A Survey of Differentially Private Generative Adversarial Networks

Liyue Fan
Department of Computer Science, UNC Charlotte
liyue.fan@uncc.edu

Abstract

Generative Adversarial Networks (GANs) have demonstrated the potential to learn the distribution of training data and generate high quality synthetic data. But recent research has shown that machine learning models, including GANs, may leak sensitive information about training samples. To protect the privacy of training samples, several approaches have been proposed to adopt *differential privacy* in the training of GANs. Moreover, as the winner of the NIST 2018 Unlinkable Data Challenge, differentially private GANs provide a promising direction for generating private synthetic data. In this paper, we survey the existing approaches for differentially private GANs, in order to facilitate the understanding of the current state of research. Specifically, we aim to provide the workshop audience with a comprehensive review of preliminaries, summaries of the approaches, characterization of evaluation criteria, as well as discussions around challenges. In addition, the existing approaches are analyzed with an emphasis on their key innovations and application domains, to facilitate the adoption of the research results. This survey will serve as a reference for future research and point out opportunities for the further development on this important topic.

Introduction

Sharing individual-level data is critical to data analysis tasks, e.g., for propensity score matching and subgroup analyses in clinical studies (Beaulieu-Jones et al. 2019), and for model training and validation in machine learning (Yoon, Jordon, and van der Schaar 2019). However, publicly available individual-level data is often scarce. For data collected from private individuals, there are privacy concerns regarding sharing their data widely. In fact, "anonymized" data can often be re-identified by linking external databases, e.g., (Sweeney 2002), or by examining unique behaviors, e.g., (Barbaro and Jr. 2006) and (De Montjoye et al. 2013). Recently, generative models have received an increasing attention for learning the data distribution from training samples. Those models provide a promising direction for sharing individual-level data, as samples can be drawn from the learned distributions for other analysis tasks.

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and variants have demonstrated superior performance in capturing the underlying data distribution. Benefiting from deep neural networks and game theory, GANs can produce high quality generated samples which are hard to distinguish from real ones. However, recent research shows that machine learning models may leak sensitive information about training samples. Attacks can be launched against target models to infer membership in the training set (Shokri et al. 2017) or to reconstruct training data (Fredrikson, Jha, and Ristenpart 2015). Similarly, GAN models do not provide guarantees on what the generated data may reveal about real, sensitive training data, e.g., real participants in a clinical study. In fact, (Hayes et al. 2019) successfully devised membership inference attacks against GANs in both white-box and black-box access settings.

To protect the privacy of training samples, many approaches adopted *differential privacy* (Dwork et al. 2006b), such as for data release (Zhang et al. 2017), clustering (Su et al. 2016), classification (Chaudhuri, Monteleoni, and Sarwate 2011), hypothesis testing (Gaboardi et al. 2016), and deep learning (Abadi et al. 2016). Differential privacy ensures that an adversary cannot effectively infer whether one record is present or absent in the input data, thus providing rigorous privacy guarantees to samples used to train models. Several recent approaches, such as `DPGAN` (Xie et al. 2018) and `DP-CGAN` (Torkzadehmahani, Kairouz, and Paten 2019), have been proposed to train GANs in a differentially private manner, in hopes of learning the data distributions without disclosing too much information about individual samples. Furthermore, differentially private GANs won the first place in the NIST Unlinkable Data Challenge (Boob et al. 2018), proving to be a promising direction for private synthetic data generation.

In this paper, we survey existing differentially private approaches for learning GANs which can serve as a reference for future research. First, we provide comprehensive preliminaries in order to facilitate the understanding of the fundamental building blocks, including GANs and variants, the definition and properties of differential privacy, and commonly used differentially private training procedures. Second, we summarize 8 previously proposed approaches on differentially private GANs while emphasizing their novelty and application domains, in order to facilitate the dissemination and adoption of the research results. Third, we present the evaluation metrics adopted by the surveyed approaches regarding quality and privacy. Although most approaches are not directly comparable to each other, we categorize the

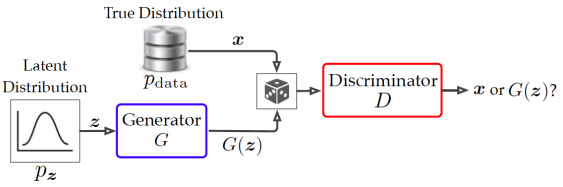

Figure 1: GAN Illustration (Zhang, Ji, and Wang 2018)

adopted quality metrics into the evaluation of *generated data* vs. *methods* trained on generated data. In doing so, we hope to provide future research with a framework for assessment. Last but not least, we identify challenges in learning differentially private GANs and present lessons learned from the surveyed approaches as well as possible avenues for future research.

Preliminary

In this section, we review Generative Adversarial Networks (GANs) and variants, differential privacy concepts and properties, and private training procedures adopted by the surveyed approaches.

GANs and Variants

There has been an increasing interest in generative models as they can produce synthetic data that have similar characteristics as real data. Generative Adversarial Networks (GANs) proposed by Goodfellow et. al (Goodfellow et al. 2014) has been the state-of-the-art method to learn generative models. An illustration of the typical architecture of GANs is depicted by (Zhang, Ji, and Wang 2018) in Figure 1. Essentially, GANs consist of two components, i.e., a generator G and a discriminator D. The generator G learns to capture the original data distribution p_{data} by mapping a latent distribution p_z. Specifically, G takes as input a random noise z and generates synthetic data. On the other hand, the discriminator D learns to discriminate between samples drawn from p_{data}, i.e., x, and those generated by G, i.e., $G(z)$. D takes a sample as input and returns a score representing whether it is real or synthetic. By generating samples that appear to come from the original data distribution, the goal of the generator is to fool the discriminator. The generator and discriminator are trained simultaneously through an adversarial process: the more the generator improves the quality of synthetic data, the harder it is for the discriminator to distinguish between original and synthetic samples.

The problem is formulated as a minimax two-player game with the following objective (Goodfellow et al. 2014):

$$\min_G \max_D \; \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]. \tag{1}$$

Variants of the original GAN formulation have been proposed to improve training and incorporate auxiliary information. Several of them have been adopted by the differentially private approaches.

Conditional GAN (CGAN) (Mirza and Osindero 2014) allows both the generator and discriminator to be conditional on some side information, denoted by y, such as class labels. As a result, CGAN can generate synthetic data for a given class. The objective for CGAN is given by:

$$\begin{aligned}\min_G \max_D \quad & \mathbb{E}_{(x,y)\sim p_{xy}}[\log D(x,y)] \\ & + \mathbb{E}_{z\sim p_z, y\sim p_y}[\log(1 - D(G(z,y),y))]\end{aligned} \tag{2}$$

where p_{xy} represents a joint distribution for real samples x and labels y, and p_y denotes the label distribution.

Auxiliary Classifier GAN (AC-GAN), proposed recently (Odena, Olah, and Shlens 2017), is an extension of the CGAN that introduces a new player C which is a classifier. In practice, the classifier C can be learned with the discriminator as an additional output. The AC-GAN objective combines the original GAN loss with the classification loss on real and synthetic data:

$$\begin{aligned}\min_{G,C} \max_D \quad & \mathbb{E}_{x\sim p_x}[\log D(x)] \\ & + \mathbb{E}_{z\sim p_z, y\sim p_y}[\log(1 - D(G(z,y)))] \\ & - \lambda_c \mathbb{E}_{(x,y)\sim p_{xy}}[\log C(x,y)] \\ & - \lambda_c \mathbb{E}_{z\sim p_z, y\sim p_y}[\log(C(G(z,y),y))]\end{aligned} \tag{3}$$

where λ_c is a hyper-parameter balancing GAN loss and auxiliary classification loss. As can be seen, the discriminator D no longer receives the label as input, which allows AC-GAN to learn a representation in the latent space p_z, independent of the class label.

Wasserstein GAN (WGAN) was proposed by Arjovsky et. al (Arjovsky, Chintala, and Bottou 2017) to improve training stability by minimizing the earth-mover distance (a.k.a., Wasserstein-1 distance) between p_z and p_{data}, as opposed to the Jensen-Shannon divergence (Goodfellow et al. 2014). Specifically, the objective of WGAN is given by:

$$\min_G \max_{w\in W} \; \mathbb{E}_{x\sim p_{data}}[f_w(x)] - \mathbb{E}_{z\sim p_z}[f_w(G(z))] \tag{4}$$

where $\{f_w\}_{w\in W}$ are a family of K-Lipschitz functions for some constant K, i.e., $\|f\|_{Lip} \leq K$. To find functions $\{f_w\}$ for Equation 4, the study (Arjovsky, Chintala, and Bottou 2017) shows that it is sufficient to train a neural network as a typical GAN, while clamping the weights w to a fixed box (e.g., $[-0.01, 0.01]$) after each gradient update. As we will see later, this operation coincides with gradient clipping in the differentially private stochastic gradient descent, for a different purpose.

Differential Privacy (DP)

Differential privacy (Dwork et al. 2006b; 2006a) has become the state-of-the-art paradigm for protecting individual privacy in statistical databases. Intuitively, it guarantees that an algorithm's output distribution will not be significantly changed by the presence (or absence) of any individual record. By observing the output, an adversary cannot learn more about individuals; therefore privacy is protected. In the context of GANs, differential privacy shows promise to enable accurate learning of the data distribution, despite adding or removing any training sample.

Specifically, two databases $\mathcal{D}$ and $\mathcal{D}'$ are neighboring databases if

$$\exists\, x \in \mathcal{D}, s.t. \mathcal{D} \setminus \{x\} = \mathcal{D}'. \tag{5}$$

Definition 1 (Differential Privacy(Dwork et al. 2006a)) *A randomized mechanism $\mathcal{M}$ is (ϵ, δ)-differentially private if for any output set S and any neighboring databases $\mathcal{D}$ and $\mathcal{D}'$*

$$\mathrm{P}(\mathcal{M}(\mathcal{D}) \in S) \leq \mathrm{e}^{\epsilon}\, \mathrm{P}(\mathcal{M}(\mathcal{D}') \in S) + \delta. \qquad (6)$$

where parameters ϵ and δ are non-negative real numbers. When $\delta = 0$, the definition is equivalent to ϵ-DP (Dwork et al. 2006b), where ϵ bounds the difference in $\mathcal{M}$'s output probabilities using $\mathcal{D}$ and $\mathcal{D}'$ everywhere. Small ϵ indicates strong privacy, and vice versa. With $\delta > 0$, (ϵ, δ)-DP allows pure ϵ-DP to fail for a small probability δ. To provide individual privacy in case of ϵ-DP failure, the recommended value for δ should be smaller than the inverse of the database size, i.e., $\frac{1}{|\mathcal{D}|}$. Recently, another relaxation of ϵ-DP, *Rényi Differential Privacy* (RDP) has been proposed. The study (Mironov 2017) shows that RDP is a strictly stronger privacy notion than (ϵ, δ)-DP, and is well-suited for expressing algorithmic privacy guarantees.

Mechanisms *Gaussian mechanism* has been widely-adopted to achieve (ϵ, δ)-DP. Specifically, for a given function $f : D \rightarrow R$, its L_2 sensitivity is defined as $\Delta_2 f = \max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2$. By adding Gaussian noise to f's output, i.e., $\mathcal{M}(\mathcal{D}) = f(\mathcal{D}) + \mathcal{N}(0, (\sigma\Delta_2 f)^2)$, satisfies (ϵ, δ)-DP for $\epsilon < 1$ and $\sigma > \sqrt{2\ln 1.25/\delta}/\epsilon$. To achieve pure ϵ-DP, the *Laplace mechanism* has been widely adopted for database queries as well as learning models. Let $\Delta_1 f = \max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$ denote the L_1 sensitivity of function f. Adding Laplacian noise to f's output, i.e., $\mathcal{M}(\mathcal{D}) = f(\mathcal{D}) + Lap(0, \lambda\Delta_1 f)$, satisfies ϵ-DP for $\lambda = 1/\epsilon$. One advantage of differential privacy is that it is resistant to *post-processing*. Given an arbitrary mapping $f : R \rightarrow R'$ and an (ϵ, δ)-differentially private mechanism $\mathcal{M} : D \rightarrow R$, $f \circ \mathcal{M} : D \rightarrow R'$ is (ϵ, δ)-differentially private.

Privacy Accounting It is possible to combine multiple differentially private mechanisms. The composition of k mechanisms that each of them are (ϵ, δ)-differentially private, is at least $(k\epsilon, k\delta)$-differentially private (Dwork et al. 2006a). The concept of *privacy accountant* (McSherry 2009; Abadi et al. 2016) has been proposed to keep track of the accumulated privacy loss in repeated execution of differentially private mechanisms.

The Moment Accountant (MA) technique (Abadi et al. 2016) has been widely adopted to account for differential privacy in composition of multiple mechanisms, which provides strong estimates of privacy loss compared to other composition theorems (Dwork et al. 2006a; Dwork, Rothblum, and Vadhan 2010). Specifically, MA (Abadi et al. 2016) tracks the overall privacy budget spent, i.e., (ϵ, δ), for composing Gaussian mechanisms with random sampling. It computes the log moments of the random variable for privacy loss and calculates the tail bound using moments bound and standard Markov inequality. On the other hand, RDP Accountant (Mironov 2017) has been shown to provide a *tighter* bound for privacy loss compared to MA (Abadi et al. 2016). In addition, RDP analysis of the Gaussian mechanism is straight-forward and the privacy budget curve for a composite mechanism is simply the sum of base mechanisms. As a result, it has been adopted by one surveyed approach `DP-CGAN` as well as Google's TensorFlow Privacy.

Training with Differential Privacy

DPSGD The differentially private stochastic gradient descent method (DPSGD) developed by (Abadi et al. 2016) is a general method to minimize loss functions empirically for complex networks with differential privacy. In order to obfuscate the influence of individual training samples on the final model, the computed gradients are clipped and random noise is added. The basic training procedure of DPSGD iterates as follows:

- A batch of samples are processed and the empirical loss is computed.
- Gradients for model weights are calculated from the loss.
- Per sample gradient is clipped to the minimum value between its L_2 norm and a clipping bound given as a hyperparameter.
- A Gaussian noise is drawn with a variance proportional to the clipping bound and is added to the clipped gradients.
- The model is updated.
- Cumulative privacy loss is computed using a privacy accountant and the training process terminates if the differential privacy budget is exhausted.

In order to learn differentially private GANs, most surveyed approaches (except for `PATE-GAN`) utilize the DPSGD method to train the discriminator network. Considering DP's resistance to post-processing and the fact that the generator does not access real data, the generator network trained with a differentially private discriminator is also differentially private.

PATE PATE (Papernot et al. 2016) provides a differentially private mechanism for learning classification models, while protecting the privacy of training samples. One surveyed approach, i.e., PATE-GAN (Yoon, Jordon, and van der Schaar 2019), adopts the PATE method to learn GANs privately. In PATE, the training set is partitioned into k disjoint subsets, i.e., $\mathcal{D}_1, ...\mathcal{D}_k$, and k classifiers, i.e., $T_1, ..., T_k$, also called *teachers*, are trained separately on the k partitions. To classify a new instance x, a differentially private output is obtained by performing a noisy aggregation of teacher outputs.

$$\mathrm{PATE}_\lambda(\mathrm{x}) = \arg \max_{\mathrm{j} \in \{1,..,\mathrm{m}\}} (\mathrm{n_j(x)} + \mathrm{Lap}(\lambda)) \qquad (7)$$

where m is the number of possible classes, and $n_j(x)$ denotes the number of teachers that output class j for x. Naturally, each query answered by PATE_λ satisfies $1/\lambda$-DP. An extension of PATE introduced in (Papernot et al. 2016) is to train a *student* model using a public dataset labeled through the PATE method. The student model itself is differentially private, and its internal parameters can be accessed freely, in terms of privacy.

Note that both DPSGD and PATE are compatible with optimization algorithms for training, such as Adam (Kingma

and Ba 2014) and RMSProp (Hinton, Srivastava, and Swersky 2012), which can operate on the clipped and perturbed gradients in DPSGD and the gradients of the student model in PATE.

Differentially Private Algorithms

This section briefly describes each surveyed approach, while highlighting their key innovations and application domains. A comparative summary of the differentially private approaches is provided in Table 1.

DPGAN, dp-GAN, GANobfuscator

DPGAN (Xie et al. 2018) and dp-GAN (Zhang, Ji, and Wang 2018) are two earlier approaches proposed to learn differentially private GANs. DPGAN has been applied to generating image and EHR data, e.g., MNIST and MIMIC III, while dp-GAN has a focus on generating image data, such as MNIST and CelebA. As for training, both DPGAN and dp-GAN follow the DPSGD method for the discriminator. DPGAN adopts the WGAN objective as in Equation 4, while dp-GAN adopts the method proposed by (Gulrajani et al. 2017), aka *improved WGAN*, which is an alternative to weight clipping (Arjovsky, Chintala, and Bottou 2017) in order to enforce the Lipschitz constraint.

DPGAN clips the model weights w, as suggested in (Arjovsky, Chintala, and Bottou 2017), to ensure discriminator network f_w is Lipschitz. Although the method does not have an explicit step that clips the gradients as in DPSGD, the authors (Xie et al. 2018) show that by clipping w to a bounded box $[-c_p, c_p]$, the gradients are automatically bounded by some constant c_g. dp-GAN explicitly clips the gradients and has proposed several optimization strategies to improve the training stability and convergence rate, including: *adaptive clipping* adjusts the parameters' clipping bounds during training by accessing a small set of public data; *parameter grouping* is to cluster those parameters with similar clipping bounds, and to achieve a trade-off between privacy loss and convergence with a uniform clipping bound for each cluster; warm starting initializes the model with a small set of public data to boost the convergence. Similarly, GANobfuscator (Xu et al. 2019) builds on the improved WGAN and adopts adaptive clipping as in dp-GAN.

dp-GAN-TSCD

We denote the approached proposed in (Frigerio et al. 2019) as dp-GAN-TSCD, where TSCD stands for "time series, continuous, and discrete" data. This approach aims to generate time series, e.g., from IoT systems, and multi-variate tabular data, such as UCI Adult and Mushroom, by establishing a flexible architecture for the generator network. Specifically, the authors adopt Long Short-Term Memories (LSTM) for time series. As for discrete variables in tabular data, a softmax layer is added which represents the probability distribution of each variable. dp-GAN-TSCD builds on the improved WGAN (Gulrajani et al. 2017) and performs training with the DPSGD method. In order to reduce the DP noise in parallel with the decreasing trend of the gradients, the authors propose a *clipping decay* strategy where the clipping bound decreases exponentially with each generator update. Empirical evaluation in (Frigerio et al. 2019) has shown improved utility as a result of the clipping decay strategy. It is worthy noting that dp-GAN-TSCD adopts a *universal* clipping bound for all gradients.

PATE-GAN

PATE-GAN (Yoon, Jordon, and van der Schaar 2019) was proposed to generate synthetic multi-variate tabular data without compromising the privacy of training data. Many datasets adopted in (Yoon, Jordon, and van der Schaar 2019) are from health domain, e.g., MAGGIC (Pocock et al. 2012), Unite Network for Organ Transplant (Cecka and Terasaki 1993), and UCI Epileptic Seizure Recognition. PATE-GAN provides an interesting approach to learning differentially private GANs. On one hand, it adopts the PATE method to achieve differential privacy, as opposed to the general DPSGD method. On the other hand, the authors proposed a novel method to train the student discriminator, without requiring publicly available dataset, as opposed to (Papernot et al. 2016). Specifically, a set of *teacher* discriminators are trained separately as vanilla models on disjoint partitions of the training set. In other words, each teacher is trained as in a standard GAN network, except the teacher only sees its own partition of the real data. A *student* discriminator is trained with *generated samples*, labeled by the teachers using the PATE method. The generator is trained to minimize its loss with respect to the student discriminator. As a result, the student model can be trained privately without public data and the generator can utilize the process to improve the generated samples.

DP-CGAN

DP-CGAN was recently proposed by (Torkzadehmahani, Kairouz, and Paten 2019) to generate synthetic data as well as corresponding labels. It adopts the CGAN objective as in Equation 2 and the DPSGD framework for training the discriminator privately. There are two key innovations of the DP-CGAN approach. First, the proposed training procedure splits the discriminator loss between real data, $\log D(x, y)$, and generated data, $\log(1 - D(G(z, y), y))$: gradients for two losses are clipped separately and then summed. This strategy would preserve more useful information from the discriminator loss on real instances, as opposed to clipping gradients from the summed loss. Second, DP-CGAN proposes to use RDP accountant to obtain a tighter estimation on the differential privacy guarantees compared to MA (Abadi et al. 2016). The authors have applied to generating visual data, i.e., MNIST. The evaluation shows DP-CGAN outperforms the baseline DP approach, i.e., without loss separation and RDP accountant, on classification tasks. The summed gradients are perturbed based on the split loss for real data, which results in lower noise and higher quality results.

SPRINT-gan

SPRINT-gan (Beaulieu-Jones et al. 2019) has a focus on sharing patient-level clinical trial data with differential privacy, such that the participants in the trial could not be

identified by accessing the synthetic data. Specifically, the study was motivated by the SPRINT (Systolic Blood Pressure Trial) trial with participants divided to intensive and standard treatment groups. The proposed approach adopts the AC-GAN objective as in Equation 3, where the generator learns to produce samples given the class label, i.e., normal or intensive treatment group, and the discriminator learns to classify real or generated samples and the class label for treatment groups. The `SPRINT-gan` method follows DPSGD framework for training the discriminator; the Adam method is applied to update the discriminator model with the clipped and perturbed gradients. Each training sample contains a participant's measurements for systolic blood pressure, diastolic blood pressure, and the number of medications prescribed, assessed at the first 12 time points in the SPRINT trial. Beside quantitative evaluation, clinicians were asked to score the realism of samples considering both blood pressure measurements and medication counts. The results show that the clinicians' scores for real data and generated data are similar, indicating the proposed method can preserve characteristics of the training data without compromising participant privacy.

DP-FedAvg-GAN

(Augenstein et al. 2019) proposed to train differentially private generative models with *federated learning*, where raw data is distributed across user devices and a central server coordinates the training of a shared global model. By inspecting the output of private models, the goal of (Augenstein et al. 2019) is to develop intuition of potential bugs in the training pipeline. The `DP-FedAvg-GAN` method has been applied to identify image pre-processing bugs (e.g., flipping pixel intensities in MNIST-like data), by examining generated examples from two user subpopulations, i.e., where the performance of a target image classification task is *high* vs. *low*. The training of `DP-FedAvg-GAN` differs from the centralized setting in the discriminator update step: in each round, the server provides the generator and discriminator models to a subset of devices; each device computes the discriminator update with its local private data and clips it, and the clipped updates are sent to the server where they are aggregated and perturbed. As privacy protection is critical in federated learning, the proposed method guarantees *user-level* differential privacy, and data from individual devices is protected from the central server as well as from other participating devices. However, the differential private models may produce low-fidelity data, and quantitative quality evaluation was omitted from the study, as the authors argue that high-fidelity data is not necessary for bug detection.

Evaluating DP Approaches

In this section, we discuss how the differentially private approaches are evaluated in terms of quality and privacy.

Quality Evaluation

As seen in Table 1, the DP approaches proposed so far have been applied to different domains, thus difficult to compare directly in quality. However, we consider it valuable to review and categorize the quality metrics adopted by the DP approaches. In general, metrics for DP approaches largely overlap with those used to assess vanilla GAN models; a metric is often applied to both vanilla GAN models and differentially private models followed by a comparison between the results.

Generated Data The quality of generated data are examined by many DP approaches. With a focus on visual data, `dp-GAN` and `GANobfuscator` adopted metrics that quantify the *realism* and *diversity* of the generated data, e.g., via the Jensen-Shannon divergence and Inception score. Other approaches evaluated the quality of the *learned distributions*, e.g., the distribution of each dimension as well as the relationship between dimensions. Specifically, `SPRINT-gan` computed summary statistics, e.g., mean and standard deviation, for three measurements; the study further evaluated pairwise Pearson correlation between measurements and compared the correlation structures of real data, synthetic data generated by vanilla models, and synthetic data generated by differentially private models. `DPGAN` adopted two metrics, i.e., DWP and DWPre first proposed in (Choi et al. 2017), to evaluate the learned distributions for each dimension and between dimensions, respectively. It is worth noting that *human experts*, i.e., clinicians, were asked to judge whether data samples are real or generated by `SPRINT-gan`, which incorporates domain knowledge into evaluating the quality of the generated data.

Methods Trained with Generated Data The second school of metrics examine the methods trained with generated data. The authors of `PATE-GAN` outlined scenarios where the synthetically generated data can be used differently by real applications. In one scenario, synthetic data is used to train methods and the *performance* of those methods on real data are evaluated and compared with the methods trained on real data. Such metrics include AUROC (area under the receiver operation characteristics curve), AUPRC (area under the precision recall curve), and accuracy. In another scenario, synthetic data is used to identify the best method(s) to be used on read data, and the performance *ranking* of the methods trained on synthetic data is compared to that of the methods trained on real data. Similarly, other approaches evaluated the performance of methods trained on generated data, e.g., for classification tasks, including `DPGAN`, `dp-GAN-TSCD`, `DP-CGAN`, and `SPRINT-gan`. In addition, `SPRINT-gan` proposed to evaluate *similarity* between models trained on synthetic data and real data, e.g., the variable importance scores for Random Forest and model coefficients for SVM and Logistic Regression. Moreover, `dp-GAN` and `GANobfuscator` used the unlabeled synthetic data to augment limited labeled real data for training *semi-supervised models*. In this setting, the performance of semi-supervised models is compared to that of the supervised models trained with limited real data.

Privacy Evaluation

Another important evaluation is conducted on the privacy offered by the differentially private approaches. While all

Method	Application Domains			Training Procedure		Evaluation Metrics		
	Computer Vision	Health	Others	DPSGD	PATE	Data	Methods	Attacks
DPGAN	✓	✓		✓		✓	✓	
dp-GAN	✓			✓		✓	✓	
GANobfuscator	✓			✓		✓	✓	✓
dp-GAN-TSCD			✓	✓		✓	✓	✓
PATE-GAN		✓	✓		✓		✓	
DP-CGAN	✓			✓			✓	
SPRINT-gan		✓		✓		✓	✓	
DP-FedAvg-GAN	✓			✓				

Table 1: Summarizing DP Generative Adversarial Networks

surveyed approaches provide the trade-off analysis between selected quality metrics and the differential privacy parameters, i.e., ϵ and δ, a few demonstrate the defense put up by the differentially private approaches against *known inference attacks*. Specifically, membership inference attacks (Shokri et al. 2017) predict whether a given record was used to train a target model, which might leak additional information about the record. GANobfuscator assessed the accuracy of membership inference attacks with different DP budgets and data sizes. The results show that differentially private GANs effectively reduce the precision of attacks, compared to vanilla GANs; the attack precision is reduced further for larger training sets. Similarly, dp-GAN-TSCD plotted the ROC curves for the membership inference attacks which demonstrate the privacy protection for smaller datasets, e.g., 300 samples. In addition, the study further analyzed the attack accuracy at different training epochs: the attack accuracy rapidly increases with more epochs for the vanilla model, while staying around 50% with differential privacy throughout the training process.

Discussion

In this section we discuss challenges in training GANs with differential privacy, summarize lessons learned from the survey approaches, and point out possible avenues for future research.

Utility Loss Due to gradient clipping and perturbation, GANs trained with differential privacy often exhibit utility loss compared to the vanilla models. Several approaches proposed techniques to overcome the loss of information during the private training process. dp-GAN proposed to separate gradients of weights and biases, and with access to some public data, to adapt the clipping bounds, cluster weight gradients, and warm starting the models without DP constraints. Similarly, dp-GAN-TSCD noted the effect of clipping bounds on utility and proposed a clipping decay technique which reduces the bounds overtime. DP-CGAN separated the discriminator loss on real and generated data to preserve useful information from the real samples and adopted RDP accountant for a tighter estimate of the privacy loss.

Utility Evaluation As presented in the previous section, the utility evaluation of the differentially private GANs adopts quality metrics for the *generated data* and the *methods* trained with the generated data. It is not straight-forward to compare all surveyed approaches directly as they were designed for different domains, e.g., computer vision vs. health, and for different tasks, e.g., labeled data vs. unlabeled data. *Human experts* were brought into the evaluation process by SPRINT-gan authors, to judge the realism of the generated clinical data. We note that the challenge remains to quantitatively evaluate the generated non-image data, e.g., for time series as mentioned by the authors of dp-GAN-TSCD.

Non-Convergence There is a chance that the generator and discriminator may not converge or converge to a noisy equilibrium, as a result of the differentially private training process. Several surveyed approaches studied the training loss over epochs, e.g., DPGAN, GANobfuscator, dp-GAN-TSCD, and SPRINT-gan. While the utility improving techniques mentioned previously may boost convergence by reducing the amount of noise introduced, they do not eliminate non-convergence as the training process must stop when the pre-defined DP budget has been exhausted. To this end, the authors of SPRINT-gan proposed to repeatedly re-run the training process and to account for the total privacy loss from all the runs, until the model converges or the privacy budget is exhausted. Furthermore, they proposed to save the generative models from all epochs and to perform model selection. This strategy was shown in (Beaulieu-Jones et al. 2019) to provide a more diverse set of models.

Privacy Risks Due to the complexity of the models, most DP approaches were evaluated with single-digit ϵ values up to 10, although for some tasks higher ϵ values were adopted, e.g., ≥ 96.5 for DWpre evaluation of DPGAN. As for known attacks, very few approaches were evaluated, e.g., in the presence of membership inference (Shokri et al. 2017) and model inversion (Fredrikson, Jha, and Ristenpart 2015). Furthermore, attacks specifically targeting GANs were proposed recently (Hayes et al. 2019) and could be studied for the differentially private approaches. In general, privacy is still an open issue regarding the adoption of differential privacy in machine learning. The reason is that the standard differential privacy, provided in DPSGD and PATE, protects individual samples in the training set, while in real applications a user may supply multiple samples e.g., for face

recognition (Fredrikson, Jha, and Ristenpart 2015), or a sensitive class as in (Hitaj, Ateniese, and Perez-Cruz 2017). As a result, it is possible to launch attacks on the privacy of the user or the class.

Hyper-parameter Tuning Beside the privacy parameters, e.g., ϵ and δ, other parameters may affect the utility of the differentially private approaches, such as batch size, learning rate, number of discriminator or generator iterations, etc. (Abadi et al. 2016) observed that the accuracy of DPSGD is more sensitive to training parameters than to the neural network structure. They suggested to choose a batch size of the same order of the number of epochs. Furthermore, it is beneficial to start with a a relatively large learning rate, decay it linearly for a few epochs, and keep it constant afterwards. While the surveyed approaches made their parameter settings available, little discussion or comparative evaluation was present to offer readers insights on how the parameters should be chosen. It may be common to tune the parameters empirically: for instance, (Yoon, Jordon, and van der Schaar 2019) reported that the number of teachers in `PATE-GAN` were selected using cross-validation; `DP-CGAN` followed the adaptive strategy for learning rate as in DPSGD with handpicked initial and final values. However, it is unclear whether the privacy loss during parameter tuning has been accounted for in all the surveyed approaches.

Future Work Several directions are open for future research on differentially private GANs. First, researchers may consider the adoption of RDP accountant to tightly bound the privacy loss. For instance, (Beaulieu-Jones et al. 2019) reported that RDP would inflict roughly one fourth of the privacy budget used in their `SPRINT-gan` approach with MA. The reported saving can be significant for utility critical applications. Second, future research could study the non-convergence issues in private GANs theoretically, by considering recent results from the machine learning community, and empirically, e.g., private model selection seems to provide better utility in (Boob et al. 2018; Beaulieu-Jones et al. 2019). Third, the definition of classic DP protects individual instances which may fail to mitigate model inversion attacks. The notion of group privacy (Dwork 2006) can be considered by future research to protect a set of instances with a given size c. Last but not least, the tuning of the hyper-parameters poses a significant challenge for wide applications of differentially private GANs. It can be seen in (Zhang, Ji, and Wang 2018; Xu et al. 2019; Boob et al. 2018) that having access to a small set of public data helps with estimating the hyper-parameters. Future research is also encouraged to explore the suggestions provided in (Abadi et al. 2016; Frigerio et al. 2019) as well as recent results on private hyper-parameter selection (Liu and Talwar 2019).

Acknowledgements

This work is supported in part by National Science Foundation under grant CNS-1949217 and grant CNS-1951430, and UNC Charlotte. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the sponsors.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. ACM.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Augenstein, S.; McMahan, H. B.; Ramage, D.; Ramaswamy, S.; Kairouz, P.; Chen, M.; Mathews, R.; et al. 2019. Generative models for effective ml on private, decentralized datasets. *arXiv preprint arXiv:1911.06679*.

Barbaro, M., and Jr., T. Z. 2006. A face is exposed for aol searcher no. 4417749. *The New York Times*.

Beaulieu-Jones, B. K.; Wu, Z. S.; Williams, C.; Lee, R.; Bhavnani, S. P.; Byrd, J. B.; and Greene, C. S. 2019. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes* 12(7):e005122.

Boob, D.; Cummings, R.; Kimpara, D.; Tantipongpipat, U. T.; Waites, C.; and Zimmerman, K. 2018. Private synthetic data generation via gans. *NIST Unlinkable Data Challenge*.

Cecka, J. M., and Terasaki, P. I. 1993. The unos scientific renal transplant registry. *Clinical transplants* 1–18.

Chaudhuri, K.; Monteleoni, C.; and Sarwate, A. D. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12(Mar):1069–1109.

Choi, E.; Biswal, S.; Malin, B.; Duke, J.; Stewart, W. F.; and Sun, J. 2017. Generating multi-label discrete patient records using generative adversarial networks. *arXiv preprint arXiv:1703.06490*.

De Montjoye, Y.-A.; Hidalgo, C. A.; Verleysen, M.; and Blondel, V. D. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3:1376.

Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; and Naor, M. 2006a. Our data, ourselves: Privacy via distributed noise generation. In Vaudenay, S., ed., *Advances in Cryptology - EUROCRYPT 2006*, 486–503. Berlin, Heidelberg: Springer Berlin Heidelberg.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006b. Calibrating noise to sensitivity in private data analysis. In Halevi, S., and Rabin, T., eds., *Theory of Cryptography*, 265–284. Berlin, Heidelberg: Springer Berlin Heidelberg.

Dwork, C.; Rothblum, G. N.; and Vadhan, S. 2010. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 51–60. IEEE.

Dwork, C. 2006. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, 1–12. Springer Verlag.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333. ACM.

Frigerio, L.; de Oliveira, A. S.; Gomez, L.; and Duverger, P. 2019. Differentially private generative adversarial networks for time series, continuous, and discrete open data. In Dhillon, G.; Karlsson, F.; Hedström, K.; and Zúquete, A., eds., *ICT Systems Security and Privacy Protection*, 151–164. Cham: Springer International Publishing.

Gaboardi, M.; Lim, H.-W.; Rogers, R. M.; and Vadhan, S. P. 2016. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*, 5767–5777.

Hayes, J.; Melis, L.; Danezis, G.; and De Cristofaro, E. 2019. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies* 2019(1):133–152.

Hinton, G.; Srivastava, N.; and Swersky, K. 2012. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on* 14:8.

Hitaj, B.; Ateniese, G.; and Perez-Cruz, F. 2017. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 603–618. ACM.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, J., and Talwar, K. 2019. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, 298–309. New York, NY, USA: ACM.

McSherry, F. D. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 19–30. ACM.

Mironov, I. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–275. IEEE.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2642–2651. JMLR. org.

Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; and Talwar, K. 2016. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*.

Pocock, S. J.; Ariti, C. A.; McMurray, J. J.; Maggioni, A.; Køber, L.; Squire, I. B.; Swedberg, K.; Dobson, J.; Poppe, K. K.; Whalley, G. A.; et al. 2012. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *European heart journal* 34(19):1404–1413.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18.

Su, D.; Cao, J.; Li, N.; Bertino, E.; and Jin, H. 2016. Differentially private k-means clustering. In *Proceedings of the sixth ACM conference on data and application security and privacy*, 26–37. ACM.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557–570.

Torkzadehmahani, R.; Kairouz, P.; and Paten, B. 2019. Dp-cgan: Differentially private synthetic data and label generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Xie, L.; Lin, K.; Wang, S.; Wang, F.; and Zhou, J. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*.

Xu, C.; Ren, J.; Zhang, D.; Zhang, Y.; Qin, Z.; and Ren, K. 2019. Ganobfuscator: Mitigating information leakage under gan via differential privacy. *IEEE Transactions on Information Forensics and Security* 14(9):2358–2371.

Yoon, J.; Jordon, J.; and van der Schaar, M. 2019. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

Zhang, J.; Cormode, G.; Procopiuc, C. M.; Srivastava, D.; and Xiao, X. 2017. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)* 42(4):25.

Zhang, X.; Ji, S.; and Wang, T. 2018. Differentially private releasing via deep generative model (technical report). *arXiv preprint arXiv:1801.01594*.