GAN-Leaks: A Taxonomy of Membership Inference Attacks against GANs

Dingfan Chen[1] **Ning Yu**[2,3] **Yang Zhang**[1] **Mario Fritz**[1]
[1] CISPA Helmholtz Center for Information Security, Germany
[2] Max Planck Institute for Informatics, Germany
[3] University of Maryland, College Park
{dingfan.chen, zhang, fritz}@cispa.saarland ningyu@mpi-inf.mpg.de

Abstract

In recent years, deep learning has achieved overwhelming success, spanning from discriminative models to generative models. In particular, generative adversarial networks (GANs) have facilitated a new level of performance in a myriad of areas, ranging from media manipulation to sanitized dataset generation. Despite the great success, the potential risks of privacy breach caused by GANs have not been analyzed systematically. In this paper, we focus on membership inference attack against GANs that reveals information about the training data used for victim models. Specifically, we present the first taxonomy of membership inference attacks, comprising not only existing attacks but also our novel ones. In addition, we propose the first generic attack model that can be instantiated in a large range of settings according to the adversary's knowledge about the victim models. We complement the systematic analysis of attack performance by a comprehensive experimental study, that investigates the effectiveness of various attacks w.r.t. model type and training configurations, over three diverse application scenarios (i.e., images, medical data, and location data).

1 Introduction

Over the past few years, generative adversarial networks (GANs) [11, 24, 26, 2, 12, 17, 6] have achieve breakthroughs in improving the realism of generated data and approximation towards training data distribution. Many IT companies and research institutes publish the pre-trained GAN models or provide users with platforms where users can apply the learned models. However, with the prevalence of publicly accessible pre-trained models and online deep learning APIs, data privacy is challenged by malicious users who intend to infer the original training data. The resulting privacy breach would raise serious issues because training data always contains sensitive attributes (e.g., a patient's disease history). One such attack is membership inference [10, 3, 27, 13, 25] which aims to identify if a query data record was used to train a deep learning model.

There are two main motivations for conducting research in membership inference attack. The first motivation is to validate and quantify the privacy vulnerability of a deep learning model. The second motivation is to establish wrongdoing, where, e.g., regulators can be clued from membership inference to propose the suspicion that a model was trained on personal data without an adequate legal basis.

Membership inference against discriminative deep learning models has been largely explored [3, 27, 1, 15, 20, 28, 7, 21], while inference against generative models is still an open question. This is more challenging from the adversary side because the victim model does not directly provide confidence values about the overfitting of data records. Recently, Hayes et al. [13] present a first approach targeting GANs, which proposes to retrain a local copy of the victim GAN in a black-box setting and to check the discriminator's confidence for membership inference in a white-box setting. Their intuition is that the overfitting of a victim GAN is either incorporated in its discriminator or

can be mimicked from a local copy of the discriminator. Hilprecht et al. [14] extend membership inference attack to both GANs and VAE via Monte Carlo integration [22]. Their intuition is that an overfitted generator tends to output data samples closer to the training data than to unseen data.

However, neither of them provides a complete, systematic, and practical analysis of membership inference attacks against GANs. For example, Hayes et al. [13] does not consider the realistic situation where the discriminator is not accessible but only the generator is deployed for query. Hilprecht et al. [14] investigate only on small-scale image datasets and does not involve white-box attack against GANs. That motivates our contributions along this direction towards a more systematic analysis:

Taxonomy of membership inference attacks against GANs. We propose a pioneering study to categorize attack settings against GANs. Given **decreasing** order of the amount of knowledge about victim GAN accessible to the adversary, the settings are benchmarked as (1) accessible discriminator, (2) white-box generator, (3) partial black-box generator, and (4) full black-box generator. In particular, two of the settings, the partial black-box and white-box settings, are newly identified for attack model design. We then establish the first taxonomy for the existing and our proposed attacks.

The first generic attack model across settings and its novel instantiated variants. We propose the first generic attack model applicable to all the settings. The instantiated attack variants in the partial black-box and white-box settings are also the first attempts. Their consistent effectiveness bridges the assumption gap and performance gap between the existing full black-box attacks in [13, 14] and discriminator-accessible attack in [13] through a complete performance spectrum.

2 Taxonomy of membership inference attack against GANs

Specifically to the attack against GANs, we distinguish the settings based on the following criteria: (1) whether the discriminator is accessible or not, (2) whether the generator is white-box or black-box, and (3) whether the latent code of a generated sample is accessible or not. We categorize the existing and the proposed attacks in Table 1.

	Latent code	Generator	Discriminator
[13] full black-box	×	■	×
[14] full black-box	×	■	×
Our full black-box	×	■	×
Our partial black-box	✓	■	×
Our white-box	✓	□	×
[13] accessible discriminator	×	×	✓

Table 1: Taxonomy of attack settings against GANs over the previous work and ours. (×: without access; ✓: with access; ■: black-box; □: white-box). The settings are more and more knowledgeable to attackers from top to bottom.

Accessible discriminator. This is the most knowledgeable setting to attackers and it converts the attack against a GAN to the attack against a classical discriminative model, no matter whether the discriminator is white-box or black-box. Existing attack methods against discriminator models can be applied to this setting. For example, Shokri et al. [27] infer membership by checking the confidence value of the discriminator. This setting is also considered in [13], corresponding to the last row in Table 1. In practice, however, the discriminator of a well-trained GAN is usually discarded without being deployed to APIs, and thus not accessible to attackers. We, therefore, devote less effort to investigating the discriminator and mainly focus on the following practical settings where the attackers only have access to the generator.

White-box generator. This is the most knowledgeable setting to attackers when the discriminator of a GAN is no longer accessible. Attackers have access to the parameters of the generator. This is a realistic open-source scenario where users publish their well-trained generator without releasing the underlying training data. This scenario is also commonly studied in the differential privacy community [9]. However, this is a novel setting for membership inference attack against GANs, which is not explored in [13] or [14]. It corresponds to the second last row in Table 1.

Partial black-box generator (known input-output pair). This is a less knowledgeable setting to attackers where they have no access to the parameters of the generator but have access to each latent code of generated samples. This is also a realistic scenario where attackers submit their latent code as input and collect corresponding generated samples from the generator. This is another novel setting and not considered in [13] or [14]. It corresponds to the third last row in Table 1.

Full black-box generator (known output only). This is the least knowledgeable setting to attackers where they are unable to provide input but just blindly query samples from the well-trained black-box generator. This corresponds to the practical scenario of closed-source GAN-based APIs.

Hayes et al. [13] investigate attacks in this setting by retraining a local copy of the API. Hilprecht et al. [14] sample data from the generator and count the number of generated samples that are inside an ϵ-ball of the query, based on an elaborate design of distance metric. Our idea is similar in spirit to Hilprecht et al. [14] but we score each query by the reconstruction error directly, which does not introduce additional hyperparameter. In short, we design a low-skill attack method with a simpler implementation that achieves comparable or better performance. Our attack and theirs correspond to the first three rows in Table 1.

3 Membership inference attack against GANs

Generic attack model. We formulate the membership inference attack as a binary classification task where we threshold the reconstruction error between a query sample x and its reconstructed copy $\mathcal{R}(x|\mathcal{G}_v)$ from the well-trained victim generator $\mathcal{G}_v$. Our intuition is that, given access to a generator, we should reconstruct samples better if they belong to the GAN training set. The attacker predicts the query sample x to be in the training set if $L_{\text{cal}}\big(x, \mathcal{R}(x|\mathcal{G}_v)\big) < \tau$, where $L_{\text{cal}}(\cdot,\cdot)$ is a calibrated distance metric between two samples and τ is a pre-defined threshold.

The attacker's goal is then to design $\mathcal{R}$ such that it activates the most accurate possible performance of $\mathcal{G}_v$ to approximate a query sample. In the following sections, we instantiate variants of $\mathcal{R}$ for different attack settings.

Full black-box attack. We start with the least knowledgeable setting where an attacker only has access to a black-box generator $\mathcal{G}_v$. The attacker is allowed no other operation but blindly collecting k samples from $\mathcal{G}_v$, denoted as $\{\mathcal{G}_v(\cdot)_i\}_{i=1}^k$. $\mathcal{G}_v(\cdot)$ indicates that the attacker has no access or control over latent code input. We then define the reconstruction of x as the nearest neighbor from $\{\mathcal{G}_v(\cdot)_i\}_{i=1}^k$. Mathematically, $\mathcal{R}(x|\mathcal{G}_v) = \arg\min_{\hat{x}\in\{\mathcal{G}_v(\cdot)_i\}_{i=1}^k} L(x,\hat{x})$.

Partial black-box attack. In the partial black-box setting where attackers have access to the latent code input z of a query sample x, we propose to establish attack exploiting z. Concretely, the attacker performs an optimization w.r.t. z in order to accurately reconstruct the query samples x. Mathematically, $\mathcal{R}(x|\mathcal{G}_v) = \mathcal{G}_v(z^*)$, where $z^* = \arg\min_z L\big(x, \mathcal{G}_v(z)\big)$.

Without knowing the parameters of $\mathcal{G}_v$, the optimization is not differentiable. As only the evaluation of function (forward-pass through the generator) is allowed by the access of $\{z, \mathcal{G}_v(z)\}$ pair, we propose to approximate the optimum can be solved via the Powell's Conjugate Direction Method [23]. For the choice of initialization, we explore three different heuristics in our experiments, including mean ($z_0 = \mu$), random ($z_0 \sim \mathcal{N}(\mu, \Sigma)$), and nearest neighbour ($z_0 = \text{argmin}_{z\in\{z_i\}_{i=1}^k} \|\mathcal{G}_v(z) - x\|_2^2$). We find that the mean and nearest neighbor initialization perform well in practice. Therefore, we apply them both in parallel, and choose the one with smaller reconstruction error for the attack.

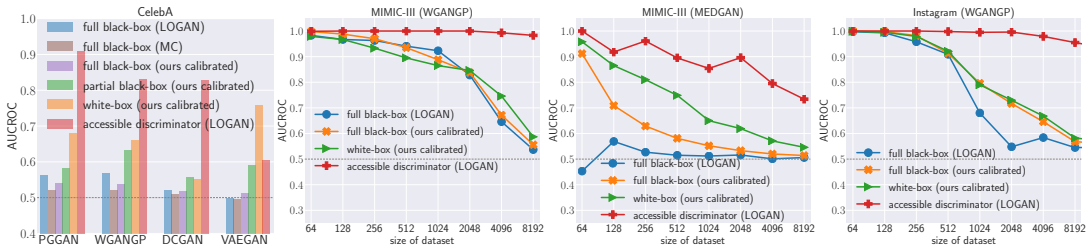
White-box attack. In the white-box setting, we have the same reconstruction formulation as in the partial black-box setting. But the reconstruction quality can be further boosted due to the access to the parameters of $\mathcal{G}_v$. This is because the optimization becomes differentiable with gradient backpropagation w.r.t. z, which can be more accurately solved by the L-BFGS [18] optimizer.

Distance metric. Our baseline distance metric $L(\cdot,\cdot)$ consists of three terms: the element-wise (pixel-wise) difference term L_2 targets low-frequency components, the deep image feature term L_{lpips} (i.e., the Learned Perceptual Image Patch Similarity (LPIPS) metric [29]) targets realism details, and the regularization term penalizes latent code far from the prior distribution. Mathematically,

$$L(x, \mathcal{G}_v(z)) = \lambda_1 L_2(x, \mathcal{G}_v(z)) + \lambda_2 L_{\text{lpips}}(x, \mathcal{G}_v(z)) + \lambda_3 L_{\text{reg}}(z)$$

where $L_2(x, \mathcal{G}_v(z)) = \|x - \mathcal{G}_v(z)\|_2^2$ and $L_{\text{reg}}(z) = \big(\|z\|_2^2 - \dim(z)\big)^2$. λ_1, λ_2 and λ_3 are used to enable/disable and balance the order of magnitude of each loss term. For non-image data, $\lambda_2 = 0$ because LPIPS is no longer applicable. For full black-box attack, $\lambda_3 = 0$ because z is not the variable to optimize. For the other cases, $\lambda_1 = 1.0$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.001$.

Attack calibration. We noticed that the reconstruction error is query-dependent, i.e., some query samples are more difficult to reconstruct due to their intrinsically more complicated representations. In this case, the reconstruction error is dominated by the representations rather than by the membership clues. Therefore, we propose to mitigate the query dependency by first training a reference GAN $\mathcal{G}_r$ with another disjoint dataset, and then calibrating our base reconstruction error according to the reference reconstruction error. Formally, $L_{\text{cal}}(x, \mathcal{G}_v(z)) = L(x, \mathcal{G}_v(z)) - L(x, \mathcal{G}_r(z))$. The optimization on the well-trained $\mathcal{G}_r$ is the same as on $\mathcal{G}_v$.



(a) (b) (c) (d)

Figure 1: Comparisons of attacks on various datasets, GANs, and w.r.t. GAN training set size.

4 Experiments

Datasets. We investigate on three modalities of datasets covering images (CelebA [19]), medical records (MIMIC-III [16]), and location check-ins (Instagram New-York [4]), which are all considered with a high risk of privacy breach.

Victim GAN models. We choose PGGAN[17], WGANGP [12], DCGAN [24], MEDGAN [8], and VAEGAN [5] into the victim model set, considering their pleasing performance on generating images and/or other data representations.

Attack evaluation. The proposed membership inference attack is formulated as a binary classification given a threshold τ. Through varying τ, we measure the area under the receiver operating characteristic curve (AUCROC) to evaluate the attack performance. With a value range of $[0, 1]$, higher value indicates better classification (attack) performance.

Comparison to baseline attacks. We compare our calibrated attack to two recent membership inference attack baselines: Hayes et al. [13] (denoted as LOGAN) and Hilprecht et al. [14] (denoted as MC, standing for their proposed Monte Carlo sampling method). LOGAN includes a full black-box attack model and a discriminator-accessible attack model against GANs. The latter is regarded as the most knowledgeable but unrealistic setting because the discriminator in GAN is usually not accessible in practice. MC only includes a full black-box attack model against GANs. Note that, to the best of our knowledge, there does not exist another attack against GANs in the partial black-box or white-box settings.

Figure 1 shows comparisons, considering several datasets, victim GAN models, and GAN training set sizes, and across different settings. Our findings are as follows. A more complete analysis is in the supplementary material.

In black-box setting, our low-skill attack consistently outperforms MC and outperforms LOGAN on the non-image datasets. It also achieves comparable performance to LOGAN on CelebA but with a much simpler implementation.

Our white-box and partial black-box attacks consistently outperform the other full black-box attacks, which indicates that the release of the generator or even just the input to the generator can lead to severe risk of privacy breach. With a complete spectrum of performance across settings, they bridge the performance gap between LOGAN black-box attack and LOGAN discriminator-accessible attack.

LOGAN with access to the discriminator is the most effective attack, except when against VAEGAN. The effectiveness can be explained by the fact that the discriminator is explicitly trained to maximize the margin between training set (membership samples) and generated set (a subset of non-membership samples), which eventually yields very accurate confidence scores for membership inference. As a consequence, it shows that releasing the discriminator results in exceptionally high risk of privacy breach. However, the discriminator score is not very effective against VAEGAN because its training relies more on sample-wise supervision than on the adversarial loss.

References

[1] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning: Revisited and enhanced. In *International Conference on Applications and Techniques in Information Security*, pages 100–110. Springer, 2017.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[3] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. Membership privacy in microrna-based studies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 319–330. ACM, 2016.

[4] Michael Backes, Mathias Humbert, Jun Pang, and Yang Zhang. walk2friends: Inferring social links from mobility profiles. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1943–1957. ACM, 2017.

[5] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. " best-of-many-samples" distribution matching. *arXiv preprint arXiv:1909.12598*, 2019.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[7] Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *arXiv preprint arXiv:1802.08232*, 2018.

[8] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. *arXiv preprint arXiv:1703.06490*, 2017.

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[10] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[13] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2019.

[14] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(4):232–249, 2019.

[15] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 603–618. ACM, 2017.

[16] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[18] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, December 2015.

[20] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

[21] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. *arXiv preprint arXiv:1805.04049*, 2018.

[22] Art B Owen. Monte carlo theory, methods and examples. *Monte Carlo Theory, Methods and Examples. Art Owen*, 2013.

[23] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 01 1964.

[24] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[25] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Proceedings of the 2019 Network and Distributed System Security Symposium*. Internet Society, 2019.

[26] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[27] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[28] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium*, pages 268–282. IEEE, 2018.

[29] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

Appendix

A more elaborated version of this work can be found in `https://arxiv.org/abs/1909.03935`.