Differential Privacy I

29.04.20
Background Story

Cafeteria wants to study eating habits of math students.
Perform a study while preserving privacy

- Meal records are uploaded into a table $t$.

<table>
<thead>
<tr>
<th>student</th>
<th>faculty</th>
<th>food</th>
<th>time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Eve</td>
<td>computer science</td>
<td>porridge</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Perform a study while preserving privacy

- Meal records are uploaded into a table $t$.

<table>
<thead>
<tr>
<th>student</th>
<th>faculty</th>
<th>food</th>
<th>time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Eve</td>
<td>computer science</td>
<td>porridge</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- The table $t$ can be secret-shared (e.g. using Sharemind).

- Aggregated statistics will be computed in a privacy-preserving manner.
Do the aggregated statistics violate privacy?

- All data is secret-shared and cannot be seen by anyone.
  - Alice | math | soup | 20
  - Bob  | math | fish | 15
  - Eve  | computer science | porridge | 25
  - ...

- The analyst will only learn the average of eating times.

SELECT AVG(time) FROM t WHERE faculty = math;
Do the aggregated statistics violate privacy?

- All data is secret-shared and cannot be seen by anyone.

<table>
<thead>
<tr>
<th>Alice</th>
<th>math</th>
<th>soup</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Eve</td>
<td>computer science</td>
<td>porridge</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- The analyst will only learn the average of eating times.

SELECT AVG(time) FROM t WHERE faculty = math;

How much can be learned about me from the average?

ALICE
Do the aggregated statistics violate privacy?

- All data is secret-shared and cannot be seen by anyone.

<table>
<thead>
<tr>
<th>Alice</th>
<th>math</th>
<th>soup</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Eve</td>
<td>computer science</td>
<td>porridge</td>
<td>25</td>
</tr>
</tbody>
</table>

- The analyst will only learn the average of eating times.

```
SELECT AVG(time) FROM t WHERE faculty = math;
```

How much can be learned about me from the average?

ALICE
How much impact Alice’s data has on the output?

SELECT AVG(time) FROM t WHERE faculty = math;

The more participants there are, the less effect has Alice’s record on the average.
Which additional knowledge the attacker has?

<table>
<thead>
<tr>
<th>Name</th>
<th>Faculty</th>
<th>Food</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Eve</td>
<td>computer science</td>
<td>porridge</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

SELECT AVG(time) FROM t WHERE faculty = math;
Which additional knowledge the attacker has?

<table>
<thead>
<tr>
<th>Alice</th>
<th>math</th>
<th>soup</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Eve</td>
<td>computer science</td>
<td>porridge</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

SELECT AVG(time) FROM t WHERE faculty = math;

Math students are: Alice, Bob, Chris

Bob ate 15 minutes
Chris ate 10 minutes
Average is 15 minutes

=> Alice ate 20 minutes!
Which additional knowledge the attacker has?

<table>
<thead>
<tr>
<th>Name</th>
<th>Faculty</th>
<th>Meal</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Eve</td>
<td>computer science</td>
<td>porridge</td>
<td>25</td>
</tr>
</tbody>
</table>

Increasing the number of participants will not help if the attacker knows all of them in advance.

Math students are: Alice, Bob, Chris

Bob ate 15 minutes
Chris ate 10 minutes
Average is 15 minutes

=> Alice ate 20 minutes!
What do we want

- Let $\mathcal{D} \in X$ be a database, and $q : X \rightarrow Y$ a query.
- Alice would feel safe participating in the dataset if:
  1. Alice knew that her answer had no impact on the released result.

$$q(\mathcal{D} \setminus AliceData) = q(\mathcal{D})$$

2. Alice knew that any attacker looking at the released result couldn’t learn any new information about herself personally.

$$\Pr[secret(Alice) \mid q(\mathcal{D})] = \Pr[secret(Alice)]$$
Why can’t we have it

1. Not only Alice cares of the privacy, but also Bob, Chris, . . . If all individuals had no impact on the released results, then the results would have no utility.

\[ q(\mathcal{D}) = q(\mathcal{D} \setminus AliceData) \]
\[ = q(\mathcal{D} \setminus (AliceData \cup BobData \cup ChrisData \cup \ldots)) \]
\[ = q(\emptyset) . \]
Why can’t we have it

1. Not only Alice cares of the privacy, but also Bob, Chris, ... If all individuals had no impact on the released results, then the results would have no utility.

\[ q(\mathcal{D}) = q(\mathcal{D} \setminus \text{AliceData}) \]
\[ = q(\mathcal{D} \setminus (\text{AliceData} \cup \text{BobData} \cup \text{ChrisData} \cup ...)) \]
\[ = q(\emptyset) . \]

Achieving either privacy or utility is easy, but getting a meaningful trade-off is the real challenge!
Why can’t we have it

2. If the result $q(D)$ shows there is a strong trend in the dataset, with high probability that trend is true for any individual, even for those who do not participate in the dataset at all.

$$\Pr[\text{secret}(Alice) \mid q(D)] > \Pr[\text{secret}(Alice)].$$

E.g. if the statistics says that math students eat a lot, then probably so does Alice.
Can we still have at least something?

Alice would feel safe participating in the dataset if the chance that the released result will be $y$ is nearly the same whether or not Alice submitted her personal information.

- **Intuition:** Perturb the result (e.g., by adding noise) such that the chance that the perturbed result will be $y$ is nearly the same, whether or not you submit your info.

- **Challenge:** Achieve privacy while minimising the utility loss.
\( \epsilon \)-differential privacy for record’s presence

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>
\( \epsilon \)-differential privacy for record’s presence

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>

We call two tables **adjacent** if they differ in presence of one row.

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>
\(\epsilon\)-differential privacy for record’s presence

<table>
<thead>
<tr>
<th></th>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>

We call two tables adjacent if they differ in presence of one row.

**Differential privacy (DP):** A function \( q : X \rightarrow D(Y) \) is called \(\epsilon\)-DP if, for all \( Y' \subseteq Y \), for all adjacent tables \( t, t' \in X \):

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon
\]

\((e \approx 2.7 \text{ is the Euler number})\)
\( \epsilon \)-differential privacy for record’s presence

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>

We call two tables **adjacent** if they differ in presence of one row.

**Differential privacy (DP):** A function \( q : X \rightarrow \mathcal{D}(Y) \) is called \( \epsilon \)-DP if, for all \( Y' \subseteq Y \), for all adjacent tables \( t, t' \in X \):

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon
\]

\( e \approx 2.7 \) is the Euler number

We need to define probability over subsets \( Y' \subseteq Y \) instead of points \( y \in Y \), since e.g. for \( Y \subseteq \mathbb{R} \) we may have \( \Pr[q(t) = y] = 0 \) for all \( y \in Y \).
Illustration of $\epsilon$-DP

Assume that the distribution $q(t)$ has a probability density function for all $t \in X$.

$$\forall Y' \subseteq Y : \frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon \iff \forall y \in Y : \frac{\text{pdf}_{q(t)}(y)}{\text{pdf}_{q(t')}(y)} \leq e^\epsilon .$$
Interpretation of $e^\epsilon$

Because we can switch $t$ and $t'$ interchangeably, $\epsilon$-DP implies

$$e^{-\epsilon} \leq \frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon.$$ 

Since $e^\epsilon \approx 1 + \epsilon$ for small $\epsilon$, then we have roughly

$$1 - \epsilon \lesssim \frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \lesssim 1 + \epsilon.$$

We see that $\epsilon$ can take any value in $[0, \infty)$. 
Sensitivity of a query

Definition

The **global sensitivity** of a function $q : X \rightarrow \mathbb{R}$ is defined as:

$$GS_q = \max_{t, t' \in X} |q(t) - q(t')|.$$  

$t, t'$ are adjacent

- The higher sensitivity is, the more impact an individual record has on the output.
- Intuitively, for higher sensitivity we need to add more noise.
- We will use notation $\Delta q = GS_q$ as well.
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

SELECT COUNT(*) FROM table;

SELECT COUNT(*) FROM table WHERE food = fish;

SELECT SUM(time) FROM table;

MAX(time) if we do not have bounds on time attribute.

SELECT AVG(time) FROM table;

MAX(time)

Let m be the number of records in t.

Let x be the time of a record that has been removed from t.

Let y be the sum of all times.

\[ |y - (y - x)/m| \leq \text{MAX}(time) \]

Where n is the least possible number of records.
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- SELECT COUNT(*) FROM table;

- SELECT COUNT(*) FROM table WHERE food = fish;

- SELECT SUM(time) FROM table;

- MAX(time) if we do not have bounds on time attribute.

- SELECT AVG(time) FROM table;

- MAX(time)

- Let \( m \) be the number of records in \( t \).

- Let \( x \) be the time of a record that has been removed from \( t \).

- Let \( y \) be the sum of all times.

\[
\left| y \frac{m}{m-1} - y - x \frac{m}{m-1} \right| = \left| x - \frac{y}{m} \right| \frac{m}{m-1} \leq \text{MAX}(time) \frac{m}{m-1} \leq \text{MAX}(time)n.
\]

- Where \( n \) is the least possible number of records.
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- SELECT COUNT(*) FROM table;
  1

- SELECT COUNT(*) FROM table WHERE food = fish;
  1

- SELECT SUM(time) FROM table;

- ∞ if we do not have bounds on time attribute.

- SELECT AVG(time) FROM table;

- Let $m$ be the number of records in $t$.
- Let $x$ be the time of a record that has been removed from $t$.
- Let $y$ be the sum of all times.

\[ \left| \frac{y}{m} - \frac{y-x}{m-1} \right| \leq \max(\text{time}) \frac{m}{m-1} \leq \max(\text{time})^n \]

Where $n$ is the least possible number of records.
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- SELECT COUNT(*) FROM table;
- SELECT COUNT(*) FROM table WHERE food = fish;

WHERE \( n \) is the least possible number of records.
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
</tbody>
</table>

- SELECT COUNT(*) FROM table;
  1
- SELECT COUNT(*) FROM table WHERE food = fish;
  1
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- SELECT COUNT(*) FROM table;
- SELECT COUNT(*) FROM table WHERE food = fish;
- SELECT SUM(time) FROM table;
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- SELECT COUNT(*) FROM table;  
  1

- SELECT COUNT(*) FROM table WHERE food = fish;  
  1

- SELECT SUM(time) FROM table;  
  MAX(time)

  - ∞ if we do not have bounds on time attribute.
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- `SELECT COUNT(*) FROM table;` 1
- `SELECT COUNT(*) FROM table WHERE food = fish;` 1
- `SELECT SUM(time) FROM table;` \( \text{MAX(time)} \)
  - \( \infty \) if we do not have bounds on `time` attribute.
- `SELECT AVG(time) FROM table;`
Some examples of sensitivity

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>math</td>
<td>soup</td>
<td>20</td>
</tr>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>faculty</th>
<th>food</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>math</td>
<td>fish</td>
<td>15</td>
</tr>
<tr>
<td>Chris</td>
<td>math</td>
<td>soup</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- SELECT COUNT(*) FROM table;
- SELECT COUNT(*) FROM table WHERE food = fish;
- SELECT SUM(time) FROM table; $\text{MAX}(\text{time})$
  - $\infty$ if we do not have bounds on time attribute.
- SELECT AVG(time) FROM table; $\frac{\text{MAX}(\text{time})}{n}$
  - Let $m$ be the number of records in $t$.
  - Let $x$ be the time of a record that has been removed from $t$.
  - Let $y$ be the sum of all times.
  - $\frac{y}{m} - \frac{y-x}{m-1} = \left| \frac{x-y/m}{m-1} \right| \leq \frac{\text{MAX}(\text{time})}{m-1} \leq \frac{\text{MAX}(\text{time})}{n}$.
  - Where $n$ is the least possible number of records.
Laplace distribution

- \( \text{Lap} (\mu, b) \propto \frac{1}{2b} \cdot e^{- \frac{|\mu - z|}{b}} \).
- Increasing \( b \) flattens the curve.
- We will only be interested in \( \mu = 0 \) case.
  - \( \text{Lap} (b) \propto \frac{1}{2b} \cdot e^{- \frac{|z|}{b}} \).
Laplace mechanism

Theorem

Let \( q : D \rightarrow \mathbb{R} \) be a function with global sensitivity \( \Delta q \). Then

\[
\mathcal{M}_q(x) := q(x) + \text{Lap} \left( \frac{\Delta q}{\epsilon} \right)
\]

is \( \epsilon \)-differentially private.

- The noisy output \( \mathcal{M}_q(t) \) is distributed according to \( \text{Lap}(q(t), \frac{\Delta q}{\epsilon}) \).
- Noise magnitude depends on \( q \) and \( \epsilon \), not on the database.
  - Can be computed once for particular \( q \) and \( \epsilon \).
  - Can be less efficient for certain databases.
- Smaller \( \Delta q \) (i.e. more similar outputs) and larger \( \epsilon \) (i.e. less privacy) means steeper curve and hence less distortion.
Proof that Laplace mechanism gives $\epsilon$-DP

We will show that $\frac{pdf_{\mathcal{M}_q(t)}(y)}{pdf_{\mathcal{M}_q(t')} (y)} \leq e^\epsilon$ for all $y \in Y$ for adjacent databases $t, t'$. 

In the derivation, we use triangle inequality $|a + b| \leq |a| + |b|$ with $a = |q(t') - z|$ and $b = |q(t) - q(t')|$. 

Proof that Laplace mechanism gives \( \epsilon \)-DP

We will show that

\[
\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')}}(y) \leq e^\epsilon
\]

for all \( y \in Y \) for adjacent databases \( t, t' \).

\[
\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')})(y)} = \frac{pdf_{Lap(q(t), \frac{\Delta q}{\epsilon})}(y)}{pdf_{Lap(q(t'), \frac{\Delta q}{\epsilon})}(y)}
\]

In the derivation, we use triangle inequality

\[
|a + b| \leq |a| + |b|
\]

with

\[
a = |q(t') - z|
\]

and

\[
b = |q(t) - q(t')|
\]
Proof that Laplace mechanism gives $\epsilon$-DP

We will show that $\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')}(y)} \leq e^\epsilon$ for all $y \in Y$ for adjacent databases $t$, $t'$.

\[
\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')}(y)} = \frac{pdf_{Lap(q(t), \frac{\Delta q}{\epsilon})}(y)}{pdf_{Lap(q(t'), \frac{\Delta q}{\epsilon})}(y)}
\]

\[
= \frac{e^{-|q(t)-z|\epsilon/\Delta q}}{e^{-|q(t')-z|\epsilon/\Delta q}}
\]

In the derivation, we use triangle inequality $|a+b| \leq |a| + |b|$ with $a = |q(t')-z|$ and $b = |q(t)-q(t')|$. 
Proof that Laplace mechanism gives $\epsilon$-DP

We will show that

$$\frac{pdf_{\mathcal{M}_q(t)}(y)}{pdf_{\mathcal{M}_q(t')} (y)} \leq e^\epsilon$$

for all $y \in Y$ for adjacent databases $t$, $t'$.

$$\frac{pdf_{\mathcal{M}_q(t)}(y)}{pdf_{\mathcal{M}_q(t')} (y)} = \frac{pdf_{\text{Lap}(q(t), \frac{\Delta q}{\epsilon})}(y)}{pdf_{\text{Lap}(q(t'), \frac{\Delta q}{\epsilon})}(y)}$$

$$= \frac{e^{-|q(t) - z|\epsilon}}{e^{-|q(t') - z|\epsilon}} \cdot \frac{e^{\frac{\epsilon}{\Delta q} \cdot (|q(t) - z| - |q(t') - z|)}}$$
Proof that Laplace mechanism gives $\epsilon$-DP

We will show that

$$\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')}^{(y)}} \leq e^\epsilon$$

for all $y \in Y$ for adjacent databases $t$, $t'$. 

\[
\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')}^{(y)}} = \frac{pdf_{Lap(q(t), \frac{\Delta q}{\epsilon})}(y)}{pdf_{Lap(q(t'), \frac{\Delta q}{\epsilon})}(y)} \\
= e^{-\frac{|q(t)-z|}{\Delta q}} e^{\frac{\epsilon}{\Delta q}} \cdot (|q(t)-z|-|q(t') - z|) \\
\leq e^{\frac{\epsilon}{\Delta q} \cdot |q(t) - q(t')|}
\]
Proof that Laplace mechanism gives $\epsilon$-DP

We will show that $\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')} (y)} \leq e^\epsilon$ for all $y \in Y$ for adjacent databases $t, t'$.

\[
\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')} (y)} = \frac{pdf_{Lap(q(t), \frac{\Delta q}{\epsilon})}(y)}{pdf_{Lap(q(t'), \frac{\Delta q}{\epsilon})}(y)}
\]

\[
e^{\frac{-|q(t) - z| \epsilon}{\Delta q}}
\]

\[
e^{\frac{-|q(t') - z| \epsilon}{\Delta q}}
\]

\[
e \frac{e^{\epsilon \cdot |q(t) - q(t')|}}{e^{\frac{\epsilon}{\Delta q} \cdot |q(t) - q(t')|}}
\]

\[
e^\epsilon
\]
Proof that Laplace mechanism gives $\epsilon$-DP

We will show that \[
\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')} (y)} \leq e^\epsilon \text{ for all } y \in Y \text{ for adjacent databases } t, t'.
\]

\[
\frac{pdf_{M_q(t)}(y)}{pdf_{M_q(t')} (y)} = \frac{pdf_{Lap (q(t), \frac{\Delta q}{\epsilon})} (y)}{pdf_{Lap (q(t'), \frac{\Delta q}{\epsilon})} (y)}
\]

\[
= e^{-\frac{|q(t) - z|}{\epsilon \Delta q}} \geq e^{-\frac{|q(t') - z|}{\epsilon \Delta q}} \leq e^\epsilon
\]

In the derivation, we use triangle inequality $|a + b| \leq |a| + |b|$ with $a = |q(t') - z|$ and $b = |q(t) - q(t')|$. 

Accuracy of DP mechanisms

We can give statements about the accuracy of our algorithms by using formulae like

$$\Pr[\text{noise} \geq \alpha] \leq \beta,$$

where

- $\text{noise} = |M_q(t) - q(t)|$ is the added noise;
- $\alpha$ is the accuracy that we want to achieve;
- $\beta$ is the probability of failure.
Accuracy of Laplace mechanism

- We need $\Pr[\text{noise} \geq \alpha] \leq \beta$.

- For this, we need the area bounded by $[-\alpha, \alpha]$ be $(1-\beta)$.

- For Laplace distribution $\text{Lap}(\frac{\Delta q}{\epsilon})$, we have

$$ \int_{-\alpha}^{\alpha} \frac{\Delta q}{2\epsilon} e^{-\frac{|z|\epsilon}{\Delta q}} \, dz = (1-\beta) \iff \alpha = \frac{\Delta q}{\epsilon} \ln \left( \frac{1}{\beta} \right) $$
Composition of DP mechanisms

- So far, we have discussed differential privacy of one query.
- Will the attacker succeed better if he makes several queries to private data?
- What if these queries have different privacy mechanisms applied to them?
Sequential composition

Theorem

Let $M_1$ and $M_2$ provide $\epsilon_1$ and $\epsilon_2$-differential privacy. The sequence of $M_1(X)$ and $M_2(X)$ provides $\epsilon_1 + \epsilon_2$-differential privacy.

Intuition: The more queries attacker makes about Alice, the more he learns about Alice. We need to strengthen the privacy of each single query to keep the same overall privacy level.
Proof of Sequential composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y'_1, Y'_2 \) be the outputs of \( \mathcal{M}_1, \mathcal{M}_2 \) respectively. We want to estimate privacy of the new mechanism \( \mathcal{M}_{12}(t) = (\mathcal{M}_1(t), \mathcal{M}_2(t)) \). Let \( Y' := (Y'_1, Y'_2) \).
Proof of Sequential composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y'_1, Y'_2 \) be the outputs of \( M_1, M_2 \) respectively. We want to estimate privacy of the new mechanism \( M_{12}(t) = (M_1(t), M_2(t)) \). Let \( Y' := (Y'_1, Y'_2) \).

\[
\frac{\Pr[M_{12}(t) \in Y']}{\Pr[M_{12}(t') \in Y']} = \frac{\Pr[(M_1(t), M_2(t)) \in (Y'_1, Y'_2)]}{\Pr[(M_1(t'), M_2(t')) \in (Y'_1, Y'_2)]}
\]
Proof of Sequential composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y'_1, Y'_2 \) be the outputs of \( \mathcal{M}_1, \mathcal{M}_2 \) respectively. We want to estimate privacy of the new mechanism \( \mathcal{M}_{12}(t) = (\mathcal{M}_1(t), \mathcal{M}_2(t)) \). Let \( Y' := (Y'_1, Y'_2) \).

\[
\frac{\Pr[\mathcal{M}_{12}(t) \in Y']}{\Pr[\mathcal{M}_{12}(t') \in Y']}
= \frac{\Pr[(\mathcal{M}_1(t), \mathcal{M}_2(t)) \in (Y'_1, Y'_2)]}{\Pr[(\mathcal{M}_1(t'), \mathcal{M}_2(t')) \in (Y'_1, Y'_2)]}
= \frac{\Pr[\mathcal{M}_1(t) \in Y'_1] \cdot \Pr[\mathcal{M}_2(t) \in Y'_2]}{\Pr[\mathcal{M}_1(t') \in Y'_1] \cdot \Pr[\mathcal{M}_2(t') \in Y'_2]} \leq e^{\epsilon_1} \cdot e^{\epsilon_2} = e^{\epsilon_1 + \epsilon_2}
\]
Proof of Sequential composition

Let $t$ and $t'$ be adjacent datasets. Let $Y'_1$, $Y'_2$ be the outputs of $\mathcal{M}_1$, $\mathcal{M}_2$ respectively. We want to estimate privacy of the new mechanism $\mathcal{M}_{12}(t) = (\mathcal{M}_1(t), \mathcal{M}_2(t))$. Let $Y' := (Y'_1, Y'_2)$.

\[
\frac{\Pr[\mathcal{M}_{12}(t) \in Y']}{\Pr[\mathcal{M}_{12}(t') \in Y']} = \frac{\Pr[(\mathcal{M}_1(t), \mathcal{M}_2(t)) \in (Y'_1, Y'_2)]}{\Pr[(\mathcal{M}_1(t'), \mathcal{M}_2(t')) \in (Y'_1, Y'_2)]}
\]

\[
= \frac{\Pr[\mathcal{M}_1(t) \in Y'_1] \cdot \Pr[\mathcal{M}_2(t) \in Y'_2]}{\Pr[\mathcal{M}_1(t') \in Y'_1] \cdot \Pr[\mathcal{M}_2(t') \in Y'_2]}
\]

\[
= \frac{\Pr[\mathcal{M}_1(t) \in Y'_1]}{\Pr[\mathcal{M}_1(t') \in Y'_1]} \cdot \frac{\Pr[\mathcal{M}_2(t) \in Y'_2]}{\Pr[\mathcal{M}_2(t') \in Y'_2]}
\]

\[
\leq e^{\epsilon_1} \cdot e^{\epsilon_2} = e^{\epsilon_1 + \epsilon_2}
\]
Proof of Sequential composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y'_1, Y'_2 \) be the outputs of \( \mathcal{M}_1, \mathcal{M}_2 \) respectively. We want to estimate privacy of the new mechanism \( \mathcal{M}_{12}(t) = (\mathcal{M}_1(t), \mathcal{M}_2(t)) \). Let \( Y' := (Y'_1, Y'_2) \).

\[
\frac{\Pr[\mathcal{M}_{12}(t) \in Y']}{\Pr[\mathcal{M}_{12}(t') \in Y']} = \frac{\Pr[(\mathcal{M}_1(t), \mathcal{M}_2(t)) \in (Y'_1, Y'_2)]}{\Pr[(\mathcal{M}_1(t'), \mathcal{M}_2(t')) \in (Y'_1, Y'_2)]}
\]

\[
= \frac{\Pr[\mathcal{M}_1(t) \in Y'_1] \cdot \Pr[\mathcal{M}_2(t) \in Y'_2]}{\Pr[\mathcal{M}_1(t') \in Y'_1] \cdot \Pr[\mathcal{M}_2(t') \in Y'_2]}
\]

\[
= \frac{\Pr[\mathcal{M}_1(t) \in Y'_1]}{\Pr[\mathcal{M}_1(t') \in Y'_1]} \cdot \frac{\Pr[\mathcal{M}_2(t) \in Y'_2]}{\Pr[\mathcal{M}_2(t') \in Y'_2]}
\]

\[
\leq e^{\epsilon_1} \cdot e^{\epsilon_2}
\]
Proof of Sequential composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y'_1, Y'_2 \) be the outputs of \( \mathcal{M}_1, \mathcal{M}_2 \) respectively. We want to estimate privacy of the new mechanism \( \mathcal{M}_{12}(t) = (\mathcal{M}_1(t), \mathcal{M}_2(t)) \). Let \( Y' := (Y'_1, Y'_2) \).

\[
\frac{\Pr[\mathcal{M}_{12}(t) \in Y']}{\Pr[\mathcal{M}_{12}(t') \in Y']} = \frac{\Pr[(\mathcal{M}_1(t), \mathcal{M}_2(t)) \in (Y'_1, Y'_2)]}{\Pr[(\mathcal{M}_1(t'), \mathcal{M}_2(t')) \in (Y'_1, Y'_2)]}
\]

\[
= \frac{\Pr[\mathcal{M}_1(t) \in Y'_1] \cdot \Pr[\mathcal{M}_2(t) \in Y'_2]}{\Pr[\mathcal{M}_1(t') \in Y'_1] \cdot \Pr[\mathcal{M}_2(t') \in Y'_2]}
\]

\[
\leq e^{\epsilon_1} \cdot e^{\epsilon_2}
\]

\[
= e^{\epsilon_1 + \epsilon_2}
\]
Parallel composition

Theorem
Let $\mathcal{M}_1$ and $\mathcal{M}_2$ provide $\epsilon_1$ and $\epsilon_2$-differential privacy. Let $X_i$ be arbitrary disjoint subsets of the input domain $X$. The sequence of $\mathcal{M}_1(X_1)$ and $\mathcal{M}_2(X_2)$ provides $\max(\epsilon_1, \epsilon_2)$-differential privacy.

Intuition: Making queries about Bob does not help the attacker to learn more about Alice.
Proof of Parallel composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y_1', Y_2' \) be the outputs of \( M_1, M_2 \) respectively. We want to estimate privacy of the new mechanism \( M_{12}(t) = (M_1(t), M_2(t)) \). Let \( Y' := (Y_1', Y_2') \).
Proof of Parallel composition

Let $t$ and $t'$ be adjacent datasets. Let $Y'_1$, $Y'_2$ be the outputs of $M_1$, $M_2$ respectively. We want to estimate privacy of the new mechanism $M_{12}(t) = (M_1(t), M_2(t))$. Let $Y' := (Y'_1, Y'_2)$.

- We have $M_i(t) = M_i(t_i)$ and $M_i(t') = M_i(t'_i)$. 

Then, we can estimate the privacy as follows:

$$\Pr[M_{12}(t) \in Y'] = \Pr[M_1(t) \in Y'_1] \cdot \Pr[M_2(t) \in Y'_2] = \Pr[M_1(t_i) \in Y'_1] \cdot \Pr[M_2(t_i) \in Y'_2].$$

Since $t$ and $t'$ are adjacent, we have either $t_1 = t'_1$ or $t_2 = t'_2$, as otherwise $t$ and $t'$ would differ in at least two records.
Proof of Parallel composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y_1', Y_2' \) be the outputs of \( \mathcal{M}_1, \mathcal{M}_2 \) respectively. We want to estimate privacy of the new mechanism \( \mathcal{M}_{12}(t) = (\mathcal{M}_1(t), \mathcal{M}_2(t)) \). Let \( Y' := (Y_1', Y_2') \).

- We have \( \mathcal{M}_i(t) = \mathcal{M}_i(t_i) \) and \( \mathcal{M}_i(t') = \mathcal{M}_i(t'_i) \).
- We have either \( t_1 = t'_1 \) or \( t_2 = t'_2 \), as otherwise \( t \) and \( t' \) would differ in at least two records.
Proof of Parallel composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y'_1, Y'_2 \) be the outputs of \( \mathcal{M}_1, \mathcal{M}_2 \) respectively. We want to estimate privacy of the new mechanism \( \mathcal{M}_{12}(t) = (\mathcal{M}_1(t), \mathcal{M}_2(t)) \). Let \( Y' := (Y'_1, Y'_2) \).

- We have \( \mathcal{M}_i(t) = \mathcal{M}_i(t_i) \) and \( \mathcal{M}_i(t') = \mathcal{M}_i(t'_i) \).
- We have either \( t_1 = t'_1 \) or \( t_2 = t'_2 \), as otherwise \( t \) and \( t' \) would differ in at least two records.

\[
\frac{\Pr[\mathcal{M}_{12}(t) \in Y']} {\Pr[\mathcal{M}_{12}(t') \in Y']} = \frac{\Pr[(\mathcal{M}_1(t_1), \mathcal{M}_2(t_2)) \in (Y'_1, Y'_2)]} {\Pr[(\mathcal{M}_1(t'_1), \mathcal{M}_2(t'_2)) \in (Y'_1, Y'_2)]}
\]

\[\Pr[\mathcal{M}_{12}(t) \in Y'] = \Pr[\mathcal{M}_1(t) \in Y'_1] \cdot \Pr[\mathcal{M}_2(t) \in Y'_2] \]

\[\Pr[\mathcal{M}_{12}(t') \in Y'] = \Pr[\mathcal{M}_1(t') \in Y'_1] \cdot \Pr[\mathcal{M}_2(t') \in Y'_2] \]

\[\Pr[(\mathcal{M}_1(t_1), \mathcal{M}_2(t_2)) \in (Y'_1, Y'_2)] = \Pr[\mathcal{M}_1(t_1) \in Y'_1] \cdot \Pr[\mathcal{M}_2(t_2) \in Y'_2] \]

\[\Pr[(\mathcal{M}_1(t'_1), \mathcal{M}_2(t'_2)) \in (Y'_1, Y'_2)] = \Pr[\mathcal{M}_1(t'_1) \in Y'_1] \cdot \Pr[\mathcal{M}_2(t'_2) \in Y'_2] \]
Proof of Parallel composition

Let \( t \) and \( t' \) be adjacent datasets. Let \( Y'_1, Y'_2 \) be the outputs of \( M_1, M_2 \) respectively. We want to estimate privacy of the new mechanism \( M_{12}(t) = (M_1(t), M_2(t)) \). Let \( Y' := (Y'_1, Y'_2) \).

- We have \( M_i(t) = M_i(t_i) \) and \( M_i(t') = M_i(t'_i) \).
- We have either \( t_1 = t'_1 \) or \( t_2 = t'_2 \), as otherwise \( t \) and \( t' \) would differ in at least two records.

\[
\frac{\Pr[M_{12}(t) \in Y']}{\Pr[M_{12}(t') \in Y']} = \frac{\Pr[(M_1(t_1), M_2(t_2)) \in (Y'_1, Y'_2)]}{\Pr[(M_1(t'_1), M_2(t'_2)) \in (Y'_1, Y'_2)]} \cdot \frac{\Pr[M_1(t_1) \in Y'_1]}{\Pr[M_1(t'_1) \in Y'_1]} \cdot \frac{\Pr[M_2(t_2) \in Y'_2]}{\Pr[M_2(t'_2) \in Y'_2]}
\]
Proof of Parallel composition

Let $t$ and $t'$ be adjacent datasets. Let $Y_1', Y_2'$ be the outputs of $M_1, M_2$ respectively. We want to estimate privacy of the new mechanism $M_{12}(t) = (M_1(t), M_2(t))$. Let $Y' := (Y_1', Y_2')$.

- We have $M_i(t) = M_i(t_i)$ and $M_i(t') = M_i(t'_i)$.
- We have either $t_1 = t_1'$ or $t_2 = t_2'$, as otherwise $t$ and $t'$ would differ in at least two records.

\[
\frac{\Pr[M_{12}(t) \in Y']}{\Pr[M_{12}(t') \in Y']} = \frac{\Pr[(M_1(t_1), M_2(t_2)) \in (Y_1', Y_2')]}{\Pr[(M_1(t'_1), M_2(t'_2)) \in (Y_1', Y_2')]} = \frac{\Pr[M_1(t_1) \in Y_{1}']}{\Pr[M_1(t'_1) \in Y_{1}']} \cdot \frac{\Pr[M_2(t_2) \in Y_{2}']}{\Pr[M_2(t'_2) \in Y_{2}']} = \begin{cases} \frac{\Pr[M_1(t_1) \in Y_{1}']}{\Pr[M_1(t'_1) \in Y_{1}']} & \text{if } t_2 = t'_2 \\ \frac{\Pr[M_2(t_2) \in Y_{2}']}{\Pr[M_2(t'_2) \in Y_{2}']} & \text{if } t_1 = t'_1 \end{cases}
\]
Proof of Parallel composition

Let $t$ and $t'$ be adjacent datasets. Let $Y'_1$, $Y'_2$ be the outputs of $\mathcal{M}_1$, $\mathcal{M}_2$ respectively. We want to estimate privacy of the new mechanism $\mathcal{M}_{12}(t) = (\mathcal{M}_1(t), \mathcal{M}_2(t))$. Let $Y' := (Y'_1, Y'_2)$.

- We have $\mathcal{M}_i(t) = \mathcal{M}_i(t_i)$ and $\mathcal{M}_i(t') = \mathcal{M}_i(t'_i)$.
- We have either $t_1 = t'_1$ or $t_2 = t'_2$, as otherwise $t$ and $t'$ would differ in at least two records.

$$\frac{\Pr[\mathcal{M}_{12}(t) \in Y']} {\Pr[\mathcal{M}_{12}(t') \in Y']} = \frac{\Pr[\mathcal{M}_1(t_1), \mathcal{M}_2(t_2)] \in (Y'_1, Y'_2)]} {\Pr[\mathcal{M}_1(t'_1), \mathcal{M}_2(t'_2)] \in (Y'_1, Y'_2)]}$$

$$= \frac{\Pr[\mathcal{M}_1(t_1) \in Y'_1]} {\Pr[\mathcal{M}_1(t'_1) \in Y'_1]} \cdot \frac{\Pr[\mathcal{M}_2(t_2) \in Y'_2]} {\Pr[\mathcal{M}_2(t'_2) \in Y'_2]}$$

$$= \begin{cases} 1 & \text{if } t_2 = t'_2 \\ \frac{\Pr[\mathcal{M}_1(t_1) \in Y'_1]} {\Pr[\mathcal{M}_1(t'_1) \in Y'_1]} & \text{if } t_1 = t'_1 \\ \frac{\Pr[\mathcal{M}_2(t_2) \in Y'_2]} {\Pr[\mathcal{M}_2(t'_2) \in Y'_2]} & \text{if } t_2 = t'_2 \end{cases}$$

$$= e^{\max(\epsilon_1, \epsilon_2)}$$
Handling Privacy Budget

- The analyst is given some privacy budget $\epsilon$.
- He can ask any number of $\epsilon_i$-DP queries $\mathcal{M}_{q_i}$.
- The system stops answering after the asked queries $\mathcal{M}_{q_1}, \ldots, \mathcal{M}_{q_n}$ altogether exceed $\epsilon$-DP.
Handling Privacy Budget

- The analyst is given some privacy budget $\epsilon$.
- He can ask any number of $\epsilon_i$-DP queries $M_{q_i}$.
- The system stops answering after the asked queries $M_{q_1}, \ldots, M_{q_n}$ altogether exceed $\epsilon$-DP.

In this example, the total spent budget is $\epsilon = \max(\epsilon_1 + \epsilon_2, \epsilon_3 + \epsilon_4 + \epsilon_5)$. 
Group privacy

- So far, we have defined differential privacy for adjacent databases (that differ in one record).

- What if databases are are $k$-adjacent (differ in $k$ records)?
- It would allow to define privacy of a group of $k$ records.
Group privacy

- Let us generalize the definition of DP.

**Differential privacy:** A function \( q : X \rightarrow Y \) is called \( \epsilon \)-DP if, for all \( Y' \subseteq Y \), for all \( k \)-adjacent tables \( t, t' \in X \):

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^{k\epsilon}
\]

- Is this definition stronger?
Let us generalize the definition of DP.

**Differential privacy:** A function $q : X \rightarrow Y$ is called $\epsilon$-DP if, for all $Y' \subseteq Y$, for all $k$-adjacent tables $t, t' \in X$:

$$\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^{k\epsilon}$$

- Is this definition stronger?
- It turns out that stating DP for $k = 1$ is sufficient.
Group privacy

- Let $t$ and $t'$ be $k$-adjacent.
- Removing rows from $t$ one by one, we get a sequence

$$t = t_0, t_1, \ldots, t_k = t'$$

such that $t_{i-1}$ and $t_i$ are 1-adjacent for all $i \in \{1, \ldots, k\}$. 

<table>
<thead>
<tr>
<th>Alice</th>
<th>Bob</th>
<th>Chris</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>$t = t_0$</td>
<td>$t_1$</td>
<td>$t_2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bob</th>
<th>Chris</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chris</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

$t_3 = t'$
Group privacy

Let $q$ be $\epsilon$-differentially private.

$$\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon$$
for $1$-adjacent tables $t, t'$.

For $k$-adjacent $t$ and $t'$, we have:

$$\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} =$$
Group privacy

- Let $q$ be $\epsilon$-differentially private.

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon \text{ for } 1\text{-adjacent tables } t, t'.
\]

- For $k$-adjacent $t$ and $t'$, we have:

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} = \frac{\Pr[q(t) \in Y']}{\Pr[q(t_1) \in Y']} \cdot \frac{\Pr[q(t_1) \in Y']}{\Pr[q(t_2) \in Y']} \cdots \frac{\Pr[q(t_{k-1}) \in Y']}{\Pr[q(t') \in Y']}
\]
Group privacy

- Let $q$ be $\epsilon$-differentially private.

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon \text{ for } 1\text{-adjacent tables } t, t'.
\]

- For $k$-adjacent $t$ and $t'$, we have:

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} = \frac{\Pr[q(t) \in Y']}{\Pr[q(t_1) \in Y']} \cdot \frac{\Pr[q(t_1) \in Y']}{\Pr[q(t_2) \in Y']} \cdots \frac{\Pr[q(t_{k-1}) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon \cdots e^\epsilon.
\]
Group privacy

- Let $q$ be $\epsilon$-differentially private.

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} \leq e^\epsilon \text{ for } 1\text{-adjacent tables } t, t'.
\]

- For $k$-adjacent $t$ and $t'$, we have:

\[
\frac{\Pr[q(t) \in Y']}{\Pr[q(t') \in Y']} = \frac{\Pr[q(t) \in Y']}{\Pr[q(t_1) \in Y']} \cdot \frac{\Pr[q(t_1) \in Y']}{\Pr[q(t_2) \in Y']} \cdots \frac{\Pr[q(t_{k-1}) \in Y']}{\Pr[q(t') \in Y']}
\leq e^\epsilon \cdots e^\epsilon
= e^{k\epsilon}
\]

- Hence, to obtain $\epsilon$-DP for a group of size $k$, we can use any $\frac{\epsilon}{k}$-DP mechanism.
Beware of the output:
- Making computation privacy-preserving is in general not sufficient.
- The final outputs may leak privacy.

No free lunch in data privacy:
- We cannot achieve absolute privacy while keeping the result reasonable.
- There will always be privacy-utility tradeoff.

Differential privacy:
- Tunes privacy-utility balance.
- Gives us provable guarantees.
- Is composable and can be used as a building block in complex systems.