Theory of Differential Privacy

Lecture 2:

DP, Randomized Response, Laplace Mechanism.

EN.601.438/638 Fall 2026

Lydia Zakynthinou
Johns Hopkins University

Sept 2, 2026

K-anonymity

= data release is k-anonymous

if ∀ record i has at least

k-1 records with identical

"non-sensitive" attributes.

	Non-Sensitive			Sensitive
i	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

[BU/NEU course]

Cons

1. Homogeneous block (e.g. 1-4 have same disease, Alice is 29yo + lives in 13030)

2. Composition of k-anonymous datasets

$\not\Rightarrow$ $K' > 1$ - anonymous.

Randomized Response (Warner '65)

Input $x \in \{0, 1\}$ (secret, "has disease or not")

Output:

$$RR(x) = \begin{cases} x & , \text{w.p. } 2/3 \\ 1-x & , \text{w.p. } 1/3 \end{cases}$$

$Y = RR(x)$ random variable.

Prob.	$Y=0$	$Y=1$
$x=0$	$2/3$	$1/3$
$x=1$	$1/3$	$2/3$

randomness in R

$$\frac{\Pr[Y=1 \mid x=0]}{\Pr[Y=1 \mid x=1]} = \frac{1}{2}, \quad \frac{\Pr[Y=0 \mid x=1]}{\Pr[Y=0 \mid x=0]} = \frac{1}{2}$$

Bayes' Rule

$$\Pr[A \mid B] = \frac{\Pr[B \mid A] \cdot \Pr[A]}{\Pr[B]}$$

Interpretation

"I believe $x=0$ ℓ times more than $x=1$,

$$\ell = \underbrace{\frac{\Pr[x=0 \mid Y=0]}{\Pr[x=1 \mid Y=0]}}_{\text{posterior belief}} \overset{BR}{=} \frac{\Pr[x=0]}{\Pr[x=1]} \cdot \frac{\Pr[Y=0 \mid x=0]}{\Pr[Y=0 \mid x=1]} = \underbrace{\frac{\Pr[x=0]}{\Pr[x=1]}}_{\text{prior}} \cdot 2$$

randomness in the world

Randomized Response

$$x \in \{0,1\}, RR_\varepsilon(x) = \begin{cases} x, & \text{w.p. } \frac{e^\varepsilon}{e^\varepsilon+1} = p_\varepsilon \\ 1-x, & \text{w.p. } \frac{1}{e^\varepsilon+1} = q_\varepsilon. \end{cases}$$

$\varepsilon \geq 0$

$$\frac{\Pr[Y=1 \mid x=1]}{\Pr[Y=1 \mid x=0]} = e^\varepsilon,$$ calculated similarly as before (where $\varepsilon = \ln 2$).

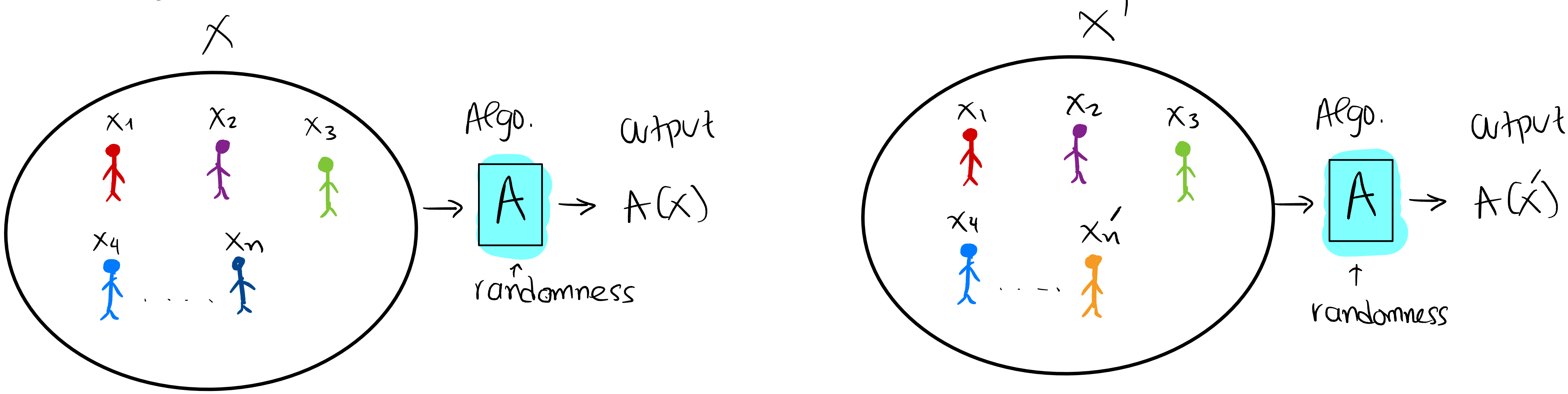
$\varepsilon \nearrow$: less privacy

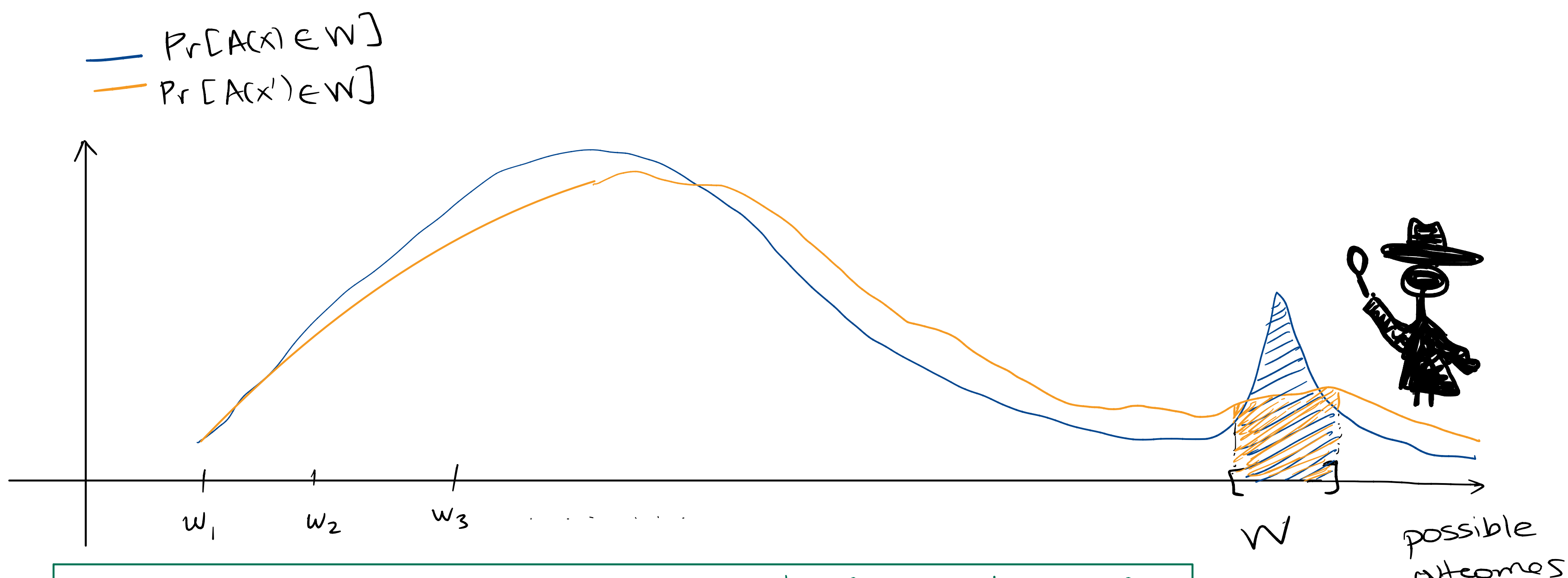
$\varepsilon = 0$: tell truth or lie w.p. 50% $\Rightarrow$ no information extracted from RR_ε
$\Rightarrow$ perfect privacy.

Distribution of output on neighboring datasets

From Lec. 1

Neighboring datasets $X \sim X'$: differ in one person's record.





Could an observer reliably tell which world generated the output?

Def. Two datasets $x=(x_1,\dots,x_n)$, $x'=(x_1',\dots,x_n')$, $x,x'\in\mathcal{X}^n$, are neighboring, denoted by:

$$x\sim x' \quad \text{if} \quad \exists\, i^* : \quad x_j = x_j' \quad \text{for every } j\neq i^*.$$

An algorithm A is ε-DP if for all pairs $x,x'\in\mathcal{X}^n$, $x\sim x'$, for all measurable* output sets S:

$$\Pr[A(x)\in S]\leq \Pr[A(x')\in S]\cdot e^{\varepsilon}$$

* mostly ignore.

Th. RR_ε is ε-DP.

Let RR_ε from before be defined as follows when applied to input $x\in\{0,1\}^n$

For $x=(x_1,\dots,x_n)\in\{0,1\}^n$: $RR_\varepsilon(x)=(RR_\varepsilon(x_1),\dots,RR_\varepsilon(x_n))\in\{0,1\}^n$.

For any pair $x\sim x'$, for any subset of outputs $E\subseteq 2^{\{0,1\}^n}$, need to show

$$\Pr[RR_\varepsilon(x)\in E]\leq e^{\varepsilon}\Pr[RR_\varepsilon(x')\in E]$$

Proof Consider fixed:

$x \sim x'$: differ in i^*, $(y_1, \dots, y_n) \in \{0,1\}^n$: output.

$$\frac{\Pr[RR_\varepsilon(x) = (y_1, \dots, y_n)]}{\Pr[RR_\varepsilon(x') = (y_1, \dots, y_n)]} = \frac{\prod_{j=1}^{n} \Pr[RR_\varepsilon(x_j) = y_j]}{\prod_{j=1}^{n} \Pr[RR_\varepsilon(x'_j) = y_j]}$$

$$\overset{x \sim x'}{=} \frac{\Pr[RR_\varepsilon(x_{i^*}) = y_{i^*}]}{\Pr[RR_\varepsilon(x'_{i^*}) = y_{i^*}]} = \begin{cases} \dfrac{p_\varepsilon = e^\varepsilon/(1+e^\varepsilon)}{q_\varepsilon = 1/(1+e^\varepsilon)} = e^\varepsilon, \text{ if } y_{i^*} = 1, x_{i^*} = 1, x'_{i^*} = 0 \\ \vdots \\ \dfrac{q_\varepsilon}{p_\varepsilon} = e^{-\varepsilon} \end{cases}$$

(case analysis)

So for a fixed $(y_1, \dots, y_n)$:

$$\Pr[RR_\varepsilon(x) = (y_1, \dots, y_n)] \le e^\varepsilon \cdot \Pr[RR_\varepsilon(x') = \underbrace{(y_1, \dots, y_n)}_{y}] \quad (1)$$

For any set of outputs E:

$$\Pr[RR_\varepsilon(x) \in E] = \sum_{y \in E} \Pr[RR_\varepsilon(x) = y] \overset{(1)}{\le} e^\varepsilon \sum_{y \in E} \Pr[RR_\varepsilon(x) = y] \le e^\varepsilon \cdot \Pr[RR_\varepsilon(x') \in E] \qquad \blacksquare$$

Laplace Mechanism

Input $x = (x_1, \dots, x_n)$.

Want to compute $f(x)$.

$$\boxed{A(x) = f(x) + \text{"noise"}}$$

Global Sensitivity

function $f: \mathcal{X}^n \to \mathbb{R}^d$

G.S. $$\Delta_f = \sup_{x \sim x'} \overbrace{\| f(x) - f(x') \|_1}^{= \sum_{i=1}^{d} |f(x)_i - f(x')_i|}$$

The Laplace Mechanism returns:

$$A_{Lap}(x) = f(x) + Lap(b) \equiv f(x) + \underset{\swarrow \text{r.v.}}{Z}, \text{ where } Z \sim Lap(\underset{\text{mean}}{0}, \underset{\text{scale}}{b})$$

$$\equiv Lap(f(x), b).$$

Th. A_{Lap} is ε-DP for $b = \frac{\Delta}{\varepsilon}$, where $\Delta_f \le \Delta$.

For simplicity: $d=1$, i.e., $f: X^n \to \mathbb{R}$.

<u>Proof sketch</u>: The probability density function of Laplace is

$$p(z) = \frac{1}{2b} e^{\left(\frac{-|z-\mu|}{b}\right)} \quad \left(\approx \Pr[Z=z] \text{ when } Z \sim Lap(\mu, b)\right)$$

Given neighboring datasets $x \sim x'$.

$$A_{Lap}(x) \sim Lap\left(f(x), \frac{\Delta}{\varepsilon}\right) \qquad A_{Lap}(x') \sim Lap\left(f(x'), \frac{\Delta}{\varepsilon}\right).$$

Their respective PDFs.

$$P_x(y) = \frac{1}{\left(\frac{\Delta}{\varepsilon}\right)} \cdot e^{-\frac{|z-f(x)|}{(\Delta/\varepsilon)}} \qquad P_{x'}(y) = \frac{1}{(\Delta/\varepsilon)} e^{-\frac{|y-f(x')|}{(\Delta/\varepsilon)}}$$

$$\Rightarrow \frac{p_x(y)}{p_{x'}(y)} = e^{\frac{-|y-f(x)|+|y-f(x')|}{(\Delta|\varepsilon)}} \overset{\text{triangle ineq.}}{\leq} e^{\frac{|f(x)-f(x')|}{(\Delta|\varepsilon)}} \overset{(*)}{\leq} e^{\frac{\Delta}{(\Delta|\varepsilon)}} = e^{\varepsilon}.$$

(*) by assumption $\Delta \geq \Delta_f = \sup_{x\sim x'} |f(x)-f(x')| \geq |f(x)-f(x')|$. $\blacksquare$

Extension to $d>1$: $A_{Lap}(x) = f(x) + (z_1, .., z_d)$, where z_i iid $Lap(\frac{\Delta}{\varepsilon})$.

For a d-dimensional $z = (z_1, .., z_d)$ the PDF is $p(z) = \frac{1}{2b} e^{-\frac{\|z-\mu\|_1}{b}}$.

Proof follows similarly since $\|f(x)-f(x')\|_1 \leq \Delta_f \leq \Delta$.

Next time:

- Accuracy of Laplace & Randomized Response
- Properties of DP that allow us to combine primitives like Laplace Mech. & RR to more complex private mech.