Today: Differential Privacy

Goal: Privacy-preserving data analysis:

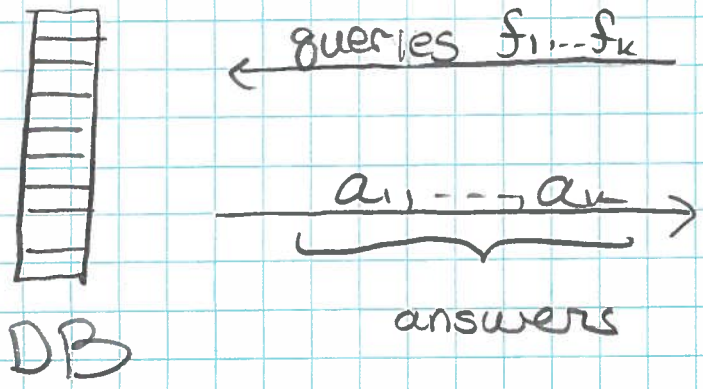
A trusted (and trustworthy) curator gathers sensitive information from a large number of respondents (the sample), and the goal is learning, and releasing to the public, statistical facts about the underlying population, without compromising the privacy of the individual respondents.

* Statistical analysis of sensitive data brings tremendous benefits, but has serious privacy risks.

Example: Statistical analysis on medical records of patients.

Challenge: Reconcile statistical analysis and privacy (even in the presence of linkage attacks)

Data Analysis Model



Two settings:

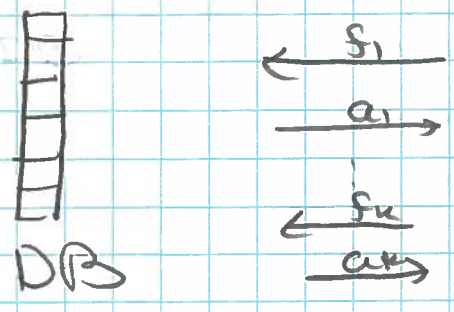
- Non-interactive setting:

  Curator computes & publishes the statistics $(a_1, \ldots, a_k)$, and the data is not used further.

- Interactive setting:

  The curator sits between DB & users

  Queries are posed adaptively.



* We won't distinguish between these two settings.

Two major privacy concerns:

- Linkage attacks: Protect against adversary that has partial information.
- Composition: Maintain privacy even when the data is used in multiple analyses (each privacy preserving in isolation).

Natural Approaches:

- Anonymization: Remove "personally identifiable information", and publish the "anonymized" dataset.
- "Just" releasing statistics: Only release aggregate statistics
- Perfect Privacy: Perfect protection of individuals?

Our focus: Differential Privacy.

Data anonymization is not safe!

Example: Netflix data release [Narayanan-Shmatikov]

Netflix released ratings for subset of movies & users, where usernames were replaced with random IDs (& some additional perturbation)

[NS08] used this data, together with public reviews from IMDB.com to de-anonymize netflix data.

⇒ Learn about movies that IMDB users didn't want to tell the world about.

This attack is a linkage attack.

"Just" releasing statistics is not safe

Suppose data was collected about the # of HIV patients among MIT students.
And suppose it is known that no-one (except me) has HIV.
In this case, this statistics violates my privacy.

* Auxiliary information is a common theme in privacy horror stories.

Lets aim for perfect privacy!

Semantic security ??

Anything that can be learned about participant from privacy-preserving data analysis can be learned without it.

Unachievable!

Example:

DB + Analysis: smoking causes cancer.

Aux info: I smoke in public.

$\Rightarrow$ I am "hurt": insurer knows that I am at risk for cancer.

Differential Privacy

[Dwork-McSherry-Nissim-Smith 06]

Intuition: The output distribution when I am in the database or out of the database, should be similar.

$\Rightarrow$ Risk incurred by participation is low.

Definition: A randomized algorithm K is ε-differentially private if $\forall S \subseteq \text{Range}(K)$ $\forall$ datasets D_1, D_2 differing on at most one element,

$$\Pr[K(D_1) \in S] \in (e^{-\varepsilon}, e^{\varepsilon}) \cdot \Pr[K(D_2) \in S]$$

$$\approx (1-\varepsilon, 1+\varepsilon)$$

where the prob is over the coin tosses of K_1.

Linkage attacks (or aux input)?

Guarantee is for any adjacent D_1 & D_2 (↑ D_1 & D_2 are adjacent if they differ in a single row)

Composition?

ℓ (adaptively chosen) algorithms, each ε-differentially private

$\Rightarrow$ Taken together is still $\ell \cdot \varepsilon$- differentially private.

Thm [Composition of differential privacy]:

If $K_1(D)$ is ε_1-diff. private, & $\forall z_1$ in range K_1 $K_2(D, z_1)$ is ε_2-diff. private,

then $K(D) \triangleq (K_1(D), K_2(D, K_1(D)))$ is $(\varepsilon_1+\varepsilon_2)$-diff private.

Pf: Fix $S \subseteq \text{RANGE}(K)$, and fix any adjacent datasets D, D'.

$$\Pr[K(D) \in S] = \sum_{(z_1,z_2)\in S} \Pr[K_1(D)=z_1] \cdot \Pr[K_2(D,z_1)=z_2]$$

$$\leq \sum_{(z_1,z_2)\in S} (e^{-\varepsilon_1}, e^{\varepsilon_1}) \Pr[K_1(D')=z_1] \cdot (e^{-\varepsilon_2}, e^{\varepsilon_2}) \Pr[K_2(D,z_1)=z_2]$$

$$= (e^{-(\varepsilon_1+\varepsilon_2)}, e^{\varepsilon_1+\varepsilon_2}) \sum_{(z_1,z_2)\in S} \Pr[K_1(D')=z_1] \cdot \Pr[K_2(D',z_1)=z_2] =$$

$\in [e^{-(\varepsilon_1+\varepsilon_2)}, e^{\varepsilon_1+\varepsilon_2}) \cdot \Pr[K(D') \in S]$ ∎

How do we construct differentially private algorithms?

[Focus: single function.]

Example: Not differentially private.

Suppose DB contains tuples (name, tag $\in \{0,1\}$)

Goal: approximate fraction of population with tag = 1

Answer: Approximate by Choosing a few random tags.

What if in DB only my tag is 1.

$$\Pr[K(D+me) \neq 0] \geq \frac{1}{n}$$

$$\Pr[K(D) \neq 0] = 0.$$

Not differentially private for any ε.

Differentially private alg:

Output #1's + noise

This is differentially private for "proper noise".

Laplacian Noise

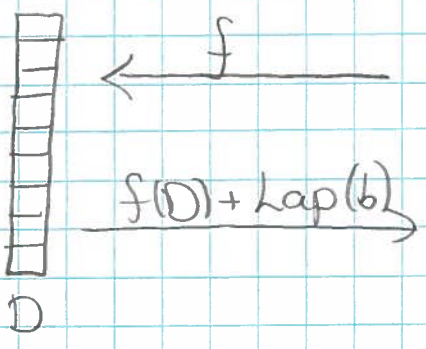
Laplace distribution $Y = Lap(b)$

density function $Pr[Y=y] = \frac{1}{2b} e^{-|y|/b}$

Standard deviation : $O(b)$

[Laplace dist. can be thought of as two exponential distributions spliced together back-to-back

ε-differentially private mechanism:



(Single function)

$f: D \longrightarrow \mathbb{R}^k$

Q : what parameter b ??

A : b depends on ε & on the "sensitivity" of f.

Def : For $f: D \longrightarrow \mathbb{R}^k$ the sensitivity of f is

$$\Delta f = \max_{\text{adjacent } D_1, D_2} \| f(D_1) - f(D_2) \|_1$$

In particular, for $k=1$ the sensitivity of f is the maximum difference in the values that the function f may take on a pair of adjacent databases.

For many types of queries Δf is quite small. For example, for counting queries ("How many rows have property P ?"), $\Delta f = 1$.

The differentially private mechanism works best (i.e., introduces the least noise) when Δf is small.

Note that sensitivity is a function of the function alone, and is independent of the database.

The sensitivity captures how great a difference (between the value of f on two adjacent databases) must be hidden by the additive noise generated by the curator.

On query f the ε-differentially private mechanism responds with:

$$\boxed{f(D) + \text{Lap}(\Delta f/\varepsilon)^k}$$

adding noise w. distribution $\text{Lap}(\Delta f/\varepsilon)$ independently to each coordinate of $f(D)$

Note: Decreasing ε yields larger expected noise magnitude. Increasing the sensitivity Δf also yields larger expected noise magnitude.

Thm: The mechanism described above is ε-differentially private.

"Pf": For $k=1$ (and single query).

Fix any $f: D \to \mathbb{R}$, any $S \subseteq \mathbb{R}$ (think of S as finite), any adjacent datasets D_1, D_2

$$\Pr[K(D_1) \in S] = \sum_{z \in S} \Pr[K(D_1) = z] = \sum_{z \in S} \Pr[f(D_1) + Lap(\Delta f/\varepsilon) = z]$$

$$= \sum_{z \in S} \Pr[Lap(\Delta f/\varepsilon) = z - f(D_1)] = \sum_{z \in S} \frac{1}{2\Delta f/\varepsilon} e^{-|z - f(D_1)| / (\Delta f/\varepsilon)}$$

$$= \sum_{z \in S} \frac{1}{2\Delta f/\varepsilon} e^{-|z - f(D_2) + f(D_2) - f(D_1)| / (\Delta f/\varepsilon)}$$

$$\in \sum_{z \in S} \frac{1}{2\Delta f/\varepsilon} e^{-|z - f(D_2)| / (\Delta f/\varepsilon)} \cdot \left[e^{-|f(D_2) - f(D_1)| / \Delta f/\varepsilon}, e^{|f(D_2) - f(D_1)| / \Delta f/\varepsilon} \right]$$

$$\in \Pr[K(D_2) \in S] \cdot \left(e^{-\varepsilon}, e^{\varepsilon} \right)$$

∎