

Differential Privacy as a Defense Mechanism: Concepts and Properties

Ayşe Ünsal

Digital Security Dept., EURECOM

Cyber in Occitanie: Summer School in Cybersecurity

July 10, 2025



Outline

- The Notion of Differential Privacy (DP)
- Various Definitions and Parameters
- Fundamentals
 - Global/Local DP
 - Mechanisms
- Components and Properties
 - Composition Theorem
 - Post-processing Invariance
- DP meets ML
- DP against Adversarial and Privacy Attacks
 - Membership Inference Attacks, Link Inference Attacks
 - Evasion Attacks-Adversarial Classification

The notion of DP

Identifier Demographic attributes			Sensitive/Confidential attributes			
Name	Gender	Age	Weight kg	Pulse/min	SpO ₂ %	BP
Alice	F	33	64	81	97	115/73
Bob	M	25	61	112	99	117/76
Arthur	M	48	65	105	90	129/77
David	M	30	73	75	100	117/76
Chloe	F	56	76	92	93	152/94
Eve	F	38	75	78	98	120/80

the use of data containing personal information has to be restricted in order to protect individual privacy

The notion of DP

- What if we simply remove the names?¹

Identifier	Demographic attributes		Sensitive/Confidential attributes			
	Gender	Age	Weight kg	Pulse/min	SpO ₂ %	BP
	F	33	64	81	97	115/73
	M	25	61	112	99	117/76
	M	48	65	105	90	129/77
	M	30	73	75	100	117/76
	F	56	76	92	93	152/94
	F	38	75	78	98	120/80

Sensitive
information
might still leak

¹ A. Narayanan et al. IEEE SP 2008, L. Sweeney et al. 2002, and so on.

The notion of DP

- Syntactic privacy

Name	Mail	Age	Profession	Diagnosis
Alice	alice@example.com	27	accountant	burn-out
Bob	abc@example.org	35	teacher	cancer

Direct identifiers Quasi identifiers Sensitive attributes

- Assumption: only these three groups (mutually exclusive) without any overlap
- Idea: Removing identifiers will prevent re-identification
- Linking attacks based on public attributes

Source: Zapatka et al., “Short Summary of Syntactic Privacy”, 2023

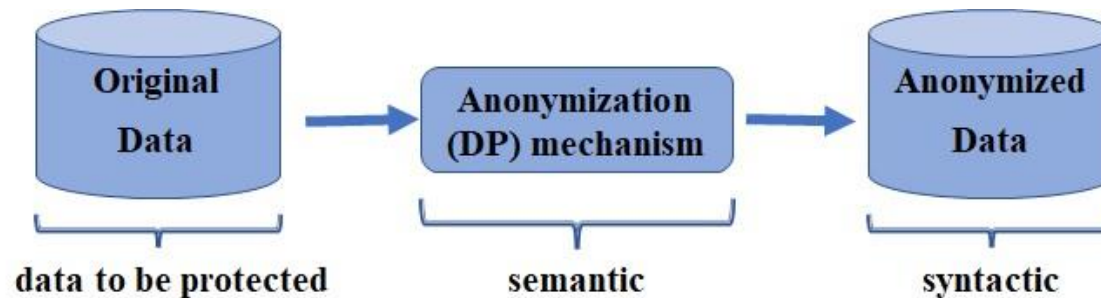
The notion of DP

- Gender, date of birth, and zip code are sufficient to uniquely identify the vast majority of Americans
- Linking these attributes in a supposedly anonymized healthcare database to public voter records, Latanya Sweeney² managed to identify the individual health record of the Governor of Massachusetts
- Need for a robust definition of privacy→Linkage attacks
 - Immune to attacks using auxiliary knowledge

² L. Sweeney, “k-anonymity: A model for Protecting Privacy”, *Int. J. Uncertainty Fuzziness and Knowledge-Based Systems* 10, 2002

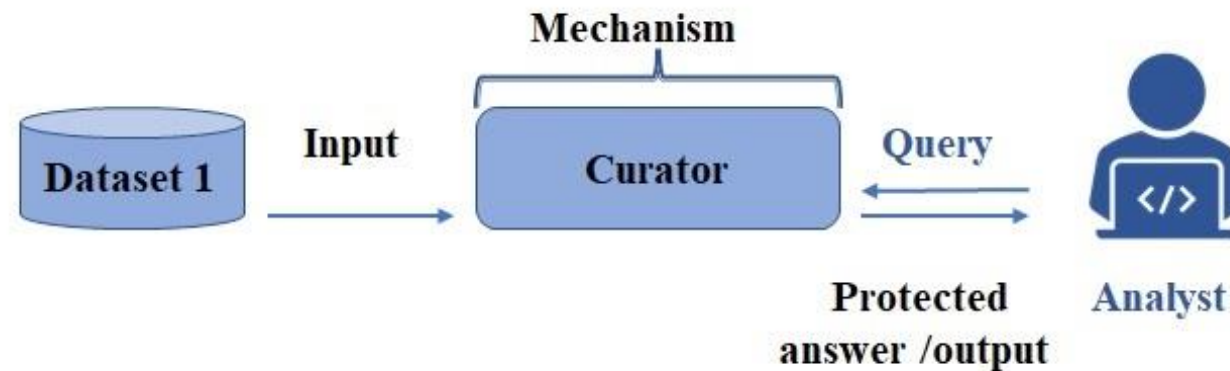
The notion of DP

- Requirement for a privacy measure:
 - Personal data processing \Leftrightarrow Right to privacy
- Privacy: Syntactic vs Semantic
 - Syntactic privacy is a property of the dataset, statistical disclosure control approach (e.g. k-anonymity)
 - Semantic privacy ensures a privacy property on the mechanism anonymizing the data (e.g., ϵ -DP)



The notion of DP

- Components:
 - Database → individual records
 - User/curator → a trusted entity to protect data privacy
 - Analyst → executes computations on the dataset



The notion of DP

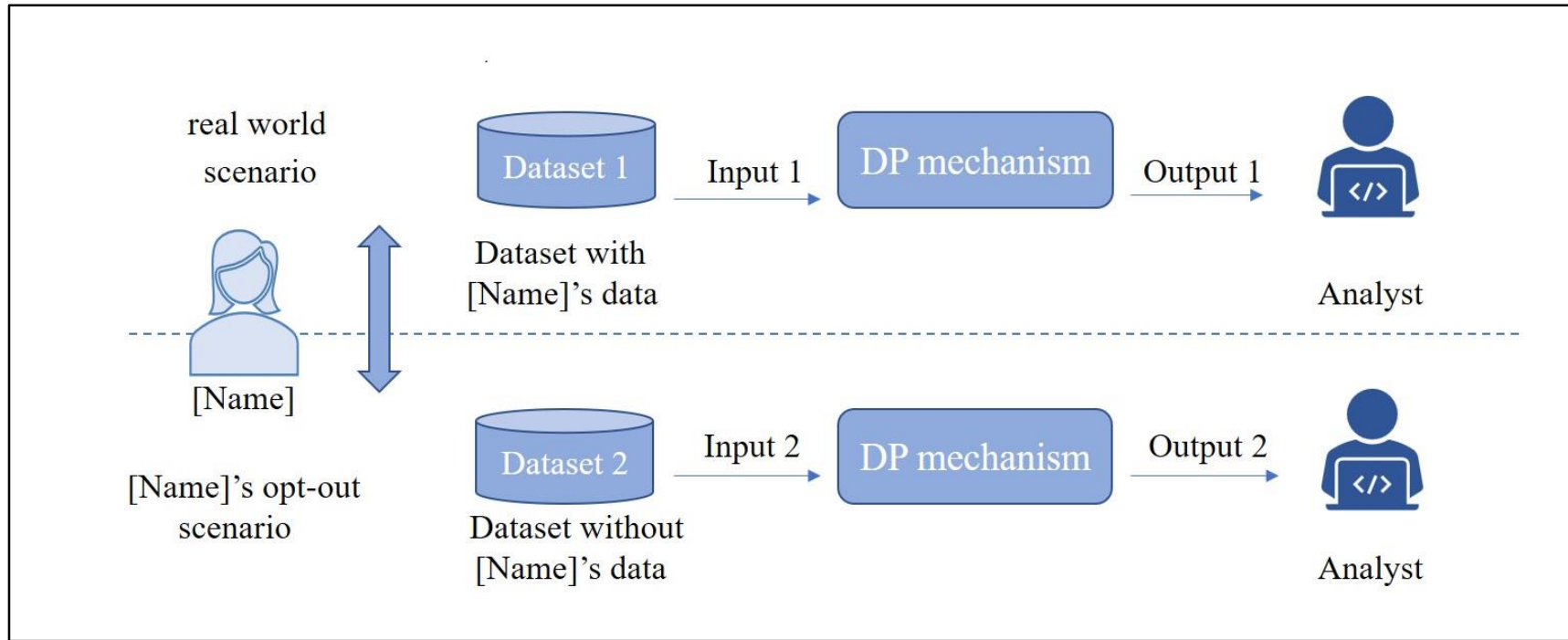
- Differential Privacy describes a promise, made by a data holder, or curator, to a data subject (owner), and the promise is like this:

“You will not be affected adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, datasets or information sources are available”³

³ C. Dwork and Aaron Roth (2014), ”The Algorithmic Foundations of Differential Privacy”, 2014

The notion of DP

- Does the protected answer disclose any information of an individual?
 - The absence or presence of a single person's information does not affect the outcome of the analysis



Definitions and Parameters

- Definition (ϵ, δ) - DP [Dwork and Roth 2014]:

A randomized algorithm Y is (ϵ, δ) - differentially private if $\forall S \subseteq \text{Range}(Y)$ and for all neighboring datasets x and x' within the domain of Y the following inequality holds.

$$\Pr[Y(x) \in S] \leq \Pr[Y(x') \in S] \exp\{\epsilon\} + \delta$$

Y	The mechanism: query(db) + noise or query(db+noise)
x and x'	Entries in neighboring databases
S	All potential output of Y that could be predicted
ϵ	Max distance between a query on databases x and x'
δ	Probability of information accidentally being leaked

Definitions and Parameters

- **Worst-case** privacy measure

A randomized algorithm Y is (ϵ, δ) - differentially private if $\forall S \subseteq \text{Range}(Y)$ and for all neighboring datasets x and x' within the domain of Y the following inequality holds.

$$\Pr[Y(x) \in S] \leq \Pr[Y(x') \in S] \exp\{\epsilon\} + \delta$$

- The adversary knows all the information but 1-entry!
- More (dp) noise lower utility \rightarrow privacy-utility trade-off

Definitions and Parameters

- (ϵ, δ) -DP is also called approximate DP
- ϵ, δ parameters are privacy loss
- The risk to one's privacy caused by a DP algorithm is bounded by ϵ, δ
- Comparison between running a query Y over database x and x'
- ϵ, δ measures how much two probabilities of random distributions of x and x' can differ
 - ϵ Privacy budget/parameter

Definitions and Parameters

- Privacy budget ϵ :
 1. A metric of privacy loss at a differential change in data; 1 entry
 2. Opposite relation with privacy
- ϵ small; higher privacy but less accurate responses
 - the inputs of the queries are very similar then the outputs will be very similar too.
- ϵ high ; lower privacy
 - An output Y is very unlikely for databases x and x'

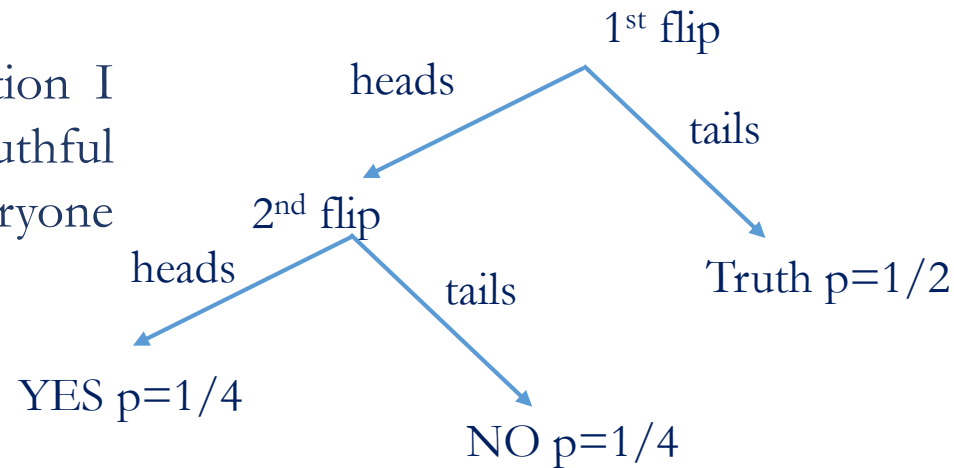
Definitions and Parameters

- Parameter δ :
 - If $\delta = 0$, Y is ϵ -DP
$$\frac{Pr[Y(x) \in S]}{Pr[Y(x') \in S]} \leq \exp\{\epsilon\}$$
 - If $\delta > 0$, with probability $1 - \delta$, we get the same guarantee of ϵ -DP.
 - Common approach is to set $\delta \leq$ the inverse of any polynomial in the size of the database
 - ϵ is independent of the database size
 - For δ , there is a higher chance of privacy leak with the database size
- No surprise: DP works better on larger databases.
 - The effect of any single individual on a given aggregate statistic diminishes as the number of individuals in a database grows

Fundamentals – Randomized Response

- Randomized response (Warner, 1965) is the forerunner to DP
- DP enforces privacy through randomization
- Used in survey interviews to determine the proportion in a group with a certain characteristic
- Individuals required to answer to sensitive queries in confidence YES/NO

If I ask this question I will not get truthful responses. I tell everyone to flip a coin!



This is an example where $p=1/2$

Fundamentals - DP

- How about DP?
- A dataset $D = \{X_1, X_2, \dots, X_i, \dots, X_n\}$, for $X_i \in \mathcal{X}$ and D is in \mathcal{X}^n and its neighbour $D' = \{X_1, X_2, \dots, X'_i, \dots, X_n\}$, we write $D \sim D'$
- We want to report $Y = f(D)$, via some randomization, so $Y \sim Q(\cdot | X_1, X_2, \dots, X_n)$
- Q satisfies ϵ -DP if

$$Q(Y \in A | D) \leq Q(Y \in A | D') \exp\{\epsilon\},$$

for all A and all pairs of $D \sim D'$.

- If Q has density q ;

$$\sup_y \frac{q(y|D)}{q(y|D')} \leq \exp\{\epsilon\}$$

- What do you see here?

Fundamentals - DP

- It means that whether or not you are in the database, this has little affect on the output Y .

- I think you are person i in the database, I want to know whether

$$X_i = a \text{ or } X_i = b$$

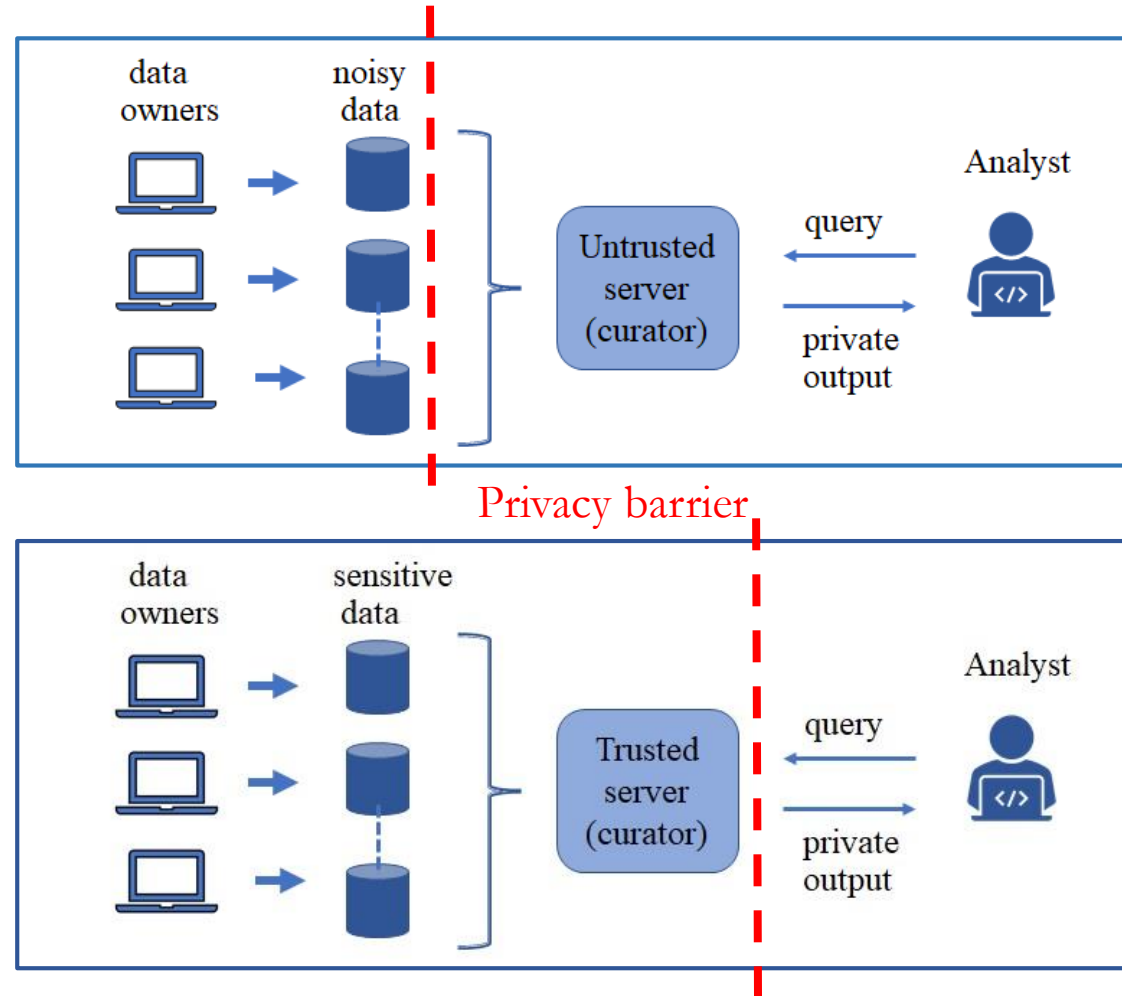
- After I see Y , my odds are

$$\frac{P(X_i = a|Y)}{P(X_i = b|Y)} = \frac{p(y|X_i = a)P(X_i = a)}{p(y|X_i = b)P(X_i = b)}$$
$$\exp\{-\varepsilon\} \frac{P(X_i = a)}{P(X_i = b)} \leq \frac{P(X_i = a|Y)}{P(X_i = b|Y)} \leq \frac{P(X_i = a)}{P(X_i = b)} \exp\{\varepsilon\}$$

- If ε is small, knowing Y does not change much since $\exp\{\varepsilon\} \approx \varepsilon + 1$.

Fundamentals – Local & Global Setting

- Local vs Global DP:



Fundamentals - Local & Global Setting

- Randomized response is a local mechanism, no need to a trusted server/curator
 - How do they differ in the context of DP?
- 1) **Local DP:** Applied on raw data at individual devices (or sensors)
Use case: RAPPOR in Google⁴, Apple iOs Private Count Mean Sketch
 - 2) **Global DP:** Applied at the central server (e.g. query output)
Use Case: US Census Bureau
 - 3) **Distributed DP:** Halfway between the two; applied when data is distributed across servers (or devices)

⁴ Erlingsson et al. “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”, 2014

Fundamentals - Local & Global Setting

Setting	Pros	Cons
Local	Raw data never shared, better privacy No need to a trusted curator	High values of ϵ or more data since total noise is high Worse utility
Global	Raw data shared Requirement of a trusted server	Less total noise Better utility : Accuracy with low values of ϵ

Fundamentals – Laplace Mechanism

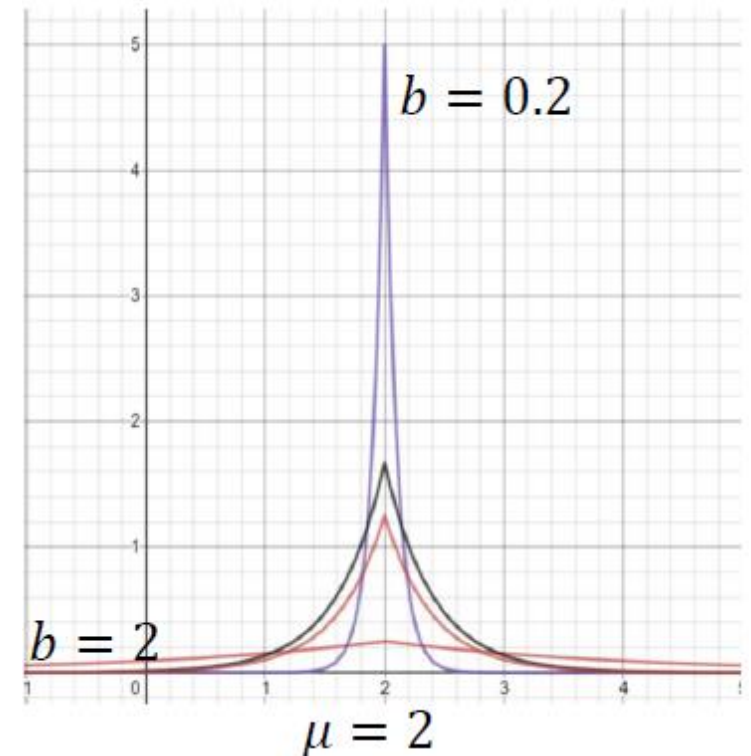
- Laplace Mechanism is defined for a function $f: \mathcal{D} \rightarrow \mathbb{R}^d$ where \mathcal{D} is the domain of the dataset D and d is the output dimension. Mechanism adds Laplace noise to the output of f as

$$\mathcal{A}(D) = f(D) + \text{Lap}(0|b)^d$$

- Laplace distribution with location and scale parameters μ and b :

$$\text{Lap}(x|b) = \frac{1}{2b} \exp\left\{-\frac{|x-\mu|}{b}\right\}$$

- Applies to any sort of numeric query



Fundamentals - Mechanism

- How do we know *how much noise* to add?
 - Sensitivity!
- Global sensitivity:

$$\Delta f = \max_{D, D'} \|f(D) - f(D')\|_1$$

- The smallest possible upper bound on the images of a query when applied to neighbours.
- Opposite relationship with the privacy $\longrightarrow b = \frac{\Delta f}{\epsilon}$
- Higher sensitivity \longrightarrow a stronger requirement for a privacy guarantee
- Consequently more noise is needed to achieve that guarantee

Fundamentals – Laplace Mechanism

Bounded vs unbounded neighbourhood

x	x'	x	x'
1	1	1	1
0	0	0	0
0	1	0	0
1	1	1	1
0	0	0	0
1	1	1	1
			1

Fundamentals – Laplace Mechanism

- Example: The sensitivity of the mean and how it applies to Laplace mechanism

- Let us start with the sensitivity!

$$f(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- # of participants is public information
 - The curator collects the true answers YES/NO to compute \bar{X}
 - Generates DP noise following $N \sim \text{Lap}(0, b)$ where $b = \Delta f / \epsilon$
 - What is Δf for the mean?
 - Say $n = 6 \rightarrow \bar{x} = \frac{3}{6}, \bar{x}' = \left\{ \frac{2}{6}, \frac{4}{6} \right\}$
 - $\Delta f = \max_{x, x'} |\bar{x} - \bar{x}'| = 1/n$

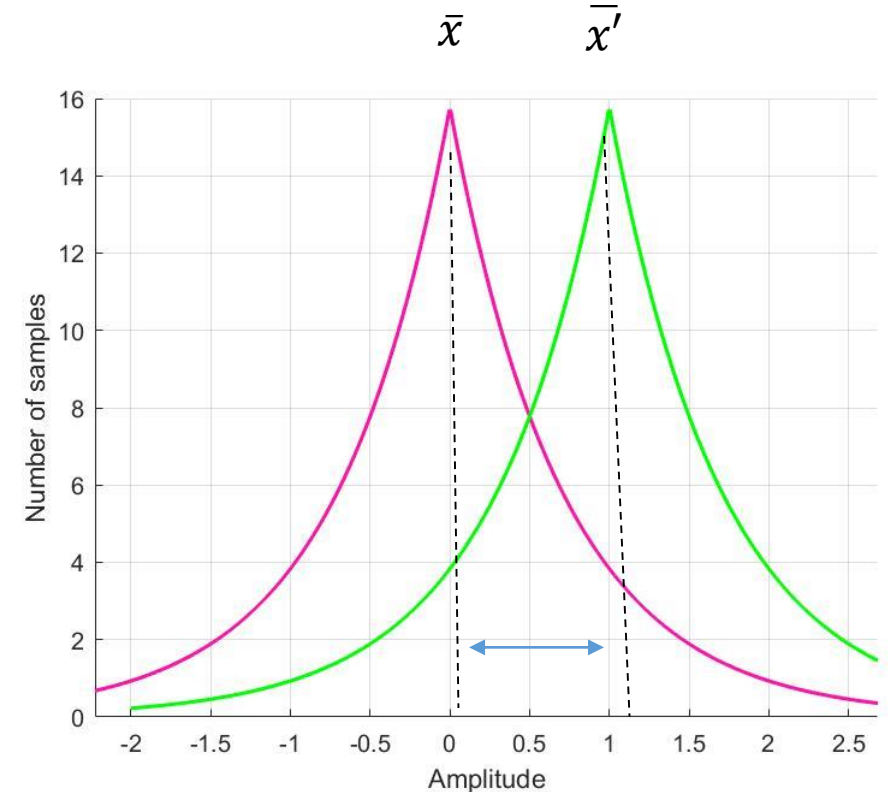
x	x'
1	1
0	0
0	1
1	1
0	0
1	1

Fundamentals – Laplace Mechanism

- Example: How it applies to Laplace mechanism, is Laplace mechanism ε -DP?

$$\begin{aligned} Y &= \bar{x} + N \\ \frac{P_Y(y|x)}{P_Y(y|x')} &= \frac{\exp\{-n\varepsilon|\bar{x} - y|\}}{\exp\{-n\varepsilon|\bar{x}' - y|\}} \\ &\leq \exp\{-n\varepsilon|\bar{x} - y - \bar{x}' + y|\} \\ &= \exp\{-\varepsilon\} \text{ since } \Delta f = 1/n \end{aligned}$$



❖ Upper bound due to $|a - b| \geq ||a| - |b||$,
 $a, b \in \mathbb{R}$



Fundamentals – Laplace Mechanism

- Utility: Error in the form of

$$\Pr[|Y - f(X)| \geq \alpha] < \beta$$

error  accuracy  tolerance

- In case of mean for Laplace mechanism, $d = 1$ with $\Delta f = \frac{1}{n}$.
- The output $Y = f(X) + N$ with $N \sim \text{Lap}(0, 1/\epsilon n)$
- The error $|Y - f(X)|$ is upper bounded by $\mathcal{O}(1/\epsilon n)$ since due to linearity

$$E[Y] = E[f(X)] \text{ and } \text{Var}[Y] = (1/\epsilon^2 n^2)$$

- Plugging in Chebyshev gives

$$|Y - f(X)| < \mathcal{O}(1/\epsilon n)$$

Fundamentals – Gaussian Mechanism

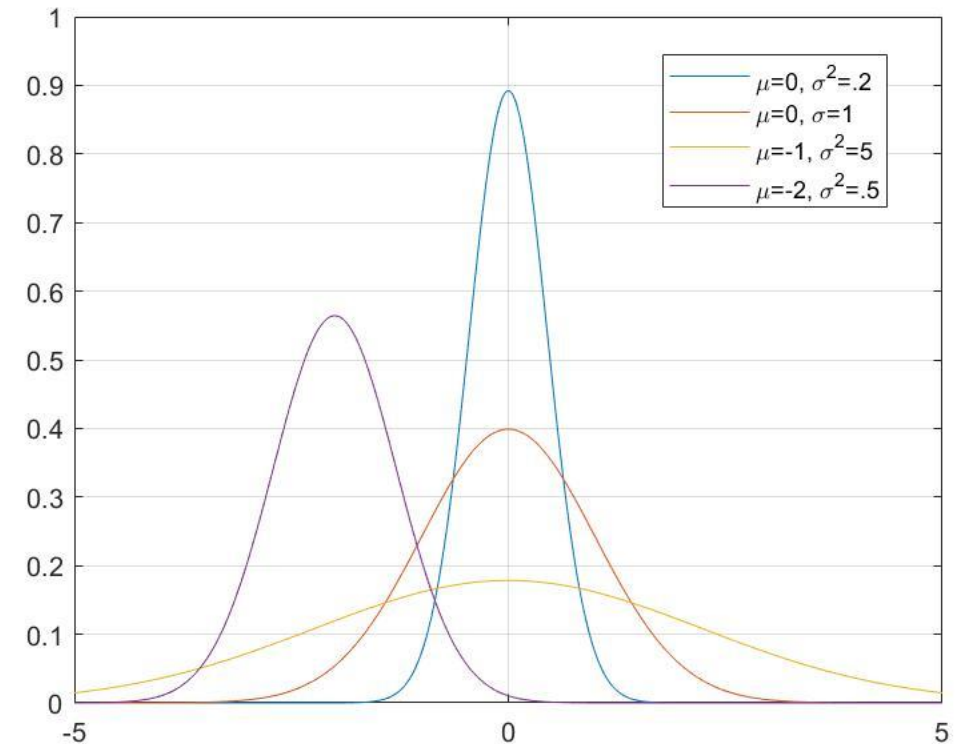
- Gaussian Mechanism defined for a function $f:\mathcal{D}\rightarrow\mathbb{R}^d$ where \mathcal{D} is the domain of the dataset D and d is the output dimension. Mechanism adds Gaussian noise to the output of f as

$$\mathcal{A}(D)=f(D)+Z^d$$

$$\text{where } Z\sim N(0, \sigma^2), \quad \sigma^2 = \frac{0.2 \ln(\frac{1.25}{\delta}) \Delta f^2}{\epsilon^2}.$$

- The Gaussian mechanism is (ϵ, δ) -DP⁵.

⁵ C. Dwork and Aaron Roth (2014), "The Algorithmic Foundations of Differential Privacy", 2014



Fundamentals – Gaussian Mechanism

- l_2 – sensitivity: $\Delta f_2 = \max_{D, D'} \|f(D) - f(D')\|_2$
- Let us take the example of mean in Gaussian mechanism $f(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ for $X \in \{0,1\}^d$.
- l_2 – sensitivity of the mean: $\frac{\sqrt{d}}{n}$ vs. l_1 – sensitivity of the mean $\frac{d}{n}$
- Maximum difference between neighbors is $\frac{1}{n} \mathbf{1}$.

Reminder: We had used l_1 – sensitivity for Laplace mechanism.

l_1 and l_2 – sensitivities are l_1 and l_2 norm of the max difference between neighbors.

Fundamentals – Gaussian Mechanism

- Why l_2 – sensitivity rather than l_1 – sensitivity for Gaussian mechanism?
 - The error!
 - Laplace mechanism is ϵ –DP with Laplace noise magnitude $d/n\epsilon$ and error $\mathcal{O}(\frac{d^{3/2}}{n\epsilon})$
 - Gaussian mechanism is (ϵ, δ) –DP with Gaussian noise magnitude $\mathcal{O}(\frac{\sqrt{d \log(\frac{1}{\delta})}}{n\epsilon})$ to each coordinate and error $\mathcal{O}(d/n\epsilon)$.
- the Gaussian mechanism can add a factor of $\mathcal{O}(\sqrt{d})$ less noise with a weaker privacy guarantee!
- In multivariate cases, it may be a better choice than Laplace!

Properties - Composability

- Composability of DP prevents accumulated privacy leakage over several independent analyses^{6,7}
 - A set of mechanisms represented by different queries each individually satisfying DP, also collectively satisfies DP

Theorem: [Sequential Composition]

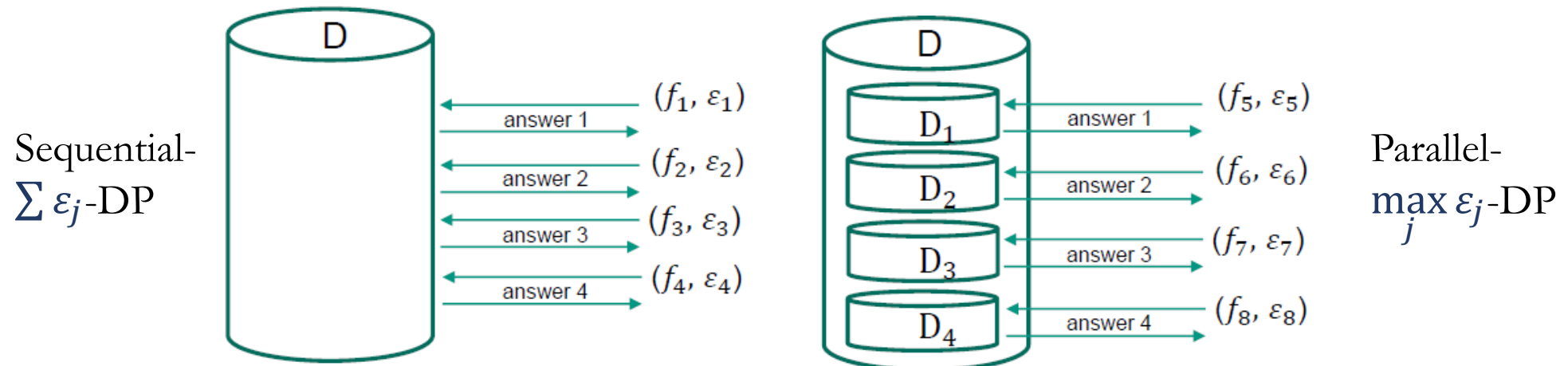
For ϵ_j -differentially private sequence of mechanisms $M = (M_1, M_2, \dots, M_m)$ defined over $\mathcal{X}^n \rightarrow \mathcal{Y}^m$, which is run (over the same input data) independently, composability of DP ensures that M satisfies $\sum \epsilon_j$ -DP.

^{6,7} Dwork et al. 2010 & Kairouz et al. 2015

Properties - Composability

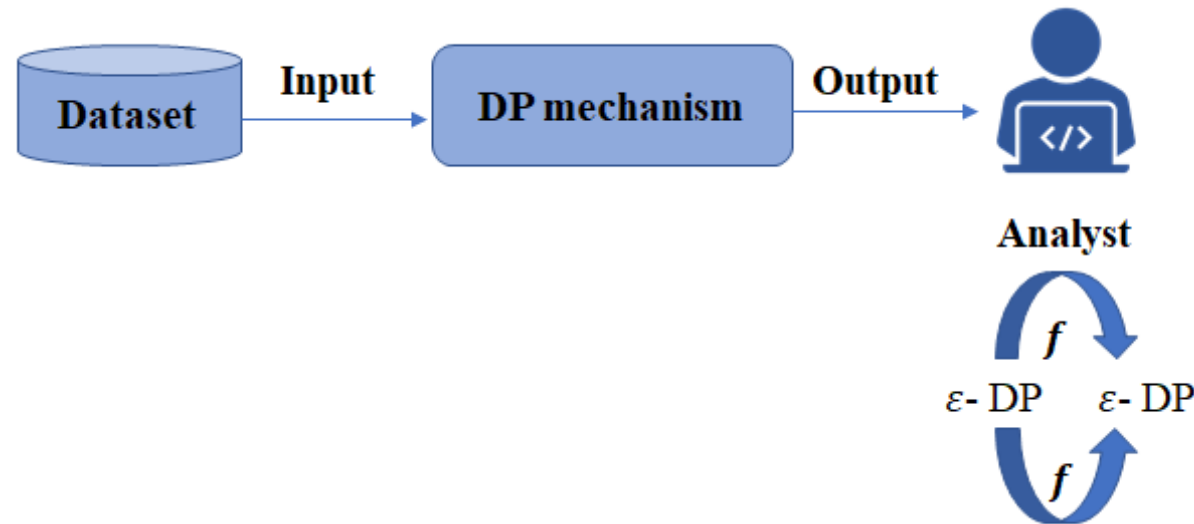
- Parallel Composition: An alternative to sequential composition - a second way to calculate a bound on the total privacy cost of multiple data releases.
- The idea is to split your data in chunks and to run DP on each chunk separately.

$M(X)$ is ε -DP with the input data X is split into k chunks s.t. $x_1 \cup x_2 \dots \cup x_k = X$. The mechanisms $M(X_1), \dots M(X_k)$ are also ε -DP.

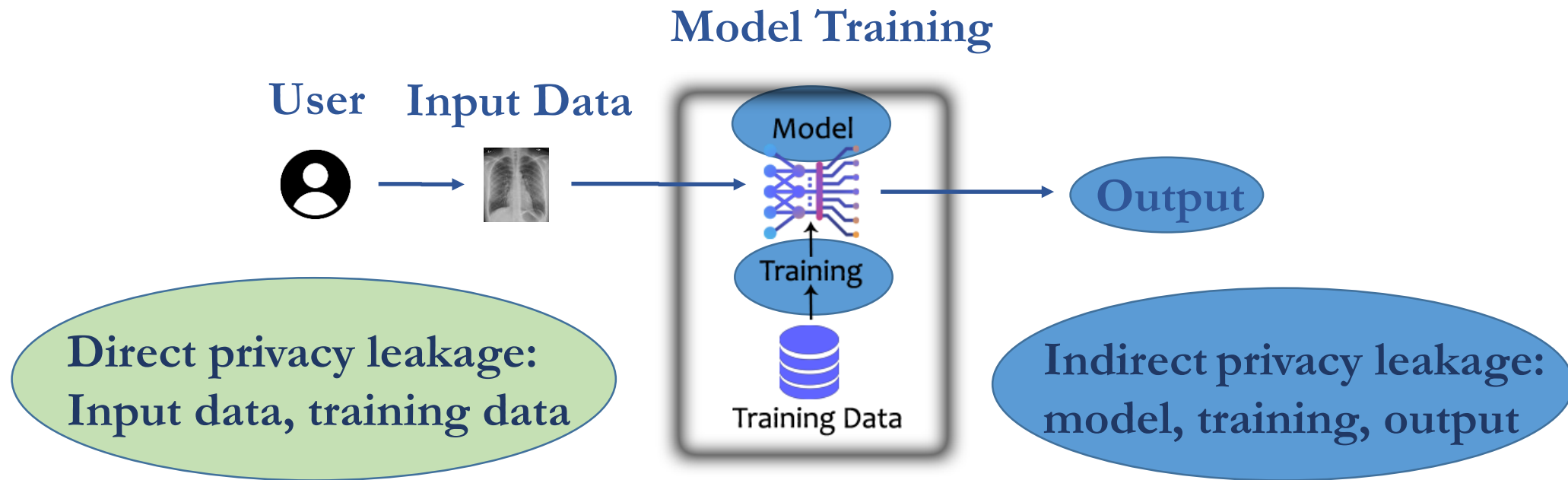


Post-processing Invariance

- Post-processing property of DP holds, that is if an algorithm is (ϵ, δ) -DP then any post-processing is also (ϵ, δ) -DP.
 - It is safe to perform arbitrary computations on a differentially private output. No danger of losing the privacy guarantee.

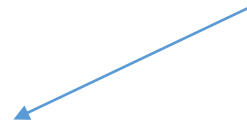


DP meets ML



DP meets ML

- Privacy-Preserving Empirical Risk Minimization (ERM)
 - ERM used to train ML models by minimizing a loss function over a dataset
 - DP can be incorporated into ERM at different stages of the ML life cycle



Local DP could be applied on the training data before the learning process begins.



Global DP could be applied to the final model parameters after training.

- Need a balance between privacy and model performance to guarantee DP with a useful model.

DP meets ML

- DP-SGD: Differentially private Stochastic Gradient Descent⁸

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \dots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N} \sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t , noise scale σ , group size L , gradient norm bound C .

Initialize θ_0 randomly

for $t \in [T]$ **do**

 Take a random sample L_t with sampling probability L/N

Compute gradient

 For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C})$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L} (\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}))$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ϵ, δ) using a privacy accounting method.

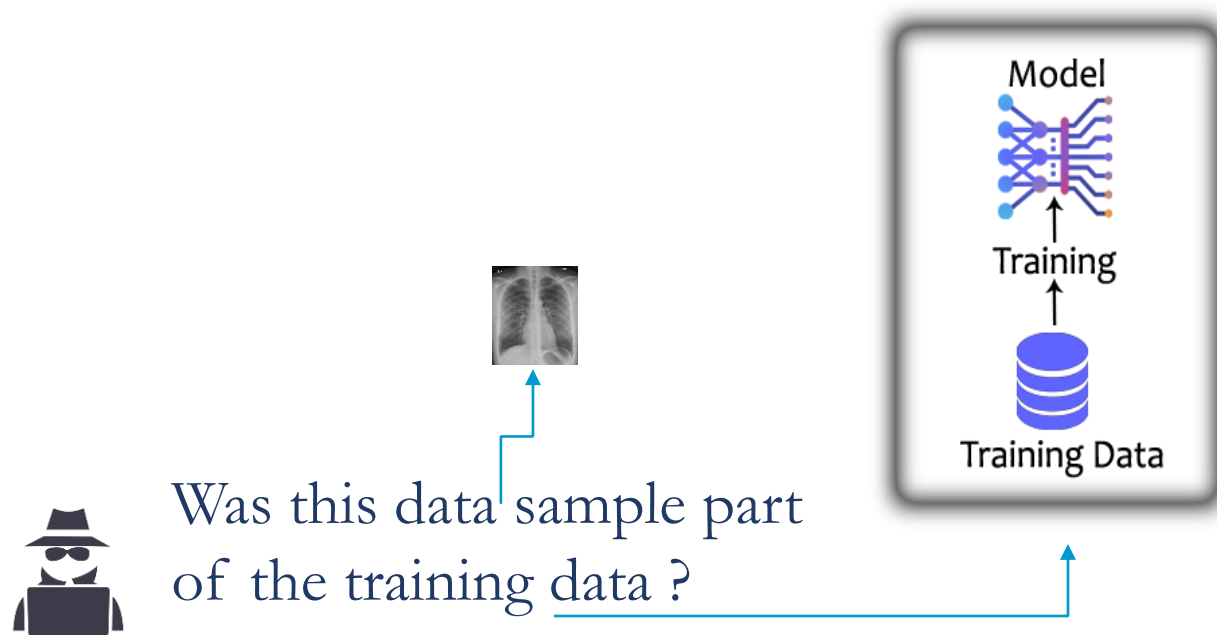
- SGD is an iterative optimization method for unconstrained optimization problems.
- Objective function with suitable smoothness properties

A key step in each private SGD update is gradient clipping that shrinks the gradient of an individual example whenever its l_2 norm exceeds some threshold.

⁸ Abadi et al. Deep Learning with Differential Privacy 2016

DP as a Defense Strategy

- Membership Inference Attacks⁹ (MIAs): indirect leakage from training data

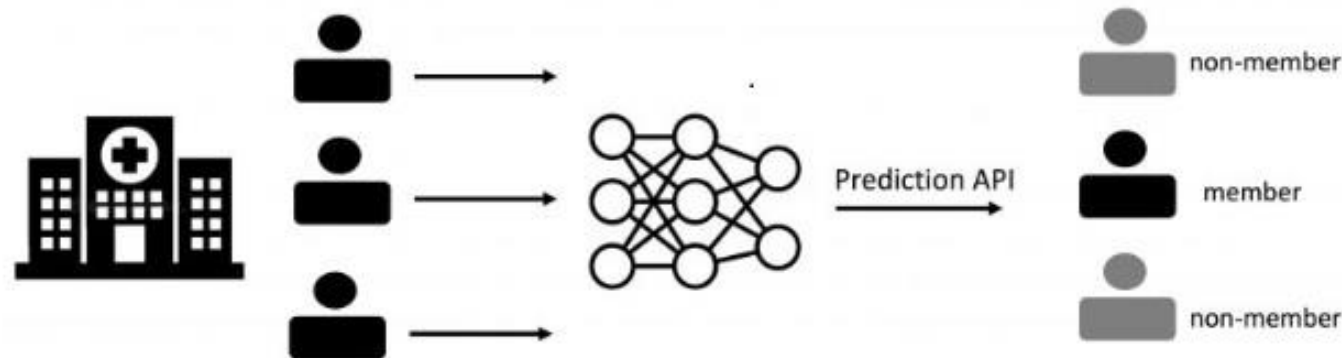


This attack is foundational privacy attack because it gives a signal about if there is some memorization or that model contains some information about the training data

⁹ R. Shokri et al. “Membership Inference Attacks against machine learning models”, 2017.

DP as a Defense Strategy

- Adversary aims to infer whether a given data point was used to train the model
- Does the sensitive training set contain a target data point?



DP as a Defense Strategy

- Membership experiment¹⁰

Experiment 1 (Membership experiment $\text{Exp}^M(\mathcal{A}, A, n, \mathcal{D})$).

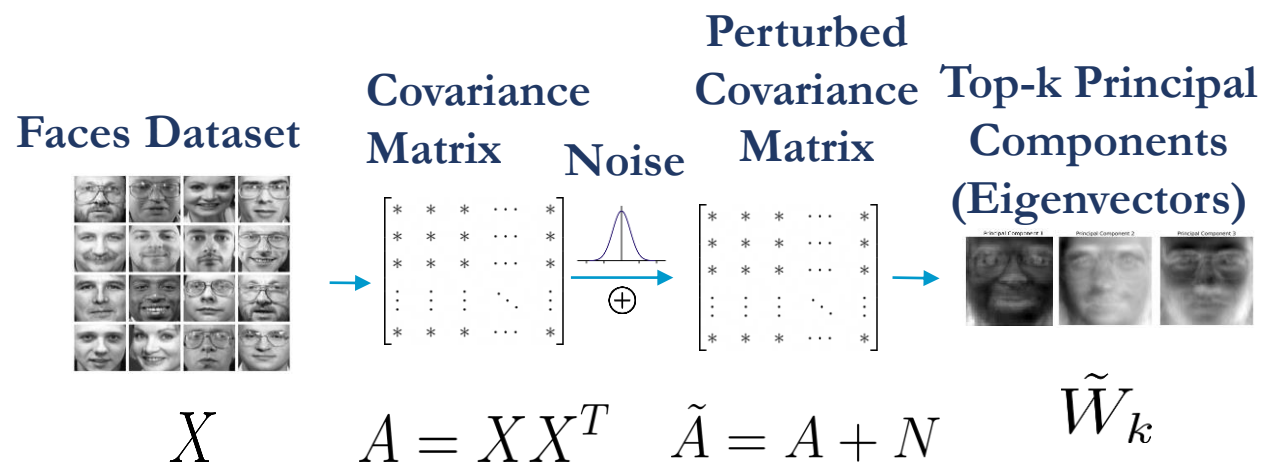
Let \mathcal{A} be an adversary, A be a learning algorithm, n be a positive integer, and \mathcal{D} be a distribution over data points (x, y) . The membership experiment proceeds as follows:

- 1) Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.*
- 2) Choose $b \leftarrow \{0, 1\}$ uniformly at random.*
- 3) Draw $z \sim S$ if $b = 0$, or $z \sim \mathcal{D}$ if $b = 1$*
- 4) $\text{Exp}^M(\mathcal{A}, A, n, \mathcal{D})$ is 1 if $\mathcal{A}(z, A_S, n, \mathcal{D}) = b$ and 0 otherwise. \mathcal{A} must output either 0 or 1.*

¹⁰ Yeom et al. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting 2018

DP as a Defense Strategy

Differentially Private PCA¹¹



Input perturbation before computing the principal components

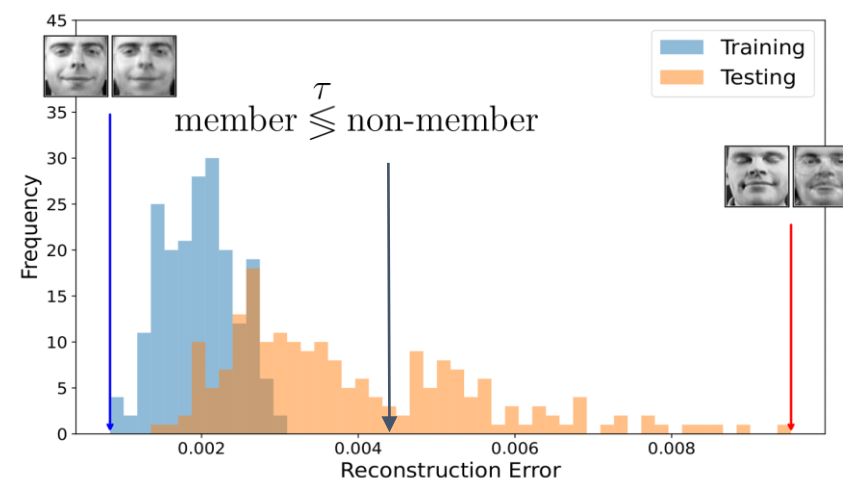


Given \tilde{W}_k , is the sample \mathbf{x}_n part of the faces dataset?

$$E = \|\mathbf{x}_n - \tilde{W}_k \tilde{W}_k^T \mathbf{x}_n\|^2$$



\mathbf{x}_n



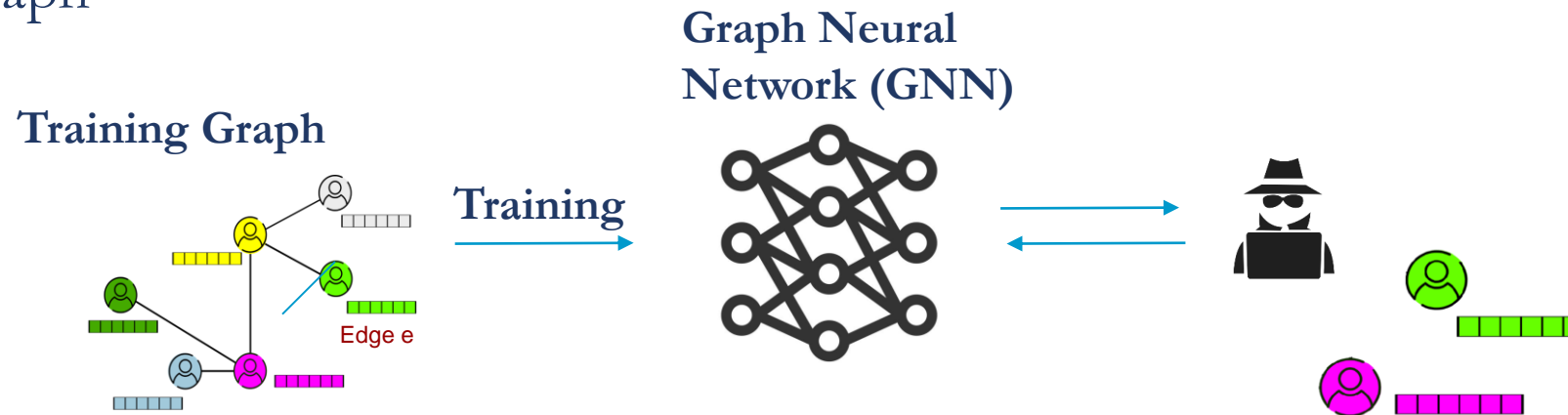
¹¹ Zari et al. Membership Inference Attacks against PCA, 2022

DP as a Defense Strategy

- The adversary computes the reconstruction error of the sample, which is the distance between the original sample and its projection into the eigenvectors
- Then compares this to a threshold
 - the sample was part of the training, if it is smaller than the threshold
 - if large, it was not part of the training data
- Why the reconstruction error is a membership signal: Histogram!
- Samples that were used in PCA training tend to have a smaller reconstruction error than the ones that are not used.
- Explanation: PCA overfits the training data, this is why the attack is successful

DP as a Defense Strategy

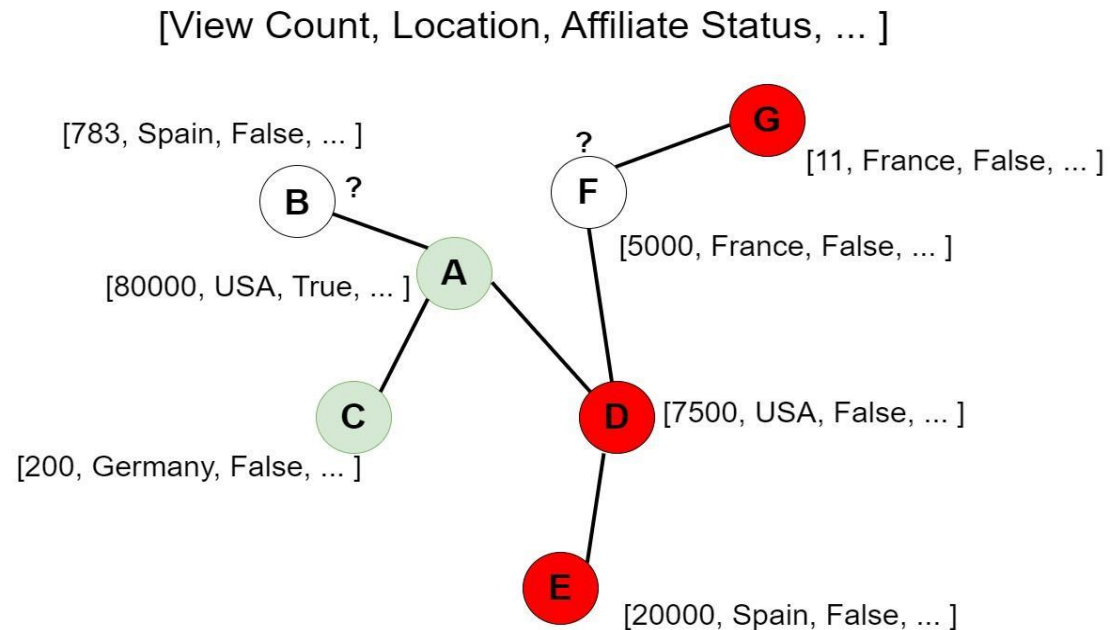
- Link Inference Attacks (LIAs) aim to infer the edges or the links of the training graph



Are these two nodes connected in the graph?

DP as a Defense Strategy

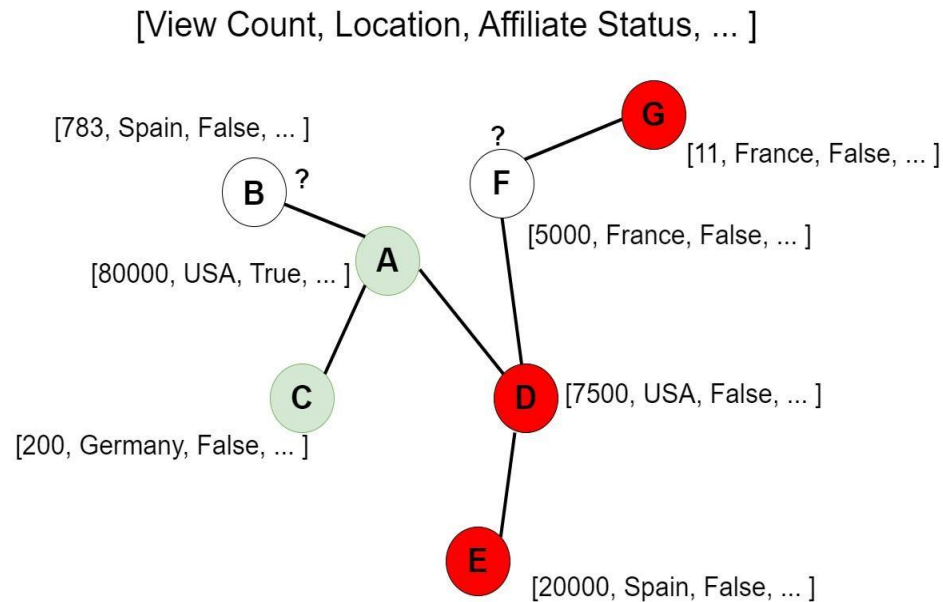
- Link Inference Attacks (LIAs) in Twitch dataset (streaming platform)



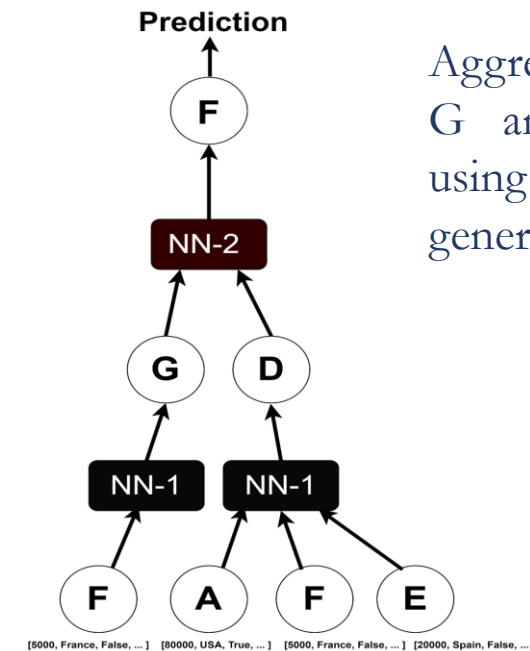
- Each node represent a user
- The edges shows the relation (follow/unfollow) which is private in this case.
- The colors are labels
- G,D,E are positive
- Is F positive?

DP as a Defense Strategy

Aggregated features of G and D instead of using only F's features to generate the prediction.



Computational graph

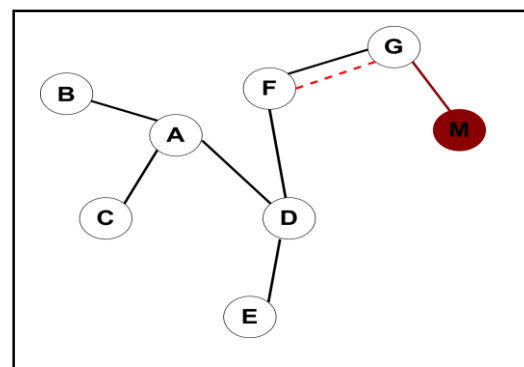
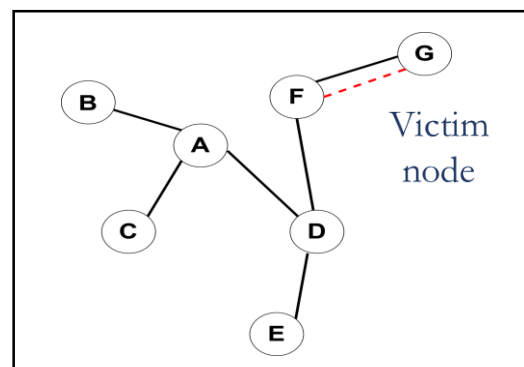


Aggregated features of G and D instead of using only F's features to generate the prediction.

DP as a Defense Strategy

- LIA¹²

Adversary's knowledge:	Adversary's active capability:
Predictions P and P'.	Ability to connect new nodes with target nodes



Query(F)
P

Connect(M, G)
Query(F)
P'

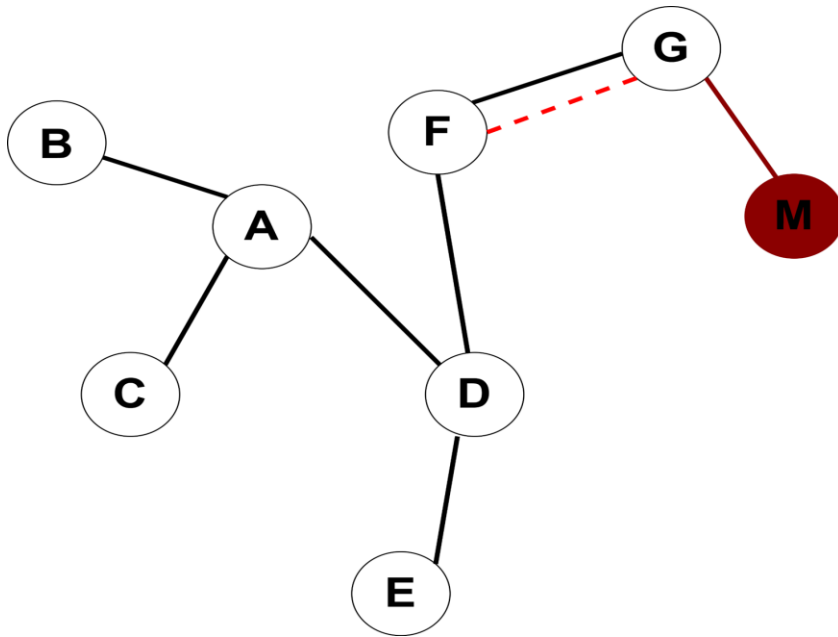


F and the target G are connected if changes of predictions of node F is greater than a threshold

If $P - P' > \tau$
F is Connected to G

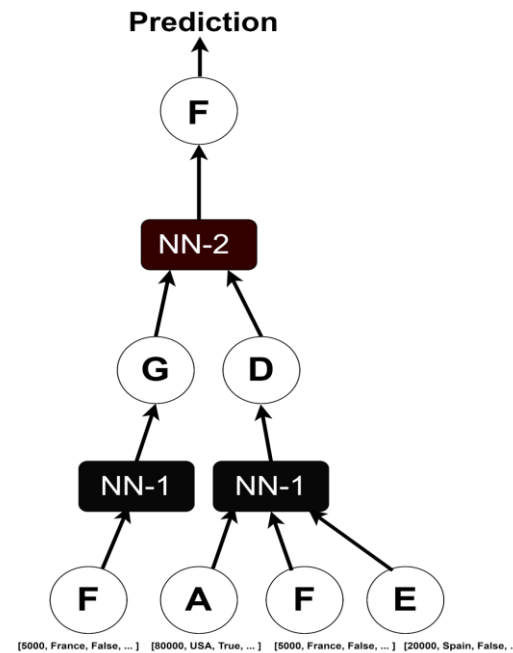
12 Zari et al. Node Injecting Link Stealing Attack, 2024

DP as a Defense Strategy

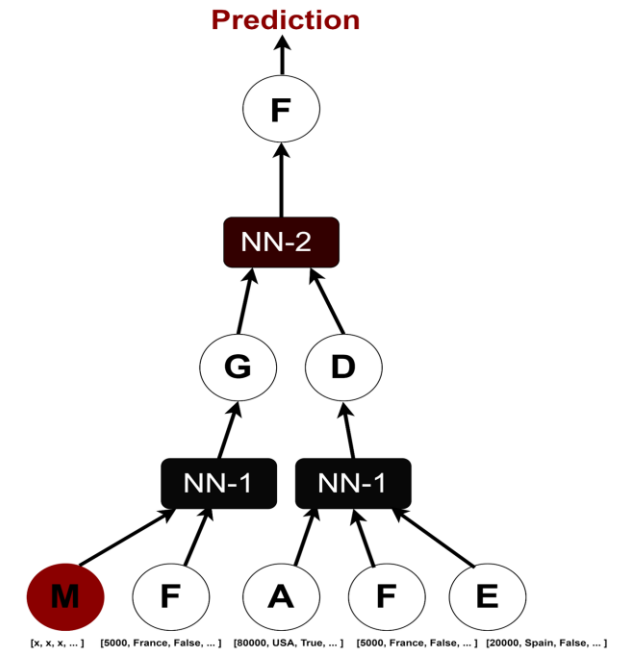


- Injection of M changes the computational graph of the prediction of node F
- The adversary needs to detect the predictions before and after the injection to infer whether F is connected to G

Before Injection



After Injection

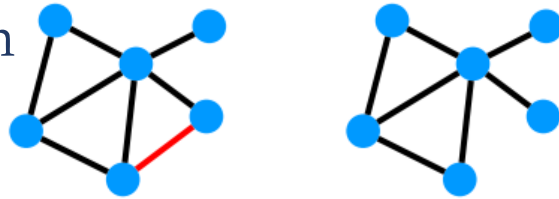


DP as a Defense Strategy

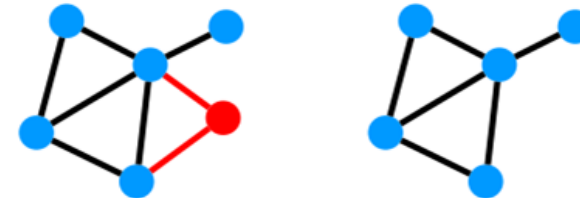
- DP-Adjacency notions for Graphs¹³

Indifferent to
addition or deletion
of a single edge

Edge-level DP



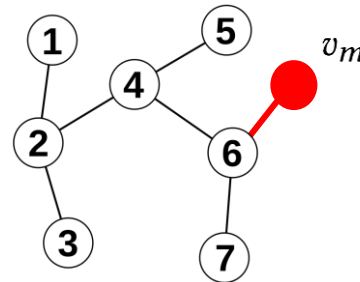
Node-level DP



Higher privacy/
degraded utility

- One-Node-One-Edge (1N1E) Level DP

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$



$$A' = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

A and A' differ
in 1 entry!

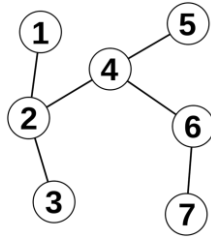
12 Zari et al. Node Injecting Link Stealing Attack, ACSAC 2024

DP as a Defense Strategy

- How LapGraph¹³ works:

Actual number of edges vs estimated number of edges

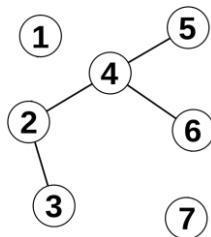
Original Graph



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\hat{A} = \begin{bmatrix} 0 & 0.99 & 0.00 & 0.10 & 0.09 & 0.03 & 0.33 \\ 0 & 0 & 1.01 & 1.01 & 0.01 & 0.10 & 0.01 \\ 0 & 0 & 0 & 0.22 & -0.19 & -0.30 & -0.02 \\ 0 & 0 & 0 & 0 & 1.07 & 1.58 & -0.02 \\ 0 & 0 & 0 & 0 & 0 & -0.03 & -0.09 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.98 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Perturbed Graph



$$\epsilon = 10 \quad |E| = 6 \quad |\hat{E}| = 4$$

$$\hat{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\hat{A} = \begin{bmatrix} 0 & 0.99 & 0.00 & 0.10 & 0.09 & 0.03 & 0.33 \\ 0 & 0 & 1.01 & 1.01 & 0.01 & 0.10 & 0.01 \\ 0 & 0 & 0 & 0.22 & -0.19 & -0.30 & -0.02 \\ 0 & 0 & 0 & 0 & 1.07 & 1.58 & -0.02 \\ 0 & 0 & 0 & 0 & 0 & -0.03 & -0.09 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.98 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

13 Fan Wuet al.. LINKTELLER: Recovering Private Edges from Graph Neural Networks via Influence Analysis.

2022

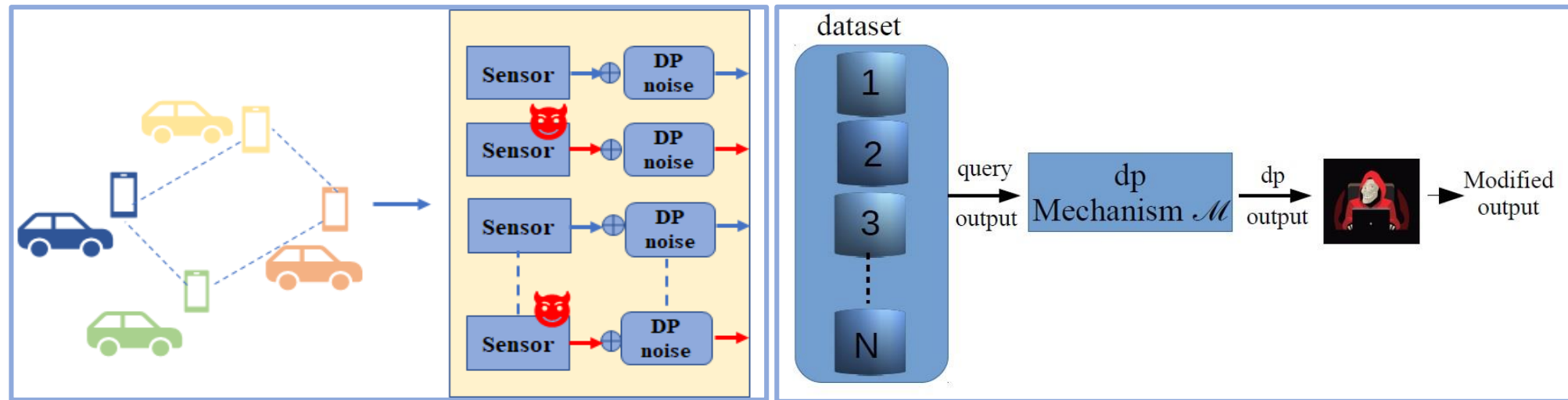
Cyber in Occitanie

A. Ünsal, EURECOM

DP for Adversarial Classification

- Adversarial Classification under DP¹⁴:

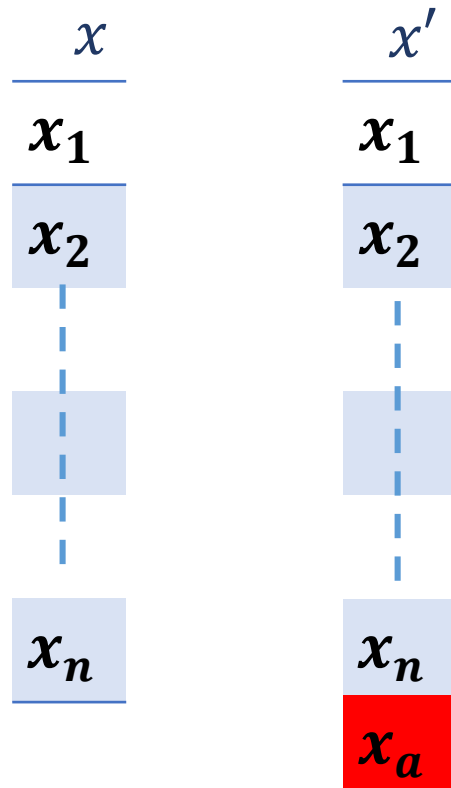
Possible scenario for Local-DP and Global DP



¹⁴ Ünsal et al. A Statistical Threshold for Adversarial Classification in Laplace Mechanisms, 2021

DP for Adversarial Classification

- Adversarial Classification under DP:



- The noisy output $Y = f(X) + N$ where $N \sim \text{Lap}(\frac{s}{\epsilon})$
- Adversary adds X_a

H_0 = Defender fails to detect X_a

H_1 = Defender detects X_a

DP for Adversarial Classification

- The corresponding likelihood ratio

$$\Lambda = \frac{\mathcal{L}(\text{Lap}(\mu_1, b_1); n)}{\mathcal{L}(\text{Lap}(\mu_0, b_0); n)} \underset{H_0}{\overset{H_1}{\geq}} \kappa$$

where κ is some positive number to be determined.

- Probability of false alarm

$$P_{FA} = \alpha = \Pr[H_0 \text{ reject} | \text{Attack}]$$

- Probability of mis-detection

$$P_{MD} = \beta = \Pr[H_1 \text{ reject} | \text{No attack}]$$

DP for Adversarial Classification

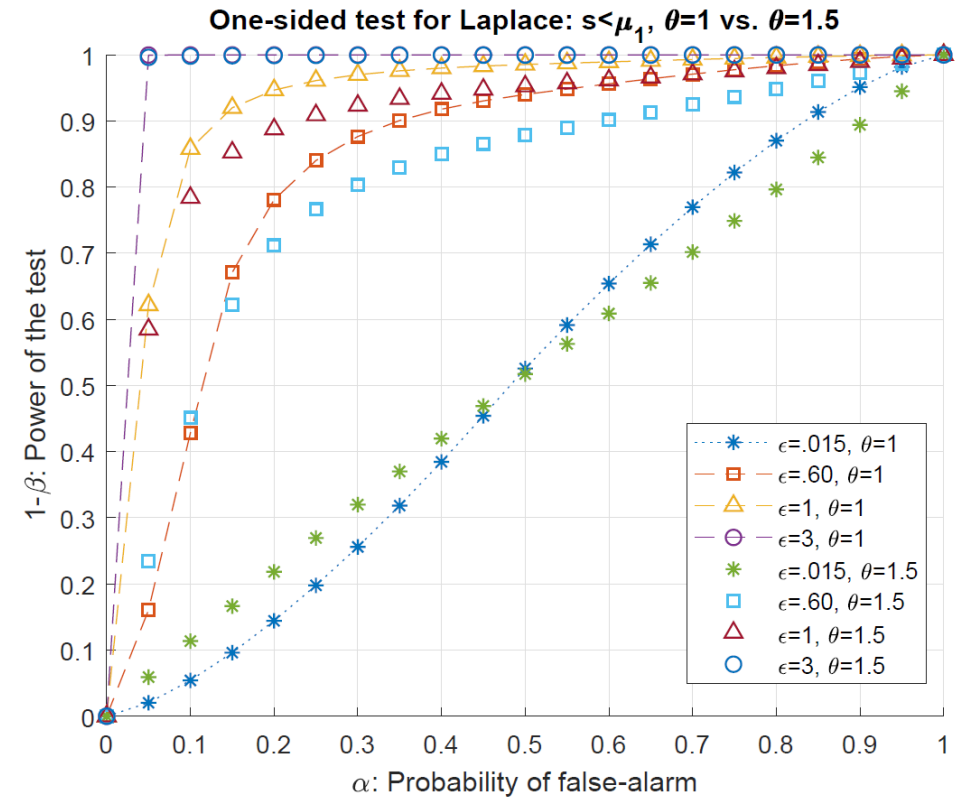
Theorem

The threshold of the best critical region of size α for deciding between H_0 and H_1 for a Laplace mechanism with the largest possible power $\bar{\beta}$ is given as a function of the probability of false alarm, privacy parameter ϵ and global sensitivity s as follows

$$k = \begin{cases} \mu_0 + \frac{s}{\epsilon} \ln(2(1 - \alpha)) & \text{if } \alpha \in [0, .5] \\ \mu_0 - \frac{s}{\epsilon} \ln(2\alpha) & \text{if } \alpha \in [.5, 1] \end{cases}$$

Then, the adversary's hypothesis testing problem for $\mu_1 - \mu_0 > 0$ is

$$Y_0 \underset{H_1}{\overset{H_0}{\leq}} k + f(x) \text{ where } f(.) \text{ denotes the query function}$$



Theoretical Resources

- DifferentialPrivacy.org
- Harvard – Privacy Tools Project

DifferentialPrivacy.org

[Home](#) [About](#) [Posts](#) [Resources](#)

Open Problem: Selection via Low-Sensitivity Queries

Two of the basic tools for building differentially private algorithms are noise addition for answering low-sensitivity queries and the exponential mechanism for selection. Could we do away with the exponential mechanism and simply use low-sensitivity queries to perform selection?

[READ MORE](#)



Courses & Educational Materials

^ Outreach

The Privacy Tools project develops open access course materials and videos on data privacy from a variety of disciplinary perspectives.

Resources

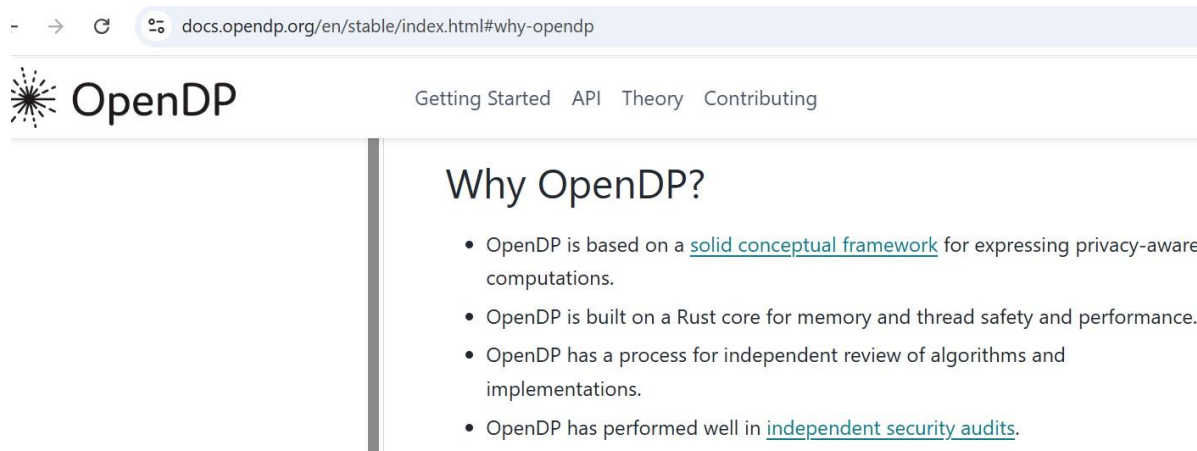
- Privacy Book

The Algorithmic Foundations of Differential Privacy

Cynthia Dwork
Microsoft Research, USA
dwork@microsoft.com

Aaron Roth
University of Pennsylvania, USA
aaroht@cis.upenn.edu

- Open DP

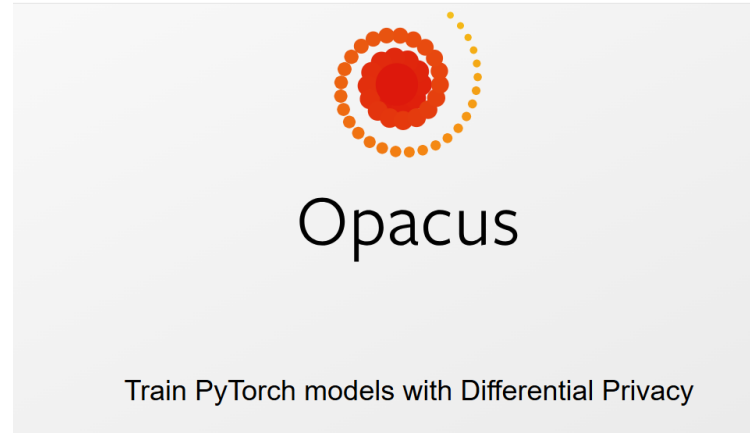


The screenshot shows a web browser with the URL `docs.opendp.org/en/stable/index.html#why-opendp`. The page features the OpenDP logo and navigation links: `Getting Started`, `API`, `Theory`, and `Contributing`. The main heading is `Why OpenDP?`, followed by a bulleted list of key features:

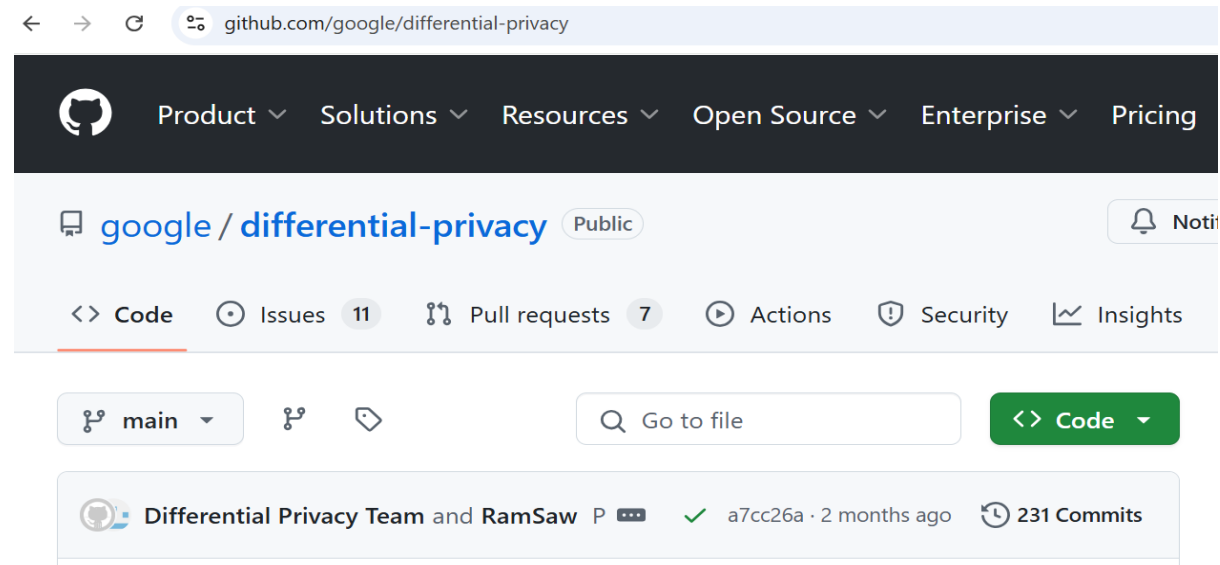
- OpenDP is based on a [solid conceptual framework](#) for expressing privacy-aware computations.
- OpenDP is built on a Rust core for memory and thread safety and performance.
- OpenDP has a process for independent review of algorithms and implementations.
- OpenDP has performed well in [independent security audits](#).

Resources

- PyTorch Opacus

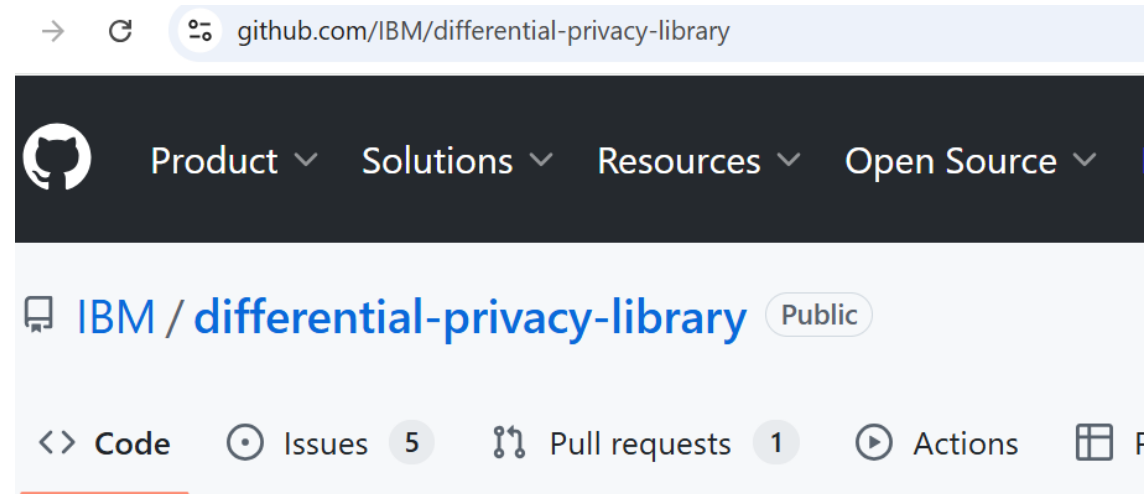


- Google DP Library



Resources

- IBM DP Library



References

- A. Narayanan and V. Shmatikov, “Robust de-anonymization of Large Sparse Datasets”, *IEEE Symposium on Security and Privacy*, 2008.
- L. Sweeney, “k-anonymity: A model for Protecting Privacy”, *Int. J. Uncertainty Fuziness and Knowledge-Based Systems* 10, 2002.
- C. Dwork and Aaron Roth (2014), ”The Algorithmic Foundations of Differential Privacy”, 2014
- Erlingsson et al. “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”, 2014
- C. Dwork, G.N. Rothblum and S. Vadhan, “Boosting and Differential Privacy”, 51st Annual Symposium on Foundations of Computer Science FOCS, pgs. 51-60, 2010.
- P. Kairouz, S. Oh and P. Viswanath, “The Composition Theorem for Differential Privacy”, 32nd International Conference on Machine Learning, 2015.
- M. Abadi, A. Chu, I. Goodfellow, H. Brendan McMahan, I. Mironov, K. Talwar, and L. Zhang. 2016. Deep Learning with Differential Privacy. In Proceedings of the 2016 ACM SIGSAC CCS '16. New York, NY, USA, 308–318.

References

- S. Yeom, I. Giacomelli, M. Fredrikson and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," in 2018 IEEE 31st Computer Security Foundations Symposium (CSF)
- O. Zari, J. Parra-Arnau, A. Ünsal, T. Strufe and M. Önen, "Membership Inference Attack against Principal Component Analysis, PSD 2022
- O. Zari, J. Parra-Arnau, A. Ünsal and M. Önen "Node Injection Link Stealing Attack". Privacy in Statistical Databases. PSD 2024. Lecture Notes in Computer Science, vol 14915. Springer, Cham.
- Fan Wu, Yunhui Long, Ce Zhang, and Bo Li. 2021. LINKTELLER: Recovering Private Edges from Graph Neural Networks via Influence Analysis. 2022 IEEE Symposium on Security and Privacy (SP) (2021), 2005–2024.
- A. Ünsal and M. Önen, "A Statistical Threshold for Adversarial Classification in Laplace Mechanisms", IEEE Information Theory Workshop, Oct. 2021
- A. Ünsal and M. Önen, "Chernoff Information as a Privacy Constraint for Adversarial Classification," 2024 60th Annual Allerton Conference, Urbana, IL, USA, 2024, pp. 1-8
- A. Ünsal and M. Önen, "Information-Theoretic Approaches to Differential Privacy", ACM Computing Surveys, 2023

Thank you!

Any questions/Comments?



unsal@eurecom.fr