Cascading and Proxy Membership Inference Attacks





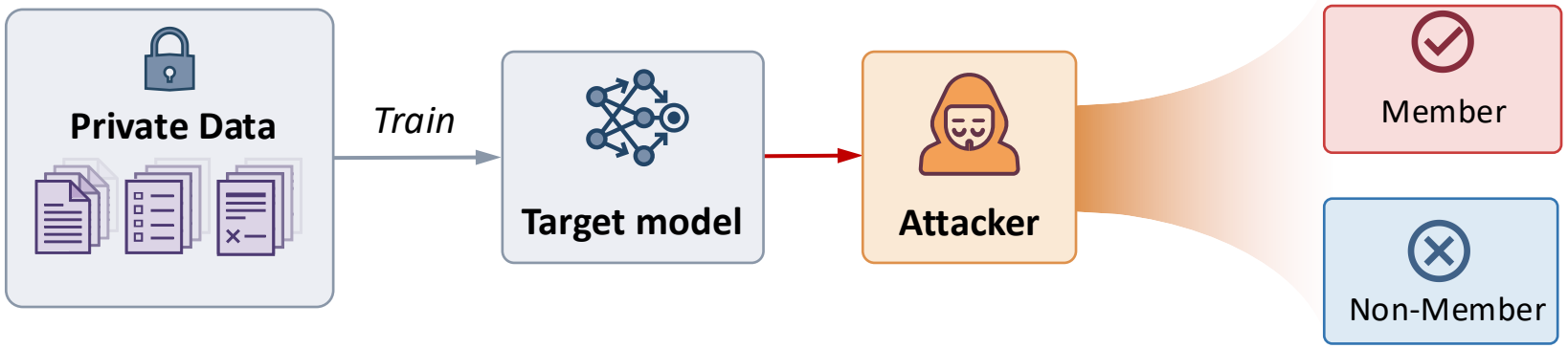



Yuntao Du, Jiacheng Li, Yuetian Chen, Kaiyuan Zhang, Zhizhen Yuan, Hanshen Xiao, Bruno Ribeiro, Ninghui Li

Purdue University

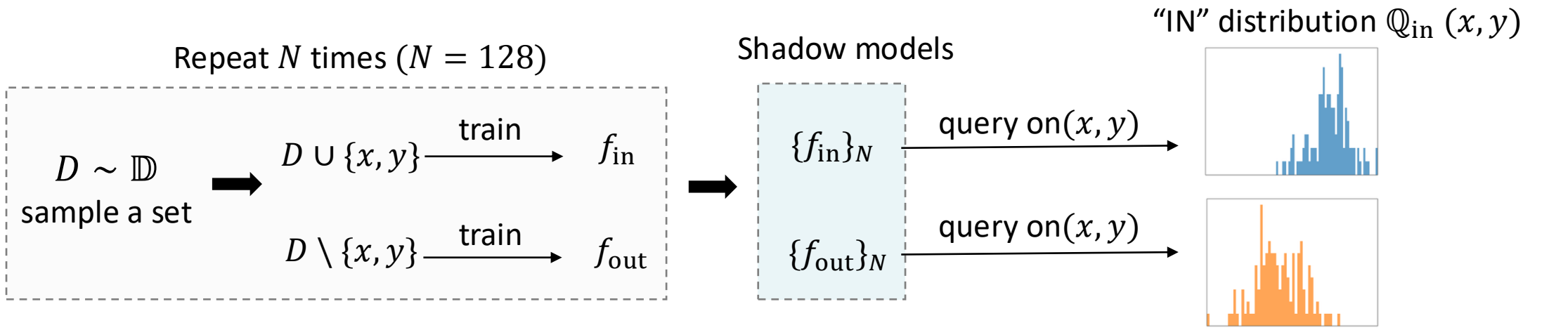
Membership Inference Attack (MIA)

- **Goal**: determine whether some specific data instances were used to train the target machine learning (ML) model
- Adversary model for MIAs
  - Types of access to the target model
    - Black-box (our focus): can query the target model with output logits/loss
    - White-box / label-only / federated
  - Auxiliary Information
    - Distribution of training data
    - Model architecture and training recipe for target model (enable training shadow models)



A Recipe for State-of-the-art MIAs

- Shadow training
  - For each membership query instance (x, y), train N shadow "IN"models and N shadow "OUT"models
  - Query the target model f_θ, compute a membership score by comparing its likelihood with the learned IN and OUT behaviors
- All SOTA MIAs use shadow training techniques

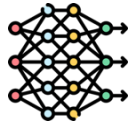


Reza Shokri, et al. ***Membership Inference Attacks Against Machine Learning Models****. IEEE S&P 2017.*
Nicholas Carlini, et al. ***Membership Inference Attacks From First Principles.*** *IEEE S&P 2022.*

Attack Settings in MIA

- MIA involves three datasets
  - Training data of the target model D
  - Membership query set D_{query}
  - Adversary's dataset D_{adv}
- MIA papers use $\mathrm{D} \subset D_{\text{query}}$
  - The membership queries contain all members and the same size of non-members
- The relationship between D_{query} and D_{adv} would result in different attack settings/strategies and disparate performance
  - Online/adaptive setting, offline/non-adaptive setting
  - Previous security games fail to elucidate the difference

Previous Membership Inference Game

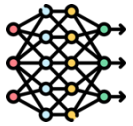
- Challenger samples a dataset $D \sim \mathbb{D}$ and train a model $D \xrightarrow{\text{train}}$  f_θ
  - Adversary is given access to the target model f_θ
- Challenger flips a bit $b \sim \{0,1\}$
  - Select $(x, y) \sim D$ if $b = 1$, otherwise $(x, y) \sim \mathbb{D}\ (x \notin D)$
- Challenger send query instance (x, y) to adversary
- Adversary responds with b' and wins if $b = b'$

Defining security game as determining the membership of **a single instance**:

- Cannot distinguish adaptive vs. non-adaptive (e.g., when shadow models are trained)
- Not clear how these datasets ($D, D_{\text{query}}, D_{\text{adv}}$) are related
- Does not match the experimental evaluation (e.g., TPR@lowFPR, AUC, etc)

Samuel Yeom, et al. ***Privacy risk in machine learning: Analyzing the connection to overfitting.*** *IEEE CSF, 2018.*
Bargav Jayaraman, et al. ***Revisiting membership inference under realistic assumptions.*** *PET, 2021.*

Our Membership Inference Game

- Challenger and Adversary both have access to distribution $\mathbb{D}$, and we assume that independently sampled datasets have few overlaps.
- Challenger samples a dataset $D \sim \mathbb{D}$ and train a model $D \xrightarrow{\text{train}}$  f_θ
  - Adversary is given access the target model f_θ
- Challenger sample a query set $D_{query} = D_a \cup D_b$, where $D_a \subseteq D$ and $D_b \sim \mathbb{D}$
- Challenger send query set D_{query} to Adversary
- Adversary responds with guesses of members $D_{ans} \subseteq D_{query}$

Shadow training

Shadow training

Adaptive setting ($D_{query} = D_{adv}$).
Adversary is allowed to train shadow models ***after*** receiving D_{query}, enable training both "IN" and "OUT" shadow models for instance in D_{query}

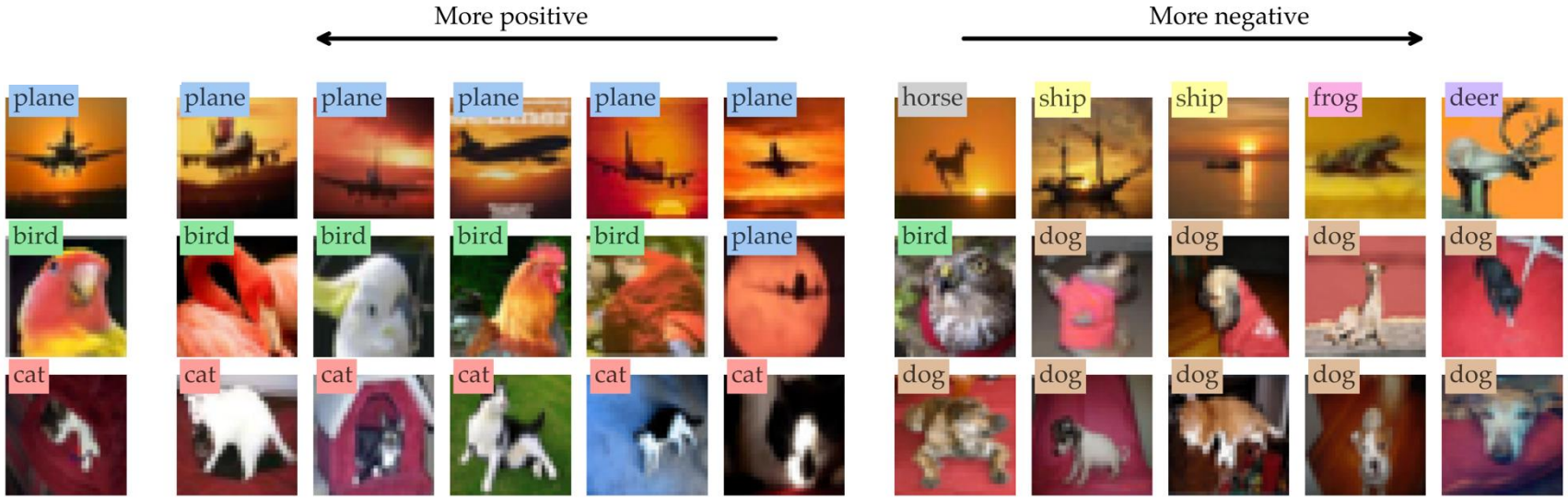
Cascading Membership Inference Attack (CMIA)

Non-Adaptive setting ($D_{query} \cap D_{adv} = \emptyset$).
Adversary is only allowed to train shadow models ***before*** receiving D_{query}, result in only "OUT" shadow models for instance in D_{query}

Proxy Membership Inference Attack (PMIA)

Cascading MIA: a new adaptive framework

- Intuition: the membership of some instances would impact the membership inference of other instances
  - Assume one instance x_2 is very similar to the target instance x_1
  - When x_1 is used for training, the model outputs low loss for x_2 even when x_2 is a non-member
  - This is known as data influence/attribution problems in ML



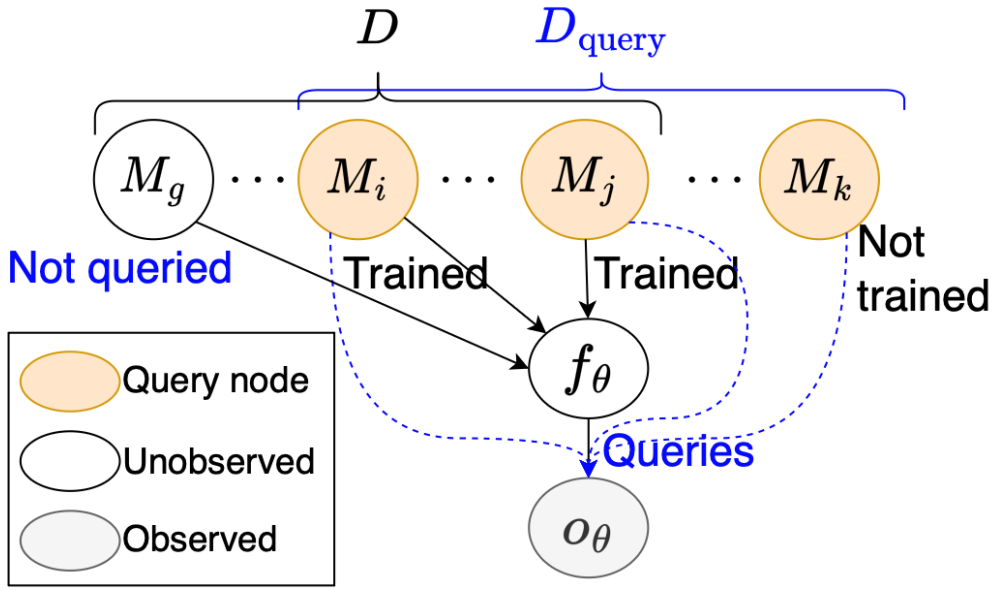
Andrew Ilyas, et al. ***Datamodels: Predicting Predictions from Training Data.*** *International Conference on Machine Learning (ICML) 2022.*

Cascading MIA: a new adaptive framework

- Key observation: membership is not independent when conditional on the outputs of target model

M_i: membership of instance x_i

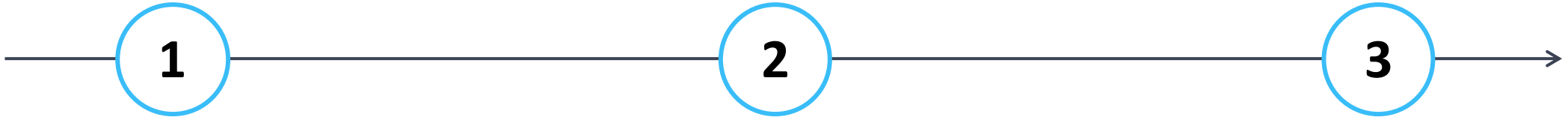
M_j: membership of instance x_j



Membership is not independent! $M_i \not\perp\!\!\!\perp M_j \mid o_\theta$

Cascading MIA: a new adaptive framework

- Can we exploit such dependence for better membership inference?
- Core idea: start from easy-to-guess instances, use their membership status as priors to help infer hard-to-guess datapoints



Initial

Perform base (static) attack to identify easy-to-guess instances

Anchoring

Fix the membership of these instances as *anchors*

Cascading

Use the anchors to help to infer remaining instances

Domino effect

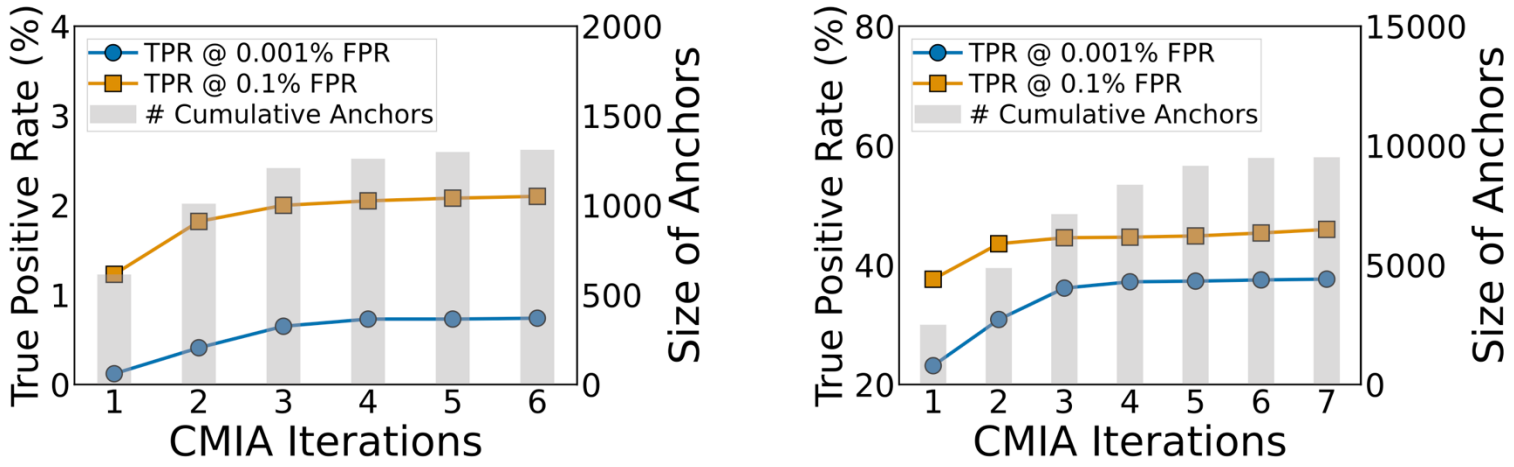
Cascading MIA: results

- It can be applied to any shadow-based MIAs and boosts performance
  - The stronger the base model, the higher performance improvement

Method		TPR @ 0.001% FPR				TPR @ 0.1% FPR				Balanced Accuracy			
		MNIST	FMNIST	C-10	C-100	MNIST	FMNIST	C-10	C-100	MNIST	FMNIST	C-10	C-100
Calibration	Base	0.01%	0.52%	0.28%	1.48%	0.19%	2.23%	1.02%	5.51%	51.05%	54.21%	54.62%	61.18%
	CMIA	**0.08%**	**1.24%**	**0.59%**	**3.81%**	**0.55%**	**4.72%**	**3.65%**	**8.52%**	**52.21%**	**55.37%**	**56.13%**	**64.09%**
	%Imp.	700.00%	138.46%	110.71%	157.43%	189.47%	111.66%	257.84%	54.63%	2.27%	2.14%	2.76%	4.76%
Attack-R	Base	0.00%	0.00%	0.21%	1.40%	0.10%	0.00%	1.30%	4.82%	52.15%	57.83%	54.26%	62.13%
	CMIA	**0.00%**	**0.00%**	**0.45%**	**2.01%**	**0.37%**	**0.00%**	**1.95%**	**6.04%**	**52.95%**	**58.48%**	**55.48%**	**63.93%**
	%Imp.	-	-	114.29%	43.57%	270.00%	-	50.00%	25.31%	1.53%	1.12%	2.25%	2.90%
LiRA	Base	0.12%	2.72%	2.64%	23.15%	1.23%	6.28%	8.45%	37.62%	51.26%	58.28%	62.52%	82.05%
	CMIA	**0.77%**	**4.42%**	**3.86%**	**36.74%**	**2.10%**	**8.34%**	**9.71%**	**45.37%**	**52.67%**	**60.91%**	**63.83%**	**84.89%**
	%Imp.	541.67%	62.50%	46.21%	58.70%	70.73%	32.80%	14.91%	20.60%	2.75%	4.51%	2.10%	3.46%
Canary	Base	0.15%	2.95%	2.36%	25.78%	1.28%	6.65%	8.12%	38.25%	53.76%	58.94%	62.60%	83.11%
	CMIA	**0.84%**	**4.73%**	**3.61%**	**37.85%**	**2.48%**	**8.47%**	**9.02%**	**45.96%**	**55.60%**	**61.07%**	**63.81%**	**84.72%**
	%Imp.	460.00%	60.34%	52.97%	46.82%	93.75%	27.37%	11.08%	20.16%	3.42%	3.61%	1.93%	1.94%
RMIA	Base	0.21%	2.05%	1.43%	10.72%	0.96%	4.71%	5.24%	30.13%	52.99%	58.16%	62.05%	80.64%
	CMIA	**0.52%**	**3.56%**	**2.05%**	**14.67%**	**1.62%**	**5.81%**	**6.05%**	**37.51%**	**53.51%**	**60.90%**	**62.49%**	**82.53%**
	%Imp.	147.62%	73.66%	43.36%	36.85%	68.75%	23.35%	15.46%	24.49%	0.98%	4.71%	0.71%	2.34%
RAPID	Base	0.23%	1.31%	0.56%	9.83%	0.79%	3.44%	3.12%	21.69%	52.44%	58.40%	59.58%	75.83%
	CMIA	**0.48%**	**2.45%**	**0.94%**	**11.83%**	**1.24%**	**4.73%**	**4.75%**	**25.90%**	**52.97%**	**58.51%**	**59.77%**	**78.52%**
	%Imp.	108.70%	87.02%	67.86%	20.35%	56.96%	37.50%	52.24%	19.41%	1.01%	0.19%	0.32%	3.55%

Cascading MIA: results

- The most significant improvements occur during the first few cascading iterations

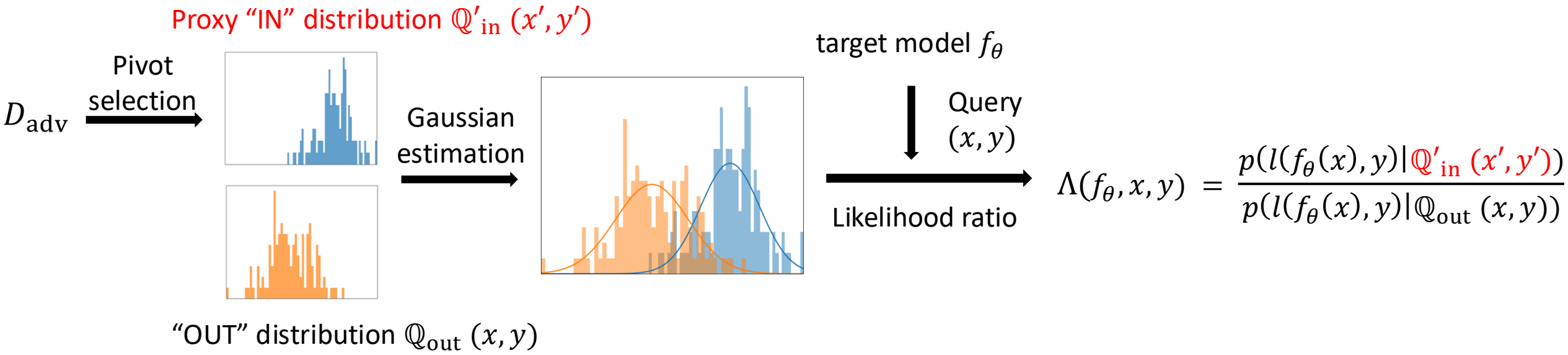


(a) MNIST (b) CIFAR-100

The impact of number of cascading iterations in CMIA when using LiRA as the base attack

Proxy MIA: a new non-adaptive attack

- Intuition: although the adversary do not have "IN" behaviors for the membership query, but they have the "IN" behaviors for the shadow' models training data (i.e., pivot data, $D_{\text{pivot}} \in D_{\text{adv}}$)
- We can use these behaviors to approximate the "IN" behaviors for membership query D_{query} and conduct likelihood ratio attack (LiRA)



Nicholas Carlini, et al. ***Membership Inference Attacks From First Principles.*** *IEEE S&P 2022.*

Proxy MIA: results

- PMIA outperform all non-adaptive attacks across benchmarks

Method	TPR @ 0.001% FPR				TPR @ 0.1% FPR				Balanced Accuracy			
	MNIST	FMNIST	C-10	C-100	MNIST	FMNIST	C-10	C-100	MNIST	FMNIST	C-10	C-100
LOSS	0.01%	0.01%	0.00%	0.00%	0.08%	0.09%	0.00%	0.00%	52.81%	61.51%	63.35%	78.20%
Entropy	0.01%	0.01%	0.00%	0.00%	0.08%	0.10%	0.00%	0.21%	52.80%	61.16%	63.08%	78.05%
Calibration	0.05%	0.06%	0.08%	1.08%	0.34%	0.45%	1.03%	2.83%	52.51%	55.10%	57.96%	66.10%
Attack-R	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	52.62%	58.47%	63.46%	77.36%
LiRA	0.09%	0.05%	0.03%	0.98%	0.30%	0.67%	0.78%	8.56%	50.54%	53.11%	58.97%	73.25%
Canary	0.11%	0.08%	0.03%	1.78%	0.30%	1.02%	0.77%	7.35%	51.01%	53.79%	58.77%	73.93%
RMIA	0.17%	0.05%	0.41%	2.73%	0.51%	1.25%	2.60%	6.64	52.78%	58.96%	62.72%	77.53%
RAPID	0.09%	0.15%	0.15%	1.16%	0.45%	0.44%	1.34%	3.14%	52.05%	58.42%	61.39%	78.49%
PMIA	**0.31%**	**0.17%**	**1.20%**	**5.90%**	**1.01%**	**2.80%**	**3.29%**	**11.5%**	**52.87%**	**61.56%**	**64.34%**	**80.4%**
%Imp.	82.35%	13.33%	192.68%	116.12%	98.04%	124.00%	26.54%	34.35%	0.11%	0.08%	1.39%	2.43%

TABLE VI: Membership inference time cost of non-adaptive attacks against a ResNet50 model on MNIST.

Attack Method	LOSS	Entropy	Calibration	Attack-R	LiRA	Canary	RMIA	RAPID	PMIA
Inference Cost/seconds	1.23	2.52	1.85	3.03	10.47	$>$ 400,000	49.5	31.5	**15.8**

Thank you !

Yuntao Du
ytdu@purdue.edu

Homepage

Code & Datasets