G-PATE: Scalable Differentially Private Data Generator via Private Aggregation of Teacher Discriminators

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advances in machine learning have largely benefited from the massive accessible training data. However, large-scale data sharing has raised great privacy concerns. In this work, we propose a novel privacy-preserving data Generative model based on the PATE framework (G-PATE), aiming to train a scalable differentially private data generator which preserves high generated data utility. Our approach leverages generative adversarial nets to generate data, combined with private aggregation among different discriminators to ensure strong privacy guarantees. Compared to existing approaches, G-PATE significantly improves the use of privacy budgets. In particular, we train a student data generator with an ensemble of teacher discriminators and propose a novel private gradient aggregation mechanism to ensure differential privacy on all information that flows from teacher discriminators to the student generator. In addition, with random projection and gradient discretization, the proposed gradient aggregation mechanism is able to effectively deal with high-dimensional gradient vectors. Theoretically, we prove that G-PATE ensures differential privacy for the data generator. Empirically, we demonstrate the superiority of G-PATE over prior work through extensive experiments. We show that G-PATE is the first work being able to generate high-dimensional image data with high data utility under limited privacy budgets ($\varepsilon \leq 1$).

1 Introduction

Machine learning has been applied to a wide range of applications such as face recognition [26], autonomous driving [22], and medical diagnoses [8, 19]. However, most learning methods rely on the availability of large-scale training datasets containing sensitive information such as personal photos or medical records. Therefore, such sensitive datasets are often hard to be shared due to privacy concerns. To handle this challenge, data providers sometimes release synthetic datasets produced by generative models learned on the original data. Though recent studies show that generative models such as generative adversarial networks (GAN) [14] can generate synthetic records that are indistinguishable from the original data distribution, there is no theoretical guarantee on the privacy protection. While privacy definitions such as differential privacy [9] and Rényi differential privacy [23] provide rigorous privacy guarantee, applying them to synthetic data generation is nontrivial.

Recently, two approaches have been proposed to combine differential privacy with synthetic data generation: DP-GAN [31] and PATE-GAN [33]. DP-GAN modifies GAN by training the discriminator using differentially private stochastic gradient descent. Though it achieves privacy guarantee due to the post processing property [10] of differential privacy, DP-GAN incurs significant utility loss on the synthetic data, especially when the privacy budget is low. In contrast, PATE-GAN trains differentially private GAN using Private Aggregation of Teacher Ensembles (PATE) [24]. Specifically, it trains a set of teacher discriminators and a student discriminator. To ensure differential privacy, the student discriminator is only trained on records that are produced by the generator and labeled by the teacher discriminators. The key limitation of this approach is that it relies on the assumption that the generator

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

would be able to generate the entire real records space to bootstrap the training process. If most of the synthetic records are labeled as fake by the teacher discriminators, the student discriminator would be trained on a biased dataset and fail to learn the true data distribution. Consequently, this trained generator would not be able to produce high-quality synthetic data. This problem does not exist for traditional GAN, where the discriminator is always able to provide useful information to the generator since they can access the real data records rather than the synthetic data only.

The main contribution of this paper is a new approach named G-PATE for training a differentially private data generator by combining the generative model with PATE mechanism. Our approach is based on the key observation that: *It is not necessary to ensure differential privacy for the discriminator in order to train a differentially private generator*. As long as we ensure differential privacy on the information flow from the discriminator to the generator, it is sufficient to guarantee the privacy property for the generator. To achieve this, we propose a private gradient aggregation mechanism to ensure differential privacy on all the information that flows from the teacher discriminators to the student generator. The aggregation mechanism applies random projection and gradient discretization to reduce privacy budget consumed by each aggregation step and to increase model scalability. Compared to PATE-GAN, our approach has three advantages. First, it improves the use of privacy budget by only applying it to the part of the model that actually needs to be released for data generation. Second, our discriminator can be trained on original data records since it does not need to satisfy differential privacy. Finally, G-PATE preserves better utility on high-dimensional data given its more efficient gradient aggregation mechanism.

Theoretically, we show that our algorithm ensures differential privacy for the generator. Empirically, we conduct extensive experiments on the Kaggle credit dataset and image datasets. To the best of our knowledge, this is the first work that is able to scale to high-dimensional face image dataset such as CelebA while still preserving high data utility. The results show that our method significantly outperforms all baselines including DP-GAN and PATE-GAN.

2 Related Work

Differential privacy [9] is a notion that ensures an algorithm outputs general information about its input dataset without revealing individual information. So far, researchers have proposed different methods to design differentially private statistical functions and machine learning models [2, 5, 1, 21, 12]. Private aggregation of teacher ensembles (PATE) is proposed to train a differentially private classifier using ensemble mechanisms. Scalable PATE [25] improves the utility of PATE with a Confident-GNMax aggregator that only returns a result if it has high confidence in the consensus among teachers. However, PATE and Scalable PATE are only applicable to categorical data (i.e., class labels) as shown in prior work. Differentially private data generative models have also been proposed. Priview [27] generates synthetic data based on marginal distributions of the original dataset, PrivBayes [34] trains a differentially private Bayesian network, and MWEM [15] uses the multiplicative weights framework us to maintain and improve a distribution approximating a given data set with respect to a set of counting queries. However, these approaches are not suitable for image datasets since the statistics they use cannot well preserve the correlations between pixels in an image.

Some recent work applies differential privacy to the training of GAN. DP-GAN [31] and DP-CGAN [29] add Gaussian noise to the gradients of the discriminators during the training process. GS-WGAN [6] reduces gradient sensitivity using the Wassertein distance and uses gradient sanitization to ensure differential privacy for the generator. PATE-GAN [33] trains a student generator using an ensemble of teacher discriminators. Compared to prior work which cannot generate high-dimensional data such as images, G-PATE improves the use of privacy budget by only ensuring differential privacy on the generator. In addition, we apply gradient discretization and random projection to reduce the privacy budget consumed by aggregating high-dimensional gradients.

3 Preliminaries

Differential Privacy. Differential privacy bounds the shift in the output distribution of a randomized algorithm that could be caused by a small input perturbation. The following definition formally describes this privacy guarantee.

Definition 1 ((ε, δ)-Differential Privacy)**.** A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{M})$ and for any neighboring datasets D and D', we have $\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta$.

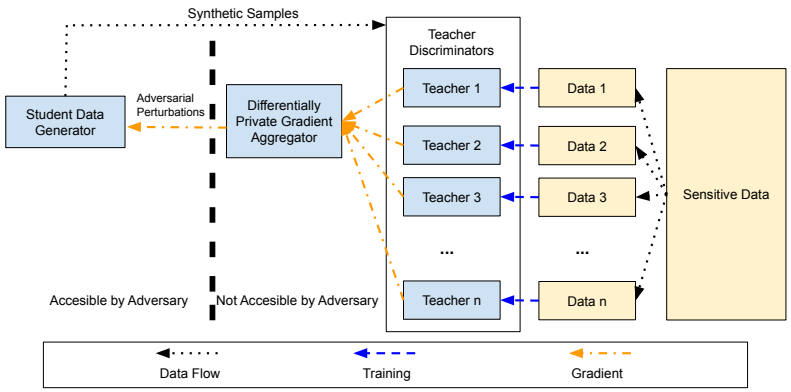


Figure 1: **Model Overview of G-PATE.** The model contains three parts: a student data generator, a differentially private gradient aggregator, and an ensemble of teacher discriminators.

Rényi Differential Privacy. Rényi differential privacy is a natural relaxation of differential privacy. Defined below, its privacy guarantee is expressed in terms of Rényi divergence.

Definition 2 ((λ, ε)-RDP)**.** A randomized mechanism $\mathcal{M}$ is said to guarantee (λ, ε)-RDP with $\lambda > 1$ if for any neighboring datasets D and D',

$$D_\lambda\left(\mathcal{M}(D) \| \mathcal{M}\left(D'\right)\right) = \frac{1}{\lambda-1} \log \mathbb{E}_{x \sim \mathcal{M}(D)}\left[\left(\frac{\mathbf{Pr}[\mathcal{M}(D)=x]}{\mathbf{Pr}\left[\mathcal{M}\left(D'\right)=x\right]}\right)^{\lambda-1}\right] \leq \varepsilon.$$

(λ, ε)-RDP implies $(\varepsilon_\delta, \delta)$-differential privacy for any given probability $\delta > 0$.

Theorem 1 (From RDP to DP)**.** *If a mechanism $\mathcal{M}$ guarantees (λ, ε)-RDP, then $\mathcal{M}$ guarantees $(\varepsilon + \frac{\log 1/\delta}{\lambda-1}, \delta)$-differential privacy for any $\delta \in (0, 1)$.*

Compared to DP, RDP supports easier composition of multiple queries and clearer privacy guarantee under Gaussian noise. Specifically, RDP could be easily composed by adding the privacy budget:

Theorem 2 (Composition of RDP)**.** *If a mechanism $\mathcal{M}$ consists of a sequence of $\mathcal{M}_1, \ldots, \mathcal{M}_k$ such that for any $i \in [k]$, $\mathcal{M}_i$ guarantees (λ, ε_i)-RDP, then $\mathcal{M}$ guarantees $(\lambda, \sum_{i=1}^k \varepsilon_i)$-RDP.*

Suppose f is a real-valued function, and the Gaussian mechanism is defined as $\mathbf{G}_\sigma f(D) = f(D) + \mathcal{N}\left(0, \sigma^2\right)$, where $\mathcal{N}\left(0, \sigma^2\right)$ is normally distributed random variable with standard deviation σ and mean 0. The Gaussian mechanism provides the following RDP guarantee:

Theorem 3 (RDP Guarantee for Gaussian Mechanism)**.** *If f has sensitivity 1, then the Gaussian mechanism $\mathbf{G}_\sigma f$ satisfies $\left(\lambda, \lambda/\left(2\sigma^2\right)\right)$-RDP.*

4 G-PATE: A Scalable Data Generative Method

In this section, we present our method named G-PATE. An overview of the method is shown in Figure 1. Unlike PATE-GAN and DP-GAN, G-PATE ensures differential privacy for the information flow from the discriminator to the generator. This improvement incurs less utility loss on the synthetic samples, so it can generate synthetic samples for higher dimensional and more complex datasets.

G-PATE makes two major modifications on the training process of GAN. First, we replace the discriminator in GAN with an ensemble of teacher discriminators trained on disjoint subsets of the sensitive data. The teacher discriminators do not need to be published, thus can be trained using non-private algorithms. In addition, we design a gradient aggregator to collect information from teacher discriminators and combine them in a differentially private fashion. The output of the aggregator is a gradient vector that guides the student generator to improve its synthetic samples.

Unlike PATE-GAN, G-PATE does not require any student discriminator. The teacher discriminators are directly connected to the student generator. The gradient aggregator adds noise in the information flow from the teacher discriminators to the student generator to ensure differential privacy. This way, G-PATE uses privacy budget more efficiently and better approximates the real data distribution to ensure high data utility.

4.1 Training the Student Generator

To achieve better privacy budget efficiency, G-PATE only ensures differential privacy for the generator and allows the discriminators to learn private information.

To ease privacy analysis, we decompose G-PATE into three parts: the teacher discriminators, the student generator, and the gradient aggregator. To prevent the propagation of private information, the student generator does not have direct access to any information in any of the teacher discriminators. Consequently, we cannot train the student generator by ascending its gradient based on loss of the discriminators. To solve this problem, we propose the use of *adversarial perturbation*, which is a small manipulation on the fake record x that causes the discriminator's loss on x to increase. The adversarial perturbation is calculated by ascending x's gradients on the loss of the discriminator. It teaches the student generator how to improve its fake records. In each training iteration, the student generator is updated in three steps: (1) A teacher discriminator generates adversarial perturbations for each record produced by the student generator. (2) The gradient aggregator takes the adversarial perturbations from all teacher models and generates a differentially private aggregation of them. (3) The student generator updates its weights based on the privately aggregated adversarial perturbation. The process is formally presented in Algorithm 1.

Algorithm 1 - Training the Student Generator.

Input: batch size m, number of teacher models n, number of training iterations N, gradient clipping constant c, number of bins B, projected dimension k, noise parameters σ_1 and σ_2, threshold T, disjoint subsets of sensitive data $S_1, S_2, \ldots, S_n$
for number of training iterations **do**
  //Phase I: Pre-Processing
  Sample m noise samples $\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_m}$
  Generate fake samples $G(\mathbf{z_1}), G(\mathbf{z_2}), \ldots, G(\mathbf{z_m})$
  for each synthetic image $G(\mathbf{z_j})$ **do**
    //Phase II: Private computation and aggregation
    for each teacher model i **do**
      Sample m data samples from S_i
      Update the teacher discriminator D_i
      Calculate the adversarial perturbation $\Delta\mathbf{x_j^{(i)}}$
    end for
    $\Delta\mathbf{X_j} \leftarrow \left(\Delta\mathbf{x_j^{(1)}}; \Delta\mathbf{x_j^{(2)}}; \ldots; \Delta\mathbf{x_j^{(n)}}\right)$
    $\Delta\mathbf{x_j^{priv}} \leftarrow \texttt{DPGradAgg}\left(\Delta\mathbf{X_j}, c, B, k, \sigma_1, \sigma_2, T\right)$
    //Phase III: Post-Processing
    $\hat{\mathbf{x}}_\mathbf{j} \leftarrow G(\mathbf{z_j}) + \Delta\mathbf{x_j^{priv}}$
  end for
  Update the student generator G by descending its stochastic gradient on $\mathcal{L}_G$ on $(\hat{\mathbf{x}}_\mathbf{1}, \hat{\mathbf{x}}_\mathbf{2}, \ldots, \hat{\mathbf{x}}_\mathbf{m})$
end for

Generating Adversarial Perturbations. Let D be a teacher discriminator. Given a fake record x, we use $\mathcal{L}_D(x)$ to represent D's loss on x. In each training iteration, the weights of D are updated by descending their stochastic gradients on $\mathcal{L}_D$.

For each input fake record x, we generate an adversarial perturbation Δx that guides the student generator on improving its output. By applying the perturbation on its output, the student generator would get an improved fake record $\hat{x} = x + \Delta x$ on which D has a higher loss. Therefore, Δx is calculated as x's gradients on $\mathcal{L}_D$:

$$\Delta x = \left.\frac{\partial \mathcal{L}_D(a)}{\partial a}\right|_{a=x}. \tag{1}$$

With the adversarial perturbation Δx, the student generator can be trained without direct access to the discriminator's loss.

Updating the Student Generator. A student generator G learns to map a random input z to a fake record $x = G(z)$ so that x is indistinguishable from a real record by D. Given an adversarial perturbation Δx, the teacher discriminators have higher loss on the perturbed fake record $\hat{x} = x + \Delta x$ compared to the original fake record x. Therefore, the student generator learns to improve its fake records by minimizing the mean squared error (MSE) between its output $G(z)$ and the perturbed fake record $\hat{x}$.

$$\mathcal{L}_G(z, \hat{x}) = \frac{1}{k}\sum_{i=1}^{k}(G(z_i) - \hat{x}_i)^2, \tag{2}$$

where k is the number of synthetic records generated per training iteration. To ensure differential privacy, instead of receiving the adversarial perturbation from a single discriminator, we train the student generator using a differentially private gradient aggregator that combines adversarial perturbations from multiple teacher discriminators. Details are provided in Section 4.2.

4.2 Differentially Private Gradient Aggregation for G-PATE

G-PATE consists of a student generator and an ensemble of teacher discriminators trained on disjoint subsets of the sensitive data. In each training iteration, each teacher discriminator generates an adversarial perturbation Δx that guides the student generator on improving its output records. Different from traditional GAN, in G-PATE, the student generator does not have access to the loss of any teacher discriminators, and the adversarial perturbation is the only information propagated from the teacher discriminators to the student generator. Therefore, to achieve differential privacy, it suffices to add noise during the aggregation of the adversarial perturbations.

However, the aggregators used in PATE and PATE-GAN are not suitable for aggregating gradient vectors because they are only applicable to categorical data. Therefore, we propose a differentially private gradient aggregator (`DPGradAgg`) based on PATE. With gradient discretization, we convert gradient aggregation into a voting problem and get the noisy aggregation of teachers' votes using PATE. Additionally, we use random projection to reduce the dimension of vectors on which the aggregation is performed. The combination of these two approaches allows G-PATE to generate synthetic samples with higher data utility, even for large scale image datasets, which is hard to be achieved by PATE-GAN. The procedure is formally presented in Algorithm 2.

Algorithm 2 - Differentially Private Gradient Aggregator (`DPGradAgg`). This algorithm takes a list of gradient vectors and returns a differentially private aggregation.

Input: Concatenated gradient vectors from each teacher model $\Delta\mathbf{X} = (\Delta\mathbf{x}^{(\mathbf{1})}, \ldots, \Delta\mathbf{x}^{(\mathbf{n})})$, gradient clipping constant c, number of bins B, projected dimension k, noise parameters σ_1 and σ_2, threshold T
$k_0 \leftarrow$ the dimension of $\Delta\mathbf{x}^{(\mathbf{1})}$
$\mathbf{R} \leftarrow$ a $k_0 \times k$ random projection matrix with each component randomly drawn from $\mathcal{N}(0, \frac{1}{k})$
$\Delta\mathbf{U} \leftarrow \Delta\mathbf{X}\mathbf{R}$
for each column $\mathbf{u_j}$ of $\Delta\mathbf{U}$ **do**
  Clip $\mathbf{u_j}$ to $(-c, c)$
  $h \leftarrow$ the histogram of $\mathbf{u_j}$ with B bins of width $\frac{2c}{B}$
  $\Delta u_j^{\text{priv}} \leftarrow$ `Confident-GNMax`$(h, \sigma_1, \sigma_2, T)$
end for
$\Delta\mathbf{u}^{\text{priv}} \leftarrow (\Delta u_1^{\text{priv}}; \ldots; \Delta u_k^{\text{priv}})$
$\Delta\mathbf{x}^{\text{priv}} \leftarrow \Delta\mathbf{u}^{\text{priv}}\mathbf{R}^\intercal$
Output: $\Delta\mathbf{x}^{\text{priv}}$

Gradient Discretization. Since PATE is originally designed for aggregating the teacher models' votes on the correct class label of an example, the aggregation mechanism in PATE only applies to categorical data. Therefore, we design a three-step algorithm to apply PATE on continuous gradient vectors. First, we discretize the gradient vector by creating a histogram and mapping each element to the midpoint of the bin it belongs to. Then, instead of voting for the class labels as in PATE, a teacher discriminator votes for k bins associated with k elements in its gradient vector. Finally, for each dimension, we calculate the bin with most votes using the `Confident-GNMax` aggregator [25] (Appendix D). The aggregated gradient vector consists of the midpoints of the selected bins.

With gradient discretization, the teacher discriminators can directly communicate with the student generator using the PATE mechanism. Since these teacher discriminators are trained on real data, they can provide much better guidance to the generator compared to the student discriminator in PATE-GAN, which is only trained on synthetic samples. Moreover, the `Confident-GNMax` aggregator ensures that the student generator would only improve its output in the direction agreed by most of the teacher discriminators.

Random Projection. Aggregation of high dimensional vectors is expensive in terms of privacy budget because private voting needs to be performed on each dimension of the vectors. To save privacy budget, we use random projection [3] to reduce the dimensionality of gradient vectors. Before the aggregation, we generate a random projection matrix with each component randomly drawn from a Gaussian distribution. We then project the gradient vector into a lower dimensional space using the random projection matrix. After the aggregation, the aggregated gradient vector is projected back to its original dimensions. Since the generation of random projection matrix is data-independent. It does not consume any privacy budgets.

Random projection is shown to be especially effective on image datasets. Since different pixels of an image are often highly correlated, the intrinsic dimension of an image is usually much lower than the number of pixels [13]. Therefore, random projection maximizes the amount of information a student generator can get from a single query to the `Confident-GNMax` aggregator, and makes it possible for G-PATE to retain reasonable utility even on high dimensional data. Moreover, random projection preserves similar squared Euclidean distance between high-dimensional vectors, therefore is beneficial to privacy protection both theoretically and empirically [32].

5 Privacy Guarantees

In this section, we provide theoretical guarantees on the privacy properties for G-PATE. To start with, we propose the following definition for a differentially private data generative model.

Definition 3 (Differentially Private Generative Model)**.** Let G be a generative model that maps a set of points Z in the noise space $\mathcal{Z}$ to a set of records X in the data space $\mathcal{X}$. Let $\mathcal{D}$ be the training dataset of G and $\mathcal{A} : \mathcal{D} \mapsto G$ be the training algorithm. We say that G is a (ε, δ)*-differentially private data generative model* if the training algorithm $\mathcal{A}$ is (ε, δ)-differentially private.

Definition 3 relaxes the definition of a DP-GAN by focusing the protection only on the generative model in a GAN. This relaxation saves privacy budget during training and improves the utility of the model. Moreover, the relaxation does not compromise the privacy guarantee for the synthetic data.

Lemma 1. *Let G be an (ε, δ)-differentially private data generative model trained on a private dataset $\mathcal{D}$. For any $Z \in \mathcal{Z}$, the synthetic dataset $X = G(Z)$ is (ε, δ)-differentially private.*

Proof Sketch. Lemma 1 is a consequence of the post-processing property of differential privacy. First, the random points Z are independent of the private dataset $\mathcal{D}$. Second, one does not need to query the discriminator during the data generation process. Therefore, the synthetic dataset is generated by post-processing G and is guaranteed to be (ε, δ)-differentially private.

Lemma 1 shows that a differentially private generative model is able to support infinite number of queries to the data generator and can be used to generate multiple synthetic datasets.

Next, we justify the privacy guarantee of the G-PATE method. The following lemma justifies RDP of the gradient aggregator (Algorithm 2).

Lemma 2 (Rényi Differential Privacy of `DPGradAgg`)**.** *The output of the gradient aggregator (`DPGradAgg`) proposed in Algorithm 2 satisfies $\left(\lambda, \sum_{1 \leq j \leq k} \varepsilon_j\right)$-RDP, where $\lambda > 1$ and ε_j is the data-dependent RDP budget with order λ for the `Confident-GNMax` aggregator on the j-th projected dimension.*

Proof Sketch. Lemma 2 can be proved by combining the RDP guarantee of the `Confident-GNMax` aggregator and the post-processing property of RDP. We first divide the input of `DPGradAgg` into two categories. The first category contains data independent parameters, including the gradient clipping constant c, the number of bins B, the projected dimension k, the noise parameters σ_1 and σ_2, and the threshold T. These parameters do not contain private information. The second category contains the gradient vectors $\Delta\mathbf{X} = (\Delta\mathbf{x}^{(\mathbf{1})}, \ldots, \Delta\mathbf{x}^{(\mathbf{n})})$, which are data-dependent and sensitive. Our privacy analysis focuses on the computation on $\Delta\mathbf{X}$. With random projection and gradient discretization, we convert $\Delta\mathbf{X}$ into k histograms and pass the histograms into the `Confident-GNMax` aggregator. Since `Confident-GNMax` satisfies data-dependent RDP [25], the privacy guarantee nicely propagates to the output of `DPGradAgg`.

We analyze the RDP guarantee of `DPGradAgg` by composing the privacy budget consumed by the `Confident-GNMax` aggregator on each projection dimension. Therefore, the Rényi differential privacy budget of the training algorithm is a composition of the data-dependent Rényi differential privacy budget of the `Confident-GNMax` aggregator over k dimensions. The data-dependent privacy budget for each `Confident-GNMax` aggregation is dependent on σ_1, σ_2, and threshold T (Appendix D). The remaining parameters (e.g. gradient clipping constant c, number of bins B) do not influence the privacy guarantee.

The next theorem justifies RDP of the G-PATE training process.

Lemma 3 (Rényi Differential Privacy of G-PATE)**.** *Let $\mathcal{A}$ be the training algorithm for the student generator (Algorithm 1) with N training iterations and k projected dimensions. The data-dependent Rényi differential privacy for $\mathcal{A}$ with order $\lambda > 1$ is $\varepsilon = \sum_{1 \leq i \leq N} \left(\sum_{1 \leq j \leq k} \varepsilon_{i,j}\right)$, where $\varepsilon_{i,j}$ is the data-dependent Rényi differential privacy for the `Confident-GNMax` aggregator in the i-th iteration on the j-th projected dimension.*

Proof Sketch. For the convenience of privacy analysis, we divide the each iteration in Algorithm 1 into three phases: pre-processing, private computation and aggregation, and post-processing. In the pre-processing phase, the generator produces fake samples without accessing the private data. In the private computation and aggregation phase, the teacher discriminators are updated based on private

data. Each teacher discriminator also generates a gradient vector. These vectors are aggregated using the DPGradAgg algorithm. Based on Lemma 2, the data-dependent RDP for this phase is $\sum_{1 \leq j \leq k} \varepsilon_{i,j}$. In the post-processing phase, the student generator is updated using the privately-aggregated gradient vector $\Delta \mathbf{x_j}^{\text{priv}}$. It satisfies RDP because of the post-processing property. Finally, the RDP of Algorithm 1 is composed over N training iterations.

The next theorem provides a theoretical guarantee on the differential privacy of G-PATE.

Theorem 4 (Differential Privacy of G-PATE)**.** *Given a sensitive dataset $\mathcal{D}$ and parameters $0 < \delta < 1$, let G be the student generator trained by Algorithm 1. There exists $\varepsilon > 0$ and $\lambda > 1$ so that G is a $(\varepsilon + \frac{\log 1/\delta}{\lambda - 1}, \delta)$-differentially private data generative model.*

Theorem 4 is the consequence of converting the Rényi differential privacy guarantee in Lemma 3 to differential privacy (Theorem 1). The proof is omitted to Appendix G.

6 Experimental Evaluation

We evaluate G-PATE against three state-of-the-art models: DP-GAN, PATE-GAN and GS-WGAN. We first perform comparative analysis with baselines on the tabular and image datasets used in the corresponding works, including the Kaggle credit tabular dataset and the grayscale image datasets (MNIST and Fashion-MNIST). In addition, we evaluate G-PATE on the privacy-sensitive large-scale high-dimensional face dataset CelebA.

6.1 Experimental Setup

Tabular Dataset. we use the same Kaggle credit card fraud detection dataset [7] (Kaggle Credit) as in [33]. The dataset contains 284,807 samples representing transactions made by European cardholders' credit cards in September 2013, and 492 (0.2%) of these samples are fraudulent transactions. Each sample consists of 29 continuous features from a PCA transformation on the original features.

Image Datasets. To demonstrate the superiority of G-PATE to PATE-GAN on high dimensional image datasets, we train G-PATE on MNIST, Fashion-MNIST [30], and the celebrity face datasets CelebA [20]. MNIST and Fashion-MNIST consist of 60,000 training examples and 10,000 testing examples. Each example is a 28×28 grayscale image, associated with a label from 10 classes. The CelebA dataset contains 202,599 images aligned and cropped based on the human face. We create three datasets: CelebA-Gender(S) is a binary classification dataset that uses the gender attributes as the labels and resizes the images to $32 \times 32 \times 3$; to demonstrate the scalability of G-PATE we also create CelebA-Gender(L) with the same label while resizing the images to $64 \times 64 \times 3$; CelebA-Hair contains images as 64x64x3 with three hair color attributes (black/blonde/brown). We follow the official training and testing partition as [20].

Implementation Details. For the Kaggle Credit dataset, both the generator and discriminator networks of G-PATE are fully connected neural network with the same architecture as PATE-GAN [33]. We use random projection with 5 projection dimensions during gradient aggregation. We use the DCGAN [28] architecture on the image datasets. We set the projection dimensions to 10 during gradient aggregation. More details are provided in Appendix E.

Evaluation Metrics. To compare the *data utility* of different data generators, we follow the standard protocol [33, 6] and train a classifier on the synthetic data and test it on the real data to benchmark the usefulness of the synthetic data for downstream tasks. Specifically, we report the **AUROC** of the classifiers for the binary-class tabular dataset, and the **classification accuracy** trained on CNN for the multi-class image datasets to measure the data utility. In addition, for image datasets, we also evaluate the *visual quality* of the synthetic images using **Inception Score (IS)** [17] and **Fréchet inception distance (FID)** [16]. More details can be found in Appendix E and F.

6.2 Evaluation Results

Kaggle Credit. The Kaggle Credit dataset is highly unbalanced. We calculates the ratio between positive and negative classes using Laplacian mechanism [10] with $\varepsilon = 0.01$. We then train a $(0.99, 10^{-5})$-differentially private data generator and sample the synthetic records according to the noisy class ratios. By the composition theorem of differential privacy [10], the data generation mechanism is $(1, 10^{-5})$-differentially private.

To compare with PATE-GAN, we select 4 commonly used classifiers evaluated in [33] and report the **AUROC** of the 4 classifiers trained on the corresponding synthetic data. We evaluate G-PATE under

Table 1: **Performance Comparison on the Tabular Dataset and Image Datasets.** We compare G-PATE with DP-GAN, PATE-GAN, GS-WGAN and vanilla DC-GAN. Vanilla DC-GAN has no privacy protection. The best results are highlighted in bold. Table (a) presents AUROC of the classifier trained on synthetic data and tested on real data. The performance satisfying $(1, 10^{-5})$-differential privacy is evaluated over 4 different classifiers: logistic regression (LR), AdaBoost, bagging, and multi-layer perceptron (MLP).

(a) **Data Utility (AUROC) on Kaggle Credit Tabular Dataset.**

	DC-GAN	PATE-GAN	DP-GAN	G-PATE
LR	0.9430	0.8728	0.8720	**0.9251**
AdaBoost	0.9416	0.8959	0.8809	**0.8981**
Bagging	0.9379	0.8877	0.8657	**0.8964**
MLP	0.9444	0.8925	0.8787	**0.9093**
Average	0.9417	0.8872	0.8743	**0.9072**

(b) **Visual Quality Evaluation on Image Datasets using Inception Score (IS)**.

Dataset	Real data	ε	DP-GAN	PATE-GAN	GS-WGAN	G-PATE
MNIST	9.86	1	1.00	1.19	1.00	**3.60**
		10	1.00	1.46	**8.59**	5.16
Fashion-MNIST	9.01	1	1.03	1.69	1.00	**3.41**
		10	1.05	2.35	**5.87**	4.33
CelebA	1.88	1	1.00	1.15	1.00	**1.17**
		10	1.00	1.16	1.00	**1.37**

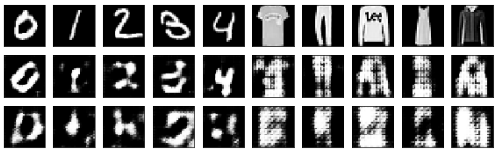
(c) **Data Utility (Accuracy) on Image Datasets.** The table presents the classification accuracy of CNN models trained on the generated data and tested on real data evaluated under two private settings: $\varepsilon = 10$ and $\varepsilon = 1$ given $\delta = 10^{-5}$.

Dataset	DC-GAN	ε	DP-GAN	PATE-GAN	GS-WGAN	G-PATE
MNIST	0.9653 ($\varepsilon = \infty$)	1	0.4036	0.4168	0.1432	**0.5880**
		10	0.8011	0.6667	0.8066	**0.8092**
Fashion-MNIST	0.8032 ($\varepsilon = \infty$)	1	0.1053	0.4222	0.1661	**0.5812**
		10	0.6098	0.6218	0.6579	**0.6934**
CelebA-Gender(S)	0.8002 ($\varepsilon = \infty$)	1	0.5201	0.4448	0.6293	**0.7016**
		10	0.5409	0.5870	0.6326	**0.7072**
CelebA-Gender(L)	0.8149 ($\varepsilon = \infty$)	1	0.5330	0.6068	0.5901	**0.6702**
		10	0.5211	0.6535	0.6136	**0.6897**
CelebA-Hair	0.7678 ($\varepsilon = \infty$)	1	0.3447	0.3789	0.3375	**0.4985**
		10	0.3920	0.3900	0.3725	**0.6217**

Table 2: **Data Utility (Accuracy) on Image Datasets given small privacy budgets.** G-PATE and baselines are evaluated following the same way as Table 1c given $\delta = 10^{-5}$ and low $\varepsilon \leq 1.0$.

ε	MNIST				Fashion-MNIST			
	DP-GAN	PATE-GAN	GS-WGAN	G-PATE	DP-GAN	PATE-GAN	GS-WGAN	G-PATE
0.2	0.1104	0.2176	0.0972	**0.2230**	0.1021	0.1605	0.1000	**0.1874**
0.4	0.1524	0.2399	0.1029	**0.2478**	0.1302	0.2977	0.1001	**0.3020**
0.6	0.1022	0.3484	0.1044	**0.4184**	0.0998	0.3698	0.1144	**0.4283**
0.8	0.3732	0.3571	0.1170	**0.5377**	0.1210	0.3659	0.1242	**0.5258**
1.0	0.4046	0.4168	0.1432	**0.5801**	0.1053	0.4222	0.1661	**0.5567**

Table 3: **Visualization of generated instances by G-PATE.** Row 1 (real image), row 2 ($\varepsilon = 10, \delta = 10^{-5}$) and row 3 ($\varepsilon = 1, \delta = 10^{-5}$) each presents one image from each class (the left 5 columns are MNIST images, and the right 5 columns are Fashion-MNIST images).



the same experimental setup as PATE-GAN for $\varepsilon = 1$. The results for baselines are following [33], and we obtain a higher baseline performance for DC-GAN compared to the results reported in [33].

Table 1a presents the data utility analysis based on AUROC between G-PATE and PATE-GAN on Kaggle Credit dataset. G-PATE outperforms both PATE-GAN and DP-GAN and is close to the vanilla DC-GAN which has no privacy protection. The high performance of G-PATE is partly due to the relatively low dimensionality of the Kaggle Credit dataset and the abundance of training examples. More experimental results on Kaggle Credit dataset are presented in Appendix A.

Image Datsets. To understand G-PATE's performance on image datasets, we evaluate the *data utility* and *visual quality* of the generated images with G-PATE, PATE-GAN, DP-GAN, and GS-WGAN on the MNIST, Fashion-MNIST, and CelebA datasets. The analysis is performed under two private settings: $\varepsilon = 1, \delta = 10^{-5}$ and $\varepsilon = 10, \delta = 10^{-5}$.

For the data utility, We report the **classification accuracy** of the synthetic data in Table 1c. G-PATE outperforms baselines under both settings, and there is a more significant improvement for the setting with a stronger privacy guarantee (i.e., $\varepsilon = 1$). Specifically, we observe that on the Fashion-MNIST dataset the synthetic records generated by DP-GAN under this setting are close to random noise, while the model trained on G-PATE generated data retains an accuracy of 58.12%.

To demonstrate the scalability of our algorithm, we also conduct experiments on the high-dimensional face dateset CelebA. The synthetic data generated by G-PATE is highly utility-preserving, while DP-GAN can barely converge given the high-dimensionality of the data. Particularly, even with the strict privacy budget $\varepsilon = 1$, the accuracy of the generated data by G-PATE on CelebA-Gender(S) is only around 10% lower than the vanilla DC-GAN. Moreover, although the dimensionality of CelebA-Gender(L) is $4\times$ larger than CelebA-Gender(S), the accuracy of both datasets is very close, which again demonstrates the scalability of G-PATE.

Table 4: **Analysis on the Hyper-parameters.** We performed comprehensive studies on the hyper-parameters of G-PATE (the number of teachers, the projection dimensions, gradient clipping constant c, and the number of bins B) for MNIST and Fashion-MNIST with $\varepsilon = 1$ and $\delta = 10^{-5}$. "N/A" means "no projections".

	Projection Dimensions k				**# of Teachers** n			**Gradient Clipping Constant** c				**# of bins** B		
	5	**10**	20	N/A	2000	3000	**4000**	5e-4	1e-4	5e-5	1e-5	5	10	20
MNIST	0.4638	**0.5880**	0.5604	0.1141	0.4240	0.5218	**0.5880**	0.4754	**0.5880**	0.5505	0.4668	**0.5880**	0.5810	0.4706
Fashion	0.5129	**0.5812**	0.5172	0.1268	0.3997	0.4874	**0.5812**	0.5140	0.5567	**0.5812**	0.5339	0.5575	**0.5812**	0.5400

To better understand different generative models, we evaluate the visual quality of the generated image data, though it is not the main focus of the differentially private data generator. In particular, we visualize the generated DP images in Table 3. We also provide an quantitative analysis based on **Inception Score** for G-PATE and baselines in Table 1b. G-PATE can consistently generate better images than baselines when $\varepsilon = 1$, for which GS-WGAN does not converge, which demonstrates the superiority of G-PATE. When $\varepsilon = 10$, G-PATE achieves the best performance on the high-dimensional face dataset CelebA. Although GS-WGAN has better visual quality on MNIST and Fashion-MNIST when $\varepsilon = 10$, G-PATE has the best data utility across different settings and datasets, which suggests that the data utility and visual quality are two orthogonal metrics, and it would be an interesting future direction to improve the visual quality. More evaluation details and evaluation of **Fréchet inception distance (FID)** can be found in Appendix F.

Analysis on the Hyper-parameters. We perform comprehensive ablation studies on the number of teachers and the the projection dimensions to gain better understanding about G-PATE. As shown in Table 4, G-PATE benefits from having more teacher discriminators. Under the same privacy guarantee, the number of noisy votes (σ_1 and σ_2) remains the same, so the output of the noisy voting algorithm is more likely to be correct, and the model would get better performance. However, this benefit diminishes as the training set for each teacher model gets smaller with the increasing number of teachers, and 4000 teachers have already achieved satisfiable results. Table 4 also demonstrates the effectiveness of the random projection method, which improves the classification accuracy by around 47%. With larger projection dimensions, the privacy consumption increase rapidly as we are accessing more private data. But if the projection dimension is too small, useful information can be lost during projection. We find the best trade-off when projection dimension equals to 10. G-PATE achieves better performance given samller bins (≤ 10). With larger bins, the teachers attain a lower agreement rate, leading to worse performance.

Analysis under Limited Privacy Budgets. We conduct another set of ablation studies given limited privacy budgets. From Table 2, we can observe that G-PATE starts to converge even under small ε on both MNIST and Fashion-MNIST datasets. The utility stably increases when the privacy budgets increase. Among different small ε, G-PATE achieves significantly higher accuracy than baselines. In particular, the accuracy of G-PATE under $\varepsilon = 0.6$ is four times higher than DP-GAN, which indicates that G-PATE is able to generate differentially private data with high utility under low privacy budgets.

Agreement of Teachers. In the gradient aggregation step, G-PATE relies on the agreement of teacher models to select the gradient direction. When there is a high agreement rate among teacher models, the gradient aggregator is more robust to noise and more likely to select a gradient direction that preserves higher utility. Intuitively, because the training partitions of teacher models come from the same dataset, we expect the teacher models to learn similar real data distribution. As a consequence, the gradients generated by the teacher models are expected to be similar. Empirically, we evaluate the agreement rate of teacher models on MNIST with 4000 teacher models and $\varepsilon = 1$. On average, the gradient aggregator achieves 60.35% agreement rate for teacher votes on the most agreed direction, and 39.11% on the second agreed direction, which suggests that teacher models have high agreement rates on the top agreement directions.

7 Conclusion

We propose G-PATE, a novel approach for training a differentially private data generator for high-dimensional data. G-PATE is enabled by a novel differentially private gradient aggregation mechanism combined with random projection. It significantly outperforms prior work on both image and non-image datasets in terms of preserving data utility for generated datasets. Beyond the high utility compared with the state-of-the-art differentially private data generative models under similar setting, G-PATE is also able to preserve high data utility even given small privacy budgets.

References

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

[2] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 464–473. IEEE, 2014.

[3] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.

[4] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

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

[6] D. Chen, T. Orekondy, and M. Fritz. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. *arXiv preprint arXiv:2006.08265*, 2020.

[7] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166. IEEE, 2015.

[8] M. de Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis, 2016.

[9] C. Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

[10] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[11] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[12] A. Friedman and A. Schuster. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–502. ACM, 2010.

[13] S. Gong, V. N. Boddeti, and A. K. Jain. On the intrinsic dimensionality of face representation. *CoRR*, abs/1803.09672, 2018.

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

[15] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. *arXiv preprint arXiv:1012.4763*, 2010.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[17] A. Khetan and S. Oh. Achieving budget-optimality with adaptive schemes in crowdsourcing. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 4844–4852. Curran Associates, Inc., 2016.

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

[19] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[20] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[21] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

[22] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.

[23] I. Mironov. Renyi differential privacy. In *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, pages 263–275. IEEE, 2017.

[24] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017.

[25] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and U. Erlingsson. Scalable private learning with PATE. In *International Conference on Learning Representations*, 2018.

[26] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[27] W. Qardaji, W. Yang, and N. Li. Priview: practical differentially private release of marginal contingency tables. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1435–1446. ACM, 2014.

[28] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[29] R. Torkzadehmahani, P. Kairouz, and B. Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[30] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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

[32] C. Xu, J. Ren, Y. Zhang, Z. Qin, and K. Ren. Dppro: Differentially private high-dimensional data release via random projection. *IEEE Transactions on Information Forensics and Security*, 12(12):3081–3093, 2017.

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

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.2, where we discuss improving the visual quality is an interesting future direction.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6.1 about the experiment setup.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We report the best test accuracy.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]