

Differential Privacy: Foundation, Applications, and Challenges

Anh-Tu Hoang

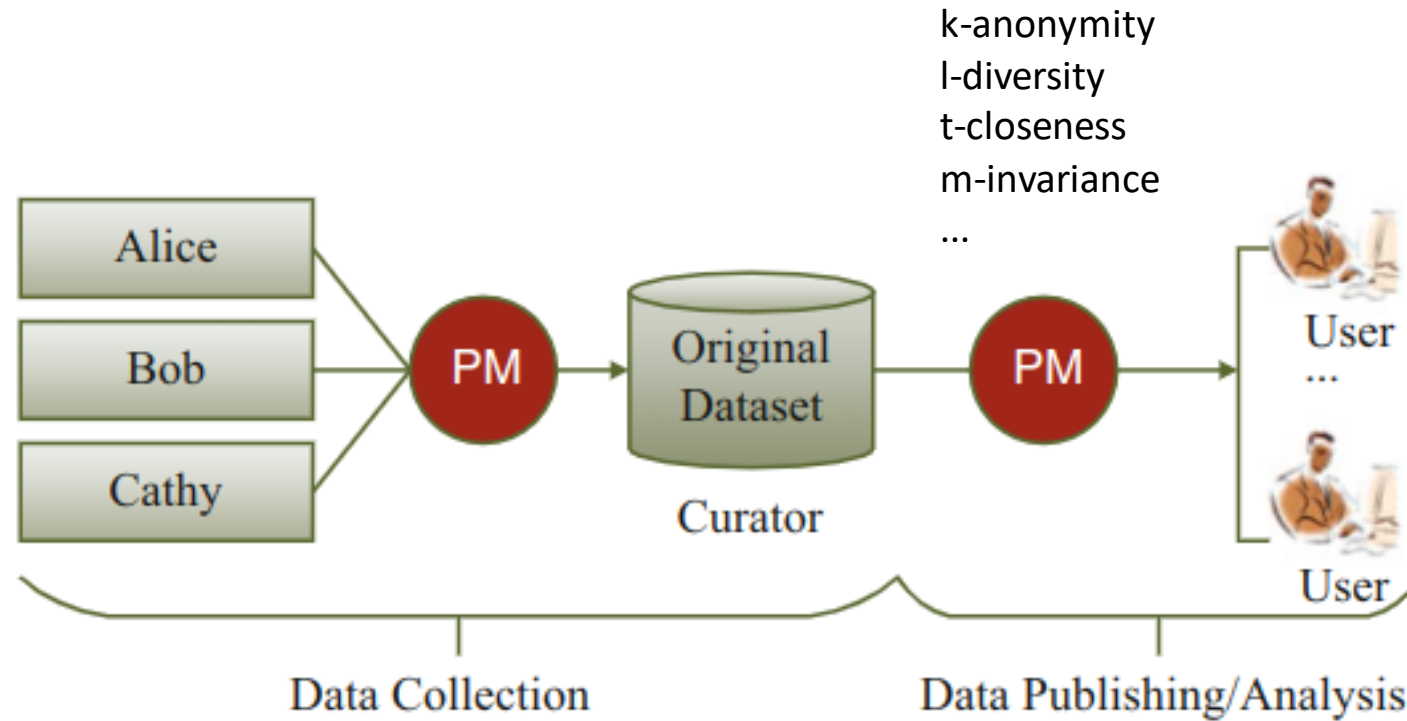
anhtu.hoang@uninsubria.it

DiSTA, University of Insubria, Italy

Agenda

- ❖ Introduction: definitions and basic concepts
- ❖ Applications
 - Interactive: Transaction/Graph Data Publishing
 - Non-Interactive: Batch Queries/ Anonymized Dataset Publishing
 - Differentially Private Data Analysis
 - Differentially Private Deep Learning
- ❖ Where to start?
- ❖ Conclusion

Privacy Preserving Data Publishing and Analysis

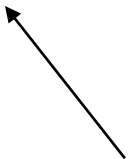


Privacy Model (PM)[1]

k-anonymity

k-anonymity: every user must have at least $k-1$ other users whose **anonymized data** are identical to his/hers.

assumed by the model
designer

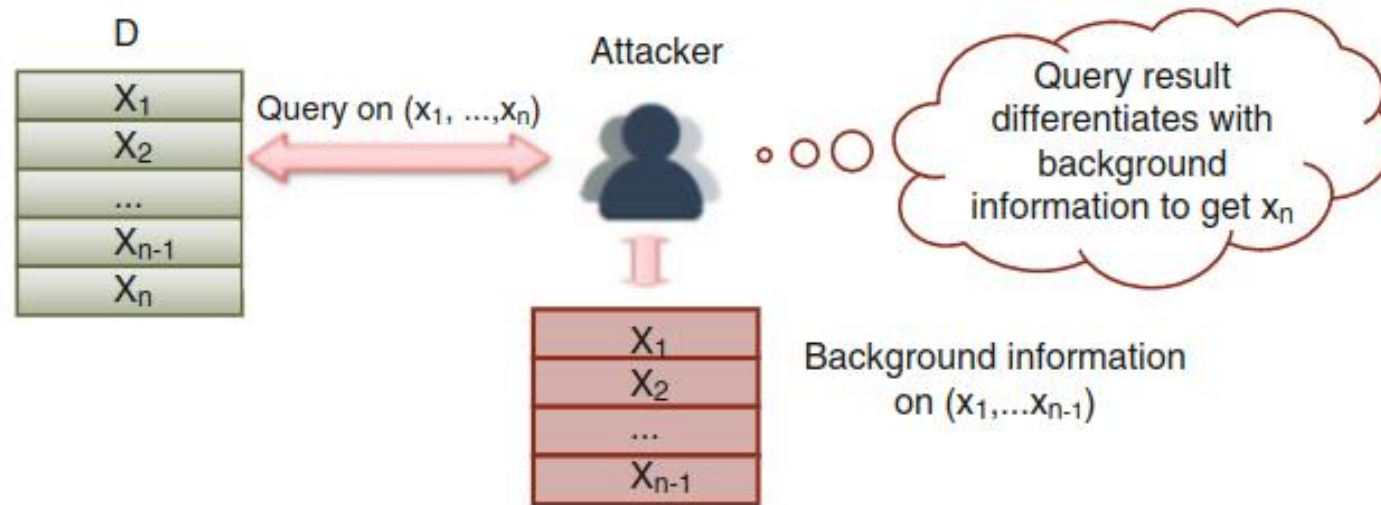


Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

3-anonymity table with quasi-identifier: {Job, Sex, Age}[2]

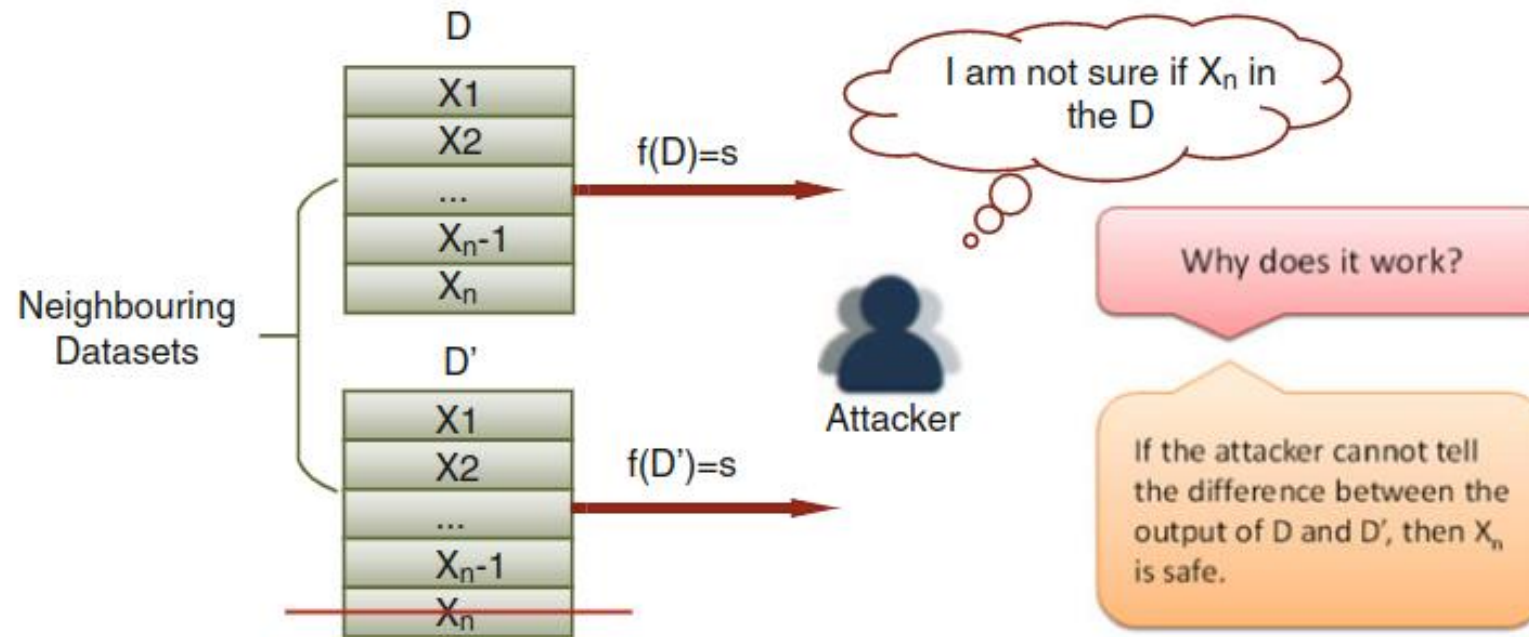
**It is hard to assume
adversary background
knowledge in practice**

Differential Privacy



Attacker Model [1]

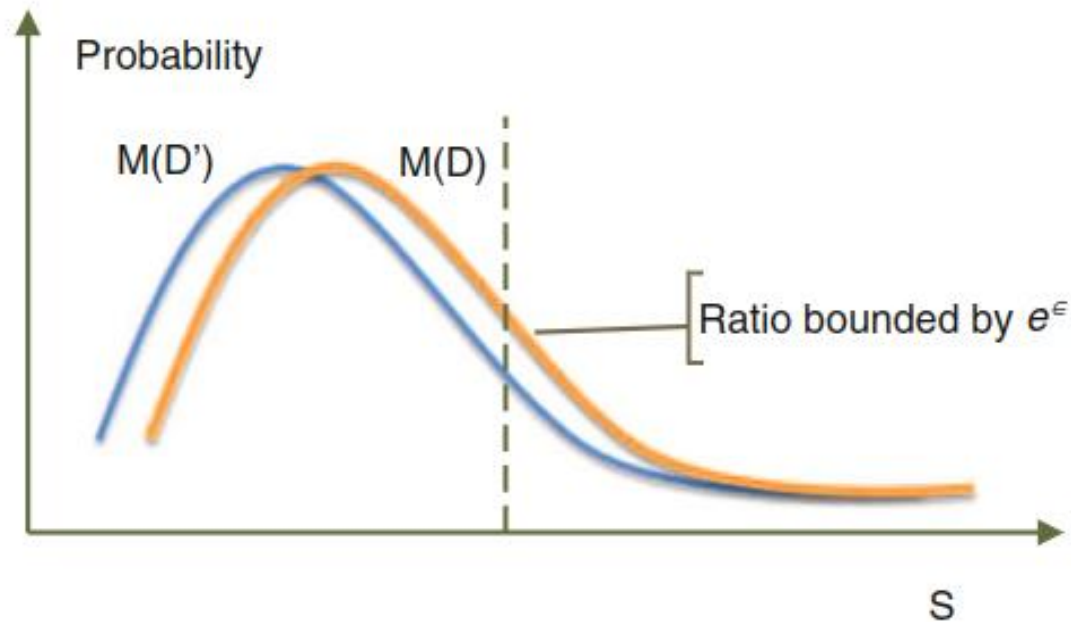
Differential Privacy (2)



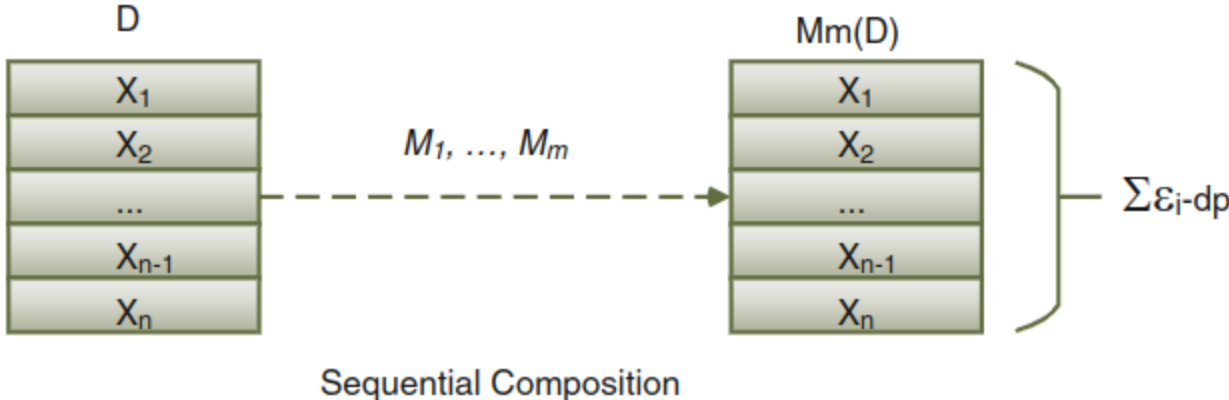
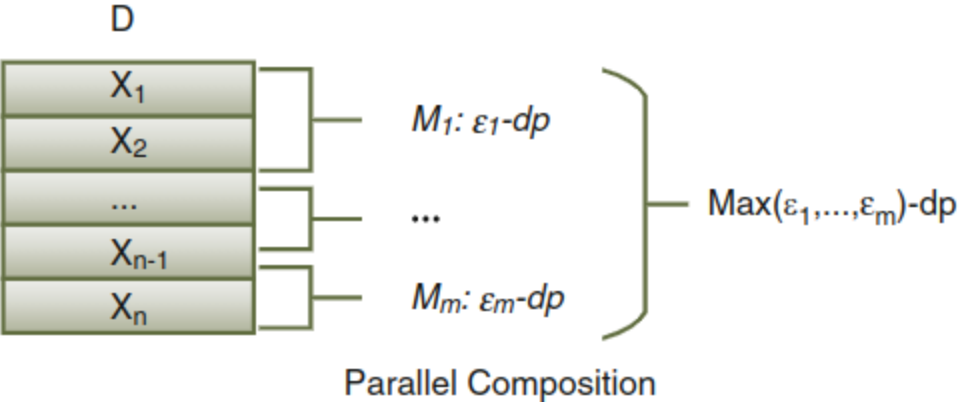
Differential Privacy Definition

❖ Definition ((ϵ , δ)-Differential Privacy [1]) A randomized mechanism M gives (ϵ , δ)-differential privacy for every set of outputs S , and for any neighbouring datasets of D and D' , if M satisfies:

$$\Pr[M(D) \in S] \leq \exp(\epsilon) \cdot \Pr[M(D') \in S] + \delta.$$



Compositions



Privacy budget composition[1]

Global Sensitivity[1]

Definition 2.2 (Global Sensitivity) For a query $f : D \rightarrow \mathbb{R}$, the *global sensitivity* of f is defined as

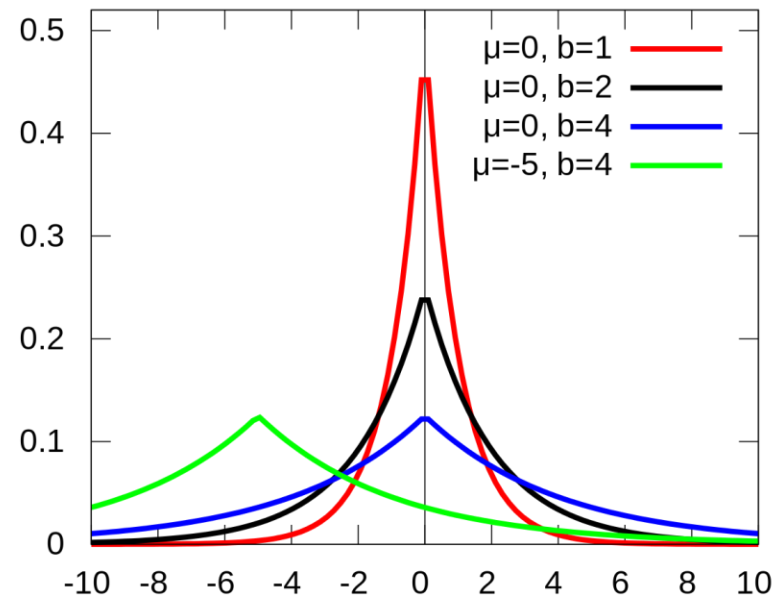
$$\Delta f_{GS} = \max_{D, D'} \|f(D) - f(D')\|_1. \quad (2.2)$$

Laplace Mechanism (numeric queries)

Definition 2.5 (Laplace Mechanism) Given a function $f : D \rightarrow \mathbb{R}$ over a dataset D , mechanism M provides the ϵ -differential privacy if it follows Eq. (2.5)

$$M(D) = f(D) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right). \quad (2.7)$$

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



Exponential Mechanism (non-numeric queries)

Definition 2.8 (Exponential Mechanism) Let $q(D, \phi)$ be a score function of dataset D that measures the quality of output $\phi \in \Phi$. Then an Exponential mechanism M is ϵ -differential privacy if

$$M(D) = \left\{ \text{return } \phi \text{ with the probability } \propto \exp\left(\frac{\epsilon q(D, \phi)}{2\Delta q}\right) \right\}. \quad (2.10)$$

where Δq represents the *sensitivity* of score function q .

Example

Name	Job	Gender	Age	Disease
Alen	Engineer	Male	25	Flu
Bob	Engineer	Male	29	HIV
Cathy	Lawyer	Female	35	Hepatitis
David	Writer	Male	41	HIV
Emily	Writer	Female	56	Diabetes
...
Emma	Dancer	Female	21	Flu

Medical Record [1]

f1: how many people in this table have HIV?

Mechanism: Laplace

Sensitivity: $\Delta_{f1} = 1$

Privacy Budget ϵ : 1.0

Output Generation: $M(D) = f1(D) + \text{Lap}(\Delta f1 / \epsilon) = f1(D) + \text{Lap}(1)$

f2: what is the most common disease?

Mechanism: Exponential

Score function q : the number of people on each disease

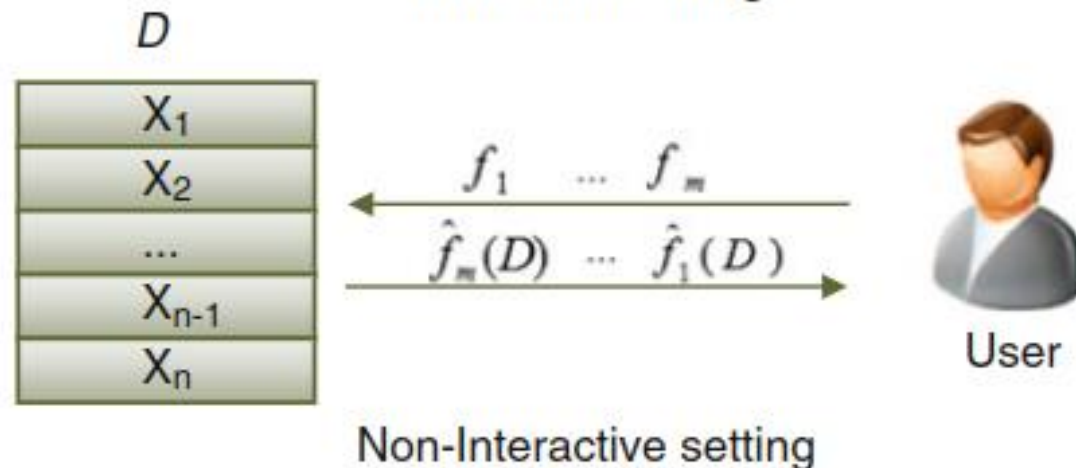
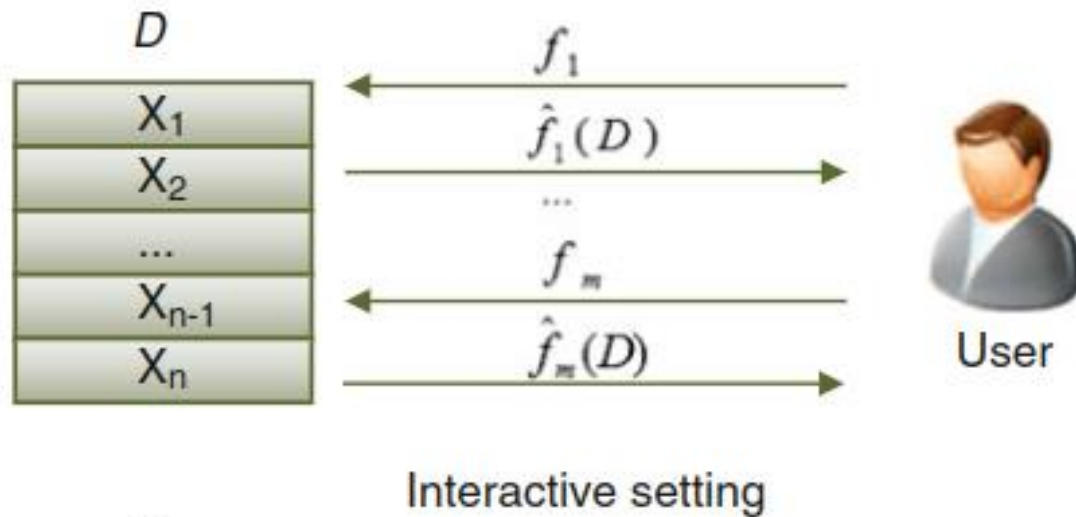
Sensitivity: $\Delta q = 1$

Output Generation: randomly choose based on probabilities of outputs

Table 2.3 Medical record exponential mechanism output

Options	Number of people	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 1$
Diabetes	24	0.25	0.32	0.12
Hepatitis	8	0.25	0.15	4×10^{-5}
Flu	28	0.25	0.40	0.88
HIV	5	0.25	0.13	8.9×10^{-6}

Publishing Settings



f_1 : How many patients have diabetes at the age of 40–79?
 f_2 : How many patients have diabetes at the age of 40–59?

Noises: $\text{Lap}(1/\epsilon) + \text{Lap}(2/\epsilon)$

If the system needs to handle a lot of queries, there are too much noises

Noises: $2 * \text{Lap}(2/\epsilon)$

Publishing Mechanisms

- ❖ Transformation: transforms original dataset to a new structure. The sensitivity of the query set will be adjusted.
- ❖ Dataset Partitioning: divides the dataset into several parts and adds noise to each part separately.
- ❖ Query Separation: add noise to small numbers of queries.

Transaction Data Publishing (Interactive)

Name	Age	Has diabetes?
Alice	35	1
Bob	50	1
Cathy	12	0
...
Eva	35	0

Laplace[3]: adds noises to real-values queries.

Query Separation[1]: given m queries, there are $O(\log m \log |X|)$ queries that can determine the answers of all other queries, where X is the domain of the queried dataset D .

Graph Data Publishing (Interactive)

Publish graphs' statistics while protecting nodes or edges

- Subgraph counts of k-star and triangle [1]
- The number of edges [1]

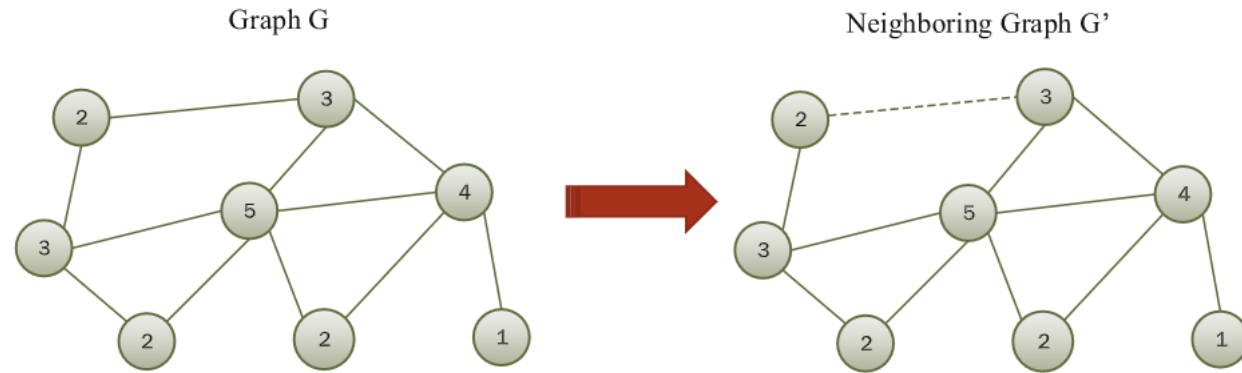


Fig. 4.4 Edge differential privacy

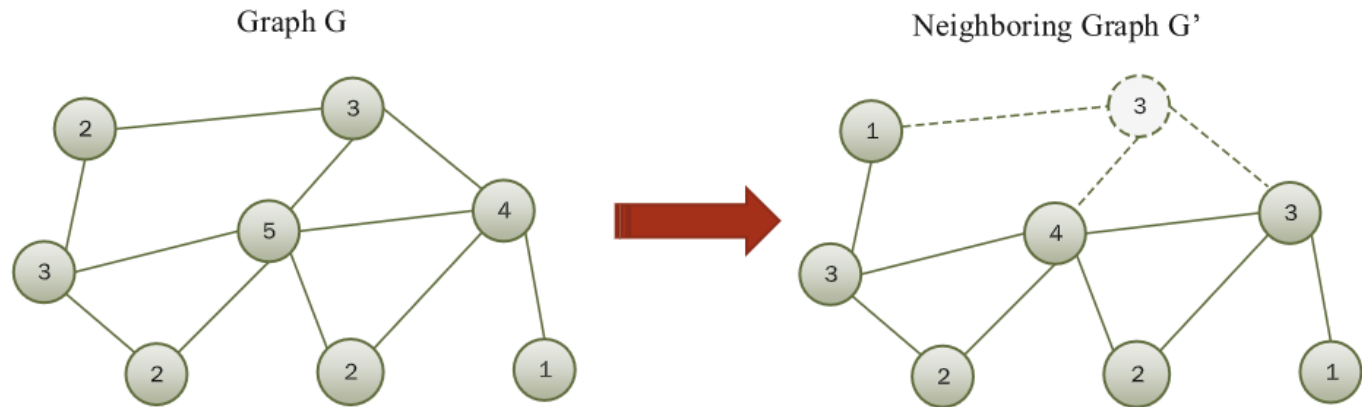


Fig. 4.5 Node differential privacy

Batch Queries Publishing (Non-Interactive)

Given m queries, how to minimize added noise?

Grade	Count	Variable
90–100	12	x_1
80–89	24	x_2
70–79	6	x_3
60–69	7	x_4

Frequency

Range query							
f_1	x_1	+	x_2	+	x_3	+	x_4
f_2	x_1	+	x_2	+	x_3		
f_3		+	x_2	+	x_3	+	x_4
f_4	x_1	+	x_2				
f_5		+	x_2	+	x_3		
f_6					x_3	+	x_4
f_7	x_1						
f_8			x_2				
f_9					x_3		
f_{10}							x_4

10 range queries

Laplace: add noises to frequency table or query results but both approaches add too much noises.

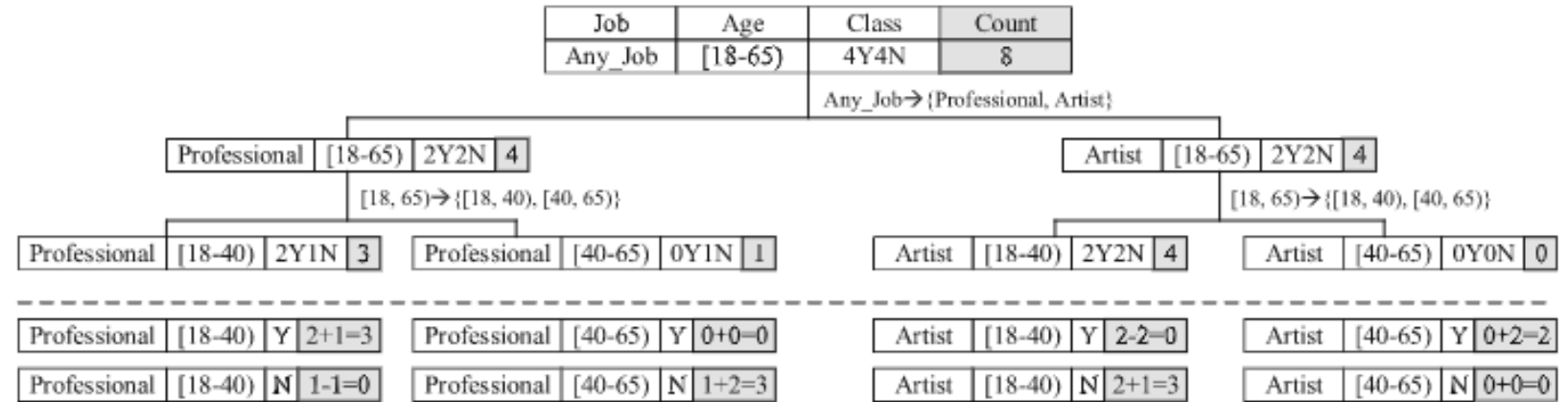
Partition of Dataset: breaks the correlations between columns to reduce noise[1]

- decompose datasets' columns into disjoint groups;
- average the counts in each group;
- add Laplace noise to each group's average count;
- the noised count is the new count of each group

Anonymized Dataset Publishing (Non-Interactive)

Job	Age	Class
Engineer	34	Y
Lawyer	50	N
Engineer	38	N
Lawyer	33	Y
Dancer	20	Y
Writer	37	N
Writer	32	Y
Dancer	25	N

dataset



- Use exponential mechanism to select candidate to be splitted in each step.
- Group records by their attributes' taxonomy trees.
- Add noise to each count.

Differentially Private Data Analysis

❖ Laplace/Exponential Framework: automatically adds noises to non-private analysis algorithms (SuLQ, PINQ)

Fig. 6.1 SuLQ interface

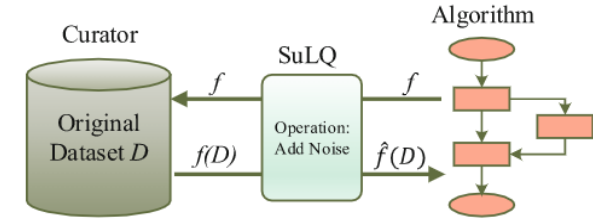
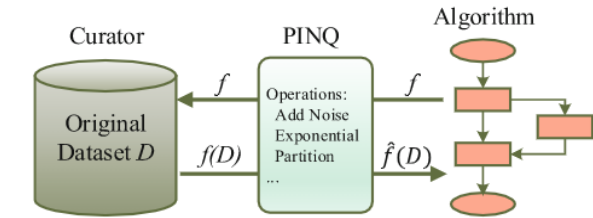


Fig. 6.2 PINQ interface



❖ Private Learning Framework: adds noise from Gamma distribution to the learned models' weights.

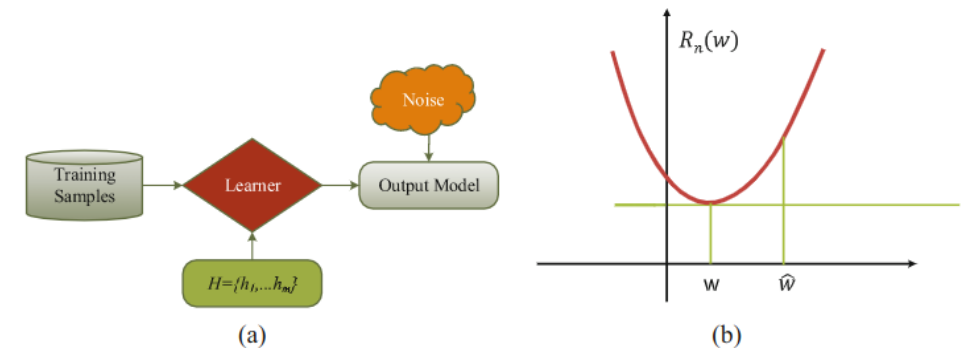


Fig. 6.4 (a) Output perturbation diagram. (b) Output perturbation

Differentially Private Deep Learning

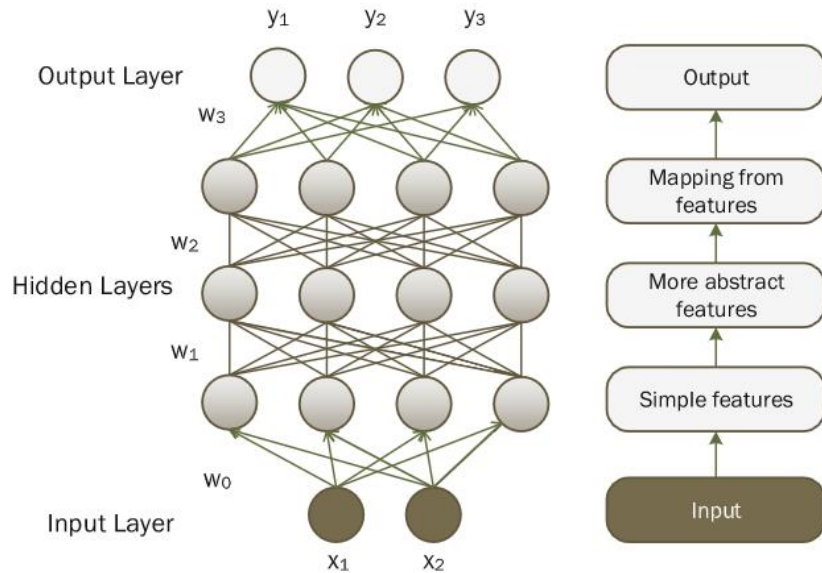
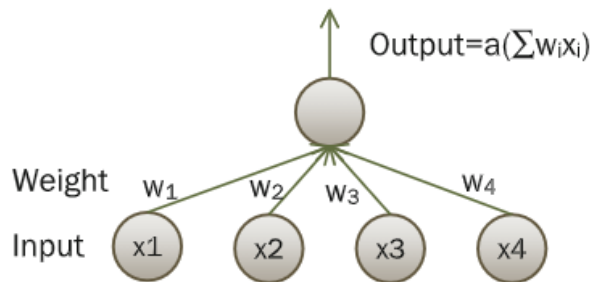


Fig. 7.1 Deep learning structure



Algorithm 1 SGD Algorithm

Require: Training dataset $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, loss function $J(\mathbf{w}, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n J(\mathbf{w}, x_i)$, learning rate α .

Ensure: \mathbf{w} .

- 1: Initialize \mathbf{w}_0 randomly;
- 2: **repeat**
- 3: Randomly take samples S_t from the training dataset \mathbf{x} ;
- 4: **for** each $i \in S_t$ **do**
- 5: compute $g_t(x_i) \leftarrow \nabla_{\mathbf{w}_t} J(\mathbf{w}_t, x_i)$; { Compute gradient }
- 6: **end for**
- 7: compute $g_t \leftarrow \frac{1}{|S_t|} \sum_{i \in S_t} g_t(x_i)$;
- 8: update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \alpha g_t$; { Descent }
- 9: **until** an approximate minimum is obtained.
- 10: **return** \mathbf{w} .

Basic Laplace Differentially Private Deep Learning

Algorithm 2 Basic Laplace Method

Require: $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, $J(\mathbf{w}, \mathbf{x})$, learning rate α , noise scale σ , batch size S , privacy budget ϵ .

Ensure: w .

1. initialize \mathbf{w}_0 randomly;

2. $\epsilon_t = \epsilon/T$;

for $t = 0, \dots, T - 1$ **do**

 3. take a random sample set S_t from D ;

 4. compute gradient: for each $i \in S_t$, compute $g_t(x_i) = \nabla_{\mathbf{w}_t} J(\mathbf{w}_t, x_i)$;

 5. add noise: $\hat{g}_t = \frac{1}{S}(\sum_i g_t(x_i) + Lap(\Delta J/\epsilon_t))$;

 6. descent: $\hat{\mathbf{w}}_{t+1} = \hat{\mathbf{w}}_t - \alpha_t \hat{g}_t$;

end for

7. $\hat{w} = \hat{\mathbf{w}}_T$.

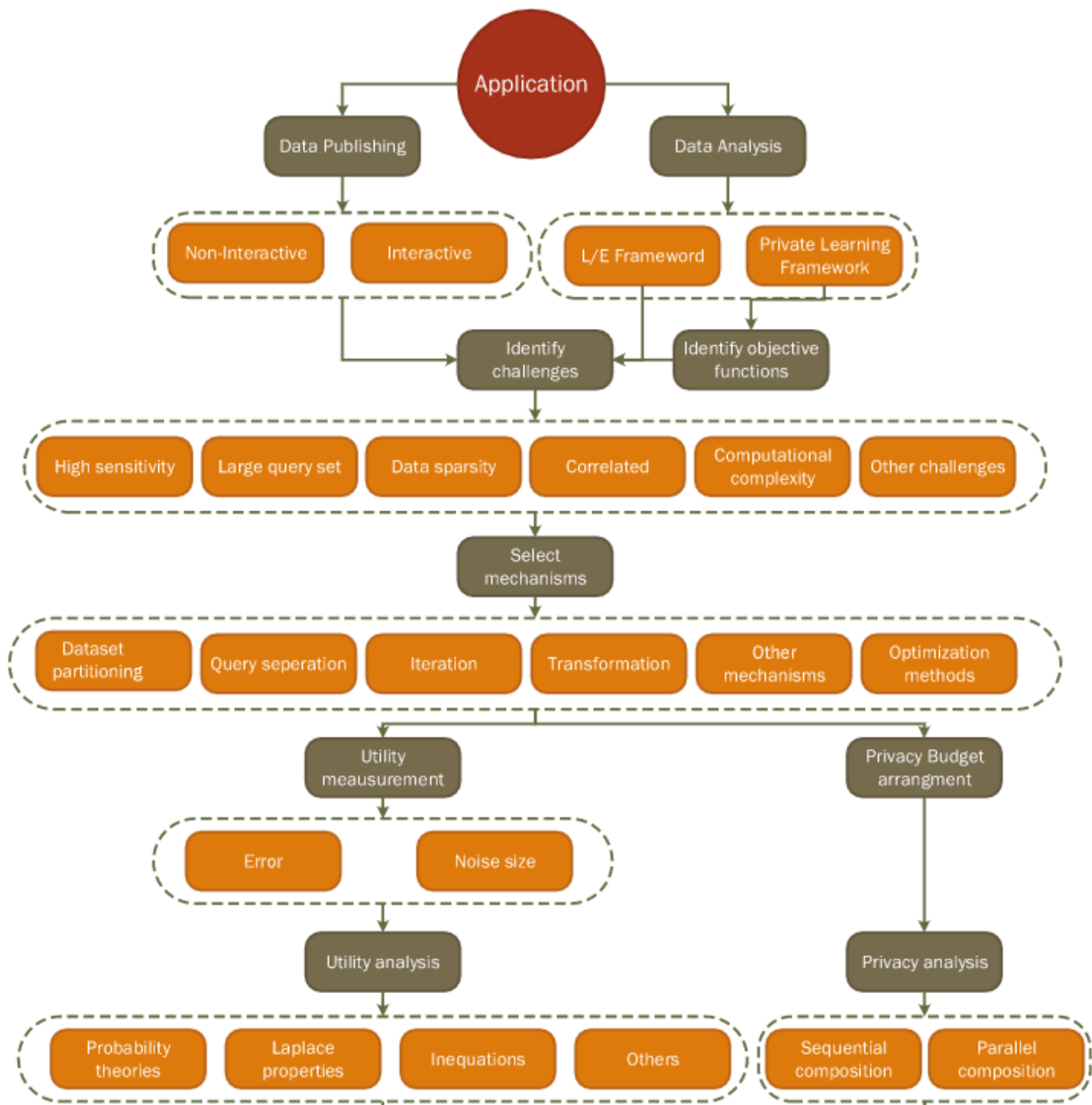
Too much added noises:

- objective function has high sensitivity,
- high number of training steps

Differentially Private Deep Learning Approaches

- ❖ Norm clipping the objective function to reduce its sensitivity.
- ❖ Group batches together and add noises to the group.

Where to start?



Conclusion & Future Work

❖ Future work:

- Personalized Privacy
- Secure Multiparty Computations with Differential Privacy
- Differential Privacy in Genetic Data
- Local Differential Privacy
- Learning Model Publishing

References

[1] Zhu T, Li G, Zhou W, Yu PS. Differential Privacy and Applications. Springer, Cham; 2017.

[2] Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey of recent developments. ACM Comput Surv. 2010;42: 1–53.

[3] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. Found Trends Theor Comput Sci. 2014;9: 211–407.

[4] Near JP, Abua C. Programming Differential Privacy. Available: <https://programming-dp.com/book.pdf>

Thank you for your attention