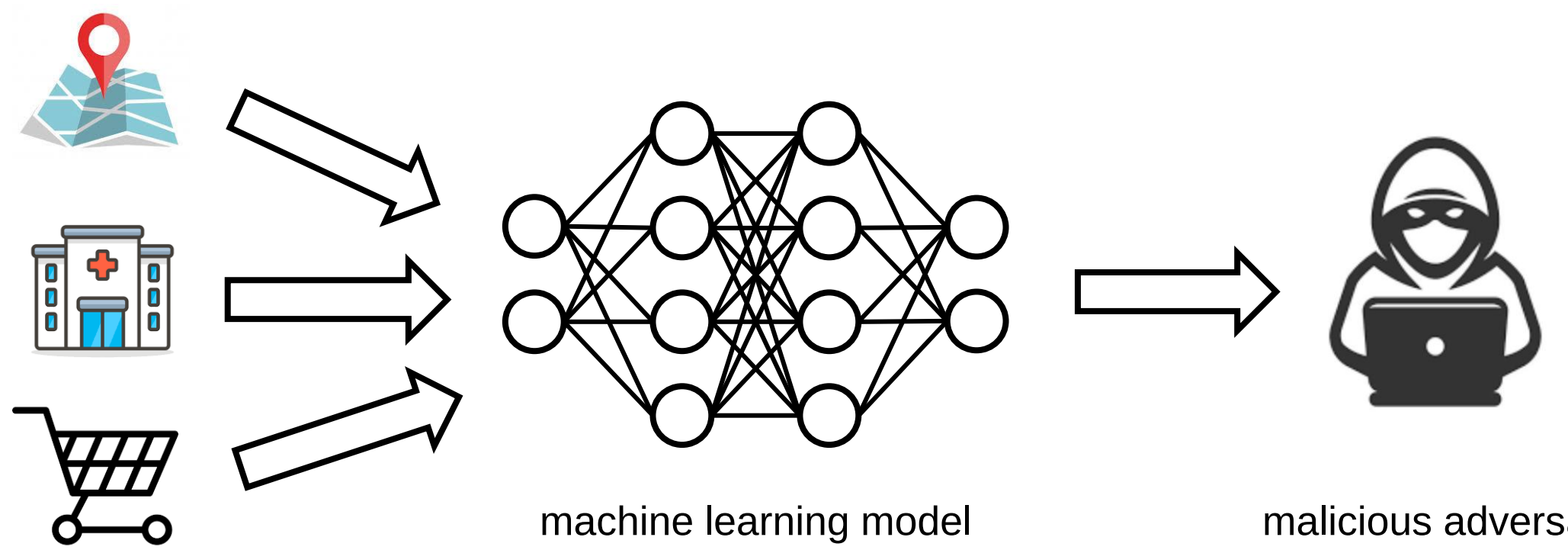
Systematic Evaluation of Privacy Risks of Machine Learning Models

Liwei Song, Prateek Mittal

PRINCETON UNIVERSITY

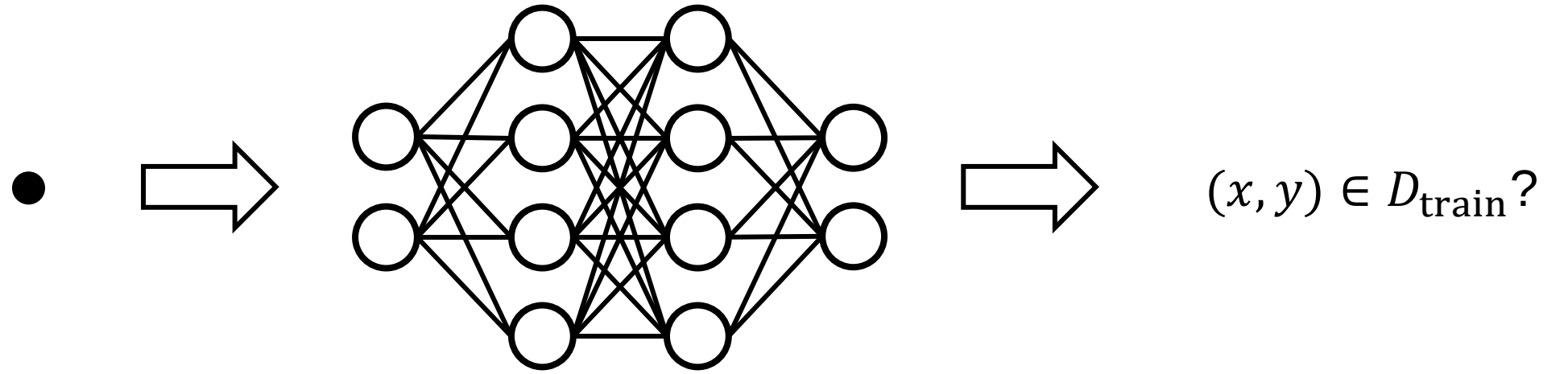
Privacy Risks in Machine Learning



Without rigorous defense methods, the malicious adversary can infer private information of users' data

Privacy Risks in Machine Learning

- Membership inference attacks
  - Guess whether a sample was used to train the target machine learning model or not
  - Distinguishability between training data (members) and test data (non-members)



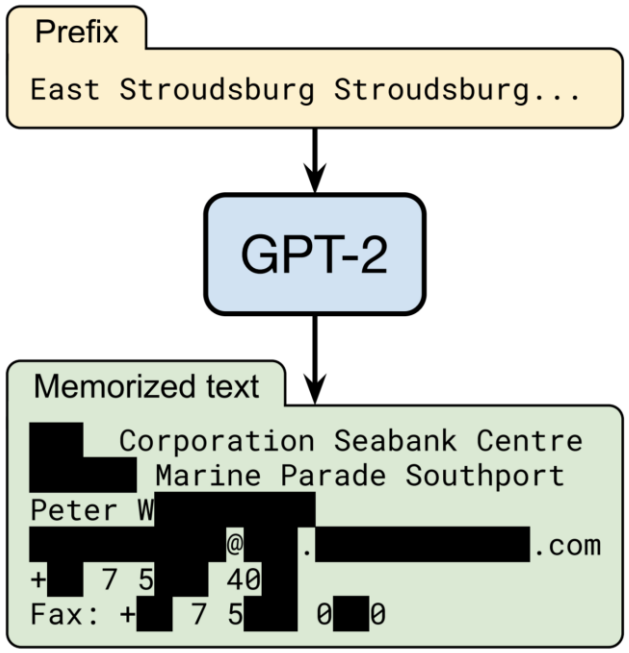
input (x, y)

target classifier F

membership inference adversary

Membership Inference Attacks

- Provide foundation for performing training data extraction attacks
- Quantify the privacy provided by differential privacy implementations and help to guide the selection of privacy parameters in statistical privacy frameworks



Carlini et al. “Extracting training data from large language models.”

Evaluating Differentially Private Machine Learning in Practice

Bargav Jayaraman and David Evans, *University of Virginia*

https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman

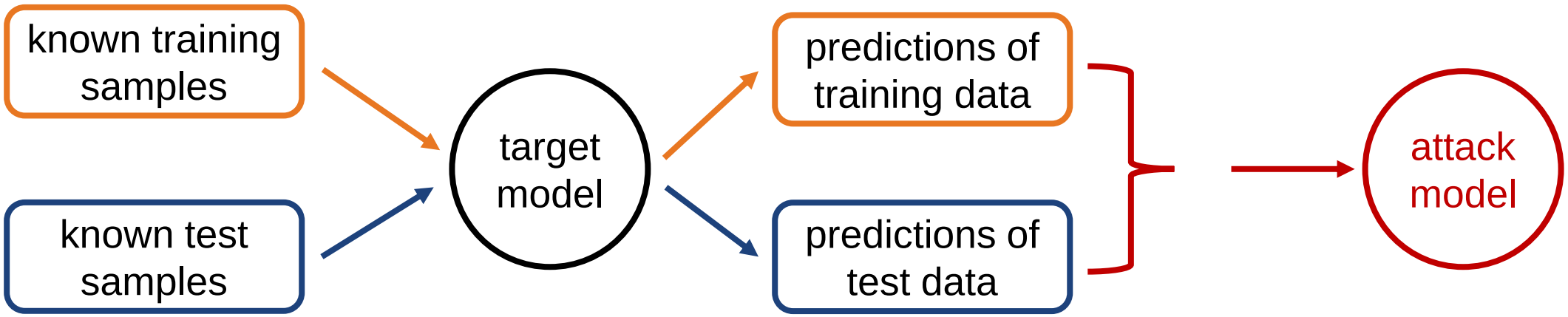
Proceedings on Privacy Enhancing Technologies | Volume 2019: Issue 3

Investigating Statistical Privacy Frameworks from the Perspective of Hypothesis Testing

Changchang Liu [1], Xi He [2], Thee Chanyaswad [3], Shiqiang Wang [4], and Prateek Mittal [5]

Membership Inference Attack Pipeline

- Consider the membership inference (MI) attack as a binary machine learning problem



Our Contributions

Benchmark aggregate membership inference attacks

- Propose a suite of metric-based attacks and use them to supplement existing neural network (NN) based MI attacks
- Evaluate multiple attack strategies and report the worst-case privacy risks

Fine-grained membership inference privacy analysis

- Define the privacy risk score to estimate each individual sample's likelihood of being a member
- Apply fine-grained analysis in conjunction with existing aggregate analysis for a thorough evaluation of privacy risks

Our evaluation methods have been integrated into Google's TensorFlow Privacy!

Our Contributions

Benchmark aggregate membership inference attacks

- Propose a suite of metric-based attacks and use them to supplement existing neural network (NN) based MI attacks
- Evaluate multiple attack strategies and report the worst-case privacy risks

Fine-grained membership inference privacy analysis

- Define the privacy risk score to estimate each individual sample's likelihood of being a member
- Apply fine-grained analysis in conjunction with existing aggregate analysis for a thorough evaluation of privacy risks

Our evaluation methods have been integrated into Google's TensorFlow Privacy!

Benchmark Aggregate MI Attacks

Existing NN-based attacks

- Train dedicated neural network (NN) classifiers to distinguish between training members and non-members
- May underestimate privacy risks due to inappropriate hyperparameter settings

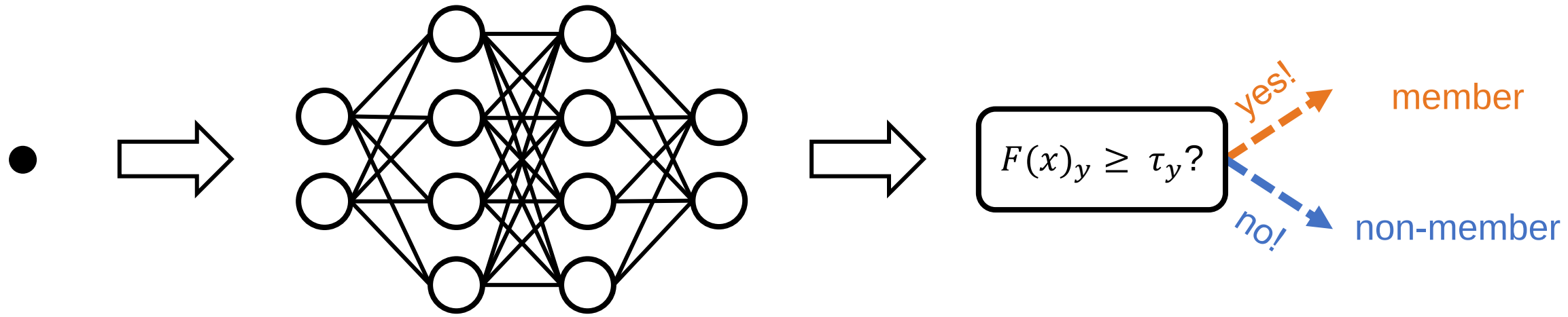
Our metric-based attacks

- Compute metrics (e.g., correctness, confidence) of model predictions, and compare them with threshold values
- Only need to tune threshold values, much easier than neural network training
- Threshold values are tuned in a class-dependent manner

Evaluate multiple (adaptive) attack strategies and report the worst-case privacy risks!

Improving Existing Attacks with Class-Dependent Thresholds

- The adversary infers a sample as a member if its prediction confidence is larger than a preset threshold, a non-member otherwise.
- Class-dependent thresholds: setting different values of τ_y for different labels y.



input (x, y)

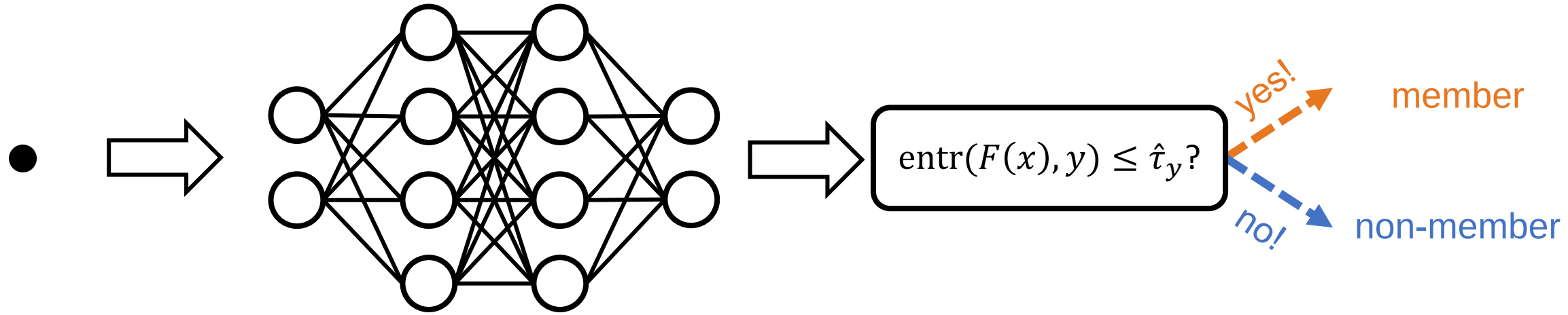
target classifier F

membership inference adversary (I_{conf})

Improving Existing Attacks with Class-Dependent Thresholds

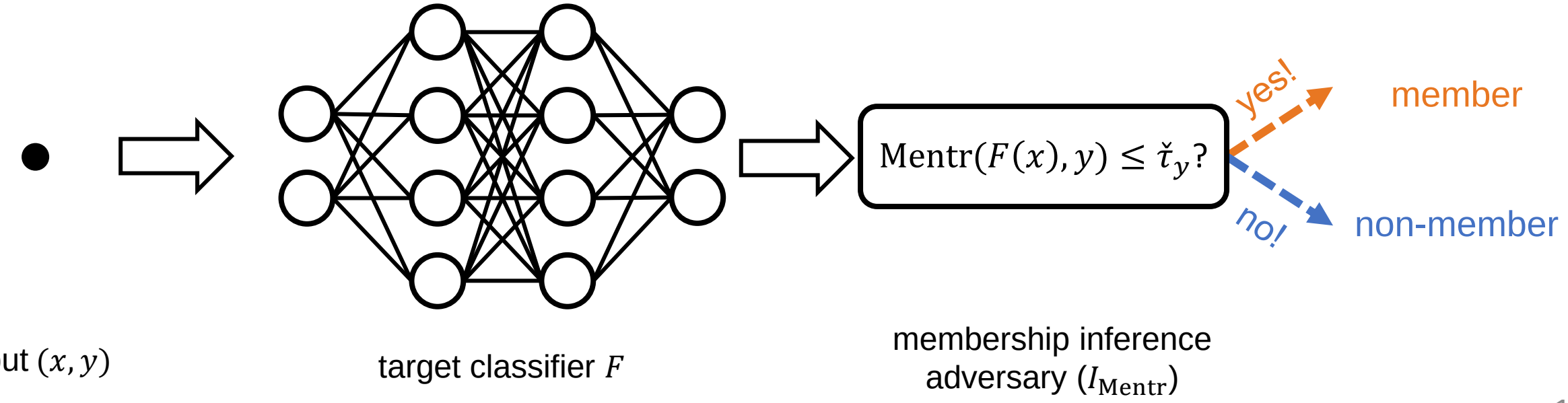
- The adversary infers a sample as a member if its prediction entropy is smaller than a preset threshold, a non-member otherwise.

- $\mathrm{entr}(F(x), y) = -\sum_i F(x)_i \log(F(x)_i)$.



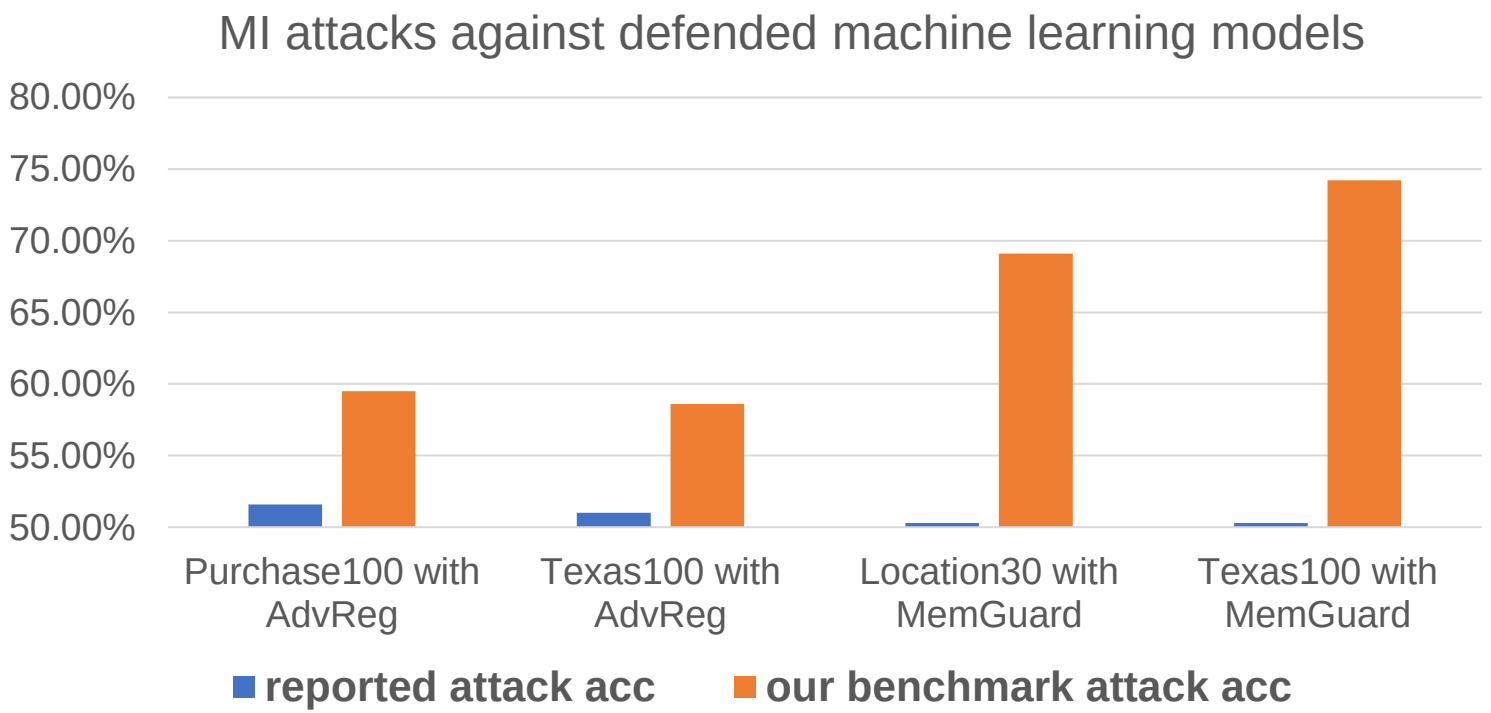
New Attack with Modified Prediction Entropy

- When having the ground truth label y, we want the entropy to be monotonically decreasing with $F(x)_y$, while monotonically increasing with $F(x)_i, i \neq y$.
- $\mathrm{Mentr}(F(x), y) = -(1 - F(x)_y)\log(F(x)_y) - \sum_{i \neq y} F(x)_i \log(1 - F(x)_i)$.
- The adversary infers a sample as a member if its modified prediction entropy is smaller than a preset threshold, a non-member otherwise.



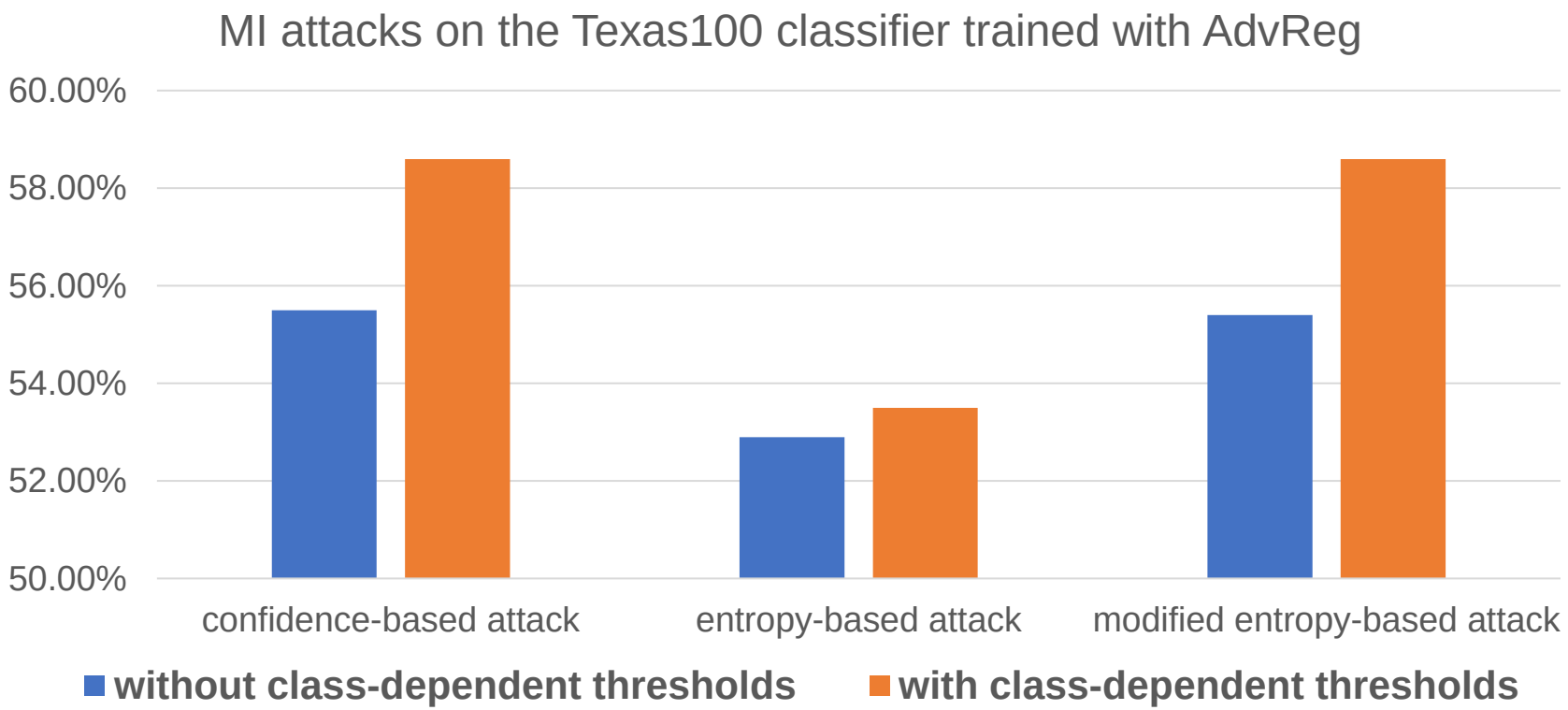
Re-evaluating state-of-the-art MI Defenses

- Apply all benchmark attack methods and report the highest attack accuracy
- Both adversarial regularization (AdvReg, *CCS'18*) and MemGuard (*CCS'19*) reported to decrease the attack success close to random guessing.
- With our benchmarks, the adversary can still achieve high attack success on the defended models.



Re-evaluating state-of-the-art MI Defenses

- By using the class-dependent thresholding technique, we increase the MI attack success by 1%~4%.
- Our new attack based on the modified entropy always outperforms the conventional entropy-based attack, and usually results in highest attack success



Our Contributions

Benchmark aggregate membership inference attacks

- Propose a suite of metric-based attacks and use them to supplement existing neural network (NN) based MI attacks
- Evaluate multiple attack strategies and report the worst-case privacy risks

Fine-grained membership inference privacy analysis

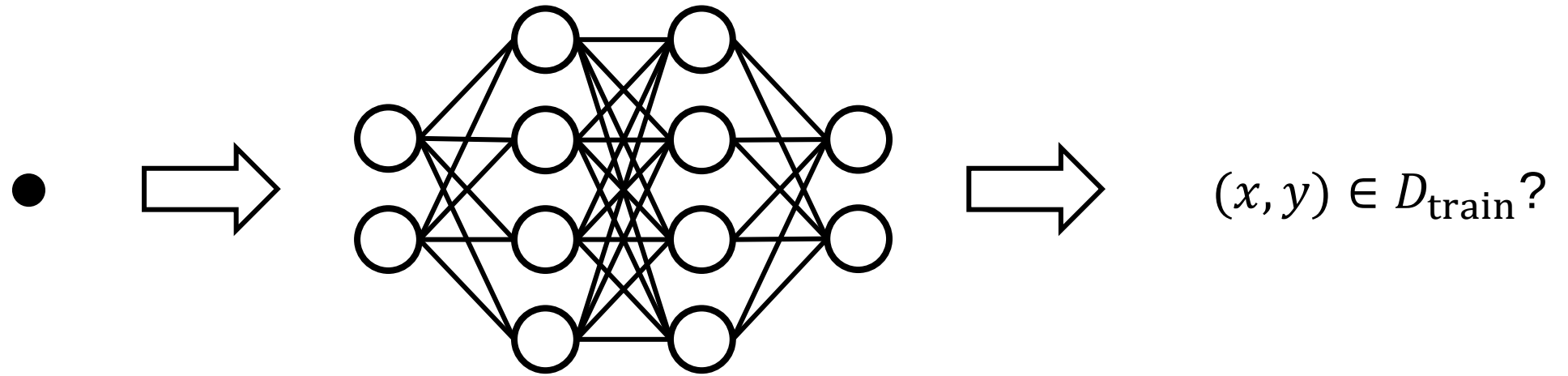
- Define the privacy risk score to estimate each individual sample's likelihood of being a member
- Apply fine-grained analysis in conjunction with existing aggregate analysis for a thorough evaluation of privacy risks

Our evaluation methods have been integrated into Google's TensorFlow Privacy!

Fine-Grained Privacy Analysis

- Definition of privacy risk score: the posterior probability of a sample $\mathrm{z} = (x, y)$ being in the training set D_{train} after observing the target model's behavior over that sample $\mathrm{O}(F, \mathrm{z})$

$$r(\mathrm{z}) = P(\mathrm{z} \in D_{\text{train}} \mid \mathrm{O}(F, \mathrm{z}))$$



Computation of Privacy Risk Score

- Use Bayes' theorem to compute the privacy risk score $r(\mathrm{z})$ based on the distribution of model's behavior conditioned on training/test set

$$r(\mathrm{z}) = \frac{P(\mathrm{z} \in D_{\text{train}}) * P(\mathrm{O}(F, \mathrm{z}) \mid \mathrm{z} \in D_{\text{train}})}{P(O(F, z))},$$

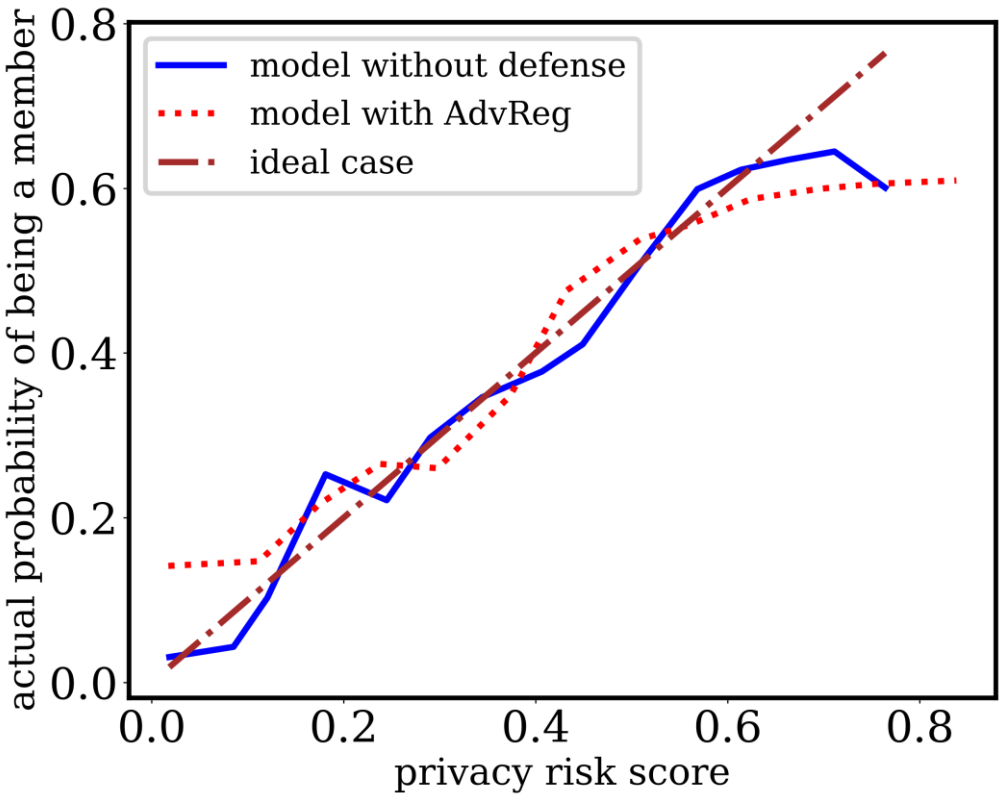
$$\text{where } P\big(O(F, z)\big) = P(\mathrm{z} \in D_{\text{train}}) * P(\mathrm{O}(F, \mathrm{z}) \mid \mathrm{z} \in D_{\text{train}}) + P(\mathrm{z} \in D_{\text{test}}) * P(\mathrm{O}(F, \mathrm{z}) \mid \mathrm{z} \in D_{\text{test}})$$

- Measure the conditional distribution in a class-dependent manner and approximate it with the distribution of modified prediction entropy

$$P(\mathrm{O}(F, \mathrm{z}) \mid \mathrm{z} \in D_{\text{train}}) \approx \begin{cases} P(\mathrm{Mentr}(F(x), y) \mid \mathrm{z} \in D_{\text{train}}, y = y_0), & \text{when } y = y_0 \\ P(\mathrm{Mentr}(F(x), y) \mid \mathrm{z} \in D_{\text{train}}, y = y_1), & \text{when } y = y_1 \\ \vdots & \\ P(\mathrm{Mentr}(F(x), y) \mid \mathrm{z} \in D_{\text{train}}, y = y_n), & \text{when } y = y_n \end{cases}$$

Validation of Privacy Risk Score

- Divide the entire range of privacy risk scores into multiple bins
- For each bin, the fraction of training samples indicates the ground-truth probability of being a member

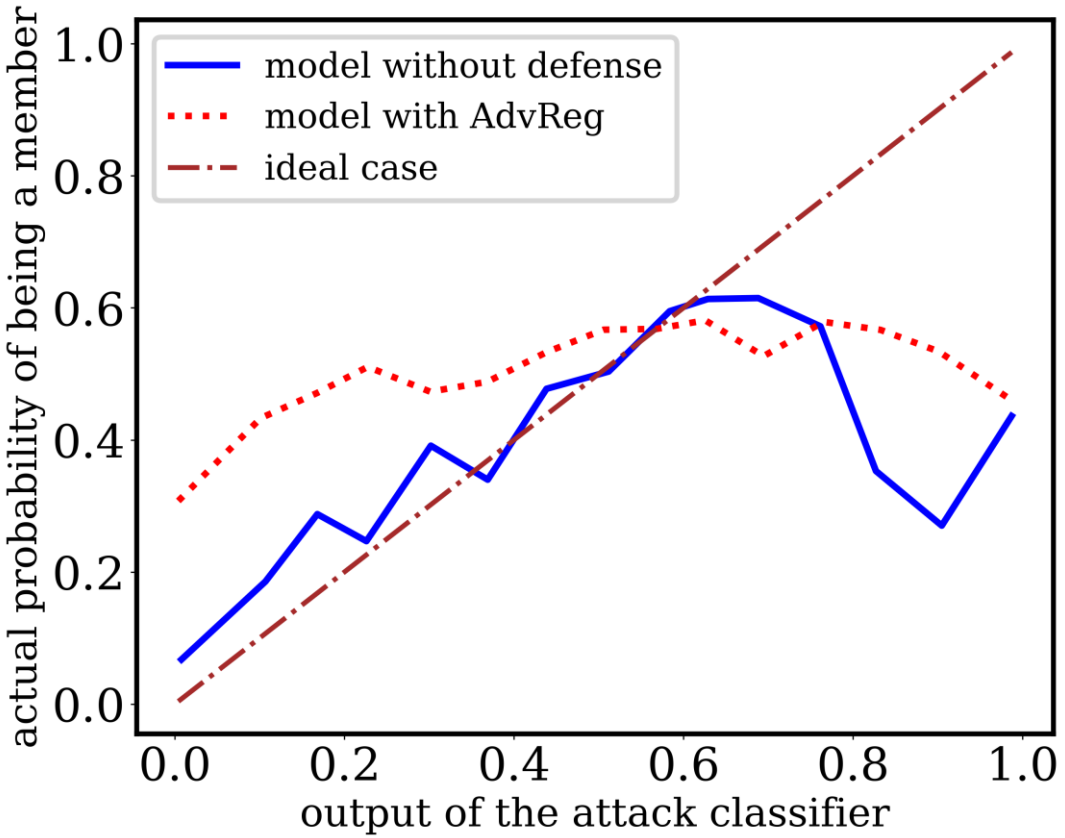


Validation of privacy risk score on Purchase100 dataset

Privacy risk score closely aligns with the actual probability of being a member!

Validation of Privacy Risk Score

- As a contrast, the output of the neural network attack classifier fails to capture the real likelihood of a sample being a member of the target model.



Usage of Privacy Risk Score

- We can perform MI attacks with high confidence: a sample is inferred as a member if and only if its privacy risk score is above a threshold value.

Threshold values on privacy risk score	1	0.9	0.7	0.5
Attack precision	88.2%	84.5%	77.0%	66.0%
Attack recall	1.4%	7.6%	43.7%	99.9%

MI attacks with high confidence against the Texas100 classifier with MemGuard

- With privacy risk score, we can identify training samples with high privacy risks.
  - Individual samples' privacy risk scores are highly correlated with their influence on the model, generalization errors, and feature embeddings. Check out our paper for more details.

Summary

- Propose metric-based MI attacks to benchmark aggregate privacy risks
  - Improve existing attacks with class-dependent threshold settings and design a new attack based on a modified entropy estimation
  - Adversarial regularization and MemGuard are not as effective as previously reported
- Propose the privacy risk score for a fine-grained privacy risk analysis
  - The privacy risk score is shown to well represent the likelihood of a sample being a member
  - We can perform attacks with high confidence and identify samples with high privacy risks
- Source code: https://github.com/inspire-group/membership-inference-evaluation
- Impact on Google's TensorFlow Privacy
  - Attack methods: https://github.com/tensorflow/privacy/pull/131
  - Fine-grained privacy analysis: https://github.com/tensorflow/privacy/pull/146

Thank You!

Contact liweis@princeton.edu for any questions