Membership Inference Attack and Differential Privacy

A Way of Attack and A Way of Defence of Computer Security

Presentation of A Brief Introduction

Changlong Ji
changlong.ji@telecom-sudparis.eu
2023.Jan.19

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

I. BACKGROUND

II. ATTACK CLASSIFICATION

1) ALL CATEGORIES
2) MEMBERSHIP INFERENCE ATTACK

III. DEFENSE CLASSIFICATION

1) ALL CATEGORIES
2) DIFFERENTIAL PRIVACY

IV. CURRENT RESEARCH DIRECTION

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

I. BACKGROUND

General perception of privacy-sensitive data

Identity

Address

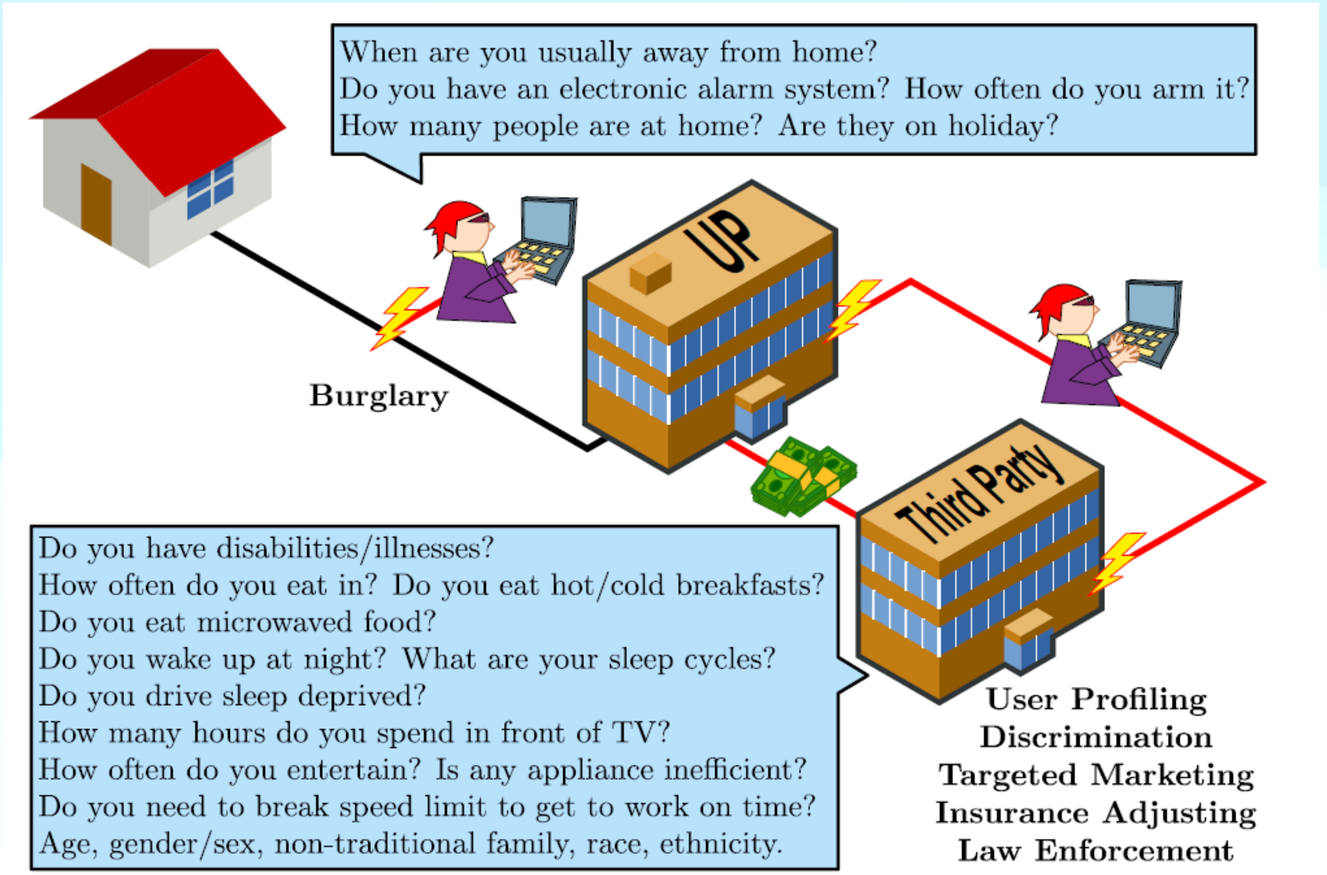
Health record

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

I. BACKGROUND

However…
No real separation between your data and your identity



TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

I. BACKGROUND

Privacy regulations everywhere

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

II. ATTACK CLASSIFICATION – ALL CATEGORIES

1. Membership Inference Attack (MIA): used to determine whether a sample is a member of the training set.

2. Model Inversion Attack (MIA): restore the model input based on the output of the model.

3. Property Inference Attack (PIA): Infer the properties of a class or properties on the entire data set.

4. Model Extraction Attack (MEA): Stealing model weights or hyperparameters.

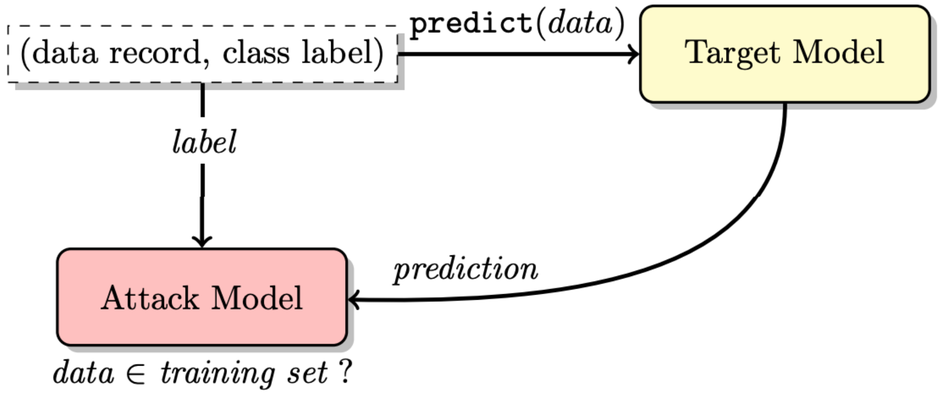
5. Functionality Extraction Attack (FEA): Through the input and output pairs of the target model, to imitate a model with similar functions.

II. ATTACK CLASSIFICATION – MEMBERSHIP INFERENCE ATTACK - BACKGROUND

Now that machine learning is more and more applied in our real life, almost all of our private data may be used in the training of machine learning models. If member reasoning attacks work, then our privacy disclosure will be unprecedented. . Think about it, when your medical records in the hospital are obtained by an attacker, the attacker may use your medical records to speculate whether you will suffer from certain diseases. And then use this to promote related products or even defraud you, which are serious leaks for our privacy.

II. ATTACK CLASSIFICATION – MEMBERSHIP INFERENCE ATTACK - DEFINITION

Define attack model $f_{attack}()$, Its input x_{attack} is a vector consisting of the correct label class and a prediction confidence vector of the target model (the attacked model) . The output of the attack model is a prediction class "in" (member) or "out" (non-member).



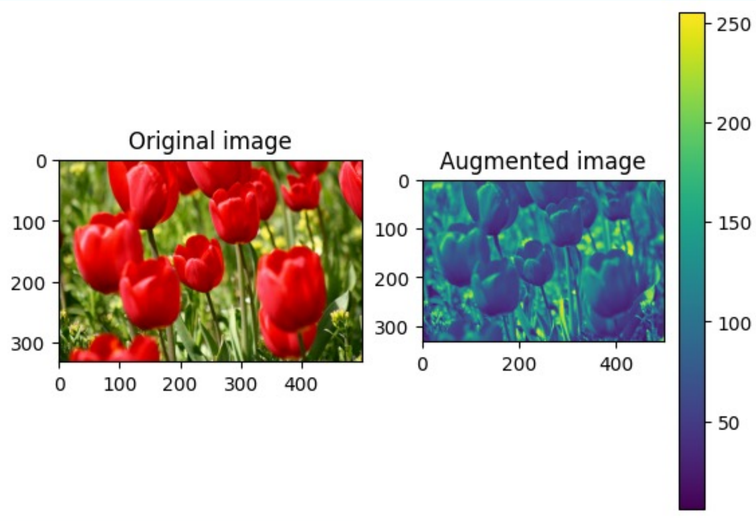
TELECOM SudParis

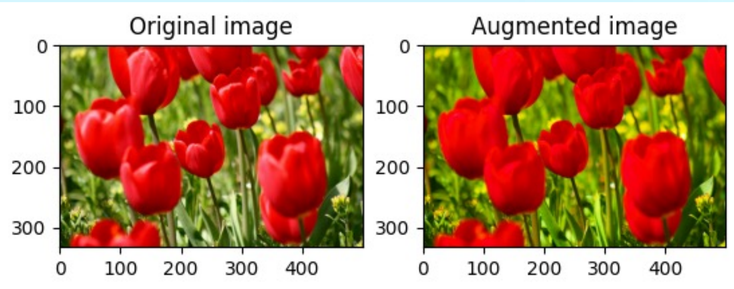
INSTITUT POLYTECHNIQUE DE PARIS

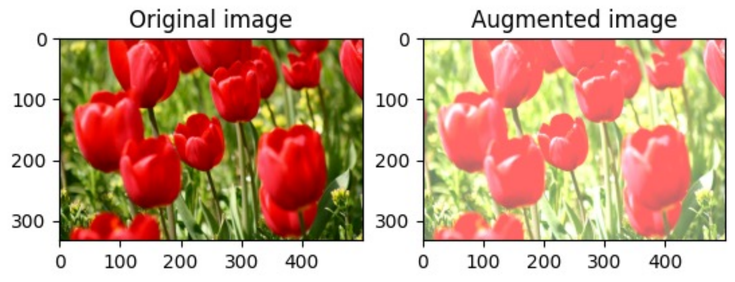
II. ATTACK CLASSIFICATION – MEMBERSHIP INFERENCE ATTACK - MAIN IDEA

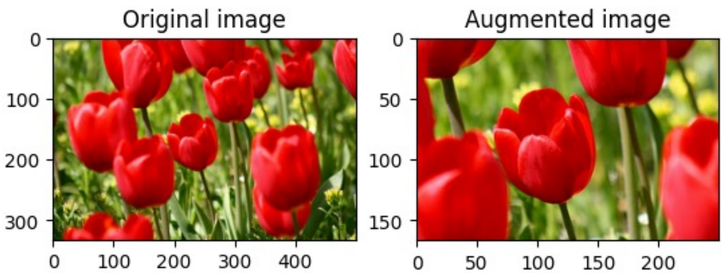
The training of this attack model requires the real label of the sample and the prediction confidence vector of the target model. Is the requirement for the attacker too harsh? You must know that too demanding requirements are difficult to achieve in reality. In order to solve these problems, Shokri and others cleverly proposed a core idea - **shadow model**. This is the key to making the member reasoning attack in the pioneering work.

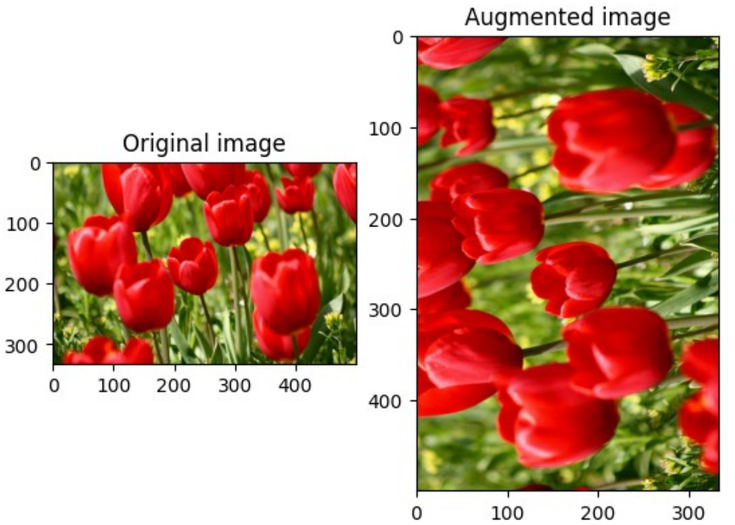
II. ATTACK CLASSIFICATION – MEMBERSHIP INFERENCE ATTACK – DATA AUGMENTATION









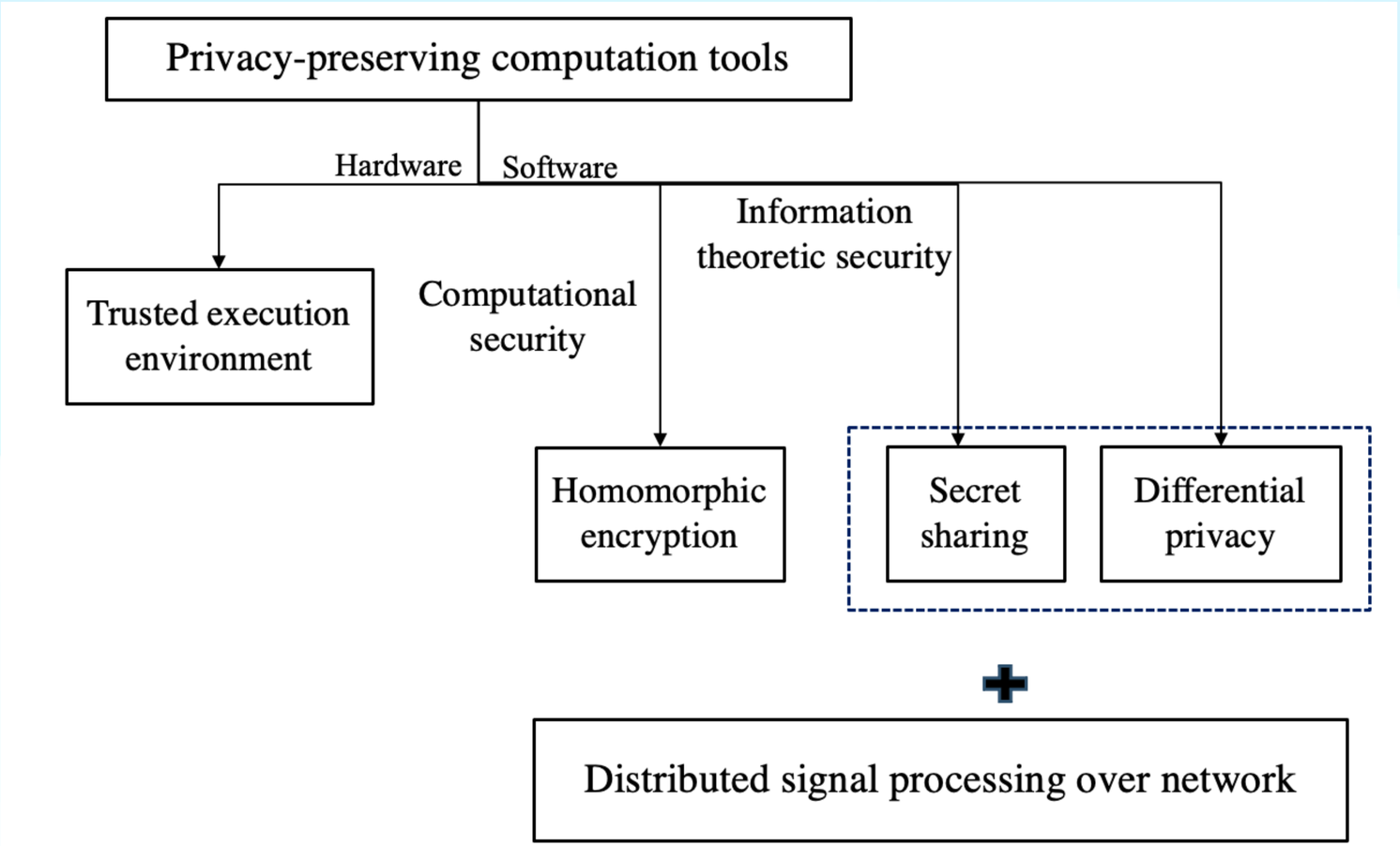


TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. DEFENSE CLASSIFICATION – ALL CATEGORIES

Overview of existing approaches

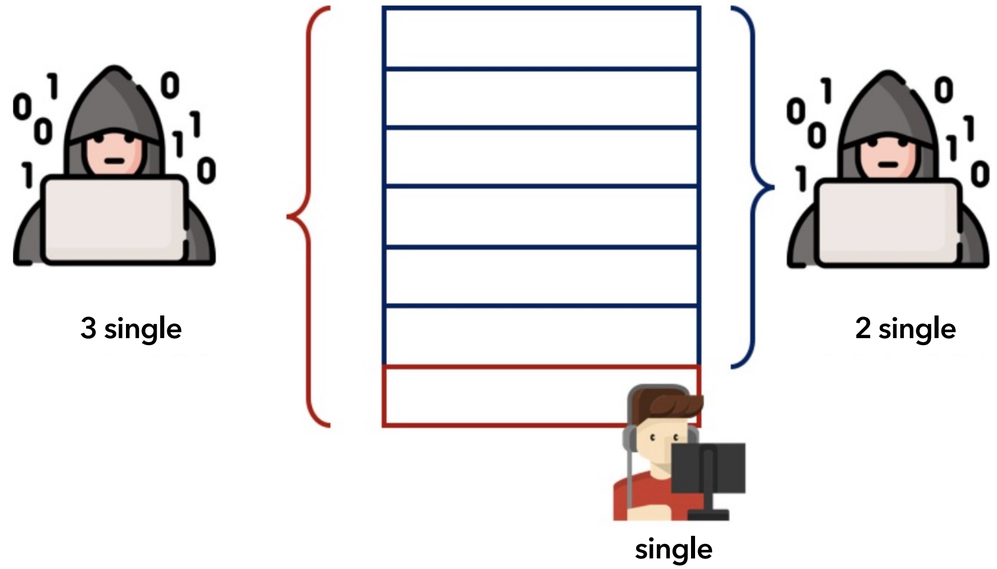


TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

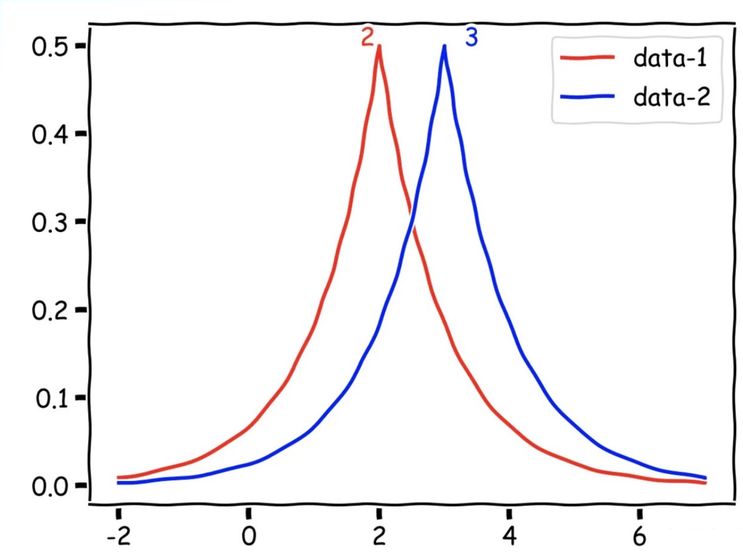
III. DEFENSE CLASSIFICATION – DIFFERENTIAL PRIVACY - BACKGROUND

Assuming that there is a marriage database, 2 are single and 8 are married, you can only check how many people are single but it is not allowed to check who is single. At the beginning of the inquiry, it was found that 2 people were single; now Bob went to register his marital status, and after checking again, he found 3 people were single. So Bob is single.



III. DEFENSE CLASSIFICATION – DIFFERENTIAL PRIVACY – MAIN IDEA

Now, if Bob is not in the database, the result may be 2.5; if Bob is, the result may be 2.5; The probability of getting a certain result from the query of the two data sets is so close that we can't tell which data set the result comes from. In this way, the knowledge of the attacker will not change due to the presence or absence of the sample Bob.



TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. DEFENSE CLASSIFICATION – DIFFERENTIAL PRIVACY – FORMULATION

The above query function can be represented by $f(x)$: $x \longrightarrow R$ (here only consider the case where the output result is 1D), and the random noise can be represented by r. The final query result is $M(x) = f(x) + r$, For two datasets x, x' with a Hamming distance of 1, for any output set a:

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

And ε is called the privacy budget. Generally speaking, the smaller ε is, the better the privacy protection is, but the greater the noise added, the lower the data availability.

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

IV. CURRENT RESEARCH DIRECTION

1. cryptographic techniques with distributed processing algorithms

2. New privacy-preserving approaches based on distributed signal processing tools

3. General metrics to relate and compare several existing information-theoretical approaches

4. …

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

Thanks

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS