PrivateClean: Data Cleaning and Differential Privacy

Data Cleaning & Integration
CompSci 590.01 Spring 2017

Presented by Xi He

Some contents were based on Sanjay Krishnan et al.’s SIGMOD 2016 slides & Machanavajjhala et al.’s VLDB 2016 DP Tutorial slides
Our world is increasingly data driven

Source (http://www.agencypja.com/site/assets/files/1826/marketingdata-1.jpg)
Aggregated Personal Data is invaluable

Dirty Privacy

Advertising

Genome Wide Association Studies

Human Mobility analysis

Source (esri.com)
Outline

• A brief tour to differential privacy

• PrivateClean: data cleaning + differential privacy
  – Private relation generation
  – Query processing
    • Correction for randomized response
    • Cleaning for dirty data
  – Evaluation

• Discussion
Personal data is ... well ... personal!

- Age
- Income
- Address
- Likes/Dislikes
- Sexual Orientation
- Medical History

Redlining

Discrimination

Physical/Financial Harm

Source (time.com)
Aggregated Personal Data …

… is made publicly available in many forms.

De-identified records
(e.g., medical)

Statistics
(e.g., demographic)

Predictive models
(e.g., advertising)
That’s fine … I am anonymous!

Source (http://xkcd.org/834/)
Anonymity is not enough …

A Face Is Exposed for AOL Searcher No. 4417749
By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Why 'Anonymous' Data Sometimes Isn't
By Bruce Schneier
12.13.07
Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

"Anonymous" Genomes Identified
The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.
By Dan Cossins | May 3, 2013
... and predictive models can breach privacy too

Marketers Can Glean Private Data on Facebook

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing
Need data analysis algorithms that can mine aggregated personal data with provable guarantees of privacy for individuals.

This is the goal of Differential Privacy.
Differential Privacy (DP)

Dwork et al. 2006

- Randomized mechanism $M$ satisfies $\epsilon$-DP if for all $D, D'$ that differ in one row, and for every output $O$
  \[
  \Pr[M(D) = O] \leq \exp(\epsilon) \Pr[M(D') = O]
  \]
Differential Privacy (DP)

- Randomized mechanism $M$ satisfies $\epsilon$-DP if for all $D, D'$ that differ in one row, and for every output $O$
  $$\Pr[M(D) = O] \leq \exp(\epsilon) \Pr[M(D') = O]$$
  - $D, D'$: Simulate the different values of a single record
  - Smaller the $\epsilon$ more the privacy (and better the utility)

Dwork et al. 2006
Basic Algorithms

• Output Perturbation to query $q(D)$
  – Laplace Mechanism: add noise $\sim \text{Lap} \left( \frac{\Delta q}{\epsilon} \right)$
  – Assume data collector is trusted

\[ \text{Laplace Distribution} \]

Output $O = q(D) + \text{noise}$
Basic Algorithms

• Input Perturbation
  – Randomized response (RR): each record is perturbed
    • By participant if data collector is not trusted
    • By data collector o.w.
  – RRCorrection is required for $q(D_{private})$
Outline

• A brief tour to differential privacy

• PrivacyClean: data cleaning + differential privacy
  – Private relation generation
  – Query processing
    • Correction for randomized response
    • Cleaning for dirty data
  – Evaluation

• Discussion
Overview of PrivateClean

Sanjay et al. 2016

• Part I: Private relation creation:
  – Dirty database $D_{\text{dirty}}$ (Categorical/ Numerical attributes)
  – Generate private dirty database $D_{\text{dirty, private}}$ with
    • RR for Categorical attributes (Cat)
    • Local version of Laplace Mechanism for Numerical attributes (Num)
Overview of PrivateClean

Sanjay et al. 2016

- Party II: Query processing:
  - \( q \) : \( \text{SELECT SUM/COUNT/AVG}(\text{Num}) \text{ Where } \text{Cond}(\text{Cat}) \)
  - local cleaner with deterministic UDF over Cat attributes in each row (Merge/Transform/Extract)

Dirty database \( D_{\text{dirty}} \)

Private dirty database \( D_{\text{dirty, private}} \)

More specific

Output based on \( D_{\text{clean, private}} \)

Private clean database \( D_{\text{clean, private}} \)
Overview of PrivateClean

Sanjay et al. 2016

- Party II: Query processing:
  - $q$: SELECT SUM/COUNT/AVG(Num) Where Cond(Cat)
  - $q_{dirty}$: local cleaner with deterministic UDF over Num attributes in each row (Merge/Transform/Extract)
  - $RR_{correction}$: correction for randomized response
Private Relation Creation

- Example dataset $D_{dirty}$
  - $q(D_{clean})$: SELECT COUNT(*) WHERE Major = EECS

<table>
<thead>
<tr>
<th>StudentId</th>
<th>Major</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IEOR</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>EECS</td>
<td>1.6</td>
</tr>
<tr>
<td>3</td>
<td>Electrical Engineering and Computer Sciences</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Nuclear Engineering</td>
<td>2.1</td>
</tr>
</tbody>
</table>
RR for Categorical Attribute $A_i$

- $\text{dom}(A_i)$: distinct values for $A_i$ in $D_{\text{dirty}}$
- For record $r$, let $r[i]$ be the value of $A_i$, RR outputs
  - the true value: $r[i]$ with prob. $1-p_i$
  - a value in $\text{dom}(A_i)$ uniformly, i.e. $U(\text{dom}(A_i))$ with prob. $p_i$
- Given $p_i$, RR satisfies $\epsilon_i$-DP, where $\epsilon_i = \ln\left(\frac{2}{p_i} - 1\right)$
  - Recall $\epsilon$-DP: $\Pr[M(D) = 0] \leq \exp(\epsilon) \Pr[M(D') = 0]$
  - $\exp(-\epsilon_i) \leq \frac{\Pr[\text{RR}(a) = a]}{\Pr[\text{RR}(a') = a]} = \frac{1-p_i+p_i/|\text{dom}(A_i)|}{p_i/|\text{dom}(A_i)|} \leq \exp(\epsilon_i)$
  - Tightest $\epsilon_i = \ln\left(\frac{|\text{dom}(A_i)|}{p_i} - |\text{dom}(A_i)| + 1\right) \leq \ln\left(\frac{2}{p_i} - 1\right)"
Lap Noise for Numerical Attribute $A_i$

- For record $r$, let $r[i]$ be the value of $A_i$, Lap outputs $r[i] + \eta$, where $\eta \sim Lap\left(\frac{\Delta_i}{\epsilon_i}\right)$,
  - $\Delta_i$ is the domain size of $A_i$ in $D_{dirty}$
  - Mean: 0, Variance: $2 \left(\frac{\Delta_i}{\epsilon_i}\right)^2$

- Satisfy $\epsilon_i$-DP
- However, this add a lot of noises for queries such as `SELECT SUM() Where Cond(Cat)`, unlike output perturbation method
DP Dirty Database

• Generates $D_{\text{dirty, private}}$

<table>
<thead>
<tr>
<th>StudentId</th>
<th>Major</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IEOR</td>
<td>1.82321</td>
</tr>
<tr>
<td>2</td>
<td>Nuclear Engineering</td>
<td>2.12212</td>
</tr>
<tr>
<td>3</td>
<td>Electrical Engineering and Computer Sciences</td>
<td>3.2412</td>
</tr>
<tr>
<td>4</td>
<td>EECS</td>
<td>2.88721</td>
</tr>
</tbody>
</table>
Overall Privacy Guarantee

• Composition: if each $A_i$ is assigned with $\epsilon_i$, the released $D_{dirty, private}$ satisfies $\sum_{A_i} \epsilon_i$ -DP.
Overall Privacy Guarantee

- Composition: if each $A_i$ is assigned with $\epsilon_i$-DP, the released $D_{\text{dirty, private}}$ satisfies $\sum_{A_i} \epsilon_i$-DP.

- Post-processing: queries/cleaning over $D_{\text{dirty, private}}$ without looking at of $D_{\text{dirty}}$ still satisfy $\sum_{A_i} \epsilon_i$-DP.

Output $O = RR\text{correction}(q_{\text{dirty}}(D_{\text{dirty, private}}))$
Outline

• A brief tour to differential privacy

• PrivateClean: data cleaning + differential privacy
  – Private relation generation
  – Query processing
    • Correction for randomized response
    • Cleaning for dirty data
  – Evaluation

• Discussion
Correction of RR

• First, we consider private clean database $D_{private}$

• $q$: SELECT **COUNT**(*) Where $A_i \in \{a_1, \ldots, a_L\}$
  
  $- \ c$: true output for $q$, $S$: size of $D_{private}$
  
  $- \ E[q(D_{private})] = c \cdot p_i^+ + (S-c) \cdot p_i^-$ where

  $- p_i^+$: Pr[including a record $r$ that satisfies the predicate] = $(1 - p_i) + p_i \cdot l/|dom(A_i)|$

  $- p_i^-$: Pr[The prob. of including a record $r$ that does not satisfy the predicate] = $p_i \cdot l/|dom(A_i)|$

  $- \text{estimate } c \text{ by } (q(D_{private})-S \cdot p_i^-)/(p_i^+ - p_i^-)$
Correction of RR

• First, we consider private clean database $D_{\text{private}}$

• $q$: SELECT $\text{COUNT}(\ast)$ Where $A_i \in \{a_1, \ldots, a_l\}$
  - estimate $c$ by $(q(D_{\text{private}}) - S p^-_i)/(p^+_i - p^-_i)$

• $q$: SELECT $\text{SUM}(\text{Num})$ Where $A_i \in \{a_1, \ldots, a_l\}$
  - estimate by $(1 - p^-_i)q(D_{\text{private}})-p^-_iq^c(D_{\text{private}}))/(p^+_i - p^-_i)$
    • where $q^c$: SELECT $\text{SUM}(\text{Num})$ Where $A_i \notin \{a_1, \ldots, a_l\}$

• Error: $O(\text{randomization} \ast \text{Sqrt}(S))$
Correction of RR

• First, we consider private clean database $D_{\text{private}}$

• $q$: SELECT COUNT(*) Where $A_i \in \{a_1, \ldots, a_l\}$
  — estimate $c$ by $(q(D_{\text{private}})-S p_i^-)/(p_i^+ - p_i^-)$

• $q$: SELECT SUM(Num) Where $A_i \in \{a_1, \ldots, a_l\}$
  — estimate by $(1-p_i^-)q(D_{\text{private}})-p_i^-q^c(D_{\text{private}}))/(p_i^+ - p_i^-)$
  • where $q^c$: SELECT SUM(Num) Where $A_i \notin \{a_1, \ldots, a_l\}$

• $q$: SELECT AVG(Num) Where $A_i \in \{a_1, \ldots, a_l\}$
  — estimate by SUM/COUNT
What Happens in Data Cleaning?

• $q$: SELECT `COUNT(*)` Where $A_i \in \{a_1, ..., a_l\}$
  
  • $p_i^+$: $\Pr[\text{including a record } r \text{ that satisfies the predicate}] = (1 - p_i) + p_i \times \frac{l}{|\text{dom}(A_i)|}$
  
  • $p_i^{-}$: $\Pr[\text{The prob. of including a record } r \text{ that does not satisfy the predicate}] = p_i \times \frac{l}{|\text{dom}(A_i)|}$

Dirty database $D_{\text{dirty}}$

Private dirty database $D_{\text{dirty, private}}$

Output $O = \text{RRCorrection}(q_{\text{dirty}}(D_{\text{dirty, private}}))$
What Happens in Data Cleaning?

$q$: SELECT **COUNT(*)** Where $A_i \in \{a_1, \ldots, a_l\}$

e.g. SELECT **COUNT(*)** Where **Major** is EECS or IEOR

$$D_{dirty, private}$$

<table>
<thead>
<tr>
<th>Id</th>
<th>Major</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IEOR</td>
<td>1.82321</td>
</tr>
<tr>
<td>2</td>
<td>Nuclear Engineering</td>
<td>2.12212</td>
</tr>
<tr>
<td>3</td>
<td>Electrical Engineering and Computer Sciences</td>
<td>3.2412</td>
</tr>
<tr>
<td>4</td>
<td>EECS</td>
<td>2.88721</td>
</tr>
</tbody>
</table>

$$D_{clean, private}$$

<table>
<thead>
<tr>
<th>Id</th>
<th>Major</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IEOR</td>
<td>1.82321</td>
</tr>
<tr>
<td>2</td>
<td>Nuclear Engineering</td>
<td>2.12212</td>
</tr>
<tr>
<td>3</td>
<td>EECS</td>
<td>3.2412</td>
</tr>
<tr>
<td>4</td>
<td>EECS</td>
<td>2.88721</td>
</tr>
</tbody>
</table>

$l = 3$

$$p_{i^+} = (1 - p_i) + p_i \times l / |\text{dom}(A_i)|, \quad p_i^- = p_i \times l / |\text{dom}(A_i)|$$

what is the right choice of $l$, $|\text{dom}(A_i)|$?
Value Provenance Graph per Attribute

\[ \text{Domain}(A_i) \text{ in } D_{\text{dirty}}/D_{\text{private,dirty}} \quad D_{\text{clean}}/D_{\text{private, clean}} \]

\[ q: \text{SELECT COUNT(*)} \]
Where \text{Major} is \text{EECS} or \text{IEOR}
From \( D_{\text{private, clean}} \)

\[ q_{\text{dirty}}: \text{SELECT COUNT(*)} \]
Where \text{Major} is \text{EECS, Electrical .. or IEOR}
From \( D_{\text{dirty, private}} \)

\[ \rightarrow RR_{\text{correction}}(q_{\text{dirty}}(D_{\text{dirty, private}})) \]
Weighted Provenance Graph for Multiple-Attribute Cleaning

• For example: <section, instructor>
  – (1,Null) \(\rightarrow\) (1, John), (2,Null) \(\rightarrow\) (2, Jane)
  – Graph for attribute instructor

• Edge with weight that represents the fraction of rows with the value in dirty database \(a\) mapped to the value in the clean database \(a'\)
  – Use this weight as the probability to include records in \(D_{dirty, private}\) that satisfy the predicate in \(q_{dirty}\)
Outline

• A brief tour to differential privacy

• PrivateClean: data cleaning + differential privacy
  – Private relation generation
  – Query processing
    • Correction for randomized response
    • Cleaning for dirty data
  – Evaluation

• Discussion
Evaluation of PrivateClean

• Baseline: **Direct** $f(D_{\text{clean, private}})$ without CorrectionRR.
• Datasets:
  – Synthetic:
    • a single Num attribute $[0,100]$, a single Cat attribute $\{1,\ldots,N\}$ drawn from a Zipfian distribution
  – TPC-DS:
    • Customer_address (ca_city, ca_county, ca_state, ca_country)
    • [ca_city, ca_county] $\rightarrow$ [ca_state], MD([ca_country]~[ca_country])
  – IntelWireless: 2.3 million
    • 68 sensor id (private), sensor measurements
  – MCAFEE: 406 records
    • Cat: contry code (private), Num: “ethusaism” with value 1-10
Varying privacy parameters $p$

- $p_i^+ = (1 - p_i) + p_i * l/|\text{dom}(A_i)|$, $p_i^- = p_i * l/|\text{dom}(A_i)|$
Varying selectivity $l/|\text{dom}(A_i)|$ of query

- $p_i^+ = (1 - p_i) + p_i \cdot l/|\text{dom}(A_i)|$, $p_i^- = p_i \cdot l/|\text{dom}(A_i)|$
Varying skewness of data

(A) COUNT Accuracy

(B) SUM Accuracy

Error (%) vs Skew z

Direct

PrivateClean
Outline

• A brief tour to differential privacy

• PrivateClean: data cleaning + differential privacy
  – Private relation generation
  – Query processing
    • Correction for randomized response
    • Cleaning for dirty data
  – Evaluation

• Discussion
Limitations

• Poor utility for perturbation per row
  – Error has a lower bound $\sqrt{|D|}$ for both RR and Lap
  – Requires # of distinct values $<< |D|$ at small privacy budget

• Poor privacy guarantee for high dimension data
  – $\sum A_i \epsilon_i$ -DP.

• Limited model:
  – Consider queries with only 1 single Cat attribute in predicate
  – Consider local cleaner that cleans a single row
  – Transformations for data cleaning are given
  – Mainly a post-processing step for DP database
In real world

• Optimal strategies to clean the data are unknown

• Data owner outsources a 3rd party (who is expert for data cleaning) to clean the data, but requires
  – the 3\textsuperscript{rd} party learns little about individual records in the data (provable privacy), but
  – with the optimal data cleaning strategy (utility)
  – with a small number of interactions (efficiency)
    → 3-way trade-offs
Summary

• Differential privacy: mine aggregated personal data with provable guarantees of privacy for individuals.

• PrivateClean: data cleaning + differential privacy
  – Private relation generation
  – Query processing

• Discussion:
  – A good start for private data cleaning
  – Limited model, lack of optimality in privacy and utility