Poster: Ultimate Power of Inference Attacks: Privacy Risks of Learning High-Dimensional Graphical Models

Sasi Kumar Murakonda
National University of Singapore
murakond@comp.nus.edu.sg

Reza Shokri
National University of Singapore
reza@comp.nus.edu.sg

George Theodorakopoulos
Cardiff University
TheodorakopoulosG@cardiff.ac.uk

I. Introduction

Models leak information about their training data. How much is the privacy risk of releasing such models which are trained on sensitive data? We focus on measuring information leakage from models about their training data, using tracing (membership inference) attacks. Given the released model and a target data sample, the adversary aims at inferring whether or not the target sample was a member of the training set. We use the term tracing attack and membership inference attack interchangeably [1], [2]. The attack is evaluated based on the power (true positive rate) and error (false positive rate), in its binary decisional task.

Tracing attacks have been extensively studied for summary statistics, where independent statistics (e.g., mean) of attributes of high-dimensional data are released. Although initial works showed the existence of powerful tracing attacks [3], more recent work provided theoretical frameworks to analyze the upper bound on the power of these inference attacks [4], and their robustness to noisy statistics [1]. Advanced machine learning models, such as deep neural networks, have recently been tested against tracing attacks [2]. However, their analysis is limited to empirical measurements of the attack success on particular data sets.

Contributions: Using the above-mentioned existing methods, it is possible to reason theoretically about tracing attacks, yet only for extremely simple models (independent statistics). In parallel, it is possible to perform empirical tracing attacks against complex models (deep neural networks), yet without much theoretical analysis on the maximum power of such attacks. In this paper, we aim at bridging this gap by providing a theoretical bound on the performance of tracing attacks against high dimensional graphical models, i.e. graphical models with many parameters, with focus on **Bayesian Networks**.

We use the likelihood ratio test (LR test) as the foundation of our tracing attack [4]. This enables us to **design the most powerful attack** against any probabilistic model. Thus, for any given error, there exists no other attack strategy that can achieve a higher power.

Our objective is to identify the elements of a model that cause membership information leakage, and measure their influence. We prove that, for a given model structure, the potential leakage of the model (the leakage that corresponds to the most powerful attack for any given error) is proportional to the square root of model's complexity (defined as the number of its independent parameters), and is inversely proportional to the square root of size of training set. Thus, the theoretical bound enables us to **quantify the potential leakage of a model before even learning the parameters of the model** on that structure.

II. Problem Statement

We consider a set of n independent m-dimensional data samples from a *population*. We refer to this set as the *pool*. Given a graphical model structure G, the pool data is used to train a graphical model, i.e., to estimate the parameters $\hat{\theta}$ of the probabilistic graphical **model** $\langle G, \hat{\theta} \rangle$. This model is *released*. Our objective is to quantify the privacy risks of releasing such models for the members of their training data.

Let us consider an adversary who observes the released model $\langle G, \hat{\theta} \rangle$. The objective of the adversary is to perform a **tracing attack** (also known as the membership inference attack) against the released model, on any target data point x: create a decision rule that determines whether x was used in the training of the parameters of $\langle G, \hat{\theta} \rangle$ or not, i.e. to classify x as being in the pool (IN) or not (OUT).

The accuracy of the tracing attack indicates the information leakage of the model about the members of its training set. We quantify the attacker's success using two evaluation metrics: the adversary's **power** (the true positive rate), and his **error** (the false positive rate). The power measures the conditional probability that the attacker classifies x as IN, given that x is indeed in the pool, i.e. $\Pr[IN|x \in \text{pool}]$. The error measures the conditional probability that the attacker classifies x as IN, given that x is not in the pool, i.e. $\Pr[IN|x \notin \text{pool}]$.

III. Framework for Attack Design

Given the released model and the target data point, the adversary aims at distinguishing between two hypothesis. Each hypothesis describes a possible world that could have resulted in the observation of the adversary, where in one world the target data is part of the training set (pool), while in the other one the target data is a random sample from the population.

- Null hypothesis ($\mathrm{H_{OUT}}$): The pool is constructed by drawing n independent samples from the general population. Parameters $\hat{\theta}$ of the model $\langle G, \hat{\theta}\rangle$ are trained on the pool data. Target data x is drawn from the general population, independently from the pool.
- Alternative hypothesis ($\mathrm{H_{IN}}$): The pool is constructed by drawing n independent samples from the general population. Parameters $\hat{\theta}$ of the model $\langle G, \hat{\theta}\rangle$ are trained on the pool data. Target data x is drawn from the pool.

We use the Likelihood Ratio test to distinguish the two hypothesis. It is worth emphasizing that according to the Neyman-Pearson lemma, the LR test achieves the **maximum power** among all decision rules with a given error (false positive rate). The only information we know about the pool is $\hat{\theta}$, the parameters of the released model learned using the pool data. Hence, we must calculate these *exact same parameters* under null hypothesis (i.e., learn the parameters using general population). Let θ be the result of this computation, i.e., the parameters of G trained on a large reference population.

We calculate L_{IN} as the likelihood of the parameters of G taking the value $\hat{\theta}$, which is equal to $\Pr[x; \langle G, \hat{\theta}\rangle]$. Similarly, we calculate L_{OUT} as the likelihood of the parameters of G taking the value θ, which is equal to $\Pr[x; \langle G, \theta\rangle]$. Hence, the log likelihood ratio statistic is computed as follows.

$$L(x) = \log\left(\frac{\Pr[x; \mathrm{H_{OUT}}]}{\Pr[x; \mathrm{H_{IN}}]}\right) = \log\left(\frac{\Pr[x; \langle G, \theta\rangle]}{\Pr[x; \langle G, \hat{\theta}\rangle]}\right) \quad (1)$$

The LR test is a comparison of the log likelihood ratio statistic $L(x)$ with a threshold. If $L(x) \leq$ threshold, then the attacker decides in favor of $\mathrm{H_{IN}}$ (rejects $\mathrm{H_{OUT}}$); otherwise, in favor of $\mathrm{H_{OUT}}$ (more precisely, in this case, he fails to reject $\mathrm{H_{OUT}}$ because there is not enough evidence to support this rejection in favor of $\mathrm{H_{IN}}$).

IV. Assumptions for theoretical analysis

To derive our main result about the best achievable power-error tradeoff, we assume that the released parameters satisfy the below conditions.

- The value of every released parameter is learned from a large enough number of samples for the central limit theorem to hold good.
- The value of every released parameter is non-trivial i.e., it is bounded away from 0 and 1 [4].

These are valid assumptions to make on part of the model publisher. In fact, the recently published methodology of learning Bayesian Networks on Cancer Analysis System (CAS) database in the National Cancer Registration and Analysis Service (NCRAS) satisfies both the assumptions (they use only the parameters that are learned using at least 50 samples) [5].

V. Main Result: Power and Error tradeoff

Our objective is to compute the maximum power β for any false positive error α of an adversary that observes the released model $\langle G, \hat{\theta}\rangle$ which has been trained on a pool of size n. In our main result, Theorem 1, we show which combinations of α and β are possible for the attacker.

Theorem 1. *Let β and α be the power and error of the LR test, for the membership inference attack, respectively. Let n be the size of the pool (model's training set), and $C(G)$ be the complexity of the released probabilistic graphical model $\langle G, \hat{\theta}\rangle$ defined as the number of independent parameters in $\hat{\theta}$. Then, the tradeoff between power and error follows the the following relation:*

$$z_\alpha + z_{1-\beta} \approx \sqrt{\frac{C(G)}{n}}, \quad (2)$$

where z_s is the quantile at level $1 - s, 0 < s < 1$ of the Standard Normal distribution.

Proof sketch: To compute β, the power of the LR test for the inference attack, for any error α, we need the distribution of $L(x)$ when x is drawn from the pool and when x is drawn from the population. Estimating the exact distribution of $L(x)$ is a hard problem. Our approach is to approximate the distributions of $L(x)$, through computing its moments $\mathrm{E}(L^k), k > 0$. To approximate the distribution using its moments, we use an established statistical principle for fitting a distribution with known moments: the maximum-entropy principle. This principle states that the probability distribution which best represents the current state of knowledge is the one with largest entropy.

■

VI. Insights from Theorem 1

Theorem 1 shows that releasing more parameters helps the attacker, and the amount of improvement depends on how large the sum already is and there are diminishing returns. In contrast, increasing the pool size n has the opposite effect to increasing C: the attack performance becomes worse. Our result generalizes that of Sankararaman et al. [4] on releasing independent marginals. The result also implies that the amount of leakage from a graphical model is decided by the number of independent parameters in the model. While learning from the training set, estimation error of a parameter provides the power for membership inference and in graphical models these estimation errors are independent for each parameter. Hence the information leakage about training set can be quantified by the number of parameters in the graphical model.

References

[1] C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan, "Robust traceability from trace amounts," in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, 2015, pp. 650–669.

[2] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017, pp. 3–18.

[3] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays," *PLoS genetics*, vol. 4, no. 8, p. e1000167, 2008.

[4] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic privacy and limits of individual detection in a pool," *Nature genetics*, vol. 41, no. 9, p. 965, 2009.

[5] "Generating the simulacrum a methodology overview," https://simulacrum.healthdatainsight.org.uk/wp/wp-content/uploads/2018/11/Methodology-Overview-Nov18.pdf.