# Towards Effective Measurement of Membership Privacy Risk for Machine Learning Models

by

Vasisht Duddu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2022

## Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Statement of Contributions

The project was done in collaboration with Sebastian Szyller (Aalto University) under the supervision of Prof. N. Asokan. All three authors wrote a technical report [18] at an earlier stage of the project. This thesis is a substantially revised and extended version of that report (which will later be updated to reflect the new work described in this thesis). We briefly describe the division of work on this project and contribution of the participants:

- Vasisht came up with the idea of using Shapley values for estimating membership privacy risk for individual training data records.
- All authors were involved in shaping the design of the experiments.
- Vasisht implemented SHAPr, setup and executed experiments, and analyzed the results.
- Sebastian provided expert advice.

# Abstract

Machine learning (ML) models are trained on data which can be sensitive. Membership inference attacks (MIAs) infer whether a particular data record was used to train an ML model. This violates the membership privacy of an individual, specially in applications where the knowledge of that individual's data record in training data is sensitive. For instance, the privacy risk of inferring an individual's health status from a model trained on a dataset containing patients with some specific disease. There is a need for a privacy metric that enables ML model builders to quantify the membership privacy risk of (a) individual training data records, (b) computed independently of specific MIAs, (c) which assesses susceptibility to different MIAs, (d) can be used for different applications, (e) efficiently. None of the prior membership privacy risk metrics simultaneously meet all of these criteria.

Ideally, a membership privacy risk metric will measure the *memorization* of individual training data records by large capacity ML models, which is the cause for membership privacy risk as suggested by prior work. In practice, this can be achieved by estimating the influence of individual training data records to a model's utility. Leave-one-out (LOO) computation, i.e., the difference in model utility with and without a data record in training dataset, can be used to measure this memorization but at high computation cost. Shapley values is an alternative LOO approach with efficient algorithms in the literature. It measures the influence of a training data record on a model's utility and thereby the extent of it being memorized by that model. Hence, we conjecture that Shapley values, can serve as a good membership privacy risk metric to indicate the susceptibility of training data records to MIAs. In this work, we explore the following research question: *can Shapley values effectively estimate the susceptibility of individual training data records to MIAs?*

We validate the above conjecture by presenting SHAPR, a membership privacy metric based on Shapely values which satisfies the desiderata (a) - (e) mentioned above. Using ten benchmark datasets and five MIAs, we show that SHAPR is indeed effective in estimating susceptibility of a training data records to different MIAs as computed using F1 scores. We then focus on recall as being more important than precision for evaluating effectiveness of membership privacy risk metrics. We find that using recall, SHAPR is effective to assess the susceptibility across different MIAs and datasets. We find that SHAPR is comparable or better than prior work for effective MIAs (good accuracy on both members and non-members).

Additionally, other than inheriting applications of Shapley values (e.g., data valuation), SHAPR is versatile and can be used for estimating the disproportionate vulnerability over different subgroups to MIAs. We apply SHAPr to evaluate the efficacy of several defenses against MIAs. First, we show that adding noise to subset of training data records lowers their privacy risk. But this comes at the cost of increasing the privacy risk for remaining training data records, making it an ineffective defence. Second, we show that the membership privacy risk of a dataset is not necessarily improved by removing high risk training data records, thereby confirming an observation from prior work in a significantly extended setting (across ten datasets, removing up to 50% of vulnerable training data records). Third, SHAPr correctly captures the decrease in MIA accuracy on using regularization based defence.

Finally, SHAPr has acceptable computational cost (compared to naïve LOO), i.e., varying from a few minutes for the smallest dataset to ≈92 minutes for the largest dataset.

## Acknowledgements

I would like to thank my advisor Prof. N. Asokan for his constant guidance and supervision throughout my master's program. Under his guidance, I learnt to think deeply about research problems and conducting rigorous research. Despite the pandemic and online meetings, he encouraged and gave me the freedom to explore new research directions which match my interests. I thank Prof. Xi He and Prof. Florian Kerschbaum for mentoring me during my coursework as well as taking time off their busy schedule to provide insightful feedback which improved the thesis.

I thank my labmates who were always available to help and discuss new research problems which eventually led to collaborations on interesting projects. My experience at University of Waterloo was enriched by the fun memories I shared with my friends.

I am thankful to Prof. Antoine Boutet (INRIA, France), Prof. Valentina E. Balas (University of Arad, Romania), Prof. Debasis Samanta (Indian Institute of Technology, Kharagpur), Prof. Reza Shokri (National University of Singapore), Dr. N. Rajesh Pillai (Defence Research and Development Organization, India) for their guidance during my undergraduate. I am grateful to my father, Dr. D. Vijay Rao, in his capacity as an academic mentor for introducing me to research and discussing ideas over dinners.

I am grateful to my parents, brother (Siddharth) and extended family for their constant love, support and motivation during the program.

## Dedication

To my family for their constant love, support and encouragement.

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

$\mathcal{A}dv$ Adversary xiv, 5–7, 11, 12, 15, 19, 22, 23, 37, 50, 51, 53

$\mathcal{M}$ Model Builder 11, 12, 37, 40, 45, 49, 50, 55

**KIFS** Koh's Influence Scores xiv, 47, 48

**$K$-NN** $K$-Nearest Neighbour 9, 53

**SPRS** Song's Privacy Risk Scores x, xiii–xv, 1, 3, 6, 7, 13, 14, 19, 20, 22–34, 36, 39, 42, 44, 45, 47, 48, 50, 51, 53, 56, 66–73

**LOO** leave-one-out iv, v, viii, 2, 3, 8, 13, 14, 16, 45, 56

**MIA** Membership Inference Attack iv, v, ix, xi, xiii–xv, 1–7, 11–15, 17–19, 22–35, 37–43, 46, 48–50, 53–56, 70, 73

**ML** Machine Learning viii, 1, 2, 4, 5, 7, 8, 11, 12, 14, 15, 31, 32, 37, 49, 50, 52–54

**SGD** Stochastic Gradient Descent 4

# List of Symbols

$z_i$ $i_{th}$ training data record with input feature $x_i$ and corresponding true label $y_i$ 6–9, 16, 17

$D_{aux}$ Auxiliary dataset available with the adversary ix, 5–7, 12, 50

$D_{te}$ Dataset to test the target model xviii, 8–11, 17, 19, 23

$D_{tr}$ Dataset to train target model ix, 7–9, 11–14, 50

$I_{conf}$ Confidence based membership inference attack xiii, 5, 6, 22–27, 29–32

$I_{corr}$ Correctness based membership inference attack xiii, 5, 6, 22–28, 30–32

$I_{ideal}$ Ideal Membership Inference Attack 14, 15, 31, 32

$I_{label}$ Label-only based membership inference attack xiii, 6, 22–32

$I_{ment}$ Modified entropy based membership inference attack xi–xiv, 6, 22–28, 30–32, 34–40, 42, 43, 47, 48, 50, 56, 67–73

$X$ Space of all possible input features corresponding to all possible data records 4

$x$ Input feature corresponding to a specific data record 4

$I_{shadow}$ Shadow model based membership inference attack xiii, 5, 22–28, 30–32, 56

$Y$ Space of all possible true label corresponding to all possible data records 4

$y$ True label for a specific data record 4

$f_\theta$ Machine learning model parameters 4

$I_{optimal}$ Optimal Membership Inference Attack 15, 17, 18, 22, 56

$\theta$ Machine learning model parameters 4

$\phi_i$ Shapley value (and by extension SHAPr score) assigned to a $i_{th}$ training data record 8, 16

$\mathcal{A}$ Specific training algorithm used to train a model (e.g., stochastic gradient descent, Adam) 8

$U(S)$ Accuracy of a machine learning model on a testing dataset $D_{te}$ when trained on $S$ 8, 16, 17

# Chapter 1

# Introduction

The assessment of data privacy risks is necessary as highlighted by several official reports from government institutions (NIST [65], the White House [26], and the United Kingdom's Information Commissioner's Office [33]). Membership inference attacks (MIAs) is a potential threat to privacy of an individual's data used for training Machine Learning (ML) models [59, 54, 61, 71]. These attacks infer whether a given data record was used to train that model. For datasets containing an individual's sensitive data, MIAs constitute a privacy threat. For instance, identifying that a randomly sampled individual's data was used to train a health-related ML model can allow an adversary to infer the health status of that individual. Hence, measuring the *membership privacy risk* of training data records is essential for data privacy risk assessment.

Several existing tools, like MLPrivacyMeter [46] and MLDoctor [42], can quantify membership privacy risk. They are based on measuring the success rate of known MIAs [59, 71, 54, 48]. In addition, these attacks use *aggregate metrics* such as accuracy, precision and recall over *all* training data records, and are not designed for quantifying *individual record-level* membership privacy risk. Record-level membership privacy allows model builder to estimate the relative risk of different training data records for a fine-grained privacy risk analysis to design privacy-preserving ML models. Additionally, this allows the user to understand the privacy risk of contributing their data to the specific ML task. Song and Mittal [61] proposed a record-level probabilistic metric (which they name "privacy risk metric" hereafter referred as SPRS) defined as the likelihood of a data record being present in the target model's training dataset. SPRS is intended to be used by adversaries rather than model builders.

Ideally, a membership privacy risk metric should capture the root cause of MIAs, namely the memorization of training data records as indicated in prior work [59, 48, 17]. Such a metric will be *independent of any specific attack* and thus be applicable to any future MIAs as well. Hence, there is a need for a membership privacy risk metric which estimates the extent to which the model memorizes an individual training data record. This memorization can be estimated by measuring the influence of each training data record on the model's utility. One potential approach is by using the leave-one-out (LOO) training approach [19, 43] where the influence of a data record is computed using the difference in model utility with and without that data record in the training dataset. Long et al. [43] proposed one such metric based on LOO computation which is independent of any specific attack. However, directly using the naïve LOO approach for each data record is computationally expensive [30, 31, 21, 32]. Therefore, we explore the research question: *Can LOO based approaches efficiently and effectively measure the susceptibility of individual training data records to MIAs?*

We conjecture that Shapley values, a well-known notion in game theory used to quantify the contributions of individuals within groups [56], can serve as a good membership privacy risk metric by *effectively* estimating the susceptibility of individual training data records to successful MIA. The application of Shapley values as membership privacy risk metric is by the virtue of approximating the LOO computation and measuring influence of individual training data records on model utility [31, 21], and thereby estimating the extent of their memorization. Crucially, Shapley values can be *efficiently* computed in one go for every training data record without having to *train two models for each training data record* (with and without that data record in the training dataset) as typically done in naïve LOO approach [30, 32]. Furthermore, Shapley values have been recently used in the context of data valuation in ML (to estimate economic value of a data record) [30, 31, 21, 20] and estimating attribute influence for explainability [44]. This makes a metric based on Shapley values more *versatile* by inheriting non-privacy applications (e.g., data valuation) where other metrics cannot work. In this work, we explore using Shapley values as the basis to quantify the membership privacy risks of individual training data records.

We make the following contributions.

1. We validate our conjecture of Shapley values being effective membership privacy risk metric by presenting SHAPr, an *LOO membership privacy risk metric using Shapley values*, with an *attack-agnostic approach* for estimating membership privacy risk for *individual training data records.* (Chapter 4)

2. We show that SHAPr is *effective* in assessing the susceptibility of training data records to state-of-the-art MIAs across ten benchmark datasets. We find that SHAPr is ei-

ther comparable or better than SPRS (adapted to the model builder's setting) against effective MIAs; (Chapter 6)

3. We demonstrate the applicability of SHAPr in Chapter 7 by showing that
   - SHAPr is versatile (Section 7.1):
     (a) it can be used to estimate the disparity of membership privacy risk across different sensitive subgroups (Section 7.1.1).
     (b) inherits other applications such as data valuation, by virtue of using Shapley Values (Section 7.1.2).
   - SHAPr is effective to evaluate defences (Section 7.2):
     (a) it can correctly estimate how adding noise to a subset of the training dataset impacts membership privacy risk (Section 7.2.1);
     (b) removing data records with high membership privacy risk as a defence does not necessarily reduce risk for the remaining data records, confirming the observation by Long et al. [43], but on a broader scale, using ten large datasets (vs. one), and exploring the effect of removing up to ∼50% of training data records (vs. <1%) (Section 7.2.2);
     (c) effective in capturing the decrease in MIA accuracy on using defences like regularization (Section 7.2.3)
4. We show that SHAPr scores can be computed more efficiently than the direct application of the LOO approach. (Chapter 8)

# Chapter 2

# Background

Consider a training dataset $D_{tr} = \{x_i, y_i\}_{i=1}^{n}$ containing input features $x_i \in X$ and corresponding classification labels $y_i \in Y$ where $X$ and $Y$ are the space of all possible inputs and corresponding labels. A Machine Learning (ML) classifier is a model $f_\theta$ which maps the inputs to the corresponding classification labels $f_\theta : X \rightarrow Y$. The function parameters $\theta$ are updated by minimizing the loss between the model's prediction $f_\theta(x)$ on input $x$ and the true labels $y$. The loss is minimized using gradient based training algorithms such as Stochastic Gradient Descent (SGD) or Adam.

An ML algorithm constitutes a space of all such ML models obtained by training them on datasets randomly sampled from the same underlying data distribution ($D_{tr} \sim P(X \times Y)$) with randomly sampled training algorithm from a set of possible training algorithms (e.g., SGD, Adam). Hence, an ML model is a specific sample of an ML algorithm trained after fixing the training dataset and training algorithm. We use this distinction between ML algorithm and ML model to quantify the privacy risk in Section 2.3 and Chapter 9: Section 9.3.

We first give a background of prior membership inference attacks (MIAs) proposed in literature (Section 2.1) followed by a description of a prior membership privacy risk metric closest to our work (Section 2.2). We then discuss the cause of such membership privacy risk (Section 2.3) followed by a discussion of Shapley values and its state-of-the-art algorithm proposed in prior work [30, 32] which we use (Section 2.4).

## 2.1 Membership Inference Attacks

MIAs exploit the difference in model behaviour on seen training data records and unseen test data records. MIAs differentiate between members and non-members of the training dataset of an ML model using the output predictions of that model, or some function of them. We identify four main types of MIAs proposed in the literature:

**Shadow Models** [59]. Shokri et al. proposed the first MIA that uses a ML attack model to distinguish between a member and non-member based on the predictions of the target model. This MIA assumes that an Adversary ($\mathcal{A}dv$) has auxiliary data ($D_{aux}$) including some training data records used by the target model. $D_{aux}$ is used to train multiple *shadow models* to mimic the utility of the target model. An attack ML model is then trained to distinguish between members and non-members using the predictions of the shadow models. Given a prediction from the target model for an arbitrary input, the attack model can classify it as a member or a non-member. This MIA has two main drawbacks: first, it assumes a strong $\mathcal{A}dv$ who has partial knowledge about the target model's training data, and second it incurs a high computational overhead due to the need to train multiple (shadow) models. We refer to this MIA as $I_{shadow}$.

**Prediction Correctness** [71]. An alternative approach, that makes weaker assumptions regarding $\mathcal{A}dv$'s capabilities, relies on the fact that models which do not generalize well make correct predictions on training data records but not on testing data records. $\mathcal{A}dv$ decides that a data record is a member if it is correctly predicted by the target model and non-member otherwise. This MIA is particularly applicable in settings where the target model outputs only a label. However, the MIA works for poorly generalizing models and assumes $\mathcal{A}dv$ knows the ground truth labels for the data records used to probe the target model. We refer to this MIA as $I_{corr}$.

**Prediction Confidence** [71, 54]. A third approach uses prediction confidence reported by the target model across *all classes*. Given an input data record, the target model outputs a vector describing the confidence that the record belongs to each class. The maximum confidence value is likely to be higher for an input data record that was also part of the training set, than for one that was not [59, 60]. The prediction confidence attack relies on this observation: it declares an input data record as a member if the associated highest confidence is higher than an adversary-chosen threshold, and as a non-member otherwise. Unlike Prediction Correctness attacks, Prediction Confidence attacks do not require $\mathcal{A}dv$ to have any knowledge of the target model's training data or the ground truth for the input data record. However it assumes that the target model outputs confidence values for all classes. We refer to this MIA as $I_{conf}$.

**Prediction Entropy** [61, 59, 54]. Rather than relying on the maximum confidence value in the output prediction, $\mathcal{A}dv$ may resort to a more sophisticated function defined over the set of confidence values in the prediction. The entropy in a model's prediction (i.e., information gain for $\mathcal{A}dv$) is the uncertainty in predictions [59, 54]. The entropy differs for training and testing data records which $\mathcal{A}dv$ can use as the basis for deciding whether an input data record was in the training set. For instance, the output for a training data record is likely to be close to a one-hot encoding, resulting in a prediction entropy close to zero. Testing data records are likely to have higher prediction entropy values. As with the previous method, $\mathcal{A}dv$ can choose a threshold for the prediction entropy to decide whether an input data record is a member or not.

A modification of prediction entropy attack was proposed by Song and Mittal [61]. The prediction entropy is low for data records with both correct or incorrect classification predicted with high confidence by the model. For a given data record $(x, y)$, the modified entropy function: $Mentr(f_\theta(x), y) = -(1-f_\theta(x)_y)log(f_\theta(x)_y) - \sum_{i \neq y}(f_\theta(x)_i log(1-f_\theta(x)_i))$, accounts for this problem. Here, $f_\theta(x)_y$ indicates the prediction on record $x$ with correct label $y$. $\mathcal{A}dv$ thresholds the modified prediction entropy to determine the membership status: $I_{ment}(f_\theta(x), y) = \mathbb{1}\{Mentr(f_\theta(x), y) \leq \tau_y\}$. We refer to this MIA as $I_{ment}$.

For $I_{ment}$ and $I_{conf}$ MIAs, instead of using a fixed threshold of 0.5 over the prediction confidence as seen in original prediction entropy attack, the thresholds $\tau_y$ are adapted for each class using the shadow models trained on $D_{aux}$ to improve the MIA accuracy. This adaptive threshold gives the best MIA accuracy [61].

**Label-Only** [15, 38]. The above MIAs (except for $I_{corr}$) assume $\mathcal{A}dv$ has access to output prediction probabilities across all classes. However, in settings where the model outputs only the most likely class label, the above MIAs are not effective. Due to better performance than Li et al [38], we use Choo et al.'s [15] label only inference MIA, which observes that the training data records have a higher distance from the decision boundary than the testing data records. $\mathcal{A}dv$ uses this difference to infer the membership status of any arbitrary data record. We refer to this MIA as $I_{label}$.

## 2.2 Song and Mittal's Privacy Risk Scores

Song and Mittal [61] describe a membership privacy risk metric[1] (which we refer to as SPRS) that defines the membership privacy risk score of $z_i$= $(x_i, y_i)$ as the posterior

[1] Source Code: https://github.com/inspire-group/membership-inference-evaluation/blob/master/privacy_risk_score_utils.py

probability that $z_i \in D_{tr}$ given the output predictions from the model $f_\theta(x_i)$. They compute the score as $r(z_i) = P(z_i \in D_{tr}|f_\theta(x_i))$. This probability is computed using Bayes' theorem as

$$\frac{P(z_i \in D_{tr})P(f_\theta(x_i)|z_i \in D_{tr})}{P(z_i \in D_{tr})P(f_\theta(x_i)|z_i \in D_{tr}) + P(z_i \in D_{te})P(f_\theta(x_i)|z_i \in D_{te})} \tag{2.1}$$

They assume that the probability of the data record belonging to the training/testing dataset is equally likely, $P(z_i \in D_{tr}) = P(z_i \in D_{te}) = 0.5$. The membership privacy risk scores rely on training shadow models on $D_{aux}$ to mimic the functionality of the target model. The conditional probabilities $P(f_\theta(x_i)|z_i \in D_{tr})$ and $P(f_\theta(x_i)|z_i \in D_{te})$ are then computed using the shadow model's output predictions on $D_{aux}$'s training and testing dataset. Further, instead of using fixed threshold based prediction entropy MIA, each class has a threshold for deciding the data record's membership which are computed using $D_{aux}$. The conditional probabilities are estimated per class $P(f_\theta(x_i)|z_i \in D_{tr}) = \{P(f_\theta(x_i)|z_i \in D_{tr}, y = y_i)\}$ across all class labels $y = y_i$.

Traditional MIAs require $\mathcal{A}dv$ to sample arbitrary data records to infer their membership status. SPRS is designed as a tool for $\mathcal{A}dv$ to identify data samples which are more likely to be members instead of sampling a large number of data records.

## 2.3 Memorization of Training Data in ML

Membership privacy risk (susceptibility to MIAs) occurs due the fact that ML models, with their inherent large capacity, tend to "memorize" training data records [48, 17]. This results in distinguishable output predictions from the ML model on seen training data records and unseen testing data records [59, 54].

To better understand "memorization" in practice, we can think of membership privacy risk as follows: consider an ML model is trained on $D_{tr}$. $\mathcal{A}dv$ samples $z_i$ from $D_{tr}$'s underlying data distribution where $z_i = (x_i, y_i)$ is the $i^{th}$ data record with input features $x_i$ and label $y_i$. $\mathcal{A}dv$ can query the model and observe the model's predictions (blackbox API access) [59, 61, 54] and parameters to compute intermediate layer output (whitebox access) [48, 36]. $\mathcal{A}dv$'s goal is to infer whether $z_i \in D_{tr}$ or $z_i \notin D_{tr}$. In practice, $\mathcal{A}dv$ can do this by estimating the *influence* of $z_i$ on model's observables (predictions or intermediate layer output) after interacting with the ML model. Hence, measuring this *influence* on the model observables acts as a signal for membership privacy risk for an individual data record $z_i$ [70].

One approach to estimate this influence was proposed by Feldman [19] which was referred to as "memorization" of a data record by an ML model. Memorization of $z_i$ can be estimated as the difference in the prediction of a model on input features $x_i$ when the model was trained with and without $z_i$ in its training set [19]. Formally, for a specific model $f_\theta$ drawn from the set of models for a training algorithm $\mathcal{A}$, Feldman [19] formulates memorization as follows:

$$mem(z_i, D_{tr}, \mathcal{A}) = |Pr_{f_\theta \sim \mathcal{A}(D_{tr} \cup z_i)}[f_\theta(x_i) = y_i] - Pr_{f_\theta \sim \mathcal{A}(D_{tr} \backslash z_i)}[f_\theta(x_i) = y_i]| \tag{2.2}$$

If $mem(z_i,\ D_{tr},\ \mathcal{A})$ is high, the model is likely to have memorized $z_i$. The above formulation of memorization is an leave-one-out (LOO) based approach which captures the extent to which the presence of a record in the training dataset *influences* the model's output predictions [19].

## 2.4 Shapley Values

An alternative approach to capture the influence of a training data record is by estimating Shapley values [21, 31, 30, 32]. Shapley values ($\phi_i$) are of the form,

$$\phi_i = \frac{1}{|D_{tr}|} \sum_{S \subseteq D_{tr} \backslash \{z_i\}} \frac{1}{\binom{|D_{tr}-1|}{|S|}} [U(S \cup \{z_i\}) - U(S)] \tag{2.3}$$

where $S$ is a randomly chosen subset of $D_{tr} \backslash \{z_i\}$ and $U(S)$ (accuracy of $f_\theta$ on a testing dataset $D_{te}$ when trained on $S$) is a utility metric. $\binom{|D_{tr}-1|}{|S|}$ denotes the binomial coefficient for choosing $|D_{tr} - 1|$ elements from a set of $|S|$ elements. Here, the Shapley value of $z_i$ is defined as the average marginal contribution of $z_i$ to $U(S)$ over all training data subsets $S \subseteq D_{tr} \backslash \{z_i\}$. Evaluating the Shapley function naïvely for all possible subsets with and without $z_i$ is computationally expensive (complexity of $O(2^{|D_{tr}|}$ for $|D_{tr}|$ data records [32]) and not scalable (leading to the same problem as with naïve LOO) [19, 43]. Note that for computing Shapley values cannot be done by training $|D_{tr}| + 1$ models - one model with all $|D_{tr}|$ samples (baseline), and then one model for each of the removed samples. Shapley value, by definition, require sampling a subset $S$ for which we train two models: one with and without $z_i$. The scores will then be computed by averaging by across multiple $S$ but for $z_i$ which adds to the computationally complexity making naïve functions expensive.

However, several prior work have proposed efficient algorithms which approximate the computation of Shapley values [31, 21, 30, 32]. We consider the most efficient algorithm in literature where Shapley values can be efficiently computed using a $K$-Nearest Neighbours ($K$-NN) classifier as a surrogate model [32]. Unlike the naïve approach to computing Shapley values which requires training two models for *each training data record*, the $K$-NN model, once trained, can be used to compute the Shapley values for *all training data records*. This improves the computational complexity to O$(|D_{tr}|log(|D_{tr}|.|D_{te}|))$ compared to exponential complexity of the formulation in Equation 2.3. We now outline this approach [32][2].

For a given $z_i$, we can first compute the *partial contribution* $\phi_i^{test}$ of a single test data record $z_{test}$ to the Shapley value $\phi_i$ of $z_i$, and then add up these partial contributions across the entire $D_{te}$.

**Step 1: "Sorting Phase".** This phase of $K$-NN classifier consists of passing $D_{tr}$ and a single testing data record $z_{test} = (x_{test}, y_{test}) \in D_{te}$, as an input to the target classifier $f_\theta^l$ which is the output of the $l^{th}$ layer in the network. $f_\theta$ denotes final layer probability scores across all classes. Following prior work on Shapley values [30, 32], the outputs $f_\theta^1(D_{tr})$ and $f_\theta^1(x_{test})$ and their corresponding true labels are used for further computation.

**Step 2: "Score Assignment".** For $z_{test}$, the $K$-NN classifier identifies the top $K$ closest training data records $(x_{\alpha_1}, \cdots, x_{\alpha_K})$ with labels $(y_{\alpha_1}, \cdots, y_{\alpha_K})$ using the distance between the predictions $(f_\theta^1(x_{\alpha_1}), \cdots, f_\theta^1(x_{\alpha_K}))$ and $f_\theta^1(x_{test})$. We use $\alpha_j(S)$ to indicate the index of the training data record, among all data records in $S$, whose output prediction is the $j^{th}$ closest to $f_\theta^1(x_{\text{test}})$. For brevity, $\alpha_j(D_{tr})$ is written simply as $\alpha_j$. Following prior work on data valuation [30, 32], we use $K = 5$.

**Step 3.** The $K$-NN classifier assigns majority label corresponding to the top $K$ training data records as the label to $x_{\text{test}}$. The probability of the classifier assigning the correct label is given as: $P[f_\theta^1(x_{\text{test}}) = y_{\text{test}}] = \frac{1}{K}\sum_{i=1}^{K} \mathbb{1}[y_{\alpha_i} = y_{\text{test}}]$. Hence, the utility of the classifier with respect to the subset $S$, and the single test data record $z_{test}$, is computed as $U^{test}(S) = \frac{1}{K}\sum_{k=1}^{\min\{K,|S|\}} \mathbb{1}[y_{\alpha_k(S)} = y_{\text{test}}]$.

**Step 4.** Consider all data records in $D_{tr}$ sorted as above $\{\cdots, z_{\alpha_{i-1}}, z_{\alpha_i}, z_{\alpha_{i+1}}, \cdots\}$. From Equation 2.3, the difference between the partial contributions for two *adjacent* data records $z_{\alpha_i}, z_{\alpha_{i+1}} \in D_{tr}$ is given by $\phi_{\alpha_i}^{test} - \phi_{\alpha_{i+1}}^{test} =$

$$\frac{1}{|D_{tr}| - 1} \sum_{S \subseteq D_{tr} \setminus \{z_{\alpha_i}, z_{\alpha_{i+1}}\}} \frac{[U^{test}(S \cup \{z_{\alpha_i}\}) - U^{test}(S \cup z_{\alpha_{i+1}})]}{\binom{|D_{tr}-2|}{|S|}} \tag{2.4}$$

Using the $K$-NN utility function: $U^{test}(S \cup \{z_{\alpha_i}\}) - U^{test}(S \cup z_{\alpha_{i+1}}) = \frac{\mathbb{1}[y_{\alpha_i} = y_{\text{test}}] - \mathbb{1}[y_{\alpha_{i+1}} = y_{\text{test}}]}{K}$.

[2] Source code: https://github.com/AI-secure/Shapley-Study

Once the label for $x_{test}$ is assigned, the partial contribution can be computed recursively starting from the farthest data record:

$$\phi^{test}_{\alpha_{|D_{tr}|}} = \frac{\mathbb{1}[y_{\alpha_{|D_{tr}|}} = y_{\text{test}}]}{|D_{tr}|} \tag{2.5}$$

$$\phi^{test}_{\alpha_i} = \phi^{test}_{\alpha_{i+1}} + \frac{\mathbb{1}[y_{\alpha_i} = y_{\text{test}}] - \mathbb{1}[y_{\alpha_{i+1}} = y_{\text{test}}]}{K} \frac{\min\{K, i\}}{i} \tag{2.6}$$

The fraction $\frac{\min\{K,i\}}{i}$ is obtained by simplifying the binomial coefficient (the full derivation can be found in Theorem 1 of Jia et al. [30]). The intuition behind Equation 2.6 is that the contribution of $z_{\alpha_i}$ is 0 if the nearest neighbor of $z_{\alpha_i}$ in $S$ is closer to $z_{test}$ than $z_{\alpha_i}$, and 1 otherwise. Using the above steps, we get $\phi^{test}$ for each $z_{test}$ of size $D_{tr} \times 1$. This recursive formulation in Equation 2.6 can be extended across all $D_{te}$ to obtain a matrix $[\phi^{test}_i]$ of size $D_{tr} \times D_{te}$. The final Shapley values can be obtained by aggregating the partial contributions $\phi^{test}_i$ across $D_{te}$.

# Chapter 3

# Problem Statement

We conjecture that Shapley values, by virtue of measuring influence on model utility, and hence the extent of memorization, can serve a good membership privacy risk metric by indicating the susceptibility of training data records to membership inference attacks (MIAs). Our goal is to verify this conjecture. To this end, we lay out the system and adversary models (Section 3.1), describe the desiderata for designing such a metric (Section 3.2), and outline the limitations of prior work (Section 3.3).

## 3.1 System and Adversary Model

**System Model.** We consider the perspective of a Model Builder ($\mathcal{M}$) who trains a model using a dataset contributed to by multiple participants. $\mathcal{M}$ wants to estimate the susceptibility of individual data records to MIAs. $\mathcal{M}$ has full access to the training ($D_{tr}$) and testing ($D_{te}$) datasets and can use them to compute membership privacy risk scores for each training data record.

**Adversary Model.** We describe the adversary model for the MIAs. The ground truth for the membership privacy risk metric for a given training data record is the degree to which an actual state-of-the-art MIA [59, 54, 61, 71] succeeds against that record. We adapt the standard adversary model for MIAs [61, 59] to $\mathcal{M}$'s perspective.

The standard adversary model from prior work [59, 54, 61] considers adversary Adversary ($\mathcal{A}dv$) has access to the prediction interface of a model $f_\theta$ built using a training dataset $D_{tr}$. $\mathcal{A}dv$ submits data records via the prediction interface and receives model outputs (this is a widely adapted setting for cloud-based Machine Learning (ML) models in the industry). Given an input data record $x$, $\mathcal{A}dv$ can only observe the final output prediction $f_\theta(x)$. The MIAs considered use the full prediction vector [61, 59] instead of the prediction labels [15, 38]. $\mathcal{A}dv$ does not know the underlying target model architecture and additionally has access to an auxiliary dataset $D_{aux}$ sampled from the same distribution as $D_{tr}$. Prior MIAs assume partial overlap between $D_{tr}$ and $\mathcal{A}dv$'s $D_{aux}$ [59, 54, 61, 47, 29].

However, we adapt the above adversary model to $\mathcal{M}$'s perspective. We assume that $\mathcal{A}dv$'s $D_{aux}$ completely overlaps with $D_{tr}$ which gives an upper bound on the membership privacy risk. This is reasonable from $\mathcal{M}$'s perspective who has complete access to $D_{tr}$ which is used to train the ML model. This implies that MIAs which rely on shadow models (to learn the characteristics to differentiate between members/non-members) are directly using the target model for as the shadow models. In other words, the underlying target model architecture is known and used as shadow models. This setting corresponds to $\mathcal{M}$ simulating the strongest possible adversary with complete knowledge of $D_{tr}$ who evaluates how accurate are MIAs by matching the MIA predictions with the ground truth membership status which is already known to $\mathcal{A}dv$.

## 3.2 Requirements for Membership Privacy Metric

We identify the following requirements which should be satisfied while designing a membership privacy risk metric:

**R1** **Fine-grained.** The metric generates scores for measuring the membership privacy risk of individual training data records. This allows for a fine-grained membership privacy risk analysis of the training data records of an ML model (Chapter 4).

**R2** **Attack-Agnostic.** Ideally, the metric should capture the root cause of all MIAs, i.e., memorization of training data records by ML models [59, 48, 19]. Hence, membership privacy risk scores resulting from the ideal metric must be computed independently of specific MIAs. This allows the scores to assess the membership privacy risks with respect to different MIAs (Chapter 4).

**R3** **Effectiveness.** The membership privacy risk scores of training data records must correlate with the likelihood of success of MIA prediction against those records. This is computed using metrics such as F1 score, precision and recall computed between the scores after applying a threshold and MIA prediction (Chapter 6). Evaluation of effectiveness presumes the availability of a reliable ground truth for computing MIA predictions. We return to this consideration in Section 3.4.

**R4** **Applicability.** The membership privacy risk scores, once computed, should be useful to estimate other characteristics of the dataset (Chapter 7). For instance, the versatility[1] of the metric to evaluate the susceptibility of sensitive subgroups to MIAs, estimating economic value (as defined in prior work on Shapley scores; Chapter 7: Section 7.1 or evaluating defences Chapter 7: Section 7.2).

**R5** **Efficiency.** Membership privacy risk scores resulting from the metric must be computed within a reasonable time and low computation overhead. (Chapter 8)

## 3.3 Limitations of Existing Metrics

Privacy assessment libraries such as MLPrivacyMeter [46] and MLDoctor [42] quantify the membership privacy risk using existing MIAs. They use aggregate metrics such as accuracy, precision and recall for MIAs across all training data records, and are not optimized for estimating the privacy risks of individual data records [61]. Hence, such metrics do not satisfy the fine-grained requirement **R1**.

Song and Mittal propose SPRS which is a probabilistic membership privacy risk metric for individual data records [61]. The more effective an MIA is against a particular data record, the higher the score. SPRS computes membership privacy risk scores for different training data records using MIA features for a specific MIA. For instance, SPRS, as indicated in the original paper, uses modified entropy over the output predictions from training and testing data records to compute the scores. This does not satisfy the attack-agnostic requirement **R2**. We later show that SPRS does not satisfy the applicability requirement **R4** and is not effective for some of the applications (c.f. Chapter 7).

Long et al. [43] propose Differential Training Privacy as a membership privacy metric based on the naïve leave-one-out (LOO) approach: computing the difference between model predictions with and without a given training record in the $D_{tr}$ and hence, the influence of that record on the model utility. However, as we saw in Chapter 2.4, direct application of the LOO approach cannot scale to large datasets and models since it requires retraining the model to estimate the score for *each* data record. Hence, such a naïve LOO approach does not satisfy the efficiency requirement **R5** (c.f. Chapter 8).

Table 3.1 summarizes the prior work with respect to the different requirements that they satisfy. None of the prior work satisfy all the requirements for an ideal membership privacy

[1] Versatility is an important design choice instead of a property of the privacy metric. This is because the deployment success of a tool depends on its costs and benefits. Given two otherwise comparable techniques, the one having additional benefits is more likely to be deployed.

Table 3.1: Summary of prior membership privacy metrics and requirements satisfied. None of the metrics satisfy all the requirements.

| **Requirements** | **MLPrivacyMeter** [46]<br>**MLDoctor** [42]<br>(Attack Based) | **SPRS** [61]<br>(Probabilistic Metric) | **Long et al.** [43]<br>(LOO Metric) |
|---|---|---|---|
| **R1 Fine-Grained** | ○ | ● | ● |
| **R2 Attack-Agnostic** | ○ | ○ | ● |
| **R3 Effectiveness** | ● | ● | ● |
| **R4 Applicability** | ○ | ○ | ● |
| **R5 Efficiency** | ● | ● | ○ |

risk metric. This begs the question of *whether LOO metric can be improved to be an efficient and effective metric for estimating susceptibility of individual training data records to MIAs.* We specifically consider Shapley values given that it has efficient algorithms proposed in prior literature and is a good approximation of naïve LOO computation [31, 30, 21, 20].

## 3.4 Challenges in Evaluating Effectiveness of Membership Privacy Risk Metric

The effectiveness of a membership privacy risk metric is evaluated with respect to an MIA's predictions (labels a data record as a member or a non-member) as a ground truth. To claim that a metric is effective, we should consider a reliable ground truth.

One possibility of such a ground truth is an Ideal MIA ($I_{ideal}$). $I_{ideal}$ predicts all training data records as members and non-training data records as non-members, i.e., it acts as a perfect distinguisher between training and non-training data records. However, we argue that $I_{ideal}$ a poor ground truth: Generally, $D_{tr}$'s distribution is long-tailed [19]. An ML model generalizes well for records that appear frequently (or are similar to learn a pattern over them). To generalize well on training data records which do not occur frequently (constitute the long tail of the $D_{tr}$'s distribution [19]) the model ends up memorizing those data records [49]. It is easier to predict the membership status of such *memorized* training data records due their higher influence on model predictions/parameters as it results in higher distinguishability between members and non-members. This distinguishability is then exploited by MIAs to identify their membership status more easily. On the other hand, it is difficult to correctly predict the membership status of generalized training data records due to low memorization (subsequently low influence on model predictions/parameters) [11]. This was shown previously, when all the prior MIAs are evaluated using $I_{ideal}$ as the ground truth, it results in high false positive rates [55, 11, 52, 67]. Hence, using $I_{ideal}$ as ground truth for evaluating the effectiveness of a membership privacy risk metric is not useful as it does not reflect what $\mathcal{A}dv$ can infer in real-world.

An alternate ground truth for evaluating membership privacy risk metric is an Optimal MIA ($I_{optimal}$). Such an MIA predicts only the highly memorized training data records (i.e., higher influence in model predictions or parameters) as members and all remaining data records as non-members. In practice, this is best that an MIA can do, as they do not exploit model's memorization of training data records directly but rather indirectly through difference in model's behaviour on seen training data records and unseen testing data records from the prediction interface of the model. While there is a known ground truth for $I_{ideal}$ , we do not have a ground truth for $I_{optimal}$ . This makes it challenging to identify whether current MIAs proposed in literature are close to $I_{optimal}$ .

**How does this impact the effectiveness evaluation of membership privacy risk metric?** Since, membership privacy risk metric capture the root cause of MIAs, i.e., memorization of training data records by ML models, their effectiveness should be computed with respect to $I_{optimal}$ as ground truth. Due to the lack of $I_{optimal}$ as ground truth for evaluating membership privacy risk metric, the effectiveness of such metrics is evaluated with respect to existing state-of-the-art MIAs proposed in the literature. Hence, for any membership privacy risk metric, we cannot generally claim their effectiveness against all MIAs but only discuss their effectiveness with respect to specific MIAs. In this work, we evaluate the effectiveness of membership privacy risk metrics with respect to specific MIAs proposed in literature.

# Chapter 4

# SHAPr: An LOO-based membership privacy risk metric

Shapley values (Chapter 2: Section 2.4) was originally designed as a game-theoretic notion to quantify the contributions of individuals within groups to the utility of a given task [56]. Recently, this was adopted in the context of machine learning for data valuation [21, 20, 31, 30] and explainability [44]. Recall that our conjecture is that leave-one-out (LOO) metric (specifically Shapley values) can be effective in estimating the membership privacy risk of individual training data records efficiently (Chapter 3). In order to validate our conjecture and evaluate the effectiveness of an LOO metric to measure membership privacy risk of individual training data records, we present SHAPr, a membership privacy risk metric that uses Shapley values. SHAPr is based on the algorithm described in Chapter 2: Section 2.4.

SHAPr scores inherit certain properties from Shapley values which allow SHAPr to satisfy the requirements introduced in Chapter 3. In the context of membership privacy risk scores, these properties can be formulated as follows:

**P1 Interpretable.** SHAPr score $\phi_i$ (Equation 2.3) of a data record $z_i = (x_i, y_i)$ is measured by how $z_i$'s addition to a training dataset $S$ influences utility $U()$ of the resulting model (Equation 2.3). Consequently, no influence (i.e., $U(S) = U(S \cup z_i)$) leads to a zero score for $z_i$. Similarly if two data records $z_i$ and $z_j$ have the same influence (i.e., $U(S \cup z_i) = U(S \cup z_j)$, then they are assigned the same score. We can identify three ranges of SHAPr scores that have associated semantics:

(a) **Case 1: $U(S \cup \{z_i\}) = U(S) \rightarrow \phi = 0$:** There is no difference in the model's output regardless of the presence of $z_i$ in the training dataset: $z_i$ has no membership privacy risk.

(b) **Case 2: $U(S \cup \{z_i\}) > U(S) \rightarrow \phi > 0$:** $z_i$ contributed to increasing the model utility. Higher scores indicate higher likelihood of memorization which increases the susceptibility to Membership Inference Attacks (MIAs).

(c) **Case 3: $U(S \cup \{z_i\}) < U(S) \rightarrow \phi < 0$:** $z_i$ was harmful to the model's utility (not learnt well by the model or is an outlier). It has a higher loss and is indistinguishable from testing data records which makes it less susceptible to MIAs.

This clear semantic association allows us to set meaningful thresholds for SHAPr scores that can be used to decide whether a data record is susceptible to MIAs. The natural choice for a threshold is zero, i.e., records with higher score are indicated as members due to higher model's memorization of those records.

**P2 Additive.** $\phi_i$ is computed using $D_{te}$ (Equation 2.3). Specifically, $\phi_i(U_k)$ represents the influence of $z_i$ on utility $U()$ w.r.t to $k^{th}$ testing data record. For two testing data records $k$ and $l$, $U_i(\{k,l\}) = U_i(k) + U_i(l)$. Hence, $\phi_i$ is the sum of the membership privacy risk scores of $z_i$ with respect to each testing data record. This property further implies *group rationality* [21, 30] where $U()$ is fairly and completely distributed amongst all the training data records.

**P3 Heterogeneous.** Different training data records influence the model's utility differently and hence, have varying susceptibility to MIAs (referred to as "heterogeneity"). SHAPr assigns scores to training data records based on their individual influence on the model's utility. This is referred to as *equitable distribution* of utility among the training data records in prior work [31].

We will refer back to these properties while interpreting the results of our experiments (Chapters 6 and 7).

By definition, SHAPr, by virtue of using Shapley values, is fine-grained as it assigns scores for individual training data records based on their influence to model utility satisfying requirement **R1**. Furthermore, the generation of SHAPr scores do not use any MIA features required for performing MIAs. Hence, this makes SHAPr an attack-agnostic metric, satisfying requirement **R2**.

**SHAPr reflects Optimal MIAs ($I_{optimal}$).** Recall from Chapter 3: Section 3.4 that we indicate that an $I_{optimal}$ predicts memorized training data records as members and remaining data records as non-members [19]. Hence, SHAPR generating positive, negative and zero scores is expected as it assigns score based on the influence of each training data record to the model utility. This by extension estimates the extent of memorization of these training data records. Having a high positive SHAPR score is indicative of high memorization being atypical samples making them vulnerable to MIAs while a model generalizes well over typical samples making them a lower privacy risk due to lower memorization. This membership privacy risk metric matches with the definition of $I_{optimal}$. Hence, we conjecture that SHAPR will reflect the susceptibility to $I_{optimal}$. The caveat is that we cannot verify this conjecture because we do not have the ground truth for $I_{optimal}$. However, the conjecture motivates our evaluation of effectiveness of SHAPR with respect to different MIAs.

# Chapter 5

# Experimental Setup

We systematically evaluate the effectiveness of SHAPr using several datasets which are described in Section 5.1. We then describe the model architecture details for training on the datasets (Section 5.2), state-of-art Membership Inference Attacks (MIAs) used for computing Adversary ($\mathcal{A}dv$)'s MIA prediction for different training data records (Section 5.3) and the metrics to evaluate the effectiveness of SHAPr with respect to these MIA predictions used as a ground truth (Section 5.4). We finally describe the model utility on $D_{te}$ and performance of different MIAs (Section 5.5).

## 5.1 Datasets

We used ten datasets for our experiments. Following prior work [59, 61], we used the same number of training and testing data records from all the datasets for computing balanced accuracy for MIAs. An exception to this is MNIST and FMNIST where we used the entire training dataset (60,000 data records) and testing dataset (10,000 data records) of different sizes to ensure the utility of the resulting model is sufficiently high. Three datasets, TEXAS, LOCATION and PURCHASE, were also used to evaluate Song's Privacy Risk Scores (SPRS) [61] – we refer to them as SPRS datasets. To facilitate comparison with SPRS, we used the same dataset partitions for the three SPRS datasets as described in [61]. We summarize the dataset partitions in Table 5.1.

**SPRS Datasets.** We briefly describe each of the ten datasets, starting with the SPRS datasets:

Table 5.1: Summary of dataset partitions for our experiments.

| **Dataset** | **Training Set Size** | **Testing Set Size** |
|---|---|---|
| **SPRS Datasets** | | |
| **LOCATION** | 1000 | 1000 |
| **PURCHASE** | 19732 | 19732 |
| **TEXAS** | 10000 | 10000 |
| **Additional Datasets** | | |
| **MNIST** | 60000 | 10000 |
| **FMNIST** | 60000 | 10000 |
| **USPS** | 3000 | 3000 |
| **FLOWER** | 1500 | 1500 |
| **MEPS** | 7500 | 7500 |
| **CREDIT** | 15000 | 15000 |
| **CENSUS** | 24000 | 24000 |

**LOCATION** contains the location check-in records of individuals [4]. We used the pre-processed dataset from [59] which contains 5003 data samples with 446 binary features corresponding to whether an individual has visited a particular location. The data is divided into 30 classes representing different location types. The classification task is to predict the location type given the location check-in attributes of individuals. As in prior work [61, 29], we used 1000 training data records and 1000 testing data records.

**PURCHASE** consists of shopping records of different users [5]. We used a pre-processed dataset from [59] containing 197,324 data records with 600 binary features corresponding to a specific product. Each record represents whether an individual has purchased the product or not. The data has 100 classes each representing the purchase style for the individual record. The classification task is to predict the purchase style given the purchase history. We used 19,732 train and test records as in prior work [61].

**TEXAS** consists of Texas Department of State Health Services' information about patients discharged from public hospitals [6]. Each data record contains information about the injury, diagnosis, the procedures the patient underwent and some demographic details. We used the pre-processed version of the dataset from [59] which contains 100 classes of patient's procedures consisting 67,330 data samples with 6,170 binary features. The classification task is to predict the procedure given patient's attributes. We used 10,000 train and test records as in prior work [29, 61].

**Additional Datasets.** We used seven other datasets described below. We rounded down the number of training data records in dataset to the nearest 1000 and split it in half between training and testing datasets.

**MNIST** consists of a training dataset of 60,000 images and a test dataset of 10,000 images that represent handwritten digits (0-9). Each data record is a 28x28 grayscale image with a corresponding class label identifying the digit. The classification task is to identify the handwritten digits. We used the entire training and testing set.

**FMNIST** consists of a training dataset of 60,000 data records and a test dataset of 10,000 data records that represent pieces of clothing. Each data record is a 28x28 grayscale image with a corresponding class from one of ten labels. The classification task is to identify the piece of clothing.

**USPS** consists of 7291 16x16 grayscale images of handwritten digits [7]. There area total of 10 classes. The classification task is to identify the handwritten digits. We used 3000 training data records and 3000 testing data records.

**FLOWER** consists of 3670 images of flowers categorized into five classes—chamomile, tulip, rose, sunflower, and dandelion—with each class having about 800 320x240 images. The dataset was collected from Flickr, Google Images and Yandex Images [3]. The classification task is to predict the flower category given an image. We used 1500 train and 1500 testing data records.

**CREDIT** is an anonymized dataset from the UCI Machine Learning dataset repository which contains 30000 records with 24 attributes for each record [2]. It contains information about different credit card applicants, including a sensitive attribute: the gender of the applicant. There are two classes indicating whether the application was approved or not. The classification task is to predict whether the applicant will default. We used 15000 training data records and 15000 testing data records.

**MEPS** contains 15830 records of different patients that used medical services, and captures the frequency of their visits. Each data record includes the gender of the patient, which is considered a sensitive attribute. The classification task is to predict the utilization of medical resources as “High” or “Low” based on whether the total number of patient visits is greater than 10. We use 7500 training data records and 7500 testing data records.

**CENSUS** consists of 48842 data records with 103 attributes about individuals from the 1994 US Census data obtained from UCI Machine Learning dataset repository [1]. It includes sensitive attributes such as gender and race of the participant. Other attributes include marital status, education, occupation, job hours per week among others. The classification task is to estimate whether the individual’s annual income is at least 50,000 USD. We used 24000 training data records and 24000 testing data records.

## 5.2 Model Architecture

While the proposed SHAPR scores are compatible with all types of machine learning models, we focus on deep neural networks in our evaluation. We used a fully connected model with the following architecture: [1024, 512, 256, 128, $n$] with tanh() activation functions where $n$ is the number of classes. This model architecture has been used in prior work on MIAs [54, 47, 48, 29, 59, 61]. SHAPR is scalable to larger models such as ResNet (previously shown for data valuation for Shapley values [32, 30]) but we focus on model architectures used previously in privacy literature.

## 5.3 Membership Inference Attacks

As discussed in Chapter 3: Section 3.4, due to the lack of an Optimal MIA ($I_{optimal}$) to be used as a ground truth, the effectiveness of SHAPR is specifically evaluated with respect to existing state-of-the-art MIAs proposed in the literature. As described in Chapter 2: Section 2.1, we consider multiple state-of-the-art MIAs: $I_{ment}$, $I_{corr}$, $I_{conf}$, $I_{label}$ and $I_{shadow}$.

$I_{ment}$ was originally used by SPRS to generate scores by using the modified entropy function over the output predictions. SPRS additionally used $I_{ment}$ as ground truth MIA predictions for individual training data records by $\mathcal{A}dv$. While this acts as a good ground truth baseline for comparison with SPRS, we extend our evaluation to remaining four MIAs as well: $I_{corr}$, $I_{conf}$, $I_{label}$ and $I_{shadow}$, and compare with SPRS.

## 5.4 Evaluation Metrics

For all the experiments, we used accuracy of MIAs as the primary metric along with the average membership privacy risk score.

**Balanced Attack Accuracy** is the number of training and testing data records, of equal dataset sizes, which are correctly distinguished as members and non-members (reported in Table 5.2). We also refer to this as simply "attack accuracy".

**Average membership privacy risk score** is the average over the membership privacy risk scores assigned to training data records by a metric to evaluate the membership privacy risk across a group of data records.

As in prior work [61], we used three additional metrics to measure the success of the SHAPr scores with respect to $I_{ment}$, $I_{corr}$, $I_{conf}$, $I_{label}$ and $I_{shadow}$ MIAs: precision, recall and F1 score.

**Precision** is the ratio of true positives to the sum of true positive and false positives. This indicates the fraction of data records inferred as members by $\mathcal{A}dv$ which are indeed members of training dataset.

**Recall** is the ratio of true positives to the sum of true positives and false negatives. This indicates the fraction of the training dataset's members which are correctly inferred as members by $\mathcal{A}dv$.

**F1 score** is the harmonic mean of precision and recall computed as $2 \times \frac{precision \times recall}{precision + recall}$. The highest values is one indicates perfect precision and recall while the minimum value of zero is when either precision or recall are zero.

In all the cases, prediction as a "member" is considered as a positive class. For evaluating the effectiveness of both the membership privacy risk metrics, SPRS and SHAPr, F1 score is used in our evaluation to account for both precision and recall. However, it is worth noting that for a metric used to assess membership privacy risk, recall is more important than precision because minimizing false negatives (i.e., failing to correctly identify a training data record at risk) is undesirable from a privacy perspective, whereas false positives (i.e., incorrectly flagging a record as risky) constitutes erring on the safe side. Hence, we focus on recall for evaluating the effectiveness of membership privacy metric and comparison.

## 5.5 Summary of Model and MIA Utility

Here, we summarize the performance of the target model utility on $D_{te}$ followed by the balanced attack accuracy of the five different MIAs.

**Target Model Utility.** We report the results obtained on training the target model Table 5.2 presents the baseline test accuracy of target models trained with each dataset. For SPRS datasets, the performance obtained are similar to the results reported in Song and Mittal [61]. While for "Additional Datasets", we the performance on these datasets is close to the best reported performance in literature.

**MIAs' Accuracy.** Table 5.3 indicates the different balanced MIA accuracy for the five different MIAs considered in this work. For the SPRS datasets, the results for $I_{ment}$, $I_{corr}$ and $I_{conf}$ are similar to the results indicated in Song and Mittal [61].

Table 5.2: Test accuracy of target models for each dataset, averaged over 10 runs.

| Dataset | Test Accuracy |
|---|---|
| **SPRS Datasets** | |
| **LOCATION** | 69.00 |
| **PURCHASE** | 84.65 |
| **TEXAS** | 49.92 |
| **Additional Datasets** | |
| **MNIST** | 98.10 |
| **FMNIST** | 89.30 |
| **USPS** | 95.50 |
| **FLOWER** | 89.60 |
| **MEPS** | 84.00 |
| **CREDIT** | 79.90 |
| **CENSUS** | 82.20 |

Table 5.3: Balanced MIAs accuracy for each dataset across different MIAs.

| Dataset | $I_{ment}$ | $I_{corr}$ | $I_{conf}$ | $I_{label}$ | $I_{shadow}$ |
|---|---|---|---|---|---|
| **SPRS Datasets** | | | | | |
| **LOCATION** | 87.70 | 71.00 | 86.50 | 85.30 | 94.95 |
| **PURCHASE** | 64.08 | 57.67 | 64.10 | 65.92 | 83.13 |
| **TEXAS** | 79.43 | 75.04 | 80.36 | 79.68 | 87.29 |
| **Additional Datasets** | | | | | |
| **MNIST** | 54.30 | 53.60 | 54.30 | 51.66 | 79.37 |
| **FMNIST** | 57.90 | 55.50 | 58.00 | 53.32 | 72.73 |
| **USPS** | 54.13 | 52.25 | 54.65 | 52.43 | 67.56 |
| **FLOWER** | 68.81 | 69.41 | 61.06 | 59.00 | 78.63 |
| **MEPS** | 61.73 | 57.74 | 61.75 | 55.92 | 79.78 |
| **CREDIT** | 57.18 | 56.39 | 57.18 | 52.67 | 78.04 |
| **CENSUS** | 55.95 | 56.14 | 55.90 | 52.91 | 77.38 |

**Remark.** *Prior work on MIAs measure success by reporting balanced accuracy [59, 61, 54] as in Table 5.3. While the balanced MIA accuracy appears to be high for several datasets, not all the above MIAs are necessarily effective in terms of precision (which measure the extent of false positive) and recall (which measure the extent of false negatives) [55]. Hence, a better understanding of the effectiveness of MIAs should consider precision and recall (c.f. Chapter 6: Section 6.3).*

# Chapter 6

# Assessing Effectiveness of SHAPr to Measure MIAs

Our goal is to evaluate how well SHAPr correlates with the success rate of Membership Inference Attacks (MIAs) (effectiveness requirement **R3**) while comparing with Song's Privacy Risk Scores (SPRS). We evaluate the effectiveness of both the membership privacy risk metrics using the MIA predictions from $I_{ment}$, $I_{corr}$, $I_{conf}$, $I_{label}$ and $I_{shadow}$ as the ground truth for computing F1 score, precision and recall.

To compute the F1 score, precision and recall, we threshold SHAPr at zero (Section 4). For SPRS, we use 0.5 as the threshold from their paper [61]. Song and Mittal [61] consider different threshold values: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0] where we note that F1 score is the best for 0.5. Furthermore, SPRS scores indicate the membership likelihood for a training data record, hence, 0.5 is a meaningful threshold for such probability scores.

For evaluating and comparing the effectiveness of SHAPr and SPRS, we first evaluate with respect to F1 score (Section 6.1) which accounts for both precision and recall (hence measures both false positive and false negatives). We then argue that recall is more important for evaluating privacy risk metrics that precision and compare SHAPr and SPRS using recall (Section 6.2). Based on the results in Section 6.1 and Section 6.2, we discuss some common observations (Section 6.3).

## 6.1 Evaluation using F1-Score

We report the mean F1 score and corresponding standard deviations across ten runs for each MIA: $I_{ment}$, $I_{corr}$, $I_{conf}$, $I_{label}$ and $I_{shadow}$ in Table 6.1.

The results are color-coded: 1) orange indicates that SPRS and SHAPr are comparable (similar mean and small standard deviation); 2) red indicates that SPRS outperformed SHAPr 3); and green indicates that SHAPr outperformed SPRS. For each dataset, we repeated the experiment ten times. We report the statistical significance of this difference (corresponding p-value of a student t-test). Our null hypothesis was that both sets of results came from the distribution with the same mean. For p-value $< 0.05$ there is enough evidence to say that one metric outperforms the other. For p-value $< 0.01$, the confidence with which we can reject the null hypothesis is even stronger. Otherwise (p-value $> 0.05$) we do not have enough evidence to say that one metric consistently outperformed the other.

We note that SPRS outperforms SHAPr on majority of the datasets for three of the MIAs: $I_{ment}$, $I_{corr}$ and $I_{conf}$. However, the performance of SHAPr is either comparable or better than SPRS for two of the attacks $I_{label}$ and $I_{shadow}$. We discuss this difference in performance between SHAPr and SPRS across the five attacks in Section 6.3.

## 6.2 Evaluation using Recall

Having shown that SHAPr's F1 score is better for majority of the MIAs on majority of the datasets compared to SPRS, we want to focus on our evaluation with recall. As indicated in Section 5.4, recall is more important than precision for evaluating membership privacy risk metrics as it constitutes erring on the safe side.

We report the mean precision, recall and their corresponding standard deviations computed over ten runs for each MIA: $I_{ment}$ in Table 6.2, $I_{corr}$ in Table 6.3, $I_{conf}$ in Table 6.4, $I_{label}$ in Table 6.5 and $I_{shadow}$ in Table 6.6. The color coding for all the tables follow the same pattern as Table 6.1 which is described above in Section 6.1.

Similar to results obtained for F1 scores (Table 6.1), we note that SPRS outperforms SHAPr on majority of the datasets for three of the MIAs: $I_{ment}$, $I_{corr}$ and $I_{conf}$. This indicates that the precision in most cases is comparable and the recall varies for the two metrics across different MIAs and different datasets. Hence, evaluating the effectiveness of the two metrics using recall is indeed helpful. Additionally, similar to results for F1 scores,

Table 6.1: F1 scores to compare effectiveness of SHAPR and SPRS with respect to different MIAs. orange indicates comparable results, red indicates SPRS outperforms SHAPr and green indicates SHAPr outperforms SPRS.

| Dataset | $I_{ment}$ | | $I_{corr}$ | | $I_{conf}$ | | $I_{label}$ | | $I_{shadow}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SPRS | SHAPR | SPRS | SHAPR | SPRS | SHAPR | SPRS | SHAPR | SPRS | SHAPR |
| **SPRS Datasets** | | | | | | | | | | |
| **LOCATION** | 0.94 ± 0.02 | 0.90 ± 0.02 | 0.97 ± 0.01 | 0.92 ± 0.01 | 0.95 ± 0.02 | 0.90 ± 0.03 | 0.93 ± 0.01 | 0.97 ± 0.00 | 0.93 ± 0.01 | 0.97 ± 0.01 |
| **PURCHASE** | 0.89 ± 0.01 | 0.89 ± 0.01 | 0.90 ± 0.01 | 0.89 ± 0.00 | 0.82 ± 0.02 | 0.89 ± 0.01 | 0.87 ± 0.01 | 0.90 ± 0.01 | 0.83 ± 0.00 | 0.86 ± 0.00 |
| **TEXAS** | 0.95 ± 0.02 | 0.83 ± 0.01 | 0.97 ± 0.00 | 0.82 ± 0.00 | 0.96 ± 0.01 | 0.83 ± 0.01 | 0.90 ± 0.01 | 0.91 ± 0.00 | 0.87 ± 0.00 | 0.86 ± 0.01 |
| **Additional Datasets** | | | | | | | | | | |
| **MNIST** | 0.72 ± 0.00 | 0.96 ± 0.00 | 0.73 ± 0.00 | 0.97 ± 0.01 | 0.73 ± 0.01 | 0.97 ± 0.00 | 0.71 ± 0.04 | 0.97 ± 0.03 | 0.82 ± 0.04 | 0.98 ± 0.00 |
| **FMNIST** | 0.98 ± 0.00 | 0.94 ± 0.00 | 0.99 ± 0.00 | 0.94 ± 0.00 | 0.99 ± 0.00 | 0.94 ± 0.00 | 0.91 ± 0.04 | 0.88 ± 0.04 | 0.85 ± 0.06 | 0.94 ± 0.01 |
| **USPS** | 0.77 ± 0.15 | 0.86 ± 0.10 | 0.82 ± 0.01 | 0.98 ± 0.00 | 0.81 ± 0.03 | 0.95 ± 0.06 | 0.89 ± 0.05 | 0.96 ± 0.02 | 0.79 ± 0.04 | 0.87 ± 0.01 |
| **FLOWER** | 0.89 ± 0.00 | 0.96 ± 0.00 | 0.87 ± 0.05 | 0.97 ± 0.00 | 0.86 ± 0.00 | 0.96 ± 0.00 | 0.98 ± 0.01 | 0.98 ± 0.01 | 0.88 ± 0.00 | 0.88 ± 0.01 |
| **MEPS** | 0.96 ± 0.01 | 0.90 ± 0.04 | 0.94 ± 0.03 | 0.93 ± 0.00 | 0.96 ± 0.01 | 0.91 ± 0.04 | 0.93 ± 0.04 | 0.93 ± 0.01 | 0.85 ± 0.03 | 0.83 ± 0.01 |
| **CREDIT** | 0.93 ± 0.03 | 0.89 ± 0.02 | 0.97 ± 0.04 | 0.93 ± 0.01 | 0.98 ± 0.05 | 0.89 ± 0.02 | 0.91 ± 0.03 | 0.93 ± 0.01 | 0.83 ± 0.02 | 0.81 ± 0.01 |
| **CENSUS** | 0.97 ± 0.02 | 0.90 ± 0.01 | 0.99 ± 0.00 | 0.91 ± 0.01 | 0.99 ± 0.00 | 0.90 ± 0.01 | 0.89 ± 0.04 | 0.90 ± 0.03 | 0.83 ± 0.01 | 0.81 ± 0.02 |

Table 6.2: Comparing effectiveness of SHAPr and SPRS with respect to $I_{ment}$. orange indicates comparable results, red indicates SPRS outperforms SHAPr and green indicates SHAPr outperforms SPRS.

| **Dataset** | **Precision** | | **Recall** | |
|---|---|---|---|---|
| | SPRS | SHAPr | SPRS | SHAPr |
| **SPRS Datasets** | | | | |
| **LOCATION** | 0.94 ± 0.06 | 0.95 ± 0.06 | 0.95 ± 0.02 | 0.87 ± 0.01 |
| **PURCHASE** | 0.98 ± 0.02 | 0.98 ± 0.03 | 0.82 ± 0.02 | 0.81 ± 0.01 |
| **TEXAS** | 0.93 ± 0.05 | 0.97 ± 0.02 | 0.96 ± 0.01 | 0.73 ± 0.03 |
| **Additional Datasets** | | | | |
| **MNIST** | 0.99 ± 0.00 | 0.99 ± 0.00 | 0.57 ± 0.01 | 0.94 ± 0.00 |
| **FMNIST** | 0.99 ± 0.01 | 0.99 ± 0.01 | 0.98 ± 0.03 | 0.89 ± 0.03 |
| **USPS** | 0.79 ± 0.20 | 0.77± 0.23 | 0.76 ± 0.07 | 0.98 ± 0.01 |
| **FLOWER** | 0.98 ± 0.01 | 0.98 ± 0.01 | 0.81 ± 0.04 | 0.94 ± 0.01 |
| **MEPS** | 0.96 ± 0.02 | 0.91 ± 0.07 | 0.96 ± 0.01 | 0.91 ± 0.01 |
| **CREDIT** | 0.88 ± 0.05 | 0.87 ± 0.04 | 0.98± 0.05 | 0.92 ± 0.02 |
| **CENSUS** | 0.94 ± 0.03 | 0.93 ± 0.02 | 0.99 ± 0.00 | 0.87 ± 0.02 |

Table 6.3: Comparing effectiveness of SHAPr and SPRS with respect to $I_{corr}$. orange indicates comparable results, red indicates SPRS outperforms SHAPr and green indicates SHAPr outperforms SPRS.

| **Dataset** | **Precision** | | **Recall** | |
|---|---|---|---|---|
| | SPRS | SHAPr | SPRS | SHAPr |
| **SPRS Datasets** | | | | |
| **LOCATION** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.94 ± 0.02 | 0.86 ± 0.01 |
| **PURCHASE** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.81 ± 0.01 | 0.80 ± 0.00 |
| **TEXAS** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.95 ± 0.01 | 0.70 ± 0.00 |
| **Additional Datasets** | | | | |
| **MNIST** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.58 ± 0.02 | 0.94 ± 0.00 |
| **FMNIST** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.98 ± 0.00 | 0.89 ± 0.00 |
| **USPS** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.69 ± 0.02 | 0.97 ± 0.00 |
| **FLOWER** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.77 ± 0.11 | 0.95 ± 0.02 |
| **MEPS** | 1.00 ± 0.00 | 0.99 ± 0.01 | 0.89 ± 0.05 | 0.88 ± 0.00 |
| **CREDIT** | 0.99 ± 0.00 | 0.98 ± 0.01 | 0.95 ± 0.07 | 0.88 ± 0.01 |
| **CENSUS** | 1.00 ± 0.00 | 0.98 ± 0.01 | 0.98 ± 0.01 | 0.84 ± 0.01 |

for two of the MIAs: $I_{label}$ and $I_{shadow}$, the performance of SHAPr is either comparable or better than SPRS across different datasets. We discuss this difference in performance of SHAPr and SPRS across different MIAs, later in Section 6.3.

Table 6.4: Comparing effectiveness of SHAPr and SPRS with respect to $I_{conf}$. orange indicates comparable results, red indicates SPRS outperforms SHAPr and green indicates SHAPr outperforms SPRS.

| **Dataset** | **Precision** | | **Recall** | |
|---|---|---|---|---|
| | SPRS | SHAPr | SPRS | SHAPr |
| **SPRS Datasets** | | | | |
| **LOCATION** | 0.94 ± 0.07 | 0.94 ± 0.07 | 0.95 ± 0.02 | 0.87 ± 0.02 |
| **PURCHASE** | 0.98 ± 0.02 | 0.98 ± 0.03 | 0.82 ± 0.02 | 0.81 ± 0.01 |
| **TEXAS** | 0.94 ± 0.05 | 0.98 ± 0.02 | 0.96± 0.01 | 0.73 ± 0.03 |
| **Additional Datasets** | | | | |
| **MNIST** | 0.99 ± 0.01 | 1.00 ± 0.00 | 0.58 ± 0.02 | 0.95 ± 0.00 |
| **FMNIST** | 1.00 ± 0.00 | 0.99 ± 0.01 | 0.99 ± 0.01 | 0.89 ± 0.00 |
| **USPS** | 0.93 ± 0.12 | 0.93 ± 0.14 | 0.71 ± 0.05 | 0.98 ± 0.01 |
| **FLOWER** | 0.96 ± 0.06 | 0.97 ± 0.05 | 0.78 ± 0.11 | 0.96 ± 0.01 |
| **MEPS** | 0.96 ± 0.02 | 0.91 ± 0.07 | 0.96 ± 0.01 | 0.91 ± 0.01 |
| **CREDIT** | 0.87 ± 0.05 | 0.87 ± 0.04 | 0.98 ± 0.05 | 0.92 ± 0.02 |
| **CENSUS** | 0.94 ± 0.03 | 0.93 ± 0.02 | 0.99 ± 0.00 | 0.87 ± 0.02 |

Table 6.5: Comparing effectiveness of SHAPr and SPRS with respect to $I_{label}$. orange indicates comparable results, red indicates SPRS outperforms SHAPr and green indicates SHAPr outperforms SPRS.

| **Dataset** | **Precision** | | **Recall** | |
|---|---|---|---|---|
| | SPRS | SHAPr | SPRS | SHAPr |
| **SPRS Datasets** | | | | |
| **LOCATION** | 0.98 ± 0.01 | 0.98 ± 0.01 | 0.89 ± 0.01 | 0.96 ± 0.01 |
| **PURCHASE** | 0.93 ± 0.01 | 0.94 ± 0.01 | 0.81 ± 0.01 | 0.87 ± 0.00 |
| **TEXAS** | 0.98 ± 0.01 | 0.99 ± 0.01 | 0.84 ± 0.02 | 0.84 ± 0.01 |
| **Additional Datasets** | | | | |
| **MNIST** | 0.99 ± 0.01 | 0.99 ± 0.01 | 0.56 ± 0.02 | 0.96 ± 0.01 |
| **FMNIST** | 0.84 ± 0.07 | 0.83 ± 0.10 | 0.99 ± 0.00 | 0.94 ± 0.04 |
| **USPS** | 0.94 ± 0.04 | 0.93 ± 0.04 | 0.85 ± 0.07 | 0.99 ± 0.01 |
| **FLOWER** | 0.97 ± 0.03 | 0.97 ± 0.03 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| **MEPS** | 0.96 ± 0.03 | 0.97 ± 0.03 | 0.90 ± 0.08 | 0.91 ± 0.02 |
| **CREDIT** | 0.95 ± 0.02 | 0.95 ± 0.03 | 0.88 ± 0.06 | 0.90 ± 0.01 |
| **CENSUS** | 0.86 ± 0.09 | 0.87 ± 0.09 | 0.93 ± 0.03 | 0.93 ± 0.05 |

While we focus on recall, we additionally check for datasets where there is degenerate precision. We note that precision is high across all the datasets and MIAs. For $I_{label}$, we note that precision is lower (around 0.77) compared to other MIAs but the numbers are still comparable to SPRS.

Table 6.6: Comparing effectiveness of SHAPr and SPRS with respect to $I_{shadow}$. orange indicates comparable results, red indicates SPRS outperforms SHAPr and green indicates SHAPr outperforms SPRS.

| Dataset | Precision | | Recall | |
|---|---|---|---|---|
| | SPRS | SHAPr | SPRS | SHAPr |
| **SPRS Datasets** | | | | |
| **LOCATION** | 0.98 ± 0.01 | 0.98 ± 0.01 | 0.89 ± 0.01 | 0.96 ± 0.00 |
| **PURCHASE** | 0.86 ± 0.00 | 0.85 ± 0.00 | 0.81 ± 0.01 | 0.86 ± 0.00 |
| **TEXAS** | 0.89 ± 0.01 | 0.89 ± 0.01 | 0.85 ± 0.01 | 0.84 ± 0.00 |
| **Additional Datasets** | | | | |
| **MNIST** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.55 ± 0.05 | 0.95 ± 0.02 |
| **FMNIST** | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.74 ± 0.09 | 0.89 ± 0.01 |
| **USPS** | 0.78 ± 0.01 | 0.78 ± 0.01 | 0.80 ± 0.09 | 0.99 ± 0.00 |
| **FLOWER** | 0.79 ± 0.01 | 0.79 ± 0.01 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| **MEPS** | 0.82 ± 0.01 | 0.79 ± 0.01 | 0.90 ± 0.07 | 0.87 ± 0.00 |
| **CREDIT** | 0.80 ± 0.01 | 0.78 ± 0.00 | 0.86 ± 0.05 | 0.84 ± 0.01 |
| **CENSUS** | 0.78 ± 0.02 | 0.77 ± 0.01 | 0.88 ± 0.03 | 0.86 ± 0.05 |

## 6.3 Observations and Discussions

Having indicated the overall results on comparing the effectiveness of SHAPr with SPRS in Section 6.1 and 6.2, we now discuss some specific observations made from both evaluations.

**Why is the performance of SPRS better than SHAPr on $I_{ment}$, $I_{corr}$ and $I_{conf}$?** SPRS's better effectiveness over SHAPr with respect to $I_{ment}$ is possibly because SPRS uses modified entropy over the model predictions to generate their membership privacy risk scores. This potentially gives them an advantage over SHAPr to correctly identify MIA predictions on $I_{ment}$. Furthermore, on using other MIAs as ground truth, we observe that SHAPr outperforms SPRS further confirming that SPRS "overfits" to $I_{ment}$ resulting in its better performance. We note that the performance of $I_{ment}$ is similar to $I_{corr}$ and $I_{conf}$ which we identify as the reason for SPRS to perform better on those MIAs as well.

**SPRS outperforms SHAPr on some datasets on $I_{ment}$, $I_{corr}$ and $I_{conf}$.** For evaluation with F1 score (Table 6.1) and recall (Tables 6.2, 6.3, 6.4, 6.5 and 6.6), we observe that on three datasets: LOCATION, TEXAS, FMNIST and CENSUS; SPRS outperforms SHAPr on $I_{ment}$, $I_{corr}$ and $I_{conf}$. Additionally, SPRS outperforms on CENSUS compared to SHAPr on all MIAs other than $I_{label}$. We believe that there are dataset specific properties which result in one metric outperforming the other over these datasets. We investigated different hypotheses to find patterns across these datasets from distributions

(Appendix A), how dataset size influences the scores, type of dataset (tabular vs. image); but we could not confirm these hypotheses. Generally, identifying which of the many dataset specific properties influence the final result, in the context of Machine Learning (ML), is time consuming and non-trivial. Addressing the issue can help identify which metric would be more appropriate on a specific dataset. We currently do not have any valid conjecture for better performance of SPRS over some of the datasets on $I_{ment}$, $I_{corr}$ and $I_{conf}$. Hence, we leave this for future work.

**Remark.** *Despite* SPRS *outperforming* SHAPr *on majority of the datasets* $I_{ment}$, $I_{corr}$ *and* $I_{conf}$, *we note that the recall of* SHAPr *on those datasets is still high for it to be used effectively as a membership privacy risk metric.*

**MIAs are poor compared to Optimal MIA ($I_{ideal}$)** Recall from Chapter 3: Section 3.4 that an $I_{ideal}$ perfectly distinguishes training data records as members and non-training data records as non-members. We briefly described how current MIAs proposed in literature are ineffective in matching $I_{ideal}$ . We now show this empirically in Table 6.7 by computing the proportion of training data records predicted as members (indicated as "Mem") and proportion of non-training data records which are predicted as non-members (indicated as "NMem"). We report the results for all the five MIAs.

Table 6.7: "Mem" indicates the % of training data records correctly predicted as members. "NMem" indicates the % of testing data records correctly predicted as non-members. Based on extent correctly predict non-training as non-members, we classify MIAs' effectiveness into three classes: red indicates "Ineffective" with <25% as "NMem", orange indicates "Moderately Effective" with 25-50% "NMem", and green indicates >50% as "NMem".

| **Dataset** | $I_{ment}$ | | $I_{corr}$ | | $I_{conf}$ | | $I_{label}$ | | $I_{shadow}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mem | NMem | Mem | NMem | Mem | NMem | Mem | NMem | Mem | NMem |
| **SPRS Datasets** | | | | | | | | | | |
| **LOCATION** | 96.03 | 81.13 | 100.00 | 42.63 | 96.00 | 79.80 | 96.45 | 76.34 | 97.36 | 91.70 |
| **PURCHASE** | 95.11 | 27.58 | 100.00 | 13.21 | 97.45 | 27.45 | 93.84 | 44.37 | 82.36 | 84.27 |
| **TEXAS** | 94.26 | 66.33 | 100.00 | 46.98 | 94.38 | 66.19 | 94.24 | 66.43 | 85.36 | 82.58 |
| **Additional Datasets** | | | | | | | | | | |
| **MNIST** | 98.85 | 3.75 | 99.99 | 1.99 | 98.93 | 4.27 | 92.56 | 5.92 | 99.45 | 51.33 |
| **FMNIST** | 99.08 | 16.48 | 100.00 | 10.55 | 99.01 | 16.55 | 93.45 | 12.48 | 97.79 | 47.67 |
| **USPS** | 69.16 | 37.83 | 99.96 | 4.54 | 67.60 | 38.45 | 63.19 | 54.48 | 77.68 | 77.37 |
| **FLOWER** | 96.88 | 24.21 | 100.00 | 11.33 | 97.06 | 23.66 | 93.41 | 26.56 | 78.44 | 77.35 |
| **MEPS** | 91.63 | 25.79 | 99.44 | 16.64 | 91.63 | 26.34 | 84.34 | 27.93 | 80.58 | 78.10 |
| **CREDIT** | 88.29 | 32.75 | 94.55 | 19.72 | 88.29 | 31.36 | 82.54 | 29.34 | 78.62 | 76.48 |
| **CENSUS** | 90.20 | 24.32 | 94.56 | 17.78 | 90.20 | 21.48 | 73.46 | 27.29 | 77.82 | 76.94 |

For $I_{ideal}$ we expect 100% "Mem" and 100% "NMem". However, we observe in Table 6.7 that none of the attacks depict the behaviour of $I_{ideal}$ : majority of the MIAs across different datasets have a high "Mem" values (except an outlier for USPS on ), the values for "NMem" are not close to 100%. Hence, we group the MIAs into three categories based on how accurately they predict non-training data records as non-members (i.e., "NMem" values):

- "Ineffective" predict almost all records as members and very few (<25%) non-training data records,
- "Moderately Effective" correctly predicts only some (25-50%) of the non-training data records as non-members,
- "Effective" correctly predicts majority (>50%) of the non-training data records as non-members, compared to other MIAs

Based on extent correctly predict non-training as non-members, we color-coded the MIAs based on their effectiveness compared to $I_{ideal}$ in Table 6.7: red indicates "Ineffective", orange indicates "Moderately Effective", and green indicates "Effective".

Based on the color-coding, we classify $I_{ment}$, $I_{corr}$ and $I_{conf}$ as "Ineffective", $I_{label}$ as "Moderately Effective" and $I_{shadow}$ as "Effective". Across all datasets other than LOCATION and TEXAS, we observe that majority of non-training data records are incorrectly predicted as members. For LOCATION and TEXAS, majority of training data records are correctly predicted as members and testing data records correctly predicted as non-members. ML models do not generalize well to these datasets due to the difficulty of the classification tasks due to which the performance of MIAs is better due to higher overfitting as noted in [52].

**Relation between the effectiveness of SHAPr with MIA effectiveness.** Ideally, membership privacy risk metrics cannot be equally effective against all MIAs and should perform better on using effective MIAs as ground truth and worse on ineffective MIAs. On comparing the performance of SHAPr across each of the three classes of MIAs with SPRS, SHAPr shows poor performance on "Ineffective" MIAs (see Table 6.1, 6.2, 6.3, 6.4), moderate performance on "Moderately Effective" MIAs (see Table 6.1 and 6.5) and good performance on "Effective" MIAs (see Table 6.1 and 6.6). Hence, we note that there is a general relationship between the MIA effectiveness and the effectiveness of either metric in predicting susceptibility to that MIA.

## Summary

SHAPr has high F1 scores and recall values across all datasets and MIAs indicating that SHAPr is indeed an effective membership privacy risk metric (sasatisfies our requirement **R3**). Furthermore, we find that the performance of SHAPr improves with the effectiveness of the MIAs on comparing with SPRS.

# Chapter 7

# Applicability of SHAPr

Having established that SHAPr is an effective membership privacy metric which satisfies requirements **R1-R3**, we evaluate SHAPr to see whether it satisfies the applicability requirement **R4**. First, we evaluate the versatility of SHAPr (Section 7.1) followed by showing how SHAPr can be used to evaluate defences (Section 7.2).

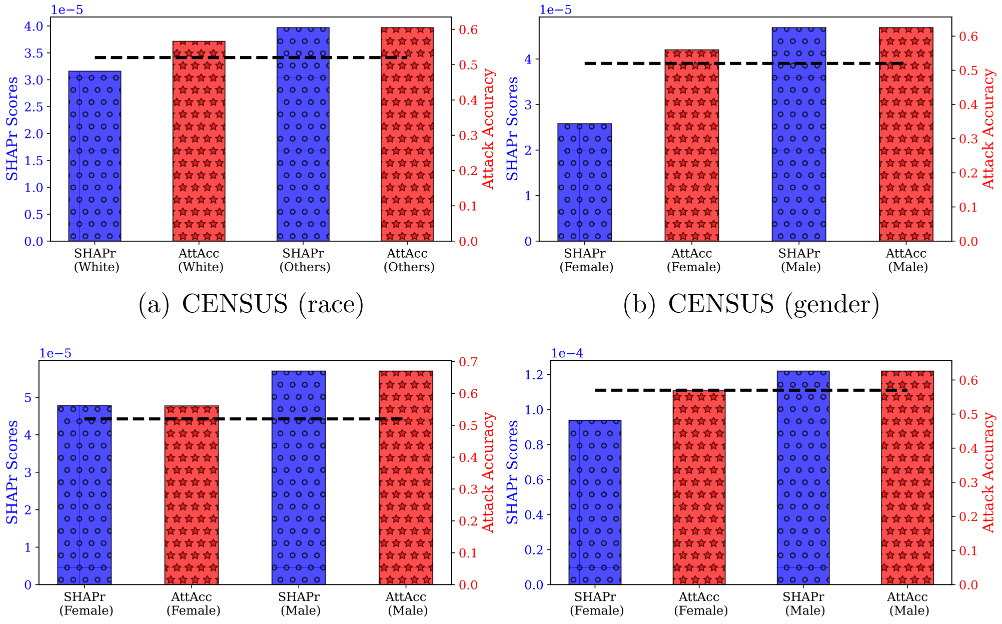
For all the evaluations from this chapter onward, we use $I_{ment}$ as SHAPr does worst on $I_{ment}$ compared to Song's Privacy Risk Scores (SPRS). Hence, choosing $I_{ment}$ gives the least advantage for SHAPr.

## 7.1 Versatility of SHAPr

For showing versatility of SHAPr, we describe two applications of SHAPr, by the virtue of properties of using Shapley values. We specifically explore the applicability in whether SHAPr can estimate the disparity of membership privacy risk across sensitive subgroups (Section 7.1.1) and using SHAPr for data valuation (Section 7.1.2).

### 7.1.1 Privacy Risk of Sensitive Subgroups

Prior work has shown that different subgroups with sensitive attribute (e.g., race or gender) have disparate vulnerability to membership inference attacks (MIAs) [68]. We evaluated whether Song's Privacy Risk Scores (SPRS) and SHAPr can correctly identify this disparity.



(a) CENSUS (race)

(b) CENSUS (gender)

(c) CREDIT (gender)

(d) MEPS (gender)

Figure 7.1: Different subgroups are vulnerable to MIAs to a different extent. blue bars indicate SHAPr scores for different groups (read values from left axis). red bars indicate $I_{ment}$ accuracy for different groups (read values from right axis). Random attack accuracy is indicated by “Black dotted line”.

From the ten datasets, we used only three datasets as they have sensitive attributes: CENSUS, CREDIT, and MEPS. CENSUS has two sensitive attributes, gender and race, while CREDIT and MEPS have gender. For gender, the majority class is “Male” and the minority class is “Female”. For race, “White” is the majority class and “Black” the minority class. We computed the ground truth $I_{ment}$ accuracy, separately for each class, using $I_{ment}$.

Figure 7.1 show that difference in SHAPr scores between different subgroups corresponds to the difference in the ground truth $I_{ment}$ accuracy. The values for SHAPr and $I_{ment}$ accuracy are indicated in Table 7.1. In Table 7.1, green is to indicate scores move in the same direction as the ground truth $I_{ment}$. red is to indicate scores either remain the same or move in opposite direction as the ground truth $I_{ment}$. We find that different subgroups are disparately vulnerable to MIA (indicated under $I_{ment}$ column). SHAPr can capture this difference (all the cells are green) – the scores are higher for subgroups with higher $I_{ment}$ accuracy. SHAPR scores are additive (Property **P2**) – we can compute the membership privacy risk over subgroups by averaging the scores within each subgroup.

Table 7.1: $I_{ment}$ accuracy and SHAPR scores corresponding to Figure 7.1. green is to indicate scores move in the same direction as the ground truth $I_{ment}$. red is to indicate scores either remain the same or move in opposite direction as the ground truth $I_{ment}$.

| **Dataset** | SHAPR | | $I_{ment}$ | |
|---|---|---|---|---|
| **CENSUS** | **Male** | **Female** | **Male** | **Female** |
| | 2.58e-5 | 4.69e-5 | 56.00 | 62.50 |
| | **White** | **Others** | **White** | **Others** |
| | 3.16e-5 | 3.97e-5 | 56.60 | 60.50 |
| **CREDIT** | **Male** | **Female** | **Male** | **Female** |
| | 4.78e-5 | 5.70e-5 | 56.10 | 67.00 |
| **MEPS** | **Male** | **Female** | **Male** | **Female** |
| | 9.39e-5 | 1.22e-4 | 56.90 | 62.60 |

**Remark.** *We additionally evaluated whether* SPRS *scores follow the trend of* $I_{ment}$ *over different subgroups (see Appendix B). We found that* SPRS *does not follow the trend as* SHAPR.

### 7.1.2 Data Valuation

We briefly discuss the application of SHAPR for data valuation. We did not carry out separate experiments but refer to the extensive prior literature on the use of Shapley values for data valuation [30, 31, 21, 32].

Two relevant properties of Shapley values are additivity (Property **P2**) which includes *group rationality*, where the complete utility is distributed among all training data records, and heterogenity (Property **P3**), which indicates *equitable assignment* of model utility to training data records based on their influence. These make Shapley values useful for data valuation [31, 21]. Since SHAPR uses Shapley values, once computed, SHAPR scores can be used directly for data valuation of both individual data records as well as groups of data records.

**Remark.** *We additionally discuss why* SPRS *cannot be used for data valuation in Appendix B.*

### Summary

SHAPr is versatile as it can effectively estimate the disparity in the membership privacy risk across different subgroups identified by a sensitive attribute. Furthermore, it inherits the applications of Shapley values such as data valuation from prior work. This satisfies the versatility requirement **R5**.

## 7.2 Effectiveness of SHAPr to Evaluate Defences

We present the application of SHAPr to evaluate different defences against membership inference attacks (MIAs) that Model Builder ($\mathcal{M}$) can deploy: 1) adding noise to training data records to lower their membership privacy risk to MIAs (Section 7.2.1), 2) retraining Machine Learning (ML) model after removal of vulnerable training data records (Section 7.2.2), and 3) evaluation of regularization based MIA defence (Section 7.2.3).

### 7.2.1 Effectiveness of Adding Noise

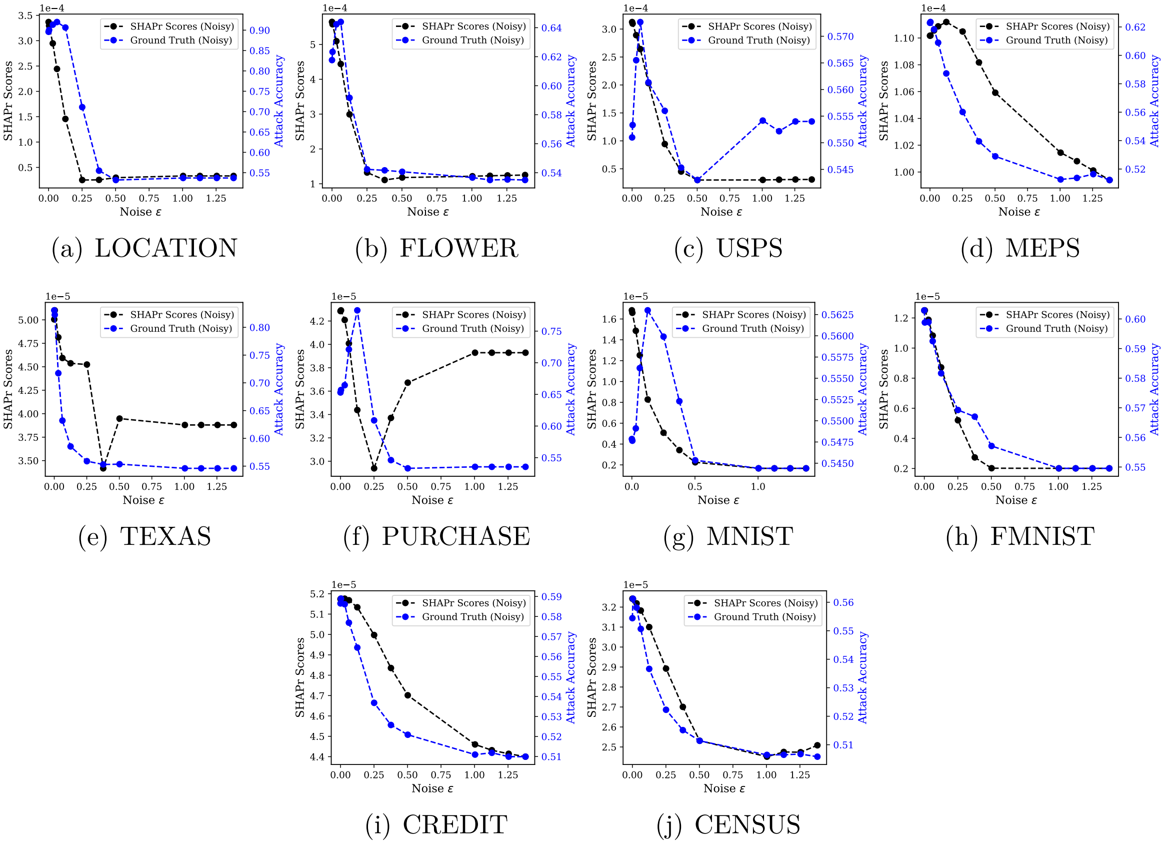
We now focus on evaluating a seemingly plausible way to thwart MIAs is to add noise to ("perturb") data records before training the model. The rationale is that Adversary ($\mathcal{A}dv$), who wants to check the presence of a data record in the training data, is likely to be thwarted because $\mathcal{A}dv$ cannot know what perturbation was added to that record.

We divided the original training set ("No Noise") into two subsets of equal size: 1) a *clean* subset without any noise and 2) a *noisy* subset with perturbed samples. We crafted the noise using FGSM [22], and tested over different values of adversarial noise perturbation budget $\epsilon$ ranging from 1/255 to 352/255 by a factor of two (under $\ell_\infty$).[1]

In the experiment, $\mathcal{A}dv$ used an auxiliary dataset $\mathcal{D}_{aux}$ that was identical to the original "No Noise" dataset used to mount $I_{ment}$. We assumed this complete overlap between $\mathcal{M}$'s and $\mathcal{A}dv$'s data as it corresponds to the scenario in which $\mathcal{M}$ is trying to estimate the privacy risk of their own model.

Our hypothesis is that adding noise to training data records lowers the $I_{ment}$ accuracy. Further, their corresponding SHAPr score is lower as the noisy samples are more difficult to learn and contribute negatively to the model utility. The more noise we add, the lower the SHAPr score, and the lower the $I_{ment}$ accuracy.

[1] Adding Gaussian noise led to similar behavior as with FGSM.



(a) LOCATION (b) FLOWER (c) USPS (d) MEPS

(e) TEXAS (f) PURCHASE (g) MNIST (h) FMNIST

(i) CREDIT (j) CENSUS

Figure 7.2: Adding noise to training data records can lower their susceptibility to MIAs: Visual trend on the noisy dataset shows SHAPR scores (black dotted line) has a strong correlation with a drop in $I_{ment}$ accuracy (blue dotted line).

We first use visualization (Figure 7.2) to obtain an indication as to whether the hypothesis is true. We find that in all datasets the ground truth $I_{ment}$ accuracy decreases (dotted blue line). The corresponding SHAPR scores indeed follow this decreasing trend as well (indicated by dotted black line). Hence, adding noise indeed lowers the risk of the noisy training data records against MIAs (here evaluated specifically for $I_{ment}$ but can be extended to other MIAs as well).

Having obtained a visual indication that the hypothesis is correct, we use quantitative metrics such as correlation coefficient to measure the similarity of SHAPR scores and ground truth $I_{ment}$ accuracy. We compute Pearson's correlation coefficient (ranges between -1 for high negative correlation and +1 for high positive correlation). As the average privacy risk scores should match the ground truth MIA predictions, we expect a positive correlation for the noisy data subset.

The results, as seen in Table 7.2, are color-coded: 1) orange indicates that the correlation is not statistically significant; 2) red indicates that the correlation is statistically significant and negative 3); and green indicates that correlation is positive and statistically significant as expected.

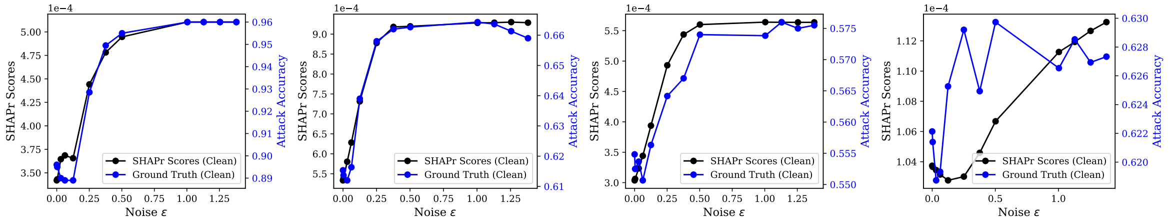
Table 7.2: SHAPr correlates with accuracy on noisy subset as seen by the positive Pearson's Correlation Coefficient (referred as "PCC"), i.e., SHAPr scores follows the decrease in accuracy. orange indicates that correlation is not significant; red indicates that the correlation is significant and negative 3); and green indicates that correlation is positive and significant.

| Dataset | PCC |
|---|---|
| **SPRS Datasets** | |
| **LOCATION** | 0.89 |
| **PURCHASE** | 0.07 |
| **TEXAS** | 0.84 |
| **Additional Datasets** | |
| **MNIST** | 0.60 |
| **FMNIST** | 0.97 |
| **USPS** | 0.43 |
| **FLOWER** | 0.94 |
| **MEPS** | 0.86 |
| **CREDIT** | 0.93 |
| **CENSUS** | 0.97 |

We indeed observe that average SHAPr scores match the ground truth MIA prediction indicated by a statistically significant positive correlation across most of the datasets for the noisy subset. SHAPr is successful in evaluating the effectiveness of adding noise to training data records as SHAPr scores are fine-grained and heterogeneous (Property **P3**) which make them sensitive to noise added to training data records.

**Remark.** *We additionally evaluate the effectiveness of* SPRS *to estimate the impact on membership privacy risk on adding noise to training data records (see Appendix C). We find that their performance is worse than* SHAPr *wherein the* SPRS *scores do not match with the ground truth MIA predictions for the noisy data subset.*

**Is adding noise to training data records an effective defence against MIAs?** In some of the smaller datasets (number of training data records <10000): LOCATION, USPS, FLOWER and MEPS, we note that the clean data points in $\mathcal{D}_{aux}$ become more vulnerable to MIAs as they become more influential to the utility of the model (Figure 7.3). We quantitatively measure the correlation with $I_{ment}$ accuracy using Pearson's correlation coefficient indicated in the parentheses of the caption in Figure 7.3.



(a) LOCATION (0.98) (b) FLOWER (0.98) (c) USPS (0.99) (d) MEPS (0.54)

Figure 7.3: For four datasets, adding noise to some training data records can make them more vulnerable to $I_{ment}$ accuracy increases for clean dataset and their corresponding SHAPr scores increases: Visual trend shows SHAPr scores (black solid line) has a strong correlation with a drop in $I_{ment}$ accuracy (blue solid line). The values in the parentheses indicate the Pearson's correlation coefficient between SHAPr scores and $I_{ment}$ accuracy.

This suggests that the use of adding noise to training data is not a robust defence as it is specific to a dataset and might not work across all the datasets. We conjecture that since these datasets are small (<10000 data records), the impact on the influence of training data records in the clean subset is significant resulting in a consistent increase of SHAPr scores. However, for larger datasets (≥10000 data records) neither the the SHAPr scores nor the $I_{ment}$ accuracy have a consistent trend (Appendix C Figure C.1).

We recommend that $\mathcal{M}$ should compute average SHAPr scores for both clean and noisy datasets separately and decide if the increase in membership privacy risk for clean dataset is acceptable for the application.

## Summary

Adding noise to some of the training data records to lower their membership privacy risk is indeed effective but could potentially result in high risk for remaining training data records depending on the dataset specific properties. This makes it challenging for addition of noise to training data records to be deployed as practical defence.

### 7.2.2 Impact of Data Removal

In data valuation research, it is well-known that removing records with high Shapley values will harm the utility of the model, and removing records with low values will improve it [31, 32]. Hence, it begs the question whether removal of records with high SHAPr scores improves the membership privacy risk of a dataset, by reducing its overall susceptibility to MIAs. To explore this question, we removed a fraction (up to 50%) of records with the highest SHAPr scores. Also, we randomly removed testing data records so as to keep the same number of member and non-member records as in previous experiments.

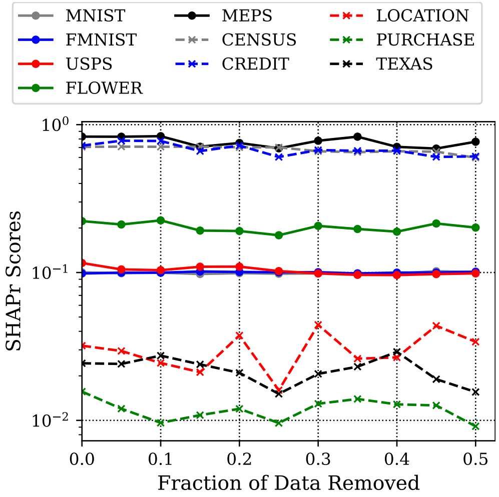


Figure 7.4: Removing a fraction of training data records with high SHAPr scores does not reduce the risk for the remaining records.

Figure 7.4 summarizes the results. Removing an increasing number of records with high SHAPr scores does not necessarily reduce the membership privacy risk for the remaining records. No consistent upward (or downward) trend was visible for the scores of the remaining records. Interestingly, depending on the number of removed samples, the scores fluctuate. A possible explanation is that once risky data records are removed, and a new model is trained using the remaining records, their contribution to the utility of the revised model changes, thereby changing their SHAPr scores.

Long et al. [43] observed a similar result. However, their experiment was limited to a single small dataset (≈ 1000 training data records), and minimal removal (only 20 records from 1.6 million records which is << 1%). Note that 1.6 million training data records was on Naive Bayes classifier and not a deep NN for which Long et al. intractable. Thanks to the superior efficiency of SHAPR, we are able to confirm that this observation holds broadly across larger datasets (10 vs. 1) and for more extensive data removal (up to 50% vs. << 1%).

### Summary

Simply removing highly vulnerable training data records to retrain the model is not an effective defence as the scores will be re-assigned for different training data records based on their influence to the new retrained model. We reiterate the results obtained by Long et al. [43] but on a broader evaluation over ten datasets with higher extent of removal.

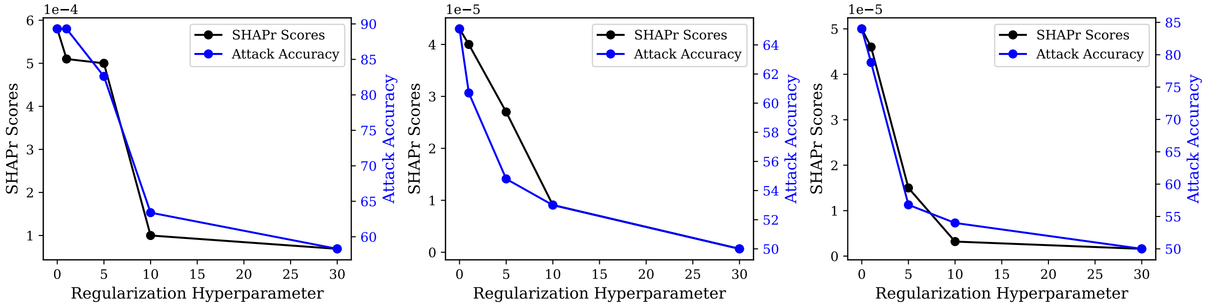
### 7.2.3 Evaluation of L2 Regularization

We evaluate whether SHAPR can be used to verify the effectiveness of L2 regularization as a defence against MIAs. Specifically, the average SHAPR scores across all training data records should decrease on applying a defence to indicate the decrease in the empirical $I_{ment}$ accuracy.

Prior defences such as adversarial regularization [47] and MemGuard [29] have been shown to be ineffective against MIAs [61]. Hence, we focus on L2 regularization previously shown as a valid defence against MIAs [72]. Following Song and Mittal [61], we consider the three datasets vulnerable to MIAs to evaluate SHAPR: LOCATION, PURCHASE and TEXAS. We choose these three SPRS datasets as they are typically used as benchmark in MIAs and defences.

We expect the balanced MIA accuracy for $I_{ment}$ (across both train and test datasets) decreases on increasing regularization. The corresponding average SHAPR scores should decrease to indicate a decrease in the privacy risk corresponding to training data records.

We find from our empirical evaluation that average SHAPR scores indeed decrease along with $I_{ment}$ accuracy. In Figure 7.5, we visually inspect to find that indeed there is a correlation between SHAPR scores and $I_{ment}$ accuracy. We find that SHAPR scores (black solid line) follows the trend with drop in $I_{ment}$ accuracy (blue solid line). In addition to the visual evidence, we quantitatively compute the Pearson's Correlation coefficient between the SHAPR scores and $I_{ment}$ accuracy. We note that there is strong positive correlation is for all the three datasets: LOCATION (0.98), PURCHASE (0.94) and TEXAS (0.99).



(a) LOCATION (b) PURCHASE (c) TEXAS

Figure 7.5: Visual trend shows SHAPr scores (black solid line) has a strong correlation with a drop in $I_{ment}$ accuracy (blue solid line).

## Summary

SHAPr can measure the effectiveness of L2 regularization as a defence against MIAs.

# Chapter 8

# Performance Evaluation of SHAPr

Having shown that SHAPr is effective in estimating the susceptibility of training data records to membership inference attacks while outperforming Song's Privacy Risk Scores (SPRS), we evaluate the efficiency of SHAPr (requirement **R5**). We evaluate whether SHAPr scores can be computed in reasonable time. We ran the evaluation on Intel Core i9-9900K CPU @ 3.60GHz with 65.78GB memory. We use the python metric time() in time library which returns the time in seconds (UTC) since epoch start.

Table 8.1: Performance of SHAPr across different datasets averaged over ten runs.

| **Dataset** | **# Records** | **# Features** | **Execution Time (s)** |
|---|---|---|---|
| **SPRS Datasets** | | | |
| **LOCATION** | 1000 | 446 | 130.77 ± 3.90 |
| **PURCHASE** | 19732 | 600 | 3065.58 ± 19.24 |
| **TEXAS** | 10000 | 6170 | 5506.79 ± 17.47 |
| **Additional Datasets** | | | |
| **MNIST** | 60000 | 784 | 2747.41 ± 22.65 |
| **FMNIST** | 60000 | 784 | 3425.90 ± 34.03 |
| **USPS** | 3000 | 256 | 238.67 ± 1.74 |
| **FLOWER** | 1500 | 2048 | 174.27 ± 11.74 |
| **MEPS** | 7500 | 42 | 732.43 ± 4.95 |
| **CREDIT** | 15000 | 24 | 1852.66 ± 30.92 |
| **CENSUS** | 24000 | 103 | 3718.26 ± 18.25 |

Table 8.1 shows the average execution time for computing SHAPr scores across datasets of different sizes over ten runs. Computation time for SHAPr scores ranges from ≈ 2 mins for LOCATION dataset to ≈ 91 mins for TEXAS. Since the scores are computed once and

designed for Model Builder ($\mathcal{M}$) with substantial computational resources (e.g., GPUs), these execution times are reasonable.

We first compare SHAPR's efficiency with the closely related leave-one-out (LOO) based metric proposed by Long et al. [43]. Long et al.'s naïve LOO scores require training $|D_{tr}|$ additional models (compared to training a single model for SHAPR) [43]. To benchmark, we used a subset of the LOCATION dataset with 100 training data records and found that SHAPR is  100x faster than a naïve LOO based approach: 3640.21 ± 244.08s (LOO) vs. 34.65 ± 1.74s (SHAPR) took only, across only five runs as the execution time for naïve LOO was high. For larger datasets LOO will take unreasonably long time to finish.

On comparing with SPRS, we acknowledge that the computation overhead of SPRS is about ~2x better than SHAPR. We report the results for a few datasets: LOCATION (59.78 ± 0.28), FLOWER (77.56 ± 10.01) and USPS (104.59 ± 6.39). Despite SPRS being faster, we note that SHAPR is more effective on majority of datasets and attacks as evaluated in Chapter 6. Next, we have shown that SHAPR is versatile and can be used to evaluate membership privacy risk with respect to sensitive subgroups (Chapter 7: Section 7.1) where SPRS does not perform well (Appendix B). We later show that for some of the defence evaluations (Chapter 7: Section 7.2.1), SPRS does not perform well (Appendix C). Finally, since both SPRS and SHAPR can be used as tools by $\mathcal{M}$, as long as both their computationally overhead are reasonable, the training time overhead will not be the main bottleneck for their deployability. The metric with more benefits will be more practical for deployment. Hence, SHAPR as membership privacy risk metric has several benefits over SPRS warranting its use for practical applications despite not being faster.

## Summary

SHAPR has reasonable execution times given the hardware resources available to $\mathcal{M}$ making it an efficient metric satisfying requirement **R5**.

# Chapter 9

# Discussions

We discuss alternatives to SHAPr, specifically Influence Functions, and their limitations to be used as a privacy risk metrics (Section 9.1). We then discuss the impact of Backdoors on SHAPr scores and use of Shapley values to detect backdoors by prior work (Section 9.2). We then discuss the connections between differential privacy and membership privacy risk and how SHAPr cannot be used for comparing privacy risk across different models or datasets (Section 9.3). We then discuss the impact of our assumption of complete overlap between training data of target model and adversary's auxiliary data (Section 9.4).

## 9.1 Comparison with Influence Functions

Influence functions [34, 51] were proposed for explaining model predictions. Since these are independent of specific MIAs (satisfying attack-agnostic requirement **R2** similar to SHAPr), they could potentially be used to design an alternative, interpretable (satisfy Property **P1** similar to SHAPr) metric for measuring membership privacy risk. We now explore the viability of such designs.

We implemented Koh et al.'s influence function [34] which assigns an influence score to each individual training data record with respect to each prediction class. We adapt it to estimate membership privacy risk by averaging the scores for each training data record across all classes. This results in computing the overall influence of that test data record as suggested by the authors in their paper [34].

We additionally implemented TracIN [51] which estimates the influence of a training data record by computing the dot product of the loss gradient for training and testing data records which in turn is computed from intermediate models saved during training. Given $|D_{tr}|$ training data records and $|D_{te}|$ testing data records, TracIN computes a score for each training data record corresponding to each testing data record resulting in a $|D_{tr}| \times |D_{te}|$ matrix. To estimate the scores of training data records across the entire test dataset, we averaged the values across all the testing data records for each training data record. This was suggested by the authors in their paper's supplementary material [51].

**Remark.** *Following the suggestion in TracIN [51] and Koh et al. [34], we average Koh's Influence Scores (KIFS) and TracIN scores to obtain individual scores for training data records similar to what we have for* SHAPR *for a fair comparison. However, it is not clear whether the averaging is the best method to obtain the final influence of individual training data records. It is possible to improve both these influence functions by considering functions other than averaging such that the mapping to membership status of training data records from influence scores is more meaningful. We leave this for future work.*

Table 9.1: Effectiveness of blackbox influence functions (KIFS [34]) and TracIN [51] as a metric for membership privacy risk scores with respect to $I_{ment}$. Comparing to SHAPR at zero threshold in Table 6.2, orange indicates comparable results, red indicates poor results and green indicates better results. Computations which took unreasonably long time were omitted indicated by "-".

| **Dataset** | **KIFS** [34] | | **TracIN** [51] | |
|---|---|---|---|---|
| | **Precision** | **Recall** | **Precision** | **Recall** |
| **SPRS Datasets** | | | | |
| **LOCATION** | 0.92 ± 0.01 | 0.48 ± 0.01 | 0.96 ± 0.00 | 0.20 ± 0.00 |
| **PURCHASE** | 0.94 ± 0.00 | 0.51 ± 0.01 | - | - |
| **TEXAS** | 0.91 ± 0.01 | 0.51 ± 0.03 | - | - |
| **Additional Datasets** | | | | |
| **MNIST** | 0.98 ± 0.02 | 0.30 ± 0.18 | - | - |
| **FMNIST** | 0.81 ± 0.08 | 0.49 ± 0.10 | - | - |
| **USPS** | 0.82 ± 0.26 | 0.33 ± 0.10 | 0.75 ± 0.23 | 0.42 ± 0.03 |
| **FLOWER** | 0.97 ± 0.02 | 0.51 ± 0.07 | 0.95 ± 0.03 | 0.46 ± 0.10 |
| **MEPS** | 0.95 ± 0.00 | 0.62 ± 0.05 | 0.96 ± 0.00 | 0.85 ± 0.00 |
| **CREDIT** | 0.85 ± 0.04 | 0.79 ± 0.03 | - | - |
| **CENSUS** | 0.94 ± 0.01 | 0.72 ± 0.12 | - | - |

For evaluation, we compute precision and recall by thresholding Koh et al.'s influence function scores (referred to as KIFS) and TracIN scores. We threshold both the scores at zero as Property **P1** for SHAPR is still applicable where different scores have different semantic meaning (i.e., positive values indicate high susceptibility to MIAs while zero and negative values indicate low susceptibility to MIAs). We then compare with MIA success of modified prediction entropy attack ($I_{ment}$) as the ground truth. Following Section 7, we use $I_{ment}$ as SHAPR does worst on $I_{ment}$ compared to Song's Privacy Risk Scores (SPRS).

We compare these results in Table 9.1 with the results of SHAPR in Table 6.2. We use orange to indicate comparable results, red to indicate poor results and green to indicate better results compared to SHAPR. Overall, both TracIN and KIFS have poor recall across all the datasets compared to SHAPR scores. Both KIFS and TracIN approximate the influence of training data records [34, 51]. KIFS is well defined for convex functions but not for large neural networks with non-convex optimization [10]. Hence, the approximation of influence scores are often erroneous. We conjecture this as a potential reason for the lower recall for KIFS compared to SHAPR.

For TracIn, the overhead of computing per-sample influence was high for large datasets (>7500 training data records) and all the datasets which took more than a day of computation were omitted indicated by "-" in Table 9.1. TracIN stores the model at different epochs during training. For each of the $N_{models}$ models saved during training, the dot product of the gradient loss over each training and testing data record is computed resulting in a complexity of O($N_{models}.|D_{tr}|.|D_{te}|$). The computational overhead for KIFS is also high and in the order of O($|D_{tr}|.|D_{te}|$).

We observe that by the KIFS and TracIN have low recall values on comparing with $I_{ment}$ predictions. Furthermore, their recall is significantly worse compared to SHAPR as well and hence does not satisfy the effectiveness requirement **R3**. Additionally, the high computationally cost of KIFS and TracIN (compared to SHAPR) does not satisfy efficiency requirement **R5**. Hence, our evaluation indicates that both the state-of-the-art influence functions (KIFS and TracIN) are not good candidates to base membership privacy risk metrics on.

## 9.2 Backdoors and Shapley Values

A backdoor to a machine learning model is a set of inputs chosen to manipulate decision boundaries of the model. Backdoors can be used for malicious purposes such as poisoning (e.g. [14]), or to embed watermarks that allow model owners to claim ownership of their model in case it gets stolen [64, 9]. A backdoor is created by changing the label of several training data records [64], by adding artifacts to the training data records themselves (e.g.

overlay text or texture to images [74]), or by introducing out-of-distribution data [9] to the training data. A successfully embedded backdoor is memorised during training, along the primary task of the model. During the verification, a verifier (can either be Model Builder ($\mathcal{M}$) or a third-party judge where $\mathcal{M}$ provides the watermark set to the judge) queries the model and expects matching backdoor predictions.

Backdoors have negative influence on model utility as they introduce noise, and make training more difficult [32]. Hence, their SHAPr scores are low. This has been used as a way for identifying and removing images with watermarks [32].

However, memorization of backdoors is required for successful verification. In other words, backdoors behave differently from other data records in the context of SHAPr: they are, by definition, memorized but unlike other memorized data records, they are likely to have low SHAPr scores. This is not a concern in our setting because $\mathcal{M}$ is the entity that computes SHAPr scores. If a backdoor is inserted intentionally by $\mathcal{M}$ (e.g., for watermarking), then $\mathcal{M}$ will know what they are. If a backdoor was inserted maliciously (e.g., by a training data provider), there is no need to provide any guarantees regarding the SHAPr scores for those records.

## 9.3 SHAPr, Differential Privacy and Membership Privacy Risk

In order to establish the connection between membership privacy risk and differential privacy, we describe the difference between estimating the privacy loss of a specific Machine Learning (ML) model and that of an ML algorithm. Recall from Chapter 2 that an ML model is a sample of an ML algorithm obtained from fixing the training data and training algorithm. Hence, there is a difference in privacy risk of specific ML model and ML algorithm [70]. Privacy risk of an ML model is with respect to the individual training data records in its specific training data. On the other hand, privacy risk of an ML algorithm depends on the randomness of sampling the training data and the training algorithm.

SHAPr is designed to estimate the membership privacy risk for an ML model trained on a *specific* training dataset obtained using a *specific* training algorithm. Hence, SHAPr cannot compare membership privacy risk across *different ML models* trained on *different datasets* with *different training algorithms*. SHAPr scores can only compare relative membership privacy risk of different training data records *within* a dataset for a *specific* model and training algorithm (recall property **P1** where different values SHAPr scores indicate varying susceptibility to MIAs). Furthermore, none of the other privacy metrics (like SPRS [61] or Long et al. [43]) designed for individual membership privacy risk scores can be used to compare the membership privacy risk of two different datasets either.

On the other hand, differential privacy can compute the privacy risk of an ML algorithm by estimating its worst case privacy loss. Differentially private algorithms enforce an upper bound on the privacy risk of an algorithm over *all* ML models with respect to *all* possible training datasets. This is different from the membership privacy risk of an ML model trained with specific training dataset for which SHAPr was designed.

Furthermore, Humphries et al. [27] show that differential privacy makes an assumption that the training data distribution is IID which is not true for real-world datasets which are non-IID. Hence, for such datasets, using differential privacy does not necessarily act as a defence against MIAs. Despite this, differential privacy can potentially be effective as a regularization technique to lower membership privacy risk. However, several prior work indicate that simple ML regularization techniques such as early stopping can achieve a better privacy-utility trade-off than using differential privacy [39]. Hence, we omit the evaluation of differential privacy as an effective defence against MIAs. For evaluation of SHAPr with respect to differential privacy, prior work has indicated theoretically that the distinguishability of Shapley values [32]. We keep the empirical evaluation as future work.

## 9.4 Assumptions of overlap between $D_{aux}$ and $D_{tr}$

In all our experiments in Chapter 6, we consider a complete overlap between $D_{aux}$ and $D_{tr}$ since we consider a $\mathcal{M}$'s perspective. We revisit this assumption, and evaluate effectiveness of SHAPr and SPRS with respect to $I_{ment}$ when there is no overlap between $D_{aux}$ and $D_{tr}$. This is a practical setting from $\mathcal{A}dv$'s perspective for which SPRS was specifically designed.

The results are shown in Table 9.2 where orange indicates comparable results, red indicates that SPRS significantly performs better and green indicates that SHAPr is significantly better. We use the p-value student t-test over ten runs to indicate statistical significance of the results.

We note that unlike in Table 9.2, SHAPr is comparable to SPRS. Specifically, SHAPr outperforms SPRS on four datasets while SPRS outperforms on four datasets. For two of the datasets, the performance is comparable between both metrics. Overall, we do not find a significant difference between the overlapping dataset assumption considered in this work and the practical non-overlapping dataset setting for $I_{ment}$ other than difference for a few datasets (FMNIST, TEXAS and CREDIT).

Table 9.2: SHAPR is comparable to SPRS (evaluated in the $\mathcal{A}dv$'s threat model of [61], representing $\mathcal{A}dv$'s perspective with no overlap between $\mathcal{D}_{aux}$ and target model's train data). Orange indicates comparable results, red indicates SPRS outperforms SHAPR and green SHAPR outperforms SPRS.

| **Dataset** | **Precision** | | **Recall** | |
|---|---|---|---|---|
| | SPRS | SHAPR | SPRS | SHAPR |
| **SPRS Datasets** | | | | |
| **LOCATION** | 0.90 ± 0.01 | 0.89 ± 0.01 | 0.94 ± 0.00 | 0.88 ± 0.01 |
| **PURCHASE** | 0.98 ± 0.02 | 0.98 ± 0.01 | 0.83 ± 0.02 | 0.81 ± 0.02 |
| **TEXAS** | 0.96 ± 0.05 | 0.97 ± 0.05 | 0.78 ± 0.11 | 0.96 ± 0.01 |
| **Additional Datasets** | | | | |
| **MNIST** | 1.00 ± 0.00 | 0.99 ± 0.01 | 0.94 ± 0.02 | 0.95 ± 0.00 |
| **FMNIST** | 1.00 ± 0.00 | 0.99 ± 0.01 | 0.91 ± 0.04 | 0.89 ± 0.01 |
| **USPS** | 0.98 ± 0.02 | 0.99 ± 0.01 | 0.60 ± 0.01 | 0.98 ± 0.00 |
| **FLOWER** | 0.97 ± 0.02 | 0.96 ± 0.04 | 0.82 ± 0.09 | 0.96 ± 0.01 |
| **MEPS** | 0.95 ± 0.02 | 0.91 ± 0.04 | 0.97 ± 0.02 | 0.92 ± 0.01 |
| **CREDIT** | 0.92 ± 0.02 | 0.85 ± 0.04 | 0.86 ± 0.07 | 0.94 ± 0.01 |
| **CENSUS** | 0.96 ± 0.02 | 0.93 ± 0.02 | 1.00 ± 0.00 | 0.86 ± 0.01 |

# Chapter 10

# Related Work

We now cover the prior work on computing the influence of training data records in Machine Learning (ML) models (Section 10.1), membership inference attacks (Section 10.2), other metrics on measuring membership privacy risk in ML (Section 10.3) and finally defences to mitigate membership inference attacks (Section 10.4).

## 10.1 Estimating influence of training data record

Data marketplaces trade training data for ML models. They assign monetary value to data by estimating the influence of each training data record to model utility. A simple approach is inspired from model explanabiility by using Influence functions [34] and TracIN [51] estimate the influence of training data records. However, these approaches are computationally expansive and do not precisely estimate memorization for membership privacy risk (Section 9.1).

Influence functions can be computed as the difference in model prediction with and without each data record in the training dataset [19]. This is computationally expensive. Alternatively, Shapley values extend the above definition by computing the difference in model utility on a test dataset [21]. This was previously explored in the context of data valuation in ML as well [30, 31, 21, 20]. Naïve approaches for computing Shapley values are computationally expensive because their require retraining for each training data record. There are two proposed variants for approximate computation of Shapley values [21]: Monte Carlo (drawing a random permutation of the training data and averaging the marginal contribution of training data records over multiple permutations) and using gradients (computing marginal contribution using gradients for each data record during training). However, these metrics are still computationally expensive and require training multiple models. $K$-NN based Shapley value computation is currently the most efficient approach proposed in literature which is used in this work [30].

## 10.2 Membership Inference Attacks

Several prior works have proposed a range of novel membership inference attacks (MIAs) with the goal of improving attack success over prior work [59, 54, 61, 11, 70]. In addition to image classification and tabular datasets, membership privacy risk has also been explored in graph models [16], text models [45] and generative models [13, 24, 41].

While we consider a practical blackbox threat model, MIAs are also proposed in whitebox threat model where Adversary ($\mathcal{A}dv$) has access to model parameters to compute intermediate outputs (e.g., federated learning) [48, 36]. SHAPR can still be applied to estimate the membership privacy risk to these MIAs as it is attack-agnostic.

Furthermore, membership privacy risk has been evaluated with respect to different aspects of trust in ML. For instance, fairness comes at the cost of membership privacy in ML where the unprivileged subgroups are at higher risk to MIAs [12]. Additionally, training ML models to resist adversarial examples using both empirical approaches (e.g., adversarial training) or certification techniques, make training data records more vulnerable to MIAs [62]. Finally, releasing model explanations can be used by $\mathcal{A}dv$ to perform successful MIAs [58].

## 10.3 Measuring Membership Privacy Risk

Other than Song's Privacy Risk Scores (SPRS) [61] and Long et al. [43] used in this work, Adversary's membership privacy advantage [71, 28] is another metric for evaluating differential privacy mechanisms. However, this is an aggregate metric and estimates membership privacy risk across all data records.

Fisher Information, originally proposed to compute the influence of attributes towards the model utility (for attribute inference attacks), was suggested as a metric to estimate membership privacy risk [23]. However, this is limited to linear models with convex loss which does not apply to the neural networks we consider. Furthermore, computing Fisher information is computationally expensive for large models as it requires inverting a Hessian.

Finally, maximal information leakage [53] was proposed as a membership privacy risk metric which is an upper bound on the privacy risks for the PATE differential privacy framework [50]. However, this information leakage metric is not designed for individual training data records.

## 10.4 Mitigating Membership Privacy Risk

Several defences have been proposed in prior work. Simple regularization for ML models (e.g., L2 regularization [72] and early stopping [40, 61]) have been shown to be effective in lowering membership privacy risk. Specific defences to thwart MIAs include adversarial regularization [47, 69] and MemGuard [29] which are ineffective [61] against stronger MIAs. Currently, knowledge distillation [57] has been shown to be effective against all MIAs that use output prediction confidence. Additionally, combining ensemble of ML models with careful knowledge-distillation resists label only MIAs in addition to confidence based MIAs [66]. In addition to these empirical defences, differential privacy is often used as a defence against MIAs to provide theoretical guarantee on the worst case privacy loss [8]. However, the current differential privacy training face privacy-accuracy trade-offs and simple techniques to improve generalization have shown to provide better trade-offs [40].

# Chapter 11

# Conclusions

We summarize our conjecture and results from our study followed by some discussion of future research directions.

## 11.1 Summary

Membership privacy risk metrics quantify the susceptibility of training data records to membership inference attacks (MIAs). We present desiderata for designing membership privacy risk metrics: **R1** *fine-grained* to assign scores to individual training data records, **R2** *attack-agnostic* where the scores are independently generated from any specific MIAs, **R3** *effective* where the record-level scores correlate to the susceptibility of specific MIAs, **R4** *applicability* of SHAPR for different tasks and evaluate defences; and **R5** *efficient* to compute scores with reasonable overhead for $\mathcal{M}$.

Our conjecture was that Shapley values computed for individual training data records, by virtue of measuring influence on model utility, and hence the extent of memorization, can serve a good membership privacy risk metric while satisfying all the requirements **R1**-**R5**.

We successfully validated our conjecture by presenting SHAPR, based on Shapley values as a membership privacy risk metric. By definition of Shapley values, SHAPR is fine-grained and assigns scores for individual training data records (**R1**) without using any specific MIA (**R2**). We confirm that SHAPR (and by extension Shapley values) indeed serve as an effective membership privacy risk metric to assess susceptibility of different training data records to MIAs as evaluated on ten benchmark datasets across five different state-of-the-art MIAs (**R3**). We find that SHAPR's performance is either comparable or better than Song's Privacy Risk Scores (SPRS) on effective MIAs. Moreover, SHAPR can be used for different applications, e.g., assess the membership privacy risk of different subgroups, estimate the value of data records and evaluate defences against MIAs (satisfying the requirement of applicability of membership privacy risk metric **R4**). Finally, SHAPR can be computed more efficiently compared to naïive leave-one-out (LOO) approach (**R5**).

## 11.2 Future Work

Based on evaluation and analysis of our work, we include some research directions for future work:

- We observed that majority of the MIAs evaluated are imperfect. The design of a practical Optimal Membership Inference Attack ($I_{optimal}$) is an open question. This warrants further study of how SHAPR fits in the context of such an MIA. Furthermore, while $I_{shadow}$ is indeed an effective MIA, whether it is an $I_{optimal}$ is yet to be investigated.

- We used $I_{ment}$ as a baseline for Chapter 7 which puts SHAPR at a disadvantage. However, we can explore the use of more effective MIA, for instance $I_{shadow}$ to re-evaluate the experiments. We expect that most of the observations will be confirmed due to higher effectiveness of SHAPR on $I_{shadow}$.

- We evaluated defences using SHAPR where there is a decrease in model utility on applying a defence. There are additional defences such as knowledge distillation [57] and SELENA [66] with a better privacy-utility trade-off where the direct application of SHAPR is not straightforward. This is due to the use of multiple models trained on disjoint public-private datasets. Using SHAPR for such defences is left for future work.

- While we focus on using peer-reviewed MIAs, there are recent MIAs which focus on better attack utility (not yet peer-reviewed at the time of writing this thesis) [11, 70]. As part of future work, we plan on using these as ground truth for evaluating the effectiveness of SHAPR with the assumption that they are closer to the $I_{optimal}$.

# References

[1] Adult income census dataset. https://archive.ics.uci.edu/ml/datasets/adult. Accessed: 2021-11-27.

[2] Credit dataset. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data). Accessed: 2021-11-27.

[3] Flowers dataset. https://www.kaggle.com/alxmamaev/flowers-recognition. Accessed: 2021-11-27.

[4] Location dataset. https://sites.google.com/site/yangdingqi/home/foursquare-dataset. Accessed: 2021-11-27.

[5] Purchase dataset. https://www.kaggle.com/c/acquire-valued-shoppers-challenge. Accessed: 2021-11-27.

[6] Texas dataset. https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm. Accessed: 2021-11-27.

[7] Usps dataset. https://www.kaggle.com/bistaumanga/usps-dataset. Accessed: 2021-11-27.

[8] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Conference on Computer and Communications Security*, page 308–318, 2016.

[9] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX Security Symposium*, pages 1615–1631, 2018.

[10] Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *arXiv 2006.14651*, 2021.

[11] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *arXiv 2112.03570*, 2021.

[12] Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroSP)*, pages 292–303, 2021.

[13] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. *GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models*, page 343–362. Association for Computing Machinery, New York, NY, USA, 2020.

[14] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[15] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR, 18–24 Jul 2021.

[16] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. Quantifying privacy leakage in graph embedding. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (Mobiquitous 2020)*, pages 1–11, Cyberspace, United States, December 2020.

[17] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. Gecko: Reconciling privacy, accuracy and efficiency in embedded deep learning. In *arXiv 2010.00912*, 2021.

[18] Vasisht Duddu, Sebastian Szyller, and N. Asokan. Shapr: An efficient and versatile membership privacy risk metric for machine learning. In *arXiv 2112.02230.*

[19] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Symposium on Theory of Computing*, STOC 2020, page 954–959, New York, NY, USA, 2020. Association for Computing Machinery.

[20] Amirata Ghorbani, Michael Kim, and James Zou. A distributional framework for data valuation. In Hal Daumé III and Aarti Singh, editors, *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3535–3544. PMLR, 13–18 Jul 2020.

[21] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR, 09–15 Jun 2019.

[22] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *arXiv 1412.6572*, 2015.

[23] Awni Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. In *arXiv 2102.11673*, 2021.

[24] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019:133–152, 01 2019.

[25] High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. Report, European Commission, Brussels, April 2019.

[26] White House. Guidance for regulation of artificial intelligence applications. In *Memorandum For The Heads Of Executive Departments And Agencies*, 2020.

[27] Thomas Humphries, Matthew Rafuse, Lindsey Tulloch, Simon Oya, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. Differentially private learning does not bound membership inference. *arXiv preprint arXiv:2010.12112*, 2020.

[28] Bargav Jayaraman, L. Wang, David E. Evans, and Quanquan Gu. Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies*, 2021:348 – 368, 2021.

[29] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Conference on Computer and Communications Security*, CCS '19, page 259–274, New York, NY, USA, 2019. Association for Computing Machinery.

[30] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. VLDB Endow.*, 12(11):1610–1623, July 2019.

[31] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the shapley value. In Kamalika Chaudhuri and Masashi Sugiyama,

editors, *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1167–1176. PMLR, 16–18 Apr 2019.

[32] Ruoxi Jia, Fan Wu, Xuehui Sun, Jiacen Xu, David Dao, Bhavya Kailkhura, Ce Zhang, Bo Li, and Dawn Song. Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification? In *Conference on Computer Vision and Pattern Recognition*, 2021.

[33] Emre Kazim, Danielle Mendes Thame Denny, and Adriano Koshiyama. AI auditing and impact assessment: according to the uk information commissioner's office. *AI and Ethics*, Feb 2021.

[34] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR, 06–11 Aug 2017.

[35] European Union Law. Art. 35 GDPR data protection impact assessment. In *General Data Protection Regulation (GDPR)*, 2018.

[36] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *USENIX Security*, pages 1605–1622, 2020.

[37] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *USENIX Security*, pages 1605–1622, 2020.

[38] Zheng Li and Yang Zhang. Membership leakage in label-only exposures. In *arXiv 2007.15528*, 2021.

[39] Jiaxiang Liu, Simon Oya, and Florian Kerschbaum. Generalization techniques empirically outperform differential privacy against membership inference. In *arXiv 2110.05524*, 2021.

[40] Jiaxiang Liu, Simon Oya, and Florian Kerschbaum. Generalization techniques empirically outperform differential privacy against membership inference. In *arXiv 2110.05524*, 2021.

[41] Xiyang Liu, Yixi Xu, Shruti Tople, Sumit Mukherjee, and Juan Lavista Ferres. Mace: A flexible framework for membership privacy estimation in generative models. In *arXiv 2009.05683*, 2021.

[42] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. Ml-doctor: Holistic risk assessment of inference attacks against machine learning models. In *arXiv 2102.02551*, 2021.

[43] Yunhui Long, Vincent Bindschaedler, and Carl A. Gunter. Towards measuring membership privacy. In *arXiv 1712.09136*, 2017.

[44] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[45] Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying privacy risks of masked language models using membership inference attacks. In *arXiv 2203.03929*, 2022.

[46] Sasi Kumar Murakonda and Reza Shokri. ML privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. In *Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs)*, 2020.

[47] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Conference on Computer and Communications Security*, CCS '18, page 634–646, New York, NY, USA, 2018. Association for Computing Machinery.

[48] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. *2019 IEEE Symposium on Security and Privacy (SP)*, May 2019.

[49] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1376–1401, Paris, France, 03–06 Jul 2015. PMLR.

[50] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the International Conference on Learning Representations*, 2017.

[51] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc., 2020.

[52] Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7900, 2021.

[53] Sara Saeidian, Giulia Cervia, Tobias J. Oechtering, and Mikael Skoglund. Quantifying membership privacy via information leakage. In *arXiv 2010.05965*, 2020.

[54] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed Systems Security*, 2018.

[55] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. Membership inference attacks are easier on difficult problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14820–14829, 2021.

[56] L. S. Shapley. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, 2016.

[57] Virat Shejwalkar and Amir Houmansadr. Membership privacy for machine learning models through knowledge transfer. *AAAI Conference on Artificial Intelligence*, 35(11):9549–9557, May 2021.

[58] Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 231–241, New York, NY, USA, 2021. Association for Computing Machinery.

[59] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[60] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 587–601, New York, NY, USA, 2017. Association for Computing Machinery.

[61] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, August 2021.

[62] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 241–257, New York, NY, USA, 2019. Association for Computing Machinery.

[63] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Conference on Computer and Communications Security*, CCS '19, page 241–257, New York, NY, USA, 2019. Association for Computing Machinery.

[64] Sebastian Szyller, Buse Gul Atli, Samuel Marchal, and N. Asokan. DAWN: dynamic adversarial watermarking of neural networks. In *arXiv 1906.00830*, 2019.

[65] Elham Tabassi, Kevin J. Burns, M. Hadjimichael, Andres Molina-Markham, and Julian Sexton. A taxonomy and terminology of adversarial machine learning. In *NIST Interagency/Internal Report*, 2019.

[66] Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. In *arXiv 2110.08324*, 2021.

[67] Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022.

[68] Mohammad Yaghini, Bogdan Kulynych, Giovanni Cherubin, and Carmela Troncoso. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. *arXiv preprint arXiv:1906.00389*, 2019.

[69] Ziqi Yang, Bin Shao, Bohan Xuan, Ee-Chien Chang, and Fan Zhang. Defending model inversion and membership inference attacks via prediction purification. In *arXiv 2005.03915*, 2020.

[70] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *arXiv 2111.09679*, 2022.

[71] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018.

[72] Zuobin Ying, Yun Zhang, and Ximeng Liu. Privacy-preserving in defending against membership inference attacks. In *Workshop on Privacy-Preserving Machine Learning in Practice*, PPMLP'20, page 61–63, New York, NY, USA, 2020. Association for Computing Machinery.

[73] Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In Hal Daumé III and Aarti Singh, editors, *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10842–10851. PMLR, 13–18 Jul 2020.

[74] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *ACM Symposium on Information, Computer and Communications Security*, pages 159–172, 2018.

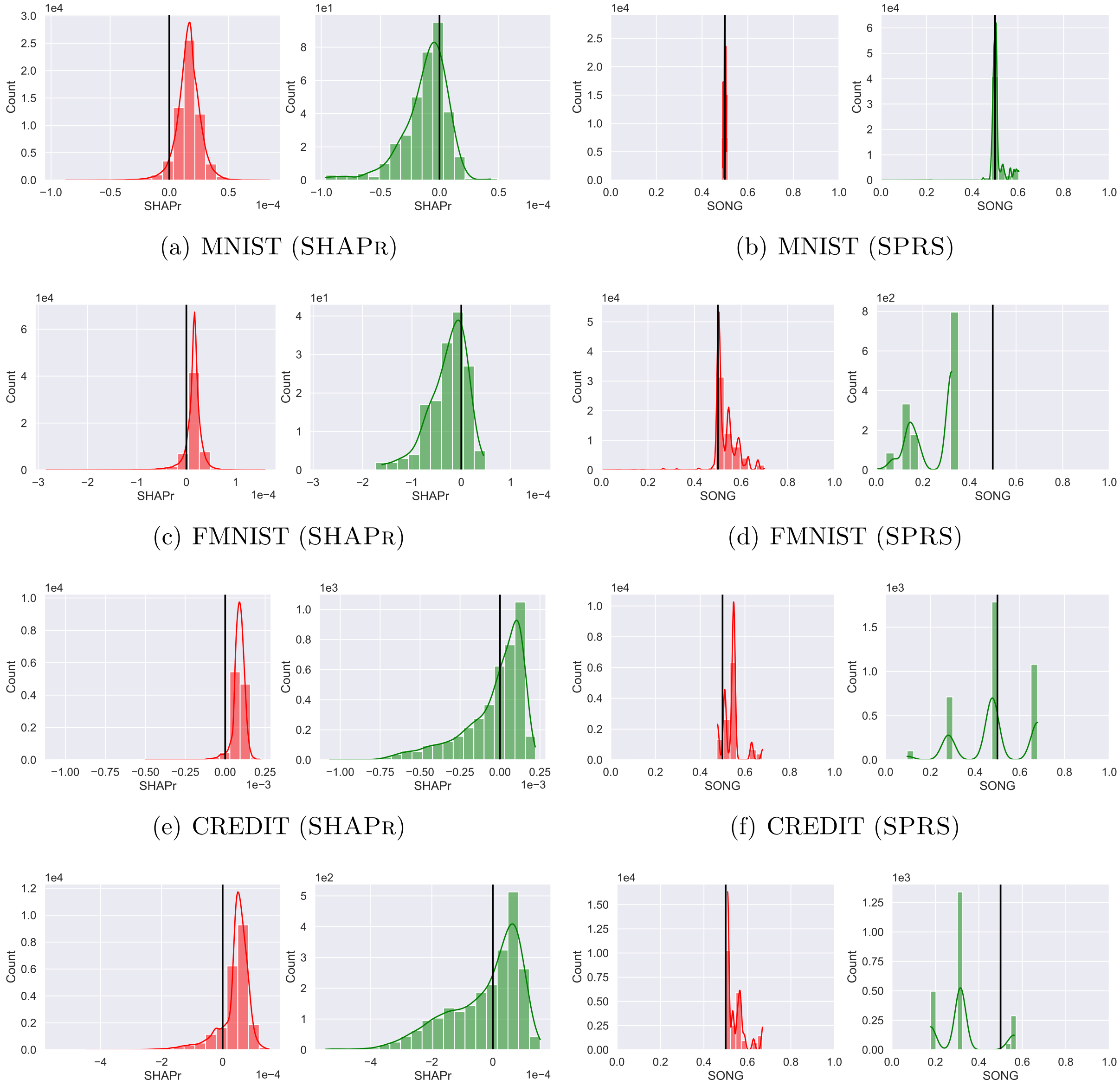
# APPENDICES

# Appendix A

# Distribution of SHAPr and SPRS Scores

We visually compare SHAPr with Song's Privacy Risk Scores (SPRS) by plotting the distribution of SHAPr (in green) and for SPRS (in red) shown in Figure A.3.

For several datasets, we observe that SPRS is centered at 0.5 indicating that the membership likelihood for a large number of training data records is inconclusive. Further, we note that the distribution of SPRS scores is not evenly distributed, with some values correspond to several records while neighboring values correspond to none. We conjecture that this is due to the fixed prior probabilities and estimating the conditional probabilities using shadow models optimized to give the same output for multiple similar data records. Compared to SPRS, SHAPr follows a more even distribution (due to the heterogeneity property **P3**).



(a) MNIST (SHAPr)

(b) MNIST (SPRS)

(c) FMNIST (SHAPr)

(d) FMNIST (SPRS)

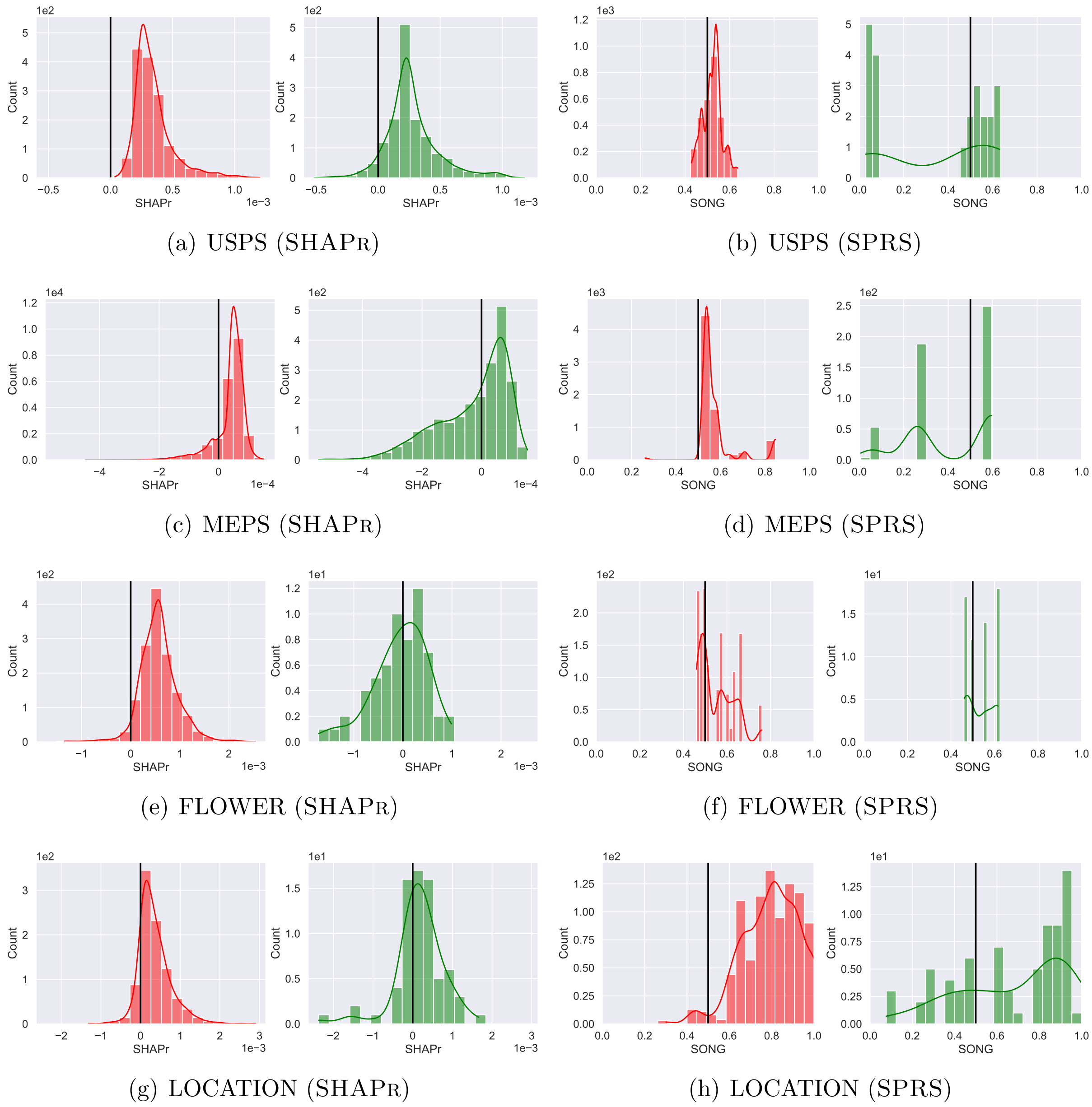
(e) CREDIT (SHAPr)

(f) CREDIT (SPRS)

(g) CENSUS (SHAPr)

(h) CENSUS (SPRS)

Figure A.1: Distributions of SHAPr scores colored based on the $I_{ment}$ attack prediction. We plot the semantic threshold at 0 (solid line).



(a) USPS (SHAPr)

(b) USPS (SPRS)

(c) MEPS (SHAPr)

(d) MEPS (SPRS)

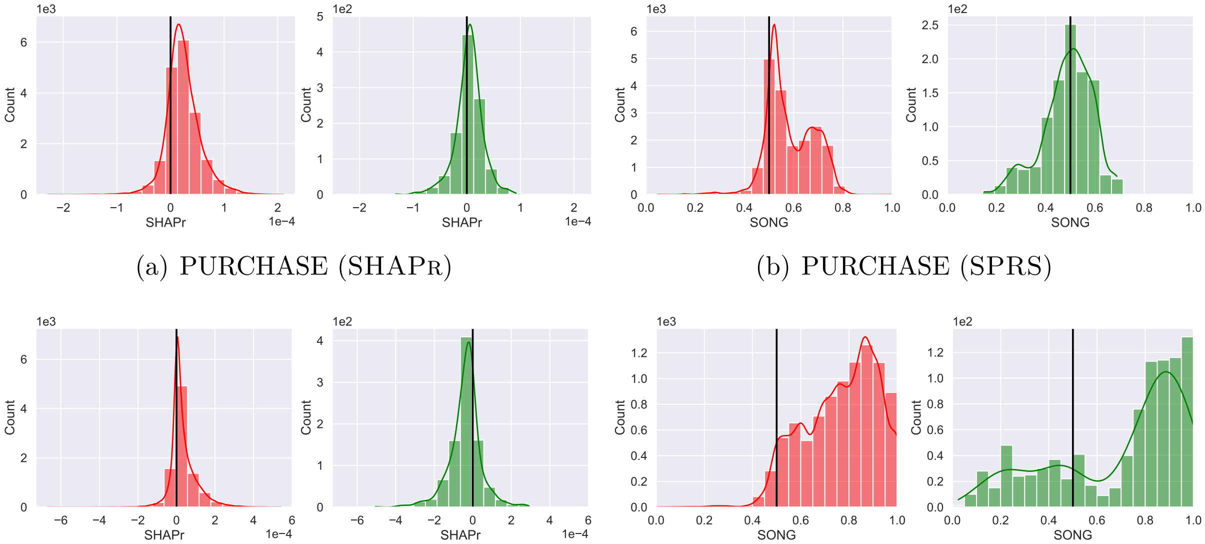
(e) FLOWER (SHAPr)

(f) FLOWER (SPRS)

(g) LOCATION (SHAPr)

(h) LOCATION (SPRS)

Figure A.2: Distributions of SHAPr scores colored based on the $I_{ment}$ attack prediction. We plot the semantic threshold at 0 (solid line).



(a) PURCHASE (SHAPr)

(b) PURCHASE (SPRS)

(c) TEXAS (SHAPr)

(d) TEXAS (SPRS)

Figure A.3: Distributions of SHAPr scores colored based on the $I_{ment}$ attack prediction. We plot the semantic threshold at 0 (solid line).

# Appendix B

# Evaluating SPRS's Versatility

Our goal is to evaluate whether SPRS is versatile. We first evaluate whether SPRS scores corresponds to change accuracy across different sensitive subgroups. We then discuss whether SPRS can be used for data valuation.

**Privacy Risks over Subgroups.** We compute attack over different sensitive subgroups ad average Song's Privacy Risk Scores (SPRS) scores for each of the subgroups. We report the results in Table B.1 which is color-coded: green indicates SPRS scores move in the same direction as the ground truth $I_{ment}$; and red indicates SPRS scores either remain the same or move in opposite direction as the ground truth $I_{ment}$.

Table B.1: Different subgroups are vulnerable to membership inference attacks (MIAs) to a different extent. green indicates SPRS scores move in the same direction as the ground truth $I_{ment}$. red indicates SPRS scores either remain the same or move in opposite direction as the ground truth $I_{ment}$.

| **Dataset** | SPRS | | $I_{ment}$ | |
|---|---|---|---|---|
| **CENSUS** | **Male** | **Female** | **Male** | **Female** |
| | 0.52 | 0.52 | 56.00 | 62.50 |
| | **White** | **Others** | **White** | **Others** |
| | 0.52 | 0.52 | 56.60 | 60.50 |
| **CREDIT** | **Male** | **Female** | **Male** | **Female** |
| | 0.52 | 0.53 | 56.10 | 67.00 |
| **MEPS** | **Male** | **Female** | **Male** | **Female** |
| | 0.57 | 0.54 | 56.90 | 62.60 |

We observe that Song's Privacy Risk Scores (SPRS) scores stay the same for both subgroups across all the datasets, regardless of the  accuracy (indicated in red ). This shows that SPRS is ineffective to estimate disparity of membership privacy risk across different sensitive subgroups.

We note that the average scores are close to 0.5 because majority of the data records have SPRS score of 0.5 (due to a lack of heterogeneity property **P3** as seen in Figure A.3). Additionally, SPRS scores do not satisfy additivity property (property **P2**) as there is no semantically meaningful notion of adding or averaging probability scores. We conjecture that the lack of both heterogeneity and additivity properties make SPRS makes ineffective at this task.

**Data Valuation.** SPRS was not designed to be additive property **P2** and hence cannot guarantee group rationality of scores among training data records. SPRS scores are not heterogeneous (property **P3**) either which does guarantee equitable assignment of privacy risk scores (as shown in Appendix A, Figure A.3). We show the lack of heterogeneity in the Appendix A, visualizing the distribution of SPRS scores (Figure A.3). Given the lack of these properties (heterogeneity, additivity, group rationality, and equitable assignment), we argue that SPRS is unlikely to be applicable for data valuation.

# Appendix C

# Evaluating Effectiveness for Adding Noise

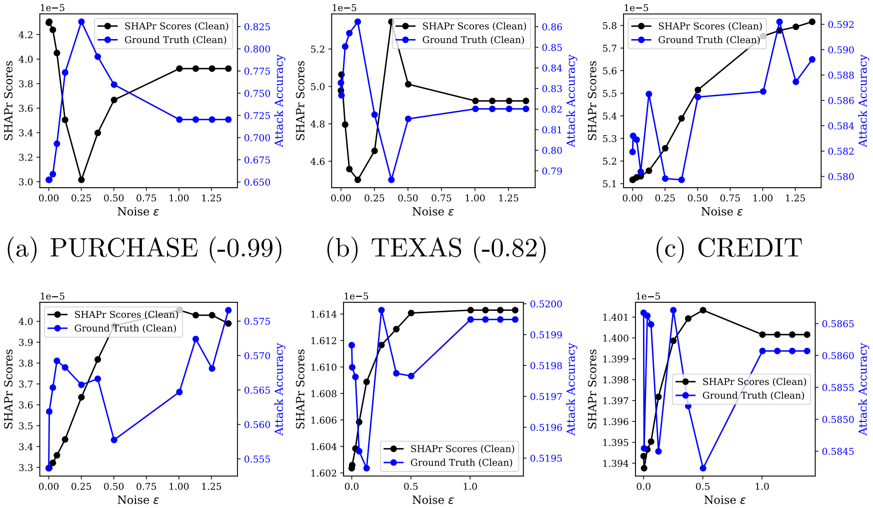
We first show the impact on the SHAPr scores for clean subset on adding noise of some of the training data records. We then discuss how SPRS scores did not follow the trend corresponding to .

**No consistent trend of SHAPr scores for clean subset for larger datasets.** For datasets with training data records ≥10000, we do not find a consistent increasing trend between both the $I_{ment}$ and method scores (Figure C.1). We conjecture that due to the large size of these datasets, the SHAPr scores do not show a significant difference to effect the average scores.

**Impact of Adding Noise to SPRS.** On adding noise to training data records, we expect the SPRS scores to decrease to match with the decrease in . Hence, on computing the Pearson's correlation coefficient, we expect the correlation to be positive.

The results, as seen in Table 7.2, are color-coded: 1) orange indicates that the correlation is not statistically significant; 2) red indicates that the correlation is statistically significant and negative 3); and green indicates that correlation is positive and statistically significant as expected.

In Table C.1, we observe that the average score is not impacted by the added noise indicated by several negative correlations (red) compared to SHAPr (Table 7.2). We observe that there is no consistent correlation between the score and the  accuracy. The lack of sensitivity of SPRS to training data noise can be attributed to clustering of SPRS scores around 0.5 indicating indecisive membership resulting in lack of heterogeneity (property **P3**) as seen Figure A.3 for SPRS's distribution.



(a) PURCHASE (-0.99)

(b) TEXAS (-0.82)

(c) CREDIT

(d) CENSUS

(e) MNIST (0.13)

(f) FMNIST (0.02)

Figure C.1: For larger datasets, visual trend shows SHAPr scores (black solid line) has a strong correlation with a drop in $I_{ment}$ accuracy (blue solid line). Pearson correlation coefficient was low or negative for these datasets as indicated in the parentheses.

Table C.1: Pearson's correlation coefficient (referred to as "PCC") between average SPRS scores and MIA accuracy on noisy subsets. orange indicates that correlation is not significant; red indicates that the correlation is significant and negative 3); and green indicates that correlation is positive and significant.

| Dataset | PCC |
|---|---|
| **LOCATION** | -0.98 |
| **PURCHASE** | -0.58 |
| **TEXAS** | 0.68 |
| **MNIST** | 0.02 |
| **FMNIST** | -0.65 |
| **USPS** | -0.90 |
| **FLOWER** | -0.90 |
| **MEPS** | -0.88 |
| **CREDIT** | -0.85 |
| **CENSUS** | -0.80 |