha \]

Accuracy Parameter

Non-Interactive: Synthetic Data

Advantage:

User can have a 'real' dataset

Disadvantage:

The synthetic dataset can only answer query set \( F \), and \( F \) should belong to a class.

The computational time is exponential to the size of the universe \( V \).

The result could not apply when the data is drawn from a continuous distribution.

Summary on Data Release

Query Type:

- Laplace mechanism can answer all types of queries, but others may not.
  - Count, Range, Sum are easier.
  - Avg, Max, Distance are difficult.

Query Number:

- Exponential of \( N \).

Accuracy:

- Depends on the query type and number

Efficiency:

- Most of mechanisms are inefficient, unless they made some assumptions.

4. Differentially Private Data Analysis

Laplace/Exponential Framework

- SuLQ
- PINQ

Private Learning Framework

- ERM
- Sample Complexity
### Background

Comparison between data analysis and release

<table>
<thead>
<tr>
<th>Data Release</th>
<th>Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanism</td>
<td>Independent Mechanism, Incorporate with particular algorithm</td>
</tr>
<tr>
<td>Output</td>
<td>A set of query answers or synthetic dataset, models</td>
</tr>
</tbody>
</table>

### Data Analysis Overview

- It aims to release an approximate accurate analysis *models*.
- The essential idea is *extending* current non-private algorithms to differential privacy algorithms.
- These extension tasks can be implemented by several *frameworks*:
  - Lap/Exp
  - Private Learning

### Laplace/Exponential Framework

- Sub-Linear Queries (SuLQ) outputted randomized continuous and Boolean values for each input. [Blum05]
  - Input Query /: \( D \rightarrow [0,1] \)
  - Output \( \Sigma_{i} f(x_{i}) + N(0, R) \)

- It can be used to implement *Decision Tree, SVD, k-means, the Perceptron Algorithm*, etc.


But SuLQ only has noise adding operations, which might be insufficient for some applications.
Laplace/Exponential Framework

**PINQ:** Privacy Integrated Queries platform (PINQ) provides more operators [McSherry10].

- It uses Laplace noise on numeric queries and Exponential mechanism on selection operators. It also provide Partition operators to deal with disjoint dataset.


#### Algorithm Examples

<table>
<thead>
<tr>
<th>Category</th>
<th>Challenges</th>
<th>Typical Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised Learning</td>
<td>Accuracy</td>
<td>SuLQ-ID3 [Blum05]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PINQ-ID3 [Friedman10]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random ID3 [Jagannathan09]</td>
</tr>
<tr>
<td>Unsupervised Learning</td>
<td>Sensitivity</td>
<td>K-means [Blum05], Corset [Feldman09]</td>
</tr>
<tr>
<td>Frequent Itemset Mining</td>
<td>Measurement</td>
<td>FIM [Feld10], PrivBasis [Li12], FIM-MCMC [Shen13]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exponential itemset candidates</td>
</tr>
</tbody>
</table>

Summary on E/L Framework

- **Advantages:**
  - Flexible: can be incorporated in various types of algorithms.
  - Simple: SuLQ and PINQ provide a possible way for non-expert.

- **Disadvantages:**
  - Accuracy is a huge challenge.
  - Many queries has large sensitivity.

- It is widely used in current application, but need further improvement.

Private Learning Framework

- Goal: Implement machine learning algorithms that protect the privacy of individuals.
  - Privacy: Differential Privacy
  - Learning: Obtain approximate accuracy models

Private Learning Framework

- It concerns about following problems:
  - How can we incorporate the differential privacy to learning algorithms?
    - Empirical Risk Minimization (ERM)
    - How many samples are needed in a bounded error?

Learning Problem

- $X = \{x_1, \ldots, x_m\}$ is a set of samples draw from a universe $U$. Suppose the total number of samples is $m$.
- Dots and triangles are denote as labels $c$.
- A concept $c$ is a function that separates one label from others.

The goal of the learning process $L$ is to find a hypothesis $h: X \rightarrow \{0, 1\}$ which agrees with $c$ on almost the entire universe.
Empirical Risk Minimization (ERM)

- How can we tell a good function (hypothesis) from a bad one?
  - The Empirical Risk Minimization (ERM) chooses a hypothesis by minimizing the Expect Loss over the training data.

Data Universe $\mathcal{H} = \{h_1, \ldots, h_n\}$

Min(Expect Loss) $\mathcal{L}$

$H(h_1, \ldots, h_n)$

Empirical Risk Minimization (ERM)

- Suppose $h$ is a hypothesis and $w$ is the output model, we define a loss function $\ell(h(w, x), y)$.
- The goal of ERM is to identify a $w$ minimizing the expectation risk:
  
  $$ R(w^*) = \min_{w^*} \mathbb{E}_{x, y \sim D} (\ell(w, x, y)) $$

- However, as the distribution of universe is unknown, we can only estimate the empirical risk $R_m(w)$ on sample set $D$.
  
  $$ R_m(w) = \min_{w} \frac{1}{m} \sum_{i=1}^{m} \ell(h(w, x_i), y_i) + \lambda r(w) $$

Empirical Risk Minimization (ERM)

- By choose different Loss Expectation, ERM implements to various learning algorithms:
  - Linear Regression [Chaudhuri2011]
  - Logistic Regression [Chaudhuri2008]
  - Kernel Method [Jain2013]
  - ...

DP ERM: Private Learner

- Two methods to produce differentially private learner via this regularized ERM [Chaudhuri11]
  - objective perturbation: added noise to the objective function prior to learning.
  - output perturbation: inserted noise to the output $w$;

Alg: Objective Perturbation

Require: Dataset $D$ and related parameters
Ensure: output private model $\hat{w}$
1. Sample noise $\eta$ from Gamma distribution
2. $\hat{w} = \arg\min_{w} \frac{1}{m} \sum_{i=1}^{m} \ell(h(w, x_i), y_i) + \lambda r(w) + \langle \eta, w \rangle$

Algorithm Output Perturbation

Require: Dataset $D$ and related parameters
Ensure: output private model $\hat{w}$

1. $w = \arg\min_{w} \frac{1}{m} \sum_{i} \ell(h(w, x_i), y_i) + \lambda r(w)$
2. Sample noise $\eta$ from Gamma distribution
3. $\hat{w} = w + \eta$

Objective vs Output Perturbation

- Sensitivity
  - Objective perturbation: easy to measure
  - Output Perturbation: hard to measure
- Performance
  - Output Perturbation has better performance
  [Chaudhuri2011]

Private Learning Framework

- It concerns about following problems:
  - How can we incorporate the differential privacy in learning algorithms?
  - How many records (sample complexity) are needed in a bounded error (accuracy)?
- PAC Learning

Probable Approximately Correct (PAC) Learning

Assumption: $X$ is sampled from a fixed but unknown probability distribution.
- Target: learner $L$ produce a hypothesis $h$ is a good approximation to $c$
- Requirement: as the number of $m$ of examples increases, the likelihood of error is small

PAC Learning

- Concept $C$ is PAC learnable if there exists an $L$ with following properties:
  - For every concept $c \in C$, for every distribution $D$ and for all $0 < \alpha < \frac{1}{2}$ and $0 < \delta < \frac{1}{2}$,
  - $L$ will with probability at least $1 - \delta$ output a hypothesis $h$ such that $err_D < \alpha$
  - in polynomial time.

\[ \Pr(err_D(c, h) \leq \alpha) \geq 1 - \delta \]
Private PAC Learning

• Every PAC learnable concept class can be learned privately, using a poly number of samples [Kasiviswanathan08].
  • Algorithm L privately PAC learns concept class C if:
    – Utility: Algorithm L PAC learns concept class C
    – Privacy: Algorithm L is $\epsilon$-differentially private


Sample Complexity

• The gap between private and non-private algorithms [Blum09].
  – Non-private: $O(VC(C))$
  – Private: $O(VC(C)\log |U|)$

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. STOC09

Sample Complexity

Sample Complexity Improvement

– Relax the search of hypothesis [Beimel13]:
  • choose a hypothesis that not in $C$, but pay the price of increasing workload on evaluation.
– Semi-supervised learning [Beimel14]:
  • starts with an unlabeled dataset and uses private data release to create a synthetic database to generate hypotheses, then uses the exponential mechanism to choose from these hypotheses using labeled examples.


Sample Complexity Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Sample Complexity</th>
<th>Privacy Level</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Exponential Mechanism to search $h$</td>
<td>$O(VC(C)\log</td>
<td>U</td>
<td>)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[Kasiviswanathan08]</td>
</tr>
<tr>
<td>Relax Privacy Level</td>
<td>From $\epsilon$ to $(\epsilon, \delta)$ or only preserve privacy of labels</td>
<td>$O(\log^{1/\delta}</td>
<td>U</td>
<td>)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[Steinke15]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[Chaudhuri11]</td>
</tr>
<tr>
<td>Relax search of Hypothesis</td>
<td>If $H \neq C$ and set a group of $H$ to privately search $h$</td>
<td>$O(\max\log</td>
<td>U</td>
<td>)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[Beimel11]</td>
</tr>
<tr>
<td>Semi-supervised learning</td>
<td>Use labeled data to search $h$</td>
<td>$O(d \cdot VC(C))$ on labeled and $O(VC(C))$ on unlabeled</td>
<td>$\epsilon$</td>
<td>[Beimel14]</td>
</tr>
</tbody>
</table>

2 How many records are needed in a bounded error?

Sample error + noise error

$m \geq \frac{1}{\alpha} \left( \ln |H| + \ln \frac{1}{\delta} \right)$
Summary on Private Learning Framework

**Advantages:**
- **Learning Private:** the gap between private and non-private learning process is quite narrow.

**Disadvantages:**
- **Inflexible:** only focuses on supervised learning problem.
- It is mainly developed in theory, but fewer applications

5. Applications on DP

**Application: Location Privacy**

Ping Xiong, Tianqing Zhu, Lei Pan, Wenjia Niu, and Gang Li. Privacy preserving in location data release: A differential privacy approach. PRICAI 2014

**Location Privacy**
- Photo location information
- IM application, such as WeChat
- Location Based Service: find a nearby restaurant, ATM, ...

Privacy Breach: LBS

Query: Nearest bus to the pub

- Special-aimed advertisement
- This person likes drinking

Collect information

Adversary

Privacy Breach: LBS

Daily trajectory

- This person’s health condition

Collect information

Adversary
Privacy Breach

Data Release:
- User may not want his/her exactly location to be identified.
- User may not want to be identified from a public dataset.

Privacy preserving

Location Preserving
- Release the coarse location, but not the exact one.
- Hide the user's location (not identity) from the service provider
  - Trajectory Preserve
    - Stream location

Traditional Location Privacy

- Location generalization
- Noise adding

Weakness

Generalization
- It is hard to control the utility
- Low level of privacy
- Noise Adding
  - Location information is associated with particular maps. When adding noise directly, some location will be changed to a weird place, such as in the middle of the sea.

Problem Definition

We consider $m$ locations, indexed by $1, 2, \ldots, m$, and model the profile of a user as the frequency of each location $u = (f_1, \ldots, f_m)$. We'd like to release each user's profile (locations) imprecisely.

- (a) **Privacy:** adversary cannot re-identify any user from the released dataset
- (b) **Utility:** accuracy and effectiveness (e.g., location-based browsing, filtering, or personalization)
What is the means of imprecisely:
- Locations
- Numbers of a certain location
• We can use the notion of Differential Privacy to protect both of them.

Challenge
• How to preserve differential privacy
  • How to define utility
  • How to maintain the Hierarchical Structure

Challenge I
Naive DP method: Only release statistical information.
  • It list all the locations, count the number and add noise to the statistical output.

Research Issue I
DP target: deleting a location or not will not change the user’s profile dramatically.
Randomize the profile
- In a user's profile, we apply the existing locations to replace the original ones.

Challenge II
- The sparse of the user's profile
- Adding noise to each bin will induce unnecessary locations.

Research Issue II
- Design a clustering-based randomization method to preserve privacy.
  - locations are clustered based on their distance.
  - For each location, randomization is performed only in the group.

Private locations clustering
- Adversary cannot figure out which cluster an location belongs to.

Challenge III
- How to maintain the Hierarchical Structure
  - Location natural has a hierarchical structure, but DP mechanism may destroy it.
Hierarchical Sensitivity

For each level, the sensitivity is the radius of clusters on that level. For example, to preserve the city level privacy, the sensitivity is measured by the radius of cities clusters.

Privacy Preserving Location Release

• Private Locations clustering
  - Adversary cannot figure out which cluster a location belongs to.
• Exponential Mechanism:
  - randomize the locations
  - Hide a user's locations: for a particular user, adversary cannot infer his locations.
• Laplace Mechanism:
  - perturb the weight.
  - Hide the distribution of locations in user's profile.

Algorithm 1: Private Location Release (PLocation) Algorithm

<table>
<thead>
<tr>
<th>Require</th>
<th>( D ), privacy parameter ( c, k ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensure</td>
<td>( \tilde{W} )</td>
</tr>
</tbody>
</table>

1. Divide privacy budget into \( c/2, e/4 \) and \( e/4 \).
2. Private Location Clustering: cluster locations into \( k \) groups in \( e/2 \) privacy budget; for each user \( u \), do
3. Cluster Weight Perturbation: add Laplace noise to the group weights with \( e/4 \) privacy budget;
4. Private Location Selection: Select locations according to the \( \tilde{W}(u) \) in \( e/4 \) privacy budget; end for
5. Output \( \tilde{W} \).

Summary on Location Privacy

• Location data are generally confronted with more serious privacy violations due to its semantic property.
  - Exponential mechanism can successfully hide the locations
  - Local Sensitivity and clustering can enhance the performance.

5. Future Directions
**Personalized Privacy**

Even $\epsilon$ indicates the privacy level of an algorithm, how to choose $\epsilon$ and adapt it according to individuals or queries still remain a challenge.

Hamid Ebadi, David Sands, and Gerardo Schneider. 
Differential privacy: Now it's getting personal. POPL 2015.

**Distributed Differential Privacy**

Most existing work is concerned with the centralized model, in which a trusted data curator holds the entire private dataset.

- But if a dataset is divided among multiple curators, how can they compute differentially private messages between themselves?


**Synthetic Dataset Release**

- A synthetic dataset can only perform a fixed learning task instead of an arbitrary one.
- How to efficiently release a synthetic dataset with arbitrary purposes remains a challenge.

**Privacy and Game Theory**

This direction involves with the purchase and sale of private data.

- a data analyst wishes to buy information from a population to estimate some information
- the owners of the private data experience some cost for their loss of privacy.


**Conclusion**

- Differential privacy is a promising privacy model that can provide provable privacy guarantee
- It also has potential development on various research communities such as data mining, machine learning, etc.
- Open problems:
  - Trade off utility and privacy
  - Applications adaptive mechanisms
  - Non-IID dataset

**References**

**Background and Preliminaries**

References

Differentially Private Data Release


Differentially Private Data Analysis


Applications