Differential Privacy in the Wild

A Tutorial on Current Practices and Open Challenges
About the Presenters

Ashwin Machanavajjhala
Assistant Professor, Duke University

“What does privacy mean … mathematically?”

Michael Hay
Assistant Professor, Colgate University

“Can algorithms be provably private and useful?”

Xi He
Ph.D. Candidate, Duke University

“Can privacy algorithms work in real world systems?”
Our world is increasingly data driven.
Aggregated Personal Data is invaluable

- Advertising
- Genome Wide Association Studies
- Human Mobility analysis

Source (esri.com)

Tutorial: Differential Privacy in the Wild
Personal data is ... well ... personal!

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

Redlining

Discrimination

Physical/Financial Harm
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.
Outline of the Tutorial

1. What is Privacy?
2. Differential Privacy
3. Answering Queries on Tabular Data
   Break
4. Applications I: Machine Learning
5. Privacy in the Real World
6. Applications II: Networks and Trajectories
Module 1: What is Privacy?

• Privacy Problem Statement

• What privacy is not …
  – Encryption
  – Anonymization
  – Restricted Query Answering

• What is privacy?
Module 2: Differential Privacy

• Differential Privacy Definition

• Basic Algorithms
  – Laplace & Exponential Mechanism
  – Randomized Response

• Composition Theorems
Module 3: Answering queries on Tabular data

• Answering query workloads on tabular databases

• Theory: two seminal results

• Survey of algorithm design ideas
  – Low dimensional range queries
  – Queries on high dimensional data

• Open Questions
Module 4: Applications I

• Private Empirical Risk Minimization
  – E.g. SVM, logistic regression
  – Make a specific learning approach private

• Private Stochastic Gradient Descent
  – E.g. Deep learning
  – Make a general purposed fitting technique private

• Other Important Problems in Private Learning
Module 5: Privacy in the Real World

• Real world deployments of differential privacy
  – OnTheMap
  – RAPPOR

• Privacy beyond Tabular Data
  – No Free Lunch Theorem
  – Customizing differential privacy using Pufferfish
Module 6: Applications II

• Pufferfish Privacy for Non-tabular Data
  
Social network

Location trajectories

• Blowfish Privacy
Scope of the Tutorial

What we do not cover:

• Securing data using encryption
• Computation on encrypted data
• Computationally bounded DP
• De-anonymization
• Anonymization schemes (k-anonymity, l-diversity, etc.)
• Access control
MODULE 1:
PROBLEM FORMULATION
Module 1: What is Privacy?

• Privacy Problem Statement

• What privacy is not …
  – Encryption
  – Anonymization
  – Restricted Query Answering

• What is privacy?
Statistical Databases

Individuals with sensitive data

Person 1 \(r_1\)

Person 2 \(r_2\)

Person 3 \(r_3\)

... 

Person N \(r_N\)

Data Collectors

Hospital

Census

Google

Economists

Information Retrieval Researchers

Data Analysts

Doctors

Medical Researchers

Recommendation Algorithms

Module 1 Tutorial: Differential Privacy in the Wild
The Massachusetts Governor Privacy Breach

- Name
- SSN
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge
- Zip
- Birth date
- Sex

Medical Data Release
The Massachusetts Governor Privacy Breach

Medical Data Release

- Name
- SSN
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

Voter List

- Name
- Address
- Date Registered
- Party affiliation
- Date last voted

Module 1

Tutorial: Differential Privacy in the Wild
Linkage Attack

Medical Data Release
- Name
- SSN
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

Voter List
- Name
- Address
- Date Registered
- Party affiliation
- Date last voted

• Governor of MA uniquely identified using ZipCode, Birth Date, and Sex.

Name linked to Diagnosis
Linkage Attack

- 87% of US population uniquely identified using ZipCode, Birth Date, and Sex.

Medical Data Release
- Name
- SSN
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

Voter List
- Name
- Address
- Date Registered
- Party affiliation
- Date last voted

Quasi Identifier
Statistical Database Privacy

Function provided by the analyst

Output can disclose sensitive information about individuals

Module 1  Tutorial: Differential Privacy in the Wild  26
Statistical Database Privacy

Privacy for individuals (controlled by a parameter $\varepsilon$)

$f_{\text{private}}(DB, \varepsilon)$

Person 1 $r_1$

Person 2 $r_2$

Person 3 $r_3$

Person N $r_N$
Statistical Database Privacy

Utility for analyst

\( f_{\text{private}}(DB) \approx f(DB) \)

\[ f_{\text{private}}(DB, \varepsilon) \]

Person 1

Person 2

Person 3

\( r_1 \)

\( r_2 \)

\( r_3 \)

Person \( N \)

\( r_N \)
Statistical Database Privacy
(untrusted collector)

Server wants to compute $f$

Individuals do not want server to infer their records

$l_1$

$l_2$

$l_3$

$l_N$

Server

$\left(\begin{array}{c}
DB
\end{array}\right)$

Module 1
Tutorial: Differential Privacy in the Wild
Statistical Database Privacy
(untrusted collector)

Perturb records to ensure privacy for individuals and Utility for server

Module 1
Tutorial: Differential Privacy in the Wild
# Statistical Databases in real-world applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Data Collector</th>
<th>Private Information</th>
<th>Analyst</th>
<th>Function (utility)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical</td>
<td>Hospital</td>
<td>Disease</td>
<td>Epidemiologist</td>
<td>Correlation between disease and geography</td>
</tr>
<tr>
<td>Genome analysis</td>
<td>Hospital</td>
<td>Genome</td>
<td>Statistician/Researcher</td>
<td>Correlation between genome and disease</td>
</tr>
<tr>
<td>Advertising</td>
<td>Google/FB/Y!</td>
<td>Clicks/Browsing</td>
<td>Advertiser</td>
<td>Number of clicks on an ad by age/region/gender ...</td>
</tr>
<tr>
<td>Social Recommendations</td>
<td>Facebook</td>
<td>Friend links / profile</td>
<td>Another user</td>
<td>Recommend other users or ads to users based on social network</td>
</tr>
</tbody>
</table>
Statistical Databases in real-world applications

• Settings where data collector may not be trusted (or may not want the liability …)

<table>
<thead>
<tr>
<th>Application</th>
<th>Data Collector</th>
<th>Private Information</th>
<th>Function (utility)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location Services</td>
<td>Verizon/AT&amp;T</td>
<td>Location</td>
<td>Traffic prediction</td>
</tr>
<tr>
<td>Recommendations</td>
<td>Amazon/Google</td>
<td>Purchase history</td>
<td>Recommendation model</td>
</tr>
<tr>
<td>Traffic Shaping</td>
<td>Internet Service Provider</td>
<td>Browsing history</td>
<td>Traffic pattern of groups of users</td>
</tr>
</tbody>
</table>
Privacy is not ...
Statistical Database Privacy is not …

• Encryption:
Statistical Database Privacy is not ...

• Encryption:
  Alice sends a message to Bob such that Trudy (attacker) does not learn the message. Bob should get the correct message ...

• Statistical Database Privacy:
  Bob (attacker) can access a database
  - Bob must learn aggregate statistics, but
  - Bob must not learn new information about individuals in database.
Statistical Database Privacy is not ... 

- Computation on Encrypted Data:
Statistical Database Privacy is not …

• Computation on Encrypted Data:
  - Alice stores encrypted data on a server controlled by Bob (attacker).
  - Server returns correct query answers to Alice, without Bob learning *anything* about the data.

• Statistical Database Privacy:
  - Bob is allowed to learn aggregate properties of the database.
Statistical Database Privacy is not ...

• The Millionaires Problem:
Statistical Database Privacy is not …

• Secure Multiparty Computation:
  - A set of agents each having a private input $x_i$ …
  - … Want to compute a function $f(x_1, x_2, \ldots, x_k)$
  - Each agent can learn the true answer, but must learn no other information than what can be inferred from their private input and the answer.

• Statistical Database Privacy:
  - Function output must not disclose individual inputs.
Statistical Database Privacy is not ... 

• Access Control:
Statistical Database Privacy is not ... 

• Access Control:
  - A set of agents want to access a set of resources (could be files or records in a database)
  - Access control rules specify who is allowed to access (or not access) certain resources.
  - ‘Not access’ usually means no information must be disclosed

• Statistical Database:
  - A single database and a single agent
  - Want to release aggregate statistics about a set of records without allowing access to individual records
Privacy Problems

• In today’s cloud context, a number of privacy problems arise:
  – Encryption when communicating data across an insecure channel
  – Secure Multiparty Computation when different parties want to compute on a function on their private data without using a centralized third party
  – Computing on encrypted data when one wants to use an insecure cloud for computation
  – Access control when different users own different parts of the data

• Statistical Database Privacy: Quantifying (and bounding) the amount of information disclosed about individual records by the output of a valid computation.
What is privacy?
Privacy Breach: Informal Definition

A privacy mechanism $M(D)$ that allows an unauthorized party to learn sensitive information about any individual in $D$, which could not have learnt without access to $M(D)$. 
Alice

Is this a privacy breach?  NO
Privacy Breach: Revised Definition

A privacy mechanism $M(D)$ that allows an unauthorized party to learn sensitive information about any individual Alice in $D$, which could not have learnt without access to $M(D)$ if Alice was not in the dataset.
K-Anonymity: Avoiding Linkage Attacks

• If every row corresponds to one individual …

… every row should look like k-1 other rows based on the quasi-identifier attributes
## K-Anonymity

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13053</td>
<td>28</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>29</td>
<td>American</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>21</td>
<td>Japanese</td>
<td>Flu</td>
</tr>
<tr>
<td>13053</td>
<td>23</td>
<td>American</td>
<td>Flu</td>
</tr>
<tr>
<td>14853</td>
<td>50</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>14853</td>
<td>55</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>14850</td>
<td>47</td>
<td>American</td>
<td>Flu</td>
</tr>
<tr>
<td>14850</td>
<td>59</td>
<td>American</td>
<td>Flu</td>
</tr>
<tr>
<td>13053</td>
<td>31</td>
<td>American</td>
<td>Cancer</td>
</tr>
<tr>
<td>13053</td>
<td>37</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>36</td>
<td>Japanese</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>32</td>
<td>American</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Problem: Background knowledge

Adversary knows prior knowledge about Umeko

Adversary learns Umeko has Cancer

<table>
<thead>
<tr>
<th>Name</th>
<th>Zip</th>
<th>Age</th>
<th>Nat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Umeko</td>
<td>13053</td>
<td>25</td>
<td>Japan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
A privacy mechanism must be able to protect individuals’ privacy from attackers who may possess background knowledge.
Welcome to H-CUPnet

H-CUPnet is a free, on-line query system based on data from the Healthcare Cost and Utilization Project (HCUP). It provides access to health statistics and information on hospital inpatient and emergency department utilization.

Statistics on Hospital Stays

- National Statistics on All Stays
  
  Create your own statistics for national and regional estimates on hospital use for all patients from the HCUP National (Nationwide) Inpatient Sample (NIS). Overview of the National (Nationwide) Inpatient Sample (NIS)

- National Statistics on Mental Health Hospitalizations
  
  Interested in acute care hospital stays for mental health and substance abuse? Create your own national statistics from the NIS.

- State Statistics on All Stays
  
  Create your own statistics on stays in hospitals for participating States from the HCUP State Inpatient Databases (SID). Overview of the State Inpatient Databases (SID)

National Statistics on Children

- National Statistics on Children
  
  Create your own statistics for national estimates on use of hospitals by children (age 0-17 years) from the HCUP Kids' Inpatient Database (KID). Overview of the Kids' Inpatient Database (KID)

- National and State Statistics on Hospital Stays by Payer - Medicare, Medicaid, Private, Uninsured
  
  Interested in hospital stays billed to a specific payer? Create your own statistics for a payer, alone or compared to other payers from the NIS, KID, and SID.

- Quick National or State Statistics
  
  Ready-to-use tables on commonly requested information from the HCUP National (Nationwide) Inpatient Sample (NIS), the HCUP Kids' Inpatient Database (KID), or the HCUP State Inpatient Databases (SID).
## Hospital discharges in NJ of ovarian cancer patients, 2009

<table>
<thead>
<tr>
<th>Age</th>
<th>#discharges</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian/Pcf</th>
<th>Native American</th>
<th>Other</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-17</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>18-44</td>
<td>70</td>
<td>40</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>45-64</td>
<td>330</td>
<td>236</td>
<td>31</td>
<td>32</td>
<td>*</td>
<td>*</td>
<td>11</td>
<td>*</td>
</tr>
<tr>
<td>65-84</td>
<td>298</td>
<td>229</td>
<td>35</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>85+</td>
<td>34</td>
<td>29</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Counts less than k are suppressed achieving k-anonymity.
#Hospital discharges in NJ of ovarian cancer patients, 2009

<table>
<thead>
<tr>
<th>Age</th>
<th>#discharges</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian/Pcf Hlnder</th>
<th>Native American</th>
<th>Other</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#discharges</td>
<td>735</td>
<td>535</td>
<td>82</td>
<td>58</td>
<td>18</td>
<td>1</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>1-17</td>
<td>3</td>
<td>1</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>18-44</td>
<td>70</td>
<td>40</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>45-64</td>
<td>330</td>
<td>236</td>
<td>31</td>
<td>32</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>65-84</td>
<td>298</td>
<td>229</td>
<td>35</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>85+</td>
<td>34</td>
<td>29</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

\[= 535 - (40 + 236 + 229 + 29)\]
# Hospital discharges in NJ of ovarian cancer patients, 2009

<table>
<thead>
<tr>
<th>Age</th>
<th>#discharges</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian/Pcf Hlnder</th>
<th>Native American</th>
<th>Other</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#discharges</td>
<td>735</td>
<td>535</td>
<td>82</td>
<td>58</td>
<td>18</td>
<td>1</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>1-17</td>
<td>3</td>
<td>1</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
</tr>
<tr>
<td>18-44</td>
<td>70</td>
<td>40</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>45-64</td>
<td>330</td>
<td>236</td>
<td>31</td>
<td>32</td>
<td>*</td>
<td>*</td>
<td>11</td>
<td>*</td>
</tr>
<tr>
<td>65-84</td>
<td>298</td>
<td>229</td>
<td>35</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>85+</td>
<td>34</td>
<td>29</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
#Hospital discharges in NJ of ovarian cancer patients, 2009

<table>
<thead>
<tr>
<th>Age</th>
<th>#discharges</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian/Pcf Hlnder</th>
<th>Native American</th>
<th>Other</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-17</td>
<td>3</td>
<td>1</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
</tr>
<tr>
<td>18-44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45-64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>85+</td>
<td></td>
<td></td>
<td>[1-3]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Can reconstruct tight bounds on rest of data

In fact, when linked with queries giving other statistics, we can figure out that exactly 1 Native American woman diagnosed with ovarian cancer went to a privately owned, not for profit, teaching hospital in New Jersey with more than 435 beds in 2009. Furthermore, the woman did not pay by private insurance, had a routine discharge, with a stay in the hospital of 33.5 days, with her home residence being in a county with 1 million plus residents (large fringe metro, suburbs), and her age was exactly 75 years.
Multiple Release problem

• Privacy preserving access to data must necessarily release some information about individual records (to ensure utility)

• However, $k$-anonymous algorithms can reveal individual level information even with two releases.
A privacy mechanism must satisfy composition …

… or allow a graceful degradation of privacy with multiple invocations on the same data.

[DN03, GKS08]
Postprocessing the output of a privacy mechanism must not change the privacy guarantee

[KL10, MK15]
Privacy must not be achieved through obscurity.

Attacker must be assumed to know the algorithm used as well as all parameters.
Summary

• Statistical database privacy is the problem of releasing aggregates while not disclosing individual records

• The problem is distinct from encryption, secure computation and access control.

• Defining privacy is non-trivial
  – Desiderata include resilience to background knowledge and composition and closure under postprocessing.
Module 2: Differential Privacy

• Differential Privacy Definition

• Basic Algorithms
  – Laplace & Exponential Mechanism
  – Randomized Response

• Composition Theorems
Differential Privacy

For every pair of inputs that differ in one row

For every output ...

If algorithm A satisfies differential privacy then

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

Intuition: adversary should not be able to use output O to distinguish between any D_1 and D_2
Why pairs of datasets *that differ in one row*?

For every pair of inputs that differ in one row:

\[
D_1 \quad D_2
\]

Simulate the presence or absence of a single record

For every output ...

\[
O
\]
Why *all* pairs of datasets ...?

For every pair of inputs that differ in one row

$D_1$  $D_2$

For every output ...

$O$

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

- \( D_1 \) and \( D_2 \) represent sets of all outputs.
- The probability of \( A(D_1) = O_1 \) is represented by the red bar.
- The probability of \( A(D_2) = O_k \) is represented by the green bar.

Set of all outputs
Should not be able to distinguish whether input was $D_1$ or $D_2$ no matter what the output.
Privacy Parameter $\varepsilon$

For every pair of inputs that differ in one row

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

For every output ...

$D_1$ $D_2$ $O$

Controls the degree to which $D_1$ and $D_2$ can be distinguished. Smaller $\varepsilon$ gives more privacy (and worse utility)
Outline of the Module 2

• Differential Privacy

• Basic Algorithms
  – Laplace & Exponential Mechanism
  – Randomized Response

• Composition Theorems
Can deterministic algorithms satisfy differential privacy?
Non trivial deterministic algorithms do not satisfy differential privacy

Space of all inputs

Space of all outputs
Non-trivial deterministic algorithms do not satisfy differential privacy

Non-trivial: at least 2 outputs in image
There exist two inputs that differ in one entry mapped to different outputs.

Pr = 1

Pr = 0
Random Sampling …

… also does not satisfy differential privacy

\[
\begin{align*}
\Pr[D_2 \rightarrow O] &= 0 & \text{implies} & \frac{\Pr[D_1 \rightarrow O]}{\Pr[D_2 \rightarrow O]} &= \infty
\end{align*}
\]
• 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 → Query q → q(D) + η

Privacy depends on the λ parameter

\[ h(\eta) \propto \exp\left(-\frac{\eta}{\lambda}\right) \]

Mean: 0,
Variance: 2 \( \lambda^2 \)

Laplace Distribution – \( \text{Lap}(\lambda) \)

[DMNS 06]
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 $D$, $D'$,

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

Thm: If sensitivity of the query is $S$, then the following guarantees $\varepsilon$-differential privacy.

$$\lambda = S/\varepsilon$$
Sensitivity: COUNT query

- Number of people having disease
- Sensitivity = 1

- Solution: $3 + \eta$
  where $\eta$ is drawn from $\text{Lap}(1/\varepsilon)$
  - Mean = 0
  - Variance = $2/\varepsilon^2$
Sensitivity: 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: $E(\text{true answer} - \text{noisy answer})^2$

  $$= \text{Var}( \text{Lap}(S(q)/\varepsilon) )$$

  $$= 2S(q)^2 / \varepsilon^2$$
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 $f$?

[MT 07]
Exponential Mechanism

- Scoring function \( w: \text{Inputs} \times \text{Outputs} \rightarrow \mathbb{R} \)

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

- \( \#(D, O) \): \# people with nationality \( O \)

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

- \( w(D, O) = |\#(D, O) - \#(D, f(D))| \)
Exponential Mechanism

• Scoring function $w: \text{Inputs} \times \text{Outputs} \rightarrow \mathbb{R}$

• Sensitivity of $w$

$\Delta_w = \max_{O \& D, D'} |w(D, O) - w(D, O')|$

where $D, D'$ differ in one tuple
Exponential Mechanism

Given an input \( D \), and a scoring function \( w \),

Randomly sample an output \( O \) from \( \text{Outputs} \) with probability

\[
\frac{e^{\frac{\epsilon}{2\Delta}} w(D, O)}{\sum_{Q \in \text{Outputs}} e^{\frac{\epsilon}{2\Delta}} w(D, Q)}
\]

- Note that for every output \( O \), probability \( O \) is output > 0.
**Randomized Response** (a.k.a. local randomization)

<table>
<thead>
<tr>
<th>Disease (Y/N)</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

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 $O$

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

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

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

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

Sampling \hspace{1cm} Variance due to coin flips
Laplace Mechanism vs Randomized Response

Privacy

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

• Laplace mechanism assumes data collected is trusted

• Randomized Response does not require data collected 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 $\mu N$ records have disease = $Y$.
• Query: # rows with Disease = $Y$

• Std dev of Laplace mechanism answer: $O(1/\varepsilon)$
• Std dev of Randomized Response answer: $O(\sqrt{N})$
Outline of the Module 2

• 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 the number of queries that can be answered.
Dinur Nissim Result

- A vast majority of records in a database of size $n$ can be reconstructed when $n \log(n)^2$ queries are answered by a statistical database …

… even if each answer has been arbitrarily altered to have up to $o(\sqrt{n})$ error
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 running all $k$ algorithms sequentially satisfies $\varepsilon$-differential privacy with $\varepsilon = \varepsilon_1 + \ldots + \varepsilon_k$.
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?
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 running all $k$ algorithms in “parallel” 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.
Case Study: K-means Clustering

Original unclustered data

Clustered data

Module 2

Tutorial: Differential Privacy in the Wild
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 $K$means

- 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 $\epsilon/T$ privacy budget, total privacy loss is $\epsilon$

• 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  

  2. Noisily compute the size of each cluster  

  3. Compute noisy sums of points in each cluster
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 $\varepsilon / T$ 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

  \[ \text{Laplace}(2T/\varepsilon) \]

  \[ \text{Laplace}(2T \mid \text{dom} \mid /\varepsilon) \]
Even though we noisily compute centers, Laplace k-means 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.
Summary

• Differentially private algorithms 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 and randomized response

• Composition rules help build complex algorithms using building blocks
MODULE 3:
ANSWERING QUERIES ON TABULAR DATA
Module 3: Answering queries on Tabular data

• Answering query workloads on tabular databases

• Theory: two seminal results

• Survey of algorithm design ideas
  – Low dimensional range queries
  – Queries on high dimensional data

• Open Questions
Problem Formulation

• **Input:**
  - Private database $D$ consisting of a single table (each tuple represents data of single individual)
  - Workload $W$ of counting queries* with arbitrary predicates

    SELECT COUNT(*) FROM D WHERE <P>;

• **Output:** (noisy) answers to $W$

• Requirement: query answering algorithm satisfies differential privacy

* Many techniques can also support linear queries: SELECT SUM($f(t)$) FROM $D$ where user-defined $f$ maps tuple to [0,1]
Analysis of temporal & spatial patterns

Counting query:

```
SELECT COUNT(*)
FROM Tweets
WHERE
moodScale=k
AND t <= time
AND time < t+1
AND UScounty = C
```
Statistical agencies: data publishing

- U.S. Census Bureau publishes statistics that can typically be derived from marginals
- A **marginal** over attributes $A_1, \ldots, A_k$ reports count for each combination of attribute values.
  - aka cube, contingency table
  - E.g. 2-way marginal on EmploymentStatus and Gender
- **Thousands** of marginals released

### Table: Employment Status

<table>
<thead>
<tr>
<th>Subject</th>
<th>Estimate</th>
<th>Margin of Error</th>
<th>Percent</th>
<th>Percent Margin of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population 16 years and over</td>
<td>6.675</td>
<td>+/-361</td>
<td>5.878</td>
<td>(X)</td>
</tr>
<tr>
<td>In labor force</td>
<td>2.715</td>
<td>+/-223</td>
<td>47.8%</td>
<td>+/-3.7</td>
</tr>
<tr>
<td>Civilian labor force</td>
<td>2.715</td>
<td>+/-223</td>
<td>47.8%</td>
<td>+/-3.7</td>
</tr>
<tr>
<td>Employed</td>
<td>2.529</td>
<td>+/-228</td>
<td>44.6%</td>
<td>+/-3.6</td>
</tr>
<tr>
<td>Unemployed</td>
<td>0</td>
<td>+/-9.2</td>
<td>0.0%</td>
<td>+/-0.6</td>
</tr>
<tr>
<td>Not in labor force</td>
<td>2.991</td>
<td>+/-266</td>
<td>52.2%</td>
<td>+/-3.7</td>
</tr>
<tr>
<td>Civilian labor force</td>
<td>2.715</td>
<td>+/-223</td>
<td>47.8%</td>
<td>+/-3.7</td>
</tr>
<tr>
<td>Percent Unemployed</td>
<td>(X)</td>
<td>(X)</td>
<td>0.0%</td>
<td>+/-0.6</td>
</tr>
<tr>
<td>Female 16 years and over</td>
<td>2.921</td>
<td>+/-216</td>
<td>2.921</td>
<td>(X)</td>
</tr>
<tr>
<td>In labor force</td>
<td>1.312</td>
<td>+/-140</td>
<td>44.9%</td>
<td>+/-4.5</td>
</tr>
<tr>
<td>Civilian labor force</td>
<td>1.312</td>
<td>+/-140</td>
<td>44.9%</td>
<td>+/-4.5</td>
</tr>
<tr>
<td>Employed</td>
<td>1.245</td>
<td>+/-135</td>
<td>42.6%</td>
<td>+/-4.3</td>
</tr>
<tr>
<td>Own children under 6 years</td>
<td>325</td>
<td>+/-117</td>
<td>325</td>
<td>(X)</td>
</tr>
<tr>
<td>All parents in family in labor</td>
<td>241</td>
<td>+/-99</td>
<td>74.2%</td>
<td>+/-17.3</td>
</tr>
<tr>
<td>Own children 6 to 17 years</td>
<td>476</td>
<td>+/-102</td>
<td>476</td>
<td>(X)</td>
</tr>
<tr>
<td>All parents in family in labor</td>
<td>389</td>
<td>+/-65</td>
<td>81.7%</td>
<td>+/-8.5</td>
</tr>
</tbody>
</table>

### Table: Commuting to Work

<table>
<thead>
<tr>
<th>Subject</th>
<th>Estimate</th>
<th>Margin of Error</th>
<th>Percent</th>
<th>Percent Margin of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workers 10 years and over</td>
<td>2.449</td>
<td>+/-217</td>
<td>2.449</td>
<td>(X)</td>
</tr>
<tr>
<td>Car, truck, or van – drove alone</td>
<td>1.219</td>
<td>+/-176</td>
<td>92.0%</td>
<td>+/-5.2</td>
</tr>
<tr>
<td>Car, truck, or van – carpooled</td>
<td>118</td>
<td>+/-73</td>
<td>4.7%</td>
<td>+/-2.9</td>
</tr>
<tr>
<td>Public transportation (excluding taxicab)</td>
<td>17</td>
<td>+/-9.8</td>
<td>0.7%</td>
<td>+/-0.6</td>
</tr>
<tr>
<td>Walked</td>
<td>531</td>
<td>+/-116</td>
<td>21.7%</td>
<td>+/-4.3</td>
</tr>
<tr>
<td>Other means</td>
<td>132</td>
<td>+/-56</td>
<td>5.4%</td>
<td>+/-2.4</td>
</tr>
<tr>
<td>Worked at home</td>
<td>135</td>
<td>+/-44</td>
<td>5.8%</td>
<td>+/-2.5</td>
</tr>
</tbody>
</table>

### Table: Mean travel time to

<table>
<thead>
<tr>
<th>Subject</th>
<th>Estimate</th>
<th>Margin of Error</th>
<th>Percent</th>
<th>Percent Margin of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Civilian employed population 16 years and over</td>
<td>2.629</td>
<td>+/-258</td>
<td>2.629</td>
<td>(X)</td>
</tr>
</tbody>
</table>

[https://factfinder.census.gov/](https://factfinder.census.gov/)
Genome Wide Association Studies

- **Goal**: to study genetic factors associated with a given disease
- Collect subsets of the population with diseases (*case*) and without (*control*)
- Extract SNPs (specific DNA subsequences)
  - For each SNP, usually find 2 *alleles* (alternative forms of the gene)

- **Counting queries**: Compute allele frequencies in both the case and control groups (marginal over SNPxDisease)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
</tr>
<tr>
<td>0 (Other allele)</td>
<td>$C_{00}$</td>
</tr>
<tr>
<td>1 (Risk allele)</td>
<td>$C_{10}$</td>
</tr>
</tbody>
</table>

- Perform association test using these frequencies (e.g., Chi Square Test) to identify SNPs highly associated with disease
Problem variant: offline vs. online

- **Offline** (batch):
  - Entire $W$ given as input, answers computed in **batch**

- **Online** (adaptive):
  - $W$ is sequence $q_1, q_2, \ldots$ that arrives online
  - **Adaptive**: analyst’s choice for $q_i$ can depend on answers $a_1, \ldots, a_{i-1}$

- Answering linear queries online is strictly harder than answering them offline [BSU16].
Important aspects of problem: Data and query complexity

• Data complexity
  – Dimensionality: number of attributes
  – Domain size: number of distinct attribute combinations
  – Many techniques specialized for low dimensional data

• Query complexity
  – Many techniques designed to work well for a specific class of queries
  – Classes (in rough order of difficulty): histograms, range queries, marginals, counting queries, linear queries
Solution variants: query answers vs. synthetic data

Two high-level approaches to solving problem

1. Direct:
   – Output of the algorithm is list of query answers

2. Synthetic data:
   – Algorithm constructs a synthetic dataset $D'$, which can be queried directly by analyst
   – Analyst can pose additional queries on $D'$ (though answers may not be accurate)
Theory

- Given negative result of Dinur-Nissim, is there any hope?
- Yes!
  - The key to Dinur-Nissim is that query answers have independent noise (which can cancel out)
  - To answer more queries, query error must be correlated
- Examples of correlation
  - Use some query answers to approximate others
  - Construct a synthetic database that is approximately accurate for queries of interest
Answering Exponentially Many Queries
Offline

• Key technical insight: For any set of count queries \( W \), there exists a small database \( D' \) consistent with \( D \) on every query in \( W \).
  – Small: \( O(\log(|W|)/\alpha^2) \)
  – Consistent: error for any \( q \) in \( W \) is at most \( \alpha \)

• Result follows from learning theory:
  – Estimates on small random sample will generalize to population
Answering Exponentially Many Queries
Offline

• Input: $W, \varepsilon$
• Output: $D'$
• The Mechanism:
  - $T = \{\text{all small databases } D'\}$
  - $f(D, D') = -\max_{q \in W} |q(D) - q(D')|$  
  - Output $D' \in T$ using Exponential Mechanism applied to $f$
• Theorem: Is $\varepsilon$-private and w.h.p. error $\alpha$ is at most
  $$O\left(\frac{\log |\text{domain}| \log |W|}{\varepsilon |D|}\right)^{1/3}$$
Answering Exponentially Many Queries
Offline

• Input: \( W, \varepsilon \)
• Output: \( D' \)
• The Mechanism:
  - \( T = \{ \text{all small databases } D' \} \)
  - \( f(D, D') = -\max_{q \in W} |q(D) - q(D')| \)
  - Output \( D' \in T \) using Exponential Mechanism applied to \( f \)
• Theorem: Is \( \varepsilon \)-private and w.h.p.
  \[
  O\left( \frac{\log |\text{domain}| \log |\varepsilon|}{\varepsilon |D|} \right)
  
  \]

Limitations
• Impractical: runtime exponential
• Offline (for online see [HR10])
Case study: range queries over spatial data

**Input: sensitive data** $D$

<table>
<thead>
<tr>
<th></th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>39.98105</td>
<td>116.30142</td>
</tr>
<tr>
<td>2</td>
<td>39.9424</td>
<td>116.30587</td>
</tr>
<tr>
<td>3</td>
<td>39.93691</td>
<td>116.33438</td>
</tr>
<tr>
<td>4</td>
<td>39.94354</td>
<td>116.3532</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BeijingTaxi dataset[1]:
4,268,780 records of (lat,lon) pairs of taxi pickup locations in Beijing, China in 1 month.

**Input: range query workload** $W$

Shown is workload of 3 range queries

**Task:** compute answers to workload $W$ over private input $D$

Baseline algorithm

1. Discretize attribute domain into cells
2. Add noise to cell counts (Laplace mechanism)
3. Use noisy counts to either…
   1. Answer queries directly (assume distribution is uniform within cell)
   2. Generate synthetic data (derive distribution from counts and sample)

Limitations

• Granularity of discretization
  – Coarse: detail lost
  – Fine: noise overwhelms signal
• Noise accumulates: squared error grows linearly with range

Scatter plot of input data
Error analysis

- Dataset is 1D: single attribute, 128 possible values
- Baseline simply adds noise to every cell count
- Incurs high error on large ranges

![Error baseline algorithm on every range query](chart)

- **[x₁, x₁]**
  - Large range
- **[x₁, x₁₂₈]**
- **[x₁₂₈, x₁₂₈]**
  - Small range

Module 3

Tutorial: Differential Privacy in the Wild
**Improved algorithm: private quad-tree**

- **Input**: maximum height $h$, minimum leaf size $L$, data set $D$
- **Recurse on nodes of tree:**
  - Add $\text{Lap}(1/\varepsilon)$ noise to node count
  - Split node domain into quadrants
  - Create child nodes
- **Stop when:**
  - Noisy count of node $\leq L$
  - Max height $h$ is reached
- **Intuition:**
  - Early stopping controls granularity of discretization
  - To answer long range queries, leverage hierarchy of noisy counts

*Tutorial: Differential Privacy in the Wild*
Improved algorithm: private quad-tree

- **Input:** maximum height $h$, minimum leaf size $L$, data set $D$

- Recurse on nodes of tree:
  - Add $\text{Lap}(1/\varepsilon)$ noise to node count
  - Split node domain into quadrants
  - Create child nodes

- Stop when:
  - Noisy count of node $\leq L$
  - Max height $h$ is reached

- **Intuition:**
  - Early stopping controls granularity of discretization
  - To answer long range queries, leverage hierarchy of noisy counts

**Exercise:** Let $h' \leq h$ be height of resulting tree. Algorithm satisfies $\varepsilon'$-differential privacy for $\varepsilon'$ equal to

1. $\varepsilon h'$
2. $\varepsilon h$
3. $4\varepsilon h$
4. $\varepsilon 4^h$
Using tree structure for range queries

• Quad-tree, each node has noisy count. How use to answer range queries?

• Idea 1: given range query q, find smallest set of noisy counts from tree that “cover” q

• Idea 2 (better):
  – Observation: node’s noisy count and sum of children’s noisy counts are two estimates of same quantity
  – Combine estimates using statistical inference
Error comparison

Error tree algorithm on every range query

Error baseline algorithm on every range query

Module 3
Tutorial: Differential Privacy in the Wild
Hierarchy lowers error on *large ranges* but incurs slightly higher error for *small ranges*.
Data transformations

- Can think of trees as transform of input

- Can apply other data transformations

- **Goal**: pick a low sensitivity transform that preserves good properties of data

Credit: Cormode
Linear transformations

• Examples
  – Hierarchical: Trees [HRMS10,QYL13], full height quadtree [CPSSY12]
  – Haar Wavelet [XWG10]
  – Discrete Fourier transform [BCDKMT07]

• Inverting transformation
  – Some transformations (e.g. tree) have redundancy (over-constrained), so require pseudo-inverse

• Matrix Mechanism [LHRMM10,LM12,LM13]
  – Formalizes problem of designing a linear transform that is tailored to the queries

• Error rates are independent of input (assumes linear transform is “full rank”)
Lossy transformations

• **Variants**
  - Drop “small” coefficients:
    - Quad-tree with early stopping (noisy count threshold)
    - Fourier coefficients: EFPA [ACC12], [RN10]
  - Data-adaptive discretization:
    - PrivTree [ZXX16], KD-Tree [CPSSY12], DAWA [LHMY14], [DNRR15], [QYL13], [BLR08]
  - Data-adaptive measurement:
    - MWEM [HLM12], DualQuery [GAHRW14]
  - Randomized transforms: sketches and compressed sensing
    - JL Transform [BBDS12], Compressive mechanism [LZWy11]

• **“Inverting” transformation**
  - Because lossy, they are under-constrained, requires estimation

• **Error rates** *depend on input*
  - Can be much lower (trades off small bias for lower variance)
  - Warrants careful empirical evaluation; algorithms are “data dependent”
High dimensional data

• Generally an under-studied area
• Two algorithms, both synthetic data generators
  – PrivBayes [ZCPSX14]
  – DualQuery [GAHRW14]
• Common properties
  – Limited to binary attributes
  – Designed to support low-order marginals
    (and other workloads well approximated by marginals, such as classification)
PrivBayes

- Method:
  - Use Bayesian network to learn data distribution
  - After BN learned, generate synthetic data by sampling from BN
- Challenge: privately choosing good decomposition

High-dimensional table $T$

Low-dimensional tables

Add noise

Noisy tables

Noisy table $T^*$

[ZCPSX14]
Dual Query

- Problem of generating synthetic data formulated as a zero sum game between
  - **Data player**: generates synthetic records to reduce query error
  - **Query player**: chooses queries with high error (on current synthetic dataset)
  - Theoretical analysis of utility comes from studying equilibrium of game

1. Let $Q^t$ be distribution over queries ($Q^1$ is uniform)
2. For $t = 1 \ldots T$
   a) $S \leftarrow$ Query player samples $s$ queries from $Q^t$
   b) Data player finds record $x_t$ that maximizes total answer on $S$
   c) $Q^{t+1} \leftarrow$ Query player updates($x_t$, $D$, $Q^t$)
3. Output synthetic database $x_1, \ldots, x_t$

- What makes it practical?
  - Unlike some prior work [HLM12], avoids storing distribution over domain (exponential)
  - Approximate solution may be good enough!
  - Optimization problem can be solved with off-the-shelf solvers
- Case study on 500K 3-way marginals over 17K binary attributes, using CPLEX solver
Empirical benchmarks

- [HMMCZ16] propose a novel evaluation framework for standardized evaluation of privacy algorithms.
- Study of algorithms for range query answering over 1 and 2D.
- Benchmark website [www.dpcomp.org](http://www.dpcomp.org)

Some data-dependent algorithms fail to offer benefits at larger scales (no. of tuples)
Open questions

• Robust and private algorithm selection
  – Pythia (Thursday 2 PM DP Session)
• Error bounds for data-dependent algorithms
• Empirical evaluation of algorithms for high dimensional data
Differential Privacy References

CACM Articles

Book

Tutorials
Module 1 References

- [S02] Sweeney, “K-anonymity”, IJFUKS 2010
- [DN03] Dinur, Nissim, “Revealing information while preserving privacy”, PODS 2003
Module 2 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
Module 3 References