PRIVACY IN ACTION
APPLYING DIFFERENTIAL PRIVACY IN DATA MINING

Jaewoo Lee and Chris Clifton
Department of Computer Science
AGENDA

• Motivation
• Background
• Challenges
  • Choosing $\epsilon$
  • Answering infinite number of queries?
• Conclusion
Problem statement

- Learn a dataset while preserving privacy
- Privacy VS Utility
- Re-identification
Data Privacy

• What is the proper measure of privacy risk?
• Context
• “Individually Identifiable Information”
  – EC95/46, HIPAA, …
• Re-identification risk
  – Given (Zip code + DOB + Gender)
  – 87% of US population can be uniquely identified*

* L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population
Safe Harbor

- removal of 18 **identifiers** (e.g., name, ssn, etc.)
- all elements of date (except year) & **age > 89**
- geographic unit (**< 20,000 people**)  

Qualified Statistician Determination

- use statistical and scientific principles
- ensure the **risk is very small** that the information could be used, by the anticipated recipient, **to identify the subject of the information**
HIPAA safe harbor implicitly enforces

- bound on the probability of identifying individuals
- 1.7% of US population is male & age $\geq 85$
- knowing the age, gender and address allows to limit them to one of 68 people

- $\Pr[I(i) \in DB \mid Release] \leq \frac{1}{68} \approx 1.5\%$

Identifiability risk

- degree to which data can be traced to an individual associated with it
• Motivation
• **Background**
• Challenges
  • Choosing $\epsilon$
  • Answering infinite number of queries?
• Conclusion
Differential privacy

- For $S \subseteq \text{Range}(f)$, an $\epsilon$-differentially private mechanism $M$ satisfies

$$\frac{\Pr[M_f(D_1) \in S]}{\Pr[M_f(D_2) \in S]} \leq e^\epsilon$$

where $D_1$ and $D_2$ differ on at most one element

- $M_f(D) = f(D) + Y$ where $Y \sim \text{Lap}(\lambda)$
Differential Privacy

Sensitivity

$$\Delta f = \max_{x, x' \in U} |f(x) - f(x')|$$

Achieving Differential Privacy

- setting $$\lambda = \frac{\Delta f}{\epsilon}$$ achieves $$\epsilon$$-differential privacy
- smaller $$\epsilon$$ means more privacy

$$\Pr[M_f(D_1) \in S] \leq e^{-\frac{|f(D_1) - f(D_2)|}{\lambda}} \leq e^\epsilon$$
AGENDA

• Motivation
• Background
• **Challenges**
  • Choosing $\epsilon$
  • Answering infinite number of queries?
• Conclusion
CHALLENGES IN APPLYING DP

Difficulties in practice

• How to set $\epsilon$
• Calculation of sensitivity
• Unbounded output range
• Answering many queries
• Complex data (e.g., graph, text)
• Budget Allocation
AGENDA

• Motivation
• Background
• Challenges
  • Choosing $\varepsilon$
  • Answering infinite number of queries?
• Conclusion
CHALLENGES IN APPLYING DP

How to set $\varepsilon$

- Open problem
- No rule of thumb or guidelines
- Physical meaning of $\varepsilon$
- indistinguishable $\neq$ unidentifiable*

* G. Cormode, Personal Privacy VS Population Privacy, KDD 2011
Example

- Release \( \text{MEAN(AD)} \)
- \( X' = \{\text{Chris, Kelly, Pat}\} \)
- Absence in \( X' \rightarrow \) academic probation
- Privacy goal \( G \)
  
  \[
  \forall i \in X, \quad \Pr[i \notin X'|\text{MEAN(AD)} = R] \leq \frac{1}{3}
  \]

- Given \( G \), what is the proper value of \( \epsilon \)?

<table>
<thead>
<tr>
<th>Name</th>
<th>School Year (SY)</th>
<th>Absence Day (AD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chris</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Kelly</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Pat</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Terry</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 1. Student database \( X \)
In many literature, the value of $\epsilon$ is

- chosen arbitrarily, or
- assumed to be given

Assume $\epsilon = 2$

- $\Delta f = \left| \frac{1+2+10}{3} - \frac{1+2}{2} \right| = \frac{17}{6}$
- $R = f(X) + \text{Lap}(\Delta f/\epsilon) = 2 + \text{Lap}(17/12)$ achieves differential privacy
- However, $\Pr[Terry \notin X' | R=2.2013] = 0.6180$
- adversary can infer Terry’s absence in $X'$ with high confidence
- The value of $\epsilon$ should be chosen carefully
What if $f = \text{MEAN(SY)}$?

- $\Delta f = \left| \frac{1+2+4}{3} - \frac{1+2}{2} \right| = \frac{5}{6}$

- Given $R = 2.2013$ and $\varepsilon = 2$,

- $\Pr[Terry \notin X' \mid R = 2.2013] = 0.3390$

- Even for the same value of $\varepsilon$ and $R$, the probability of identifying individuals varies with the domain
Differential Identifiability [KDD 2012]

- Provide a privacy definition that directly relates to legal privacy requirements
- As strong as differential privacy
- Intuitive parameterization
- Limit the adversary’s confidence $\rho$
- Quantifying risk of identification
FRAMEWORK

HOW DOES IT WORK?

\[ f = \text{mean} \]

\[ R = 2 + \]

\[ \Pr[\text{user} \in DB| R] = 5 \]

U

\[ U = 1 \]

\[ U = 2 \]

\[ U = 3 \]

\[ U = 12 \]
ADVERSARY MODEL

ASSUMPTIONS ABOUT THE ADVERSARY

What the adversary knows

• \( U \): people in the universe & their data
• \( D' \): \(|D|\)-1 rows
• \( \mathcal{M} \): privacy mechanism (+ noise distribution)

What the adversary does

• generates a set of possible worlds \( \Psi \)
• \( \Psi = \{D' \cup \{i\} | i \in U\} \)
• \( \forall \omega \in \Psi, \Pr[\omega = D | R = \mathcal{M}_f(D)] \)
How much noise is enough?

- enough to hide the impact of any individuals
- **largest contribution** that an individual can make to the output

Individual contribution

- \( C_f(i) = \max_{\omega \in \psi_i, \omega' \in \psi_i} |f(\omega) - f(\omega')| \)

Sensitive range

- \( S(f) = \max_{\omega, \omega' \in \Psi} |f(\omega) - f(\omega')| \)
- Note that \( S(f) = \max_i C_f(i) \)
Setting

- \( f = \text{mean} \)
- \( D' = \{\text{Chris, Jaewoo, Koray} \} \)
- \( D = D' \cup \{\text{Winston} \} \)
- \( R = 5.041 \)
- \( \rho = 0.4 \)

- \( \lambda = \frac{S(f)}{\ln(m-1)\rho} = 5.21 \)
Possible worlds

- $\omega_1 = \{1, 2, 3, 4\}$
- $\omega_2 = \{1, 2, 3, 5\}$
- $\omega_3 = \{1, 2, 3, 10\}$
- $\Pr[M_f(\omega_1) = R] = 0.0589$
- $\Pr[M_f(\omega_2) = R] = 0.0618$
- $\Pr[M_f(\omega_3) = R] = 0.0785$

- $Pr[\omega_3 = D|R] = \frac{0.0785}{0.0589 + 0.0618 + 0.0785} = 0.3941$
Definition

\[ \forall D' = D - \{i\}, \forall i \in U - D' \]

\[ \Pr[I(i) \in I_D | M_f(D) = R, D'] \leq \rho \]

- Limit the probability of identifying an individual in the database to \( \rho \)
Risk of disclosure

\[ \Gamma(i) = \Pr[I(i) \in I_D | \mathcal{M}_f(D) = R, D'] \]

- Compute upper bound on \( \Gamma(i) \), \( \max_i \Gamma(i) \)
- Enforce \( \max_i \Gamma(i) \leq \rho \)
\[ \Gamma(i) = \Pr[I(i) \in I_D \mid \mathcal{M}_f(D) = R, D'] = \frac{\Pr[D = D' \cup \{i\}] \Pr[\mathcal{M}_f(D' \cup \{i\}) = R]}{\Pr[\mathcal{M}_f(D) = R]} \]

\[ = \frac{\Pr[D = D' \cup \{i\}] \Pr[\mathcal{M}_f(D' \cup \{i\}) = R]}{\sum_{\omega \in \Psi} \Pr[\omega] \Pr[\mathcal{M}_f(\omega) = R]} \]

\[ = \frac{\Pr[\mathcal{M}_f(D' \cup \{i\}) = R]}{\sum_{\omega \in \Psi} \Pr[\mathcal{M}_f(\omega) = R]} \]

\[ = e^{-\frac{|f(D) - R|}{\lambda}} \]

\[ = \sum_{\omega \in \Psi} e^{-\frac{|f(\omega) - R|}{\lambda}} \]

\[ = e^{-\frac{|f(D) - R|}{\lambda}} + e^{-\frac{|f(\omega_1) - R|}{\lambda}} + \cdots + e^{-\frac{|f(\omega_{m-1}) - R|}{\lambda}} \]
UPPER BOUND ON $\Gamma (1)$

DERIVATION OF SCALE FACTOR

\[
\leq \frac{1}{1 + e^{-\frac{|f(\omega_1) - f(\omega_j)|}{\lambda}} + \cdots + e^{-\frac{|f(\omega_{m-1}) - f(\omega_j)|}{\lambda}}}
\]

\[
= \frac{1}{1 + (m - 1) e^{-\frac{S(f)}{\lambda}}} \leq \rho
\]

\[
\lambda \geq \frac{S(f)}{\ln \frac{(m - 1)\rho}{1 - \rho}}
\]
Claim: Any $\epsilon$-differential private mechanism satisfies

$$\frac{1}{1 + (m-1)e^{-\epsilon}}$$ - differential identifiability

$$\Gamma(i) = \frac{\Pr[\mathcal{M}_f(D' \cup \{i\}) = R]}{\sum_{\omega \in \Psi} \Pr[\mathcal{M}_f(\omega) = R]}$$

$$= \frac{\Pr[\mathcal{M}_f(D) = R]}{\Pr[\mathcal{M}_f(D) = R] + \sum_{j=1}^{m-1} \Pr[\mathcal{M}_f(\omega_j) = R]}$$

If $\mathcal{M}$ is $\epsilon$-differentially private

$$\forall j, \Pr[\mathcal{M}_f(\omega_j) = R] \geq e^{-\epsilon} \Pr[\mathcal{M}_f(D) = R]$$

$$\leq \frac{1}{1 + (m-1)e^{-\epsilon}}$$
## Adult dataset

- UCI Machine learning repository
- 1994 US Census data,
- Training + Test data, 48,842 individuals

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Max.</th>
<th>Min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age(AG)</td>
<td>90</td>
<td>17</td>
</tr>
<tr>
<td>Education-Number (EN)</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Capital-gain (CG)</td>
<td>99999</td>
<td>0</td>
</tr>
<tr>
<td>Capital-loss (CL)</td>
<td>4356</td>
<td>0</td>
</tr>
<tr>
<td>Hours-per-week (HW)</td>
<td>99</td>
<td>1</td>
</tr>
</tbody>
</table>
Differential Privacy

(a) Capital-gain

(b) Capital-loss

Differential Identifiability

(a) Capital-gain

(b) Capital-loss
• Motivation
• Background
• Challenges
  • Choosing $\epsilon$
  • Answering infinite number of queries?
• Conclusion
• Counting Query based
  – Candidate generation: no DB access
  – Support verification: for each $X \in C_k$, ask $\text{count}(X)$
  – Each access spends a small privacy budget
• **PrivBasis [VLDB 2012]**
  - treat transactional DB as tabular data
  - project the data into lower dimensional space

• **SmartTruncation [VLDB 2012]**
  - tries to minimize the sensitivity
  - truncate transaction

Search space reduction causes information loss
### Frequent Itemset Discovery

- **Threshold query**: is the support of itemset $X$ greater than $\tau$?
- Find frequent itemsets without worrying about their supports
- Can we provide **exponentially many** differentially private answers using a small privacy budget?

### Support Derivation

- Given $\mathcal{L} = \{X_1, ..., X_{|\mathcal{L}|}\}$, build an FP-tree
- Know which itemsets are frequent will help us in building a compact tree
- **Sensitivity** of insert operation?
- Imposing consistency
Algorithm

- Compute $\hat{t} = t + \text{noise}$
- Compute $\hat{X} = \sigma(X) + \text{noise}$
- If $\hat{X} \geq \hat{t}$ (X is frequent) then, output 1
- Otherwise (X is infrequent), output 0
- The output of algorithm is a vector $v = (v_1, v_2, ..., v_t)$, not the count of each candidate itemset
WHAT TO PERTURB?

Perturbing support only

Noise changes the answer!

\[
\frac{\Pr[V_D = 1]}{\Pr[V_{D'} = 1]} = \infty
\]
CONCLUSION

• Many data mining tasks are proven to be done in a privacy preserving manner with small loss in accuracy
• Differential privacy is promising but needs care when it comes to its application