CS208: Applied Privacy for Data Science
Implementing (Centralized) Differential Privacy: Programming Interfaces for DP

James Honaker & Salil Vadhan
School of Engineering & Applied Sciences
Harvard University

March 15, 2019
PSI Library
PSI Library

- Modular architecture
- Class based, using R
- Contains common exploratory statistics
  - Univariate descriptive statistics, such as means, quantiles, histograms, and approximate cumulative distribution functions.
  - Basic statistical estimators, matching algorithms and difference-of-means tests for causal inference, and low-dimensional linear, logit, probit and poisson regression.
- Reusable
# Generate and print dp release of mean of age

dp.release <- dpMean$new(mechanism='mechanismLaplace',
    var.type='numeric',
    variable="age",
    n=10000,
    epsilon=0.1,
    delta=10^-6,
    rng=c(0,110))
dp.release$release(my_data)
Example tree, showing leaf values a through h below, node labels above, and differentially private values to the right with noise ν.
Eight of the possible ways of using the nodes on the tree to estimate $a + b + c + d$, the sum of the left four leaves. Nodes in green are added, and nodes in red subtracted to create an estimate.
PSI Budgeter
“Accessible and reusable data are fundamental to science in order to continuously validate and build upon previous research. Progressive expansive scientific advance rests upon access to data accompanied with sufficient information for reproducible results, a scientific ethic to maximize the utility of data to the research community, and a foundational norm that scientific communication is built on attribution.”

- Crosas et al. 2015
Open Science: Discovery, Replication, Reuse

Tl;dr: Replication is the foundation of science; reuse is the building block.

“Accessible and reusable data are fundamental to science in order to continuously validate and build upon previous research. Progressive expansive scientific advance rests upon access to data accompanied with sufficient information for reproducible results, a scientific ethic to maximize the utility of data to the research community, and a foundational norm that scientific communication is built on attribution.”

- Crosas et al. 2015
Overview

Our goal is Privacy Protection that is:

1. **Accessible:** no DP expert optimizing algorithms for a particular dataset or application
2. **General:** works across the spectrum of social science datasets
3. **Workflow-Embracing:** fits into workflow of practicing social scientists, using familiar concepts and tools
Overview

Our goal is Privacy Protection that is:

1. **Accessible:** no DP expert optimizing algorithms for a particular dataset or application

2. **General:** works across the spectrum of social science datasets

3. **Workflow-Embracing:** fits into workflow of practicing social scientists, using familiar concepts and tools

4. **Additive:** strong privacy protection *increases* the ability to explore and share and open data

(cf. Census PUMS vs RDC's)
Overview

Our goal is Privacy Protection that is:

1. **Accessible**: no DP expert optimizing algorithms for a particular dataset or application
2. **General**: works across the spectrum of social science datasets
3. **Workflow-Embracing**: fits into workflow of practicing social scientists, using familiar concepts and tools
4. **Additive**: strong privacy protection increases the ability to explore and share and open data
5. **Tiered Access**: DP interface for wide access to rough statistical information; users can still apply for raw data (cf. Census PUMS vs RDC’s)
27 Instances around the world
Used by researchers from > 500 institutions
75,000 datasets in Harvard Dataverse repository
http://dataverse.org
Datasets are restricted due to privacy concerns

Goal: enable wider sharing while protecting privacy
Actors

**data depositors:** Come to deposit their sensitive dataset in a repository, and may wish to make DP access available.
Actors

**data depositors:** Come to deposit their sensitive dataset in a repository, and may wish to make DP access available.

**data curators:** Maintain the hardware and software on which PSI runs and the accompanying repository infrastructure.
Actors

**data depositors:** Come to deposit their sensitive dataset in a repository, and may wish to make DP access available.

**data curators:** Maintain the hardware and software on which PSI runs and the accompanying repository infrastructure.

**data analysts:** Come to access sensitive datasets in the repository, often with the goal of data exploration.
Actors

Data depositors:
- Level of Trust: Trusted

Data curators:
- Level of Trust: Trusted

Data analysts:
- Level of Trust: Untrusted
  - Semi-Trusted

Responsibilities

- Data depositors:
  - Define the Appropriate Privacy Protection

- Data curators:
  - Maintain Integrity of Software System

- Data analysts:
  - None
  - Abide by DUA
workflow

Upload ➞ Budget ➞ Release ➞ Explore ➞ Query

http://psiprivacy.org/about
workflow

Upload ➔ Budget ➔ Release ➔ Explore ➔ Query
workflow

Upload ➔ Budget ➔ Release ➔ Explore ➔ Query
workflow

Upload -> Budget -> Release -> Explore -> Query
workflows:

1. Upload
2. Budget
3. Release
4. Explore
5. Query
Upload \rightarrow Budget \rightarrow Release \rightarrow Explore \rightarrow Query
http://psiprivacy.org/
http://psiprivacy.org/about
http://2ra.vn/demos/PSIdemoCS208.mp4
Ektelo
Ektelo

- Ektelo supports a library of vetted operators

- Operators encode best practices from literature
  - Allow sophisticated algorithms to be expressed

- Algorithm expressed as plan: a sequence of operators

- Every plan automatically satisfies differential privacy.

[SIGMOD 2018] Zhang, McKenna, Kotsogiannis, Hay, M, Miklau
“Ektelo: A framework for describing differentially private algorithms”
Ektelo Implementation

Plan Authors (non-experts) — Firewall — Privacy Engineers (experts)

Plan Executor — Operator Library — Privacy Engine

Client Space — Protected Kernel

D
**Operator classes and types**

<table>
<thead>
<tr>
<th>Transform</th>
<th>Partition selection</th>
<th>Query selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>T-Vectorize</td>
<td>PA: AHPpartition</td>
</tr>
<tr>
<td>TP</td>
<td>V-SplitByPartition</td>
<td>PG: Grid</td>
</tr>
<tr>
<td>TR</td>
<td>V-ReduceByPartition</td>
<td>PD: Dawa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PW: Workload-based</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PS: Stripe(attr)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PM: Marginal(attr)</td>
</tr>
<tr>
<td>Inference</td>
<td>Query</td>
<td></td>
</tr>
<tr>
<td>NLS</td>
<td>Nneg Least squares</td>
<td></td>
</tr>
<tr>
<td>MW</td>
<td>Mult Weights</td>
<td></td>
</tr>
<tr>
<td>HR</td>
<td>Thresholding</td>
<td></td>
</tr>
</tbody>
</table>

**Theorem:** If red and orange operators are vetted, then any Ektelo plan satisfies DP
**Code Reuse:** Existing algorithms implemented with reusable operators. **Unified 16 implementations of Laplace mechanism**

<table>
<thead>
<tr>
<th>ID</th>
<th>Cite</th>
<th>Algorithm name</th>
<th>Plan signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[8]</td>
<td>Identity</td>
<td>SI LM</td>
</tr>
<tr>
<td>2</td>
<td>[39]</td>
<td>Privelet</td>
<td>SP LM LS</td>
</tr>
<tr>
<td>3</td>
<td>[17]</td>
<td>Hierarchical (H2)</td>
<td>SH2 LM LS</td>
</tr>
<tr>
<td>4</td>
<td>[34]</td>
<td>Hierarchical Opt (HB)</td>
<td>SHB LM LS</td>
</tr>
<tr>
<td>5</td>
<td>[22]</td>
<td>Greedy-H</td>
<td>SG LM LS</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Uniform</td>
<td>ST LM LS</td>
</tr>
<tr>
<td>7</td>
<td>[15]</td>
<td>MWEM</td>
<td>I: ( SW LM MW  )</td>
</tr>
<tr>
<td>8</td>
<td>[42]</td>
<td>AHP</td>
<td>PA TR SI LM LS</td>
</tr>
<tr>
<td>9</td>
<td>[22]</td>
<td>DAWA</td>
<td>PD TR SG LM LS</td>
</tr>
<tr>
<td>10</td>
<td>[6]</td>
<td>Quadtrees</td>
<td>SQ LM LS</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>UniformGrid</td>
<td>SU LM LS</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>DAWA-Striped</td>
<td>PS TP[ PD TR SG LM] LS</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>HB-Striped</td>
<td>PS TP[ SHB LM] LS</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>PrivBayesLS</td>
<td>SPB LM LS</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>MWEM variant b</td>
<td>I: ( SW SH2 LM MW )</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>MWEM variant c</td>
<td>I: ( SW LM NLS )</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>MWEM variant d</td>
<td>I: ( SW SH2 LM NLS )</td>
</tr>
</tbody>
</table>

**New variants of algorithms are easy to construct**
Covert Channels
Other Issues in Programming DP

- Multi-relational databases
  - Standard joins have unbounded Lipschitz constant, so need to truncate
- Side-channel attacks
  - Info can be leaked through timing, approx. of real numbers, global state, exceptions, etc.
  - Constrain language & implementation to match model better.
- Verifying DP building blocks for more complex DP algs.
  - Specialized programming languages
  - Annotate programs with types to assist automated verification of DP
  - Trade-off between usability and expressiveness
  - Now can even synthesize DP algorithms from examples!
- Guidance on Privacy Budgeting
- Choice of Programming Model (eg. SQL vs. MapReduce vs. R)
Covert Channels/Side Channels

- Timing
- Floating Point
- Randomization
- Globals