Lecture: Heuristic notions of privacy and an introduction to DP

In the last two lectures, we saw examples of reconstruction attacks that allow for (sometimes efficient) reconstruction of sensitive attributes from queries of bounded error. If the goal, however, was simply the "hiding" of the sensitive attribute of a single individual, whose quasi-identifiers are known, then there are simple mechanisms that provide "anonymity" in a heuristic sense.

1. **K-Anonymity** [Sweeney (2002) (see also Sweeney (1997))]:

A dataset $\mathcal{D} = \{(q_i, s_i) : i \in [N]\}$ respects K-anonymity, if for any

(s_i: sens. attribute; q_i: quasi-identifiers)

given $q \in Q$, we have

$$\#\{i : q_i = q\} = 0 \quad \text{or} \quad \#\{i : q_i = q\} \geq K.$$

Clearly, a dataset by itself may not respect K-anonymity, and one typically carries out "generalization" (or binning) or "suppression" (or replacing with '*') of attributes to give rise to "equivalence classes" that respect K-anonymity.

generalized quasi-identifiers.

Example: Setting of hospital records.

Gender (Quasi-identifiers)	Postal Code (Quasi-identifiers)	Disease (Sensitive attribute)
Male	600020	Heart disease
Male	600018	Lung infection
Female	600036	Osteoporosis
Female	600036	Cervical cancer
Female	600036	Osteoporosis

Upon generalization, this yields a dataset $\overline{D}$ that respects 2-anonymity:

(Gender, Postal Code)	Disease
(Male, 600018) or (Male, 600020)	Heart disease
	Lung infection
(Female, 600036)	Osteoporosis
	Osteoporosis
	Cervical cancer

Eq. class

However, K-anonymity has an obvious pitfall: What if all the sensitive attributes in a given eq. class are equal? Then, it is easy for an adversary to identify the sensitive attribute of a user, given knowledge of his/her quasi-identifiers.

To ameliorate this issue:

2. **ℓ-Diversity** [Machanavajjhala, Kifer, Gehrke, Venkitasubramaniam (2007)]:

A dataset $\mathcal{D} = \{(q_i, s_i) : i \in [N]\}$ respects ℓ-diversity if each quasi-identifier has, associated with it, at least ℓ distinct sensitive attributes.

HW: Does the dataset $\mathcal{D}$ above respect ℓ-diversity? If so, what is the largest such ℓ?

However, in very recent work [Rameshwar, Tandon (2025)], we argued that in the "worst-case", "composing" together / "linking" datasets could rapidly degrade anonymity in the sense of ℓ-diversity.

... see also "t-closeness"

[Li, Li, Venkatasubramaniam, "t-Closeness: Privacy beyond k-anonymity and ℓ-diversity," (2007)]

However, we are interested in rigorous notions of privacy, which also "compose" gracefully, and are backed by formal results in probability / information theory.

Differential Privacy

The chief idea in defining a formal notion of privacy is to address the problem of tracing: an adversary must not be able to distinguish b/w whether a dataset contained/did not contain a sample from a given user, purely based on the outputs of a privacy-preserving mechanism.

Def 1: Datasets $\mathcal{D}, \mathcal{D}'$ with the same collection of users $[N]$ are said to be neighbouring if $\mathcal{D} = (x_1, x_2, \dots, x_N)$ and $\mathcal{D}' = (x_1', x_2', \dots, x_N')$, with $x_j \neq x_j'$, for some $j \in [N]$ and $x_i = x_i'$, for all $i \neq j$.

Def 2 (DP): A randomized algorithm (or mechanism) $A : \mathbb{D}_n \to \mathcal{Y}$ is ϵ-differentially private (or ϵ-DP), for $\epsilon > 0$, if, for every pair $\underline{x}, \underline{x}'$ of neighbouring datasets, and every $S \subseteq \mathcal{Y}$, we have

$$\mathbb{P}_R[A(\underline{x}) \in S] \leq e^{\epsilon} \mathbb{P}_R[A(\underline{x}') \in S].$$

Remarks: (i) Note that, by the definition above, we must have, $\forall \underline{x}, \underline{x}'$ neighbouring & all $S \subseteq \mathcal{Y}$,

$$e^{-\epsilon} \mathbb{P}_R[A(\underline{x}') \in S] \leq \mathbb{P}_R[A(\underline{x}) \in S] \leq e^{\epsilon} \mathbb{P}_R[A(\underline{x}') \in S].$$

(ii) The definition above is inspired by the structure of the tracing attack we saw last class: we want log-likelihoods to be "small".

(iii) How does one set ϵ? Typically, one would like ϵ to be "small" (≈ 0.1 or ≈ 1), but nothing stops us from setting $\epsilon \gg 1$. Clearly, in the latter case, the privacy guarantees (on small datasets) are unlikely to be meaningful.

[HW: Look up the values of ϵ used in the 2020 US Census releases, the 2024 Israel release of live birth statistics, and in real-world deployments by big tech firms. What do you think of these values?]

Achieving DP: A First Pass on Binary-Valued Data

Suppose that $\mathcal{D} = (x_1, \dots, x_N)$ is such that $x_i \in \{0,1\}$, $\forall i \in [N]$. A simple way of "privatizing" this dataset is to randomize each bit so that it is biased towards the true value.

This leads to RR [Randomized Response, Warner (1965)]:

→ For each $i \in [N]$, set $Y_i = \begin{cases} x_i, & \text{w.p. } 1-p, \\ 1-x_i, & \text{w.p. } p \end{cases}$ for $p \in (0, 1/2)$.

→ Return $(Y_1, \dots, Y_N)$.

HW: ① Argue that $\sum_{i=1}^{N} Y_i$ is a "good estimate" of the statistic $\sum_{i=1}^{N} x_i$, i.e.,

$$\left(\mathbb{E}\left[\left(\sum Y_i - \sum x_i\right)^2\right]\right)^{1/2} = O(\sqrt{n}).$$

② For what value of ϵ does RR(p) satisfy ϵ-DP?

③* Get rid of the "bias" term in ①, by constructing a new estimator that is an unbiased estimate of $\sum x_i$. Can you use a Chernoff bound to show that this estimate is "close" to $\sum x_i$, with high prob.? (at most α away)