MEMBERSHIP INFERENCE ATTACKS

Pierre Joly
pierre.joly@etu.ec-lyon.fr

ABSTRACT

Membership inference attacks enable an adversary to determine whether a given data point was part of a model's training set, solely from model queries. This document reviews and analyzes two representative methods: "Membership Inference Attacks From First Principles" (LiRA), which models logit distributions using IN/OUT shadow models, and "Low-Cost High-Power Membership Inference Attacks" (RMIA), which relies on pairwise likelihood-ratio comparisons with population data. We contrast their methodologies, highlight their offline vs. online variants, and discuss trade-offs in accuracy, reference-model count, and computational overhead.

1 MEMBERSHIP INFERENCE ATTACKS FROM FIRST PRINCIPLES

Carlini et al. (2022) introduce the *Likelihood Ratio Attack (LiRA)*. In principle, for a data point (x, y), one compares the likelihood of observing a trained model f under two hypotheses: (1) (x, y) was in f's training set Q_{in}, versus (2) it was not Q_{out}. Formally:

$$\Lambda(f; x, y) \;=\; \frac{p\big(f \,\big|\, Q_{\text{in}}(x, y)\big)}{p\big(f \,\big|\, Q_{\text{out}}(x, y)\big)}, \tag{1}$$

but directly modeling $p(f|Q_{\text{in/out}})$ is intractable. Instead, the authors focus on the distribution of losses (or equivalently, logit-scaled confidences) for (x, y), which is well-approximated by a Gaussian.

Shadow Models. LiRA trains IN shadow models (which include (x, y) in their training subsets) and OUT shadow models (excluding (x, y)), typically using the same architecture and hyperparameters as the target model but retrained from scratch on different data partitions. For each shadow model f_i, one records the logit $\phi\big(f_i(x)_y\big)$. The mean and variance of these logits across all IN (resp. OUT) models yield $\mu_{\text{in/out}}$ and $\sigma^2_{\text{in/out}}$. Given a target model's logit $\phi\big(f(x)_y\big)$, LiRA applies:

$$\Lambda \;=\; \frac{p\Big(\phi(f(x)_y) \;\Big|\; \mathcal{N}(\mu_{\text{in}}, \sigma^2_{\text{in}})\Big)}{p\Big(\phi(f(x)_y) \;\Big|\; \mathcal{N}(\mu_{\text{out}}, \sigma^2_{\text{out}})\Big)}, \tag{2}$$

so a higher ratio indicates that (x, y) is likelier a training member.

Online vs. Offline LiRA. In the online version, each target point is explicitly included in half of the shadow models. By contrast, the offline variant omits separate IN models, using only OUT statistics. Formally:

$$\Lambda \;=\; 1 - \Pr\big[\, Z > \phi\big(f(x)_y\big)\big], \quad Z \sim \mathcal{N}(\mu_{\text{out}}, \sigma^2_{\text{out}}). \tag{3}$$

While more efficient, this offline approach is slightly less accurate.

Evaluation at Low FPR. Crucially, the authors measure true-positive rates at very low false-positive rates (e.g., 0.1%), rather than only aggregate metrics like AUC. Even a few high-confidence member identifications constitute a major privacy leak. Empirically, LiRA yields up to an order-of-magnitude improvement over prior methods in this low-FPR regime.

2 LOW-COST HIGH-POWER MEMBERSHIP INFERENCE ATTACKS

Zarifzadeh et al. (2024) introduce the *Robust Membership Inference Attacks (RMIA)* by framing the hypothesis test as many pairwise likelihood-ratio comparisons. For a data point x, it estimates $\Pr(\theta|x)$ versus $\Pr(\theta|z)$ for random population points z. The membership score is then the fraction of $z \in \pi$ such that x "dominates" z:

$$\text{ScoreMIA}(x;\theta) \;=\; \Pr_{z\sim\pi}\big[\,\text{LR}_\theta(x,z) \;\geq\; \gamma\big] \;=\; \frac{1}{|Z|}\sum_{z\in Z} 1\big(\,\text{LR}_\theta(x,z) \;\geq\; \gamma\big),$$

where $\text{LR}_\theta(x,z) \;=\; \dfrac{\Pr(x|\theta)}{\Pr(x)} \Big/ \dfrac{\Pr(z|\theta)}{\Pr(z)}$.

Computation of $\Pr(x)$. By Bayes' rule, $\Pr(x)$ can be interpreted as an integral over all possible models θ' that could be trained from the population distribution π. In practice, RMIA approximates this integral by training reference models θ' on subsets of the population data. One averages $\Pr(x|\theta')$ over these reference models. Half of these models explicitly include x in training (IN-models) while half exclude x (OUT-models). This yields an unbiased estimate of $\Pr(x)$.

Online vs. Offline RMIA. In the offline mode, one uses only OUT models and a linear correction to boost $\Pr(x|\text{IN})$. Concretely, if $\Pr(x)_{\text{OUT}}$ is the average over OUT models, one sets

$$\Pr(x) \;\approx\; \tfrac{1}{2}\Big[(1+a)\;\Pr(x)_{\text{OUT}} \;+\; (1-a)\Big], \tag{4}$$

where $a \in [0,1]$ is a small parameter that adjusts how aggressively to shift the OUT estimate upward. This allows RMIA to operate with very few reference models in an offline setting while retaining high power.

3 CONCEPTUAL COMPARISON

Both LiRA and RMIA interpret membership inference as a likelihood-ratio test, but they differ in their core statistical modeling and the use of reference data:

- **Likelihood-ratio interpretation.** LiRA and RMIA both estimate the probability of the observed model parameters given "IN" vs. "OUT" hypotheses. LiRA does so by fitting separate distributions of logit (or loss) values for IN and OUT shadow models, whereas RMIA directly forms pairwise likelihood ratios of (x, z).
- **Reference Models and Data.**
  - LiRA trains two sets of shadow models on random splits of population data: one set that includes x, one that excludes x. These produce a Gaussian approximation of logits for each scenario.
  - RMIA also trains reference (shadow) models but uses them to compute $\Pr(x)$ by averaging $\Pr(x|\theta')$. Crucially, it then compares x against a population set of data points z to perform a pairwise test.
- **Offline vs. Online Modes.** In both attacks, an online mode requires training new "IN" models for each query x, yielding higher accuracy but large compute costs. By contrast, their offline modes rely on only "OUT" shadow models:
  - Offline LiRA removes the IN shadow set entirely and approximates the likelihood ratio from the OUT distribution alone.
  - Offline RMIA relies on linear scaling of $\Pr(x)_{\text{OUT}}$ to compensate for lacking IN models, then defines membership as the fraction of z that x dominates in the ratio test.
- **Sensitivity to Model Count.** LiRA generally benefits from numerous shadow models to stabilize the mean/variance estimates. RMIA remains effective with fewer reference models because it leverages many z points in the final pairwise test, gaining a robust sample of comparisons.

Overall, both methods provide powerful membership tests. LiRA frames the attack around statistical fitting of IN vs. OUT logit distributions, while RMIA focuses on per-data-point pairwise dominance and a more granular exploitation of population data.

REFERENCES

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

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *Forty-first International Conference on Machine Learning*, 2024.