Semi-Supervised Knowledge Transfer For Deep Learning From Private Training Data

Nicolas Papernot[1] Martn Abadi[2] Ifar Erlingsson[2] Ian Goodfellow[2] Kunal Talwar [2]

[1]Pennsylvania State University

[2]Google Brain

ICLR, 2017

Presenter: Xueying Bai

Outline

Outline

Challenge of Learning from Private Data

Learning algorithm should protect the privacy of user's private training data (eg. private photographs).
However, there are some problems revealed when learning from private data using traditional machine learning methods.

- **Some examples are implicitly memorized in the model.** Experiments on deep neural network show this phenomenon.
- **Attacks can recover the sensitive training data from models.** For example, Fredrikson used output probabilities of a computer vision classifier to reveal individual faces from the training data.

The goal of this paper is to preserve the privacy of training data when learning classifiers.

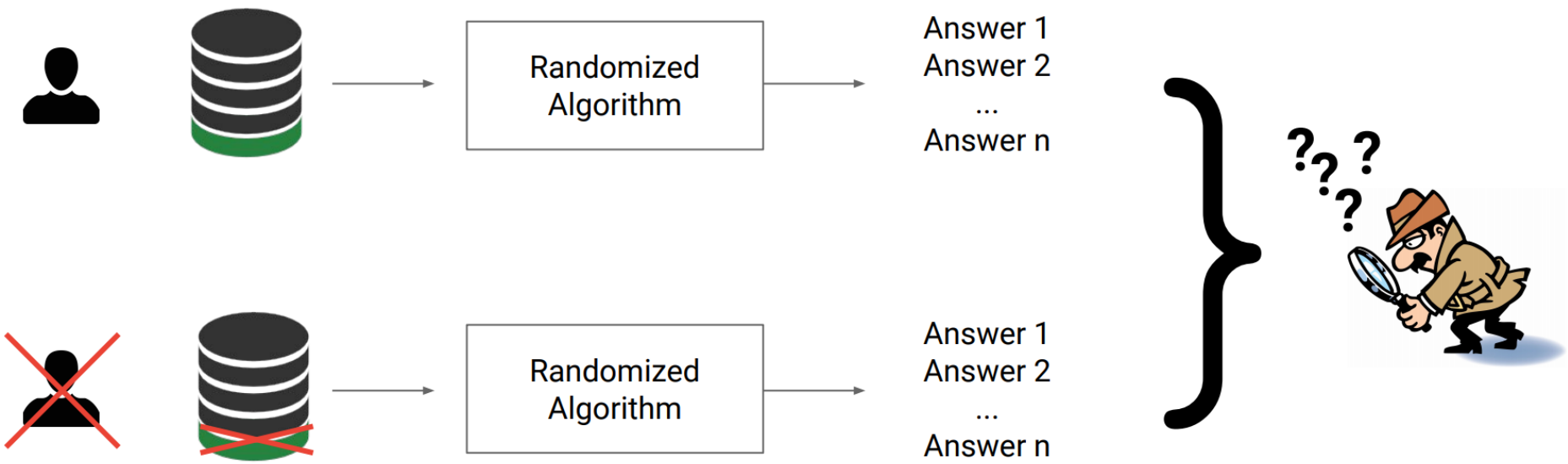
Attack

Two types of attacks:

- **Black-box Adversary:** Recover private training data through model quering, without knowing the parameters and the structure of the model.
- **White-box Adversary:** Know the model structure and parameters.
- **Strong attack assumption in this paper:** Adversary can make unbounded number of queries (black-box); adversary can access to model internals (white box).

Privacy

Strategy which has privacy guarantee:



Outline

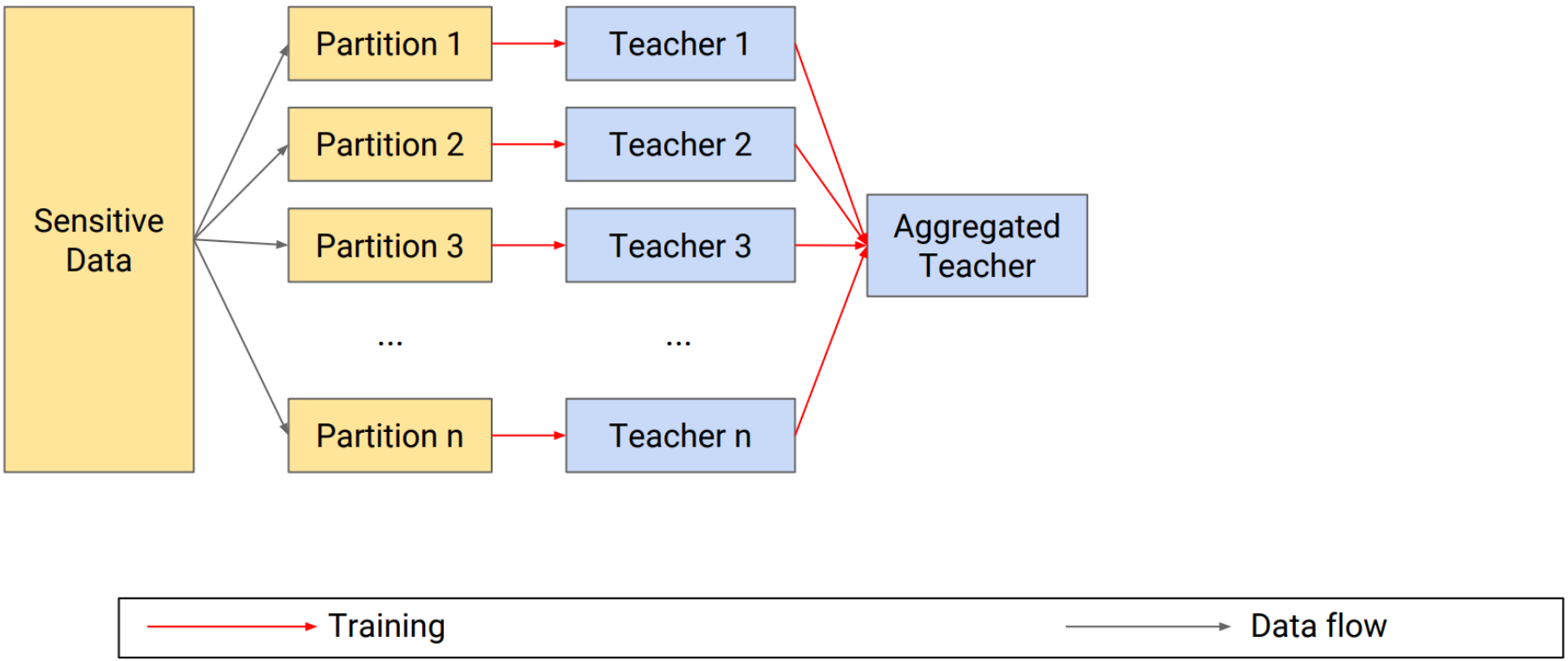
Overview

- **Approach.** First, an ensemble of teacher models is trained on disjoint subsets of the sensitive data. Then, using auxiliary, unlabeled non-sensitive data, a student model is trained on the aggregate output of the ensemble.
- **Privacy analysis of the approach.**

Outline

Data Partitioning and Teachers

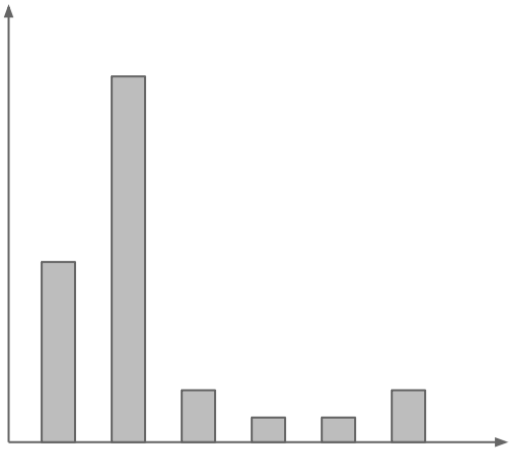
Partition the data into n disjoint sets, then train a model separately on each set.



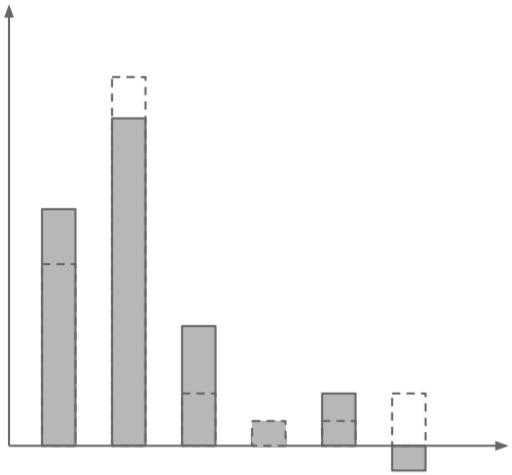
Aggregation

- If most teachers agree on the label, it does not depend on specific partitions, so the privacy cost is small.
- If two majority classes have close vote counts, the disagreement may reveal private information.
- Add random noise to the vote counts to introduce ambiguity.

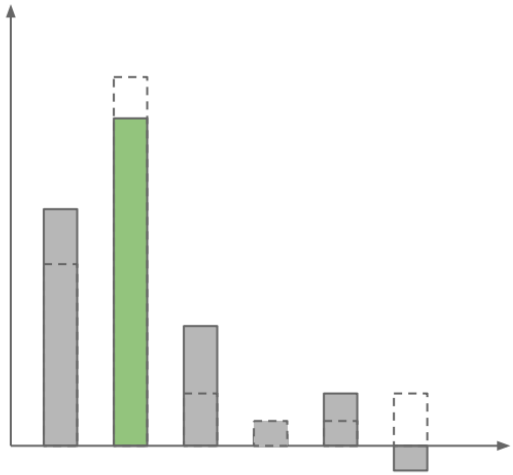
Aggregation



Count votes



Add Laplacian noise



Take maximum

- $j \in [m]$ is the class label, $i \in [n]$ is the ith teacher, $\overrightarrow{x}$ is the input.
- Count votes: $n_j(\overrightarrow{x}) = |\{i : i \in [n], f_i(\overrightarrow{x}) = j\}|$
- Laplacian noise: $Lap(\frac{1}{\gamma})$
- Output label: $f(x) = \arg\max_j \{n_j(\overrightarrow{x}) + Lap(\frac{1}{\gamma})\}$

Outline

Reason to Use the Student Model

There are threats in teacher ensemble:

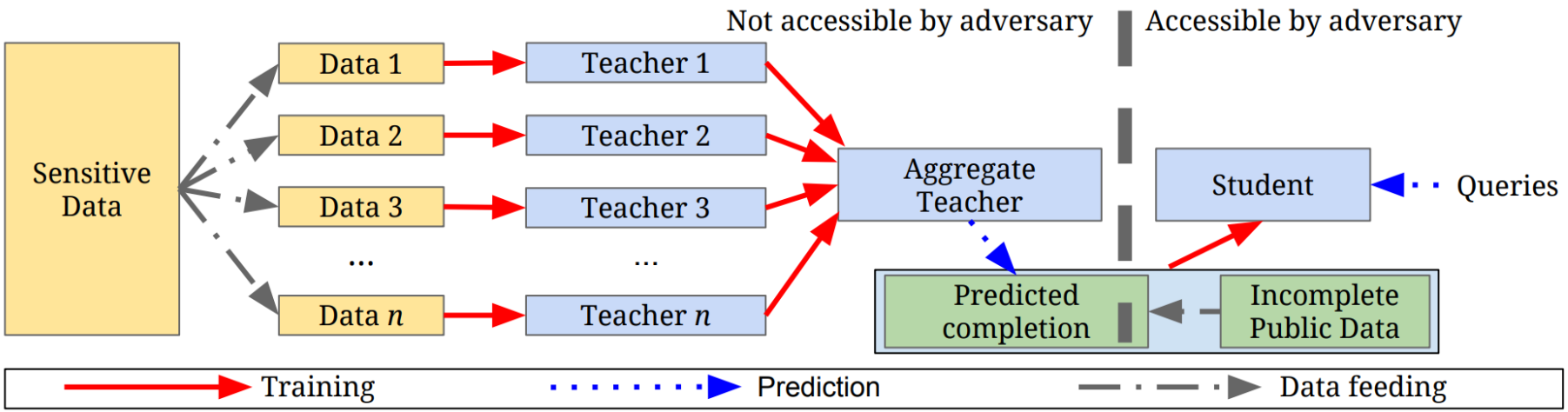
- **Each prediction increases privacy cost.** For each query to the teacher, the aggregated result will reveal information of the data. So need the student model.
- **Inspection of internals may reveal private data.** Later analysis will prove the privacy guarantee.

The student model is the one deployed, in lieu of the teacher ensemble.

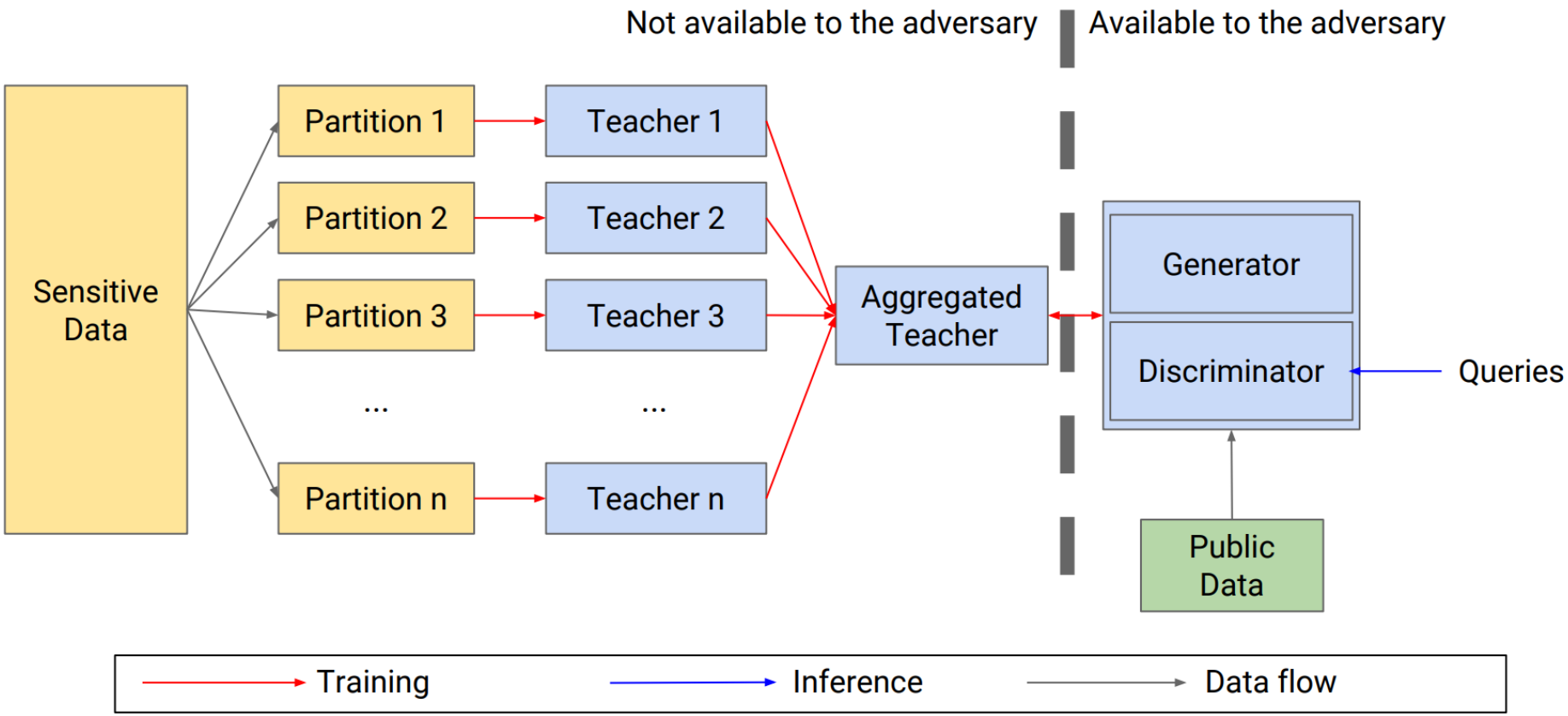
- Student finishes training after fixed number of queries to the teacher. Then no matter how many queries the user gives to the student, the privacy loss is fixed.

The Student Model

The student model is trained on unlabeled public data. To do semi-supervised learning, part of the data are labeled using the teacher aggregation result.



Train the Student with GANs



- The discriminator is extended to multi-class classifier with $m + 1$ classes (m classes plus a generated class).
- Only trained discriminator will be used after training.

Outline

Differential Privacy Preliminaries

Differential privacy is a strong standard. It's defined using pairs of adjacent databases: d and d', which only differ by only 1 training example.

- **Definition 1.** A randomized mechanism M with domain D and range R satisfies (ε, δ)-differential privacy if for any two adjacent input $d, d' \in D$ and for any subset of outputs $S \subset R$ it holds that:

$$\Pr[M(d) \in S] \leq e^{\varepsilon} \Pr[M(d') \in S] + \delta$$

  Smaller ε, stronger privacy guarantee. δ is the tolerant bias.

- **Definition 2.** Let $M : D \to R$ be a randomized mechanism and d, d' a pair of adjacent databases. Let aux denote an auxiliary input. For an outcome $o \in R$, the privacy loss at o is defined as:

$$c(o; M, aux, d, d') \triangleq \log \frac{\Pr[M(aux, d) = o]}{\Pr[M(aux, d') = o]}$$

  The privacy random variable is defined as:

$$C(M, aux, d, d') \triangleq c(M(d); M, aux, d, d')$$

The Moments Accountant

A natural way to bound the approach's privacy loss is to first bound the privacy cost of each label queried by the student, and then use the composition theorem to derive the total cost of training the student. So it's better to track each step's privacy cost.

- **Definition 3.** Let $M : D \to R$ be a randomized mechanism and d, d' a pair of adjacent databases. Let aux denote the auxiliary input. The moments accountant is defined as:

$$\alpha_M(\lambda) \stackrel{\Delta}{=} \max_{aux,d,d'} \alpha_M(\lambda; aux, d, d')$$

$$\alpha_M(\lambda; aux, d, d') \stackrel{\Delta}{=} \log E[\exp(\lambda C(M, aux, d, d'))]$$

$\alpha_M(\lambda; aux, d, d')$ is a moment generating function of the privacy loss random variable.

Composability and Tail bound

Theorem 1. *1.* ***[Composability]*** *Suppose that a mechanism $\mathcal{M}$ consists of a sequence of adaptive mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_k$ where $\mathcal{M}_i \colon \prod_{j=1}^{i-1} \mathcal{R}_j \times \mathcal{D} \to \mathcal{R}_i$. Then, for any output sequence $o_1, \ldots, o_{k-1}$ and any λ*

$$\alpha_{\mathcal{M}}(\lambda; d, d') = \sum_{i=1}^{k} \alpha_{\mathcal{M}_i}(\lambda; o_1, \ldots, o_{i-1}, d, d') \,,$$

where $\alpha_{\mathcal{M}}$ is conditioned on $\mathcal{M}_i$'s output being o_i for $i < k$.

2. ***[Tail bound]*** *For any $\varepsilon > 0$, the mechanism $\mathcal{M}$ is (ε, δ)-differentially private for*

$$\delta = \min_{\lambda} \exp(\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon) \,.$$

- This theorem enables adding bounds over continuous steps, and derive an (ϵ, δ) guarantee for the final α.

Privacy Analysis of PATE

Theorem 2. *Suppose that on neighboring databases d, d', the label counts n_j differ by at most 1 in each coordinate. Let $\mathcal{M}$ be the mechanism that reports $\arg\max_j \left\{ n_j + Lap(\frac{1}{\gamma}) \right\}$. Then $\mathcal{M}$ satisfies $(2\gamma, 0)$-differential privacy. Moreover, for any l, aux, d and d',*

$$\alpha(l; \textsf{aux}, d, d') \leq 2\gamma^2 l(l+1)$$

Theorem 3. *Let $\mathcal{M}$ be $(2\gamma, 0)$-differentially private and $q \geq \Pr[\mathcal{M}(d) \neq o^*]$ for some outcome o^*. Let $l, \gamma \geq 0$ and $q < \frac{e^{2\gamma}-1}{e^{4\gamma}-1}$. Then for any aux and any neighbor d' of d, $\mathcal{M}$ satisfies*

$$\alpha(l; \textsf{aux}, d, d') \leq \log((1-q)\left(\frac{1-q}{1-e^{2\gamma}q}\right)^l + q\exp(2\gamma l)).$$

Lemma 4. *Let $\mathbf{n}$ be the label score vector for a database d with $n_{j^*} \geq n_j$ for all j. Then*

$$\Pr[\mathcal{M}(d) \neq j^*] \leq \sum_{j \neq j^*} \frac{2 + \gamma(n_{j^*} - n_j)}{4\exp(\gamma(n_{j^*} - n_j))}$$

- These theorems enable the bound of specific moments.

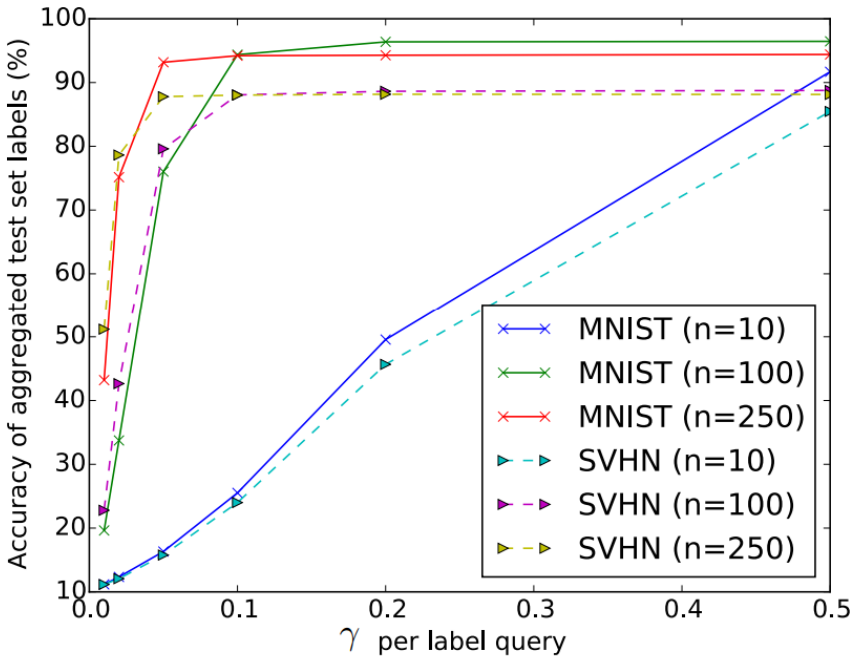
Outline

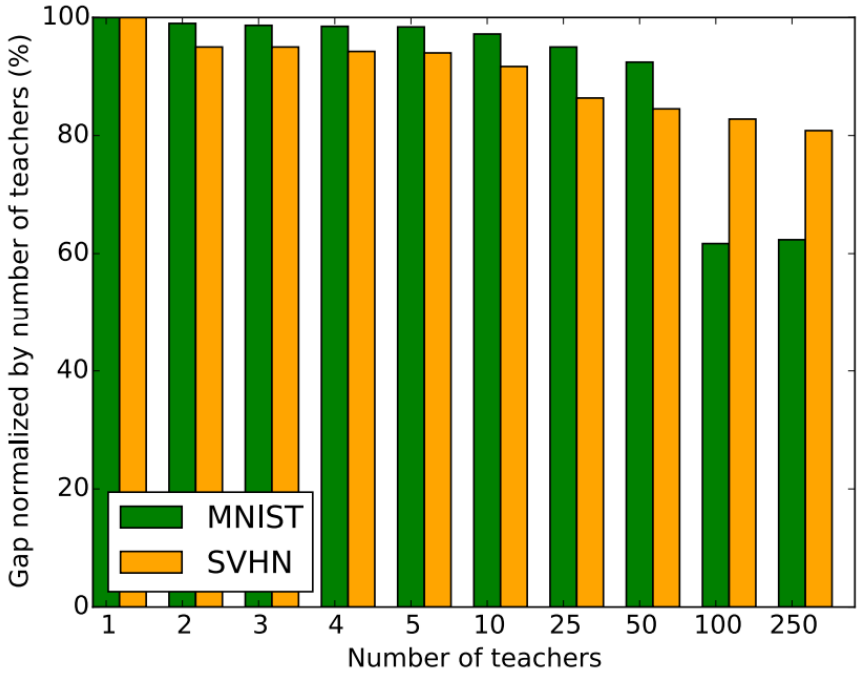
Settimgs

Dataset	Teacher Model	Student Model	Student Public Data	Testing Data
MNIST	2 conv + 1 relu	GANs (6 fc layers)	test[:1000]	test[1000:]
SVHN	2 conv + 2 relu	GANs (7 conv + 2 NIN)	test[:1000]	test[1000:]
UCI Adult	RF (100 trees)	RF (100 trees)	test[:500]	test[500:]
UCI Diabetes	RF (100 trees)	RF (100 trees)	test[:500]	test[500:]

Outline

Training an Ensemble of Teachers and Privacy





- Correct label: $\frac{1}{\gamma}$ should be small.
- Strong privacy: γ should be small.
- Large gap will result to strong privacy. When increasing teachers, the gap will increase.

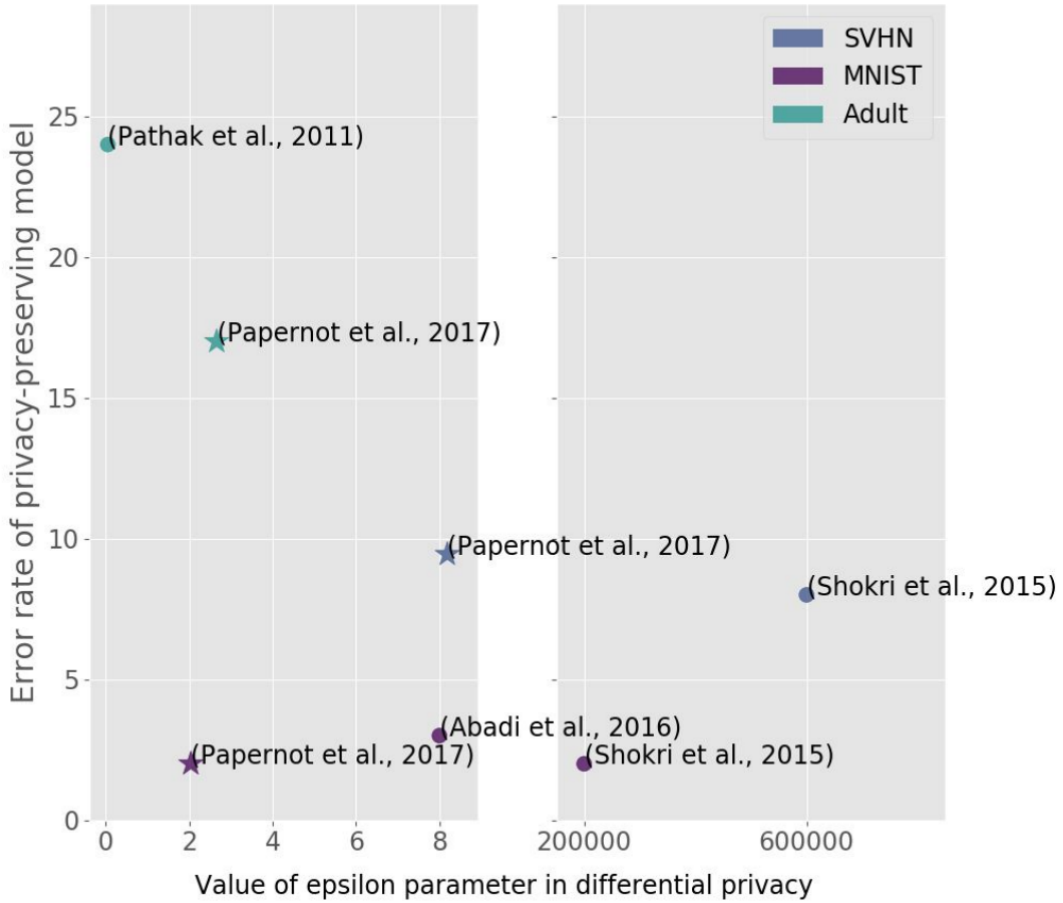
Outline

Semi-supervised Training of the Student

Dataset	ε	δ	Queries	Non-Private Baseline	Student Accuracy
MNIST	2.04	10^{-5}	100	99.18%	98.00%
MNIST	8.03	10^{-5}	1000	99.18%	98.10%
SVHN	5.04	10^{-6}	500	92.80%	82.72%
SVHN	8.19	10^{-6}	1000	92.80%	90.66%

Outline

Comparison



UCI Diabetes	
ε	1.44
δ	10^{-5}
Non-private baseline	93.81%
Student accuracy	93.94%

Conclusions

Main contributions in this paper:

- Combine semi-supervised learning with precise, data-dependent privacy analysis.
- Establish a precise guarantee of training data privacy.
- The model is independent of the learning algorithm for either teachers or students, very generic.
- Experiments show it can achieve comparable result with the state-of-the-art under the privacy guarantee.