Quantifying Membership Privacy via Information Leakage

Sara Saeidian[*], Giulia Cervia[†], Tobias Oechtering[*], Mikael Skoglund[*]

[*]{saeidian, oech, skoglund}@kth.se

[†]giulia.cervia@imt-nord-europe.fr

Outline

- Introduction & Context
- Maximal Leakage
- Entrywise Information Leakage
- Privacy Case Study: PATE
- Privacy Analysis of PATE

Introduction & Context

Privacy-preserving Machine Learning

- Machine learning models known to memorize unique properties of individual data points
- This can be exploited by several types of privacy attacks such as
  - reconstruction attacks
  - model inversion attacks
  - *membership inference attacks*

Membership Inference Attacks

- **Goal**: whether or not a sample was used in the training
  - **Example**: Was Alice's data used to train a model for detecting cancer?
- Requires only black-box access to the machine learning model
  - **Example**: shadow models [1]
- Differential privacy [2] by definition neutralizes the attack
- *Information theoretic view of membership privacy?*

[1] Reza Shokri et al. "Membership inference attacks against machine learning models". In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18

[2] Cynthia Dwork, Aaron Roth, et al. "The algorithmic foundations of differential privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407

Maximal Leakage

Maximal Leakage: Setup

- Assume X is a private random variable and Y is the public output of a channel with input X

How much information does Y leak about X?

- Consider a threat model where the adversary
  - observes Y
  - is interested in guessing some discrete function of X, called U

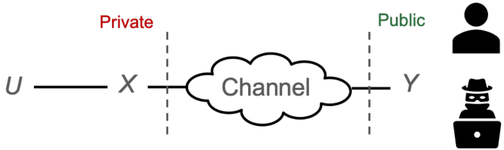

Figure 1: Threat model

Maximal Leakage: Definition

Definition: Maximal Leakage[3]

The maximal leakage from X to Y is defined as

$$\mathcal{L}(X \to Y) = \sup_{U: U - X - Y} \log \frac{\mathbb{P}\left(U = \hat{U}(Y)\right)}{\max_{u \in \mathcal{U}} P_U(u)},$$

where $\hat{U}$ is the optimal (MAP) estimator of U.

Maximal leakage

- captures the multiplicative increase in the probability of correctly guessing U, upon observing Y
- is an operationally meaningful measure of privacy

[3] Ibrahim Issa, Aaron B Wagner, and Sudeep Kamath. "An operational approach to information leakage". In: *IEEE Transactions on Information Theory* (2019)

Maximal Leakage: Properties

- For **finite** alphabets, maximal leakage takes the simple form
$$\mathcal{L}(X \to Y) = \log \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}: P_X(x) > 0} P_{Y|X}(y \mid x).$$
- Two important properties:
  - **Composition**: if the Markov chain $Y_1 - X - Y_2$ holds
$$\mathcal{L}(X \to (Y_1, Y_2)) \leq \mathcal{L}(X \to Y_1) + \mathcal{L}(X \to Y_2).$$
  - **Data-processing inequality**: if the Markov chain $X - Y_1 - Y_2$ holds
$$\mathcal{L}(X \to Y_2) \leq \min\{\mathcal{L}(X \to Y_1), \mathcal{L}(Y_1 \to Y_2)\}.$$

Entrywise Information Leakage

Entrywise Information Leakage

- Maximal leakage quantifies the information leaking about the whole dataset
- We want to measure the information leakage about individual data entries

> What if we assume the adversary knows *all the entries except for a single data entry*?

- In this setup, observations leak information only about the unknown entry
- *But how do we model the adversary's side information?*

Pointwise Conditional Maximal Leakage: Definition

Definition: Pointwise Conditional Maximal Leakage[4]

Suppose the value of the random variable Z is a priori given as $z \in \mathcal{Z}$. The pointwise conditional maximal leakage from X to Y given $Z = z$ is defined as

$$\mathcal{L}(X \to Y|Z = z) := \sup_{U: U-(X,Z)-Y} \log \frac{\mathbb{P}\left(U = \hat{U}(Y, Z = z)\right)}{\mathbb{P}\left(U = \tilde{U}(Z = z)\right)},$$

where both $\hat{U}$ and $\tilde{U}$ are optimal estimators of U.

- For **finite** alphabets, pointwise conditional maximal leakage takes the simple form

$$\mathcal{L}(X \to Y|Z = z) = \log \sum_{y \in \mathcal{Y}} \max_{x:\, P_{X|Z}(x|z) > 0} P_{Y|XZ}(y|x, z).$$

[4] Cf. Issa, Wagner, and Kamath, "An operational approach to information leakage", Def. 6

Pointwise Conditional Maximal Leakage: Properties

Same useful properties as maximal leakage:

- **Composition**: if the Markov chain $Y_1 - (X,Z) - Y_2$ holds
$$\mathcal{L}(X \to (Y_1,Y_2) \mid Z=z) \leq \mathcal{L}(X \to Y_1 \mid Z=z) + \mathcal{L}(X \to Y_2 \mid Z=z).$$
- **Data-processing inequality**: if the Markov chain $(X,Z) - Y_1 - Y_2$ holds
$$\mathcal{L}(X \to Y_2 \mid Z=z) \leq \min\{\mathcal{L}(X \to Y_1 \mid Z=z), \mathcal{L}(Y_1 \to Y_2 \mid Z=z)\}.$$

Privacy Case Study: PATE

Private Aggregation of Teacher Ensembles (PATE)

- PATE [5,6] is a framework for privacy-preserving **classification** of sensitive data
- Three main components:
  - ensemble of teacher models
  - aggregation mechanism
  - student model

[5] Nicolas Papernot et al. "Semi-supervised knowledge transfer for deep learning from private training data". In: *arXiv preprint arXiv:1610.05755* (2016)

[6] Nicolas Papernot et al. "Scalable private learning with pate". In: *arXiv preprint arXiv:1802.08908* (2018)

PATE: System Model

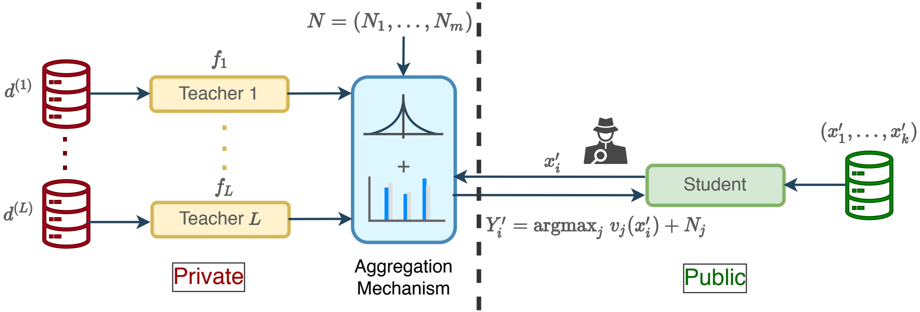

Figure 2: PATE System Model

PATE: Teacher Models

- Training data is divided into disjoint partitions
- Each teacher is a classification model trained on one of the partitions
- Teachers predict labels independently of each other

PATE: Aggregation Mechanism

- Adds noise to the histogram of teachers' votes and returns the class with the largest (noisy) value

- **Example**:
  - $L = 4$ teachers and $m = 3$ classes
  - $f_1(x'_i) = 0,\ f_2(x'_i) = 2,\ f_3(x'_i) = 2$, and $f_4(x'_i) = 0$.

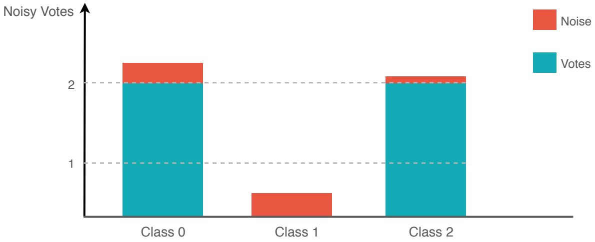

Figure 3: Example illustrating the aggregation mechanism

PATE: Student Model

- A classification model trained using a **public unlabeled** dataset that is labeled by the teachers' ensemble through the aggregation mechanism
- Must be trained with as few queries as possible

PATE: Advantages

- No need to centrally store sensitive data
- Privacy guarantees independent of the machine learning techniques used to train the teachers/student
- Privacy-accuracy synergy: increased agreement among teachers in labeling a query lowers its privacy cost

Privacy Analysis of PATE

Some Notation

Notation	Meaning
D	training data
D^*	**unknown** data entry
$D^- = D \setminus D^*$	**known** data entries
$(x'_1, \ldots, x'_k)$	student's unlabeled dataset
$(Y'_1, \ldots, Y'_k)$	predicted labels
$V(x'_i) = \big(V_1(x'_i), \ldots, V_m(x'_i)\big)$	histogram of votes for x'_i
$V^-(x'_i) = \big(V^-_1(x'_i), \ldots, V^-_m(x'_i)\big)$	histogram of **known** votes for x'_i
$N = (N_1, \ldots, N_m)$	sequence of noise

Table 1: Notation

Overview of Approach (1/2)

- Assume the adversary knows $D^- = d^-$ and wants to guess D^*
- Evaluate
$$\mathcal{L}(D^* \to (Y_1', \ldots, Y_k') \mid D^- = d^-) = \mathcal{L}(D \to (Y_1', \ldots, Y_k') \mid D^- = d^-)$$
- Use the composition lemma for pointwise conditional maximal leakage
$$\mathcal{L}(D \to (Y_1', \ldots, Y_k') \mid D^- = d^-) \leq \sum_{i=1}^{k} \mathcal{L}(D \to Y_i' \mid D^- = d^-)$$

Overview of Approach (2/2)

- Use the data-processing inequality for pointwise conditional maximal leakage

$$\mathcal{L}(D \to Y_i' \mid D^- = d^-) \leq \min\{\underbrace{\mathcal{L}(D \to V(x_i') \mid D^- = d^-)}_{\text{leakage of training}}, \underbrace{\mathcal{L}(V(x_i') \to Y_i' \mid D^- = d^-)}_{\text{leakage of aggregation}}\}.$$

- Evaluate leakage of aggregation (leakage of training is difficult to analyze and is usually very large)

$$\mathcal{L}(V(x_i') \to Y_i' \mid D^- = d^-) = \mathcal{L}(V(x_i') \to Y_i' \mid V^-(x_i') = v^-)$$

Some Definitions: Majorization

Definition: Majorization[7]

Consider $p, q \in \mathbb{R}^n$ with non-increasingly ordered elements, i.e., $p_1 \geq p_2 \geq \ldots \geq p_n$ and $q_1 \geq q_2 \geq \ldots \geq q_n$. We say that p majorizes q, and write $p \succ q$ if

$$\sum_{i=1}^{m} p_i \geq \sum_{i=1}^{m} q_i, \quad \text{for} \quad m = 1, \ldots, n-1 \quad \text{and} \quad \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i.$$

Examples: define $\mathcal{Q} = \{(q_1, q_2, q_3) \in \mathbb{R}^3 : \sum_{i=1}^{3} q_i = 9\}$

- $(5, 3, 1) \succ (4, 4, 1)$
- $(4, 4, 1)$ and $(5, 2, 2)$ cannot be compared using majorization
- $(3, 3, 3)$ is majorized by all $q \in \mathcal{Q}$
- $(9, 0, 0)$, $(0, 9, 0)$ and $(0, 0, 9)$ majorize all $q \in \mathcal{Q}$

[7] Albert W Marshall, Ingram Olkin, and Barry C Arnold. *Inequalities: theory of majorization and its applications*. Vol. 143. Springer, 1979

Some Definitions: Schur-concave Function

Definition: Schur-concave Function

Consider a real-valued function Φ defined on $\mathcal{I}^n \subset \mathbb{R}^n$. Φ is said to be Schur-concave on $\mathcal{I}^n$ if $p \succ q$ on $\mathcal{I}^n$ implies $\Phi(p) \leq \Phi(q)$.

Some Definitions: Log-concave Function

Definition: Log-concave Function

A non-negative function $f : \mathbb{R}^n \to \mathbb{R}_+$ is said to be log-concave if it can be written as $f(x) = \exp \phi(x)$ for some concave function $\phi : \mathbb{R}^n \to [-\infty, \infty)$.

Examples:

- Gaussian probability density $\quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
- Laplace probability density $\quad f(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$
- ...

Results: Schur-concavity of the Entrywise Leakage

Theorem 1

Consider the aggregation mechanism in PATE where the noise has a **log-concave** probability density. Then, $\mathcal{L}(V(x_i') \to Y_i' \mid V^{-}(x_i') = v^{-})$ is Schur-concave in v^{-}.

This implies that

- leakage is maximized when
$$v^{-} = v^{-}_{max} = \left(\frac{L-1}{m}, \ldots, \frac{L-1}{m}\right),$$
- leakage is minimized when
$$v^{-} = v^{-}_{min} = (0, \ldots, 0, L-1, 0, \ldots, 0).$$
- *stronger agreement among teachers lowers the privacy cost of a query*

Results: Bounds using Laplace Noise (1/3)

Proposition 1

Consider the PATE framework with **Laplace** distributed noise. Then,

$$\mathcal{L}(V(x_i') \to Y_i' \mid V^-(x_i') = v^-) \leq \frac{1-m}{m} 2^{-m} e^{-\gamma} + \frac{1}{m}\left[1 - \left(1 - \frac{1}{2}e^{-\gamma}\right)^m\right] e^{\gamma}$$
$$+ \frac{1}{2}\left(1 - \frac{1}{2}e^{-\gamma}\right)^{m-1} - \frac{m-1}{4} e^{-\gamma} H(m-2),$$

where

$$H(m) := \gamma + \sum_{k=1}^{m} \frac{2^{-k} - \left(1 - \frac{1}{2}e^{-\gamma}\right)^k}{k} \quad \text{for } m \geq 1 \quad \text{and} \quad H(0) := \gamma,$$

and equality holds for $v^- = v^-_{max}$.

Results: Bounds using Laplace Noise (2/3)

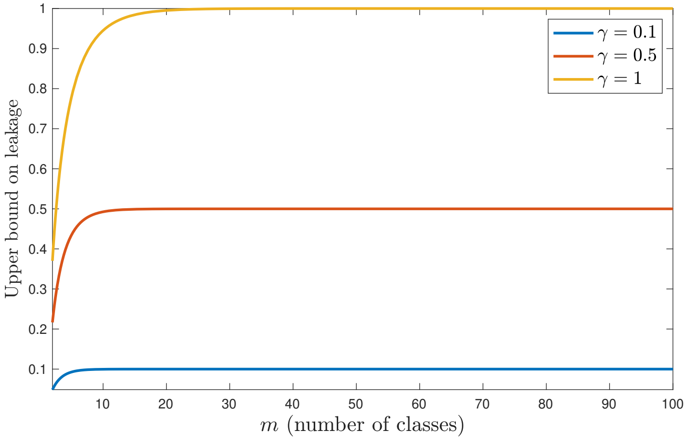

Figure 4: Upper bound on the entrywise leakage for different m and γ

Results: Bounds using Laplace Noise (3/3)

- Can we simplify the bound in Proposition 1?

Theorem 2

Consider the PATE framework with **Laplace** distributed noise. Then,

$$\mathcal{L}(D^* \to Y_i' \mid D^- = d^-) = \mathcal{L}(D \to Y_i' \mid D^- = d^-) \leq \gamma.$$

Summary

- We showed that the entrywise leakage of the aggregation mechanism in PATE is Schur-concave when the noise has log-concave pdf
- We derived bounds on the leakage with Laplace noise