A Pragmatic Approach to Membership Inferences on Machine Learning Models

Yunhui Long[1], Lei Wang[2], Diyue Bu[2], Vincent Bindschaedler[3],
Xiaofeng Wang[2], Haixu Tang[2], Carl A. Gunter[1], and Kai Chen[4,5]

[1]University of Illinois at Urbana-Champaign
[2]Indiana University Bloomington
[3]University of Florida
[4]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences
[5]School of Cyber Security, University of Chinese Academy of Sciences

Outline

- Pragmatic MIA
  - Step1: select vulnerable target records
  - Step2: identify vulnerable models
  - Step3: infer positive memberships
- Experiments

Pragmatic MIA: intro

- Motivation:

  Prior work focused on an adversary that indiscriminately attacks all the records without regards to the cost of false positive or negatives. But false positive membership can be costly. The author assumes that some records are more vulnerable to MIA than others. Privacy is violated if the adversary is confident about the membership of even a small amount of samples.

- e.g. An adversary has an attack precision of 51.7%. Two scenarios:
  - 1.7% of individuals have their membership status permanently and unequivocally at risk.
  - all individuals have a probability of 51.7% of having their membership correctly guessed.

  The first scenario is much more serious.
- Black-box setting:
  - Architecture of the target model is known, but the weights is unknown.
  - the adversary can access a set of records that are drawn independently from the data distribution, which may or may not overlap with the actual training data for the target models.

Pragmatic Attack vs Indiscriminate Attack

- Indiscriminate attack:

  let D be a set of records, A be the target model training algorithm.

  1) The user randomly splits D into a training set S_{train} and a testing set S_{test} of the same size.
  2) The user trains a model $M = \mathcal{A}(S_{\text{train}})$. The adversary has black-box access to M.
  3) $\forall r \in D$, $x_r = 1$ if $r \in S_{\text{train}}$, otherwise $x_r = 0$.
  4) $\forall r \in D$, the adversary obtains a guess $x'_r \in \{0, 1\}$.
  5) $\forall r \in D$, the adversary succeeds if $x'_r = x_r$, otherwise the adversary fails.

- Pragmatic Attack:

  1) The adversary chooses a target $r \in D$.
  2) The user randomly splits D into a training set S_{train} and a testing set S_{test} of the same size.
  3) The user trains a model $M = \mathcal{A}(S_{\text{train}})$. The adversary has black-box access to M.
  4) $x_r = 1$ if $r \in S_{\text{train}}$, otherwise $x_r = 0$.
  5) The adversary produces a guess $x'_r \in \{1, \perp\}$ and performs an attack only if $x'_r = 1$.
  6) If $x'_r = 1$ and $x_r = 1$, the adversary succeeds. If $x'_r = 1$ and $x_r = 0$, the adversary fails.

The probability is calculated over the randomness of the training dataset and the randomness of training algorithm A. (2)-(6) are repeated to uniformly sample the random space and estimate the success probability of the attack.

- Difference between pragmatic attack and indiscriminate attack:
  - Instead of attacking all records, a pragmatic adversary carefully selects the attack target records that are vulnerable to MIA.
  - A pragmatic adversary tries to minimize the false positives: the adversary makes a positive inference only if it has high confidence that the target record in in the training set, otherwise it makes on inferences.

Pragmatic attack overview

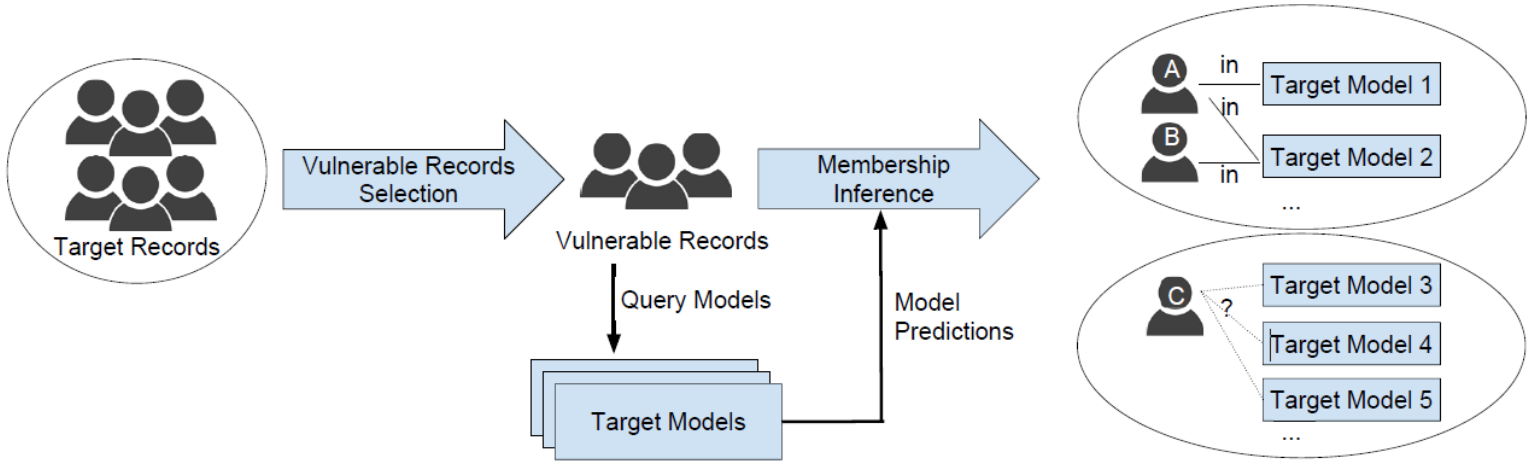


Figure 1: Attack Overview. Our attack consists of 3 steps. First, the adversary selects vulnerable records from a group of target records. Second, the adversary generate queries related to the vulnerable records and query the target models. Finally, the adversary identifies target models that are trained with the vulnerable records. The adversary only makes a positive membership inference if a target model's predictions strongly indicate the presence of a target record in its training dataset. Otherwise, the adversary makes no inferences. In the above example, the adversary would infer that target model 1 is trained with record A and target model 2 is trained with records A and B. However, the adversary would not make any inferences on record C and target models 4,5,6 because the predictions of the models do not give it high confidence for positive inferences.

Step1: Identify vulnerable target records

- Idea: a vulnerable record has very few neighbor records and thus has high influence on the target model.
- Reference model: similar to shadow model, trained with the same algorithm as the target model and trained on reference datasets that are sampled from the same space as the target training set, but not containing the target record.
- Feature vector:

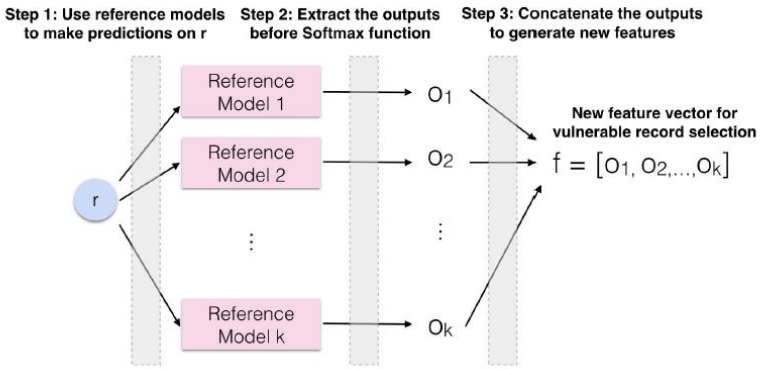


Figure 3: Features for vulnerable records selection. We concatenate intermediate outputs of locally trained reference models and use them as features for vulnerable records selection.

- Select vulnerable records: Distance in the feature vector space defined as cosine distance. Two records r1 and r2 are neighbors if the cosine distance between their feature vectors are below a threshold α. A record r is considered to be vulnerable if its number of neighbors is below a threshold β.

Identify vulnerable models – direct inference

- Given a target model M, a target record r, and k reference models.
- First we obtain the loss of all reference models on r as $L_1, L_2, \ldots, L_k$. we view these losses as samples i.i.d. drawn from a distribution D_L and estimate the CDF of D_L as F(L), which takes a real-valued loss L as input.
- Based on the loss of r on target model M, we estimate the confidence of r to be present in the training set:

$$p = F\left(\mathcal{L}\left(M, r\right)\right)$$

If p is smaller than a threshold (e.g. 0.01), r is in training set;

Otherwise, r in not in training set.

Identify vulnerable models – indirect inference

- Indirect inference determine the membership of the target record r without querying r itself, but by querying records(enhancing records) seemingly uncorrelated with r.
- Enhancing record: the record whose outputs from the target model are expected to be influence largely by the target model.
- Method to find enhancing records:
  1. Random record generation as initialization
  2. Enhancing record selection:

$$I\left(r,q\right)=\frac{1}{k}\sum_{i=1}^{k}\mathrm{t}\left(M_{\mathrm{ref}_i}^{r}\left(q,y_r\right)-M_{\mathrm{ref}_i}\left(q,y_r\right)\right)\ , \quad (1)$$

, where $\mathrm{M}_{\mathrm{ref_i}}$ is the original reference models (shadow models training without the target record), $M_{ref_i}^{r}$ is trained using original reference models with the target record. y_r is the label of the target record r.

t is the following threshold function:

$$\mathrm{t}\left(x\right)=\begin{cases}1 & \text{if} \quad x>0,\\ 0 & \text{otherwise.}\end{cases}$$

If $I(r,q)$ is above a threshold, then q is a enhancing record.

The intuition behind equation (1): q is a enhancing record if adding r to the training dataset almost always increase the model's output probability on the class label y_r for the query q.

Algorithm 1 Enhancing Records Selection Algorithm

procedure $\mathrm{select}_\theta(q)$ ▷ Input a random query
  $I\left(r,q\right)\leftarrow\sum_{i=1}^{k}\mathrm{t}\left(M_{\mathrm{ref}_i}^{r}\left(q,y_r\right)-M_{\mathrm{ref}_i}\left(q,y_r\right)\right)/k$
  if $I>\theta$ **then**
    Accept q ▷ Use q in MIA
  else
    Reject q

Identify vulnerable models – indirect inference

Execute algorithm 1 for a large amount of random records is slow, to speed up, two methods can be used :

- Enhancing record optimization: search for enhancing records by optimizing the following objective

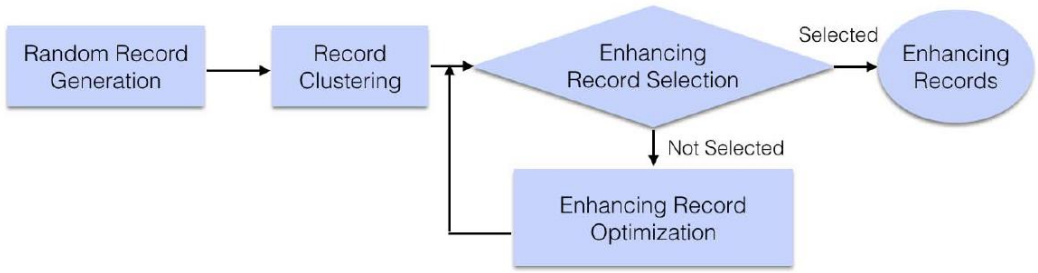
$$\max_{q} I\left(r, q\right), \tag{2}$$

equation (2) is hard to optimize and can be approximated by the following objective:

$$\min_{q} \sum_{i=1}^{k} \max\left(0, \gamma - \left(M^{r}_{\mathrm{ref}_i}\left(q, y_r\right) - M_{\mathrm{ref}_i}\left(q, y_r\right)\right)\right), \tag{3}$$

where γ is a parameter indicating the margin width of hinge loss.

- Record clustering: improve the random initialization step of enhancing record generation:
  - cluster the random selected records into k disjoint clusters based on their pairwise cosine distance.
  - in each cluster, select the record with least average cosine distance to all other records in the same cluster.
- Steps for generating enhancing records:



- After generated multiple enhancing records, we repeat the direct inference step for the enhancing records.

Experiments

- Datasets:
  - MNIST
  - Adult
  - Cancer

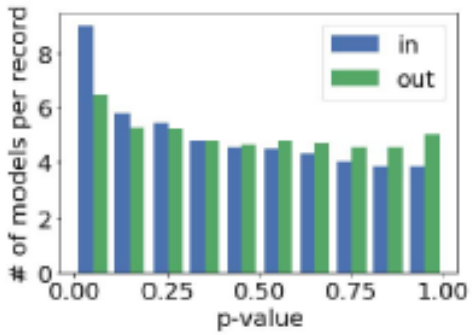
- Models:
  - Google ML Engine (no knowledge of the model architecture)
  - MLP with 3 hidden layers(for Adult), 4-layer CNN(for MNIST), logistic regression(for Cancer)

- Metric of success of attack:
  - precision
  - recall (i.k.a. coverage)

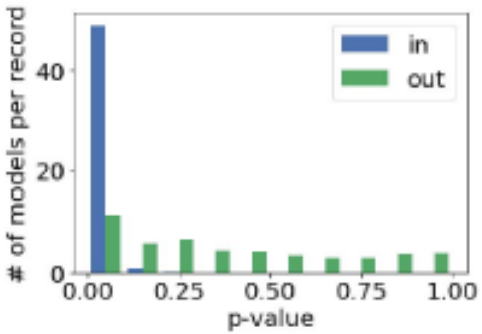
Dataset (Model)	Training Accuracy	Testing Accuracy
Adult	0.85 ± 0.01	0.85
Cancer	0.95 ± 0.04	0.94 ± 0.03
MNIST	0.99	0.98
Adult (Google)	0.84 ± 0.03	0.84 ± 0.02
MNIST (Google)	0.90	0.90

TABLE 2: Training and Testing Accuracy of Target Models. All the target models were well-generalized models with difference between training and testing accuracy smaller than 0.01.

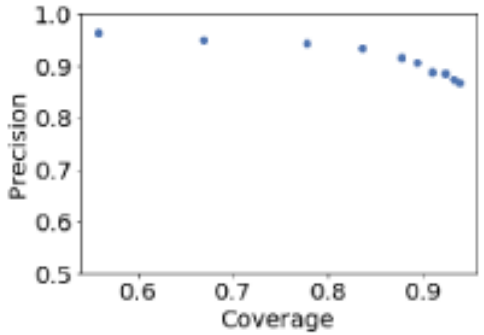
Effectiveness of pragmatic attack with directive inference



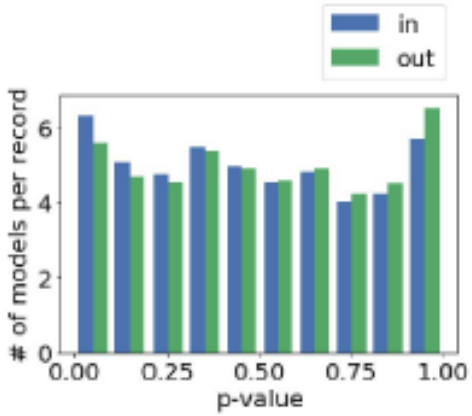
(a) p-values for all MNIST records



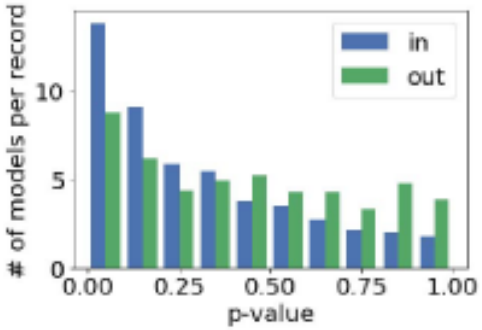
(b) p-values for selected MNIST records



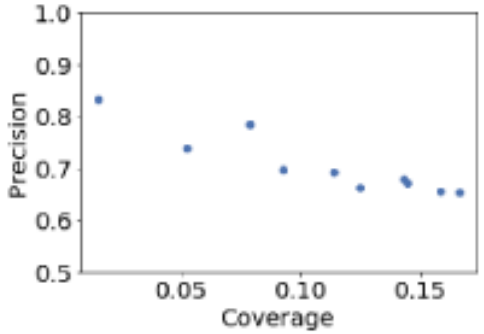
(c) attack performance on MNIST



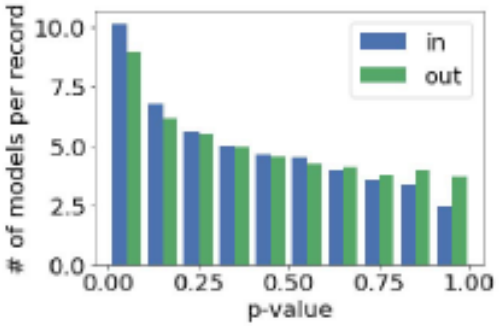
(d) p-values for all Adult records



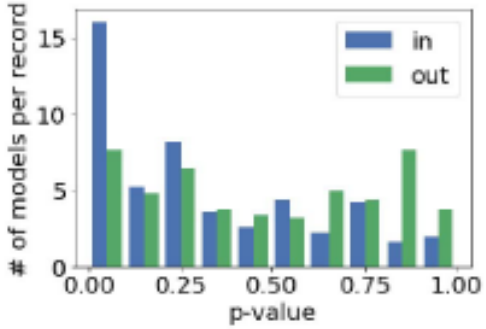
(e) p-values for selected Adult records



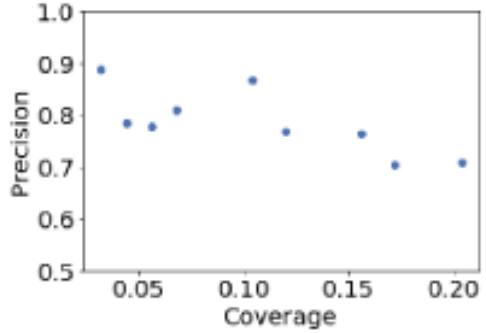
(f) attack performance on Adult



(g) p-values for all Cancer records



(h) p-values for selected Cancer records



(i) attack performance on Cancer

Dataset (Model)	Vulnerable Records	Precision	Coverage
MNIST	27	95.05%	66.89%
Adult	13	73.91%	5.23%
Cancer	5	88.89%	3.20%
MNIST (Google)	1	100%	4%
Adult (Google)	7	80%	2.67%

TABLE 1: Performance of Direct Inference. We measured the performance of a direct inference attack by its precision and coverage. To achieve a high precision, we selected a few vulnerable records (neighbor threshold $\alpha = 0.2$ for MNIST, 0.4 for Adult, and 0.1 for Cancer; probability threshold $\beta = 0.1$), and made positive inferences only when attack confidence is high ($p \leq 0.01$).

Direct Inference vs Indirect Inference

Dataset	Cut-off p-value	Prec. (direct)	Coverage (direct)	Prec. (indirect)	Coverage (indirect)
Adult	0.01	-	0	1	14%
	0.1	70.83%	34%	75%	24%
Cancer	0.01	1	6%	-	0
	0.1	66.67%	52%	88.89%	16%
MNIST	0.01	96.15%	1	1	2%
	0.1	89.29%	1	52.38%	22%

TABLE 3: Comparison between direct and indirect inferences. We performed the attack on the same selected record with direct inference and indirect inference. The result indicates that membership inference attack is feasible without directly querying the target record. On the Adult dataset, indirect inferences even outperformed direct inferences.

Compare with Shokri MIA

Dataset	Attack Confidence Threshold	Precision	Coverage
Cancer (3 records)	0.8 0.9	50.25% -	40% 0
Adult (13 records)	0.6 0.7	66.67% -	4.92% 0
MNIST (27 records)	0.6 0.7 0.8	50% 19.6% -	56.25% 6.25% 0

TABLE 5: Performance of the attack of Shokri et al. [30] on the same target models and the same target records. To imitate the attack strategy of a pragmatic adversary, we performed prior attack on the selected target records and made predictions only when the attack classifier has high confidence. However, the prior indiscriminative attack could not achieve high precision even under a low coverage.

Visualization of some vulnerable records

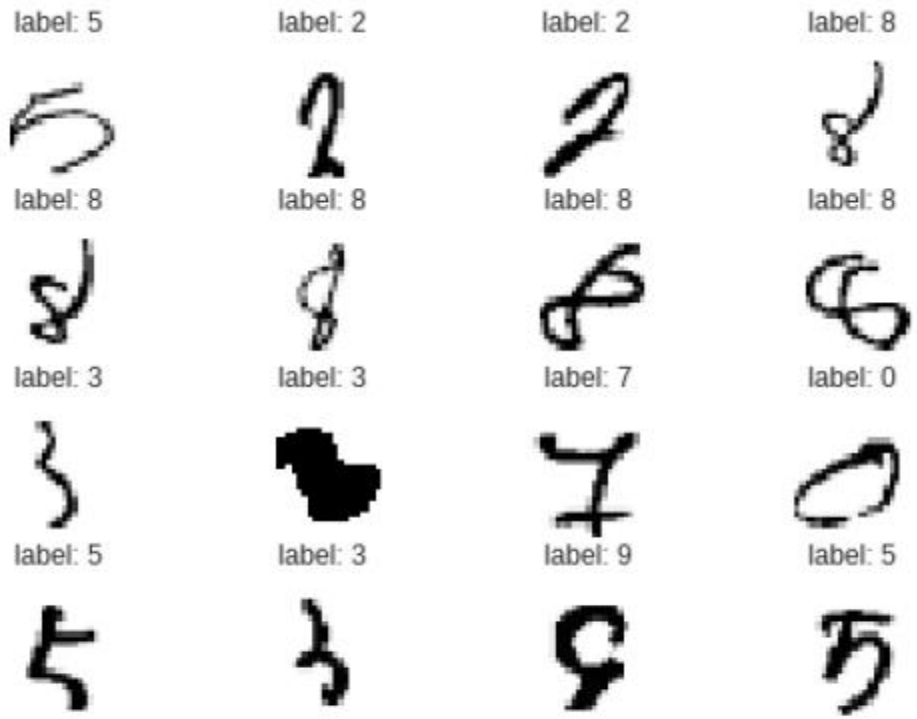


Figure 9: Vulnerable Examples in MNIST Dataset

Visualization of enhancing records of a target record

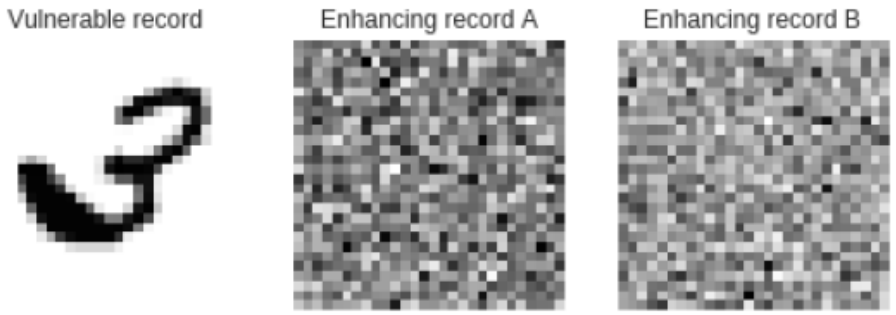


Figure 10: A vulnerable record from MNIST with its two enhancing records. In practice, it is difficult to find out what the target record is, by looking at the enhancing records used by an adversary.

Attack performance w.r.t. regularization

Regularization Coefficient λ	Training Acc.	Test Acc.	# of Target Records	Prec.	Coverage
0	0.99	0.98	52	90.84%	68.31%
0.001	0.99	0.99	1	1	54.8%
0.01	0.98	0.98	1	93.36%	4%

TABLE 4: Attack Performance w.r.t. Regularization ($\alpha = 0.2$, $\beta = 2$, $p \leq 0.01$) on MNIST dataset. We applied L2 regularization with varying coefficients λ. Experiment results show that applying regularization reduced, but did not fully eliminate the privacy risk of a pragmatic adversary.