Sample Complexity of Differential Privacy

Mark Bun*
Harvard University (2nd year Ph.D., supported by an NDSEG fellowship)

MAIN QUESTION

How many data samples do we need to achieve both differential privacy and statistical accuracy?

i.e. How big a study do we need to conduct to answer our questions and preserve privacy?

DIFFERENTIAL PRIVACY

<table>
<thead>
<tr>
<th>x₁</th>
<th>x₂</th>
<th>x₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>San</td>
<td>San</td>
<td>x₁</td>
</tr>
<tr>
<td>x₂</td>
<td>x₃</td>
<td>x₄</td>
</tr>
</tbody>
</table>

D and D' are neighbors if they differ only on one user's data

An algorithm San is (ε, δ)-differentially private if for all neighbors D, D' and every S ⊆ Range(San),

Pr[San(D) ∈ S] ≤ e^ε Pr[San(D') ∈ S] + δ

Think of ε = Θ(1) and δ = o(1/n)

ACCURACY FOR COUNTING QUERIES

Counting queries: What fraction of rows in a database satisfy property q?

e.g. q(x) = LikesBread AND LikesToast

<table>
<thead>
<tr>
<th>LikesBread</th>
<th>LikesToast</th>
<th>LikesWheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d (~4) attributes per record</th>
</tr>
</thead>
<tbody>
<tr>
<td>q(x)=0</td>
</tr>
<tr>
<td>q(x)=1</td>
</tr>
<tr>
<td>q(x)&gt;2/3</td>
</tr>
</tbody>
</table>

Answers aᵦ are α-accurate if |aᵦ - q(D)| < α for every q ∈ Q

SAMPLE COMPLEXITY UPPER BOUNDS

For general queries Q,

O(√d log |Q| / α²)

samples suffice [HR10], using the analysis in [GRU12]

But for certain Q, the sample complexity can be much lower:

Point queries: POINTᵦ(x) = {1 if x = y
|                  |
|                  |
|                  |
|                  |
| 0                |
| 0               -
| 1               -
| 11              -

(_.1) POINTᵦ = 0.5

Threshold queries: THRESHᵦ(x) = {1 if x ≥ y
|                  |
|                  |
|                  |
|                  |
| 0                |
| 0               -
| 1               -
| 11              -

(_.2) THRESHᵦ = 0.5

Again, log |Q| = d, but just O(1/α) samples suffice [BNS13]

SAMPLE COMPLEXITY LOWER BOUNDS

Our contributions [BUV13]

- To answer arbitrary queries, Ω(√d log |Q| / α²) samples are necessary (nearly tight)

- If α is a constant, this lower bound still holds for conjunction queries

Tool 1: Fingerprinting Codes

Coalition of users S ⊆ [n]

Goal is to answer (most) I-way conj's on at least one Dᵦ
⇒ privacy breach

Goal is to answer (most) (k+1)-way conj's compute “subset sums of (k+1)-way conj’s”

COMPOSITION OF LOWER BOUNDS

 log |Q| / α²

REFERENCES


ACKNOWLEDGMENTS

*Based mostly on joint work with Jonathan Ullman and Salil Vadhan [BUNV13]

Most slides are courtesy of Jonathan Ullman

CONTACT

mbun@seas.harvard.edu
Graduate School of Engineering and Applied Sciences
Cambridge, MA 02138

REFERENCES


