Today: Differential Privacy

Confidentiality - The goal is to hide everything

Privacy - Delicate balance between confidentiality & utility.

Suppose: A hospital (or census bureau) wish to release statistical information about their data.

Two conflicting goals:

Utility: Release (valuable) aggregate statistics

Privacy: Individual info remains hidden

Differential Privacy [Dwork 06, Dwork-McSherry-Nissim-Smith 06]

Rigorous definition of privacy.

Privacy was studied since 1960.

First attempt: Anonymize - Remove obvious identifiers.

Known to be insecure due to external information.

Indeed, anonymization schemes are known to be often broken.

Eg. MA Group Insurance Commission (GIC) released "anonymized" data on state employees that showed every single hospital visit. The goal was to help researchers.

They removed all obvious identifiers, such as name, address, social security number, etc.

When GIC data was released, William Weld, who was the governer of MASS, assured the public that GIC protected the privacy of individuals.

Latanya Sweeney, a PhD student at MIT at the time, proved him wrong, and quickly reidentified Governer Weld's personal info (which included diagnoses & perscriptions).

This was done by using external info (she knew he lived in Cambridge, etc...)

What if we only release aggregate info?

Problematic!

What if a company releases the avg salary of its employees before & after a certain employee's resignation?

* Even a couple of "too accurate" statistics reveal individual information.

Eg: 2008 Homer et al. showed that given a persons DNA they could predict with high confidence whether that person belongs to the study group.

This caused the NIH to change their data sharing practices.

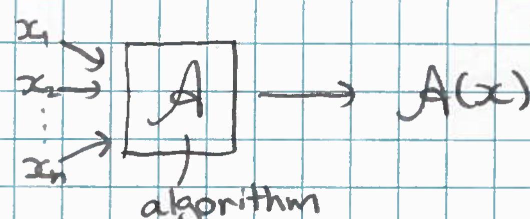
Differential Privacy (DP)

Provides rigorous guarantees against arbitrary external information.

Adopted by Apple, US Census Bureau, etc..

Given a Data set $x = (x_1, \dots, x_n) \in D^n$

D is the domain (can be numbers, medical data, tax forms, ...).



A is a (randomized) alg releasing info about the data.

DP guarantee: $\forall x \in D^n \quad \forall i \in [n] \quad \forall x_i' \in D$

$$A(x_1 \dots x_n) \approx A(x_1 \dots x_{i-1}, x_i', x_{i+1} \dots x_n)$$

Intuition: Changing one person's data does not change the outcome by much.

Formalizing this intuition:

Def: x & x' are neighboring datasets if they differ at one data point.

Def: Alg A is ε-DP if $\forall$ neighboring $x, x' \in D^n$ $\forall S$

$$\frac{\Pr[A(x) \in S]}{\Pr[A(x') \in S]} \leq e^{\varepsilon} \approx 1+\varepsilon$$

Note: May seem more natural to require $\forall S$

$$\left| \Pr[A(x) \in S] - \Pr[A(x') \in S] \right| \leq \varepsilon \quad \text{(statistical distance)}$$

The reason for the specific choice of def is that it is easier to achieve (and is a close variant to the "natural" one above), and provides stronger privacy guarantees.

How Do we get DP ?

Idea: Add noise to the response.

Eg: Suppose we want to estimate the proportion of diabetics in a dataset. Each person is associated with a bit $x_i \in \{0,1\}$

Idea: Randomize response:

with prob $1-\varepsilon$ report true val x_i

& with prob ε report $\bar{x}_i$

Say we want to release a summary $f(x)$

eg. proportion of diabetics $f(x) = \frac{1}{n} \sum x_i$.

Idea: Add noise, and publish $f(x) + \underbrace{\text{noise}}_{?}$.

How do we sample the noise?

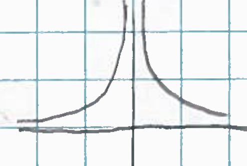
Natural to add Gaussian noise, but analyzing becomes complicated.

Instead: Laplace Mechanism

Noise is sampled from the Laplace distribution

$$h_{\mu,b} = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

← very similar to a Gaussian but has $|x-\mu|$ as opposed to $(x-\mu)^2$.



Denote $\text{Lap}(b) = h_{0,b}$ (mean $= 0$).
var $= 2b^2$

Which b should we use??

Def: Global sensitivity of f

$$GS_f = \max_{\substack{\text{neighboring} \\ x, x'}} \| f(x) - f(x') \|_1$$

Thm: If $A(x) = f(x) + \text{Lap}\left(\frac{GS_f}{\varepsilon}\right)$ then

A is ε-DP.

"Pf"

Fix any neighboring datasets x & x'

Assuming for simplicity A releases only one value

$$\Pr[A(x) = y]$$
$$\parallel$$
$$\Pr[\, f(x) + \underset{\substack{\wr \\ \text{Lap}(b)}}{\text{noise}} = y \,] \qquad b = \frac{GS_f}{\varepsilon}$$
$$\parallel$$
$$\frac{1}{2b} e^{\frac{|y - f(x)|}{b}}$$

$$\Longrightarrow \quad \frac{\Pr[A(x) = y]}{\Pr[A(x') = y]} = e^{\frac{1}{b}\left(|y - f(x)| - |y - f(x')|\right)} \leq$$

$$\leq e^{\frac{1}{b}|f(x) - f(x')|}$$

$$= e^{\varepsilon \frac{|f(x) - f(x')|}{GS_f}}$$

$$\leq e^{\varepsilon}$$

■

Note: $\text{Lap}\left(\frac{GS_f}{\varepsilon}\right)$ adds noise w. variance $2\frac{GS_f^2}{\varepsilon^2}$. Agrees w. intuition

more sensitivity $\longrightarrow$ more noise, smaller $\varepsilon \Longrightarrow$ more noise.

More generally: $f : \mathbb{D}^n \longrightarrow \mathbb{R}^\ell$

$$A(D) = f(D) + \mathrm{Lap}\left(\frac{GS(f)}{\varepsilon}\right)^\ell$$

Composition: If t ε-DP statistics are released, then the joint information is $t \cdot \varepsilon$ diff private.

"Thm" Suppose

$$\left.\begin{aligned} z_1 &= A_1(D) \\ z_2 &= A_2(D, z_1) \\ &\vdots \\ z_t &= A_t(D, z_{t-1}) \end{aligned}\right\} \text{all DP alg's}$$

Then $A(D) = (z_1, \ldots, z_t)$ is $t\varepsilon$-DP.

"Pf": Fix any neighboring D, D'

$$\begin{aligned} \Pr[A(D) = (z_1 \ldots z_t)] &= \Pr[A_1(D) = z_1] \cdot \Pr[A_2(D, z_1) = z_2] \cdot \ldots \\ &\le \Pr[A_1(D') = z_1] \cdot e^\varepsilon \cdot \Pr[A_2(D', z_1) = z_2] \cdot e^\varepsilon \cdot \ldots \\ &= \Pr[A(D') = z_1 \ldots z_t] \cdot e^{t\varepsilon} \end{aligned}$$

∎

What about accuracy ?

Eg. : Noisy summation

Suppose $f(x) = \sum_{i=1}^{n} d_i \quad d_i \in \{0,1\}$

Clearly $GS_f = 1$. Therefore,

$$A(D) = \sum_{i=1}^{n} d_i + Lap\left(\frac{1}{\varepsilon}\right) \quad \text{is } \varepsilon\text{-diff private.}$$

Note : $\mathbb{E}(A(D)) = f(D)$ since the mean of $Lap(\frac{1}{\varepsilon})$ is 0.

By Chebyshev's inequality :

$$\Pr\left[\,|A(D) - f(D)| > k\,\right] \leq \frac{2}{\varepsilon^2 k^2}$$

[Chebyshev : $\Pr[|X-\mu| \geq k] < \frac{\sigma^2}{k^2}$
In $Lap(\frac{1}{\varepsilon})$, $\sigma^2 = 2/\varepsilon^2$]

So if $\varepsilon = 0.1$ we can be 95% sure that $A(D)$ does not deviate from the actual answer by roughly 64

$$\left[\frac{2}{\varepsilon^2 k^2} = \frac{200}{k^2} \leq \frac{5}{100} \iff k^2 \geq 4000 \quad \text{which holds for } k \geq 64 \right]$$

If n is large enough $n \pm 64$ may be a "reasonable" error.

Note : We should think of $\varepsilon \approx 0.1$,
as opposed to 2^{-80}...

Eg. Noisy average

Suppose $f(D) = \frac{1}{n}\sum d_i$ $\quad d_i \in \{0,1\}$.

Then $GS_f = \frac{1}{n}$

$$A(D) = \frac{1}{n}\sum d_i + Lap\left(\frac{1}{n\varepsilon}\right) \qquad \Pr[|A(D)-f(D)|>k] \leq \frac{2}{n^2\varepsilon^2 k^2}$$

$\sigma^2 = \frac{2}{n^2\varepsilon^2}$

Chebyshev

$\Rightarrow$ The more data the more accuracy we get!
(for the same level of privacy)

Interpreting DP

A naive (incorrect) hope:

Your beliefs about me are the same after seeing the output as they were before.

Incorrect hope:

Eg: Suppose you know I smoke.

A study shows smoking causes cancer.

$\Rightarrow$ You learn something about me!

DP Guarantee: No matter what you know ahead of time: You learn (almost) the same things about me whether or not my data is used.