J. Mantas et al. (Eds.)

doi:10.3233/SHTI240634

# Tunable Privacy Risk Evaluation of Generative Adversarial Networks

Bayrem KAABACHI* [a,1], Farah BRIKI* [a], Bogdan KULYNYCH [a], Jérémie DESPRAZ [a], Jean Louis RAISARO [a]

[a] *Biomedical Data Science Center, Lausanne University Hospital (CHUV) and University of Lausanne, Switzerland*

ORCiD ID: Bayrem Kaabachi https://orcid.org/0009-0002-7534-8493, Farah Briki https://orcid.org/0009-0003-0517-0121, Bogdan Kulynych https://orcid.org/0000-0001-5923-3931, Jean Louis Raisaro https://orcid.org/0000-0003-2052-6133

**Abstract.** Generative machine learning models such as Generative Adversarial Networks (GANs) have been shown to be especially successful in generating realistic synthetic data in image and tabular domains. However, it has been shown that such generative models, as well as the generated synthetic data, can reveal information contained in their privacy-sensitive training data, and therefore must be carefully evaluated before being used. The gold standard method through which such privacy leakage can be estimated is simulating membership inference attacks (MIAs), in which an attacker attempts to learn whether a given sample was part of the training data of a generative model. The state-of-the art MIAs against generative models, however, rely on strong assumptions (knowledge of the exact training dataset size), or require a lot of computational power (to retrain many "surrogate" generative models), which make them hard to use in practice. In this work, we propose a technique for evaluating privacy risks in GANs which exploits the outputs of the discriminator part of the standard GAN architecture. We evaluate our attacks in terms of performance in two synthetic image generation applications in radiology and ophthalmology, showing that our technique provides a more complete picture of the threats by performing worst-case privacy risk estimation and by identifying attacks with higher precision than the prior work.

**Keywords.** Synthetic Data, Privacy, Membership Inference Attacks

## 1. Introduction

The sensitive nature of healthcare data presents a critical barrier to biomedical research. Synthetic data generation is one of the possible technical solutions to this problem which promises to enable privacy-preserving data sharing by releasing synthetic data instead of the real records. When it comes to synthetic data generation, Generative Adversarial Networks [1] have been shown to be effective tools that can recreate realistic samples from image data (e.g., X-rays), or tabular data. The practical application of sharing syn-

[1] Corresponding Author: Bayrem Kaabachi, Rue du Bugnon 21, 1011 Lausanne, Switzerland. Email: mohamed-beyrem.kaabachi@chuv.ch. The authors marked with * have contributed equally.

thetic data in healthcare, however, is limited due to the difficulties in effectively measuring residual privacy risks as required by strict regulatory frameworks such as GDPR and HIPAA. Indeed, previous research indicates [2] that synthetic data is vulnerable to membership inference attacks (MIAs): attacks in which an adversary aims to determine whether a given record was part of the training dataset or not by observing the generative model or the synthetic data. For example, if we have published a synthetic dataset generated from a cohort of patients with a sensitive health condition, an adversary could run a MIA to determine if some targeted individuals were likely part of the training dataset. If successful, this could lead to the unintended disclosure of the sensitive health condition characterizing the dataset.

In this paper, we propose an efficient and flexible method to estimate such privacy risks by building on an existing method called LOGAN [2], a GAN-specific attack that also provides an efficient assessment of inference risks. LOGAN comes with two drawbacks: (1) it assumes knowledge of the model's training data size, and (2) allows neither to set the false-positive rate (FPR) of an attack in advance, nor to evaluate the receiver-operating characteristic (ROC) curve of the attack, which are the standard ways to assess the privacy risk in the machine learning privacy community [3, 4]. In addition, MIAs require a low FPR to be a realistic privacy threat [3] as a high FPR would mean the attack often incorrectly identifies "non-member" data as "member" data, thus producing unreliable results and leading to an overly conservative evaluation where the privacy risk of releasing a synthetic dataset is not significant.

We thus extend LOGAN by introducing a tunable method for privacy risk assessment. Our approach simulates various attack strategies at different levels of attack FPR, allowing for the computation of the attack's ROC curve as opposed to only a single number at an FPR that is unknown in advance. As such, our attack enables a data custodian to choose the most appropriate strategy for estimating the risk of membership inference based on the specific data sharing scenario.

Although there exists a growing body of work that studies MIAs on synthetic data and synthetic data generation, we focus on LOGAN as other methods can be extremely computationally expensive [4], or are effective only on tabular data due to the challenges of applying density estimation to high-dimensional data [5].

## 2. Method

*Threat Model* We perform our evaluation from the point of view of a data custodian who wants to release synthetic data. We aim to estimate privacy risk by simulating an attack where the adversary can query the discriminator of the target GAN model. Additionally, the attacker has access to a population $\mathscr{D}$ which closely mirrors the data distribution of the training dataset.

*Attack* A GAN consists of two subnetworks: a generator which generates a synthetic sample, and a discriminator which distinguishes between real and generated samples. We propose a membership inference attack based on the framework introduced by Ye et al. [6], which, like LOGAN [2], uses the discriminator output but leverages it in a different way. The discriminator's output is a useful feature to evaluate for MIAs, as it can "overfit" to inputs that were part of training data. Following Ye et al. [6], we define two possible worlds that any sample can belong to: the in world if it is a real sample from the training set, and out world if it is not. The attack runs a hypothesis test which aims to determine to which of the two worlds a given sample belongs:

$$H_0 : (\theta, z) \text{ where } z \sim S_{\mathsf{train}}, \quad H_1 : (\theta, z) \text{ where } z \sim \mathscr{D}, \tag{1}$$

with $S_{\mathsf{train}}$ denoting the training dataset, and $\theta = \mathscr{T}(S_{\mathsf{train}})$ the parameters of the trained model. The attacker queries the discriminator output on a sample $D_\theta(z)$ and chooses which hypothesis to reject if $\phi(\theta,z) \triangleq 1 - D_\theta(z) \leq c_\alpha(\theta)$, where $c_\alpha(\theta)$ is a threshold function that depends on the target model $\theta$ and the parameter $\alpha$, such that $0 \leq \alpha \leq 1$. Notably, $\alpha$ can be determined solely by examining the discriminator output on samples from the out world, by averaging over many $H_1$ (out world) samples and relying on an assumption that for the $H_0$ (in world) samples the values of $\phi(\theta,z) = 1 - D_\theta(z)$ are lower. Concretely, the attacker sets $c_\alpha(\theta)$ such that the rate at which non-members are incorrectly identified as members (false positives) equals the chosen FPR $\alpha$:

$$\Pr_{z \sim \mathscr{D}}[\phi(\theta,z) \leq c_\alpha(\theta)] = \alpha \tag{2}$$

The parameter $\alpha$ controls the FPR of the attack and ensures that the attack is tunable, which is a crucial difference to LOGAN. In Algorithm 1, we present the procedure which evaluates the privacy risk of a GAN model by simulating the attack for $k$ target FPR values $\alpha_1^*, \alpha_2^*, \ldots, \alpha_k^*$. During the attack phase, every parameter $\alpha_i^*$ is mapped to the corresponding threshold $c_{\alpha_i^*}(\theta)$ computed to approximately satisfy Eq. (2) using a reference dataset sampled from $\mathscr{D}$. During the evaluation phase, we compute the false negative rate (FNR) $\beta_i$ of the attack for each threshold $c_{\alpha_i^*}(\theta)$ as well as the effective FPR $\alpha_i$ computed on a test dataset $S_{\mathsf{train}}$, an independent sample from the data distribution $\mathscr{D}$. Having measured both FNR and FPR of the attack, we can report its success metrics such as accuracy and precision.

**Algorithm 1** Tunable MIA against GANs

```
$S_{\mathsf{ref}} \sim \mathscr{D}^n$                                  ▷ **Attack Phase**
$L \leftarrow \{\phi(\theta,z) \mid z \in S_{\mathsf{ref}}\}$
**for** $i$ from 1 to $k$ **do**
    $c_{\alpha_i^*} \leftarrow \alpha_i^*$-quantile of $L$
$S_{\mathsf{test}} \sim \mathscr{D}^n$                                 ▷ **Evaluation Phase**
**for** $i$ from 1 to $k$ **do**
    $\alpha_i = \Pr_{z \sim S_{\mathsf{test}}}[\phi(\theta,z) \leq c_{\alpha_i^*}(\theta)]$
    $\beta_i = 1 - \Pr_{z \sim S_{\mathsf{train}}}[\phi(\theta,z) \leq c_{\alpha_i^*}(\theta)]$
```

## 3. Experimental Evaluation

*Datasets* We test our privacy risk evaluation on the MedMNIST+ dataset which is a collection of pre-processed medical image datasets. Specifically, we use ChestMNIST, which contains chest X-ray images, OCTMNIST, which contains retinal optical coherence tomography (OCT) images, and PneumoniaMNIST, which consists of chest X-ray images specifically for pneumonia classification. Each subset is formatted into $64 \times 64$ grayscale images for consistency and ease of use.

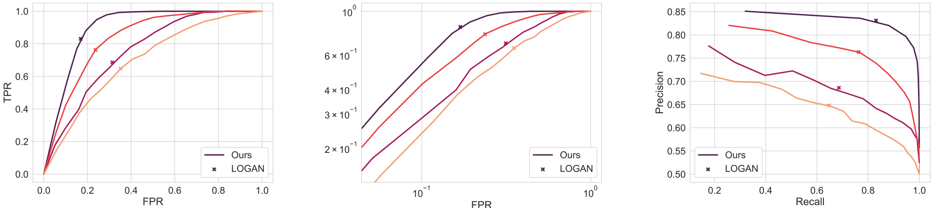
*Generative Method and Baselines* In our study, we focus on image data, and use a Deep Convolutional GAN (DCGAN) [7]. This model extends the original GAN framework by integrating convolutional layers that are well-suited for generating image data. On each dataset, we train several DCGAN models with different random seeds, and compare our privacy evaluation method to LOGAN [2].

**Table 1.** Comparison of the highest achievable MIA risk across different datasets (±$\sigma$ denotes std.)

| Dataset | Attack | Balanced Accuracy | Precision |
|---|---|---|---|
| PneumoniaMNIST | LOGAN | $0.732 \pm 0.081$ | $0.732 \pm 0.081$ |
| | **Ours** | $\mathbf{0.740 \pm 0.089}$ | $\mathbf{0.791 \pm 0.058}$ |
| ChestMNIST | LOGAN | $0.883 \pm 0.164$ | $0.883 \pm 0.164$ |
| | **Ours** | $\mathbf{0.896 \pm 0.171}$ | $\mathbf{0.962 \pm 0.021}$ |
| OCTMNIST | LOGAN | $0.821 \pm 0.109$ | $0.821 \pm 0.109$ |
| | **Ours** | $\mathbf{0.831 \pm 0.111}$ | $\mathbf{0.873 \pm 0.069}$ |

*Results* Table 1 shows the effectiveness of the LOGAN attack against our method across various medical imaging datasets, where we are selecting the best-performing values across multiple False Positive Rate (FPR) thresholds. We compute these scores with four different GANs trained with different seeds, which results in different quality of generation. In all cases, our approach enables us to identify an attack which achieves significantly higher precision than LOGAN by controlling the false positive rate.

The comparison between our privacy risk estimation and LOGAN, as depicted in Fig. 1 for PneumoniaMNIST, is representative of similar trends observed across other datasets. It shows that our approach serves as a generalization of the LOGAN framework: it delivers comparable attack performance at specific FPR thresholds that match the FPR of LOGAN while also being able to simulate a broader spectrum of attack scenarios.



(a) ROC curve: Overall Performance.

(b) Performance in low FPR region.

(c) Precision and recall.

**Figure 1.** Performance comparison of our privacy risk evaluation to LOGAN [2] on the PneumoniaMNIST dataset. The × marks represent the performance of LOGAN—it is a special case of our generalized method. Each line represents attack performance against a single generative model trained with a different random seed.

## 4. Discussion

A limitation of our attack is assuming that the adversary can query the discriminator of the generative model. Indeed, the common practice is to release only the generator subnetwork (in the case of GANs), or limiting sharing to the synthetic dataset. The effectiveness of our attack also depends on the quality and representativeness of the population pool that the attacker has access to. If the attacker's sample pool diverges significantly from the model's training data, the efficacy of the attack in identifying membership may not be representative of the real risk generated by the model. Assuming strong adversarial capabilities such access to the discriminator and a representative “out” records pool, however, is a standard approach [2] for evaluating privacy leakage.

## 5. Conclusions

Our approach presents a scalable attack methodology for estimating membership inference risk against GANs which is more flexible than prior work. It is computationally efficient as it does not require training in any additional models. This efficiency enables us to provide a comprehensive overview of multiple attack scenarios. We focus on low false positive rate regions, which have been identified in prior research as critical for the effective success of membership inference attacks [3]. Moreover, the adversarial model in our attack has a reduced amount of prior knowledge compared to the baseline, as we assume that the attacker does not know the exact number of samples, thereby simulating a more realistic scenario.

## References

[1] 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*, volume 27, 2014.

[2] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models. *Proceedings on Privacy Enhancing Technologies*, 2019(1), January 2019.

[3] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership Inference Attacks From First Principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, May 2022.

[4] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data - anonymisation groundhog day. In Kevin R. B. Butler and Kurt Thomas, editors, *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*. USENIX Association, 2022.

[5] Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. In *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 3493–3514. PMLR, 2023.

[6] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced Membership Inference Attacks against Machine Learning Models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Los Angeles CA USA, November 2022. ACM.

[7] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.