Quantifying Membership Privacy via Information Leakage

Sara Saeidian, *Member, IEEE*, Giulia Cervia, *Member, IEEE*, Tobias J. Oechtering, *Senior Member, IEEE*, and Mikael Skoglund, *Fellow, IEEE*

***Abstract*— Machine learning models are known to memorize the unique properties of individual data points in a training set. This memorization capability can be exploited by several types of attacks to infer information about the training data, most notably, membership inference attacks. In this paper, we propose an approach based on information leakage for guaranteeing membership privacy. Specifically, we propose to use a conditional form of the notion of *maximal leakage* to quantify the information leaking about *individual* data entries in a dataset, i.e., the entrywise information leakage. We apply our privacy analysis to the *Private Aggregation of Teacher Ensembles* (PATE) framework for privacy-preserving classification of sensitive data and prove that the entrywise information leakage of its aggregation mechanism is Schur-concave when the injected noise has a log-concave probability density. The Schur-concavity of this leakage implies that increased consensus among teachers in labeling a query reduces its associated privacy cost. Finally, we derive upper bounds on the entrywise information leakage when the aggregation mechanism uses Laplace distributed noise.**

***Index Terms*— Privacy-preserving machine learning, membership inference, maximal leakage, log-concave probability density.**

I. Introduction

In recent years, many useful machine learning applications have emerged that require training on sensitive data. Such applications span across a diverse range of fields such as medical imaging [1], rumor identification in social media [2], or financial fraud detection [3]. While all machine learning applications by definition reveal some information about the training data, privacy concerns arise when machine learning models memorize properties that are *unique* to individual data entries. In fact, a variety of privacy attacks have demonstrated that it is indeed possible to exploit this "memorization" capability of models to infer information about data entries in the training set [4].

Arguably, the simplest type of privacy attacks against machine learning models is *membership inference* attacks in which an adversary infers whether or not a certain data point was used in the training [5], [6]. In response to such attacks, a number of mitigation techniques have been proposed in the literature, with *differential privacy*-based methods being the most commonly studied. Differential privacy [7] provides provable and operationally meaningful privacy guarantees, and by definition neutralizes membership inference attacks. Roughly speaking, differential privacy ensures that all datasets differing in only one entry (i.e., adjacent datasets) produce an output with similar probabilities. Moreover, it has several useful properties, such as satisfying data-processing inequalities and composition theorems [7].

The standard definition of differential privacy (i.e., pure differential privacy) uses a parameter ϵ to define a multiplicative upper bound on the changes in the probability of an output for all adjacent datasets in the input [8]. However, this definition is known to be very strict, and has limited applicability. As such, several relaxations of differential privacy have been proposed, the most notable of which is (ϵ, δ)-differential privacy [9]. A common interpretation of (ϵ, δ)-differential privacy is that the guarantees of ϵ-differential privacy hold except with probability δ. Thus, it provides the necessary flexibility for studying a larger class of privacy-preserving mechanisms such as the Gaussian mechanism [8].

Despite the advantages of (ϵ, δ)-differential privacy, one should note that its privacy guarantees are qualitatively different from those of pure differential privacy (see [10] for illustrative examples). On this account, recently Rényi differential privacy [10] was proposed as an alternative relaxation of pure differential privacy. While Rényi differential privacy satisfies the same useful properties as pure differential privacy, it does not offer any intuitive operational meaning, and its privacy guarantees are usually translated into (ϵ, δ)-differential privacy for interpretation.

In this paper, we propose to use (a conditional form of) the notion of *maximal leakage* [11] to measure the amount of information leaking about any single data entry in a dataset, i.e., the entrywise information leakage. Maximal leakage [11] is an operationally meaningful privacy metric that captures the inference capabilities of an adversary trying to deduce some information about the input data by observing the

Manuscript received October 12, 2020; revised January 21, 2021; accepted April 4, 2021. Date of publication April 19, 2021; date of current version May 6, 2021. This work was supported by the Strategic Research Agenda Program, Information and Communication Technology–The Next Generation (SRA ICT–TNG) funded by the Swedish Government and KTH Digital Futures Center within the project DataLEASH. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Matthieu R. Bloch. *(Corresponding author: Sara Saeidian.)*

Sara Saeidian, Tobias J. Oechtering, and Mikael Skoglund are with the Division of Information Science and Engineering, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: saeidian@kth.se; oech@kth.se; skoglund@kth.se).

Giulia Cervia was with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden. She is now with the Centre for Digital Systems, IMT Lille Douai, Institut Mines-Télécom, University of Lille, 59000 Lille, France (e-mail: giulia.cervia@imt-lille-douai.fr).

Digital Object Identifier 10.1109/TIFS.2021.3073804

output. Specifically, maximal leakage *quantifies* the maximal gain in an adversary's ability to correctly guess any arbitrary discrete function of the input data by observing the output (as opposed to making a guess with no observations). Note that the original definition of maximal leakage quantifies the information leaking about the *whole dataset*, whereas we are interested in measuring the information leaking about *single data entries* in the dataset. As such, similarly to [12], we consider an adversary who knows the values of all the entries in the dataset, except for a single data entry of interest. Intuitively, in this setup, observations only convey the *unique* information contributed by the unknown data entry since all other entries are already known to the adversary. To quantify this entrywise information leakage, we propose a conditional form of maximal leakage, namely the *pointwise conditional maximal leakage*, which is also a special case of the event-conditional Sibson mutual information introduced in [13]. Then, by allowing the unknown entry to be any of the entries in the dataset, we can derive upper bounds on the entrywise information leakage, and provide meaningful worst-case privacy guarantees.

Maximal leakage satisfies several useful properties, most notably a data-processing inequality and a composition lemma [11]. The data-processing inequality ensures that no manipulation of the output can increase the information leakage, while the composition property characterizes the information leaked through multiple observations. Here, we show that the same properties hold for pointwise conditional maximal leakage, rendering it suitable for privacy analysis of more complex information systems.

We apply our privacy analysis to the *Private Aggregation of Teacher Ensembles* (PATE) framework [14], [15]. PATE is a general framework for privacy-preserving classification of sensitive data, and operates by transferring the knowledge of an ensemble of models (called *teachers*) trained on disjoint partitions of the sensitive data to a *student* classifier. Specifically, the student is trained using a public unlabelled dataset which will be labelled by the teachers through an *aggregation mechanism*. The aggregation mechanism is essentially the *Report-Noisy-Max mechanism* [7] which adds noise to the teachers' predictions to enable derivation of privacy guarantees.

PATE has several advantages as a privacy-preserving machine learning framework. First, the privacy guarantees result solely from the aggregation mechanism and are agnostic to the specific machine learning techniques used by each teacher. This is because the modular structure of PATE enables us to invoke the data-processing inequality to uncouple the information leaked through the training and aggregation, and guarantee that the overall leakage is less than both. Second, PATE lends itself well to distributed learning by allowing data owners to separately train their own predictors, hence mitigating the need for centralized storage of the sensitive data. Finally, the aggregation mechanism induces a favorable *synergy between privacy and accuracy* such that increased agreement among the teachers in labelling a query lowers its associated privacy cost. This synergy is one of the main focuses of this paper, and will be extensively studied.

The privacy guarantees established by PATE are characterized in [14], [15] in terms of differential privacy, and results from experiments are reported. However, these works do not analytically prove the aforementioned synergy between privacy and accuracy observed in the framework. Here, we will analyze the privacy of the framework in terms of the entrywise information leakage, and prove the privacy-accuracy synergy using analytical arguments in order to provide deeper insights into the workings of the framework, especially the Report-Noisy-Max mechanism used for aggregating teachers' predictions. As [14], [15] present a thorough experimental study, here we refrain from repeating the experiments but focus on giving a rigorous theoretical analysis of the framework.

A. Contributions

Our contributions can be summarized as follows:

i) **Introducing pointwise conditional maximal leakage.** We approach membership privacy from a novel angle by studying the information leakage of individual data entries in a database. We begin by deriving a data-processing inequality and a composition lemma for pointwise conditional maximal leakage, and then apply them to the problem of studying the entrywise information leakage in PATE.

ii) **Proving the privacy-accuracy synergy in PATE.** We show that the entrywise information leakage of the aggregation mechanism in PATE (i.e., the Report-Noisy-Max mechanism) is *Schur-concave* [16], [17] when the injected noise has a *log-concave* [18], [19] probability density. As we will see, this implies that increased consensus among teachers lowers the privacy cost of labelling a query. Note that many commonly used probability distributions including the Laplace and Gaussian distributions are log-concave rendering this result fairly general.

iii) **Deriving membership privacy guarantees for PATE with Laplace noise.** We derive upper bounds on the entrywise information leakage when the noise injected in the aggregation mechanism has Laplace distribution. We present two types of bounds: a data-independent bound, which holds uniformly for all training datasets and is tight in the sense that the bound holds with equality when the information leakage is maximized. Our other bound is data-dependent in that it depends on the training data through the teachers' predictions. The data-dependent bound can be tighter than the data-independent bound when there is a large consensus among the teachers in predicting the label of a query.

B. Other Related Work

Information leakage metrics. In recent years, a large body of work has been dedicated to studying various information-theoretic privacy metrics. Most notably, mutual information has been frequently proposed and studied as such a metric (see e.g., [20]–[22]) by appealing to its operational meaning in communication theory. Similarly, in [23] another information-theoretic quantity namely the total variation distance is studied as a privacy metric in an information disclosure scenario.

More closely related to our approach, several information leakage metrics have recently emerged that aim to capture the inference abilities of an adversary trying to guess a secret. For instance, [24] proposes to use the probability of correctly guessing the secret as a privacy metric. In [25] a class of tunable loss functions are introduced to capture a range of adversarial objectives, e.g., refining a belief or guessing the most likely value for the secret. Other methods include posing the privacy problem as a hypothesis test, e.g., in [26]. It is worth mentioning that some of the proposed privacy metrics (such as mutual information and total variation distance) have no clear operational meaning in the privacy setting, which limits their applicability. A systematic survey of privacy metrics is provided in [27].

Privacy-preserving machine learning. Several centralized and decentralized solutions have been proposed in the literature that provide privacy guarantees in terms of differential privacy. To give a few examples, [28] proposes a collaborative framework for privacy-preserving deep learning where the guarantees of differential privacy are obtained by perturbing the gradients. Another example is [29] where the privacy analysis of gradient perturbations are improved by introducing the *moments accountant* framework. Other methods include privacy-preserving logistic regression [30], [31], support vector machines [32] and empirical risk minimization [33], [34].

C. Outline of the Paper

The rest of the paper is organized as follows: in Section II we will review the definition of maximal leakage and give a short summary of the operation of the PATE framework. In Section III we will present the definition of pointwise conditional maximal leakage, and state a few of its key properties. In Section IV we will present our privacy analysis of the framework and state our results. Section V concludes the paper.

II. Background

Throughout the paper, upper-case letters are used to represent discrete random variables, upper-case calligraphic letters represent their corresponding alphabets and lower-case letters represent the elements of the alphabets. We will use $[[1, n]] = \{1, \dots, n\}$ to denote the set of integers between one and n. Let $A = (A_1, \dots, A_n)$ be a sequence of n elements. We will use the notation $A \setminus A_j$ to denote the sequence of $n-1$ elements obtained by removing the jth element in A for some $j \in [[1, n]]$. Furthermore, we will use $|| \cdot$ to denote the cardinality of a set, and $\log(\cdot)$ to denote the natural logarithm. Finally, all sets considered in this paper are assumed to be finite.

We begin by reviewing a few key concepts.

A. Maximal Leakage

Let X be a random variable representing the data containing sensitive information, and Y be the publicly observed output of a probability kernel $P_{Y|X}$ with input X. Suppose that an adversary observes Y and wishes to guess an arbitrary discrete function of X, denoted by U.

Definition 1 (Maximal leakage [11]): Suppose P_{XY} is a joint distribution defined on the alphabets $\mathcal{X}$ and $\mathcal{Y}$. The maximal leakage from X to Y is defined as

$$\mathcal{L}(X \to Y) := \sup_{U: U - X - Y} \log \frac{\mathbb{P}\left(U = \hat{U}(Y)\right)}{\max_{u \in \mathcal{U}} P_U(u)}, \tag{1}$$

where $\hat{U}$ is the optimal estimator (i.e., MAP estimator) taking values from the same alphabet as U.
Maximal leakage quantifies the maximal gain in the adversary's ability to correctly guess U after observing Y (compared to correctly guessing U with no observations). It is shown in [11, Theorem 1] that for finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, (1) simplifies to

$$\mathcal{L}(X \to Y) = \log \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}: P_X(x) > 0} P_{Y|X}(y \mid x). \tag{2}$$

B. The PATE Framework

PATE [14], [15] is a general framework for privacy-preserving classification of sensitive data. It operates by transferring the knowledge of an ensemble of classifiers, called *teachers*, trained on (disjoint) partitions of the sensitive data to a *student* classifier. More specifically, the PATE framework consists of the following three main components:

Teacher models. A teacher is a classification model trained on one of the disjoint partitions of the sensitive training data, and can use any classification algorithm suited for the task. At inference, each teacher predicts a label independently of others, to which we will refer as that teacher's *vote*. Thus, partitioning data into L sets (and correspondingly training L teachers) produces L primary votes for predicting the label of any new data point.

Aggregation mechanism. To predict the label of a new data point, the aggregation mechanism (i.e., the Report-Noisy-Max mechanism [7]) constructs the histogram of teachers' votes, adds calibrated noise to each of the bins, and outputs the class label with the maximum noisy vote as the final aggregate prediction. Note that the overall privacy guarantees of the framework result from the addition of noise in the aggregation mechanism.

Student model. The student model is trained using a public unlabelled dataset which will be labelled by the teachers' ensemble through the aggregation mechanism. Note that to limit the privacy cost of the overall system, the student must be trained with as few queries to the teachers as possible.

III. Pointwise Conditional Maximal Leakage

In this section, we introduce the notion of pointwise conditional maximal leakage, and state two of its important properties. Recall that maximal leakage is defined in a setup where an adversary wishes to guess an arbitrary discrete function U of the private input data X by observing the output Y. Here, we consider the case where the adversary has some *a priori* knowledge about X. We model this a priori knowledge as the outcome of a random variable, and accordingly define a *conditional* form of maximal leakage. Consider an adversary that knows the outcome of a random variable Z.

Definition 2 (Pointwise conditional maximal leakage): Suppose P_{XYZ} is a joint distribution defined on the alphabets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, and that 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

$$\mathcal{L}(X \to Y | Z = z) := \sup_{U : U - (X,Z) - Y} \log \frac{\mathbb{P}\left(U = \hat{U}(Y, Z = z)\right)}{\mathbb{P}\left(U = \tilde{U}(Z = z)\right)}, \tag{3}$$

where $\hat{U}$ is the optimal estimator of U given Y and $Z = z$, and $\tilde{U}$ is optimal estimator of U given only $Z = z$.

Proposition 3: For finite alphabets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, the pointwise conditional maximal leakage can be expressed as

$$\mathcal{L}(X \to Y | Z = z) = \log \sum_{y \in \mathcal{Y}} \max_{x:\, P_{X|Z}(x|z) > 0} P_{Y|XZ}(y|x, z). \tag{4}$$

The proof is given in Appendix A-A.

Pointwise conditional maximal leakage is an adaptation of conditional maximal leakage proposed in [11] and differs slightly from it. The definition in [11] conditions the leakage on the random variable Z itself, which translates into a maximization over the outcomes of Z in (4). We, on the other hand, are conditioning the leakage directly on the outcomes of Z since we are interested in characterizing the leakage for all outcomes, not just the one with the highest leakage. Moreover, as we will see later, the pointwise definition allows us to obtain a data-dependent bound on the leakage which can be tighter than the data-independent bound. More discussions on the comparison of the two bounds can be found in Section IV-B.

Remark 4: If the Markov chain $Z - X - Y$ holds, (4) becomes

$$\mathcal{L}(X \to Y | Z = z) = \log \sum_{y \in \mathcal{Y}} \max_{x:\, P_{X|Z}(x|z) > 0} P_{Y|X}(y|x). \tag{5}$$

Similarly to [11], we now state two important properties of the pointwise conditional maximal leakage: a data-processing inequality and a composition lemma. These properties will be used in the next section to analyze the entrywise information leakage of the PATE framework.

Lemma 5 (Composition): If the Markov chain $Y_1 - (X, Z) - Y_2$ holds, then,

$$\begin{aligned} &\mathcal{L}(X \to (Y_1, Y_2) \mid Z = z) \\ &\leq \mathcal{L}(X \to Y_1 \mid Z = z) + \mathcal{L}(X \to Y_2 \mid Z = z). \end{aligned} \tag{6}$$

More generally, for $k \geq 1$ it holds that

$$\begin{aligned} &\mathcal{L}(X \to (Y_1, \ldots, Y_k) \mid Z = z) \\ &\leq \mathcal{L}(X \to Y_1 \mid Z = z) + \ldots + \mathcal{L}(X \to Y_k \mid Z = z). \end{aligned} \tag{7}$$

Lemma 5 states that the information leaked to multiple observations is upper bounded by the sum of the information leaked through each of the observations.

Lemma 6 (Data-processing inequality): If the Markov chain $(X, Z) - Y_1 - Y_2$ holds, then,

$$\begin{aligned} &\mathcal{L}(X \to Y_2 \mid Z = z) \\ &\leq \min\{\mathcal{L}(X \to Y_1 \mid Z = z), \mathcal{L}(Y_1 \to Y_2 \mid Z = z)\}. \end{aligned} \tag{8}$$

Lemma 6 states that all processing of the output can only decrease the information leakage. Further, it allows us to upper bound the end-to-end leakage of a complex mechanism in terms of the leakages of its smaller intermediate mechanisms. The proofs of Lemma 5 and Lemma 6 are given in Appendix A-B and A-C, respectively.

IV. Information Leakage Analysis of PATE

In this section, we will use the pointwise conditional maximal leakage to measure the information leaking about individual data entries in the PATE framework. We will begin by describing our system model in Section IV-A. Then, in Section IV-B we will first prove that increased consensus among teachers in answering queries induces a lower privacy cost (i.e., the privacy-accuracy synergy), and then, state bounds on the entrywise leakage when noise with Laplace distribution is used in the aggregation.

A. System Model

Suppose $d = ((x_1, y_1), \ldots, (x_n, y_n)) \in \mathcal{X}^n \times \mathcal{Y}^n$ represents the training data where $\mathcal{X}$ is the arbitrary but finite domain set and $\mathcal{Y} = [[1, m]]$ is the label set. The pairs (x_i, y_i) are sampled independently according to some distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, i.e., $D \sim \mathcal{P}^n$. We use the training data d to train L teachers for a classification task with $m \geq 2$ classes in the PATE framework. Let $(d^{(1)}, \ldots, d^{(L)})$ represent a disjoint partitioning of the training set such that $d^{(i)} \neq \emptyset$ for all $i \in [[1, L]]$, $\bigcup_{i=1}^{L} d^{(i)} = d$ and $d^{(i)} \cap d^{(j)} = \emptyset$ for all $i \neq j$. Each partition $d^{(i)}$ is used to train a teacher model $f_i : \mathcal{X} \to [[1, m]]$. This results in a total of L teacher models, classifying queries independently of each other.

The student model is trained using a public and unlabelled dataset, which will be labelled by the teachers ensemble in a privacy-preserving manner. Let $(x'_1, \ldots, x'_k) \in \mathcal{X}^k$ be the independently sampled unlabelled dataset and suppose that the student queries the ensemble about the label of x'_i. Each teacher separately predicts a label for x'_i, referred to as a *vote*. Let $\upsilon(x'_i) = (\upsilon_1(x'_i), \ldots, \upsilon_m(x'_i))$ be the histogram of teachers' votes, where $\upsilon_j(x'_i) = ||\{l : l \in [[1, L]], f_l(x'_i) = j\}$ corresponds to the number of teachers who classified x'_i as belonging to class j. Note that $\sum_{j=1}^{m} \upsilon_j(x'_i) = L$.

The aggregation mechanism in PATE is essentially the *Report-Noisy-Max mechanism* [7] which operates by adding i.i.d. noise samples to the bins of the votes' histogram, and returning the class label with the highest (noisy) value. Let Lap(b) denote the Laplace distribution with location 0 and scale b. Suppose $N = (N_1, \ldots, N_m)$ is a sequence of i.i.d. Laplace random variables, where $N_j \sim \text{Lap}(\frac{1}{\gamma})$ for $j \in [[1, m]]$ represents the noise added to the jth bin. Note that γ determines the dispersion of the noise, and thus, affects the privacy guarantees of the system. Roughly speaking, smaller values of γ correspond to larger noise, and in turn, stronger

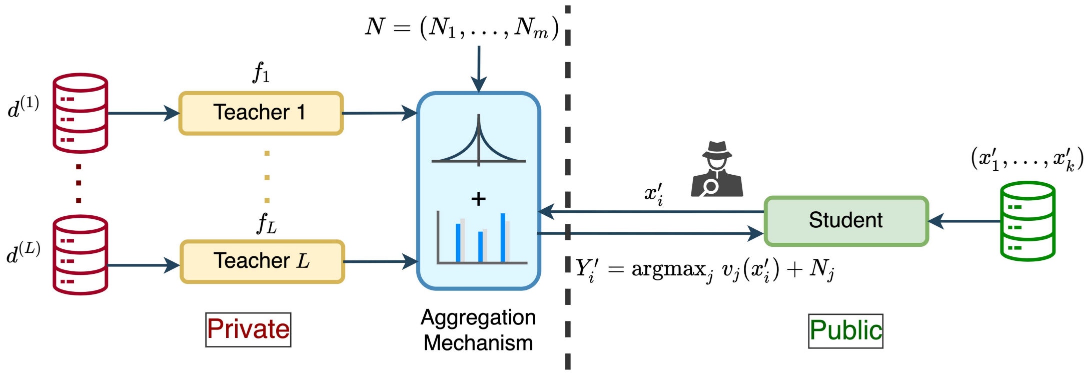

Fig. 1. PATE system model [14]: each partition of the sensitive training data is used to train a teacher. A student model is then trained using a public data-set labelled by the noise-perturbed predictions of the teachers. An adversary who knows all the data-entries except for D^* is trying to guess D^* by observing teachers' responses to queries made by the student.

privacy guarantees. Finally, let $Y'_i = \arg\max_j v_j(x'_i) + N_j$ be the random variable denoting the predicted label for x'_i returned by the aggregation mechanism. Labelling the entire dataset $(x'_1, \dots, x'_k)$ produces k such predictions, each of which entailing a privacy cost. The system model is depicted in Figure 1.

B. Measuring the Entrywise Information Leakage

In this section, we will lay out the details of how we quantify membership privacy through measuring the information leaking about individual data entries in the training set using the notion of pointwise conditional maximal leakage. In order to evaluate the entrywise leakage, let us consider the following scenario: assume an adversary knows the values of all the entries in the teachers' training set (i.e., the private training set) except for a single entry denoted by $D^* = (X^*, Y^*)$. The adversary tries to guess the value of D^* (or any arbitrary discrete function of it) by observing the queries made by the student and their corresponding labels returned by the aggregation mechanism. Clearly, in this setup, observations leak information only about the unknown entry D^* since the adversary already knows all the other entries.

Now, suppose (1) the adversary has perfect knowledge of the algorithms used to train each teacher, and that (2) the training is done deterministically. That is, we will assume that all classification algorithms and the resulting teacher models (i.e., predictors) are deterministic. Note that the first assumption allows us to remain very conservative about the capabilities of the adversary in order to derive privacy guarantees that remain valid even against highly knowledgeable adversaries. Furthermore, we are using the second assumption to consider a scenario in which the training leaks a lot of information about D^*, and the overall privacy guarantees stem only from the aggregation mechanism. As such, our privacy analysis remains valid for all PATE structures regardless of how the teachers are trained, or what classification algorithms are used.

It follows naturally from the previous assumptions that, in principle, the adversary knows all the votes except for the vote of the teacher whose training partition includes D^*. Note that we are considering a general setup in which any single data entry can arbitrarily affect the vote of its teacher, resulting in observations which are highly informative for inferring the data entry of interest (as an extreme example, consider a teacher whose vote depends only on D^*). In other words, if the adversary can already predict the last vote there is no information left to be leaked.

Based on the scenario described, let $D^- = D \setminus D^*$ be the random vector representing the portion of the training set known to the adversary, and let $V^-(x'_i) = (V^-_1(x'_i), \dots, V^-_m(x'_i))$ be the random variable representing the histogram of the known votes for input x'_i. Note that $\sum_{j=1}^m V^-_j(x'_i) = L - 1$ for all $x'_i \in \mathcal{X}$. For simplicity, let $Y' = (Y'_1, \dots, Y'_k)$ denote the sequence of random variables representing the predicted labels for the queries $(x'_1, \dots, x'_k)$. We are interested in quantifying the information leaking about D^* to Y' given that the adversary knows d^- (i.e., the outcome of D^-). We have

$$\begin{aligned}\mathcal{L}(D^* \to Y' \mid D^- = d^-) &= \log \sum_{y' \in \mathcal{Y}^k} \max_{\substack{d^* \in \mathcal{X}\times\mathcal{Y}:\\ \mathbb{P}(d^*|d^-)>0}} \mathbb{P}(y' \mid d^*, d^-)\\ &= \log \sum_{y' \in \mathcal{Y}^k} \max_{\substack{d \in \mathcal{X}^n\times\mathcal{Y}^n:\\ \mathbb{P}(d|d^-)>0}} \mathbb{P}(y' | d)\\ &\overset{\text{(a)}}{=} \mathcal{L}(D \to Y' \mid D^- = d^-), \qquad (9)\end{aligned}$$

where (a) follows from (5) since the Markov chain $D^- - D - Y'$ holds. Using Lemma 5 we can upper bound the information leaked through multiple queries by writing

$$\mathcal{L}(D \to Y' \,|\, D^- = d^-) \le \sum_{i=1}^{k} \mathcal{L}(D \to Y'_i \,|\, D^- = d^-), \qquad (10)$$

that is, the information leaked to the output of multiple queries is upper bounded by the sum of the information leaked through individual queries. Further, using Lemma 6 we can upper bound the information leaked to the output of a single query as

$$\begin{aligned}&\mathcal{L}(D \to Y'_i \mid D^- = d^-) \le \min\{\mathcal{L}(D \to V(x'_i) \mid D^- = d^-),\\ &\mathcal{L}(V(x'_i) \to Y'_i \mid D^- = d^-)\}, \qquad (11)\end{aligned}$$

i.e., the information leaked to the output of a single query is upper bounded by the smallest of the information leaked through the training and the information leaked through the aggregation mechanism.

As we do not want to make any assumptions about how privately the teachers are trained, we now turn to evaluating the information leaked through the aggregation mechanism. Let $\delta_j = (0, \dots, 0, 1, 0, \dots, 0)$ be a sequence with all elements equal to 0, except for the jth element which equals 1. We will use δ_j to represent a single vote for class j. Then, we have

$$\begin{aligned}\mathcal{L}(V(x_i') \to Y_i' \,|\, D^- = d^-)&\\ &= \mathcal{L}(V(x_i') \to Y_i' \,|\, V^-(x_i') = v^-)\\ &\overset{\text{(a)}}{=} \log \sum_{j=1}^{m} \max_{\substack{v = v^- + \delta_{j'}:\\ j' \in [[1,m]]}} \mathbb{P}(Y_i' = j \,|\, V(x_i') = v)\\ &\overset{\text{(b)}}{=} \log \sum_{j=1}^{m} \mathbb{P}(Y_i' = j \,|\, V(x_i') = v^- + \delta_j), \qquad (12)\end{aligned}$$

where (a) follows from (5), and (b) follows from the fact that the probability of outputting class j is maximized when the last vote (i.e., the vote of the teacher whose training partition includes D^*) is placed for class j.

1) The Privacy-Accuracy Synergy: Now, we will evaluate the leakage of the aggregation mechanism as described by (12) using ideas from majorization theory [16], [17] and assuming that the noise used in the mechanism has a log-concave probability density [18], [19]. Specifically, we will find the v^- maximizing or minimizing (12) for any noise with log-concave probability density.

Definition 7 (Majorization): Consider $p, q \in \mathbb{R}^n$ with non-increasingly ordered elements, i.e., $p_1 \geq p_2 \geq \dots \geq p_n$ and $q_1 \geq q_2 \geq \dots \geq q_n$. We say that p majorizes q, and write $p \succ q$ if

$$\sum_{i=1}^{m} p_i \geq \sum_{i=1}^{m} q_i, \text{ for } m = 1, \dots, n-1 \text{ and } \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i. \qquad (13)$$

Note that majorization only describes a partial ordering. For example, $(4, 4, 1)$ and $(5, 2, 2)$ cannot be compared in terms of majorization. On the other hand, if we define $\mathcal{Q} = \{(q_1, q_2, q_3) \in \mathbb{R}_+^3 : \sum_{i=1}^{3} q_i = 9\}$, then $(3, 3, 3)$ is majorized by all $q \in \mathcal{Q}$ while $(9, 0, 0)$, $(0, 9, 0)$ and $(0, 0, 9)$ majorize all $q \in \mathcal{Q}$.

Definition 8 (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 $\Phi(p) \leq \Phi(q)$.

Definition 9 (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 $\phi : \mathbb{R}^n \to [-\infty, \infty)$.

Note that many commonly used probability density functions (and their corresponding CDFs) are log-concave, such as the Laplace and the Gaussian distributions [18].

Theorem 10: Consider the aggregation mechanism in PATE (i.e., the Report-Noisy-Max mechanism) where the noise has a log-concave probability density. Then, $\mathcal{L}(V(x_i') \to Y_i' \,|\, V^-(x_i') = v^-)$ is Schur-concave in v^-. Thus, assuming that $L - 1$ is divisible by m, $\mathcal{L}(V(x_i') \to Y_i' \,|\, V^-(x_i') = v^-)$ is maximized when

$$v^- = v^-_{max} = \left(\frac{L-1}{m}, \dots, \frac{L-1}{m}\right), \qquad (14)$$

and is minimized when

$$v^- = v^-_{min} = (0, \dots, 0, L-1, 0, \dots, 0) = (L-1)\,\delta_j, \qquad (15)$$

for some $j \in [[1, m]]$.

The proof of the theorem is given in Appendix B.

Remark 11: The Schur-concavity of the entrywise information leakage of the aggregation mechanism $\mathcal{L}(V(x_i') \to Y_i' \,|\, V^-(x_i') = v^-)$ implies that stronger consensus among teachers lowers the amount of information leaked about any individual data entry.

The preceding remark points to one of the main advantages of the PATE framework: increased accuracy of the teacher models results in stronger consensus in predicting the label of a given query, which, in turn, results in stronger privacy guarantees. Note that [14], [15] intuitively come to the same conclusions regarding the synergy between privacy and accuracy for the case of Laplace and Gaussian noise distributions, whereas here we have analytically proved this property and generalized it to the class of log-concave probability densities.

2) Data-Independent Bound: Now, we will apply Theorem 10 to (12) to get a bound on the leakage of the aggregation mechanism with Laplace noise.

Proposition 12: Consider the PATE framework where noise with Laplace distribution is used in the aggregation mechanism. For all v^-, the information leaked to the output of a single query is upper bounded by

$$\mathcal{L}(V(x_i') \to Y_i' \mid V^-(x_i') = v^-) \leq \log(B_1), \qquad (16)$$

where

$$\begin{aligned}B_1 :=\,& (1-m)\, 2^{-m} e^{-\gamma} + e^{\gamma}\left(1 - (1 - \frac{1}{2}e^{-\gamma})^{m}\right)\\ &+ \frac{m}{2}(1 - \frac{1}{2}e^{-\gamma})^{m-1} - \frac{m(m-1)}{4} e^{-\gamma} H(m-2). \qquad (17)\end{aligned}$$

Also, $H(0) := \gamma$ and

$$H(m) := \gamma + \sum_{k=1}^{m} \frac{2^{-k} - (1 - \frac{1}{2}e^{-\gamma})^{k}}{k} \quad \text{for } m \geq 1, \qquad (18)$$

The bound is attained at $v^- = v^-_{max}$ defined in (14).

The proof of this result is given in Appendix C-A. Proposition 12 describes a data-independent bound that holds uniformly for all v^- (and consequently all d^-) but depends on m, the number of classes. It can be verified through simple calculations that the bound is non-decreasing in m. Therefore, by letting m tend to infinity, we get the following simpler bound which holds for all d^- and all $m \geq 2$.

Theorem 13: Consider the setting of Proposition 12. For all d^- and all $m \geq 2$, the information leaked about D^* as a result of labelling a single query is upper bounded by

$$\begin{aligned}\mathcal{L}(D^* \to Y_i' \mid D^- = d^-) &= \mathcal{L}(D \to Y_i' \mid D^- = d^-)\\ &\leq \mathcal{L}(V(x_i') \to Y_i' \mid V^-(x_i') = v^-)\\ &\leq \gamma. \qquad (19)\end{aligned}$$

The proof of this result is given in Appendix C-B.

Note that the bounds stated in Proposition 12 and Theorem 13 give a more accurate characterization of the leakage as consensus among teachers decreases. This is demonstrated in the following example where we calculate the leakage in (12) directly using the conditional probabilities, and compare it with the bounds.

Example 14: Suppose the PATE framework has been implemented with $L = 11$ teachers to classify queries into $m = 4$ classes. Further, suppose that for a given query x'_i, the histogram of teachers' votes is (some permutation of) $\upsilon = (5, 3, 2, 1)$, and that Laplace noise with $\gamma = 0.1$ is used in the aggregation mechanism. Depending on which partition of the training set includes D^*, the adversary has obtained one of the following values: $\upsilon^- \in \{(4, 3, 2, 1), (5, 2, 2, 1), (5, 3, 1, 1), (5, 3, 2, 0)\}$. We can now directly use (12) to calculate the leakage of the aggregation mechanism using the probability density function of the Laplace distribution for each υ^-. For simplicity of notation we define $\mathcal{L}(\upsilon^-) := \mathcal{L}(V(x'_i) \to Y'_i \mid V^-(x'_i) = \upsilon^-)$. Then, we have one of the following four cases:

- $\upsilon^- = (4, 3, 2, 1) \implies \mathcal{L}(\upsilon^-) = 8.50 \times 10^{-2}$.
- $\upsilon^- = (5, 2, 2, 1) \implies \mathcal{L}(\upsilon^-) = 8.40 \times 10^{-2}$.
- $\upsilon^- = (5, 3, 1, 1) \implies \mathcal{L}(\upsilon^-) = 8.37 \times 10^{-2}$.
- $\upsilon^- = (5, 3, 2, 0) \implies \mathcal{L}(\upsilon^-) = 8.35 \times 10^{-2}$.

Therefore, $\mathcal{L}(\upsilon^-) \leq 8.50 \times 10^{-2}$ while Proposition 12 predicts $\mathcal{L}(\upsilon^-) \leq \log(B_1) = 8.61 \times 10^{-2}$ and Theorem 13 predicts $\mathcal{L}(\upsilon^-) \leq 0.1$. Note that due to the Schur-concavity of $\mathcal{L}(\upsilon^-)$ it was already expected that information leakage would be largest for $(4, 3, 2, 1)$, and it would have sufficed to just consider this case. Now, suppose $\upsilon = (3, 3, 3, 2)$. Calculating the leakage using the corresponding conditional probabilities gives $\mathcal{L}(\upsilon^-) \leq 8.58 \times 10^{-2}$, which is closer to the value predicted by Proposition 12 and Theorem 13.

Our final data-independent bound describes the information leaked through multiple queries.

Corollary 15: Consider the setting of Theorem 13. The information leaked about D^* as the result of training a student model on k samples is upper bounded by

$$\mathcal{L}(D^* \to Y' \mid D^- = d^-) \leq k\gamma. \tag{20}$$

This result is a direct consequence of Theorem 13 and Lemma 5, and characterizes the overall information leaked about a single data entry as a result of training a student classifier using k queries to the teachers.

3) Data-Dependent Bound: In the previous section, we presented bounds on the leakage that hold uniformly regardless of the data used in the training. Here, we present a bound that depends on the training data through υ^-.

Proposition 16: Consider the PATE framework where noise with Laplace distribution is used in the aggregation mechanism. Suppose υ^- is sorted in non-increasing order and that the first r coordinates have equal votes, that is, $\upsilon_1^- = \ldots = \upsilon_r^- > \upsilon_{r+1}^- \geq \ldots \geq \upsilon_m^-$ for some $1 \leq r \leq m$. Then, we have

$$\mathcal{L}(V(x'_i) \to Y'_i \mid V^-(x'_i) = \upsilon^-) \leq \log(B_2), \tag{21}$$

where

$$B_2 := r\left(1 - \frac{2 + \gamma\,(\upsilon_1^- + 1 - \upsilon_2^-)}{4\exp\left(\gamma\,(\upsilon_1^- + 1 - \upsilon_2^-)\right)}\right) + \sum_{j=r+1}^{m} \frac{2 + \gamma\,(\upsilon_1^- - 1 - \upsilon_j^-)}{4\exp\left(\gamma\,(\upsilon_1^- - 1 - \upsilon_j^-)\right)}. \tag{22}$$

The proof of this result is given in Appendix C-C.

In practice, in order to calculate the information leaked through a query response, one has to take the minimum of the data-dependent bound in Proposition 16 and the data-independent bound in Proposition 12. Roughly speaking, the data-dependent bound is tighter than the data-independent bound when the teachers have strong agreement over the label of a query. This is illustrated in the numerical example below.

Example 17: Suppose the PATE framework has been implemented for a classification task with $m = 4$ classes and that Laplace noise with $\gamma = 0.1$ is used in the aggregation mechanism. First, consider the case where $L = 11$ and $\upsilon^- = (4, 3, 2, 1)$. Then, $\log(B_1) = 8.61 \times 10^{-2}$ while $\log(B_2) = 6.81 \times 10^{-1}$, so the data-independent bound is much tighter. Now, suppose $L = 101$ and $\upsilon^- = (90, 5, 5, 0)$. Then, $\log(B_2) = 1.05 \times 10^{-3}$, while the data-independent remains as before. Therefore, the data-dependent bound is tighter when there is a strong consensus among teachers.

V. Conclusion

In this paper, we have proposed an approach based on information leakage for quantifying membership privacy. Particularly, we showed that the pointwise conditional maximal leakage, a conditional form of maximal leakage, can be used to measure the information leaking about individual data entries in a dataset. We applied our privacy analysis to PATE and derived novel privacy guarantees for this privacy-preserving classification framework in the form of upper bounds on its entrywise information leakage when the injected noise has Laplace distribution. We also showed that the privacy-accuracy synergy of PATE can be explained by studying the entrywise information leakage of the framework while it was only intuitively justified through the lens of differential privacy.

As our work has taken a step towards gaining a deeper understanding of some underlying privacy principles in the PATE framework, our results can be used in the design of machine learning algorithms that preserve both privacy and utility. For example, we can consider a situation in which we have a fixed privacy budget per query. Then, using the data-dependent bound of Proposition 16, one can adjust the noise parameter γ in order to achieve the budget for each query. We except that this will improve the utility of the system since, for example, less noise will be required when there is a strong consensus over the label of a query. Another potential application is in privacy thresholding schemes where queries which are expensive in terms of privacy will not be answered at all. Once again this method will improve both the privacy and the utility of the system since the expensive queries are precisely those which were not labelled with certainty by the teachers.

Appendix A
Proofs of the Results in Section III

A. Proof of Proposition 3

This result follows readily from [11, Theorem 1] by considering $\mathcal{L}(X' \to Y)$ such that $P_{X'} = P_{X|Z=z}$. Nevertheless, we provide an alternative proof.

Upper Bound: First, we prove the upper bound on $\mathcal{L}(X \to Y \mid Z = z)$. Consider any discrete U satisfying $U - (X, Z) - Y$ and define

$$\mathcal{L}_U(X \to Y \mid Z = z) := \log \frac{\mathbb{P}\left(U = \hat{U}(Y, Z = z)\right)}{\mathbb{P}\left(U = \tilde{U}(Z = z)\right)}, \tag{23}$$

where $\hat{U}$ and $\tilde{U}$ are MAP estimators of U. Then, $\mathcal{L}(X \to Y \mid Z = z) = \sup_{U : U-(X,Z)-Y} \mathcal{L}_U(X \to Y \mid Z = z)$.

For each $z \in \mathcal{Z}$, define $\mathcal{U}_z := \{u : P_{U|Z}(u \mid z) > 0\}$. The two probabilities in $\mathcal{L}_U(X \to Y \mid Z = z)$ are

$$\mathbb{P}\left(U = \tilde{U}(Z = z)\right) = \max_{u \in \mathcal{U}_z} P_{U|Z}(u \mid z), \tag{24}$$

and

$$\begin{aligned}
&\mathbb{P}\left(U = \hat{U}(Y, Z = z)\right) \\
&= \sum_{y \in \mathcal{Y}} \max_{u \in \mathcal{U}_z} P_{UY|Z}(u, y \mid z) \\
&= \sum_{y \in \mathcal{Y}} \max_{u \in \mathcal{U}_z} \sum_{x : P_{X|Z}(x|z) > 0} P_{U|XZ}(u \mid x, z) P_{Y|XZ}(y \mid x, z) P_{X|Z}(x \mid z) \\
&\leq \sum_{y \in \mathcal{Y}} \left(\max_{x' : P_{X|Z}(x'|z) > 0} P_{Y|XZ}(y \mid x, z) \right) \max_{u \in \mathcal{U}_z} \sum_{x : P_{X|Z}(x|z) > 0} P_{UX|Z}(u, x|z) \\
&= \max_{u \in \mathcal{U}_z} P_{U|Z}(u \mid z) \sum_{y \in \mathcal{Y}} \left(\max_{x' : P_{X|Z}(x'|z) > 0} P_{Y|XZ}(y \mid x, z) \right). \qquad (25)
\end{aligned}$$

Thus,

$$\mathcal{L}_U(X \to Y \mid Z = z) \leq \log \sum_{y \in \mathcal{Y}} \max_{x : P_{X|Z}(x|z) > 0} P_{Y|XZ}(y \mid x, z) \tag{26}$$

for all U such that $U - (X, Z) - Y$ holds. Then,

$$\mathcal{L}(X \to Y \mid Z = z) \leq \log \sum_{y \in \mathcal{Y}} \max_{x : P_{X|Z}(x|z) > 0} P_{Y|XZ}(y|x, z). \tag{27}$$

Lower bound: To prove the lower bound on $\mathcal{L}(X \to Y \mid Z = z)$, we will consider a discrete U for which $\mathcal{L}_U(X \to Y \mid Z = z)$ attains the bound. We fix a U' such that $U' - (X, Z) - Y$ holds and $H(X \mid U') = 0$, that is, the value of X is completely determined by the value of U'. Further, we assume that $U' \mid Z = z$ is uniformly distributed, i.e., $P_{U'|Z}(u \mid z) = \frac{1}{|\mathcal{U}_z|}$ for all $z \in \mathcal{Z}$ and $u \in \mathcal{U}_z$. Then,

$$\begin{aligned}
&\mathcal{L}_{U'}(X \to Y \mid Z = z) \\
&= \log \sum_{y \in \mathcal{Y}} \frac{\max_{u \in \mathcal{U}_z} P_{U'|Z}(u \mid z) P_{Y|U'Z}(y \mid u, z)}{\max_{u \in \mathcal{U}_z} P_{U'|Z}(u \mid z)} \\
&= \log \sum_{y \in \mathcal{Y}} \frac{\max_{u \in \mathcal{U}_z} P_{U'|Z}(u \mid z) P_{Y|XZ}(y \mid x, z)}{\max_{u \in \mathcal{U}_z} P_{U'|Z}(u \mid z)} \\
&= \log \sum_{y \in \mathcal{Y}} \frac{\frac{1}{|\mathcal{U}_z|} \max_{x : P_{X|Z}(x|z) > 0} P_{Y|XZ}(y \mid x, z)}{\frac{1}{|\mathcal{U}_z|}} \\
&= \log \sum_{y \in \mathcal{Y}} \max_{x : P_{X|Z}(x|z) > 0} P_{Y|XZ}(y|x, z). \qquad (28)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{L}(X \to Y \mid Z = z) &= \sup_{U : U-(X,Z)-Y} \mathcal{L}_U(X \to Y \mid Z = z) \\
&\geq \mathcal{L}_{U'}(X \to Y \mid Z = z) \\
&= \log \sum_{y \in \mathcal{Y}} \max_{x : P_{X|Z}(x|z) > 0} P_{Y|XZ}(y|x, z). \qquad (29)
\end{aligned}$$

Hence, from (27) and (29) it follows that

$$\mathcal{L}(X \to Y \mid Z = z) = \log \sum_{y \in \mathcal{Y}} \max_{x : P_{X|Z}(x|z) > 0} P_{Y|XZ}(y|x, z). \tag{30}$$

B. Proof of Lemma 5

Consider the Markov chain $Y_1 - (X, Z) - Y_2$. Then,

$$\begin{aligned}
&\mathcal{L}(X \to (Y_1, Y_2) \mid Z = z) - \mathcal{L}(X \to Y_1 \mid Z = z) \\
&= \log \frac{\sum_{y_1, y_2} \max_{x : P_{X|Z}(x|z) > 0} P_{Y_1 \ Y_2|XZ}(y_1, y_2 \mid x, z)}{\sum_{y_1} \max_{x : P_{X|Z}(x|z) > 0} P_{Y_1|XZ}(y_1 \mid x, z)} \\
&= \log \sum_{y_2} \frac{\sum_{y_1} \max_x P_{Y_1|XZ}(y_1 \mid x, z) P_{Y_2|XZ}(y_2 \mid x, z)}{\sum_{y_1} \max_{x : P_{X|Z}(x|z) > 0} P_{Y_1|XZ}(y_1 \mid x, z)} \\
&\leq \log \sum_{y_2} \frac{\sum_{y_1} \max_x P_{Y_1|XZ}(y_1 \mid x, z) \left(\max_{x'} P_{Y_2|XZ}(y_2 \mid x', z)\right)}{\sum_{y_1} \max_{x : P_{X|Z}(x|z) > 0} P_{Y_1|XZ}(y_1 \mid x, z)} \\
&= \log \sum_{y_2} \max_{x' : P_{X|Z}(x'|z) > 0} P_{Y_2|XZ}(y_2 \mid x', z) \\
&= \mathcal{L}(X \to Y_2 \mid Z = z). \qquad (31)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathcal{L}(X \to (Y_1, Y_2) \mid Z = z) \\
&\leq \mathcal{L}(X \to Y_1 \mid Z = z) + \mathcal{L}(X \to Y_2 \mid Z = z). \qquad (32)
\end{aligned}$$

C. Proof of Lemma 6

Our proof follows the same reasoning as the proof of [11, Lemma 1]. For all discrete U satisfying $U - (X, Z) - Y_1 - Y_2$ it holds that

$$\mathcal{L}_U(X \to Y_2 \,|\, Z = z) \le \mathcal{L}_U(X \to Y_1 \,|\, Z = z), \tag{33}$$

where $\mathcal{L}_U$ is defined in (23). Therefore,

$$\begin{aligned}\mathcal{L}(X \to Y_2 \mid Z = z) &= \sup_{U:\, U-(X,Z)-Y_1-Y_2} \mathcal{L}_U(X \to Y_2 \mid Z = z)\\ &\le \sup_{U:\, U-(X,Z)-Y_1} \mathcal{L}_U(X \to Y_1 \mid Z = z)\\ &= \mathcal{L}(X \to Y_1 \mid Z = z).\end{aligned} \tag{34}$$

Similarly,

$$\begin{aligned}\mathcal{L}(X \to Y_2 \,|\, Z = z) &= \sup_{U:\, U-(X,Z)-Y_1-Y_2} \mathcal{L}_U(X \to Y_2 \,|\, Z = z)\\ &\le \sup_{U:\, U-Z-Y_1-Y_2} \mathcal{L}_U(Y_1 \to Y_2 \,|\, Z = z)\\ &= \mathcal{L}(Y_1 \to Y_2 \,|\, Z = z).\end{aligned} \tag{35}$$

Thus,

$$\begin{aligned}\mathcal{L}(X \to Y_2 \,|\, Z = z) \le \min\{&\mathcal{L}(X \to Y_1 \,|\, Z = z),\\ &\mathcal{L}(Y_1 \to Y_2 \,|\, Z = z)\}.\end{aligned} \tag{36}$$

Appendix B
Proof of Theorem 10

Before stating the proof, let us recall some concepts/results from majorization theory.

Definition 18 (Symmetric function): Let $x = (x_1, \ldots, x_n) \in \mathcal{I}^n \subset \mathbb{R}^n$ and consider a real-valued function $\Phi : \mathcal{I}^n \to \mathbb{R}$. The function $\Phi(x)$ is said to be symmetric if x can be arbitrarily permuted without changing the value of $\Phi(x)$.

Lemma 19 (Schur's condition): Let $x = (x_1, \ldots, x_n) \in \mathcal{I}^n \subset \mathbb{R}^n$ and consider a continuously differentiable function $\Phi : \mathcal{I}^n \to \mathbb{R}$. $\Phi(x)$ is Schur-concave on $\mathcal{I}^n$ if and only if it is symmetric on $\mathcal{I}^n$ and

$$(x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) \le 0 \quad \text{for all} \quad 1 \le i, j \le n. \tag{37}$$

Since $\Phi(x)$ must be symmetric, it is sufficient to verify the reduced condition

$$(x_1 - x_2)\left(\frac{\partial f}{\partial x_1} - \frac{\partial f}{\partial x_2}\right) \le 0. \tag{38}$$

Proposition 20 ([17, Theorem 2.21]): Let $x = (x_1, \ldots, x_n) \in \mathbb{R}_+^n$ and let $f : \mathbb{R}_+^n \to \mathbb{R}_+$ be a Schur-concave function. Consider the following problems

$$\max_x f(x) \quad \text{such that} \quad \sum_{i=1}^{n} x_i = S, \tag{39}$$

and

$$\min_x f(x) \quad \text{such that} \quad \sum_{i=1}^{n} x_i = S. \tag{40}$$

Then, the global maximum is achieved by

$$x_{max} = \frac{S}{n}(1, \ldots, 1), \tag{41}$$

and the global minimum is achieved by

$$x_{min} = (0, \ldots, 0, S, 0, \ldots, 0). \tag{42}$$

We now prove that the entrywise information leakage of the aggregation mechanism is Schur-concave when the injected noise has a log-concave probability density. In order to simplify the proof, we will assume that the elements of υ^- (i.e., the histogram of known votes) can take non-negative real values. The results of the proof, however, will be readily applicable to histograms of non-negative integers.

Using (12) we define

$$f_j(\upsilon^-) := \mathbb{P}(Y_i' = j \mid V(x_i') = \upsilon^- + \delta_j), \tag{43}$$

where $\delta_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ represents a single vote for class j. Then,

$$\begin{aligned}\mathcal{L}(V(x_i') \to Y_i' \mid V^-(x_i') = \upsilon^-) &= \log \sum_{j=1}^{m} f_j(\upsilon^-)\\ &= \log f(\upsilon^-),\end{aligned} \tag{44}$$

where $f(\upsilon^-) = \sum_{j=1}^{m} f_j(\upsilon^-)$. It is clear from (44) that the leakage does not depend on the order of elements in υ^-, thus $\mathcal{L}(V(x_i') \to Y_i' \,|\, V^-(x_i') = \upsilon^-)$ is symmetric. Moreover, according to [16, 3.B.1], the composition of an increasing function and a Schur-concave function remains Schur-concave. Since $\log(\cdot)$ is an increasing function, to prove the Schur-concavity of the entrywise leakage we only need to verify Schur's condition for $f(\upsilon^-)$.

Without loss of generality assume that $\upsilon^- = (\upsilon_1^-, \ldots, \upsilon_m^-)$ is non-increasingly ordered, i.e., $\upsilon_1^- \ge \ldots \ge \upsilon_m^-$. Let $N = (N_1, \ldots, N_m)$ denote the tuple of noise, where the elements are independent, identically distributed and have a log-concave probability density. We write

$$\begin{aligned}&f_j(\upsilon^-)\\ &= \mathbb{P}(Y_i' = j \,|\, V(x_i') = \upsilon^- + \delta_j)\\ &= \mathbb{P}\{\upsilon_j^- + N_j + 1 > \upsilon_1^- + N_1, \ldots, \upsilon_j^- + N_j + 1 > \upsilon_m^- + N_m\}\\ &= \int_{-\infty}^{\infty} \left[\prod_{\substack{l=1\\ l \ne j}}^{m} \mathbb{P}\{N_l < (\upsilon_j^- - \upsilon_l^- + t + 1)\}\right] g(t) dt\\ &= \int_{-\infty}^{\infty} \left[\prod_{\substack{l=1\\ l \ne j}}^{m} G(\upsilon_j^- - \upsilon_l^- + t + 1)\right] g(t)\, dt,\end{aligned} \tag{45}$$

where $g(t)$ is the probability density function of N_j and $G(t) = \int_{-\infty}^{t} g(t')\, dt'$ is its corresponding cumulative distribution function. According to [19, Proposition 1], if g is log-concave, then G is also log-concave. We now check Schur's condition by writing

$$\frac{\partial f(\upsilon^-)}{\partial \upsilon_1^-} - \frac{\partial f(\upsilon^-)}{\partial \upsilon_2^-} = \sum_{j=1}^{m} \frac{\partial f_j(\upsilon^-)}{\partial \upsilon_1^-} - \frac{\partial f_j(\upsilon^-)}{\partial \upsilon_2^-}, \tag{46}$$

where we have one of the following three cases:
if $j = 1$, then,

$$\begin{aligned}&\frac{\partial f_1(v^-)}{\partial v_1^-} - \frac{\partial f_1(v^-)}{\partial v_2^-}\\ &= \int_{-\infty}^{\infty} \left[\sum_{l=2}^{m} g(v_1^- - v_l^- + t + 1) \prod_{\substack{k=2\\k\neq l}}^{m} G(v_1^- - v_k^- + t + 1) \right]\\ &\quad \times g(t)\, dt\\ &\quad - \int_{-\infty}^{\infty} [-g(v_1^- - v_2^- + t + 1)] \left[\prod_{l=3}^{m} G(v_1^- - v_l^- + t + 1) \right]\\ &\quad \times g(t)\, dt, \end{aligned} \tag{47}$$

if $j = 2$, then,

$$\begin{aligned}&\frac{\partial f_2(v^-)}{\partial v_1^-} - \frac{\partial f_2(v^-)}{\partial v_2^-}\\ &= \int_{-\infty}^{\infty} [-g(v_2^- - v_1^- + t + 1)] \left[\prod_{l=3}^{m} G(v_2^- - v_l^- + t + 1) \right]\\ &\quad \times g(t)\, dt\\ &\quad - \int_{-\infty}^{\infty} \left[\sum_{\substack{l=1\\l\neq 2}}^{m} g(v_2^- - v_l^- + t + 1) \prod_{\substack{k=1\\k\neq 2,l}}^{m} G(v_2^- - v_k^- + t + 1) \right]\\ &\quad \times g(t)\, dt, \end{aligned} \tag{48}$$

and if $j \neq 1, 2$, then,

$$\begin{aligned}&\frac{\partial f_j(v^-)}{\partial v_1^-} - \frac{\partial f_j(v^-)}{\partial v_2^-}\\ &= \int_{-\infty}^{\infty} -\Big[g(v_j^- - v_1^- + t + 1) G(v_j^- - v_2^- + t + 1)\\ &\qquad + g(v_j^- - v_2^- + t + 1)\, G(v_j^- - v_1^- + t + 1) \Big]\\ &\qquad \cdot \left[\prod_{\substack{l=3\\l\neq j}}^{m} G(v_j^- - v_l^- + t + 1) \right] g(t)\, dt. \end{aligned} \tag{49}$$

Then,

$$\begin{aligned}&\frac{\partial f(v^-)}{\partial v_1^-} - \frac{\partial f(v^-)}{\partial v_2^-}\\ &= A_1 - A_2 + \sum_{j=3}^{m} B_{(1,j)} - B_{(2,j)} + B_{(3,j)} - B_{(4,j)}, \end{aligned} \tag{50}$$

where

$$\begin{aligned}A_1 &= 2\int_{-\infty}^{\infty} g(v_1^- - v_2^- + t + 1) \left[\prod_{k=3}^{m} G(v_1^- - v_k^- + t + 1) \right]\\ &\quad \times g(t) dt, \end{aligned} \tag{51}$$

$$A_2 = 2\int_{-\infty}^{\infty} g(v_2^- - v_1^- + t + 1) \left[\prod_{k=3}^{m} G(v_2^- - v_k^- + t + 1) \right] g(t) dt, \tag{52}$$

$$\begin{aligned}B_{(1,j)} &= \int_{-\infty}^{\infty} g(v_1^- - v_j^- + t + 1) G(v_1^- - v_2^- + t + 1)\\ &\quad \cdot \left[\prod_{\substack{k=3\\k\neq j}}^{m} G(v_1^- - v_k^- + t + 1) \right] g(t) dt, \end{aligned} \tag{53}$$

$$\begin{aligned}B_{(2,j)} &= \int_{-\infty}^{\infty} g(v_j^- - v_1^- + t + 1) G(v_j^- - v_2^- + t + 1)\\ &\quad \cdot \left[\prod_{\substack{k=3\\k\neq j}}^{m} G(v_j^- - v_k^- + t + 1) \right] g(t) dt, \end{aligned} \tag{54}$$

$$\begin{aligned}B_{(3,j)} &= \int_{-\infty}^{\infty} g(v_j^- - v_2^- + t + 1) G(v_j^- - v_1^- + t + 1)\\ &\quad \times \left[\prod_{\substack{k=3\\k\neq j}}^{m} G(v_j^- - v_k^- + t + 1) \right] g(t) dt, \end{aligned} \tag{55}$$

$$\begin{aligned}B_{(4,j)} &= \int_{-\infty}^{\infty} g(v_2^- - v_j^- + t + 1) G(v_2^- - v_1^- + t + 1)\\ &\quad \times \left[\prod_{\substack{k=3\\k\neq j}}^{m} G(v_2^- - v_k^- + t + 1) \right] g(t) dt. \end{aligned} \tag{56}$$

We now show that both $A_1 - A_2$ and $B_{(1,j)} - B_{(2,j)} + B_{(3,j)} - B_{(4,j)}$ are non-positive. However, let us first recall some properties of log-concave functions.

Proposition 21 ([19, Lemma 1]): Consider $g : \mathbb{R} \to \mathbb{R}_+$ and suppose that $\{x : g(x) > 0\} = (a, b)$. Then, $g(x)$ is log-concave if and only if for all $a < x_1 \leq x_2 < b$ and all $\delta \geq 0$ it holds that

$$g(x_1 + \delta) g(x_2) \geq g(x_1) g(x_2 + \delta). \tag{57}$$

Proposition 22 ([18, Remark 2]): Suppose $g : \mathbb{R} \to \mathbb{R}_+$ is a continuously differentiable function and let $\{x : g(x) > 0\} = (a, b)$. Then, $g(x)$ is log-concave if and only if $\frac{g'(x)}{g(x)}$ is a non-increasing function of x in (a, b).

We now prove that $A_1 - A_2 \leq 0$. By a change of variable in A_1 we let $\upsilon_1^- - \upsilon_2^- + t = u$. Then,

$$A_1 - A_2 = \int_{-\infty}^{\infty} \prod_{k=3}^{m} G(u + \upsilon_2^- - \upsilon_k^- + 1) \cdot \\ \times \left[g(u+1)g(u+\upsilon_2^- - \upsilon_1^-) - g(u)g(u+\upsilon_2^- - \upsilon_1^- + 1)\right] du. \tag{58}$$

We now apply Proposition 21 to the preceding equation by noting that $u \geq u + \upsilon_2^- - \upsilon_1^-$ (due to the non-increasing order of the elements in υ^-), and write

$$g(u+1)g(u + \upsilon_2^- - \upsilon_1^-) - g(u)g(u+\upsilon_2^- - \upsilon_1^- + 1) \leq 0. \tag{59}$$

Since $\prod_{k=3}^{m} G(u + \upsilon_2^- - \upsilon_k^- + 1) \geq 0$, we conclude that

$$A_1 - A_2 \leq 0. \tag{60}$$

Similarly, Proposition 21 and Proposition 22 can be used to show that $B_{(1,j)} - B_{(2,j)} + B_{(3,j)} - B_{(4,j)} \leq 0$ for all $j = 3, \ldots, m$. Therefore, we have verified Schur's condition for $f(\upsilon^-)$, and conclude that $\mathcal{L}(V(x_i') \to Y_i' \mid V^-(x_i') = \upsilon^-)$ is Schur-concave. Finally, by Proposition 20, the entrywise leakage is maximized by

$$\upsilon^- = \upsilon_{max}^- = \left(\frac{L-1}{m}, \ldots, \frac{L-1}{m}\right), \tag{61}$$

and is minimized by

$$\upsilon^- = \upsilon_{min}^- = (0, \ldots, 0, L-1, 0, \ldots, 0) = (L-1)\,\delta_j, \tag{62}$$

for each $j \in [[1, m]]$.

Appendix C
Proofs for the Leakage With Laplace Noise

A. Proof of Proposition 12

Let $N = (N_1, \ldots, N_m)$ be the sequence of i.i.d. Laplace random variables, where $N_j \sim \mathrm{Lap}(\frac{1}{\gamma})$ for all $j \in [[1, m]]$. To find an upper bound on the leakage, we will apply Theorem 10 and calculate $\mathcal{L}(V(x_i') \to Y_i' \mid V^-(x_i') = \upsilon^-)$ for $\upsilon^- = \upsilon_{max}^- = (\frac{L-1}{m}, \ldots, \frac{L-1}{m})$. We write

$$\mathcal{L}(V(x_i') \to Y_i' \mid V^-(x_i') = \upsilon_{max}^-) \\ = \log \sum_{j=1}^{m} \mathbb{P}(Y_i' = j \mid V(x_i') = \upsilon_{max}^- + \delta_j), \tag{63}$$

where

$$\mathbb{P}(Y_i' = j \mid V(x_i') = \upsilon_{max}^- + \delta_j) \\ = \mathbb{P}\{N_j + 1 > N_1, \ldots, N_j + 1 > N_m\} \\ = \int_{-\infty}^{\infty} \left[\prod_{\substack{l=1 \\ l \neq j}}^{m} \mathbb{P}\{N_l < (t+1)\}\right] \cdot \frac{\gamma}{2} e^{-\gamma |t|} dt, \tag{64}$$

and

$$\mathbb{P}\{N_l < (t+1)\} = \begin{cases} \frac{1}{2} e^{\gamma(t+1)} & t \leq -1, \\ 1 - \frac{1}{2} e^{-\gamma(t+1)} & t \geq -1. \end{cases} \tag{65}$$

Thus, we have

$$\mathbb{P}(Y_i' = j \mid V(x_i') \\ = \upsilon_{max}^- + \delta_j) = \underbrace{\frac{\gamma}{2} \int_{-\infty}^{-1} \left[\frac{1}{2} e^{\gamma(t+1)}\right]^{m-1} \cdot e^{\gamma t} dt}_{A} \\ + \underbrace{\frac{\gamma}{2} \int_{-1}^{0} \left[1 - \frac{1}{2} e^{-\gamma(t+1)}\right]^{m-1} \cdot e^{\gamma t} dt}_{B} \\ + \underbrace{\frac{\gamma}{2} \int_{0}^{\infty} \left[1 - \frac{1}{2} e^{-\gamma(t+1)}\right]^{m-1} \cdot e^{-\gamma t} dt}_{C}. \tag{66}$$

It is straightforward to calculate integrals A and C as

$$A = \frac{2^{-m}}{m} e^{-\gamma} \quad \text{and} \quad C = \frac{1 - \left[1 - \frac{1}{2} e^{-\gamma}\right]^m}{m} e^{\gamma}. \tag{67}$$

Integral B can be written as

$$B = \frac{1}{2}\left(1 - \frac{1}{2} e^{-\gamma}\right)^{m-1} - 2^{-m} e^{-\gamma} \\ - \frac{\gamma(m-1)}{4} e^{-\gamma} \int_{-1}^{0} \left(1 - \frac{1}{2} e^{-\gamma(t+1)}\right)^{m-2} dt. \tag{68}$$

We define

$$H(m) := \gamma \int_{-1}^{0} \left(1 - \frac{1}{2} e^{-\gamma(t+1)}\right)^{m} dt \\ = \gamma \sum_{k=0}^{m} \binom{m}{k} \left(-\frac{1}{2}\right)^k e^{-\gamma k} \int_{-1}^{0} e^{-\gamma k t} dt \\ = \sum_{k=0}^{m} \binom{m}{k} \left(-\frac{1}{2}\right)^k \frac{1}{k} \left(1 - e^{-\gamma k}\right). \tag{69}$$

Using $\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}$ for $m \geq 1$, we get

$$H(m) = \sum_{k=0}^{m-1} \binom{m-1}{k} \left(-\frac{1}{2}\right)^k \frac{1}{k} \left(1 - e^{-\gamma k}\right) \\ + \sum_{k=0}^{m} \binom{m-1}{k-1} \left(-\frac{1}{2}\right)^k \frac{1}{k} \left(1 - e^{-\gamma k}\right) \\ = H(m-1) + \frac{1}{m}\left(2^{-m} - (1 - \frac{1}{2} e^{-\gamma})^m\right), \tag{70}$$

and $H(0) = \gamma$. Thus,

$$H(m) = \begin{cases} \gamma & m = 0, \\ \gamma + \sum_{k=1}^{m} \frac{2^{-k} - (1 - \frac{1}{2} e^{-\gamma})^k}{k} & m \geq 1, \end{cases} \tag{71}$$

Note that $H(m)$ is non-negative and monotonically decreasing in m. Since $\sum_{k=1}^{\infty} \frac{t^k}{k} = \log \frac{1}{1-t}$ for $||t < 1$, we have $\lim_{m \to \infty} H(m) = 0$. Hence, integral B can be written as

$$B = \frac{1}{2}\left(1 - \frac{1}{2} e^{-\gamma}\right)^{m-1} - 2^{-m} e^{-\gamma} - \frac{m-1}{4} e^{-\gamma} H(m-2). \tag{72}$$

Finally, we have

$$\begin{aligned}&\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-)\\ &\quad\le \mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-_{max})\\ &\quad=\log(B_1),\end{aligned} \tag{73}$$

where

$$\begin{aligned}B_1 &:= (1-m)\,2^{-m}e^{-\gamma}+e^{\gamma}\left(1-(1-\frac{1}{2}e^{-\gamma})^{m}\right)\\ &\quad+\frac{m}{2}(1-\frac{1}{2}e^{-\gamma})^{m-1}-\frac{m(m-1)}{4}e^{-\gamma}H(m-2).\end{aligned} \tag{74}$$

B. Proof of Theorem 13

In order to prove the bound, we will show that $k(m) := \exp\left(\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-_{max})[m]\right)$ is concave in m and that

$$\lim_{m\to\infty}\exp\left(\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-_{max})\right)=e^{\gamma}. \tag{75}$$

Since m is an integer, we will check the second-order difference of the leakage with respect to m. The first-order difference is

$$\begin{aligned}\Delta k(m) &= k(m+1)-k(m)\\ &=(1-\frac{1}{2}e^{-\gamma})^{m}-\frac{1}{2}e^{-\gamma}\left(2^{-(m-1)}+mH(m-1)\right),\end{aligned} \tag{76}$$

and the second-order difference is

$$\begin{aligned}\Delta^2\,k(m) &= \Delta k(m+1)-\Delta k(m)\\ &=-\frac{1}{2}e^{-\gamma}H(m)\overset{(a)}{\le}0,\end{aligned} \tag{77}$$

where (77) follows from the fact that $H(m)$ is non-negative. Thus, we have shown that $\exp\left(\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-_{max})\right)$ is concave in m. Furthermore, it is straightforward to verify that (75) holds. Hence, we have

$$\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-_{max})\le\gamma. \tag{78}$$

Finally, we get

$$\begin{aligned}\mathcal{L}(D^*\to Y_i' \mid D^-=d^-)&=\mathcal{L}(D\to Y_i' \mid D^-=d^-)\\ &\le\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-)\\ &\le\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-_{max})\le\gamma.\end{aligned} \tag{79}$$

C. Proof of Proposition 16

Similarly to the proof of Proposition 12, we can write

$$\begin{aligned}&\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-)\\ &\quad=\log\sum_{j=1}^{m}\mathbb{P}(Y_i'=j \mid V(x_i')=\upsilon^-+\delta_j),\end{aligned} \tag{80}$$

where

$$\begin{aligned}&\mathbb{P}(Y_i'=j \mid V(x_i')=\upsilon^-+\delta_j)\\ &=\mathbb{P}\{N_j+\upsilon_j^-+1>N_1+\upsilon_1^-,\ldots,N_j+\upsilon_j^-+1>N_m+\upsilon_m^-\}.\end{aligned} \tag{81}$$

For $1\le j\le r$, we have

$$\begin{aligned}&\mathbb{P}(Y_i'=j \mid V(x_i')=\upsilon^-+\delta_j)\\ &\quad=\mathbb{P}(Y_i'=1 \mid V(x_i')=\upsilon^-+\delta_1)\\ &\quad\le\mathbb{P}\{N_1+\upsilon_1^-+1>N_2+\upsilon_2^-\}\\ &\quad=\mathbb{P}\{N_2-N_1<\upsilon_1^-+1-\upsilon_2^-\},\end{aligned} \tag{82}$$

and for $r+1\le j\le m$, we have

$$\begin{aligned}&\mathbb{P}(Y_i'=j \mid V(x_i')=\upsilon^-+\delta_j)\\ &\quad\le\mathbb{P}\{N_j+\upsilon_j^-+1>N_1+\upsilon_1^-\}\\ &\quad=\mathbb{P}\{N_1-N_j<\upsilon_j^-+1-\upsilon_1^-\}.\end{aligned} \tag{83}$$

It is straightforward to see that the random variable described as the difference of two Lap($\frac{1}{\gamma}$) random variables has the following CDF:

$$\begin{aligned}&\mathbb{P}\{N_1-N_2\le x\}\\ &\quad=\begin{cases}\frac{1}{4}\exp(\gamma x)(2-\gamma x) & x\le 0,\\ 1-\frac{1}{4}\exp(-\gamma x)(2+\gamma x) & x\ge 0.\end{cases}\end{aligned} \tag{84}$$

Then, by noting that $\upsilon_1^-+1-\upsilon_2^->0$ and $\upsilon_j^-+1-\upsilon_1^-\le 0$ for $r+1\le j\le m$, we get

$$\mathcal{L}(V(x_i')\to Y_i' \mid V^-(x_i')=\upsilon^-)\le\log(B_2), \tag{85}$$

where

$$\begin{aligned}B_2 &:= r\left(1-\frac{2+\gamma(\upsilon_1^-+1-\upsilon_2^-)}{4\exp\left(\gamma(\upsilon_1^-+1-\upsilon_2^-)\right)}\right)\\ &\quad+\sum_{j=r+1}^{m}\frac{2+\gamma(\upsilon_1^--1-\upsilon_j^-)}{4\exp\left(\gamma(\upsilon_1^--1-\upsilon_j^-)\right)}.\end{aligned} \tag{86}$$

References

[1] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[2] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, “Rumor identification in microblogging systems based on users’ behavior,” *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 99–108, Sep. 2015.

[3] J. West and M. Bhattacharya, “Intelligent financial fraud detection: A comprehensive review,” *Comput. Secur.*, vol. 57, pp. 47–66, Mar. 2016.

[4] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, “SoK: Security and privacy in machine learning,” in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS P)*, Apr. 2018, pp. 399–414.

[5] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18.

[6] Y. Long *et al.*, “Understanding membership inferences on well-generalized learning models,” 2018, *arXiv:1802.04889*. [Online]. Available: http://arxiv.org/abs/1802.04889

[7] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Proc. Theory Cryptogr. Conf.* Berlin, Germany: Springer, 2006, pp. 265–284.

[9] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptograph. Techn.* Berlin, Germany: Springer, 2006, pp. 486–503.

[10] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Aug. 2017, pp. 263–275.

[11] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[12] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi, "Differential privacy: On the trade-off between utility and information leakage," in *Proc. Int. Workshop Formal Aspects Secur. Trust*. Berlin, Germany: Springer, 2011, pp. 39–54.

[13] J. Liao, L. Sankar, O. Kosut, and F. P. Calmon, "Robustness of maximal α-Leakage to side information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 642–646.

[14] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *Proc. ICLR*, 2017, pp. 1–16.

[15] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and U. Erlingsson, "Scalable private learning with pate," in *Proc. ICLR*, 2018, pp. 1–34.

[16] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and its Applications*, vol. 143. New York, NY, USA: Academic, 1979.

[17] E. Jorswieck and H. Boche, *Majorization and Matrix-Monotone Functions in Wireless Communications*, vol. 3. Boston, MA, USA: Now, 2007.

[18] M. Bagnoli and T. Bergstrom, "Log-concave probability and its applications," *Econ. Theory*, vol. 26, no. 2, pp. 445–469, Aug. 2005.

[19] M. Y. An, "Log-concave probability distributions: Theory and statistical testing," Dept. Econ., Duke Univ., Durham, NC, USA, Work. Paper 3-95, 1997.

[20] V. Prabhakaran and K. Ramchandran, "On secure distributed source coding," in *Proc. IEEE Inf. Theory Workshop*, Sep. 2007, pp. 442–447.

[21] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 6, pp. 838–852, Jun. 2013.

[22] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy, and mutual-information privacy," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 5018–5029, Sep. 2016.

[23] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 594–603, 2020.

[24] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Privacy-aware guessing efficiency," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 754–758.

[25] J. Liao, O. Kosut, L. Sankar, and F. du Pin Calmon, "Tunable measures for information leakage and applications to privacy-utility tradeoffs," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8043–8066, Dec. 2019.

[26] Z. Li, T. J. Oechtering, and D. Gündüz, "Privacy against a hypothesis testing adversary," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 6, pp. 1567–1581, Jun. 2019.

[27] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey," *ACM Comput. Surveys*, vol. 51, no. 3, pp. 1–38, Jul. 2018.

[28] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2015, pp. 1310–1321.

[29] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[30] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 289–296.

[31] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett, "Functional mechanism: Regression analysis under differential privacy," 2012, *arXiv:1208.0219*. [Online]. Available: http://arxiv.org/abs/1208.0219

[32] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: Privacy-preserving mechanisms for SVM learning," 2009, *arXiv:0911.5708*. [Online]. Available: http://arxiv.org/abs/0911.5708

[33] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Mach. Learn. Res.*, vol. 12, pp. 1069–1109, Mar. 2011.

[34] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization, revisited," 2014, *arXiv:1405.7085*. [Online]. Available:https://arxiv.org/abs/1405.7085