Security and Safety, Vol. 4, 2024017 (2025)
https://doi.org/10.1051/sands/2024017
sands.edpsciences.org

Security and Safety of Data in Cloud Computing

Commentary

Open Access

Other Fields

Advancing membership inference attacks: The present and the future

Zheng Li and Yang Zhang*

CISPA Helmholtz Center for Information Security, Saarbruken 66123, Germany

Received: 30 July 2024 / Revised: 21 October 2024 / Accepted: 24 October 2024 / Published online: 30 January 2025

Citation Li Z and Zhang Y. Advancing membership inference attacks: The present and the future. Security and Safety 2025; **4**: 2024017. https://doi.org/10.1051/sands/2024017

1 Introduction

Machine Learning (ML) has made significant progress in various tasks. This success is driven by powerful computational resources, expert experience, and, crucially, large-scale data. High-performance ML models typically require millions or even billions of samples. Training on such extensive datasets allows algorithms to capture complex patterns and improve generalization. Thus, large-scale data is essential for developing advanced ML models, driving the impressive advancements and implementations in the field today.

However, large-scale datasets often contain sensitive information about individuals, such as health status, bank details, and shopping preferences, raising significant privacy risks. One major privacy threat is membership inference attacks (MIAs) [1–7], where an adversary determines whether a specific data sample was used to train an ML model. This can lead to serious data leaks; for example, if an ML model is trained on data from individuals with a certain disease, an adversary can infer a victim's health status if they know the victim's data is in the training set. Given the importance of data privacy and widespread ML applications, MIAs have gained significant attention in academia and industry. This paper reviews the current state of MIAs, focusing on their principles, threat models, and methodologies. We also discuss the limitations and challenges in current research and propose future directions to enhance ML model robustness and privacy.

2 Current status of MIAs

2.1 Fundamentals

Membership leakage in ML models occurs when an adversary determines if a specific data sample was used to train the model. Formally, given a data sample x, a trained model $\mathcal{M}$, and the adversary's knowledge Ω, a membership inference $\mathcal{A}$ is defined as: $\mathcal{A} : x, \mathcal{M}, \Omega \rightarrow \{0, 1\}$ Here, 0 indicates that x is not a member of $\mathcal{M}$'s training set, and 1 indicates that it is. The attack model $\mathcal{A}$ is a binary classifier. MIAs exploit differences in model behavior on training data versus unseen data, as models typically perform better on training data. An adversary can use this performance discrepancy to infer the membership status of specific data points. Depending on the threat model, MIAs can be constructed in various ways, which will be discussed in later sections.

2.2 Threat models

The effectiveness of MIAs varies significantly based on the attacker's knowledge and capabilities. We discuss different threat models:

* Corresponding author (email: zhang@cispa.de)

White-box Scenario. The attacker has full access to the model's parameters, architecture, and possibly the training algorithm. This allows detailed analysis of the model's workings, making it easier to identify discrepancies between training and unseen data. White-box attacks are the most powerful due to the extensive information available.

Score-box Scenario. The attacker can query the model and observe outputs, such as confidence scores, without accessing internal parameters or architecture. By analyzing these outputs, the attacker can infer if a data sample was part of the training set. This relies on the model producing higher confidence scores for training data.

Label-only Scenario. The attacker can only query the model and see the final predicted labels, without access to confidence scores or other detailed outputs. Despite limited information, attackers can use statistical techniques on the predicted labels to detect patterns indicative of training data.

2.3 Methodologies

Currently, there are three representative attacks for different scenarios: white-box attacks, score-based attacks, and label-only attacks.

White-box Attacks. In white-box attacks [8, 9], the attacker obtains the prediction score and intermediate computations (features) of a data sample on the target model. Since there are no criteria or rules for distinguishing between members and non-members of the target model, the most common approach is for the attacker to train the attack model using its local model and dataset (called shadow model and shadow dataset). Concretely, the attacker splits a shadow dataset into training and testing sets. By querying the shadow model with training samples, the attacker computes prediction scores and features, concatenates them into a vector, and labels it as a member if the sample is from the training set, otherwise as a non-member. This creates an attack training dataset, which is used to train a binary classifier. The attacker then queries the target model to differentiate members from non-members.

Score-based Attacks. Score-based attacks [1, 2, 5, 10–13] also require training a shadow model. Unlike white-box attacks, these rely solely on output scores from the target model rather than intermediate features. The attacker queries the shadow model with the shadow training dataset (members) and shadow test dataset (non-members) to create an attack training dataset. This dataset is used to construct an attack model. Specifically, LiRA [12] trains N reference models and then measures the likelihood of the target sample's loss and determines membership. TrajectoryMIA [5] and SeqMIA [13] are state-of-the-art attacks that exploit membership information leaked from the training process of the target model.

Label-only Attacks. In label-only attacks [4, 14], the target model reveals only predicted labels without intermediate features or output scores. These attacks depend on using predicted labels for their inputs. The attacker trains a shadow model and queries it with data samples, observing changes in predicted labels caused by perturbations. Membership is determined based on whether the perturbations exceed a predetermined threshold. Recently, YOQO [15] is an advanced attack method that does not use perturbations and focuses on reducing the number of queries to the target model.

3 Limitations and challenges

3.1 Technical challenges

This section delves into the limitations and challenges, highlighting technical challenges, defense mechanisms, and practical considerations.

3.2 Technical challenges

Model-Specific Vulnerabilities. Different models have varying susceptibility to MIAs. For instance, image classifiers often overfit training data, making them more prone to MIAs. However, their complex architecture can introduce noise that complicates the attack. Similarly, natural language processing models can also memorize training data, especially when using small datasets, leading to higher vulnerability. Many emerging models need to adopt specific attack methods.

Scalability Issues. As datasets and models grow in size and complexity, both attacks and defenses face scalability challenges. Attacking models trained on large datasets requires extensive computational resources

and sophisticated techniques. Models with numerous parameters and layers, such as deep neural networks, pose difficulties for both attack accuracy and defense implementation due to their intricate behavior and high-dimensional space.

3.3 Defense mechanisms and their limitations

Current representative defenses against MIAs include:
Differential Privacy. Differential privacy (DP) [16–18] is a robust defense providing formal privacy guarantees. However, it has challenges: (1) Adding noise to data or model parameters can decrease accuracy and utility. (2) Setting the privacy budget ϵ is crucial: a lower ϵ improves privacy but reduces utility, while a higher ϵ compromises privacy.

Regularization Techniques. Regularization methods [2, 19] such as dropout, L2 regularization, and early stopping can help mitigate overfitting, thereby reducing susceptibility to MIAs. However, these techniques also have their limitations: (1) Regularization can lead to underfitting if applied too aggressively, thereby reducing the model's performance on legitimate tasks. (2) Tuning regularization parameters requires careful experimentation and validation, which can be resource-intensive.

Adversarial Training. Adversarial training [20, 21], where models are trained with adversarial examples to improve robustness, can also offer some protection against MIAs. However, (1) Adversarial training requires generating adversarial examples and iterating through multiple training cycles, which is computationally expensive. (2) While effective against certain types of attacks, adversarial training may not generalize well to all possible membership inference scenarios.

Output Perturbation. Adding calibrated noise directly to the model's predictions, especially in black-box settings, can confuse attackers by masking the true confidence levels that are typically used to infer membership status.

3.4 Practical challenges in real-world applications

Legal and Ethical Considerations. The implementation of defenses against MIAs must also consider legal and ethical implications: (1) Compliance with data privacy laws like GDPR and CCPA requires careful data handling and explicit user consent, complicating defense implementation. (2) Balancing model performance and privacy can create ethical dilemmas, especially in critical areas like healthcare, where accuracy is crucial.

Balancing Usability and Security. Striking a balance between maintaining model usability and ensuring security against MIAs is a persistent challenge: (1) Defenses that reduce model performance can harm user experience, causing resistance from stakeholders. (2) Robust defenses may increase costs in terms of computational resources and manpower, which must be justified by the level of risk mitigation.

4 Future directions

This section outlines promising future directions to enhance the understanding, detection, and mitigation of MIAs.

4.1 Innovative attack techniques

Exploring New Attack Vectors. Future research should explore new attack vectors that target various aspects of machine learning models. For instance, recent work on the Pattern of Metric Sequence [13] has shown that it can more effectively capture differences between members and non-members.

Leveraging Auxiliary Information. Incorporating auxiliary information to improve attack accuracy is a promising direction: (1) Utilizing side-channel information such as model update patterns, timing information, or system-level data to infer membership status. (2) Combining attack strategies with external data sources or knowledge bases to increase the confidence and precision of inferences.

4.2 Enhanced defense strategies

Robust Privacy-Preserving Techniques. Developing advanced privacy-preserving techniques that offer strong protections without significantly compromising model performance is critical: (1) Refining differential privacy algorithms to reduce the trade-off between privacy and accuracy, possibly through adaptive noise addition or context-aware privacy budgets. (2) Combining multiple defense strategies, such as differential privacy with adversarial training or regularization, to provide layered security.

Adaptive Defense Mechanisms. Creating adaptive defenses that can dynamically respond to evolving threats is an essential direction: (1) Developing mechanisms that automatically adjust defense parameters based on real-time monitoring of model performance and detected threats. (2) Implementing systems that analyze model behavior to detect anomalies, such as abnormal query patterns or output distributions, which could indicate the presence of a membership inference attack.

Author contribution statement

The initial idea for this paper was proposed by Dr. Yang Zhang, who then discussed the logical structure with Dr. Zheng Li. Both authors contributed to the writing and reviewing of the paper.

References

[1] Shokri R, Stronati M and Song C et al. Membership inference attacks against machine learning models. In: IEEE Symposium on Security and Privacy (S&P), IEEE, 2017, 3–18.
[2] Salem A, Zhang Y and Humbert M et al. ML-leaks: model and data independent membership inference attacks and defenses on machine learning models. In: Network and Distributed System Security Symposium (NDSS), Internet Society, 2019.
[3] He X, Wen R and Wu Y et al. Node-level membership inference attacks against graph neural networks. CoRR abs/2102.05429, 2021.
[4] Li Z and Zhang Y. Membership leakage in label-only exposures. In: ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2021,880–895.
[5] Liu Y, Zhao Z and Backes M et al. Membership inference attacks by exploiting loss trajectory. In: ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2022, 2085–2098.
[6] He X, Li Z and Xu W et al. Membership-doctor: comprehensive assessment of membership inference against machine learning models. CoRR abs/2208.10445, 2022.
[7] Wu Y, Yu N and Li Z et al. Membership inference attacks against text-to-image generation models. CoRR abs/2210.00968, 2022.
[8] Nasr M, Shokri R and Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In: IEEE Symposium on Security and Privacy (S&P), IEEE, 2019, 1021–1035.
[9] Leino K and Fredrikson M. Stolen memories: leveraging model memorization for calibrated white-box membership inference. In: USENIX Security Symposium (USENIX Security), USENIX, 2020, 1605–1622.
[10] Li J, Li N and Ribeiro B. Membership inference attacks and defenses in classification models. In: ACM Conference on Data and Application Security and Privacy (CODASPY), ACM, 2021, 5–16.
[11] Song L and Mittal P. Systematic evaluation of privacy risks of machine learning models. In: USENIX Security Symposium (USENIX Security). USENIX, 2021.
[12] Carlini N, Chien S and Nasr M et al. Membership inference attacks from first principles. In: IEEE Symposium on Security and Privacy (S&P), IEEE, 2022, 1897–1914.
[13] Li H, Li Z and Wu S et al. SeqMIA: sequential-metric based membership inference attack. CoRR abs/2407.15098, 2024.
[14] Choquette Choo CA, Tramèr F and Carlini N et al. Label-only membership inference attacks. In: International Conference on Machine Learning (ICML), PMLR, 2021, 1964–1974.
[15] Wu Y, Qiu H and Guo S et al. You only query once: an efficient label-only membership inference attack. In: The Twelfth International Conference on Learning Representations, 2024.
[16] Chaudhuri K, Monteleoni C and Sarwate AD. Differentially private empirical risk minimization. J Mach Learn Res, 2011, 1069–1109.
[17] Dwork C, McSherry F and Nissim K et al. Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography Conference (TCC), Springer, 2006, 265–284.
[18] Iyengar R, Near JP and Song DX et al. Towards practical differentially private convex optimization. In: IEEE Symposium on Security and Privacy (S&P), IEEE, 2019, 299–316.
[19] Truex S, Liu L and Emre Gursoy M et al. Towards demystifying membership inference attacks. CoRR abs/1807.09173, 2018.
[20] Nasr M, Shokri R and Houmansadr A. Machine learning with membership privacy using adversarial regularization. In: ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2018, 634–646.
[21] Jia J, Salem A and Backes M et al. MemGuard: defending against black-box membership inference attacks via adversarial examples. In: ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2019, 259–274.