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

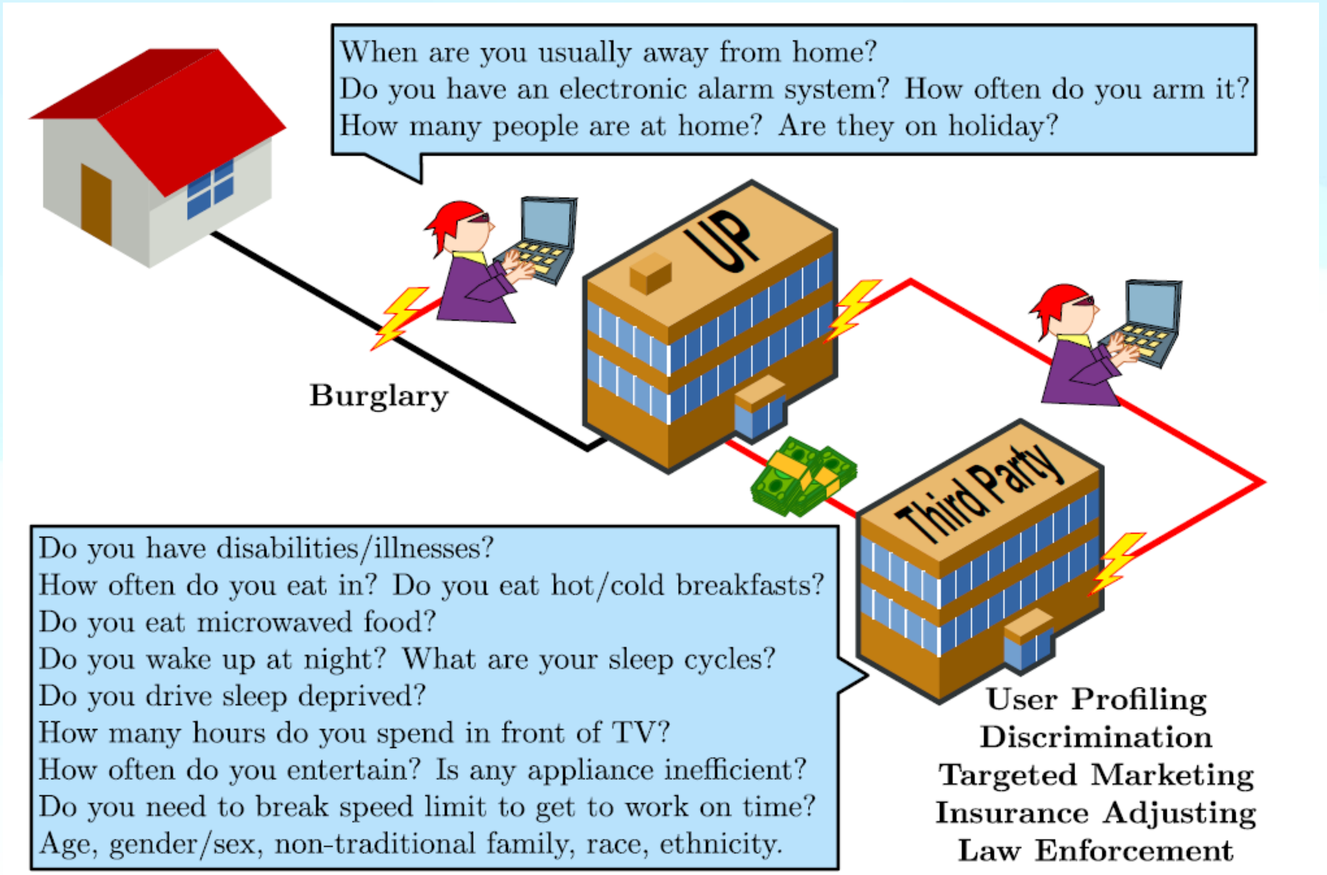
Health record

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.

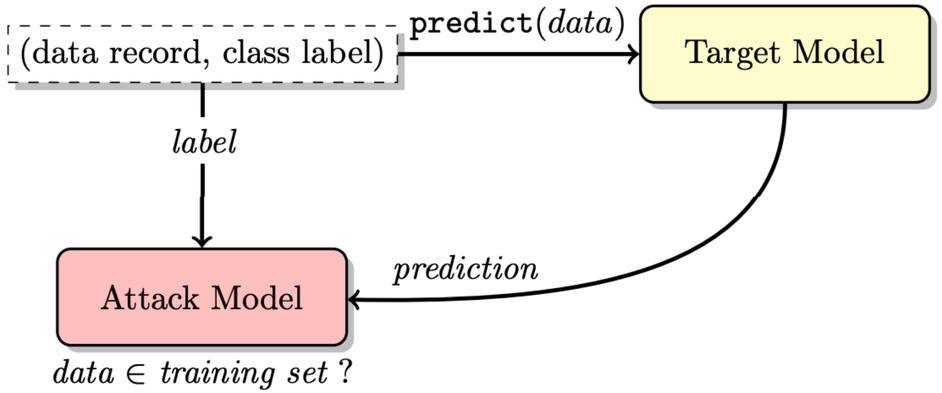
5. Functionality Extraction Attack (FEA): Through the input and output pairs of the target model, to imitate a model with similar functions.

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).



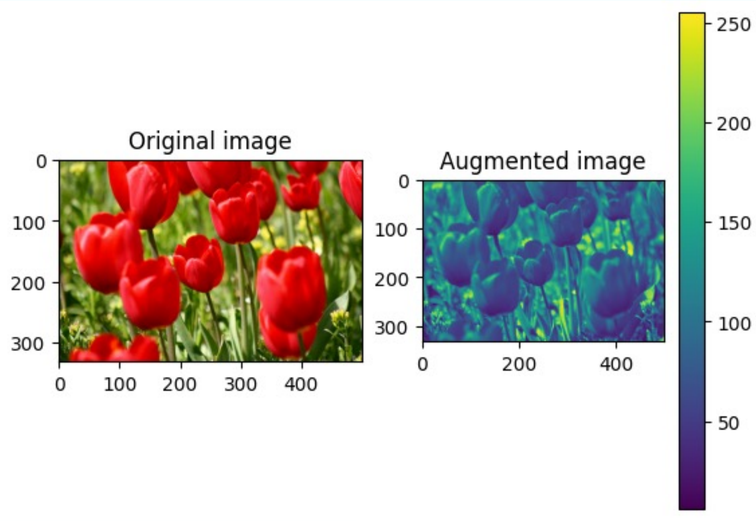
TELECOM SudParis

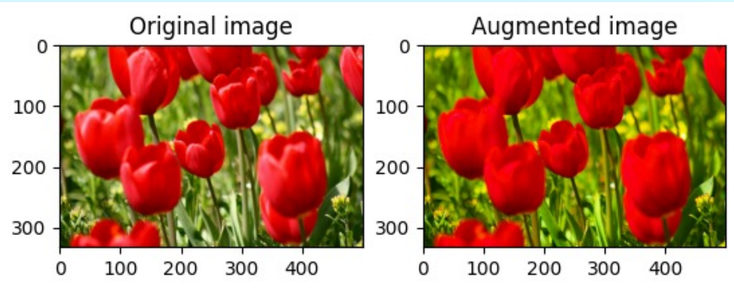
INSTITUT POLYTECHNIQUE DE PARIS

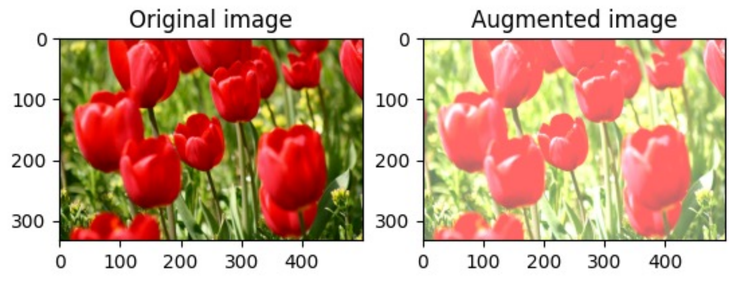
II. ATTACK CLASSIFICATION – MEMBERSHIP INFERENCE ATTACK - MAIN IDEA

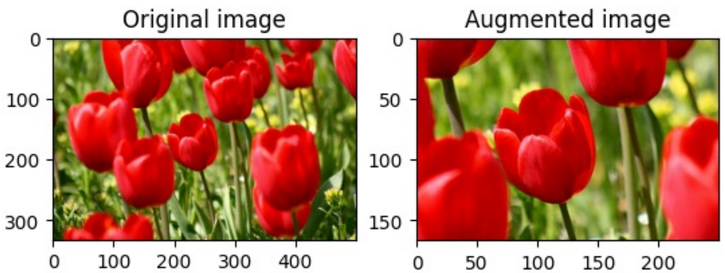
The training of this attack model requires the real label of the sample and the prediction confidence vector of the target model. Is the requirement for the attacker too harsh? You must know that too demanding requirements are difficult to achieve in reality. In order to solve these problems, Shokri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

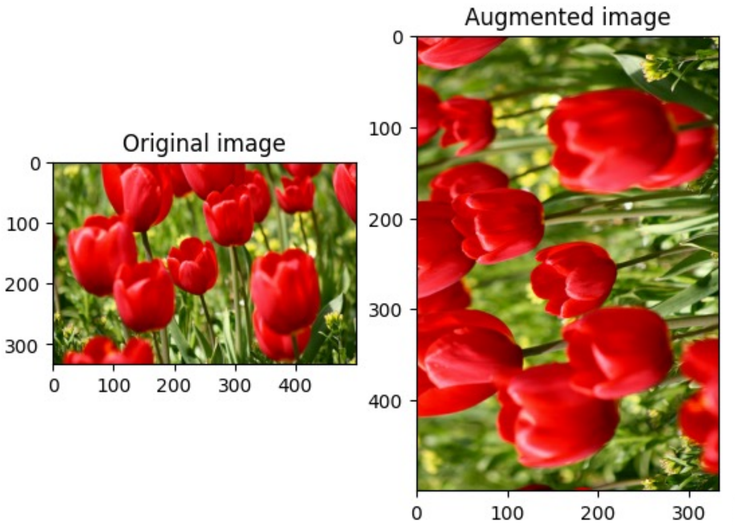
II. ATTACK CLASSIFICATION – MEMBERSHIP INFERENCE ATTACK – DATA AUGMENTATION









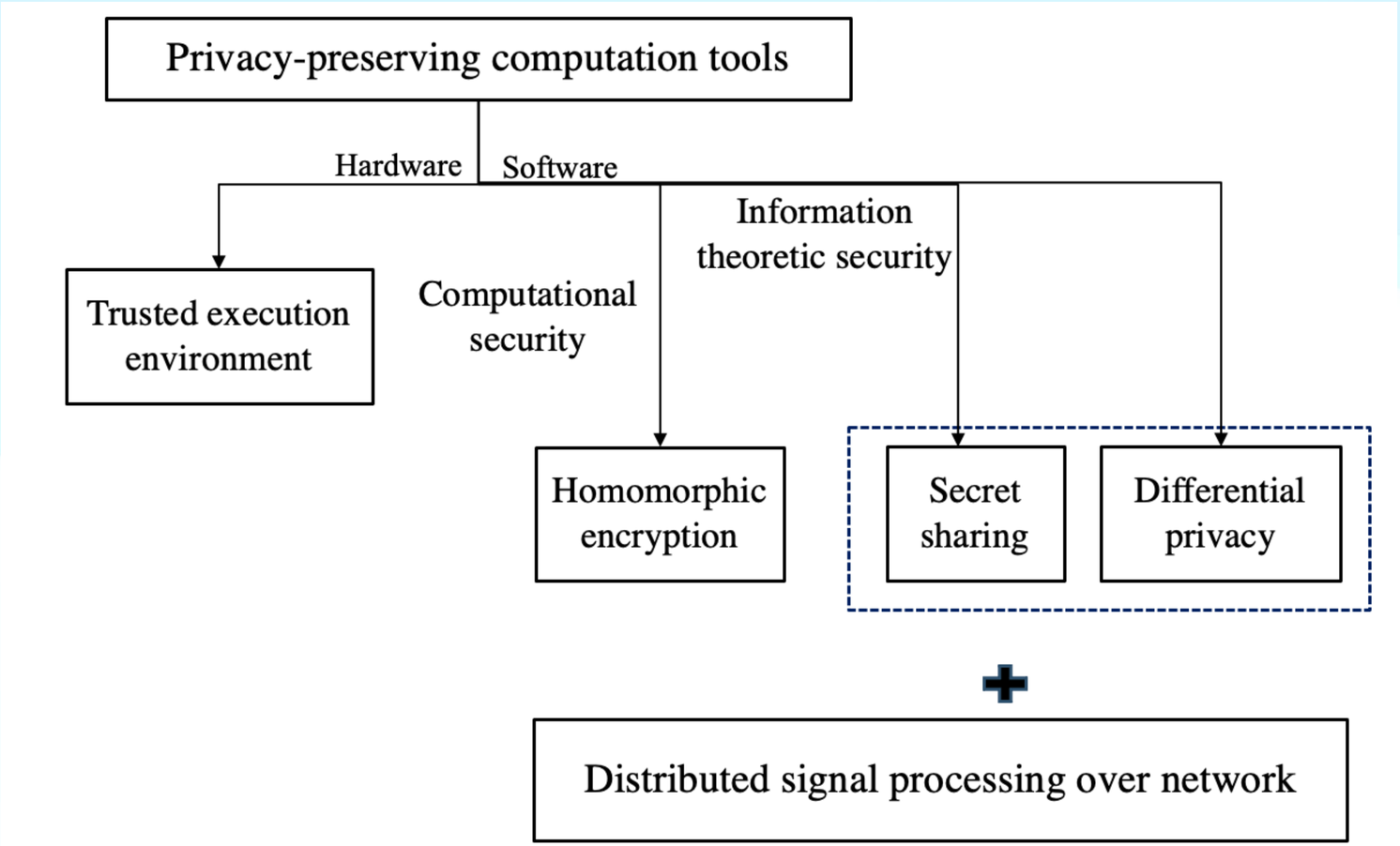


TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. DEFENSE CLASSIFICATION – ALL CATEGORIES

Overview of existing approaches

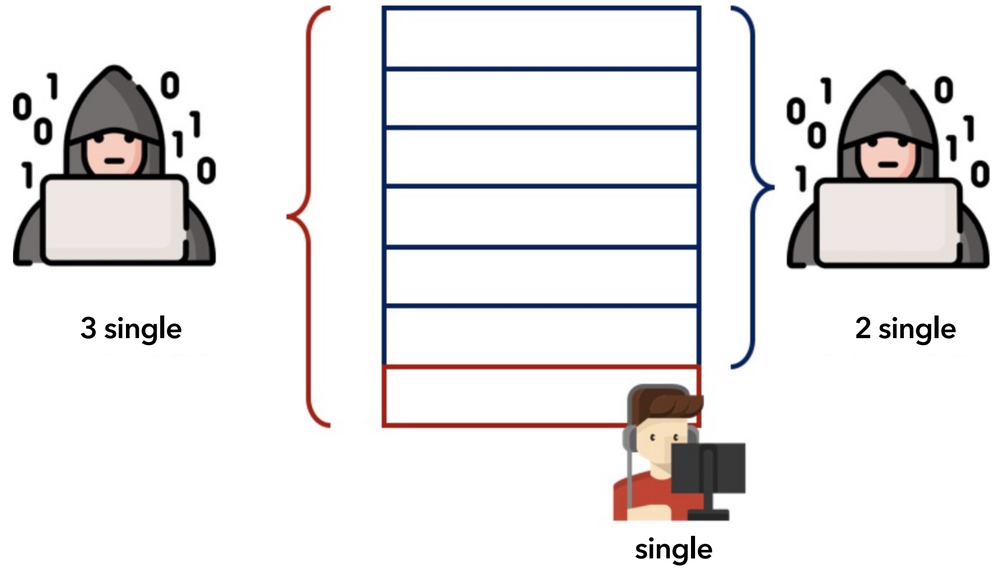


TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

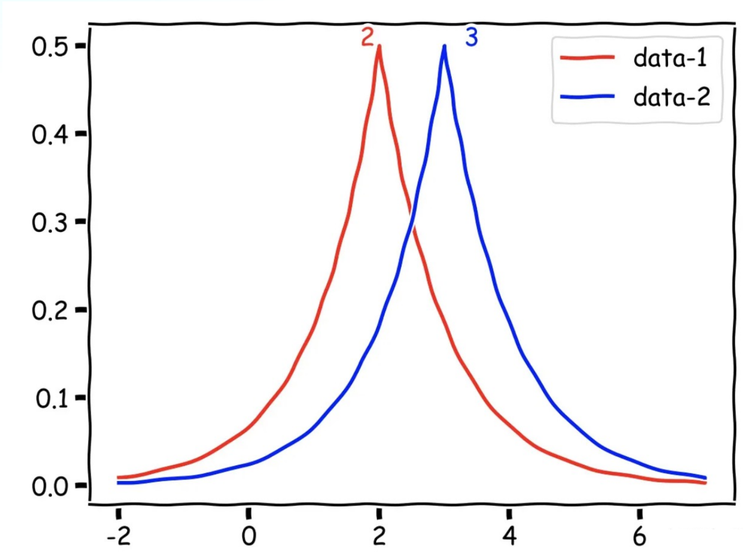
III. DEFENSE CLASSIFICATION – DIFFERENTIAL PRIVACY - BACKGROUND

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

Method	Training Acc	Test Acc	TPR @ 0.1% FPR	TPR @ 0.001% FPR	Log-scale AUC	MIA Balanced Acc
Base	100.0 ± 0.0	92.8 ± 0.2	8.20 ± 0.45	2.45 ± 0.93	0.815 ± 0.007	63.34 ± 0.26
Smooth	100.0 ± 0.0	92.9 ± 0.3	5.22 ± 0.66	0.14 ± 0.07	0.734 ± 0.012	62.28 ± 0.86
Disturblabel	99.9 ± 0.0	92.7 ± 0.3	5.88 ± 0.83	0.70 ± 0.45	0.775 ± 0.013	61.69 ± 0.24
Noise	100.0 ± 0.0	92.6 ± 0.2	$\mathbf{8.33} \pm 0.26$	$\mathbf{2.79} \pm 0.68$	$\mathbf{0.819} \pm 0.004$	$\mathbf{63.56} \pm 0.24$
Cutout	100.0 ± 0.0	93.1 ± 0.4	7.71 ± 0.39	$\mathbf{2.48} \pm 1.03$	0.811 ± 0.010	63.23 ± 0.26
Mixup	99.7 ± 0.1	93.0 ± 0.2	5.17 ± 0.51	1.31 ± 0.40	0.779 ± 0.008	60.05 ± 0.53
Jitter	100.0 ± 0.0	92.7 ± 0.2	$\mathbf{8.24} \pm 0.35$	$\mathbf{2.97} \pm 0.76$	$\mathbf{0.819} \pm 0.004$	$\mathbf{63.41} \pm 0.31$
Distillation	99.9 ± 0.0	93.2 ± 0.2	7.04 ± 0.33	2.19 ± 0.70	0.805 ± 0.005	61.57 ± 0.39
PGD-AT	99.2 ± 0.1	82.2 ± 0.2	$\mathbf{23.78} \pm 0.89$	$\mathbf{10.52} \pm 2.30$	$\mathbf{0.897} \pm 0.005$	$\mathbf{78.82} \pm 0.37$
TRADES	96.2 ± 0.2	80.0 ± 0.4	$\mathbf{17.88} \pm 1.56$	$\mathbf{8.14} \pm 1.12$	$\mathbf{0.881} \pm 0.006$	$\mathbf{77.21} \pm 0.65$
AWP	93.2 ± 2.0	82.6 ± 0.9	$\mathbf{10.58} \pm 3.48$	$\mathbf{3.06} \pm 1.81$	$\mathbf{0.828} \pm 0.045$	$\mathbf{72.13} \pm 3.76$
TRADES-AWP	91.9 ± 0.5	80.5 ± 0.2	$\mathbf{12.43} \pm 0.89$	$\mathbf{3.48} \pm 1.36$	$\mathbf{0.848} \pm 0.006$	$\mathbf{74.86} \pm 0.80$

Table 1. Attack success rates of different data enhancement on CIFAR-10. The 2nd and 3rd columns show the training and test accuracies of each method, respectively. The 4th - 7th columns show four metrics to evaluate the extent of privacy leakage. We highlight the MIA success rates for different DA and AT methods that are larger than that for Base.

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

Thanks

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

II. SPECIFIC WORK - DETAILED WORK

The above query function can be represented by $f(x)$: $x \longrightarrow R$ (here only consider the case where the output result is 1D for the time being), 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:

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. EXPERIMENT EVALUATION - EXPERIMENT SETTING

1. Data Collection: Foursquare dataset and Gowalla dataset

2. Evaluation Plan: Training set: validation set: test set = 8:1:1

3. Evaluation Metric: Since our application scenario focuses on recommending activities to users, our primary evaluation objective is to see whether a user's interested activity appears at the top of the returned list

Dataset	New York (Foursquare)	Tokyo (Foursquare)	New York (Gowalla)
Users	824	1,939	244
Venues	38,336	61,858	9,352
Check-ins	227,428	573,703	85,010
Average number of activity categories per user	38.37	31.39	55.58

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. EXPERIMENT EVALUATION - IMPACT OF PARAMETERS

The STAP model separately considers spatial and temporal features of user activity preference. From spatial perspective, a user's PFRs are determined by four parameters, i.e., l, d, s_{freq}, and $s_{ratioPB}$.

1. Spatial Parameter Setting

2. Temporal Parameter Setting

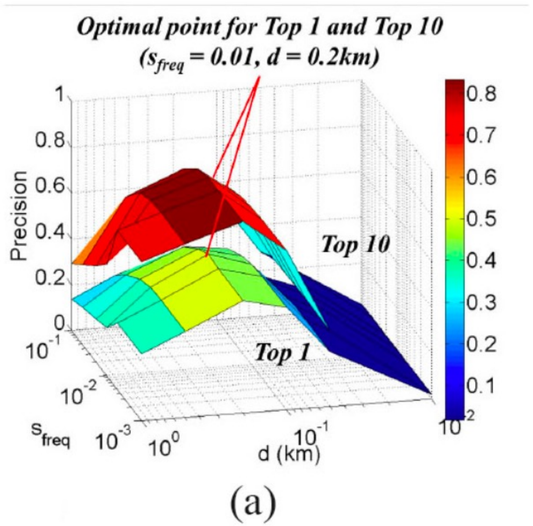
TELECOM SudParis

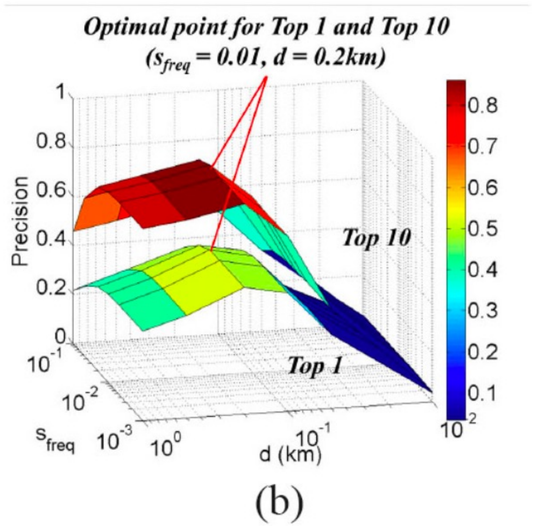
INSTITUT POLYTECHNIQUE DE PARIS

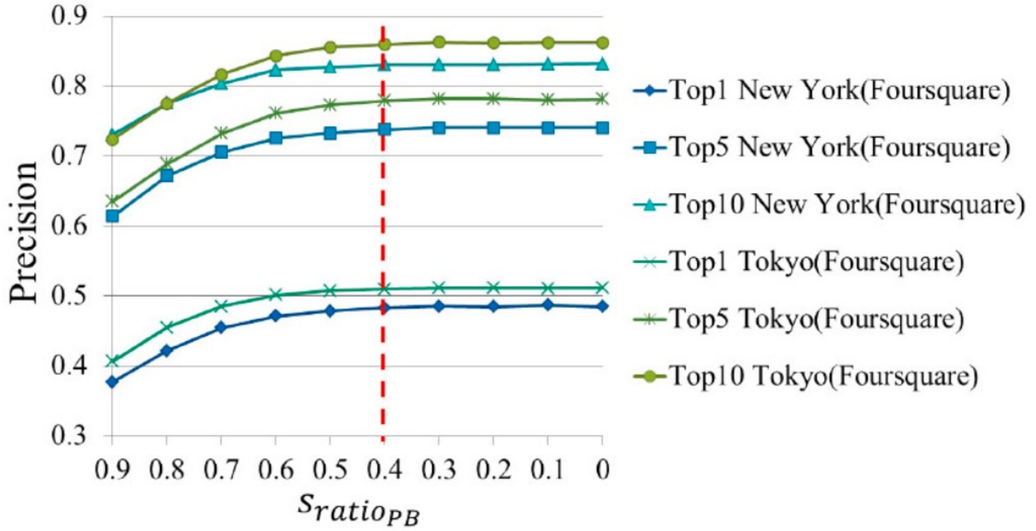
III. EXPERIMENT EVALUATION - IMPACT OF PARAMETERS

1. Spatial Parameter Setting

$d = 0.2$ km, $s_{freq} = 0.01$, and $s_{ratioPB} = 0.4$







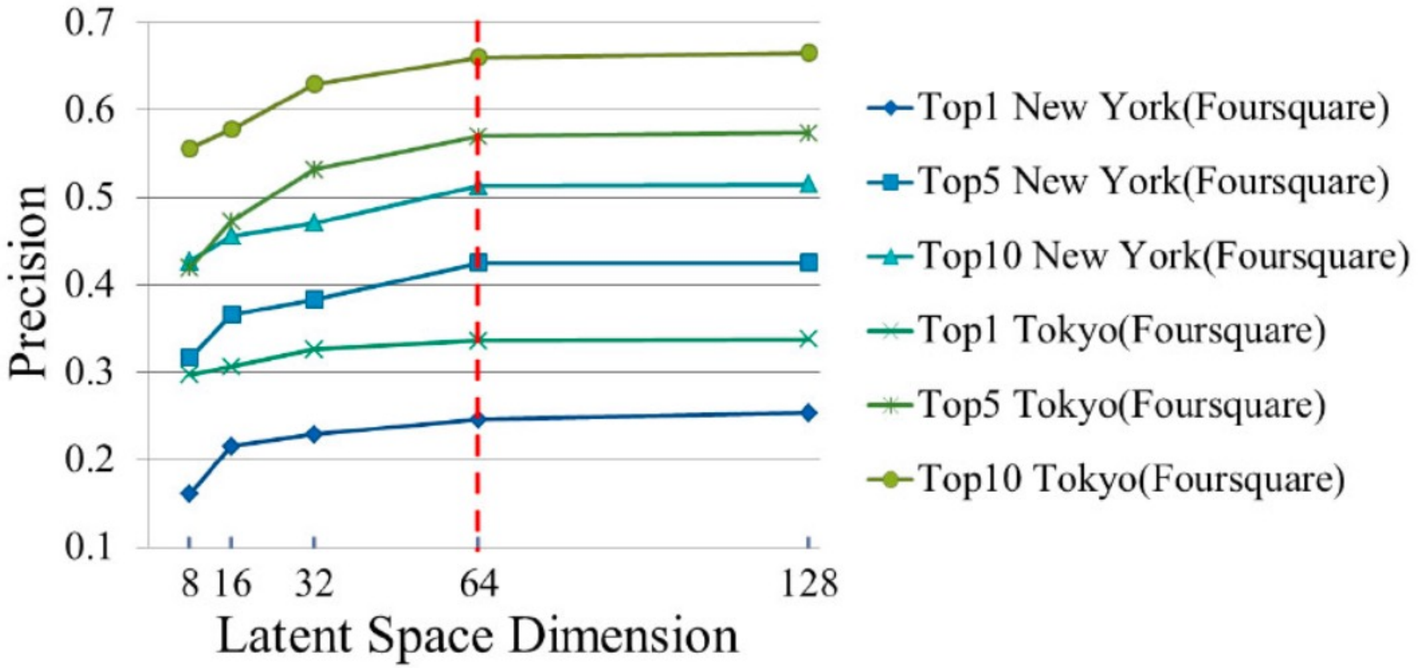
TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. EXPERIMENT EVALUATION - IMPACT OF PARAMETERS

2. Temporal Parameter Setting

The latent space dimension is set to 64.



TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. EXPERIMENT EVALUATION - COMPARISONS

Three Comparisons

A. Comparison With Baseline Approaches

B. Comparison Between Different Datasets

C. Comparison Between Different Activity Categories

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. EXPERIMENT EVALUATION - COMPARISONS

A. Comparison With Baseline Approaches - Different Approaches

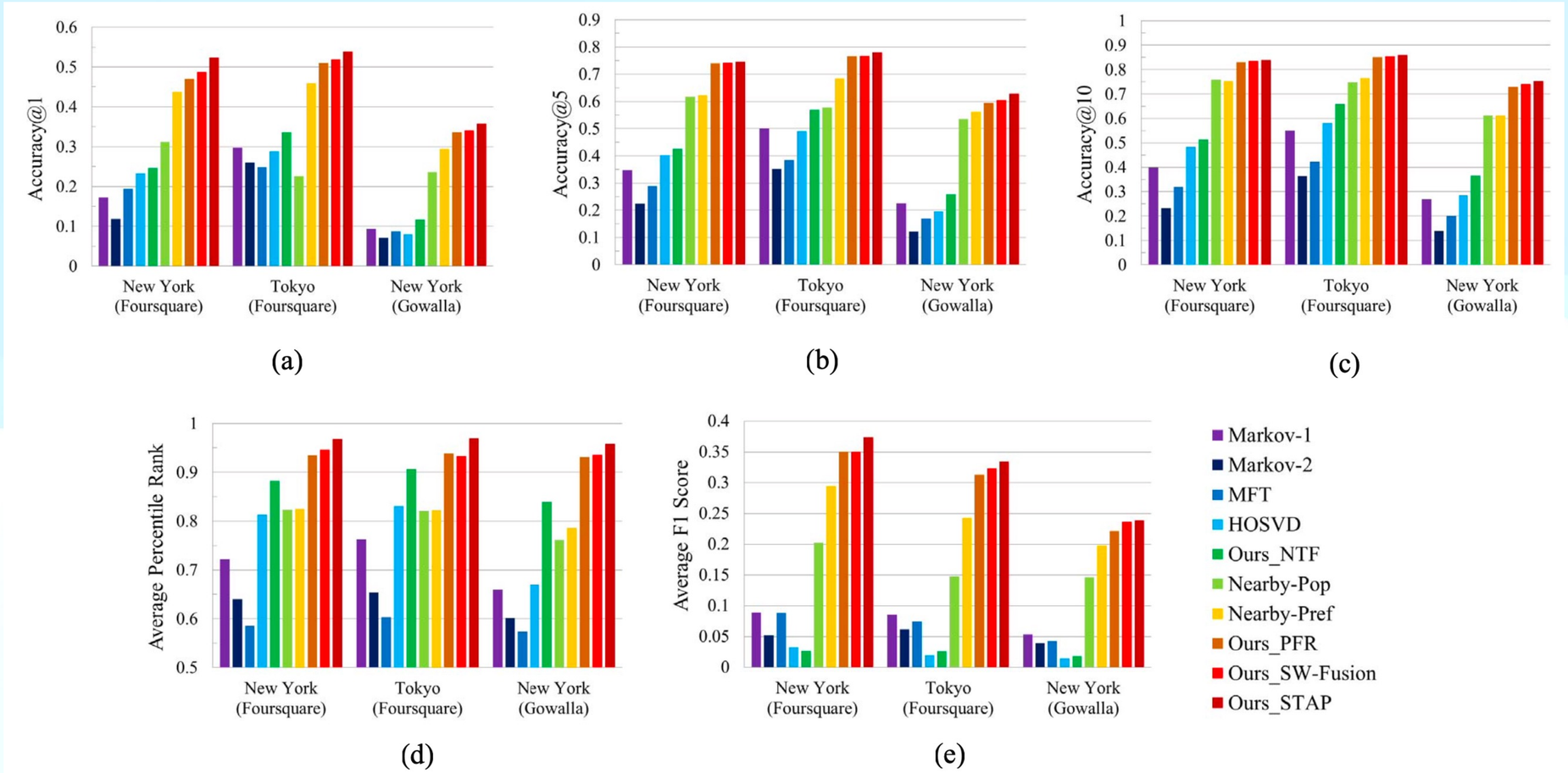
1. Sequential pattern mining approaches
    ① Order-K Markov model (Markov-K)
2. Temporal based approaches
    ① Most frequent activity by time (MFT)
    ② High order singular vector decomposition (HOSVD)
    ③ Temporal model of our STAP model (Ours_NTF)
3. Spatial based approaches
    ① Most popular activity around (nearby-pop)
    ② Most preferred activity around (nearby-pref)
    ③ Spatial model of our STAP model(Ours_PFR)
4. Spatial temporal based approaches
    ① Static weighted fusion (Ours_SW-fusion)
    ② Ours_STAP

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. EXPERIMENT EVALUATION - COMPARISONS

A. Comparison With Baseline Approaches - Results Figures



TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. EXPERIMENT EVALUATION - COMPARISONS

A. Comparison With Baseline Approaches - Performance Analysis

1. For sequential pattern mining approaches, both order 1 and order 2 Markov model obtain unsatisfied results.

2. For temporal based approaches, tensor factorization methods, i.e., NTF and HOSVD, can better capture user activity preference than the frequency based approach, i.e., MFT.

3. Spatial based approaches lead to better performance than temporal based methods.

4. Compared to the static weighted fusion method SW- Fusion, the context-aware fusion framework achieves the best performance.

III. EXPERIMENT EVALUATION - COMPARISONS

B. Comparison Between Different Datasets - Analysis

1. Tokyo users' activities have stronger temporal regularities than those of New York users, because the accuracy difference between temporal based approaches and spatial based approaches is relatively small with the Tokyo dataset than that with the New York dataset.

2. The improvement of PFR-based approaches over the nonpersonalized functional region based approach (Nearby-Pop) is larger with the Tokyo dataset than with the New York dataset.

3. comparing the results on the New York (Foursquare) and New York (Gowalla) datasets, we find that our solution consistently achieves better performance than the baselines.

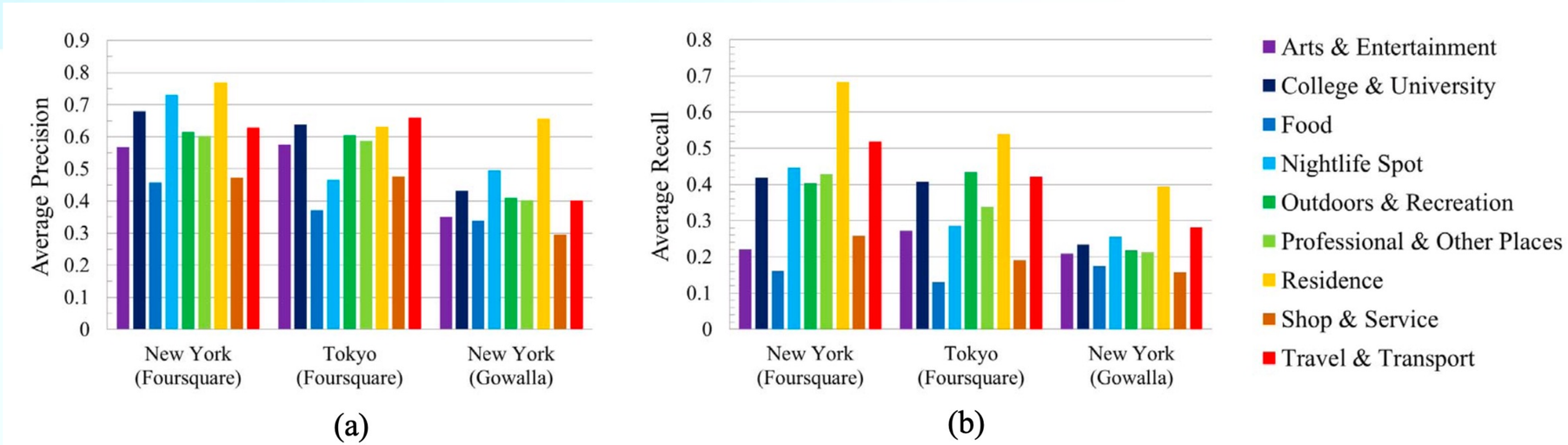
TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

III. EXPERIMENT EVALUATION - COMPARISONS

C. Comparison Between Different Activity Categories - Analysis

1. In all the three datasets, we observe that categories in residence and college & university yield good precision and recall, which implies that users in LBSNs exhibit strong spatial temporal regularities in activities like going home and going to school.
2. There are also some activity categories yield different results with different datasets.



TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

IV. CONCLUSIONS

The experiment results show that the STAP model achieves consistently good performance with all three datasets and outperforms various baseline approaches, which verifies the generality and advantages of our solution in modeling spatial-temporal activity preference with sparse check-in data.

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS