Current Developments in Differential Privacy

Salil Vadhan
Center for Research on Computation & Society
School of Engineering & Applied Sciences
Harvard University
Def [DMNS06]: A randomized algorithm $C$ is $\epsilon$-differentially private iff for all databases $D, D'$ that differ on one row, and all query sequences $q_1, \ldots, q_t$

$$\text{Distribution of } C(D, q_1, \ldots, q_t) \approx \perp \epsilon \text{ Distribution of } C(D', q_1, \ldots, q_t)$$

“My data has little influence on what the analysts see”
Def [DMNS06]: A randomized algorithm $C$ is $\varepsilon$-differentially private iff for all databases $D, D'$ that differ on one row, all query sequences $q_1, \ldots, q_t$, and all sets $T \subseteq \mathbb{R}^t$,

$$\Pr[C(D, q_1, \ldots, q_t) \in T] \leq (1 + \varepsilon) \cdot \Pr[C(D', q_1, \ldots, q_t) \in T]$$

$\varepsilon$ small constant, e.g. $\varepsilon = 0.01$
Differential Privacy: Key Points

![Diagram showing database D' and function C]

Distribution of $C(D,q_1,\ldots,q_t) \approx \downarrow \varepsilon$ Distribution of $C(D',q_1,\ldots,q_t)$

- **Idea:** inject random noise to obscure effect of each individual
  - Not necessarily by adding noise to answer!

- **Good for Big Data:** more utility and more privacy as $n \to \infty$. 
Distribution of $C(D, q_1, \ldots, q_t) \approx \downarrow \varepsilon$ Distribution of $C(D', q_1, \ldots, q_t)$

- **Strong guarantees:** for all databases, regardless of adversary’s auxiliary knowledge

- **Scalable:** don’t require privacy expert in the loop for each database or release
Some Differentially Private Algorithms

- histograms [DMNS06]
- contingency tables [BCDKMT07, GHRU11, TUV12, DNT14],
- machine learning [BDMN05, KLNRS08],
- regression & statistical estimation [CMS11, S11, KST11, ST12, JT13]
- clustering [BDMN05, NRS07]
- social network analysis [HLMJ09, GRU11, KRSY11, KNRS13, BBDS13]
- approximation algorithms [GLMRT10]
- singular value decomposition [HR12, HR13, KT13, DTTZ14]
- streaming algorithms [DNRY10, DNPR10, MMNW11]
- mechanism design [MT07, NST10, X11, NOS12, CCKMV12, HK12, KPRU12]
- ...

See Simons Institute Workshop on Big Data & Differential Privacy 12/13
Amazing Possibility I: Synthetic Data

Theorem [BLR08,HR10]: If $n \gg d$, can generate differentially private synthetic data preserving exponentially many statistical properties of dataset (e.g. fraction of people w/each set of attributes).

- Computational complexity is a challenge [DNRRV09,UV11,U13]
- Practical implementations in [HLM12,GGHRW14]
Q: How could this be possible?

Would be easy if we could compromise privacy of “just a few” people.

- A few random rows preserves many statistical properties.
- Differential privacy doesn’t allow this.
Amazing Possibility I: Synthetic Data

Q: How could this be possible?

Construct a “smooth” distribution on synthetic datasets (via [MT07])

- Put higher probability on synthetic datasets that agree more with real dataset on statistics of interest.
- Ensure (Probability of each inaccurate synthetic dataset) \( \times (\# \text{ of synthetic datasets}) \) is very small.
### Amazing Possibility II: Statistical Inference & Machine Learning

#### Theorem [KLNRS08,S11]: Differential privacy for vast array of machine learning and statistical estimation problems with little loss in convergence rate as $n \to \infty$.

- Optimizations & practical implementations for logistic regression, ERM, LASSO, SVMs in [RBHT09,CMS11,ST13,JT14].

<table>
<thead>
<tr>
<th>$\geq 40$</th>
<th>Smoke?</th>
<th>Cancer?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Hypothesis $h$ or model $\theta$ about world, e.g. rule for predicting disease from attributes.

- Smoke?
- Cancer?
- $n$
Challenges for DP in Practice

• Accuracy for “small data” (moderate values of $n$)

• Modelling & managing privacy loss over time
  – Especially over many different analysts & datasets

• Analysts used to working with raw data
  – One approach: “Tiered access” – DP for wide access, raw data only by approval with strict terms of use (cf. Census PUMS vs. RDCs)

• Cases where privacy concerns are not “local” (e.g. privacy for large groups) or utility is not “global” (e.g. targeting)

• Matching guarantees with privacy law & regulation

• ...

Some Efforts to Bring DP to Practice

- CMU-Cornell-PennState “Integrating Statistical and Computational Approaches to Privacy”
  - See http://onthemap.ces.census.gov/
- UCSD “Integrating Data for Analysis, Anonymization, and Sharing” (iDash)
- UT Austin “Airavat: Security & Privacy for MapReduce”
- UPenn “Putting Differential Privacy to Work”
- Stanford-Berkeley-Microsoft “Towards Practicing Privacy”
- Duke-NISSS “Triangle Census Research Network”
- Harvard “Privacy Tools for Sharing Research Data”
- ...
A project at Harvard: “Privacy Tools for Sharing Research Data”

http://privacytools.seas.harvard.edu/

- Computer Science, Social Science, Law, Statistics

- **Goal:** to develop technological, legal, and policy tools for sharing of personal data for research in social science and other fields.

- Supported by an NSF Secure & Trustworthy Cyberspace “Frontier” grant and seed funding from Google.
### Goal:

Use differential privacy to widen access.

### Many datasets are restricted due to privacy concerns:

- **006271H-ID-InterGenerational-Coded-Data.pdf**
  - Description of coded data variables

- **006271H-ID-InterGenerational-BerkSpou-Data.sav**
  - Data on Spouses in Berkeley Sample in SPSS Portable Format

- **006271H-ID-InterGenerational-BerkSubj-Data.sav**
  - Data on Subjects in Berkeley Sample in SPSS Portable Format

- **006271H-ID-InterGenerational-BerkSpou-Data.tab**
  - Data on Spouses in Berkeley Sample in Tab Portable Format

- **006271H-ID-InterGenerational-BerkSubj-Data.tab**
  - Data on Subjects in Berkeley Sample in Tab Portable Format
For non-restricted datasets, can run many statistical analyses ("Zelig methods") through the Dataverse interface, without downloading data.
• We’d make PrivateZelig an option, the interface would stay roughly the same
• For sensitive datasets PrivateZelig might be the only option
Dataverse Analysis

The following are the results of your requested analysis.

Summary Results

privatezelig(formula=..., model="logit", DPalg="smith", eps=0.1)
- Call: zelig(formula = sex ~ class + age + ed1hour + ed2hour, model = "logit", data = data)

Deviance Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-8.4904</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0001</td>
<td>8.4904</td>
</tr>
</tbody>
</table>

Coefficients:

| Term                 | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------------|----------|------------|---------|---------|
| (Intercept)          | 2.0761e+13| 2.5442e+13 | 0.8160  | 0.4145  |
| class                | 5.9152e-03| 3.9310e-01 | 0.0150  | 0.9880  |
| age                  | -2.0761e+13| 2.5442e+13 | -0.8160 | 0.4145  |
| ed1hour10012835      | 4.1522e+13| 5.0883e+13 | 0.8160  | 0.4145  |
| ed1hour100285552     | 8.3044e+13| 1.0177e+14 | 0.8160  | 0.4145  |
| ed1hour1004600704    | 6.2283e+13| 7.6325e+13 | 0.8160  | 0.4145  |
| ed1hour100926200     | 6.2283e+13| 7.6325e+13 | 0.8160  | 0.4145  |
| ed1hour1011177792    | 1.0381e+14| 1.2721e+14 | 0.8160  | 0.4145  |
| ed1hour1011535104    | 1.0381e+14| 1.2721e+14 | 0.8160  | 0.4145  |

You could get information about what alg we ran, the privacy param, etc.

Analysis would come back in the same format
Conclusions

Differential Privacy offers

• Strong, scalable privacy guarantees
• Compatibility with many types of “big data” analyses
• Amazing possibilities for what can be achieved in principle

There are some challenges, but reasons for optimism

• Intensive research effort from many communities
• Some successful uses in practice already
• Differential privacy easier as $n \to \infty$