Setting

Medical data
Query logs
Social network data
...

Data mining
Statistical queries
Setting – Data Release

Main concern: Do not violate user privacy!

Publish:
Aggregated data, e.g., outcome of medical study, research paper, ...
Example – Linkage Attack

- The Commonwealth of Massachusetts Group Insurance Commission (GIC) releases 135,000 records of patient encounters, each with 100 attributes
  - Relevant attributes removed, but ZIP, birth date, gender available
  - Considered “safe” practice
- Public voter registration record
  - Contain, among others, name, address, ZIP, birth date, gender
- Allowed identification of medical records of William Weld, governor of MA at that time
  - He was the only man in his zip code with his birth date …

More attacks! (cf. Netflix grand prize challenge!)
One way out? Differential Privacy

• A formal definition of privacy
  – Satisfied in systems deployed by Google, Uber, Apple, ...

• Used by 2020 census

• Idea: Any information-related risk to a person should not change significantly as a result of that person’s information being included, or not, in the analysis.
  – Even with side information!
Ideal Individual’s Privacy

For every individual A whose record in DB

DB w/ A’s data

\[ \hat{x} \]

Analysis

Output

DB w/o A’s data

\[ \hat{x}' \]

Analysis

Output’

Very good for privacy. But the output would be useless as it does not depend on any individual’s record!

Ideally: Should be identical!

Common Theme:
- Tension / Balance between privacy & utility
- Privacy is not a 0 / 1 property.
More Realistic Privacy Goal

$\vec{x}$

DB w/ A’s data

$\vec{x}'$

DB w/o A’s data

Analysis

Output

Should be “similar”

Output’
We say that $\vec{x}, \vec{x}'$ differ at exactly one entry.

Here, $M$ is randomized, i.e., it makes random choices.
Setting – Mechanism

**Definition.** A mechanism $\mathbf{M}$ is $\epsilon$-differentially private if for all subsets $T \subseteq \mathbb{R}$, and for all databases $\mathbf{x}, \mathbf{x}'$ which differ at exactly one entry,

$$P(M(\mathbf{x}) \in T) \leq e^\epsilon P(M(\mathbf{x}') \in T)$$

Dwork, McSherry, Nissim, Smith, ‘06

Think: $\epsilon = \frac{1}{100}$ or $\epsilon = \frac{1}{10}$

$e^\epsilon \approx 1 + \epsilon$ for small $\epsilon$
Example – Counting Queries

• DB is a vector $\bar{x} = (x_1, ..., x_n)$ where $x_1, ..., x_n \in \{0,1\}$
  - $x_i = 1$ if individual $i$ has disease
  - $x_i = 0$ means patient does not have disease or patient data wasn’t recorded.

• Query: $q(\bar{x}) = \sum_{i=1}^{n} x_i$

Here: $\bar{x}$ and $\bar{x}'$ differ at one entry means they differ at one single coordinate, e.g., $x_i = 1$ and $x'_i = 0$
A solution – Laplacian Noise

Mechanism $M$ taking input $\mathbf{x} = (x_1, \ldots, x_n)$:

- Return $M(\mathbf{x}) = \sum_{i=1}^{n} x_i + Y$

Here, $Y$ follows a **Laplace distribution** with parameter $\epsilon$

\[ f_Y(y) = \frac{\epsilon}{2} e^{-\epsilon |y|} \]

\[ \mathbb{E}[Y] = 0 \]

\[ \text{Var}(Y) = \frac{2}{\epsilon^2} \]
Better Solution – Laplacian Noise

Mechanism $M$ taking input $\vec{x} = (x_1, \ldots, x_n)$:
- Return $M(\vec{x}) = \sum_{i=1}^{n} x_i + Y$

Here, $Y$ follows a Laplace distribution with parameter $\epsilon$

$$f_Y(y) = \frac{\epsilon}{2} e^{-\epsilon |y|}$$

Key property: For all $y, \Delta$

$$\frac{f_Y(y)}{f_Y(y + \Delta)} \leq e^{\epsilon \Delta}$$
Theorem. The Laplacian Mechanism with parameter $\epsilon$ satisfies $\epsilon$-differential privacy

Show: \( \forall \vec{x}, \vec{x}' \) differ at one entry, \([a, b] \)

\[
P(M(\vec{x}) \in [a, b]) \leq e^{\epsilon} \cdot P(M(\vec{x}') \in [a, b])
\]

\[
\Delta = \sum_{i=1}^{n} x'_i - \sum_{i=1}^{n} x_i \quad |\Delta| \leq 1
\]

\[
P(M(\vec{x}) \in [a, b]) = P(\vec{x}' + \vec{Y} \in [a, b]) = \int_{a-s}^{b-s} f_Y(y)dy = \int_{a}^{b} f_Y(y' - s)dy'
\]

\[
= \int_{a}^{b} f_Y(y - s') + \Delta dy \leq e^{\epsilon} \int_{a}^{b} f_Y(y - s')dy \leq e^{\epsilon} \int_{a}^{b} f_Y(y - s')dy
\]

\[
= e^{\epsilon} P(M(\vec{x}') \in [a, b])
\]
How Accurate is Laplacian Mechanism?

Let’s look at $\sum_{i=1}^{n} x_i + Y$

- $\mathbb{E}[\sum_{i=1}^{n} x_i + Y] = \sum_{i=1}^{n} x_i + \mathbb{E}[Y] = \sum_{i=1}^{n} x_i$

- $\operatorname{Var}(\sum_{i=1}^{n} x_i + Y) = \operatorname{Var}(Y) = \frac{2}{\epsilon^2}$

This is accurate enough for large enough $\epsilon$!
Differential Privacy – What else can we compute?

• **Statistics:** counts, mean, median, histograms, boxplots, etc.
• **Machine learning:** classification, regression, clustering, distribution learning, etc.
• ...
Differential Privacy – Nice Properties

• **Group privacy:** If $M$ is $\epsilon$-differentially private, then for all $T \subseteq \mathbb{R}$, and for all databases $\mathbf{x}, \mathbf{x}'$ which differ at (at most) $k$ entries,

$$P(M(\mathbf{x}) \in T) \leq e^{k\epsilon} P(M(\mathbf{x}') \in T)$$

• **Composition:** If we apply two $\epsilon$-DP mechanisms to data, combined output is $2\epsilon$-DP.
  – How much can we allow $\epsilon$ to grow? (So-called “privacy budget.”)

• **Post-processing:** Postprocessing does not decrease privacy.
Local Differential Privacy

Laplacian Mechanism

What if we don’t trust aggregator?

Solution: Add noise locally!
Example – Randomize Response

Mechanism $M$ taking input $\vec{x} = (x_1, \ldots, x_n)$:

- For all $i = 1, \ldots, n$:
  
  - $y_i = x_i$ w/ probability $\frac{1}{2} + \alpha$, and $y_i = 1 - x_i$ w/ probability $\frac{1}{2} - \alpha$.
  
  - $\hat{x}_i = \frac{y_i - \frac{1}{2} + \alpha}{2\alpha}$

- Return $M(\vec{x}) = \sum_{i=1}^{n} \hat{x}_i$

For a given parameter $\alpha$

**Example – Randomize Response**

**Mechanism** $M$ taking input $\mathbf{x} = (x_1, \ldots, x_n)$:

- For all $i = 1, \ldots, n$:
  - $y_i = x_i$ w/ probability $\frac{1}{2} + \alpha$, and $y_i = 1 - x_i$ w/ probability $\frac{1}{2} - \alpha$.
  - $\hat{x}_i = \frac{y_i - \frac{1}{2} + \alpha}{2\alpha}$

- Return $M(\mathbf{x}) = \sum_{i=1}^{n} \hat{x}_i$

**Theorem.** Randomized Response with parameter $\alpha$ satisfies $\epsilon$-differential privacy, if $\alpha = \frac{e^\epsilon - 1}{e^\epsilon + 1}$.

**Fact 1.** $\mathbb{E}[M(\mathbf{x})] = \sum_{i=1}^{n} x_i$

**Fact 2.** $\text{Var}(M(\mathbf{x})) \approx \frac{n}{\epsilon^2}$
Differential Privacy – Challenges

• **Accuracy vs. privacy:** How do we choose $\epsilon$?
  – Practical applications tend to err in favor of accuracy.
  – See e.g. [https://arxiv.org/abs/1709.02753](https://arxiv.org/abs/1709.02753)

• **Fairness:** Differential privacy hides contribution of small groups, by design
  – How do we avoid excluding minorities?
  – Very hard problem!

• **Ethics:** Does differential privacy incentivize data collection?
• Cynthia Dwork and Aaron Roth. “The Algorithmic Foundations of Differential Privacy”.
  – https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf
• https://privacytools.seas.harvard.edu/