

SMOTE-DP: Improving Privacy-Utility Tradeoff with Synthetic Data

Yan Zhou¹, Bradley Malin², and Murat Kantarcioglu³

¹ University of Texas at Dallas
Richardson, TX 75082
yan.zhou2@utdallas.edu

² Vanderbilt University Medical Center
Nashville, TN 37232
b.malin@vumc.org

³ Virginia Tech
Blacksburg, VA 24061
muratk@vt.edu

Abstract. Privacy-preserving data publication, including synthetic data sharing, often experiences trade-offs between privacy and utility. Synthetic data is generally more effective than data anonymization in balancing this trade-off, however, not without its own challenges. Synthetic data produced by generative models trained on source data may inadvertently reveal information about outliers. Techniques specifically designed for preserving privacy, such as introducing noise to satisfy differential privacy, often incur unpredictable and significant losses in utility. In this work we show that, with the right mechanism of synthetic data generation, we can achieve strong privacy protection without significant utility loss. Synthetic data generators producing contracting data patterns, such as Synthetic Minority Over-sampling Technique (SMOTE), can enhance a differentially private data generator, leveraging the strengths of both. We prove in theory and through empirical demonstration that this SMOTE-DP technique can produce synthetic data that not only ensures robust privacy protection but maintains utility in downstream learning tasks.

Keywords: Privacy · Synthetic Data · Utility

1 Introduction

Privacy-preserving data publishing has important practical implications for decision-making in both academic and commercial applications. The main challenge of data sharing is minimizing the risk of privacy leakage so that the benefit of publicly sharing data valuable for the society will outweigh the cost of individual privacy. Traditional methods for injecting privacy into publicly available data are based on the principle of anonymization [38,19]. Yet these types of techniques have been shown to be vulnerable to privacy attacks such as linkage based on residual quasi-identifiers to resources that contain the identity of the corresponding records [24]. More recently, privacy-preserving data publishing methods have

focused on synthetic data generation. Synthetic data has the potential to preserve statistical correlations while hiding personal information [40,1].

However, a recent study challenges such claims about the privacy protective nature of synthetic data [36]. The main disagreement lies in the perception of synthetically generated data as artificial data, and whether real data can be recovered. Although synthetic data from generative models largely preserves the statistical properties of the original data, it is susceptible to the risk of linkability which allows the attacker to re-identify records in the sensitive dataset. Individuals most vulnerable to linkage attacks are those present in the outlier group, statistically lying an abnormal distance from other samples in the population [36]. Using this statistical infrequency as prior knowledge, the attacker can strategically determine the membership information of a given target record in a sensitive dataset. In both image and tabular data domains, synthetic data has demonstrated certain vulnerabilities to private information extraction attacks [4]. Quantifying the protection synthetic data can provide remained largely unexplored until differential privacy (DP) emerged. Differential privacy bounds the maximum impact a single data entry can have on the output of a computation. Although differentially private models can provide formal guarantees for privacy, it is shown that they tend to incur significant utility loss [22], and it remains unclear how the trade-off affects the overall statistical data properties [36].

This begs the question of if the two types of synthetic data approaches can be used in conjunction to complement each other, so that privacy protection and utility preservation can be achieved simultaneously. In this paper, we present a synthetic data generator that can substantially improve privacy-utility trade-off, referred to as *SMOTE-DP*. The *SMOTE-DP* generator is composed of an end-to-end construct that enables data flow from a non-DP data generator G_1 to a DP generator G_2 to produce the final synthetic dataset (as illustrated in Figure 1):

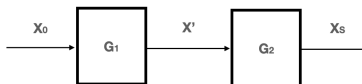


Fig. 1. SMOTE-DP: a synthetic data generator pipeline.

- a.) For the non-DP generator G_1 , we employ a well-known synthetic data generator SMOTE [3]. SMOTE is frequently relied upon to rebalance an imbalanced dataset to improve model accuracy on minority classes. When compared to other generative models, SMOTE has been shown to achieve a more reasonable balance between privacy and utility [27]. To our knowledge, its power to reinforce an existing differentially private mechanism for better risk-utility trade-off has neither been studied theoretically nor empirically. Note that our goal is not to promote the use of SMOTE. As we will explain later, SMOTE can be replaced by any data generation/transformation techniques that produce contracting data patterns.

- b.) For the DP generator, we use a differentially private Bayesian network as the differential private model satisfying ϵ -differential privacy [45]. An excessively large privacy budget ϵ , which induces very little utility loss, is often insufficient for achieving a desired privacy guarantee [36,33]. This scenario can be resolved by SMOTE because of the contracting nature of SMOTE-generated data patterns, which constitutes the underlying principle of SMOTE-DP for privacy preserving synthetic data generation. As with SMOTE, any differentially private data generator can be used as a substitute for the DP Bayesian network used in this study. There is no theoretical dependence between specific DP and non-DP generator pairs.

An important note is that *SMOTE-DP is different from sequential composition where source data is consulted more than once and multiple results of differential privacy are released. SMOTE transforms source data prior to DP.*

We demonstrate, both in theory and through empirical analysis, that SMOTE-DP provides a more balanced risk-utility trade-off than either technique can achieve in isolation. The main contributions of this work are as follows:

- We propose a synthetic data generation mechanism, composed of both DP and non-DP generators, that can radically improve utility while providing strong privacy protection.
- We provide a theoretical justification that SMOTE can reduce sensitivity, that is, the maximum change in the output caused by a single change in the input, consequently strengthen privacy protection provided by the DP generator.
- We provide empirical results to demonstrate that our SMOTE-DP data generator provides strong privacy protection without sacrificing utility.
- We recommend to be cautiously skeptical, because of the complexity of the real world, that higher privacy levels must be achieved with smaller privacy parameter values, and hence at the expense of significant utility loss. At least, this assumption should be verified before it is taken for granted.

The remainder of the paper is organized as follows. Section 2 puts our investigation in context of the related literature. Section 3 presents the theory that supports our proposed technique, and Section 4 presents our empirical results. Finally, Section 5 concludes our work.

2 Related Work

Data anonymization techniques such as generalization and deletion [38,19,24] were developed in response to the need of hiding personally identifiable information and hence preventing privacy leakage. These anonymization techniques were quickly shown to be vulnerable to linkage (also known as re-identification) and inference attacks [23,25,30,34,4,32,43]. To overcome the limitations of traditional anonymization methods, synthetic data generation techniques have been proposed [40,1]. The notion of fully synthetic data was proposed [31] to facilitate the

release of data of “no actual individual”. Exactly how much protection synthetic data can provide remained unanswered in techniques covered under the conventional statistical disclosure limitation framework. And, over time, nearly all such techniques, including “suppression, coarsening, swapping, shuffling, sampling, and most noise-infusion techniques”, were shown to have failed to satisfy differential privacy [1], especially when data is high-dimensional and sparse and the validity of protection can only be guaranteed when the minimum prior sample size is much larger than practically possible [11].

Differential privacy [7,14,20,21,39] (DP), compared to traditional statistical disclosure limitation, provides theoretical guarantees against inferential disclosure for synthetic data. DP bounds the maximum amount of information an attacker can infer from the presence of an individual. Despite its great popularity, DP’s risk-utility trade-off, where privacy risk and data utility are inversely correlated, has led to some concerns [12,22,36]. Recently, Muralidhar et al. [22] raised a strong objection to the use of DP as a privacy model on microdata by demonstrating that it does not guarantee both confidentiality and utility, and the amount of information loss as a result of privacy protection remains unknown. DP synthetic data has also been shown to be vulnerable to privacy attacks such as linkage attacks [35]. More important, it has been observed that some individuals receive substantially less protection than others and, perhaps of greater concern, it is difficult to predict which individuals records are vulnerable to attack. Stadler et al. [36] restated similar concerns and proposed new privacy protection measures. They point out that differentially private synthetic data protects individuals at a significant cost in utility, and it lacks transparency about the trade-off. As a result, it remains unclear what data properties have been preserved and what have been suppressed. The risk-utility trade-off also creates a debate about what the ideal privacy parameter ϵ should be. NIST’s 2022 challenge on the selection of ϵ reported that organizations choose a wide range of ϵ values, all significantly exceeding the value recommended by the DP researchers [26].

Regardless of the existing open challenges, DP models have been quickly adopted by an increasing number of organizations in industry [5]. Addressing the risk-utility issue remains a highly non-trivial matter. In this paper, we present a synthetic data generation mechanism that radically improves the privacy-utility tradeoff. The proposed technique is composed of a differential private data generator and the oversampling technique SMOTE [3]. The idea is motivated by the recent findings that new samples synthesized with SMOTE demonstrate better utility and good privacy compared to other synthetic data generators [27]. In addition, SMOTE-based oversampling has also been demonstrated as an effective technique for mitigating biases in AI algorithms [46,2,13,41,17]. On the one hand, SMOTE demonstrates characteristics in line with differential privacy models that are protective of the private/sensitive data to an extent; on the other hand, SMOTE has the potential to complement differential privacy by keeping utility loss bounded.

SMOTE is widely used in machine learning to oversample a minority class to improve classification performance when data is imbalanced. SMOTE first

selects an instance x from the minority class, then creates synthetic samples from x by performing linear interpolation between x and one of its K nearest neighbors $x_i \in x_1, \dots, x_K$. Informally, SMOTE generates data in accordance with the same assumption that the k -nearest neighbor algorithm is based on. In this respect, similar points are located near one another and share the same density function value. One disadvantage of SMOTE-generated synthetic data is a potential to deviate from the true distribution of the underlying group [29]. SMOTE-generated samples tend to be more contracted and form unrealistic graph patterns where edges are densely populated with synthetic samples due to linear interpolation [9,8]. However, SMOTE and its extensions in general produce better classification accuracy. This benefit is not limited to minority data. In fact, both accuracy and distribution deviation improve when the size of data increases [8].

SMOTE with differential privacy was empirically confirmed for its positive effect on downstream classification where a privately generated noisy histogram was used [18]. To develop a deeper understanding of the effect of SMOTE on a DP generator, in this paper we directly tackle the underlying mechanism that connects to the improved risk-utility trade-off in the context of privacy attacks such as linkage attacks.

3 Methodology

In this section, we provide theoretical justification for our data generation mechanism, SMOTE-DP, that improves privacy and preserves utility. In essence, SMOTE generates contracting data patterns with smaller covariances that allow differentially private mechanisms to choose “unreasonably” large ϵ values that would normally be considered as meaningless. Large ϵ values typically correspond to an underestimation in sensitivity, raising the potential for breaking differential privacy guarantees. We explain how SMOTE assists in mitigating this problem.

3.1 Background

Our proof is built on two important foundations. One is the sensitivity of query output. The other is the distribution of SMOTE-generated data, which verifies contracting data patterns generated by SMOTE.

Sensitivity of Query Output A differentially private algorithm achieves its bound by applying an amount of noise that scales with the sensitivity of the query output. Given a query function f , sensitivity measures the maximum change in the output of f caused by a single change to an instance in the dataset. Formally, sensitivity is defined as follows:

Definition 1. *Given a query function $f : D^n \rightarrow R$, let the sensitivity of f be Δf :*

$$\Delta f = \sup_{X, X': d(X, X') \leq 1} |f(X) - f(X')|$$

where $d(X, X')$ is the distance between two datasets X and X' , and two neighboring datasets have a distance no greater than one.

SMOTE-generated Data Patterns Theoretical analyses have been developed to help understand the underlying workings of SMOTE. Elreedy and Atiya [8] explore SMOTE-generated data patterns by investigating the mean and covariance of the new distribution. Let $x_0 \in R^d$ be a data point in the original distribution, and x be one of its K nearest neighbors. Let their difference $\Delta = x - x_0$. The SMOTE-generated sample z is $z = x_0 + w\Delta$, where w is a random number from the uniform distribution in $[0, 1]$. Assuming the probability density of the original distribution is multivariate Gaussian, the mean and the covariance of the SMOTE-generated data distribution are [8]:

$$E(z) \approx \mu_{x_0} \quad (1)$$

$$\Sigma_z = \Sigma_{x_0} + [(2\pi)^{\frac{1-d}{2}} \frac{C}{3} |\Sigma_{x_0}|^{\frac{1-d}{2d}} \Sigma_{x_0} \left(\frac{d}{2d-1}\right)^{\frac{d}{2}} - 2\pi C |\Sigma_{x_0}|^{\frac{1}{d}} \Sigma_{x_0} \left(\frac{d}{d-2}\right)^{\frac{d+2}{2}}] I \quad (2)$$

where C is a random variable parameterized in terms of the dimension d and the number of chosen neighbors K , $|\Sigma_{x_0}|$ is the determinant, and I is the identity matrix. Note that the mean of the SMOTE-generated samples is close to the true mean. Algebraic computation results in a negative value from the term inside the square brackets in the covariance, suggesting the contraction of the SMOTE-generated data patterns compared to the original distribution, with the diagonal variances becoming smaller, as illustrated in Figure 2. Hence,

$$tr(\Sigma_z) < tr(\Sigma_{x_0}) \quad (3)$$

Other probability density assumptions such as multivariate Laplace also suggest such contracting effect on SMOTE-generated data. Later, Elreedy et al. formulate the probability distribution of the SMOTE-generated samples [9], which is also in consistent with the contracting phenomenon derived in [8].

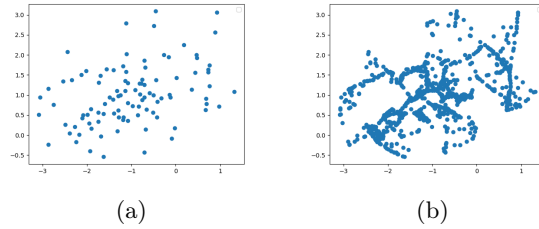


Fig. 2. Original data distribution vs. SMOTE-generated sample distribution.

3.2 SMOTE-DP: Theory and Proof

Given a dataset $X \sim \mathcal{N}(\mu, \Sigma)$, let $X' \sim \mathcal{N}(\mu, \Sigma)$ be a dataset such that X and X' differ by one instance, denoted as $\|X - X'\|_1 \leq 1$. We define the sensitivity Δf_s of the synthetic output as follows:

Definition 2. For any datasets $X, X' : \|X - X'\|_1 \leq 1$, a generative model f_S , and two synthetic data sets, $X_S = f_S(X)$ and $X'_S = f_S(X')$, generated from X and X' with f_S , the sensitivity of the synthetic output is:

$$\Delta f_s = \max_{X_S, X'_S} \mathbb{E}(\|X_S - X'_S\|) = \max_{X_S, X'_S} (\|\mu_{X_S} - \mu_{X'_S}\| + \text{tr}(\Sigma_{X_S} + \Sigma_{X'_S}))$$

where $X_S, X'_S \sim \mathcal{N}(\mu_S, \Sigma_S)$, μ_{X_S} and $\mu_{X'_S}$ are the mean of X_S and X'_S , and Σ_{X_S} and $\Sigma_{X'_S}$ are the covariance matrices.

Sensitivity of SMOTE-transformed Data Given a dataset X , we apply SMOTE on X to generate synthetic data sets X_S and X'_S where $\|X_S - X'_S\|_1 \leq 1$. A DP algorithm achieves ϵ -differential privacy by adding noise to X_S according to a Laplace distribution:

$$X^\epsilon = f_S(X) + Y, \text{ where } Y \sim \text{Lap}(\Delta f_s / \epsilon) \quad (4)$$

where Δf_s is the sensitivity defined on the synthetic data generation mechanism $f_S()$. In this paper, $f_S()$ is specifically defined as SMOTE. Notice the amount of noise introduced scales proportionately to sensitivity.

Theorem 1. (SMOTE Sensitivity Reduction) Let Δf be the sensitivity defined on dataset X , Δf_s be the sensitivity on the SMOTE-transformed dataset X_S with $f_S : X \mapsto X_S$, $\Delta f_s < \Delta f$ given $X \sim \mathcal{N}(\mu, \Sigma)$ and $X_S \sim \mathcal{N}(\mu_S, \Sigma_S)$.

Proof. Given any datasets $X, X' : \|X - X'\|_1 \leq 1$, according to Definition 2, its sensitivity is:

$$\Delta f = \max_{X, X'} \mathbb{E}(\|X - X'\|) = \max_{X, X'} (\|\mu_X - \mu_{X'}\| + \text{tr}(\Sigma_X + \Sigma_{X'}))$$

where μ_X, Σ_X are the mean and covariance of X and $\mu_{X'}, \Sigma_{X'}$ are the mean and covariance of X' . Hence,

$$\Delta f_s - \Delta f = \max_{X_S, X'_S} (\|\mu_{X_S} - \mu_{X'_S}\| + \text{tr}(\Sigma_{X_S} + \Sigma_{X'_S})) - \max_{X, X'} (\|\mu_X - \mu_{X'}\| + \text{tr}(\Sigma_X + \Sigma_{X'}))$$

According to Eq. (1) and the inequality (3) in Section 3.1,

$$\begin{aligned} \max_{X_S, X'_S} \|\mu_{X_S} - \mu_{X'_S}\| - \max_{X, X'} \|\mu_X - \mu_{X'}\| &\approx 0, \\ \max_{X_S, X'_S} \text{tr}(\Sigma_{X_S} + \Sigma_{X'_S}) - \max_{X, X'} \text{tr}(\Sigma_X + \Sigma_{X'}) &< 0. \end{aligned}$$

Therefore, $\Delta f_s < \Delta f$, that is, the sensitivity is reduced after applying SMOTE.

We now prove SMOTE can enhance privacy by essentially reducing the sensitivity of the original dataset Δf , that is, $\Delta f_s = \alpha \cdot \Delta f$ where $0 < \alpha < 1$ because of SMOTE's ability to produce contracting data patterns.

Privacy of SMOTE-DP With the definition of sensitivity for a synthetic generator, we prove a theorem that states SMOTE can enhance privacy given a privacy budget ϵ in DP post-processing.

Theorem 2. (*SMOTE Enhanced Privacy*) *Given a SMOTE-transformed dataset X_S , X_S^ϵ —generated with an ϵ -differential private mechanism—is differentially private with respect to $\alpha \cdot \epsilon$ in the Σ_{X_S} -transformed metric space, where $0 < \alpha < 1$, and Σ_{X_S} is the covariance matrix of SMOTE-generated data.*

Proof. From the discussion in Section 3.2, we know SMOTE-generated patterns are contracting with diagonal variances $(\sigma_1, \dots, \sigma_n)$ becoming smaller in Σ_{X_S} than in Σ_X . Hence, with SMOTE preprocessing, the amount of noise for achieving ϵ differential privacy should be calibrated with respect to the updated *sensitivity* Δf_s . Since $\Delta f_s < \Delta f$, or equivalently, $\Delta f_s = \alpha \cdot \Delta f$ where $0 < \alpha < 1$, it follows:

$$\frac{\Delta f}{\epsilon} = \frac{\Delta f_s}{\alpha \cdot \epsilon}$$

Mapping the ϵ -differentially private data X^ϵ to the SMOTE-transformed data space X_S^ϵ , the noise introduced to the original data X to achieve ϵ -differential privacy effectively becomes the noise added to X to achieve $(\alpha \cdot \epsilon)$ -differential privacy on X_S :

$$X_S^\epsilon = f_S(X) + Y + w\Delta, \text{ where } Y \sim \text{Lap}(\Delta f_s / (\alpha \cdot \epsilon)) \quad (5)$$

which essentially decreases the privacy budget from ϵ to $\alpha \cdot \epsilon$ for $0 < \alpha < 1$.

Theorem 2 justifies that an unreasonably large ϵ chosen as a compromise for better utility for generating X^ϵ (Eq. 4) can become a reasonable choice on the SMOTE preprocessed substitute X_S (Eq. 5), demonstrating the benefit of post-processing DP on the SMOTE-transformed data in synthetic data publishing.

4 Experimental Results

All experiments presented in this section were conducted on an Intel[®] Xeon[®] machine with a 2.30GHz CPU, 256GB memory, and four 8GB GeForce RTX 2080 GPUs. We first verify SMOTE can enhance DP with better utility on a simple artificial dataset for which we know its population mean and covariance.

4.1 Experiments on Artificial Data

We generate a simple two-dimensional dataset X with two classes, each following a normal distribution. We investigate whether applying a DP mechanism on SMOTE-preprocessed data X_S produces better estimate of the mean μ and covariance Σ than applying DP on the original non-private X . We measure the difference in mean and covariance using Frobenius norm. Figure 3 shows the data distributions before and after synthetic generation with DP, SMOTE,

and SMOTE-DP, and the Frobenius norm (denoted as F-norm in the figure) of the mean and covariance differences between the original and the synthetic data generated using the three mechanisms. As can be seen, DP-generated data has the highest covariance difference (4.154) from the original, SMOTE-processed data has the smallest covariance difference (0.291), while SMOTE-DP discovers the middle ground (2.727) between DP and SMOTE. All three result in approximately the same norm for the mean. The experiment demonstrates that SMOTE enhances DP with better data utility by respecting the true mean and covariance. Similar experiments were repeated on 10-dimensional artificial datasets with a various number of informative attributes. The same conclusion can be drawn.

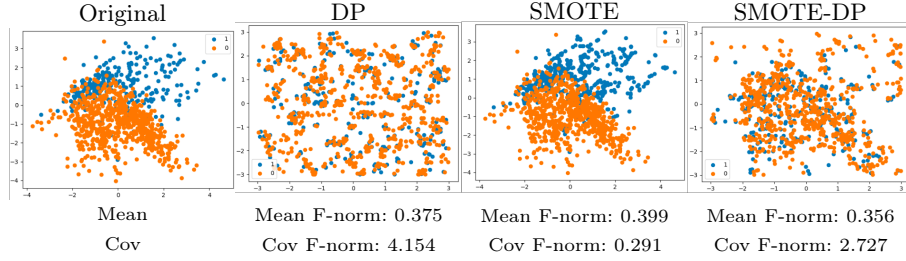


Fig. 3. Data distributions before and after synthetic generation with DP, SMOTE, and SMOTE-DP.

4.2 Experiments on Real Data

We now focus on test cases reported in [36] that were claimed to be most challenging where DP has experienced significant utility loss while protecting privacy, and other generative models with no explicit privacy protection fail to protect outlier data points in the face of linkage attacks. We adapted their source code ⁴ to repeat their experiment and performed additional experiments with SMOTE for comparison.

Experiment Setup Details of datasets, generative models, and metrics used in this experiment are discussed below.

Datasets We include three datasets *Texas* [37], *German Credit Risk*, and *Employee* [16] in the experiments. The *Texas* dataset is the hospital discharge data file released by the Texas Department of State Health Services. As in [36], the dataset used in this experiment consists of 50,000 randomly sampled patient data. There are 18 attributes including 11 categorical and 7 continuous attributes. The *German Credit* data, for greater clarity, contains a smaller set of attributes compared to the original source in the UCI repository [15]. It has ten attributes

⁴ https://github.com/spring-epfl/synthetic_data_release

including the class label. The total number of entries in the dataset is 1000, and each entry represents a person labeled as good or bad credit risk. The *Employee* data contains the information of 4,653 employees in a company. There are nine attributes including the class label.

Generative Models We use the same generative models, both private and non-private, as in [36]. GreedyBayes [28] is a non-private implementation of Bayesian network. The sanitization procedure by NHS England [10] implemented in [36] is used for comparing the effectiveness of synthetic data and data sanitization. Differentially private generative models used in our experiment include PrivBayes [45] and PATEGAN [44]. An extension of SMOTE, SMOTENC, implemented in the Python library [6] is used both standalone and in SMOTE-DP with parameter $K = 1$.

Privacy Gain Protection is measured in terms of privacy gain defined in [36] as the reduction in advantage when the adversary is provided with synthetic data S instead of the real data X : $PG = A(X, x_t) - A(S, x_t)$ where A estimates the adversary’s advantage and x_t is the target of protection.

Attack Target As in [36], five outlier records (outside the 95% quantile range) that are most vulnerable to privacy attacks are chosen for our experiments. Besides outliers, random samples are also selected and studied for comparison.

Linkage (Membership Inference) Attack and Utility Linkage attacks are modeled as a game played between a data publisher and an adversary [42,30,36]. Given publicly released data S and some prior knowledge, the adversary aims to decide whether a target x_t is in the source data X .

In a linkage attack, if the source data X is released, the advantage of the adversary is assigned the highest value of one since the adversary only needs to verify the presence of x_t in the released X . The privacy gain is defined as $PG = 1 - A(S, x_t)$ where $A(S, x_t)$ is the adversary’s advantage if S is released, and is defined as the difference in the probability that the adversary’s guess is “yes”: $A(S, x_t) = P(Yes|S, x_t \in X) - P(Yes|S, x_t \notin X)$. Hence, if the privacy gain $PG \geq 1$ by releasing the synthetic data S , the target x_t is protected; otherwise, there is privacy risk. $PG > 1$ suggests, occasionally, training a generator without the target x_t actually improves the adversary’s chance of predicting “Yes”.

The work reported in [36] argued that DP algorithms failed to protect outliers and privacy was achieved at the expense of utility. Our experiment suggests more careful evaluations, or at least extensive studies are needed to warrant such a claim. The left plot in Figure 4 illustrates the privacy gain (PG) on five manually selected outliers, identical to the ones reported in [36], when different data release techniques are used to publish the *Texas* dataset. The right plot shows the *utility in terms of classification accuracy with the Random Forest classifier*. The adversary is assumed to apply a feature selection technique using correlation analysis [36]. In this experiment, besides setting $\epsilon \in \{0.1, 1.0\}$ as normally being done, we added an additional study where we set an outrageously large $\epsilon = 50$. Quite counterintuitively, the privacy of the selected outliers are equally well protected ($PG \approx 1$) with such a high ϵ value using the PrivBayes

algorithm (2^{nd} & 3^{rd} from left to right in the plots). The corresponding utility, as a result of such high ϵ , is nearly as good as what has been achieved on the non-private raw data. Hence, it is likely that choosing much higher ϵ values is sufficient to both protect privacy and retain utility on this group of outliers. PateGAN also appeared to be insensitive to the ϵ value on the outliers (4^{th} & 5^{th} in the plots).

When SMOTE is used to preprocess the non-private data with $\epsilon = 50$, SMOTE-PateGAN (2^{nd} from the right) achieved much better privacy than PateGAN ($\epsilon = 1.0$). SMOTE helped improve the utility of PateGAN, however, the loss is still significant potentially because of mode collapse introduced by PateGAN. SMOTE-PrivBayes (SMOTEPB, 3^{rd} from the right) achieved similar privacy protection as PrivBayes ($\epsilon = 0.1$) except for one outlier, however, SMOTE-PrivBayes retained much higher utility than PrivBayes did. All DP generators provided better privacy than the non-DP generators (1^{st} , 6^{th} & last).

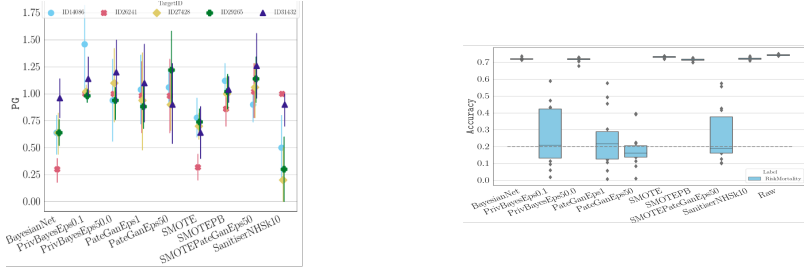


Fig. 4. Adversary's privacy gain & utility on five outliers in the *Texas* dataset.

We also recommend to be cautious when considering DP for privacy protection. Figure 5 shows the results of repeating the previous experiment using 10 randomly selected instances to replace the five outliers. *The non-DP generator BayesianNet* (1^{st} from the left) *provided much better privacy protection than all DP generators* (2^{nd} – 4^{th}) *without incurring any utility loss*. Note that SMOTE-preprocessing helped DP generators achieve comparable levels of privacy to BayesianNet.

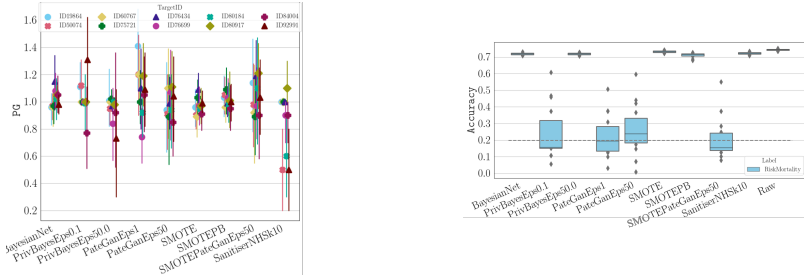


Fig. 5. Adversary's privacy gain & utility on 10 random samples in the *Texas* dataset.

As seen in Figure 5, smaller ϵ values (2^{nd} and 4^{th}) provided better privacy for non-outliers than the larger $\epsilon = 50$ (3^{rd} and 5^{th}) as we normally expect, at the expense of utility. When PrivBayes with $\epsilon = 50$ was applied to the SMOTE-preprocessed data, better privacy ($PG \approx 1$) and stability (low variance) were observed, while utility was largely retained with very little loss. This confirms our hypothesis that SMOTE can enhance both privacy (tolerating large DP ϵ values) and utility (negligible utility loss) for the post-processing DP mechanism.

On a side note, we do not have to designate SMOTE to transform the non-private data space. As in our proof, any synthetic data generators producing contracting data patterns, i.e., data with smaller covariance, can serve as a privacy/utility enhancement mechanism for any DP algorithms, by setting a much less aggressive privacy budget (large ϵ values).

Similar observations can be made on the *GermanCredit* dataset, shown in Figure 6. When DP generators are applied to the SMOTE-preprocessed data, stable privacy protection (low variance on all samples) can be achieved without high utility loss. Pate-GAN again experienced mode collapse, making it a less ideal candidate for privacy protection on this dataset.

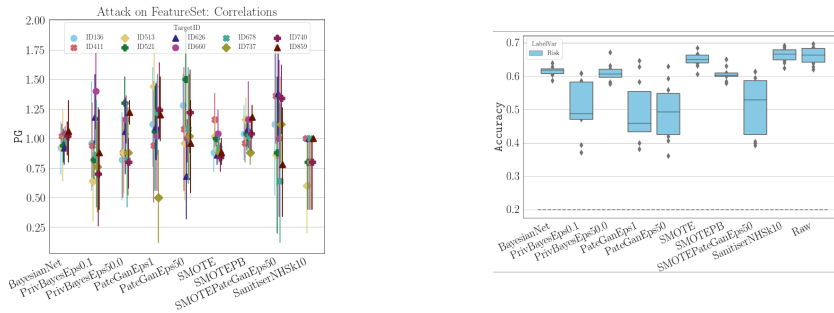


Fig. 6. Adversary’s privacy gain & utility on 10 random samples in the *German Credit* dataset with different publishing methods.

Results on the *Employee* dataset present another counterintuitive aspect of the DP algorithms on the randomly selected samples, as shown in Figure 7. With the large $\epsilon = 50$, the DP algorithms actually achieved better privacy than with a much smaller $\epsilon \in \{0.1, 1.0\}$. In addition, Non-private BayesianNet appeared to preserve privacy better than the two DP algorithms. Again, when the DP algorithms are applied to SMOTE-preprocessed data, they provide better stability and privacy-utility tradeoff.

All our experiments suggest DP alone cannot provide consistent and stable privacy protection for all individuals in real world applications. It is not uncommon to experience worse privacy protection for some individuals with the use of DP mechanisms. However, our SMOTE-DP technique has been shown to consistently provide better and stable privacy protection without significantly hurting utility.

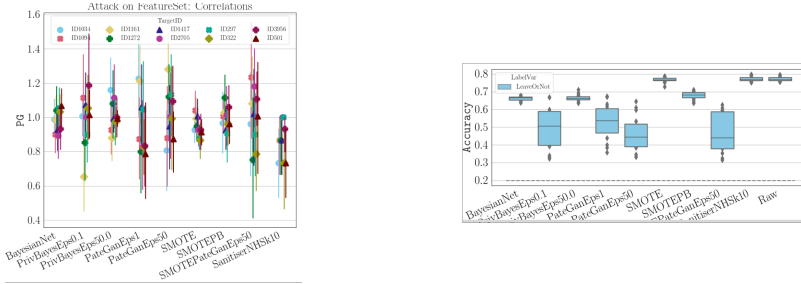


Fig. 7. Adversary’s privacy gain & utility on random samples in the *Employee* dataset with different publishing methods.

5 Conclusions

The release of synthetic data has been touted as offering enhanced privacy protection compared to traditional data sanitization techniques. However, recent research has begun to question the prevailing optimism surrounding the use of synthetic data for privacy-preserving data publishing. One issue is that synthetic data produced by non-private generative models may fail to protect the privacy of outlier records in the dataset. Conversely, while differentially private generators offer strong protection against privacy breaches, they often do so at the expense of significant utility loss. To address this issue, we propose a data generation mechanism that combines the existing non-private synthetic data generator SMOTE with differentially private generative models, aiming to improve utility without compromising privacy protection. The idea is to set a very large privacy budget ϵ for the differentially private generator. With the contracting nature of data processed by SMOTE (smaller covariance), allocating a larger privacy budget becomes justifiable, effectively preserving both privacy and high utility of the data. Through both theoretical investigation and empirical demonstration, we show that our proposed data generation mechanism, SMOTE-DP, significantly enhances data utility while still safeguarding sensitive individual information.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abowd, J.M., Vilhuber, L.: How protective are synthetic data? In: Domingo-Ferrer, J., Saygin, Y. (eds.) *Privacy in Statistical Databases*. pp. 239–246. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
2. Chakraborty, J., Majumder, S., Menzies, T.: Bias in machine learning software: Why? how? what to do? In: *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. pp. 429–440. ESEC/FSE 2021 (2021). <https://doi.org/10.1145/3468264.3468537>, <https://doi.org/10.1145/3468264.3468537>

3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* **16**(1), 321–357 (jun 2002)
4. Chen, D., Yu, N., Zhang, Y., Fritz, M.: Gan-leaks: A taxonomy of membership inference attacks against generative models. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. pp. 343–362. CCS '20, Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3372297.3417238>, <https://doi.org/10.1145/3372297.3417238>
5. Desfontaines, D.: A list of real-world uses of differential privacy. <https://desfontain.es/privacy/real-world-differential-privacy.html> (10 2021), ted is writing things (personal blog)
6. Developers, T.I.L.: Smotenc (2024), https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTENC.html
7. Dwork, C.: Differential privacy. In: *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*. pp. 1–12. ICALP'06, Springer-Verlag, Berlin, Heidelberg (2006). https://doi.org/10.1007/11787006_1, https://doi.org/10.1007/11787006_1
8. Elreedy, D., Atiya, A.F.: A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences* **505**, 32–64 (2019). <https://doi.org/https://doi.org/10.1016/j.ins.2019.07.070>, <https://www.sciencedirect.com/science/article/pii/S0020025519306838>
9. Elreedy, D., Atiya, A.F., Kamalov, F.: A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Machine Learning* (2023). <https://doi.org/10.1007/s10994-022-06296-4>, <https://doi.org/10.1007/s10994-022-06296-4>
10. England, N.: A and e synthetic data, <https://data.england.nhs.uk/dataset/a-e-synthetic-data>
11. Gehrke, J., Kifer, D., Machanavajjhala, A., Abowd, J., Vilhuber, L.: Privacy: Theory meets practice on the map. In: *International Conference on Data Engineering*. p. 10. Cornell University Computer Science Department, Cornell, USA (April 2008), <http://www.truststc.org/pubs/463.html>
12. Ghatak, D., Sakurai, K.: A survey on privacy preserving synthetic data generation and discussion on privacy-utility trade-off problem. In: Su, C., Sakurai, K. (eds.) *Science of Cyber Security - SciSec 2022 Workshops*. pp. 167–180. Springer Nature Singapore, Singapore (2022)
13. Gonzalez Zelaya, C.V., Missier, P., Prangle, D.: Parametrised Data Sampling for Fairness Optimisation. In: *Proceedings of Explainable AI for Fairness, Accountability & Transparency Workshop (KDD XAI)*. ACM (2019)
14. Hall, R., Wasserman, L., Rinaldo, A.: Random differential privacy. *Journal of Privacy and Confidentiality* **4**(2) (Mar 2013). <https://doi.org/10.29012/jpc.v4i2.621>, <https://journalprivacyconfidentiality.org/index.php/jpc/article/view/621>
15. Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994), DOI: <https://doi.org/10.24432/C5NC77>
16. Kaggle: Datasets, <https://www.kaggle.com/datasets>
17. Lavallo, A., Maté, A., Trujillo, J., García Carrasco, J., et al.: A methodology based on rebalancing techniques to measure and improve fairness in artificial intelligence algorithms (2022)
18. Lut, Y.: Privacy-Aware Data Analysis: Recent Developments for Statistics and Machine Learning. Phd thesis, Columbia University (October 2022), available at <https://doi.org/10.7916/k9cc-v818>

19. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1**(1), 3–es (mar 2007). <https://doi.org/10.1145/1217299.1217302>, <https://doi.org/10.1145/1217299.1217302>
20. Mahawaga Arachchige, P.C., Bertok, P., Khalil, I., Liu, D., Camtepe, S., Atiquz-zaman, M.: Local differential privacy for deep learning. *IEEE Internet of Things Journal* **7**(7), 5827–5842 (2020). <https://doi.org/10.1109/JIOT.2019.2952146>
21. Mironov, I.: Rényi differential privacy. In: *IEEE 30th Computer Security Foundations Symposium (CSF)*. pp. 263–275. IEEE (2017)
22. Muralidhar, K., Domingo-Ferrer, J., Martínez, S.: epsilon-differential privacy for microdata releases does not guarantee confidentiality (let alone utility). In: *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23-25, 2020, Proceedings*. pp. 21–31. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-57521-2_2, https://doi.org/10.1007/978-3-030-57521-2_2
23. Narayanan, A., Shi, E., Rubinstein, B.: Link prediction by de-anonymization: How we won the kaggle social network challenge. pp. 1825–1834. *Proceedings of the International Joint Conference on Neural Networks (Oct 2011)*. <https://doi.org/10.1109/IJCNN.2011.6033446>, 2011 International Joint Conference on Neural Network, IJCNN 2011 ; Conference date: 31-07-2011 Through 05-08-2011
24. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. pp. 111–125 (2008). <https://doi.org/10.1109/SP.2008.33>
25. Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: *2009 30th IEEE Symposium on Security and Privacy*. pp. 173–187 (2009). <https://doi.org/10.1109/SP.2009.22>
26. Near, J., Darais, D.: Differential privacy: Future work & open challenges. <https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-future-work-open-challenges> (1 2022), cybersecurity Insights: a NIST blog
27. NIST: Hlg-mos synthetic data challenge submissions (2022), https://media.githubusercontent.com/media/usnistgov/HLG-MOS_Synthetic_Data_Test_Drive/nist-pages/submissions/cra/cra_report.pdf
28. Ping, H., Stoyanovich, J., Howe, B.: Datasynthesizer: Privacy-preserving synthetic datasets. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management. SSDBM '17, Association for Computing Machinery, New York, NY, USA (2017)*. <https://doi.org/10.1145/3085504.3091117>, <https://doi.org/10.1145/3085504.3091117>
29. Prati, R.C., Batista, G.E.A.P.A., Monard, M.C.: Learning with class skews and small disjuncts. In: Bazzan, A.L.C., Labidi, S. (eds.) *Advances in Artificial Intelligence – SBIA 2004*. pp. 296–306. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
30. Pyrgelis, A., Troncoso, C., Cristofaro, E.D.: Knock knock, who's there? membership inference on aggregate location data. In: *25th Annual Network and Distributed System Security Symposium, NDSS 2018. The Internet Society (2018)*, https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_05B-5_Pyrgelis_paper.pdf
31. Rubin, D.B.: Statistical disclosure limitation. *Journal of official Statistics* **9**(2), 461–468 (1993)
32. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: ML-leaks: Model and data independent membership inference attacks and defenses on machine

- learning models. In: Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS) (2019)
33. Seeman, J., Susser, D.: Between privacy and utility: On differential privacy in theory and practice. *ACM J. Responsib. Comput.* (oct 2023). <https://doi.org/10.1145/3626494>, <https://doi.org/10.1145/3626494>, just Accepted
 34. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 3–18 (2017). <https://doi.org/10.1109/SP.2017.41>
 35. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data - A privacy mirage. *CoRR abs/2011.07018* (2020), <https://arxiv.org/abs/2011.07018>
 36. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data – anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 1451–1468. USENIX Association, Boston, MA (Aug 2022), <https://www.usenix.org/conference/usenixsecurity22/presentation/stadler>
 37. Texas Department of State Health Services, Austin, T.: Texas hospital inpatient discharge public use data file 2013 q1-q4, <https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm>
 38. Sweeney, L.: K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**(5), 557–570 (oct 2002). <https://doi.org/10.1142/S0218488502001648>, <https://doi.org/10.1142/S0218488502001648>
 39. Winograd-Cort, D., Haeberlen, A., Roth, A., Pierce, B.C.: A framework for adaptive differential privacy. *Proc. ACM Program. Lang.* **1**(ICFP) (aug 2017). <https://doi.org/10.1145/3110254>, <https://doi.org/10.1145/3110254>
 40. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling Tabular Data Using Conditional GAN. Curran Associates Inc., Red Hook, NY, USA (2019)
 41. Yan, S., Kao, H.t., Ferrara, E.: Fair class balancing: Enhancing model fairness without observing sensitive attributes. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management. pp. 1715–1724. CIKM '20 (2020). <https://doi.org/10.1145/3340531.3411980>, <https://doi.org/10.1145/3340531.3411980>
 42. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: Analyzing the connection to overfitting. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). pp. 268–282 (2018). <https://doi.org/10.1109/CSF.2018.00027>
 43. Yeom, S., Giacomelli, I., Menaged, A., Fredrikson, M., Jha, S.: Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *J. Comput. Secur.* **28**(1), 35–70 (jan 2020). <https://doi.org/10.3233/JCS-191362>, <https://doi.org/10.3233/JCS-191362>
 44. Yoon, J., Jordon, J., van der Schaar, M.: PATE-GAN: Generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations (2019), <https://openreview.net/forum?id=S1zk9iRqF7>
 45. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.* **42**(4) (oct 2017). <https://doi.org/10.1145/3134428>, <https://doi.org/10.1145/3134428>
 46. Zhou, Y., Kantarcioglu, M., Clifton, C.: On Improving Fairness of AI Models with Synthetic Minority Oversampling Techniques, pp. 874–882 (2023). <https://doi.org/10.1137/1.9781611977653.ch98>, <https://epubs.siam.org/doi/abs/10.1137/1.9781611977653.ch98>