Motivation

- **Last time:** on a dataset with $n$ individuals, releasing $m = n$ counts with error $E = o(\sqrt{n})$ allows for reconstructing $1 - o(1)$ fraction of sensitive attributes.

- **Q:** what happens if we allow error $\Omega(\sqrt{n}) \leq E \leq o(n)$?

- **A (today):** if we release $m = n^2$ counts, can be vulnerable to “membership attacks”.
What is this $\sqrt{n}$ threshold?

- If $X = X_1 + \cdots + X_n$ for independent random variables $X_i$ each with standard deviation $\sigma$, then the standard deviation of $X$ is $\sigma \cdot \sqrt{n}$.
- So the “sampling error” for a sum is typically $\Theta(\sqrt{n})$.
- If the $X_i$’s are bounded (or “subgaussian”), then $X$ will have Gaussian-like concentration around its mean $\mu$: 
  \[ \Pr[|X - \mu| > t \cdot \sqrt{n}] \leq e^{-\Omega(t^2)} \] [Chernoff-Hoeffding Bound]
Normalized Counts (i.e. Averages)

- If \( X = (X_1 + \cdots + X_n)/n \) for independent random variables \( X_i \) each with standard deviation \( \sigma \), then the standard deviation of \( X \) is \( \sigma/\sqrt{n} \).
- So the “sampling error” for a sum is typically \( \Theta(1/\sqrt{n}) \).
- If the \( X_i \)'s are bounded (or “subgaussian”), then \( X \) will have Gaussian-like concentration around its mean \( \mu \): 
  \[
  \Pr[|X - \mu| > t/\sqrt{n}] \leq e^{-\Omega(t^2)} \quad \text{[Chernoff-Hoeffding Bound]}
  \]

This is why subsampling \( k \) out of \( n \) rows allows us to approximate \( m \) averages each to within \( \pm O \left( \left( \frac{1}{\sqrt{k}} \right) \cdot \sqrt{\log m} \right) \)
Motivation

- **Last time:** on a dataset with $n$ individuals, releasing $m = n$ averages with error $E = o\left(\frac{1}{\sqrt{n}}\right)$ allows for reconstructing $1 - o(1)$ fraction of sensitive attributes.

- **Q:** what happens if we allow error $\Omega\left(\frac{1}{\sqrt{n}}\right) \leq E \leq o(1)$?

- **A (today):** if we release $m = n^2$ counts, can be vulnerable to “membership attacks”.
Membership Attacks: Setup

Attacker gets:
- Access to mechanism outputs
- Alice’s data
- (Possibly) auxiliary info about population

Then decides: if Alice is in the dataset $X$

[slide based on one from Adam Smith]
Membership Attacks: Examples

- Genome-wide Association Studies [Homer et al. `08]
  - release frequencies of SNP’s (individual positions)
  - determine whether Alice is in “case group” [w/a particular diagnosis]
- ML as a service [Shokri et al. `17]
  - apply models trained on X to Alice’s data

[slide based on one from Adam Smith]
### Membership Attacks from Means

#### Data Table

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Attributes/Predicates

- \( d \) attributes/predicates

#### Population

0.5 0.75 0.5 0.5 0.75 0.5 0.25 0.25 0.5

#### Alice’s Data

0 1 0 1 1 1 1 0 1 0

#### Attacker

0.5 0.75 0.5 0.75 0.5 0.25 0.25 0.5

#### “In”/“Out”

[slide based on one from Adam Smith]
**Membership Attacks from Means**

- Population = [vector $p = (p_1, \ldots, p_d)$ of probabilities]
  - $j$’th attribute = iid Bernoulli($p_j$), independent across $j$
  - Adversary gets $p$ (or a few random draws) given to adversary

[slide based on one from Adam Smith]
• Population = vector \( p = (p_1, \ldots, p_d) \) of probabilities
  – \( j \)'th attribute = iid Bernoulli(\( p_j \)), independent across \( j \)
  – Adversary gets \( a \approx \bar{x} \) and \( p \) (or a few random draws)
  – Only assume that \( a = M(x) \) has \( |a_j - \bar{x}_j| \leq \alpha \) whp.
    (“Noise” need not be independent or unbiased.)

[slide based on one from Adam Smith]
We are interested in $\alpha > 1/\sqrt{n}$.
In this regime, if $p$ known to mechanism, can prevent attack. (Q: Why?)
So we will assume random $p_j$'s (e.g. iid uniform in $[0,1]$).
Theorem [Dwork et al. `15]: There is a constant $c$ and an attacker $A$ such that when $d \geq cn$ and $\alpha < \min \left\{ \sqrt{d/O(n^2 \log(1/\delta))}, 1/2 \right\}$:

- If Alice is IN, then $\Pr[A(y, a, p) = \text{IN}] \geq \Omega \left( \frac{1}{\alpha^2 n} \right)$.
- If Alice is OUT, then $\Pr[A(y, a, p) = \text{IN}] \leq \delta$. 

[slide based on one from Adam Smith]
**Theorem [Dwork et al. `15]:** There is an attacker $A$ such that when $d \geq O(n)$ and $\alpha < \min \left\{ \sqrt{d/\Theta(n^2 \log(1/\delta))}, 1/2 \right\}$:

- If Alice is IN, then $\Pr[A(y, a, p) = \text{IN}] \geq \Omega \left( \frac{1}{\alpha^2 n} \right)$. (true positive)
- If Alice is OUT, then $\Pr[A(y, a, p) = \text{IN}] \leq \delta$. (false positive)

**Remarks:**

- Only interesting when $\delta < \Omega \left( \frac{1}{\alpha^2 n} \right)$.
- On average, successfully trace $\Omega \left( \frac{1}{\alpha^2} \right)$ members of dataset. This is the best possible. (Why?)
- Can safely release at most $\tilde{O}(n^2)$ means!
The Attacker

\[
A(y, a, p) = \begin{cases} 
\text{IN} & \text{if } \langle y, a \rangle - \langle p, a \rangle > T \\
\text{OUT} & \text{if } \langle y, a \rangle - \langle p, a \rangle \leq T 
\end{cases}
\]

Note: given \( p, a \), can choose \( T = T_{p,a} = \frac{1}{\sqrt{d \log(1/\delta)}} \) to make false positive probability exactly \( \delta \).

[slide based on one from Adam Smith]
Attacks on Aggregate Stats

- What error $\alpha$ makes sense?
  - Estimation error due to sampling $\approx 1/\sqrt{n}$
  - Reconstruction attacks require $\alpha \lesssim 1/\sqrt{n}$, $d \geq n$
  - Robust membership attacks: $\alpha \lesssim \sqrt{d}/n$

- Lessons
  - “Too many, too accurate” statistics reveal individual data
  - “Aggregate” is hard to pin down
Membership Attacks on ML as a Service

[Shokri et al. 2017]
Switch to slides from Reza Shokri’s talk
Another Attack on ML?

[Frederickson et al. `14, cf. McSherry `16]

$n$ people

Data set $X$

Mechanism (stats, ML model, …)

Population

Alice’s (known) data

Attacker gets:

- Access to mechanism outputs
- Some of Alice’s data
- (Possibly) auxiliary info about population

Then computes: a sensitive attribute of Alice
Another Attack on ML?

[Frederickson et al. `14, cf. McSherry `16]

\[ \text{Data set X} \]

Population

\[ \text{Alice's (known) data} \]

\[ \text{Mechanism (stats, ML model, ...)} \]

\[ n \text{ people} \]

Difference from reconstruction attacks:

- Above attack works even if Alice not in dataset. Based on correlation between known & sensitive attributes.
- Reconstruction attacks work even when sensitive bit uncorrelated.
Goals of Differential Privacy

• **Utility:** enable “statistical analysis” of datasets
  – e.g. inference about population, ML training, useful descriptive statistics

• **Privacy:** protect individual-level data
  – against “all” attack strategies, auxiliary info.

**Q:** Can it help with privacy in microtargetted advertising? [Korolova attacks]
  – inference from impressions?
  – inference from clicks?
  – displaying intrusive ads?
Further Discussion

Reactions to the “Five Views” responses to membership attacks on genomic data?