Draft Concept for Topic 2.2.5: Membership Inference on Synthetic Data

Summer 2025: 194.055 - Security, Privacy, and Explainability in ML

***Group 9**: Marko Georgiev - 12442103; Thomas Brezina - 11778916* ***Mentor**: Prof. Rudolf Mayer*

Topic Description

We begin with a description of our topic, *Topic 2.2.5: Membership Inference on Synthetic Data*. The goal is to determine whether a known record was part of the real training data used to generate a synthetic dataset. We focus on the no-box attack variant, meaning we only have access to the synthetic data and possibly some auxiliary knowledge. However, we cannot query or retrain the generator. Our project will implement and evaluate two no-box MIA techniques on synthetic datasets. We will assess the effectiveness of each method using Attack Success Rate and AUC, as well as its efficiency through runtime performance.

Project Setup

After researching and reviewing relevant literature, we outline the synthetic data generators, dataset choice, and evaluation strategy that we will rely on.

Synthetic Data Generators

We plan to use CTGAN (Conditional Tabular GAN), a widely used GAN-based generator available through SDV or SynthCity. For comparison, we will also use either TVAE (Tabular VAE) or a Gaussian Copula model, both of which are non-GAN and easy to train. Both generators will be trained on the same real dataset for consistency. Using two different generators will enable us to determine whether certain attacks are more successful against one model type than another (e.g., a deep generative model versus a simpler statistical model). Both chosen generators will be used on the same real dataset for fairness.

Datasets

Since the focus is on attack methodology, the choice of dataset is secondary. We will use a public tabular dataset with mixed attributes, such as the UCI Adult Income dataset or a medical dataset. The data will be split into training and hold-out sets for training the generator and for non-member instances, respectively.

Evaluation Metrics

As suggested, we use three metrics to assess our attacks. Attack Success Rate measures the accuracy with which the method identifies membership, using indicators such as accuracy, precision, or recall at a specified threshold. A correct "positive" means the attack correctly flags a true member. The AUC (Area Under the ROC Curve) evaluates how well the attack distinguishes between members and non-members across all thresholds, making it ideal for binary classification tasks like ours. A higher AUC implies stronger separation ability. Runtime reflects efficiency, captured as total execution time or time per 1,000 membership queries, offering insight into the attack's scalability; however, this might change depending on the project flow.

No-Box MIA Approaches

In this section, we outline the most interesting papers on which we base our project. Given the abundance of research on Membership Inference Attacks (MIAs), we found that the treatment of auxiliary information is a key differentiating factor among existing

approaches. For this reason, we focus on three papers that provide contrasting perspectives in that regard.

GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models

This work presents a comprehensive taxonomy of MIAs and introduces the concept of full black-box attacks, which assume minimal knowledge on the part of the attacker [1]. Later papers criticize this approach as naive for overlooking correlations and relying on unrealistic assumptions. Relevant sections include:

Membership Inference Attacks against Synthetic Data through Overfitting Detection

This paper questions the assumptions made by GAN-Leaks and similar works [1], [2], particularly their disregard for auxiliary information. It emphasizes that ignoring such data leads to overly simplistic threat models [3].

Synthetic Is All You Need: Removing the Auxiliary Data Assumption for Membership Inference Attacks Against Synthetic Data

Interestingly, this paper turns the criticism of the previous one around by demonstrating that auxiliary data is not strictly necessary for effective machine learning applications in MIAs [4]. This presents a compelling contrast to [3], challenging the notion that real data or external knowledge is required.

Project Objective

Lastly, our goal is to implement and compare two no-box MIAs that differ in their reliance on auxiliary information auxiliary data affects attack success and efficiency. Each attack will be tested on two synthetic datasets, generated by two models (a deep CTGAN and a simpler model, e.g., a Gaussian Copula) from a single initial dataset. This would result in four complete evaluations in total, where we report the Attack Success Rate, AUC, and runtimes for each. The objective is to provide clear, empirical insights into how auxiliary knowledge influences the effectiveness of MIAs on synthetic tabular data. Depending on the results, we may also consider using multiple initial datasets, resulting in additional evaluations.

References

[1] D. Chen, N. Yu, Y. Zhang, and M. Fritz, “GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models,” in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event USA: ACM, Oct. 2020, pp. 343–362. doi: 10.1145/3372297.3417238.

[2] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, “LOGAN: Membership Inference Attacks Against Generative Models,” Aug. 21, 2018, *arXiv*: arXiv:1705.07663. doi: 10.48550/arXiv.1705.07663.

[3] B. van Breugel, H. Sun, and Z. Qian, “Membership Inference Attacks against Synthetic Data through Overfitting Detection”.

[4] F. Guépin, M. Meeus, A.-M. Cretu, and Y.-A. de Montjoye, “Synthetic is all you need: removing the auxiliary data assumption for membership inference attacks against synthetic data,” Sep. 21, 2023, *arXiv*: arXiv:2307.01701. doi: 10.48550/arXiv.2307.01701.