Data Privacy Examination against Semi-Supervised Learning

Jiadong Lou
jiadong.lou1@louisiana.edu
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

Xu Yuan*
xu.yuan@louisiana.edu
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

Miao Pan
mpan2@uh.edu
University of Houston
Houston, Texas, USA

Hao Wang
haowang@lsu.edu
Louisiana State University
Baton Rouge, Louisiana, USA

Nian-Feng Tzeng
nianfeng.tzeng@louisiana.edu
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

ABSTRACT

Semi-supervised learning, which learns with only a small amount of labeled data while collecting voluminous unlabeled data to aid its training, has achieved promising performance lately, but it also raises a serious privacy concern: Whether a user's data has been collected for use without authorization. In this paper, we propose a novel membership inference method against semi-supervised learning, serving to protect user data privacy. Due to involving both the labeled and unlabeled data, the membership patterns of semi-supervised learning's training data cannot be well captured by the existing membership inference solutions. To this end, we propose two new metrics, i.e., inter-consistency and intra-entropy, tailored specifically to the semi-supervised learning paradigm, able to respectively measure the similarity and calculate the cross-entropy among prediction vectors from the perturbed versions. By exploiting the two metrics for membership inference, our method can dig out membership patterns imprinted on prediction outputs of semi-supervised learning models, thus facilitating effective membership inference. Extensive experiments have been conducted for comparing our method with five rectified baseline inference techniques across four datasets on six semi-supervised learning algorithms. Experimental results exhibit that our inference method achieves over 80% accuracy under each experimental setting, substantially outperforming all baseline techniques. Moreover, our results also reveal that a semi-supervised learning model 1) trained by more effective learning algorithms, 2) possessing better performance, or 3) trained with less labeled data, is more vulnerable to our membership inference.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings**; • **Security and privacy** → **Domain-specific security and privacy architectures**.

Corresponding author.

ASIA CCS '23, July 10–14, 2023, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0098-9/23/07…$15.00
https://doi.org/10.1145/3579856.3590333

KEYWORDS

Semi-supervised learning, membership inference attack, data privacy

ACM Reference Format:
Jiadong Lou, Xu Yuan*, Miao Pan, Hao Wang, and Nian-Feng Tzeng. 2023. Data Privacy Examination against Semi-Supervised Learning. In *ACM ASIA Conference on Computer and Communications Security (ASIA CCS '23), July 10–14, 2023, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3579856.3590333

1 INTRODUCTION

While machine learning has exhibited a great potential in handling different categories of tasks, its reliance on vast labeled data typically hinders widespread deployment in the large-scale application scenarios. Reliable large-sized ground-truth datasets are necessary for supervised learning to achieve decent performance, but data labeling tasks are well known as expensive, labor-intensive, and time-consuming. So far, there is no effective method for labeling large-sized data reliably. On the other hand, semi-supervised learning, which was considered little useful years ago, has made new breakthroughs lately, greatly lessening its label reliance by learning with only a small amount of labeled data while acquiring knowledge from voluminous unlabeled data to aid its training. Since the unlabeled data can be easily collected, the semi-supervised learning paradigm has the potential to compete with its supervised learning counterpart.

Many studies have been conducted to gradually improve the learning performance of semi-supervised paradigms [5, 6, 21, 22, 32, 39, 43, 46], resulting in promising results. But their collection and use of large amounts of unlabeled data will also raise certain security and privacy concerns [1, 2]. In [7], the authors have unleashed the security issue by proposing the first poisoning attack on the unlabeled data for disturbing the classification performance of semi-supervised learning. In contrast, this work focuses on the data privacy issue to answer the question: if a user can examine whether his/her private data has been unauthorizedly collected for use. This is referred to as membership inference in general.

Extensive membership inference techniques [10, 12, 16, 17, 23–25, 33–35, 40–42] have been proposed for targeting different learning paradigms (e.g., online learning, contrastive learning, etc.) or considering the special scenario (e.g., label-only predictions). One may wonder if those membership inference techniques proposed previously for inferring the training dataset can be adopted for use. Due to the special design of using both labeled and unlabeled data,

the semi-supervised learning paradigm in fact cannot have its membership patterns inferred properly by applying previous techniques. Investigations into membership inference attacks in the realm of semi-supervised learning are thus rather limited so far. In [15], a membership inference solution toward semi-supervised learning has been proposed, by evaluating the similarity between augmented query data instances. Although this solution focuses on data augmentation techniques typically employed in semi-supervised learning algorithms, the exploration of more apparent membership patterns stemming from such leading approaches as consistency regularization and low-entropy pseudo-labeling, remains open. To this end, enabling users to examine if their data have been collected illicitly for semi-supervised training use is still challenging and remains mostly unaddressed.

In this paper, we propose a novel method for membership inference against the semi-supervised learning models, for use in data privacy examination. We consider the inference scenario where a user, called an inferrer hereafter, aims to examine whether his/her data, referred to as the probing data, have been adopted illicitly to train a target semi-supervised model, no matter weather it is used as labeled or unlabeled data. For practicality, we consider the black-box scenario in that an inferrer has no background knowledge about the target model, including learning algorithms, model structure, parameters, data distribution, etc., but the inferrer can query the target model for obtaining prediction vectors of the probing data. By exploring prominent semi-supervised learning, we have discovered recently two improvements, which motivate our inference design strategies. In particular, to effectively learn correct knowledge from the unlabeled data, two special designs of consistency regularization and low-entropy pseudo-labeling, are considered as the leading semi-supervised learning algorithms. The former enforces algorithms to output similar prediction vectors for the perturbed versions, usually generated by the augmentation approach so as to have stable predictions for similar data samples. The latter requires algorithms to output high confident predictions under unlabeled data, measured by the low-entropy prediction vectors to avoid ambiguous predictions on unlabeled data. Such two designs imprint the apparent patterns of the training data on their prediction vectors, which can then be extracted for membership inference. Inspired by two special designs, we propose two metrics, called *inter-consistency* and *intra-entropy*, with the former to measure the similarity degrees among the prediction vectors of perturbed probing data and the latter to calculate the cross-entropy of a prediction vector and its one-hot vector. Such two new metrics can help to unveil membership patterns, facilitating our inference.

Overall, our inference involves three steps. First, the inferrer divides his/her local dataset into two disjoint parts, called the local member dataset and the local non-member dataset. The former is used to train a shadow model to fit the output pattern of the target model. Notably, the inferrer is unaware of the dataset distribution of the target model. Instead, the inferrer just applies his/her local datasets to train the shadow model. If the target model has collected his/her data for training use, the local dataset will naturally have the same distribution as the training set of the target model. Second, the inferrer generates multiple perturbed versions of each data sample in the local dataset and inputs them into the shadow model to obtain prediction vectors. The two metrics, i.e., inter-consistency and intra-entropy, are calculated based on these prediction vectors to create the membership metrics for each data in the local dataset. In the end, the inferrer assembles the membership patterns according to these membership metrics while labeling them with 1 or 0 based on whether they are derived from the training data of the shadow model or not, and uses the dataset of membership patterns to train a membership classifier. When conducting membership inference, the inferrer queries the target model with the perturbed versions of probing data, extracts its membership metrics, and uses the membership classifier to predict whether the probing data belongs to the training set of the target model or not.

Extensive experiments have been conducted on four datasets across six semi-supervised learning algorithms to evaluate the performance of our method. Experimental results exhibit that our method can achieve inference accuracy over 80% under each setting, demonstrating its superiority to be adopted for user data privacy examination against semi-supervised learning models. In addition, our results also reveal that 1) the leading learning algorithms, 2) the models with higher testing performance, and 3) the models trained with less labeled data, are more vulnerable to inference.

The remainder of this paper is organized as follows. In Section 2, we present the preliminary knowledge of the semi-supervised learning and discuss related work on membership inference attack. In Section 3, we illustrate our problem, state our threat model, and define our goal. Section 4 elaborates on the detailed solution for our membership inference against semi-supervised learning. In Section 5, we outline extensive experiments for evaluating the performance of our inference method and present experimental results. Section 6 lists our future work. Section 7 concludes this paper.

2 PRELIMINARIES & RELATED WORK

In this section, we first present the background knowledge of semi-supervised learning and then discuss prior studies on the membership inference toward machine learning models.

2.1 Semi-Supervised Learning

Considering that the data labeling task is expensive and labor-intensive while unlabeled data can be collected much easier, semi-supervised learning is proposed to lift learning performance by relying on unlabeled data. In general, it aims at improving the performance of a model trained from the labeled data and learns the knowledge acquired from the unlabeled data. Denote $\mathcal{X} = \{(x_1, l_1), (x_2, l_2), \cdots, (x_m, l_m)\}$ as the labeled dataset, where x_i and l_i indicate a data sample and its label, respectively. Denote $\mathcal{U} = \{u_1, u_2 \cdots, u_n\}$ as the unlabeled dataset, where u_i is an unlabeled data sample. Semi-supervised learning tries to train a model $f_{semi} \leftarrow \mathcal{A}(\mathcal{X}, \mathcal{U})$, where $\mathcal{A}$ is the semi-supervised learning algorithm. Although the idea of semi-supervised learning has been proposed for a long time [26, 36, 50], its learning performance is usually mediocre. A series of recent studies [5, 6, 8, 21, 22, 27, 32, 43, 45–48] gradually made improvements to the semi-supervised learning methods, promoting them to become a promising learning paradigm. Two lines of technologies contribute to their performance improvement, summarized as follows.

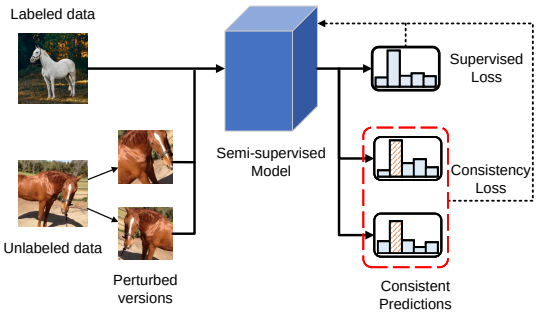


Figure 1: The semi-supervised learning model with the consistency regulation.

Consistency Regularization. Intuitively, a good learning model should be stable in producing the prediction results of a batch of similar data samples. To this end, consistency regularization is proposed to minimize the difference among predictions of the perturbed versions from each unlabeled data sample. As shown in Figure 1, the semi-supervised model is trained with both the standard supervised loss on the labeled data and the consistency regularization loss on the unlabeled data. The latter, denoted by $loss_{UL}(\mathcal{U})$, can be defined as:

$$loss_{UL}(\mathcal{U}) = \sum_{i=1}^{n} D\Big(f_{semi}(\alpha(u_i)) - f_{semi}(\alpha(u_i))\Big), \quad (1)$$

where $f_{semi}()$ is the semi-supervised model and $\alpha()$ represents a stochastic function to produce the perturbed version of unlabeled data u_i, so that the two terms in Eqn. (1) are different. D represents a function to calculate the distance between two prediction results and usually adopts the cross-entropy loss. Different solutions have been proposed to create the perturbed versions of a data sample, such as adopting adversarial transformation [27], leveraging data augmentations [5, 6, 39, 46], or using model predictions in the early training epochs [21, 43]. With this consistency regularization loss, the model trained on the labeled data will be improved with the additional knowledge acquired from the unlabeled data, making it achieve much better performance.

Low Entropy Pseudo-labeling. The second technology comes from the pseudo-labeling, where the model itself produces the artificial labels of the unlabeled data for improving the model. The pseudo-label is encouraged to have a low entropy, to be more reliable and more effective in helping semi-supervised model training. Similarly, the model is trained with the standard supervised loss on labeled data and the loss on unlabeled data as follows:

$$\sum_{i=1}^{n} \mathbb{I}(max(f_{semi}(u_i)) > \theta) D\Big(O(f_{semi}(u_i)), f_{semi}(u_i)\Big), \quad (2)$$

where θ is the threshold to ensure that the model only adopts a pseudo-label whose largest class probability (i.e., $max(f_{semi}(u_i))$ surpasses this threshold and $O()$ represents a function to generate the pseudo-label. For example, in [22], $O()$ is a function to transfer the prediction vector into the "one-hot" form. As shown in Fig. 2, two prediction vectors, which are $\{0, 0.5, 0.2, 0.1, 0.2\}$ and $\{0, 0.7, 0.1, 0.2, 0\}$, can have the same one-hot pseudo-label $\{0, 1, 0, 0, 0\}$; however, despite the same pseudo-label $\{0, 1, 0, 0, 0\}$, the latter prediction is more confident. By calculating the loss between the prediction vectors and one-hot vectors, models are regulated to produce lower entropy pseudo-labels. Besides, [5, 6] also adopt the sharpening function in $O()$. That is, given a prediction vector $\mathbf{q} = \{q_1, q_2, \cdots, q_n\}$, a sharpening calculation is applied as follows:

$$Sharpen(\mathbf{q}, T)_i := q_i^{\frac{1}{T}} \Big/ \sum_{j=1}^{n} q_j^{\frac{1}{T}}, \quad (3)$$

where T is a hyperparameter referred to as temperature. When $T \to 0$, the output vectors will approach a Dirac ("one-hot") distribution and be applied as the pseudo-label for calculating the loss with their original prediction vectors.

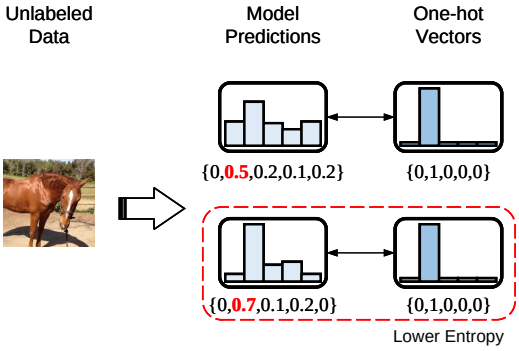


Figure 2: Low entropy pseudo-labeling for lower entropy model predictions of unlabeled data.

Holistic Approach. Recently, the leading semi-supervised learning methods, e.g., MixMatch [6], Re-MixMatch [5], and FixMatch [39], proposed the holistic strategy by combining the idea of both consistency regularization and pseudo-labeling on both labeled and unlabeled data. For example, in MixMatch, for each labeled data $\{x, l\}$, authors generate an augmented version of it, denoted as $\{\hat{x}, l\}$, and calculate the supervised loss as follows:

$$Loss_L(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x_i \in \mathcal{X}} D(l_i, f_{semi}(\hat{x}_i)). \quad (4)$$

This loss lets the model output the correct predictions on the augmented version of the labeled data, meeting the idea of consistency regulation. For each unlabeled data, say u_i, authors average the model's prediction results across its K augmentations, as follows:

$$\bar{q}_i = \frac{1}{K} \sum_{i=1}^{K} f_{semi}(\alpha(u_i)) \,. \quad (5)$$

This averaged prediction vector for an unlabeled data naturally becomes similar to the predictions of all its perturbed versions. Such a design also meets the idea of consistency regularization that makes similar data, i.e., unlabeled data and its augmented versions, have similar predictions. Besides, a sharpening function is adopted to the prediction vectors $\bar{q}_i$ for producing the pseudo-label q_i, encouraging the model to produce lower entropy predictions

on unlabeled data. At last, q_i is assigned as the pseudo-label of unlabeled data u_i for calculating the loss:

$$L_{\mathcal{U}} = \frac{1}{|\mathcal{U}|} \sum_{u_i \in \mathcal{U}} D||q_i - f_{semi}(u_i)||_2^2 \,. \tag{6}$$

MixMatch, Re-MixMatch, and FixMatch all adopt such a similar idea with different details. By combining the idea of consistency regularization and low-entropy pseudo-labeling, the holistic semi-supervised learning approach achieves significant performance improvement and has yielded leading semi-supervised learning algorithms.

2.2 Related Work

Many studies have been undertaken for membership inference, but they are mainly toward the supervised learning paradigms. The first membership inference attack toward the supervised machine learning model aimed to determine if some data samples belong to the training dataset or not [38]. It leverages the fact that the prediction vectors of a target model may preserve certain patterns, which can be learned by multiple shadow models to mimic such patterns. A membership classifier is trained to differentiate the prediction vectors from the member and from non-member data. Based on this idea, plentiful solutions [10, 12, 16, 17, 23, 24, 33–35, 40–42, 48] have been proposed to improve [38]'s work or extend it to target some specific supervised learning algorithms or scenarios. For example, [35] considered a solution that relied only on one shadow model to conduct the membership inference. An online learning algorithm was treated in [34], where the attack was conducted by measuring the derivation of output vectors when the model is updated with online data. In addition, [10, 16, 40] have proposed inference solutions toward the transfer learning, federated learning, and natural language domains under supervised learning algorithms. Besides, [33] has demonstrated that black-box attacks can achieve similar inference performance as that of white-box attacks, whereas [42] has exploited the adversarially robust models, for lifting the risk of membership inference attacks. [12, 24] have explored the membership inference attack to the label-only models, in which the target supervised models only output the classification results instead of prediction vectors. On the other hand, some countermeasure solutions have also been proposed for defending the inference attack on supervised models. They can be categorized into two directions: 1) preventing overfitting on the prediction results [19, 23, 28, 35, 41] and 2) adopting differential privacy [3, 4, 18, 37, 44, 49].

In the unsupervised scope, [25] has proposed the EncoderMI to exploit the membership inference of an image encoder towards its training data under the contrastive learning paradigm [11, 14, 31]. It leverages the characteristic of image augmentation, where an encoder is likely to output similar feature vectors corresponding to two augmented versions from the same input, for conducting inference attack. The membership inference attack on semi-supervised learning has rarely been explored yet. In [15], the authors proposed the first and the only method on the semi-supervised learning, where the inference is conducted by measuring the similarity between the augmented versions of the query data. While this approach effectively leverages data augmentation techniques commonly applied in semi-supervised learning algorithms, the more apparent membership patterns resulted from state-of-the-art strategies, such as consistency regularization and low-entropy pseudo-labeling, have yet to be investigated. Besides, existing membership inference solutions cannot perform well here, since the separate regulations on labeled and unlabeled data raise new challenges in conducting the membership inference attack. Our experiments in Section 5 will validate that the performance of existing membership inference solutions are mediocre when applying directly to the semi-supervised learning regime.

3 PROBLEM STATEMENT

This work undertakes the novel effort toward the membership inference of semi-supervised learning models, enabling users or companies to be capable of verifying if their data are collected illicitly for use by the semi-supervised models. This section first illustrates the threat model and then formally defines our membership inference scope, followed by the encountered challenges.

3.1 Threat Model

Target Model. We target classification models trained by semi-supervised learning algorithms with both labeled and unlabeled data. For each input data, such a model outputs a prediction vector, with each entry representing the probability that the input data belongs to each corresponding class.

Background Knowledge. We consider the black-box scenario where the attacker has no background knowledge of the target models and the datasets (both labeled and unlabeled). That is, the specific semi-supervised learning algorithms, model architectures, hyperparameters, both labeled and unlabeled data, and other information, are all unknown. The only knowledge that our inferrer knows is that the model of interest is trained by semi-supervised learning.

Inferrer Ability. The inferrer possesses some local datasets and can use them to query the target model to obtain the corresponding prediction vectors. He/She is also able to construct a shadow model and train a classifier for conducting the membership inference. Notably, we don't enforce the inferrer to know the dataset distribution of the target model. Instead, the inferrer just applied his local datasets to perform the inference. If his data has been used by the target model, this local dataset naturally has the same distribution as the dataset of target model, while the inference can be successful. In addition, the inferrer only focuses on if his/her data has been used, but never steals additional information from, or impose any negative effects on, the target model.

Our Membership Inference Goal. The inferrer aims at examining whether his/her data samples have been used by the suspect model, no matter used as the labeled or unlabeled data. Denote $f_{semi}()$ as the target model, which is trained with both the labeled dataset $\mathcal{X} = \{(x_1, l_1), (x_2, l_2), \cdots, (x_m, l_m)\}$ and the unlabeled dataset $\mathcal{U} = \{u_1, u_2 \cdots, u_n\}$. The inferrer will use his/her local dataset $\mathcal{D}^{local}$ to train a shadow model and query this shadow model for obtaining the prediction vectors of data in the local dataset. The inferrer builds a classifier $f_{mem}()$, with the input of the prediction vector of probing data queried from the target model, it will output "1", if the probing data is in the target model's training dataset; and "0",

otherwise. Formally, corresponding to each probing data d_p, our membership inference can be defined as follows,

$$f_{mem}\Big(\mathcal{G}\big(f_{semi}(d_p)\big)\Big) = \begin{cases} 1 & \text{if } d_p \in \mathcal{X} \cup \mathcal{U}\ ; \\ 0 & \text{otherwise}\ , \end{cases} \tag{7}$$

where $\mathcal{G}()$ represents the function for extracting patterns from the prediction vectors of the probing data and $d_p \in \mathcal{X} \cup \mathcal{U}$ represents that the probing data sample is in the training dataset, regardless of being labeled or unlabeled data.

3.2 Challenges

Two challenges are encountered in our design of effective membership inference solutions, elaborated as follows.

First, the membership inference requires an examiner to obtain the correlation between the training data and their corresponding prediction vectors for learning membership patterns for inference. However, due to black-box access to the target model, it is impractical for an examiner to get such correlations. Since [38], designing a shadow model that aims to imitate the target model has been a popular way to learn such correlations for mining membership patterns. But simply applying this method to the complex learning paradigms typically suffers an apparent performance degradation, as demonstrated in [25, 34], due to that the specific algorithm's characteristic or property is difficult to be completely captured by a shadow model. Hence, how to fully extract membership patterns in semi-supervised learning regime is yet explored, remaining a challenging problem.

Second, the semi-supervised learning involves both the labeled and unlabeled data in the training dataset, which are treated differently during the training process in the semi-supervised learning algorithms. This will naturally lead to the difference in the prediction vectors of labeled and unlabeled training data. As a result, the binary classification task of distinguishing training and non-member data will become a more complex classification task of identifying both labeled and unlabeled as the training data against the non-member data. It is required to extract the commonality of labeled and unlabeled data for differentiating them from the non-member data, lifting the membership inference difficulty compared to other learning regimes.

4 MEMBERSHIP INFERENCE TO SEMI-SUPERVISED LEARNING MODELS

In this section, we illustrate our membership inference approach targeting the semi-supervised learning models by overcoming the aforementioned challenges. Our motivation and overview are presented first, followed by detailed descriptions.

4.1 Motivation

The plausibility of membership inference comes from the fact that the prediction results behave differently on the training data and on the non-member data. While previous studies on supervised learning demand capturing the special membership patterns hidden in prediction vectors of the target model, we observe that the semi-supervised learning algorithms themselves impose apparent patterns on prediction vectors during the training phase. Recall that semi-supervised learning has to train a high-quality model with the limited labeled data and acquire additional knowledge from unlabeled data, so special regulations are required to prevent the model from amplifying its error knowledge learned from the unlabeled data. The consistency regularization and low-entropy pseudo-labeling are two most advanced solutions, respectively for minimizing the differences among prediction results of perturbed versions from the same data sample and for encouraging confident predictions by lowing the entropy of prediction vectors. When lifting model performance by such delicate solutions, semi-supervised learning inevitably imprints certain patterns on the prediction vectors of its training data, i.e., the high similarity of different perturbed versions from one data sample and low entropy. By taking into account such patterns, we are able to propose our powerful membership inference method.

4.2 Overview

The central theme of our method is to build a membership inference classifier, which is capable of distinguishing the member and non-member data. Our general idea can be summarized as follows. The inferrer first divides his/her local data into two disjointed datasets, with one used for training a shadow model, called the member data, and another one referred to as the non-member data. Then, the inferrer generates the perturbed versions of each data sample from two datasets, and queries the trained shadow model to get their prediction vectors. Two metrics are then designed to extract the membership patterns from the prediction vectors corresponding to each data sample, which are then used to train a membership inference classifier for distinguishing the member and non-member data. With this classifier, the inferrer can query the target model to get the prediction vector and then input it to the inference classifier to determine if this data belongs to the training set of target model or not.

4.3 Shadow Model Construction

Since the inferrer has no background knowledge about the target model and its training dataset, he/she will train a shadow model with his/her local dataset, enforcing it to be capable of distinguishing the membership and non-membership patterns. We divide the local dataset into two disjointed parts, denoted as the $\mathcal{D}_M^{local}$ and $\mathcal{D}_{NM}^{local}$, respectively representing the local member and non-member datasets. Formally, we have

$$\mathcal{D}_M^{local} \cup \mathcal{D}_{NM}^{local} = \mathcal{D}^{local},$$
$$\mathcal{D}_M^{local} \cap \mathcal{D}_{NM}^{local} = \emptyset.$$

The local member dataset $\mathcal{D}_M^{local}$ is used to train the shadow model. Notably, since the target semi-supervised algorithm is unknown, here we can arbitrarily select one semi-supervised model to serve as our shadow model. Our experiments in Section 5.3 confirm that the different choice of shadow model has insignificant impacts on our inference results. We denote the trained shadow model as $f_{shadow}()$ and use the data from $\mathcal{D}_M^{local}$ and $\mathcal{D}_{NM}^{local}$ to separately query this model to get the corresponding prediction vectors for further extracting the membership patterns and training our inference classifier.

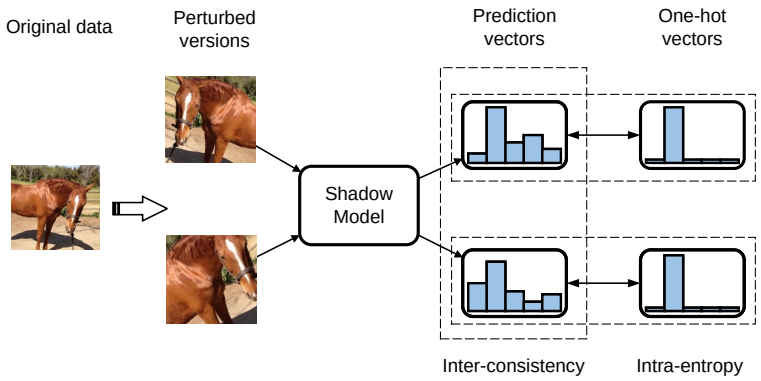


Figure 3: The overview of membership pattern extraction for Inter-consistency and Intra-entropy.

4.4 Membership Pattern Extraction

Instead of directly training the inference classifier with the prediction vectors from the shadow model, we devise two new metrics tailored to semi-supervised learning, able to facilitate membership inference. Inspired by the fact the advanced semi-supervised learning algorithms ensure the consistency regularization and low-entropy pseudo-labeling by minimizing the difference among prediction results of perturbed versions from one data and by lowering the entropy of each prediction vector, respectively, two metrics *Inter-consistency* and *Intra-entropy* are devised to capture the above two properties as shown in Fig. 3. For each data d from local member dataset $\mathcal{D}_M^{local}$ or non-member dataset $\mathcal{D}_{NM}^{local}$, we generate its K perturbed versions, denoted as $\{d_1, d_2, \cdots, d_K\}$. The data augmentation functions, which have been widely adopted in the leading semi-supervised learning algorithms and have demonstrated to have the best training performance [46], are applied to generate the perturbed versions. The K perturbed data versions $\{d_1, d_2, \cdots, d_K\}$ will query the shadow model to obtain their respective prediction vectors, denoted as $\{\mathbf{p_1}, \mathbf{p}_2, \cdots, \mathbf{p_K}\}$. Then, the two metrics are defined as follows.

Inter-consistency. The first is the Inter-consistency, denoted as M_d^{Inter}, which is to measure the similarity degree of those prediction vectors corresponding to each data d, yielding

$$M_d^{Inter} = \left\{S(\bar{\mathbf{p}}, \mathbf{p_i}) | i \in [1, K]\right\}, \text{where } \bar{\mathbf{p}} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{p_i}\,, \tag{8}$$

where $S()$ represents cross-entropy calculation to measure the similarity of two terms $\bar{\mathbf{p}}$ and $\mathbf{p_i}$. This formula captures the similarity between each prediction vector and the averaged prediction vector. Since consistency regularization enables the model to have a stable prediction on similar data samples, the inter-consistency, which measures the similarity degree among prediction vectors of perturbed versions, should have a high value.

Intra-entropy. We next calculate the intra-entropy of each input data to capture the low-entropy pseudo-labeling characteristics. The intra-entropy, denoted by M^{Intra}, is expressed as follows:

$$M_d^{Intra} = \left\{S(\hat{\mathbf{p}}_\mathbf{i}, \mathbf{p_i}) | i \in [1, K]\right\}, \tag{9}$$

where $\hat{\mathbf{p}}_\mathbf{i}$ is the "one-hot" vector of prediction result $\mathbf{p_i}$. For example, if $\mathbf{p_i} = \{0.1, 0.6, 0.3\}$, then $\hat{\mathbf{p}}_\mathbf{i} = \{0, 1, 0\}$. M^{Intra} measures the entropy of the prediction vectors, with the lower entropy signifying the higher prediction confidence. Since a semi-supervised learning algorithm encourages low entropy prediction, M^{Intra} can help describe the membership pattern from this perspective.

4.5 Membership Inference Classifier

With the local member dataset $\mathcal{D}_M^{local}$, the non-member dataset $\mathcal{D}_{NM}^{local}$, and their inter-consistency and intra-entropy metrics available, we next train the inference classifier. We first assemble the membership pattern dataset with two metrics, formally defined as follows:

$$\{(M_d^{Inter}, M_d^{Intra}), 1 | d \in \mathcal{D}_M^{local}\} \cup$$
$$\{(M_d^{Inter}, M_d^{Intra}), 0 | d \in \mathcal{D}_{NM}^{local}\}\,,$$

That is, if the input data d belongs to the local member dataset, we assign the label "1" to its metric vector (M^{Inter}, M^{Intra}); otherwise, we assign the label "0". This dataset includes both the two metrics representing the membership patterns and the label for indicating a member or non-member data. We use this membership pattern dataset and adopt a fully connected neural network by following the standard supervised learning procedure to train the inference classifier $f_{mem}()$, for classifying the member and non-member labels. Ideally, we expect the inference classifier $f_{mem}()$ to have the prediction results as follows

$$f_{mem}\left(M_d^{Inter}, M_d^{Intra}\right) = \begin{cases} 1 & \text{if } d \in \mathcal{D}_M^{local} \\ 0 & \text{if } d \in \mathcal{D}_{NM}^{local} \end{cases}. \tag{10}$$

Since a classifier $f_{mem}()$ typically outputs an inference score, which is in the range of $[0, 1]$, for the input data d. If the inference score is larger than a threshold (for example 0.5 in general), we consider the predicted label to be 1, i.e., the input data belongs to a member data.

4.6 Membership Inference Pipeline

With the trained inference classifier, we are ready to perform the membership inference to the target model. That is, we use a data sample d to generate its perturbed versions and query the target model to get the corresponding prediction vectors. We apply the

Table 1: Experiment settings on target models and shadow models across four datasets

	Target Model			Shadow Model		
	Labeled Data	Unlabeled Data	Structure	Labeled Data	Unlabeled Data	Structure
CIFAR-10	250/4000	30000	ResNet-16/WRN-16	2000	10000	ResNet-16
CIFAR-100	2500/10000	25000	ResNet-16/WRN-16	5000	10000	ResNet-16
SVHN	250/1000	25000	ResNet-16/WRN-16	1000	10000	ResNet-16
STL-10	1000	25000	ResNet-16/WRN-16	1000	10000	ResNet-16

inter-consistency and intra-entropy to calculate the corresponding $(M_d^{Inter}, M_d^{Intra})$ as defined in Section 4.4, and then input them to the inference classifier $f_{mem}()$. If the output label is "1" ("0"), this probing data belongs (does not belong) to the training set of the target model.

5 EVALUATION

In this section, we implement our proposed semi-supervised membership inference method and conduct extensive experiments for performance evaluation.

5.1 Experiment Setup

5.1.1 Dataset. Our experiments are conducted on four datasets, which are widely adopted in the leading semi-supervised learning studies, described in sequence next:

- **CIFAR-10** [20]. CIFAR-10 dataset contains 60,000 color images (i.e., 50,000 training images and 10,000 testing images) from 10 object categories. The size of each image is 32×32.
- **CIFAR-100** [20]. CIFAR-100 has the same format as CIFAR-10 and it contains 100 classes with 500 training images and 100 testing images for each class.
- **SVHN** [29]. SVHN is a real-world image dataset for developing machine learning and object recognition algorithms. It contains 73257 images for training and 26032 images for testing over 10 classes. The size of each image is 32×32.
- **STL-10** [13]. The STL-10 dataset is an image recognition dataset for developing the unsupervised feature learning. There are 500 training images and 800 test images per class, having the total of 10 classes. Besides, it includes 100,000 unlabeled images for unsupervised learning. The size of each image is 96×96.

For CIFAR-10, CIFAR-100, and SVHN, we remove some images' labels to consider them as the unlabeled data. We conduct our evaluation on image classification, which is the primary domain of semi-supervised learning algorithms [30].

5.1.2 Target Model. We adopt the ResNet-16 and WRN-16 as the model structure to train the target model, by following the settings as in [39]. Six semi-supervised learning algorithms are employed to help the target model's training, i.e., Pseudo-Labeling [22], Mean Teacher [43], UDA [46], MixMatch [6], Re-MixMatch [5], and FixMatch [39]. Among them, Mean Teacher and UDA apply the consistency regularization, Pseudo-Labeling mainly adopts the low entropy Pseudo-Labeling, while MixMatch, Re-MixMatch, and FixMatch are holistic algorithms that combine both regulation techniques together. Notably, UDA, MixMatch, Re-MixMatch, and FixMatch are new leading semi-supervised methods which can achieve promising classification performance, while the Pseudo-labeling and Mean Teacher are the traditional semi-supervised algorithms.

We follow the dataset settings given in [39] to train each target model, with details shown in Table 1. Specifically, for CIFAR-10, we adopt 250 or 4000 labeled data and 30000 unlabeled data. For CIFAR-100, we adopt 2500 or 10000 labeled data plus 25000 unlabeled data. For SVHN, we use 250 or 1000 labeled data plus 25000 unlabeled data. For STL-10, we use 1000 labeled data and 25000 unlabeled data.

Due to the computation resource limitations, we reduce the unlabeled data amounts and the number of layers of the model structure compared to the original setting in [39]. So, the testing performance of these target model is lower around 5% compared to the performance reported in the original paper. By employing the dropout operation and early stop, the overfitting degrees of these target models, measured as the training-testing accuracy gap, are lower than 8%.

5.1.3 Shadow Model. We take the ResNet-16 as our shadow model structure, and train it with the inferrer's local dataset under the semi-supervised paradigm. For CIFAR-10, we randomly sample 2000 labeled data and 10000 unlabeled data as the local member dataset for training the shadow model. Meanwhile, another 5000 data are sampled as the local non-member dataset. Similarly, for CIFAR-100, 5000 labeled data and 10000 unlabeled data are sampled as the local member dataset, and another 5000 data are taken as the local non-member dataset. For both SVHN and STL-10, we choose 1000 labeled data and 10000 unlabeled data as the local member dataset, and take another 5000 images as the local non-member dataset. Table 1 lists the detailed settings.

5.1.4 Membership Inference Classifier. Our membership inference classifier is trained with a fully connected neural network having five hidden layers, with the number of neurons in each layer being 128. Adma is used as the optimizer and the classifier is trained for 200 epochs. Six perturbed versions are generated for each data sample. The choices of perturbed versions amounts will be discussed in Section 5.5.

5.1.5 Performance Metrics. To evaluate the performance of our proposed membership inference method, we randomly sample data, including labeled data and unlabeled data from the both training dataset , and non-member dataset of the target model, as the probing dataset. Here, the non-member dataset is also disjointed from the training dataset of the shadow model. In particular, we sample 25% data from the training dataset of the target model (keeping the ratio between label and unlabeled data unchanged) and sample the same amount of non-member data, mixing them up as the probing dataset. For example, for the target model trained on CIFAR-10

with 34000 data samples (4000 labeled and 30000 unlabeled) in Table 1, we sample 8500 member data samples (1000 labeled and 7500 unlabeled) and 8500 non-member data samples as the probing set. Then, we query our membership classifier for obtaining the inference results.

We first calculate the precision, recall, and accuracy as the evaluation metrics, which follow previous work [25, 38], defined as:

$$Precision = \frac{TP}{TP + FP}, \; Recall = \frac{TP}{TP + FN},$$
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP (or TN) represents the number of member data (or non-member data) that are correctly predicted to be the member (or non-member), and FN (or FP) indicates the number of member data (or non-member) that are wrongly predicted to be the non-member (or member). Second, we take AUC (i.e., area under the ROC curve) as the evaluation metric. For a randomly selected member data sample and a random non-member data sample, the AUC value reflects the probability of how well our inference classifier correctly judges the membership likelihoods of those two samples, with the higher AUC value for the member data sample, the better. Hence, a classifier with a larger AUC value signifies better inference performance, with an AUC value of 0.5 equating to a random guess.

5.2 Comparison to Existing Methods

We rectify some existing membership inference methods to fit into the semi-supervised learning paradigm, as our counterparts for comparison. The goal is to exhibit that the existing inference solutions cannot perform well enough for semi-supervised learning, thus calling for our new solution.

5.2.1 Counterpart Methods. Five counterparts originally targeting supervised and unsupervised learning are taken into account, outlined as below.

Baseline-1. For membership inference on the supervised learning paradigm, most existing methods can be considered as the extension of the first membership inference work [38], where the authors provided the essential idea of identifying the difference between the target model's prediction results of training data and unseen data (i.e., non-member data). To rectify it for comparison, we first query the semi-supervised model with the local dataset to get the prediction results and then follow the method proposed in [38] to train multiple shadow models. After that, we input the local training dataset and non-member dataset to the shadow models for getting the prediction vectors, which are then applied to build the membership classifier. For the probing data, we first query the target model to obtain its prediction vectors and then input them to the classifier for determining the membership.

Baseline-2. Similar to Baseline-1, in [41], the authors proposed a shadow model-based membership inference method based on measuring the entropy of prediction vectors. Considering that the model is trained to minimize the loss of the training data by fitting the ground truth label of the training data sample, it is prone to produce the prediction results with low entropy in the training data. Thus, by training a membership classifier to learn the difference in the entropy of prediction vectors of training and non-member data, one can conduct the membership inference attack. To fit it into the semi-supervised learning paradigm, we train the shadow model and the classifier to learn such entropy on prediction results of the training and the non-member data samples.

Baseline-3. In the unsupervised learning regime, EncoderMI proposed in [25] infers the training dataset of an encoder which is trained with the contrastive learning technique. That work is motivated by the observation that when a single data sample and its multiple augmented versions are inputted into the model for the training task, their encoding vectors will be similar (dissimilar) if they are positive (negative) samples. Then the membership inference is conducted based on identifying the difference of encoding vectors for augmented versions between the training and the non-member data samples. To fit it to the semi-supervised learning paradigm, we train the classifier based on prediction results of augmented versions from the training and the non-member data samples.

Baseline-4. In [17], authors proposed a membership inference solution without the need of the shadow model. Given two probing datasets, one with more training data of the target model, and the other with more non-member data, their data distributions become more similar by exchanging the data samples from the two datasets, leading to similar prediction results corresponding to the two datasets. Hence, by calculating the similarity variations among prediction results of two probing datasets after exchanging their data samples, the authors can conduct the membership inference. Since that solution is not limited to a special learning paradigm, we can directly apply it to semi-supervised learning paradigms.

Baseline-5. The label-only membership inference proposed in [24] tackles the scenario that a target model only outputs the classification label rather than the prediction vectors. It constructs a membership feature vector for data based on the classification label of augmented feature vector versions, in which each entry in a vector represents whether the corresponding augmented version is predicted correctly by the target classifier. A membership inference classifier is then trained to learn the difference between feature vectors derived from training data and from non-member data. This inference technique is not limited to any learning paradigm, so we directly apply it to semi-supervised learning. Since it requires the ground-truth label of probing data while STL-10 only contains very few labels, its evaluation on STL-10 is excluded in our experiments.

5.2.2 Comparison to Baseline Methods. We assume the target model employs the ResNet-16 as its structure and FixMatch as its semi-supervised learning algorithm. For baselines 1, 2, 4, and our methods, we train the shadow models with the same structure, i.e., ResNet-16, and the same semi-supervised learning algorithm, i.e., FixMatch. Table 2 lists the accuracy, precision, recall and AUC value of the five baselines and our method on four datasets.

From this table, we observe that our method solidly outperforms all five counterparts on four datasets in terms of three performance metrics. In particular, the AUC performance results of our method are in the range of 86.3% to 90.1%, demonstrating that our inference method based on the inter-consistency and intra-entropy can well capture the membership patterns for correct prediction. For the five counterparts, Baseline-3 achieves the best performance, but

Table 2: Accuracy (%), precision (%) , recall (%), and AUC (%) of the five baseline counterparts and our method in four datasets with FixMatch learning algorithm

		CIFAR-10	CIFAR-100	SVHN	STL-10
Baseline-1	Accuracy	61.4	60.4	58.7	62.3
	Precision	57.8	62.5	61.6	59.7
	Recall	60.0	60.5	60.6	58.7
	AUC	61.2	61.7	60.9	59.8
Baseline-2	Accuracy	68.7	67.5	68.7	65.2
	Precision	68.5	65.9	67.8	68.7
	Recall	67.2	69.0	64.3	66.7
	AUC	67.9	68.7	65.7	70.0
Baseline-3	Accuracy	71.9	70.1	71.4	67.6
	Precision	70.0	68.7	70.3	71.0
	Recall	70.7	71.0	67.1	68.6
	AUC	70.5	69.8	69.5	72.0
Baseline-4	Accuracy	62.4	62.1	64.2	66.9
	Precision	61.3	67.2	67.0	66.5
	Recall	66.3	61.7	63.5	61.2
	AUC	65.3	69.0	66.8	64.4
Baseline-5	Accuracy	62.2	62.3	62.6	––
	Precision	61.4	67.2	60.4	––
	Recall	65.0	66.3	66.2	––
	AUC	64.7	67.1	66.0	––
Our Method	Accuracy	87.9	89.4	88.3	89.5
	Precision	87.4	89.3	88.9	89.2
	Recall	89.0	89.1	89.4	90.3
	AUC	86.3	90.1	88.2	89.5

at only around 70% in AUC, while Baseline-1 performs the worst, typically with the performance results of just around 60% in AUC. These performance results are all starkly lower than what were reported in their original articles (near 90%), indicating that those existing counterparts cannot be directly transferred to the semi-supervised learning paradigm. The reason can be summarized as follows. Regarding Baseline-1, directly inferring the membership pattern on prediction results of the probing data is ineffective, validated in many prior studies. Regarding Baseline-2, it is an extension of baseline-1 while considering the entropy metrics. However, such entropy will behave differently on the labeled and unlabeled training data, thereby lowering its performance. Regarding Baseline-3, although the augmented versions of the probing data can enrich membership knowledge, they are still insufficient to work in the semi-supervised learning paradigm without the aid of two new metrics we devise to extract more membership patterns. Baseline-4 leverages the similarity variations between prediction vectors of two probing datasets; however, exchanging labeled data and unlabeled data rather than training non-member data also causes the similarity variations, hurting inference accuracy. Regarding Baseline-5, the absence of the prediction vector cannot capture the sufficient membership patterns, hindering the inference capability. In contrast, our method, which measures the two metrics based on prediction vectors, can better extract membership patterns, excelling in membership inference.

5.3 Evaluation toward Various Semi-supervised Learning Algorithms

This section will first evaluate the performance of our method on models trained under different semi-supervised learning algorithms and then explore better methods for training the shadow model.

We conduct the membership inference on target models trained by WRN-16 structure under six different semi-supervised learning algorithms. Since we consider the black-box settings where the inferrer has no knowledge about the target model, six semi-supervised learning algorithms trained with RestNet-16 are chosen to serve as our shadow model, for evaluation. Due to the page limits, we only present the results of the STL-10 dataset, as listed in Table 3. Obviously, when both target and shadow models are trained with the same semi-supervised algorithms, our inference method achieves the best performance. When they employ different learning algorithms, performance degrades but is still more than 80% always. For example, if we take UDA as the target learning algorithm, when shadow models are trained with other five algorithms, the inference accuracy values equal 81.1%, 83.2%, 86.4%, 87.8.8%, and 87.1%, respectively. Such results demonstrate that our method can always achieve impressive performance under the black-box setting.

Leading learning algorithms are vulnerable to our inference method. When comparing the last four rows (i.e., with target models being UDA, MixMatch, Re-MixMatch, and FixMatch, respectively) to the second and the third rows (i.e., with target models being Pseudo-Labeling and Mean Teacher, respectively) in Table 3, we observe that the leading learning methods (i.e., UDA, MixMatch, Re-MixMatch, and FixMatch) are more vulnerable to our membership inference method. The reason is that, while these leading learning algorithms take advantage of both consistency regularization and low-entropy pseudo-labeling to design their loss function for better training performance, more apparent membership patterns are imprinted on the prediction vectors of training data. Hence, they are more vulnerable to our inference method. Such an observation indicates that delicate learning algorithms for improving training performance yield loss functions that leave more membership patterns on the prediction vectors, better facilitating an inferrer to examine whether their private data are used by the suspect model.

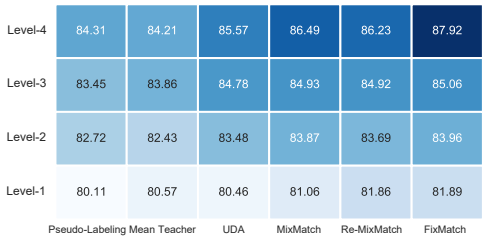
Holistic algorithms are better choice for shadow model. When comparing the last three columns (i.e., with shadow models being MixMatch, Re-MixMatch, and FixMatch, respectively) to the second, third, and fourth columns (i.e., with shadow models being Pseudo-Labeling, Mean Teacher, and UDA, respectively) in Table 3, we observe that employing holistic learning algorithms (i.e., MixMatch, Re-MixMatch, and FixMatch) to train the shadow model can achieve better inference performance. Such an observation stems from the fact that a holistic algorithm implements both consistency regulation and low entropy pseudo labeling together during the training process, so those prediction vectors produced by its resulting shadow models contain more apparent characteristics for the membership patterns. Hence, the classifier trained from metrics extracted from these shadow models can better infer the training dataset of the target model.

5.4 Evaluation across Target Models

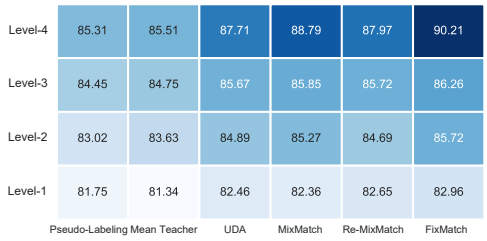
We next evaluate the performance of our membership inference toward different testing performance levels, the amount of labeled data, and the structures of target models. **Better models are more vulnerable to our inference method.** We evaluate our membership inference on target models with different testing performance

Table 3: Accuracy (%), precision (%), recall (%), and AUC (%) of our membership inference methods on target model with Psuedo-Labeling, Mean Teacher, UDA, MixMatch, Re-MixMatch, and FixMatch algorithms on STL-10

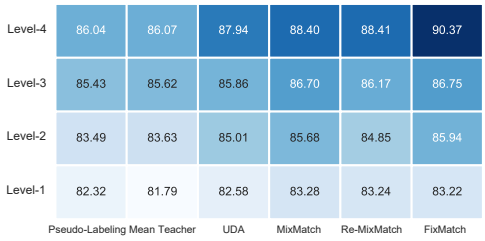
Target Model		Psuedo-Labeling	Mean Teacher	UDA	MixMatch	Re-MixMatch	Fixmatch
Psuedo-Labeling	Precision	85.1	80.7	83.6	84.3	84.6	84.3
	Recall	84.7	80.5	83.5	84.6	84.1	86.8
	Accuracy	84.9	80.5	82.7	84.8	84.3	84.6
	AUC	84.2	81.3	82.6	85.0	84.7	85.0
Mean Teacher	Precision	81.6	85.5	82.5	84.5	84.6	84.9
	Recall	80.9	84.4	83.8	84.3	84.5	84.8
	Accuracy	80.3	84.9	82.9	84.6	84.6	84.5
	AUC	80.9	84.7	82.7	84.0	85.2	84.7
UDA	Precision	81.2	82.8	89.7	86.0	86.3	88.2
	Recall	84.7	82.9	88.7	86.9	86.9	86.6
	Accuracy	81.1	83.2	89.5	86.4	87.8	87.1
	AUC	83.5	83.8	89.0	85.1	82.3	83.7
MixMatch	Precision	84.5	84.5	84.6	91.2	88.3	87.8
	Recall	85.1	83.9	84.1	90.3	87.4	87.6
	Accuracy	84.0	84.3	84.6	91.8	88.3	90.1
	AUC	84.9	84.7	84.3	92.0	88.2	88.7
Re-MixMatch	Precision	84.7	84.8	84.5	87.5	91.2	88.7
	Recall	84.2	84.6	84.6	88.2	90.9	90.0
	Accuracy	84.3	84.1	84.3	87.6	92.1	89.0
	AUC	84.9	83.8	84.7	88.0	91.7	89.7
Fixmatch	Precision	84.4	82.9	84.4	88.9	88.1	91.7
	Recall	82.3	84.8	84.6	87.7	88.8	92.1
	Accuracy	84.5	84.6	84.5	87.9	87.1	91.9
	AUC	83.9	84.9	85.7	88.1	88.7	92.7



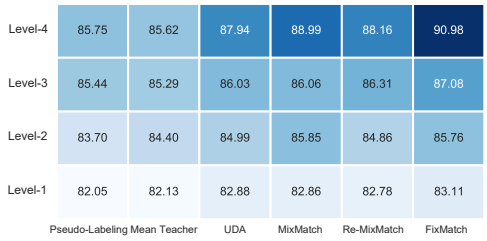
(a) Performance on CIFAR-10 dataset



(b) Performance on CIFAR-100 dataset



(c) Performance on SVHN dataset



(d) Performance on STL-10 dataset

Figure 4: Accuracy (%) of our membership inference method on the target models across different learning algorithms with four testing accuracy levels.

levels. Notably, the testing performance of a target model differs from the learning ability of the semi-supervised learning algorithms. The target models are trained with WRN-16 structure under six different semi-supervised learning algorithms. Notably, different learning algorithms cannot achieve the same testing accuracy; for example, FixMatch achieves around 90% while Pseudo-Labeling

Table 4: Accuracy (%) of our membership inference method on the target models across various learning algorithms with different amounts of labeled data

	CIFAR-10		CIFAR-100		SVHN	
	250 labeled data	4000 labeled data	2500 labeled data	10000 labeled data	250 labeled data	1000 labeled data
Psuedo-Labeling	82.13	80.14	82.99	81.49	82.97	80.20
Mean Teacher	82.12	80.52	82.01	81.67	82.29	82.06
UDA	83.41	80.74	84.13	82.09	84.60	81.40
MixMatch	87.25	83.24	87.29	83.41	86.24	83.62
Re-MixMatch	87.09	83.54	86.12	83.85	86.92	83.81
Fixmatch	89.34	86.11	89.78	86.18	90.25	86.85

Table 5: Accuracy (%) of our membership inference method on the target models across different model structures

	CIFAR-10		CIFAR-100		SVHN		STL-10	
	ResNet-16	WRN-16	ResNet-16	WRN-16	ResNet-16	WRN-16	ResNet-16	WRN-16
Psuedo-Labeling	87.96	82.36	85.09	82.72	83.99	82.99	82.54	81.63
Mean Teacher	87.27	83.04	85.18	81.66	82.73	81.83	83.69	81.32
UDA	88.52	82.81	87.65	83.46	84.41	84.66	86.86	83.31
MixMatch	87.72	84.93	87.24	85.06	85.48	84.55	85.53	82.27
Re-MixMatch	88.92	85.54	87.12	86.04	87.97	85.82	87.24	83.20
Fixmatch	89.49	86.10	89.96	86.19	89.65	86.50	89.23	87.94

reaches only 70%. Hence, for each semi-supervised learning algorithm, we consider four levels (i.e., Levels 1 to 4) of testing accuracy, where the Level-$(i+1)$ can achieve around 4% better performance than Level-i. We train the target model on each dataset to achieve different testing accuracy levels by manually controlling the training epochs. Our shadow model is trained with ResNet-16 plus FixMatch. Our membership inference results under six semi-supervised learning algorithms with different testing performance levels are shown in Figures 4(a), (b), (c), and (d), corresponding to the datasets of CIFAR-10, CIFAR-100, SVHN, and STL-10, respectively. In each figure, the x-axis lists six semi-supervised learning algorithms used by the target model, and the y-axis denotes the four testing performance levels. The value in each box indicates our inference accuracy result.

When comparing the results among four levels, we observe our membership inference achieves higher accuracy for better testing performance. For example, in Figure 4(d), when the target model is trained by the MixMatch, the values of inference accuracy are 88.99%, 86.06%, 85.85%, and 82.86%, respectively, on Level-4, Level-3, Level-2, and Level-1. The reason is that a model trained with better performance imprints more apparent patterns on the prediction vectors, better facilitating our membership inference. This observation also raises the challenge that a more accurate model, which is typically what we expect for conducting the classification task, will suffer higher risks for inference attacks. But from a user's perspective, more evidence is left on the well-trained learning model that can help a user verify if his/her data has been illicitly employed for model training use.

Model trained with less labeled data are more vulnerable to our inference method. We now train the target model with different amounts of labeled data while fixing the number of unlabeled data samples to show our inference performance. Notably, more labeled data help to hike model performance, but we can manually control the training epochs for target models to have the same testing performance under different amounts of labeled data. We train the shadow model with ResNet-16 and FixMatch algorithm. Table 4 shows the accuracy of our membership inference to the target models trained under WRN-16 with six different semi-supervised learning algorithms and different amounts of labeled data.

Table 6: Different treatments on the labeled and unlabeled data across different semi-supervised learning algorithms

	Labeled data	Unlabeled data
MixMatch	1 augmented version	K augmented version + sharpening
Re-Mixmatch	1 augmented version	K augmented version + sharpening
Fixmatch	1 augmented version	2 augmented versions + sharpening
UDA	No augmented version	K augmented version + sharpening
Mean Teacher	No augmented version	1 augmented version
Pseudo-Labeling	No augmented version	No augmented version+one-hot vector

We observe that our inference performs better when the target model uses less labeled data. Such a performance improvement is more apparent for target models with UDA, MixMatch, Re-MixMatch, and FixMatch. For example, on CIFAR-10, our inference performance increases by 1.99%, 1.62%, 2.67%, 4.01%, 3.55%, and 3.23%, respectively, when reducing the labeled data from 4000 to 250, corresponding to target models with Pseudo-Labeling, Mean Teacher, UDA, MixMatch, Re-MixMatch, and FixMatch. The same trend is also observed across the datasets of CIFAR-100 and SVHN. The reason can resort to the inherent design of semi-supervised learning algorithms, which treat the labeled and unlabeled data differently, as shown in Table 6. The employment of augmentation on the data related to the inter-consistency while sharpening operations as calculated in Eqn. (3) can result in the apparent intra-entropy of the prediction vectors. From Table 6, we find that the sharpening operations only operate on the unlabeled data, so that their prediction vectors possess more apparent intra-entropy patterns. Therefore, the inference accuracy of unlabeled data is higher than that of labeled data, resulting in the phenomenon that less labeled data will lift the portion of unlabeled data, thus helping to boost membership inference accuracy.

Model structures do not hinder the inference. Since we consider the black-box scenario where an inferrer has no background knowledge of the target model structure, we next evaluate our inference performance across different structures. Here the shadow mode employs the ResNet-16 with FixMatch, while the target models are trained respectively with six semi-supervised learning algorithms under the ResNet-16 and WRN-16. Table 5 shows the accuracy of our membership inference on different target models. We observe that the inference accuracy of all target models with ResNet-16 outperforms that with WRN-16. It is natural that our inference performs the best when the target and shadow models share the same structure. When the target models adopt a different structure, i.e., WRN-16, the inference performance decreases, but the accuracy values are all still more than 81%. They are considered to be good enough from this perspective.

5.5 Impact of Augmentation Operations

Our membership inference method requires creating multiple augmented versions of the probing data. Here, we explore the influence of the number of augmented versions of probing data on the membership inference performance.

We construct the shadow model with the ResNet-16 and FixMatch while training the target model with six different semi-supervised learning algorithms under the ResNet-16. We take the STL-10 dataset for experiments while varying the amount of augmented versions from 2 to 10. Fig. 5 shows the inference performance variations across six learning algorithms. From this figure, we observe that corresponding to each semi-supervised learning algorithm, our inference performance increases with more augmented versions of probing data. Considering that some strong augmented versions can have a big semantic deviation compared to the original probing data, therefore, inter-consistency and intra-entropy calculated on more perturbed versions can alleviate the influence of such deviation, providing more typical values to better depict the membership pattern.

On the other hand, when the number of perturbed versions reaches a certain value, for example, 6 for FixMatch and MixMatch, 7 for Mean Teacher and Pseudo-Labeling, inference performance becomes almost stable with only a slight increase. Since more perturbed versions of probing data incur a more calculation time in extracting the two membership patterns and training the membership classifier, it is recommended to choose around 6 perturbed versions when implementing our inference method to achieve effective and time-efficient inference performance.

6 FUTURE WORK

In this paper, we provide the novel exploration of membership inference toward the semi-supervised learning regime. However, there are some problems that need to be further explored in our future work, as stated in sequence below.

First, plenty of new semi-supervised learning algorithms are expected to be proposed every year while they may have different designs or hyperparameter settings. While these settings can improve the performance of the semi-supervised models, they definitely impose certain impacts on the prediction vectors. As a result, our membership inference attack performance based on the membership patterns on the prediction vectors may also be influenced. Hence, more research toward an in-depth exploration of how these new settings in emerging semi-supervised learning algorithms impact the membership inference performance are expected and exhibited as open problems.

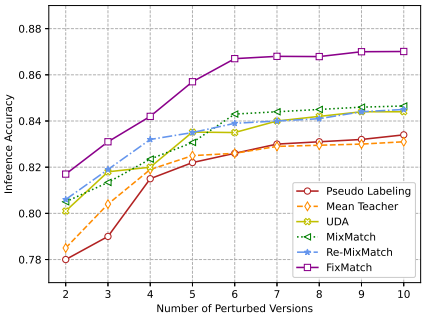


Figure 5: The inference accuracy variation with different amounts of perturbed versions of a probing data.

Second, an emerging series of membership inference solutions, i.e., per-example membership inference attacks [9], raised new concerns in designing inference methods around low false-positive rates. Our future work will make efforts on refining our membership inference attack targeting these new concerns toward the semi-supervised learning regime to further improve the practicality of our attack.

7 CONCLUSION

This paper has presented a novel membership inference study on semi-supervised learning for users' data privacy protection. We have devised two metrics, called inter-consistency and intra-entropy, for fully extracting membership patterns in the prediction vectors of training data, targeting two key inference designs, i.e., the consistency regularization and low-entropy pseudo-labeling of semi-supervised learning algorithms. We have built the shadow model, extracted the two membership metrics, and trained the membership classifier to determine whether the probing data is employed illicitly to train a target model. Extensive experiments demonstrate that our membership inference method can achieve the accuracy of over 80% under different settings, solidly outperforming all its counterparts. In addition, we also discover that the semi-supervised models are more vulnerable to our inference if they are trained by more effective learning algorithms, with fewer labeled data, or to have better testing performance.

ACKNOWLEDGMENTS

This work was supported in part by NSF under Grants 1763620, 1948374, 2019511, 2146447, and 2153502. Any opinion and findings expressed in the paper are those of the authors and do not necessarily reflect the view of funding agency.

REFERENCES

[1] 2020. Twitter demands AI company stops 'collecting faces'. https://www.bbc.com/news/technology-51220654.

[2] 2021. FTC settlement with Ever orders data and AIs deleted after facial recognition pivot. https://techcrunch.com/2021/01/12/ftc-settlement-with-ever-orders-data-and-ais-deleted-after-facial-recognition-pivot/.

[3] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC conference on computer and communications security (CCS)*. 308–318.

[4] Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE, 464–473.

[5] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. 2019. ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring. In *International Conference on Learning Representations (ICLR)*.

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems (NIPS)* 32 (2019).

[7] Nicholas Carlini. 2021. Poisoning the Unlabeled Dataset of {Semi-Supervised} Learning. In *Proceedings of USENIX Security Symposium*. 1577–1592.

[8] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2021. Membership Inference Attacks From First Principles. *arXiv preprint arXiv:2112.03570* (2021).

[9] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*. 1897–1914.

[10] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *Proceedings of USENIX Security Symposium*. 2633–2650.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[12] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *Proceedings of International Conference on Machine Learning (ICML)*. 1964–1974.

[13] Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 215–223.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[15] Xinlei He, Hongbin Liu, Neil Zhenqiang Gong, and Yang Zhang. 2022. Semi-Leak: Membership Inference Attacks Against Semi-supervised Learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*. Springer, 365–381.

[16] Seira Hidano, Takao Murakami, and Yusuke Kawamoto. 2021. TransMIA: membership inference attacks using transfer shadow training. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*. 1–10.

[17] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. 2021. Practical Blind Membership Inference Attack via Differential Comparisons. In *Proceedings of Network and Distributed Systems Security Symposium (NDSS)*.

[18] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. 2019. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 299–316.

[19] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the ACM SIGSAC conference on computer and communications security (CCS)*. 259–274.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[21] Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242* (2016).

[22] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proceedings of Workshop on challenges in representation learning, (ICML)*, Vol. 3. 896.

[23] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2021. Membership Inference Attacks and Defenses in Classification Models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy (CODASPY)*. 5–16.

[24] Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 880–895.

[25] Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. 2021. EncoderMI: Membership Inference against Pre-trained Encoders in Contrastive Learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2081–2095.

[26] Geoffrey J McLachlan. 1975. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *J. Amer. Statist. Assoc.* 70, 350 (1975), 365–369.

[27] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1979–1993.

[28] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 634–646.

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).

[30] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems* 31 (2018).

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[32] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. *Advances in neural information processing systems (NIPS)* 28 (2015).

[33] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of International Conference on Machine Learning (ICML)*. 5558–5567.

[34] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. 2020. Updates-leak: Data set inference and reconstruction attacks in online learning. In *Proceedings of 29th USENIX Security Symposium*. 1291–1308.

[35] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Proceedings of Network and Distributed Systems Security Symposium (NDSS)*.

[36] Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory* 11, 3 (1965), 363–371.

[37] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the ACM SIGSAC conference on computer and communications security (CCS)*. 1310–1321.

[38] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of IEEE Symposium on Security and Privacy (S$P)*. 3–18.

[39] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems (NIPS* 33 (2020), 596–608.

[40] Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 377–390.

[41] Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *Proceedings of USENIX Security Symposium*.

[42] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 241–257.

[43] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems (NIPS)* 30 (2017).

[44] Di Wang, Minwei Ye, and Jinhui Xu. 2017. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems* 30 (2017).

[45] Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. 2021. On the Importance of Difficulty Calibration in Membership Inference Attacks. *arXiv preprint arXiv:2111.08440* (2021).

[46] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems (NIPS)* 33 (2020), 6256–6268.

[47] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, and Reza Shokri. 2021. Enhanced Membership Inference Attacks against Machine Learning Models. *arXiv preprint arXiv:2111.09679* (2021).

[48] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *computer security foundations symposium (CSF)*. IEEE, 268–282.

[49] Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. 2019. Differentially private model publishing for deep learning. In *2019 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 332–349.

[50] Xiaojin Jerry Zhu. 2005. Semi-supervised learning literature survey. (2005).