CS573 Data Privacy and Security

Differential Privacy

Li Xiong
Outline

• Differential Privacy Definition
• Basic techniques
• Composition theorems
Statistical Data Privacy

- Non-interactive vs interactive
- Privacy goal: individual is protected
- Utility goal: statistical information useful for analysis

Original Data → Privacy Mechanism → Statistics/Synthetic data

Queries

Data curator

Data analyst

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Age</th>
<th>Marital Status</th>
<th>Education</th>
<th>Occupation</th>
<th>Income (K)</th>
<th>Available</th>
<th>Number of Children</th>
<th>Available</th>
<th>At Risk</th>
<th>Available</th>
<th>At Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>56</td>
<td>1</td>
<td>Single</td>
<td>College</td>
<td>Student</td>
<td>1200</td>
<td>No</td>
<td>2</td>
<td>Yes</td>
<td>100</td>
<td>Yes</td>
<td>100</td>
</tr>
<tr>
<td>NA</td>
<td>39</td>
<td>2</td>
<td>Married</td>
<td>High School</td>
<td>Nurse</td>
<td>1050</td>
<td>Yes</td>
<td>2</td>
<td>No</td>
<td>100</td>
<td>No</td>
<td>100</td>
</tr>
<tr>
<td>NA</td>
<td>48</td>
<td>3</td>
<td>Divorced</td>
<td>College</td>
<td>Teacher</td>
<td>1500</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>100</td>
<td>Yes</td>
<td>100</td>
</tr>
<tr>
<td>NA</td>
<td>12</td>
<td>4</td>
<td>Widowed</td>
<td>High School</td>
<td>Artist</td>
<td>850</td>
<td>No</td>
<td>3</td>
<td>Yes</td>
<td>100</td>
<td>Yes</td>
<td>100</td>
</tr>
</tbody>
</table>
Recap

• Anonymization or de-identification (input perturbation)
  – Linkage attacks, homogeneity attacks

• Query auditing/restriction
  – Query denial is itself disclosive, computationally infeasible

• Summary statistics
  – Differencing attacks
Differential Privacy

• Promise: an individual will not be affected, adversely or otherwise, by allowing his/her data to be used in any study or analysis, no matter what other studies, datasets, or information sources, are available”

• Paradox: learning nothing about an individual while learning useful statistical information about a population
Differential Privacy

- Statistical outcome is indistinguishable regardless whether a particular user (record) is included in the data
Differential Privacy

- Statistical outcome is indistinguishable regardless whether a particular user (record) is included in the data
Differential privacy: an example

Original records

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>HIV+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank</td>
<td>42</td>
<td>Y</td>
</tr>
<tr>
<td>Bob</td>
<td>31</td>
<td>Y</td>
</tr>
<tr>
<td>Mary</td>
<td>28</td>
<td>Y</td>
</tr>
<tr>
<td>Dave</td>
<td>43</td>
<td>N</td>
</tr>
</tbody>
</table>

... ... ...

Original histogram

Perturbed histogram with differential privacy

3 if we add Alice
Differential Privacy

For every pair of inputs that differ in one row

\[ D_1 \quad \quad D_2 \]

For every output ...

\[ O \]

Adversary should not be able to distinguish between any \( D_1 \) and \( D_2 \) based on any \( O \)

\[
\log \left( \frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]} \right) < \varepsilon \quad (\varepsilon > 0)
\]
Why pairs of datasets *that differ in one row*?

For every pair of inputs that differ in one row

\[ D_1 \quad D_2 \]

For every output ...

\[ O \]

Simulate the presence or absence of a single record
Why *all* pairs of datasets ...?

For every pair of inputs that differ in one row

\[ D_1 \quad D_2 \]

For every output ...

\[ O \]

Guarantee holds no matter what the other records are.
Why *all* outputs?

Should not be able to distinguish whether input was $D_1$ or $D_2$ no matter what the output

Worst discrepancy in probabilities
Privacy Parameters

For every pair of inputs that differ in one row

\[ D_1 \quad D_2 \]

For every output ...

\[ O \]

\[ \Pr[A(D_1) = O] \leq e^\varepsilon \Pr[A(D_2) = O] \]

Controls the degree to which \( D_1 \) and \( D_2 \) can be distinguished. Smaller the \( \varepsilon \) more the privacy (and better the utility)
Differential Privacy: Some Qualitative Properties

- Protection against presence/participation of a single record
- Quantification of privacy loss
- Composition
- Post-processing
Differential Privacy: Additional Remarks

• Correlations between records
• Granularity of a single record (difference for neighboring database)
  – Group privacy
  – Graph database (eg social networks): node vs edge
  – Movie rating database: user vs event (movie)
Outline

• Differential Privacy Definition
• Basic techniques
  – Laplace mechanism
  – Exponential mechanism
  – Random Response
• Composition theorems
Differential Privacy

For every pair of inputs that differ in one row

\[ D_1 \quad D_2 \]

Adversary should not be able to distinguish between any \( D_1 \) and \( D_2 \) based on any \( O \)

\[
\log \left( \frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]} \right) < \varepsilon \quad (\varepsilon > 0)
\]

[Dwork ICALP 2006]
Can deterministic algorithms satisfy differential privacy?
Non trivial deterministic algorithms do not satisfy differential privacy

Space of all inputs

Space of all outputs
(at least 2 distinct outputs)
Non-trivial deterministic algorithms do not satisfy differential privacy

Each input mapped to a distinct output.
There exist two inputs that differ in one entry mapped to different outputs.
Output Randomization

- Add noise to answers such that:
  - Each answer does not leak too much information about the database.
  - Noisy answers are close to the original answers.
Laplace Mechanism

Database → True answer → q(D) → q(D) + η → Researcher

Laplace Distribution – Lap(S/ε)

Tutorial: Differential Privacy in the Wild

Module 2
Laplace Distribution

- PDF: 
  \[ f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \]
- Denoted as Lap(b) when \( \mu=0 \)
- Mean \( \mu \)
- Variance \( 2b^2 \)
How much noise for privacy?

[Sensitivity] Consider a query $q: I \rightarrow R$. $S(q)$ is the smallest number s.t. for any neighboring tables $D$, $D'$,

$$|q(D) - q(D')| \leq S(q)$$

[Theorem]: If sensitivity of the query is $S$, then the algorithm $A(D) = q(D) + \text{Lap}(S(q)/\varepsilon)$ guarantees $\varepsilon$-differential privacy.
Example: COUNT query

- Number of people having disease
- Sensitivity = 1

Solution: $3 + \eta$, where $\eta$ is drawn from Lap($1/\varepsilon$)
  - Mean = $0$
  - Variance = $2/\varepsilon^2$
Example: SUM query

- Suppose all values $x$ are in $[a,b]$  
- Sensitivity = $b$
Privacy of Laplace Mechanism

• Consider neighboring databases D and D’
• Consider some output O

\[
\frac{\Pr [A(D) = O]}{\Pr [A(D') = O]} = \frac{\Pr [q(D) + \eta = O]}{\Pr [q(D') + \eta = O]}
\]

\[
= \frac{e^{-|O-q(D)|/\lambda}}{e^{-|O-q(D')|/\lambda}}
\]

\[
\leq e^{q(D) - q(D')}/\lambda \leq e^{S(q)/\lambda} = e^\varepsilon
\]
Utility of Laplace Mechanism

• Laplace mechanism works for any function that returns a real number

• Error: $E(\text{true answer} - \text{noisy answer})^2$
  $$= \text{Var}(\text{Lap}(S(q)/\epsilon))$$
  $$= 2*S(q)^2 / \epsilon^2$$

• Error bound: very unlikely the result has an error greater than a factor (Roth book Theorem 3.8)
Outline

• Differential Privacy Definition
• Basic techniques
  – Laplace mechanism
  – Exponential mechanism
  – Random Response
• Composition theorems
Exponential Mechanism

• For functions that do not return a real number ...
  – “what is the most common nationality in this room”: Chinese/Indian/American...

• When perturbation leads to invalid outputs ...
  – To ensure integrality/non-negativity of output
Exponential Mechanism

Consider some function $f$ (can be deterministic or probabilistic):

How to construct a differentially private version of $f$?
Exponential Mechanism

**Theorem** For a database $D$, output space $R$ and a utility score function $u : D \times R \to R$, the algorithm $A$

$$\Pr[A(D) = r] \propto \exp \left( \frac{\varepsilon \times u(D, r)}{2\Delta u} \right)$$

satisfies $\varepsilon$-differential privacy, where $\Delta u$ is the sensitivity of the utility score function

$$\Delta u = \max_{r \in R, D, D'} |u(D, r) - u(D', r)|$$
Example: Exponential Mechanism

• Scoring/utility function \( w: Inputs \times Outputs \rightarrow R \)

• \( D \): nationalities of a set of people

• \( f(D) \): most frequent nationality in \( D \)

• \( u(D, O) = \#(D, O) \) the number of people with nationality \( O \)
Privacy of Exponential Mechanism

The exponential mechanism outputs an element $r$ with probability

$$\Pr[A(D) = r] \propto \exp \left( \frac{\epsilon \times u(D, r)}{2\Delta u} \right)$$

$$\Delta u = \max_{r, D, D'} |u(D, r) - u(D', r)|$$

Approximately $\Pr[A(D) = r] / \Pr[A(D') = r] \leq \epsilon$

(Exact proof with normalization factor: Roth Book page 39)
Privacy of Exponential Mechanism

\[ \frac{\Pr[\mathcal{M}_E(x, u, \mathcal{R}) = r]}{\Pr[\mathcal{M}_E(y, u, \mathcal{R}) = r]} = \frac{\left( \frac{\exp\left( \frac{\varepsilon u(x, r)}{2\Delta u} \right)}{\sum_{r' \in \mathcal{R}} \exp\left( \frac{\varepsilon u(x, r')}{2\Delta u} \right)} \right)}{\left( \frac{\exp\left( \frac{\varepsilon u(y, r)}{2\Delta u} \right)}{\sum_{r' \in \mathcal{R}} \exp\left( \frac{\varepsilon u(y, r')}{2\Delta u} \right)} \right)} \\
= \left( \frac{\exp\left( \frac{\varepsilon u(x, r)}{2\Delta u} \right)}{\exp\left( \frac{\varepsilon u(y, r)}{2\Delta u} \right)} \right) \cdot \left( \frac{\sum_{r' \in \mathcal{R}} \exp\left( \frac{\varepsilon u(y, r')}{2\Delta u} \right)}{\sum_{r' \in \mathcal{R}} \exp\left( \frac{\varepsilon u(x, r')}{2\Delta u} \right)} \right) \\
= \exp\left( \frac{\varepsilon (u(x, r') - u(y, r'))}{2\Delta u} \right) \cdot \left( \frac{\sum_{r' \in \mathcal{R}} \exp\left( \frac{\varepsilon u(y, r')}{2\Delta u} \right)}{\sum_{r' \in \mathcal{R}} \exp\left( \frac{\varepsilon u(x, r')}{2\Delta u} \right)} \right) \\
\leq \exp\left( \frac{\varepsilon}{2} \right) \cdot \exp\left( \frac{\varepsilon}{2} \right) \cdot \left( \frac{\sum_{r' \in \mathcal{R}} \exp\left( \frac{\varepsilon u(x, r')}{2\Delta u} \right)}{\sum_{r' \in \mathcal{R}} \exp\left( \frac{\varepsilon u(x, r')}{2\Delta u} \right)} \right) \\
= \exp(\varepsilon). \]
Utility of Exponential Mechanism

• Can give strong utility guarantees, as it discounts outcomes exponentially based on utility score
• Highly unlikely that returned element $r$ has a utility score inferior to $\max_r u(D,r)$ by an additive factor of $O((\Delta u/\varepsilon) \log |\mathcal{R}|)$. (Theorem 3.11 Roth book)
Outline

• Differential Privacy Definition
• Basic techniques
  – Laplace mechanism
  – Exponential mechanism
  – Random Response
• Composition theorems
Randomized Response (a.k.a. local randomization)

With probability $p$, Report true value

With probability $1-p$, Report flipped value
Differential Privacy Analysis

• Consider 2 databases $D$, $D'$ (of size $M$) that differ in the $j^{th}$ value
  – $D[j] \neq D'[j]$. But, $D[i] = D'[i]$, for all $i \neq j$

• Consider some output $O$

\[
\frac{P(D \rightarrow O)}{P(D' \rightarrow O)} \leq e^\varepsilon \iff \frac{1}{1 + e^\varepsilon} < p < \frac{e^\varepsilon}{1 + e^\varepsilon}
\]
Utility Analysis

- Suppose $n_1$ out of $n$ people replied “yes”, and rest said “no”
- What is the best estimate for $\pi = \text{fraction of people with disease} = Y$?

$$\hat{\pi} = \frac{n_1/n - (1-p)}{2p-1}$$

- $E(\hat{\pi}) = \pi$
- $\text{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{1}{n(16(p-0.5)^2 - 0.25)}$

  **Sampling**  
  **Variance due to coin flips**
Laplace Mechanism vs Randomized Response

Privacy

• Provide the same $\varepsilon$-differential privacy guarantee

• Laplace mechanism assumes data collector is trusted

• Randomized Response does not require data collector to be trusted
  – Also called a *Local* Algorithm, since each record is perturbed
Laplace Mechanism vs Randomized Response

Utility

• Suppose a database with N records where μN records have disease = Y.

• Query: # rows with Disease=Y

• Std dev of Laplace mechanism answer: O(1/ε)

• Std dev of Randomized Response answer: O(√N)
Outline

• Differential Privacy

• Basic Algorithms
  – Laplace
  – Exponential Mechanism
  – Randomized Response

• Composition Theorems
Why Composition?

• Reasoning about privacy of a complex algorithm is hard.

• Helps software design
  – If building blocks are proven to be private, it would be easy to reason about privacy of a complex algorithm built entirely using these building blocks.
A bound on the number of queries

• In order to ensure utility, a statistical database must leak some information about each individual

• We can only hope to bound the amount of disclosure

• Hence, there is a limit on number of queries that can be answered
Composition theorems

Sequential composition
\[ \sum_i \varepsilon_i \text{– differential privacy} \]

Parallel composition
\[ \max(\varepsilon_i) \text{– differential privacy} \]
Sequential Composition

• If $M_1, M_2, \ldots, M_k$ are algorithms that access a private database $D$ such that each $M_i$ satisfies $\varepsilon_i$-differential privacy, then the combination of their outputs satisfies $\varepsilon$-differential privacy with $\varepsilon = \varepsilon_1 + \ldots + \varepsilon_k$
Parallel Composition

• If $M_1, M_2, ..., M_k$ are algorithms that access disjoint databases $D_1, D_2, ..., D_k$ such that each $M_i$ satisfies $\varepsilon_i$-differential privacy,

then the combination of their outputs satisfies $\varepsilon$-differential privacy with $\varepsilon = \max\{\varepsilon_1, ..., \varepsilon_k\}$
Postprocessing

• If $M_1$ is an $\varepsilon$ differentially private algorithm that accesses a private database $D$, then outputting $M_2(M_1(D))$ also satisfies $\varepsilon$-differential privacy.
Summary

• Differential privacy ensure an attacker can’t infer the presence or absence of a single record in the input based on any output.

• Building blocks
  – Laplace, exponential mechanism, (local) randomized response

• Composition rules help build complex algorithms using building blocks
Case Study: K-means Clustering

Original unclustered data

Clustered data
Kmeans

• Partition a set of points $x_1, x_2, \ldots, x_n$ into $k$ clusters $S_1, S_2, \ldots, S_k$ such that the following is minimized:

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} \| x_j - \mu_i \|_2^2$$

Mean of the cluster $S_i$
Kmeans

Algorithm:

• Initialize a set of k centers
• Repeat
  Assign each point to its nearest center
  Recompute the set of centers
  Until convergence ...

• Output final set of k centers
Differentially Private Kmeans

- Suppose we fix the number of iterations to $T$
- In each iteration (given a set of centers):
  1. Assign the points to the new center to form clusters
  2. Noisily compute the size of each cluster
  3. Compute noisy sums of points in each cluster
Differentially Private Kmeans

• Suppose we fix the number of iterations to T
  
  Each iteration uses $\varepsilon/T$ privacy budget, total privacy loss is $\varepsilon$

• In each iteration (given a set of centers):
  
  1. Assign the points to the new center to form clusters
  
  2. Noisily compute the size of each cluster
  
  3. Compute noisy sums of points in each cluster
Differentially Private Kmeans

Exercise: Which of these steps expends privacy budget?

- In each iteration (given a set of centers):
  
  1. Assign the points to the new center to form clusters
  
  2. Noisily compute the size of each cluster

  3. Compute noisy sums of points in each cluster
Differentially Private Kmeans

Exercise: Which of these steps expends privacy budget?

- In each iteration (given a set of centers):
  1. Assign the points to the new center to form clusters  **NO**
  2. Noisily compute the size of each cluster  **YES**
  3. Compute noisy sums of points in each cluster  **YES**
Differentially Private Kmeans

What is the sensitivity?

- In each iteration (given a set of centers):
  1. Assign the points to the new center to form clusters
  2. Noisily compute the size of each cluster
  3. Compute noisy sums of points in each cluster
Differentially Private Kmeans

• Suppose we fix the number of iterations to $T$

  Each iteration uses $\frac{\varepsilon}{T}$ privacy budget, total privacy loss is $\varepsilon$

• In each iteration (given a set of centers):
  1. Assign the points to the new center to form clusters
  2. Noisily compute the size of each cluster
     \[ \text{Laplace}(2T/\varepsilon) \]
  3. Compute noisy sums of points in each cluster
     \[ \text{Laplace}(2T |\text{dom}| /\varepsilon) \]
• Even though we noisily compute centers, Laplace kmeans can distinguish clusters that are far apart.

• Since we add noise to the sums with sensitivity proportional to $|\text{dom}|$, Laplace k-means can’t distinguish small clusters that are close by.
Privacy as Constrained Optimization

• Three axes
  – Privacy
  – Error
  – Queries that can be answered

• E.g.: Given a fixed set of queries and privacy budget $\varepsilon$, what is the minimum error that can be achieved?

• E.g.: Given a task and privacy budget $\varepsilon$, how to design a set of queries (functions) and allocate the budget such that the error is minimized?
References


[DN03] Dinur, Nissim, “Revealing information while preserving privacy”, PODS 2003

[BDMN05] Blum, Dwork, McSherry, Nissim, “Practical privacy: the SuLQ framework”, PODS 2005


[DMNS06] Dwork, McSherry, Nissim, Smith, “Calibrating noise to sensitivity in private data analysis”, TCC 2006