Fingerprinting Codes and the Price of Approximate Differential Privacy

June 1, 2014

Mark Bun
Jonathan Ullman
Salil Vadhan

Harvard University
Privacy-Preserving Data Analysis

Want curators that are:  
- Private  
- Accurate  
- Efficient
Privacy-Preserving Data Analysis

Want curators that are:  ◆ Private ◆ Accurate ◆ Efficient
Privacy-Preserving Data Analysis

Want curators that are:  
- Differentially Private  
- Statistically Accurate  
- Sample Efficient
What This Talk is About

- **Sample complexity** for approx. differential privacy

- **MAIN RESULT:** For high-dimensional data, **Privacy + Accuracy** requires more samples than **Accuracy** alone
  
  e.g. $d$ attribute means
  
  - Accuracy: $\Theta(\log d)$
  - Privacy + Accuracy: $\tilde{\Theta}(d^{1/2})$

- New techniques for privacy lower bounds
Differential Privacy

[DN03+Dwork, DN04, BDMN05, DMNS06, DKMMN06]

**D**

| $x_1$ | $x_2$ | $\vdots$ | $x_n$ |

**$D'$**

| $x_1$ | $x_2'$ | $\vdots$ | $x_n$ |

$D$ and $D'$ are neighbors if they differ on one row.

$M$ is $(\varepsilon, \delta)$-differentially private if for all neighbors $D$, $D'$ and $T \subseteq \text{Range}(M)$:

$$\Pr[M(D') \in T] \leq (1 + \varepsilon)\Pr[M(D) \in T] + \delta$$

small const., e.g. $\varepsilon = 0.1$

“cryptographically small” need $\delta \ll 1/n$, often $\delta = \text{negl}(n)$

- Privacy
- Accuracy
- Sample Complexity
Counting Queries

“What fraction of the rows of $D$ satisfy some property $q$?”

E.g. attribute means
$q = \text{Skywalker?}$
$q(D) = 3/4$

$M$ is $\alpha$-accurate for $Q$ if
$|a_i - q_i(D)| < \alpha$ for every $i$

Privacy, Accuracy, Sample Complexity
(Privately) Answering Attribute Means

[DN03, DN04, BDMN05, DMNS06]

$d$ binary attributes

<table>
<thead>
<tr>
<th></th>
<th>DarkSide?</th>
<th>Twin?</th>
<th>Skywalker?</th>
<th>&lt; 3ft?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$n$ rows

3/4
+ Noise(O(1/n))

(\(\alpha\)-accuracy requires \(n \geq 1/\alpha\))

❖ Privacy  ❖ Accuracy  ❖ Sample Complexity
(Privately) Answering Attribute Means
[DN03, DN04, BDMN05, DMNS06]

$d$ binary attributes

<table>
<thead>
<tr>
<th></th>
<th>DarkSide?</th>
<th>Twin?</th>
<th>Skywalker?</th>
<th>&lt; 3ft?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$n$ rows

\[
\begin{align*}
\text{Noise(O}(d^{1/2}/n)\text{)} & \quad + & \quad \text{Noise(O}(d^{1/2}/n)\text{)} & \quad + & \quad \text{Noise(O}(d^{1/2}/n)\text{)} & \quad + & \quad \text{Noise(O}(d^{1/2}/n)\text{)} \\
1/4 & \quad + & \quad 1/2 & \quad + & \quad 3/4 & \quad + & \quad 1/4
\end{align*}
\]

(\(\alpha\)-accuracy requires \(n \geq d^{1/2}/\alpha\))

- Privacy
- Accuracy
- Sample Complexity
Sample Complexity

How big does \( n \) have to be to guarantee statistical accuracy on the population?

<table>
<thead>
<tr>
<th>( q_1, q_2, \ldots \in \mathbb{Q} )</th>
<th>( a_1, a_2, \ldots )</th>
<th>( \alpha )-accurate answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 1 1 0</td>
<td>0 0 1 1 1 0</td>
<td>0 0 1 1 1 0</td>
</tr>
<tr>
<td>0 1 1 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 1 1 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 0 1 0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( d \) binary attributes

\( n \) rows
Sample Complexity

Answer: $n = \Theta(\log |Q| / \alpha^2)$ [Vap98]

e.g. $\Theta(\log d)$ for attribute means with $\alpha = 0.05$
How big does $n$ have to be to guarantee accuracy and privacy?

$M_{q_1, q_2, \ldots} \in \mathbb{Q}$

$d$ binary attributes

$n$ rows

$(1, o(1/n))$-diff. priv.

$\alpha$-accurate answers

Privacy

Accuracy

Sample Complexity
Sample Complexity for Diff. Privacy

Question: Is there an additional price of diff. privacy over statistical accuracy alone?

\[ (1, o(1/n))-\text{diff. priv.} \]

\[ q_1, q_2, \ldots \in Q \]

\[ a_1, a_2, \ldots \]

\[ \alpha\text{-accurate answers} \]
## Sample Complexity for Diff. Privacy

### No privacy

<table>
<thead>
<tr>
<th>Q = attribute means</th>
<th>Q, α arbitrary</th>
</tr>
</thead>
<tbody>
<tr>
<td>α = 0.05</td>
<td>α = arbitrary</td>
</tr>
</tbody>
</table>

| n = Θ(log d)        | n = Θ(log |Q|/α²) |
|---------------------|----------------|
| [Vap98]             | [Vap98]        |

### (0.1, o(1/n))-diff. privacy

#### Upper bound:

| Ő(d^{1/2})          | ∀ Q: Ő(log |Q| • d^{1/2}/α²) |
|---------------------|----------------------------|
| [...DMNS06]         | [HR10]                      |

#### Lower bound:

| Ō(log d)            | ∀ Q: max ŕ(log |Q|/α), ŕ(1/α²) |
|---------------------|-----------------------------|
| [DN03, Rot10]       | [DN03]                      |

#### OUR WORK:

| ŕ(d^{1/2})          | ∀ Q: ŕ(log |Q| • d^{1/2}/α²) |
|---------------------|----------------------------|

### Privacy • Accuracy • Sample Complexity
Beyond Reconstruction Attacks

• Tight lower bounds known for (ε, 0)-diff. privacy [HT10, Har11], but break even for δ = negl(n) [De11, BNS13]

• Prior lower bounds for (ε, δ)-diff. privacy gave reconstruction attacks [DN03, Rot10], which hold even for δ = constant

• **This work:** Fingerprinting codes enable optimal lower bounds for (ε, δ=o(1/n))-diff. privacy (followed by [DTTZ14, BST14])
New Techniques

• Fingerprinting codes $\Rightarrow$ diff. privacy lower bounds
  $\tilde{\Omega}(d^{1/2})$ for attribute means ($\alpha$ const.)

• Composition of sample complexity lower bounds
  $\tilde{\Omega}(kd^{1/2})$ for $k$-way conjunctions ($\alpha$ const.)
  $\tilde{\Omega} (\log |Q| \cdot d^{1/2}/\alpha^2)$ for arbitrary queries
Fingerprinting Codes [BS95]

I want to distribute my new movie

...but the galaxy is full of pirates!
Fingerprinting Codes [BS95]

I want to distribute my new movie

...but the galaxy is full of pirates!

Who collude against me!
Fingerprinting Codes [BS95]
Fingerprinting Codes \[\text{[BS95]}\]

Gen(1^n) outputs $C \subseteq \{0,1\}^d$^n

For all coalitions $S$ and all pirate alg. for producing $w$,

$\text{Pr}[\text{Trace}(w, C) \in S] \approx 1$
FP Codes vs. Diff. Privacy

Coalition of $n$ pirates

Feasible codeword $w$

Pr[Trace($w, C$) = ] $\geq 1/n$
FP Codes vs. Diff. Privacy

Coalition of $n$ pirates

Feasible codeword $w$

Pr[Trace($w, C$) = $\ldots$] $\ll 1/n$
FP Codes vs. Diff. Privacy

Trace behaves very differently depending on whether is in the coalition

Fingerprinting codes are the “opposite” of differential privacy!

(Parallels computational lower bounds via traitor-tracing schemes [DNRRV09, U13])
Lower Bound for Attribute Means

Database of $n$ users

Suppose (for contradiction) we have

- A FP code of length $d$ for $(n+1)$ users
- A diff. private $M$ that is accurate for attribute means on $(\{0,1\}^d)^n$

Reduction: Use $M$ to break security of the FP code
**Lower Bound for Attribute Means**

Database of $n$ users = Coalition of $n$ pirates

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Suppose (for contradiction) we have

- A FP code of length $d$ for $(n+1)$ users
- A diff. private $M$ that is accurate for attribute means on $(\{0,1\}^d)^n$

Reduction: Use $M$ to break security of the FP code

| .23 | .87 | .95 | .03 | .79 |

Round answers $w$

$M$ accurate $\Rightarrow w$ feasible
Lower Bound for Attribute Means

Database of $n$ users = Coalition of $n$ pirates

$\Pr[\text{Trace}(w) = \text{M}] \geq 1/n$
Lower Bound for Attribute Means

Database of $n$ users = Coalition of $n$ pirates

Contradicts security of FP code!

$$\Pr[\text{Trace}(w) = \text{False}] \geq \frac{(1/n) - \delta}{1 + \epsilon} \geq \frac{1}{3n}$$

M private => Trace fails

M accurate => $w$ feasible

Round answers
Lower Bound for Attribute Means

• \( \exists \) FP code for \( n \) users with length \( d \)
  \[ \Rightarrow d \] attribute means require \( n \) samples

• [Tar03] \( \exists \) FP code for \( \tilde{\Omega}(d^{1/2}) \) users of length \( d \)
  \[ \therefore \] attribute means require \( n \geq \tilde{\Omega}(d^{1/2}) \)
Sample Complexity for Diff. Privacy

No privacy

Q = attribute means
α = 0.05

Q, α arbitrary

\[ n = \Theta(\log d) \]
[Vap98]

\[ n = \Theta(\log |Q|/\alpha^2) \]
[Vap98]

(1, o(1/n))-diff. privacy

Upper bound:

\[ \tilde{O}(d^{1/2}) \]
[...DMNS06]

\[ \forall Q: \tilde{O}(\log |Q| \cdot d^{1/2}/\alpha^2) \]
[HR10]

Lower bound:

\[ \tilde{\Omega}(\log d) \]
[DN03, Rot10]

\[ \exists Q: \max \tilde{\Omega}(\log |Q|/\alpha), \tilde{\Omega}(1/\alpha^2) \]
[DN03]

OUR WORK:

\[ \tilde{\Omega}(d^{1/2}) \]

\[ \exists Q: \tilde{\Omega}(\log |Q| \cdot d^{1/2}/\alpha^2) \]
Conclusions

• Fingerprinting codes yield privacy violations beyond reconstruction attacks

• Price of $(\epsilon,\delta)$-diff. privacy for high-dimensional data

• Open questions:
  – Sample complexity of computationally efficient algorithms for $k$-way conjunctions?
    [e.g. BCD+07, GHRU11, UV11, TUV12, DNT13, CTUW14]
  – Combinatorial characterization of sample complexity?
    [e.g. HT10, Har11, NTZ13, BNS13]
Thank you!