Recap

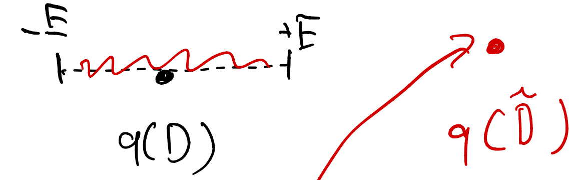
- $y = q(D) \rightsquigarrow$ what y reveals about D?
- k-anonymity, aggregation, ...

If you reveal too many accurate statistics, attacker can reconstruct D

- Only few $\tilde{D}$ are consistent with the statistics of D

Example: Dinur-Nissim

Real D



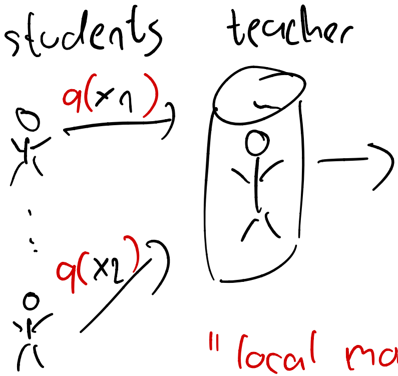
if you observe this you know 100% that $\tilde{D} \neq D$

How to avoid?

Random statistic $q(D)$ that is consistent with every $\tilde{D}$ and is useful!

Randomized Response

Example:



How many cheated?

"local model"

(temporary)

Plausible deniability

adversary can't reliably distinguish if a student cheated based on their response $q(x_i)$

Definition: RR (1965)

Database $D = \{x_1, \dots, x_n\}$ $\quad x_i \in \{0, 1\}$ cheated

Goal: $p = \frac{1}{n} \sum_{i=1}^{n} x_i$ fraction cheated

Algorithm: flip a coin

$$y_i = \begin{cases} x_i & \text{w.p. } \frac{1}{2} + \gamma \\ 1 - x_i & \text{w.p. } \frac{1}{2} - \gamma \end{cases} \qquad \gamma \in [0, \tfrac{1}{2}]$$

- $\gamma = \frac{1}{2} \Rightarrow$ no privacy
- $\gamma = 0 \Rightarrow$ perfect privacy, useless
- $0 < \gamma < \frac{1}{2} \Rightarrow$ if $y_i = 1$, maybe cheated, or just flipped response

$\Rightarrow$ plausible deniability

Goal: $p = \frac{1}{n}\sum_{i=1}^{n} x_i$ fraction cheated

Algorithm: flip a coin

$$y_i = \begin{cases} x_i & \text{w.p. } \frac{1}{2}+\gamma \\ 1-x_i & \text{w.p. } \frac{1}{2}-\gamma \end{cases} \qquad \gamma \in [0, \tfrac{1}{2}]$$

Naive: $\frac{1}{n}\sum_{i=1}^{n} y_i$

Biased: $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right]$

$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[y_i]$

$= \frac{1}{n}\sum_{i=1}^{n} (\frac{1}{2}+\gamma)x_i + (\frac{1}{2}-\gamma)(1-x_i)$

$= \frac{1}{n}\sum_{i=1}^{n} 2\gamma x_i + (\frac{1}{2}-\gamma)$

$= 2\gamma p + (\frac{1}{2}-\gamma) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right]$

$\hat{p} = \frac{1}{2\gamma}\left(\frac{1}{n}\sum_{i=1}^{n} y_i - (\frac{1}{2}-\gamma)\right)$ RR estimator

unbiased: $\mathbb{E}[\hat{p}] = p$

How accurate is this?

Concentration bounds

"How far is $\hat{p}$ from $\mathbb{E}[\hat{p}]$ (*)"

(*) with high probability

(Chernoff, Chebyshev, Hoeffding)

$|p-\hat{p}| \le \tilde{O}\left(\frac{1}{\gamma\sqrt{n}}\right)$ w.h.p.

ignore logs

RR error implications

$|p - \tilde{p}| \leq \tilde{O}\left(\frac{1}{\gamma\sqrt{n}}\right)$ whp

- Goes to 0 if n grows
- Want $|p - \tilde{p}| \leq \alpha$

$\rightsquigarrow n = \tilde{O}\left(\frac{1}{\gamma^2 \alpha^2}\right)$ $\quad \gamma \in [0, \frac{1}{2}]$

$\Rightarrow$ privacy-utility tradeoff!

RR provides plausible deniability

Plausible deniability $\approx$ privacy

Differential privacy

- generalizes this notion
- makes it formal!

Differential privacy

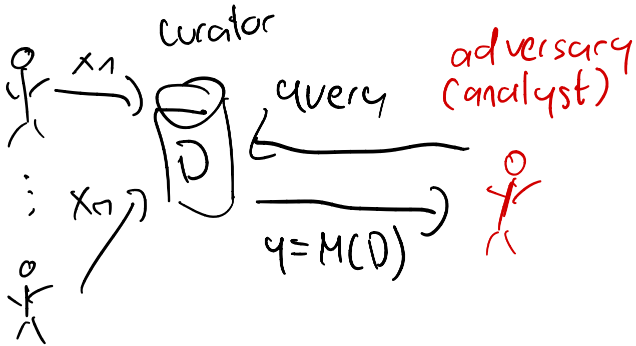
Setting

Database $D \in \mathcal{X}^n$ — arbitrary sets

Mechanism $M : \mathcal{X}^n \to \mathcal{Y}$

(prob. also)

"Central model"



(RR is "local model")

Definition:

Two databases $D, D' \in \mathcal{X}^n$ are <u>neighboring</u> if "replacement"

i) $|D| = |D'|$

ii) D and D' <u>differ in at most 1 row</u>

(Alternative: define in terms of addition/removal. Similar results up to constant factors)

Plausible deniability:

Adversary cannot distinguish $M(D)$ or $M(D')$

$M(D \cup \{you\})$ vs. $M(D \cup \{someone\ else\})$

Definition (ε-DP)

A (randomized) mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is ε-differentially private if

- $\forall D, D' \in \mathcal{X}^n$, $D \sim D'$ and
- $\forall S \subseteq \mathcal{Y}$ we have

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S]$$

Intuition: If you change one row of your database, the output you get is "basically the same".
↳ Probabilistically!

Initial discuss

1. DP is property of an algorithm, not database/output.
2. Worst-case definition
3. No attacker modeling!