Differential Privacy and PINQ

Benjamin Kilimnik, Paul Xu
Why care? What’s the problem?

Why Differential Privacy?

How does D.P. work?

What is PINQ?

Discussion
Why care?

“Of course, I don't even get out of bed for less than a petabyte”

Keenan
Why care?

“Of course, I even get out of bed for less than a petabyte”

Banji
Activity - Reconstruction from Aggregates

1. There are four people in total
2. Two have age 17
3. Two love Squid-Game
4. Two love Spongebob
5. Average age of Squid-Game lovers is 30
6. Average age of Spongebob lovers is 32
Activity - Reconstruction from Aggregates

1. There are four people in total
2. Two have age 17
3. Two love Squid-Game
4. Two love Spongebob
5. Average age of Squid-Game lovers is 30
6. Average age of Spongebob lovers is 32

<table>
<thead>
<tr>
<th>Age</th>
<th>Fav TV show</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
1. There are four people in total
2. Two have age 17
3. Two love Squid-Game
4. Two love Spongebob
5. Average age of Squid-Game lovers is 30
6. Average age of Spongebob lovers is 32

- From (1), (3), (4): given two people aged 17, either
  - they both love Squid Game (X since avg age would be 17)
  - they both love Spongebob (X since avg age would be 17)
  - or one loves Squid Game, the other Spongebob
1. There are four people in total
2. Two have age 17
3. Two love Squid-Game
4. Two love Spongebob
5. Average age of Squid-Game lovers is 30
6. Average age of Spongebob lovers is 32

<table>
<thead>
<tr>
<th>Age</th>
<th>Fav TV show</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>Squid Game</td>
</tr>
<tr>
<td>17</td>
<td>Spongebob</td>
</tr>
<tr>
<td>43</td>
<td>Squid Game</td>
</tr>
<tr>
<td>47</td>
<td>Spongebob</td>
</tr>
</tbody>
</table>

- From (3), (5): given that avg age of Squid-Game lovers is 30
  - Other Squid-Game lover must be 43
- From (4), (6): given that avg age of Spongebob lovers is 32
  - Other Spongebob lover must be 47
Do the following strategies work?

1. Aggregation
2. Removing personal identifying information (PII)
3. Adding random noise
Do the following strategies work?

1. Aggregation
2. Removing personal identifying information (PII)
3. Adding random noise
   a. Can be averaged across instances to get original value
The Tradeoffs

Information Disclosure vs. Privacy Loss

Accuracy vs. Privacy Loss
Differential Privacy: Assumptions

Assumption about data:
- Very few assumptions other than it can be organized by records
- No assumptions about type or content in records

Assumption about adversary:
- Very rigorous assumption: your adversary has all of your data but one record
Differential Privacy: Intuition

Almost the same

Result

Result

Image credit: https://desfontain.es/privacy/friendly-intro-to-differential-privacy.html
These two intuitions are equivalent

- For any record in the dataset, the result of the noisy computation with and without this record should be almost identical

- An attacker who has all but any one record in the dataset should not be able to figure out with the result of the computation whether that record is in the data or not

“Almost” the same: how small is the difference?
“Almost” the same, you say?

\( \varepsilon \)-differential privacy

\( \varepsilon \): “privacy budget” that quantifies privacy loss. (trade off!)

Note that:

- When \( A \) and \( B \) differ by one record, \( |A \oplus B| = 1 \)
- When \( \varepsilon \) is much less than 1, \( \exp(\varepsilon) \approx 1 + \varepsilon \)
- Bounds the change (ratio) in probability, but does not concern the probability of the event itself
- Nice properties when coupled with Laplace noise

**Definition 1.** We say a randomized computation \( M \) provides \( \varepsilon \)-differential privacy if for any two data sets \( A \) and \( B \), and any set of possible outputs \( S \subseteq \text{Range}(M) \),

\[
\Pr[M(A) \in S] \leq \Pr[M(B) \in S] \times \exp(\varepsilon \times |A \oplus B|).
\]
ε-Differential Privacy

\[
\frac{\Pr[M(A) \in S]}{\Pr[M(B) \in S]} \leq e^\epsilon
\]

Almost the same in probability!

Image credit: https://desfontain.es/privacy/friendly-intro-to-differential-privacy.html
An example: Noisy Count

**Theorem 1.** The mechanism $M(X) = |X| + \text{Laplace}(1/\epsilon)$ provides $\epsilon$-differential privacy.

The Laplace Distribution:

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

$$= \frac{1}{2b} \begin{cases} 
\exp\left(-\frac{\mu-x}{b}\right) & \text{if } x < \mu \\
\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu 
\end{cases}$$

Source: https://en.wikipedia.org/wiki/Laplace_distribution
An example: Noisy Count

Suppose you choose $\varepsilon = 0.5$ and report a count of 100.

What does the attacker see?

Theorem 1. The mechanism $M(X) = |X| + \text{Laplace}(1/\varepsilon)$ provides $\varepsilon$-differential privacy.

Source: https://github.com/OpenMined/PyDP/blob/dev/examples/laplace_demo/laplace.ipynb
The effect of $\epsilon$ - the privacy budget
What is PINQ?

- Language to make transformations and aggregates differentially private (think Pandas, but for .NET)

<table>
<thead>
<tr>
<th>What PINQ can do</th>
<th>What it can’t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safely perform calculations with configurable privacy</td>
<td>Protect against intentional data leaks or malicious code</td>
</tr>
</tbody>
</table>
What is PINQ?

Example 1 Counting searches from distinct users in PINQ.

```csharp
var data = new PINQueryable<SearchRecord>( ... ... );

var users = from record in data
              where record.Query == argv[0]
              groupby record.IPAddress

Console.WriteLine(argv[0] + " : " + users.NoisyCount(0.1));
```
What is PINQ?

Example 5 Measuring query frequencies in PINQ.

```csharp
// prepare data with privacy budget
var agent = new PINQAgentBudget(1.0);
var data = new PINQuery<slice>(rawdata, agent);

// break out fields, filter by query, group by IP
var users = data.Select(line => line.Split(','))
    .Where(fields => fields[20] == args[0])
    .GroupBy(fields => fields[0]);

// output the count to the screen, or anywhere else
Console.WriteLine(args[0] + " : " + users.NoisyCount(0.1));
```
**Stable Transformations**

**Definition 2.** We say a transformation $T$ is $c$-stable if for any two input data sets $A$ and $B$,

$$|T(A) \oplus T(B)| \leq c \times |A \oplus B|.$$  

<table>
<thead>
<tr>
<th>Transformation</th>
<th>C-Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHERE</td>
<td>1</td>
</tr>
<tr>
<td>SELECT</td>
<td>1</td>
</tr>
<tr>
<td>JOIN*</td>
<td>2</td>
</tr>
<tr>
<td>GROUPBY</td>
<td>2</td>
</tr>
</tbody>
</table>
Stable Transformations

**Definition 2.** We say a transformation $T$ is $c$-stable if for any two input data sets $A$ and $B$,

$$|T(A) \oplus T(B)| \leq c \times |A \oplus B|.$$  

<table>
<thead>
<tr>
<th>Transformation</th>
<th>C-Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHERE</td>
<td>1</td>
</tr>
<tr>
<td>SELECT</td>
<td>1</td>
</tr>
<tr>
<td>JOIN</td>
<td>2</td>
</tr>
<tr>
<td>GROUPBY</td>
<td>2</td>
</tr>
</tbody>
</table>

$c^*\epsilon$-differential privacy
Composition

**Sequential:** chaining computations increases privacy cost additively by the sum of epsilon for each sub-computation

**Parallel:** running computations on disjoint sets reduces privacy budget by the worst privacy guarantee of the parallel parts
Why PINQ?

- Non-expert analysts gain privacy guarantees
- Enables arbitrary, composed computations with measurable privacy

“Why Gramma, what big data you have!”
Hypothesis

A data analysis platform like pinq provides non-experts with formal differential privacy guarantees, enabling operations over raw data via a secure and simple API while guarding against unintentional disclosure of sensitive information.
BACKUP SLIDES
Hold up, what is differential privacy?
Netflix Activity: De-anonymization
Netflix Activity: De-anonymization

![Diagram showing the process of de-anonymization with symbols representing likes and dislikes. The anonymized Netflix data is combined with public, incomplete IMDB data to reveal the identities of Molte Keenan, Banji Justos Roj, and Coltun.](image-url)
Netflix Activity: De-anonymization

Identified Netflix Data

Molte
Keenan
Banji
Justos
Roj
Coltun