On the Formal Privacy Guarantees of Synthetic Data

Marcel Neunhoeffer[1,2], Jonathan Latner[1], and Jörg Drechsler[1, 2, 3]

[1] *Institute for Employment Research, Nuremberg, Germany*
[2] *Ludwig-Maximilians-Universität, Munich, Germany*
[3] *University of Maryland, College Park, USA*

Abstract

What privacy guarantees can synthetic data satisfy even without formal guarantees during the training of the synthesizer? In this paper, we explore this question using synthesizers under simplified settings to show that the privacy guarantees offered by these synthesizers can be directly translated into a ρ-zCDP guarantee. We further explore the conditions under which this equivalence holds and show that it is significantly harder to get formal privacy guarantees for more realistic synthetic data models. Furthermore, we discuss under which conditions such synthetic data can be used to draw valid statistical inferences.

1 Introduction

Synthetic data was first introduced as an idea for statistical disclosure limitation (SDL) in 1993 (Rubin, 1993; Drechsler & Haensch, 2023) based on Rubin's seminal ideas on multiple imputation for nonresponse (Rubin, 1978, 1987). From the beginning, the primary motivation for the approach was to offer a high level of data protection while making it as simple as possible for the data users to obtain valid inferences, including valid confidence intervals or significance tests. The methodology was proposed as an alternative to other protection strategies such as noise infusion or swapping that require complicated adjustment procedures to enable valid inferences from the protected data. The necessary theory to obtain valid variance estimates based on the synthetic data was derived in Reiter (2002) and Raghunathan et al. (2003). We will review these procedures in Section 3.

On the other hand, for many years, most of the published papers on synthetic data, especially those focusing on fully synthetic data, i.e., datasets in which all records have been replaced by synthetic values, did not try to formally quantify the privacy guarantees offered by the synthetic data approach. In his original paper, Rubin seemed to indicate that it would be safe to release synthetic data simply because "not one unit of which [the synthetic data] is an actual unit" (Rubin, 1993, p. 462). This strong belief changed in recent years, and more and more metrics to measure the level of protection of synthetic data generated without formal privacy guarantees are being proposed in the literature (Reiter et al., 2014; Taub & Elliot, 2019; van Breugel et al., 2023; Boudewijn et al., 2023). However, most of these measures are post-hoc and instance-specific, i.e., they only try to measure the risks for a specific generated dataset based on certain assumptions about the background knowledge of an attacker and on the strategy that they would use to learn information from the released synthetic data.

A completely different approach to synthetic data that gained popularity in recent years is routed in the concept of differential privacy (Dwork et al., 2006c). Differential privacy (DP) is a framework that offers formal privacy guarantees by allowing to quantify the privacy loss of a data release mechanism mathematically. With DP, the focus shifts from quantifying the level of protection of the data to measuring the privacy guarantees of the mechanism that produces these data. The major attractiveness of this framework lies in the fact that it quantifies the privacy guarantees ex-ante, i.e., those guarantees always hold irrespective of the input data. It also makes (almost) no assumptions about the attacker, implying that the guarantees will still hold even if more information becomes available that might be used to attack the data. Compared to the more hand-wavy approaches to measure the level of protection of classical synthetic datasets generated without formal guarantees, DP offers an elegant, theoretical approach to measure the privacy guarantees that can be given. In addition, DP has some attractive properties, such as immunity to post-processing and composition properties that we will review in Section 2, which allow us to quantify the accumulation of the privacy loss over multiple data releases.

There is a vast literature on differentially private synthetic data, starting with the work of Barak et al. Barak et al. (2007) and Blum et al. Blum et al. (2008). Ullman and Vadhan Ullman & Vadhan (2011) showed that releasing differentially private synthetic is generally computationally intractable. Hence, most of the literature has focused on methods that are efficient in practical scenarios Hardt & Rothblum (2010); Gaboardi et al. (2014); Zhang et al. (2017); Vietri et al. (2020); McKenna et al. (2021); Neunhoeffer et al. (2021); McKenna et al. (2022); Liu et al. (2023); Wang et al. (2023).

However, a significant downside of all these approaches is that it is tough to quantify the amount of uncertainty and error introduced to ensure DP. Ignoring this extra uncertainty will lead to invalid inferences for most analyses. Still, this critical downside of these procedures received little attention in the literature so far. Except for the paper of Charest (2011), which suggested strategies for obtaining valid inferences from a DP synthesizer for count data, we have yet to be aware that this problem has ever been addressed in the literature. Thus, we currently live in a world in which we have synthesis strategies based on the ideas of Rubin that –assuming correctly specified synthesis models–allow us to obtain valid inferences that fully account for any errors introduced during synthesis but offer no quantifiable privacy guarantees. While on the other hand, we have synthesizers based on the concept of differential privacy that offer very strong formal privacy guarantees, but make it almost impossible to obtain estimates based on the synthetic data that fully capture the uncertainty in those estimates and thus will generally lead to invalid inferences based on the synthetic data.

In this paper, we aim to address this gap by investigating which formal privacy guarantees classical synthesizers based on Rubin's approach can offer and under which circumstances. We only focus on simple artificial examples to illustrate that it is possible, at least in principle, to quantify the formal guarantees of these synthesizers. We explore different data synthesizers and show that the privacy guarantees offered by these synthesizers can be directly translated into a ρ-zCDP guarantee (Bun & Steinke, 2016). We further explore the conditions under which this equivalence holds. We hope that our research will stimulate further research on this topic, allowing quantification of the formal guarantees under more realistic settings in the future.

The remainder of this paper is organized as follows. In Section 2, we review some background material on DP that will be relevant for the remainder of the paper. Section 3 provides a brief overview of the inferential procedures developed to enable valid inferences based on synthetic data generated following Rubin's proposal. In Section 4, we present our findings regarding the privacy guarantees of traditional synthesizers generated without formal guarantees. We limit our evaluations to univariate data and parametric synthesis models based on a normality assumption. The paper concludes with discussions on important research questions that still need to be addressed to quantify the privacy guarantees of traditional synthesizers in realistic settings.

2 Preliminaries on Differential Privacy

Differential privacy is a mathematical notion of privacy for statistical data analysis that bounds the amount of information an adversary can learn about any individual. Given a space of datasets $\mathcal{X}^T$, we say that two datasets D, D' are *neighboring* if they differ in one individual's information.

Definition 2.1 (Differential Privacy Dwork et al. (2006b,a)) *A randomized algorithm $\mathcal{M} : \mathcal{X}^T \to \mathcal{R}$ is* (ε, δ)-differentially private *if for every pair of neighboring datasets $D, D' \in \mathcal{X}^T$, and for every subset of possible outputs $\mathcal{S} \subseteq \mathcal{R}$,*

$$\mathbb{P}[\mathcal{M}(D) \in \mathcal{S}] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(D') \in \mathcal{S}] + \delta.$$

Zero-concentrated differential privacy (zCDP) Dwork & Rothblum (2016); Bun & Steinke (2016) is a variant of differential privacy that quantifies the closeness of distributions differently and is especially amenable to analyzing Gaussian noise addition.

Definition 2.2 (Zero-Concentrated Differential Privacy (zCDP) Bun & Steinke (2016)) *A randomized algorithm $\mathcal{M} : \mathcal{X}^T \rightarrow \mathcal{R}$ is ρ-*zCDP *if for every pair of neighboring datasets $D, D' \in \mathcal{X}^T$, and for all $\alpha \in (1, \infty)$, $\mathrm{D}_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \rho\alpha$, where D_α denotes the Rényi divergence of order α.*

A ρ-zCDP guarantee can be translated to a (ϵ, δ)-DP guarantee using the following result from Bun & Steinke (2016): If $\mathcal{M}$ provides ρ-zCDP, then $\mathcal{M}$ is $(\rho + 2\sqrt{\rho \log(\frac{1}{\delta})}, \delta)$-differentially private for every $\delta > 0$.

Like ε-DP, zCDP offers two desirable properties: immunity to post-processing and composition. Immunity to post-processing implies that any function of an output satisfying ρ-zCDP also satisfies zCDP with the same level of ρ. The composition property characterizes how privacy costs increase as more statistics are released about the data.

Theorem 2.3 (zCDP Post-processing) *Let $M : \mathcal{X}^n \rightarrow \mathcal{Y}$ and $f : \mathcal{Y} \rightarrow \mathcal{Z}$ be randomized algorithms. Suppose that M satisfies ρ-zCDP. Then $f(M(x))$ satisfies ρ-zCDP.*

Theorem 2.4 (zCDP Composition) *Assume $\mathcal{M}_1 : \mathcal{X}^T \rightarrow \mathcal{R}$ is ρ-zCDP and $\mathcal{M}_2 : \mathcal{X}^T \rightarrow \mathcal{R}$ is ρ'-zCDP, then the mechanism defined as $(\mathcal{M}_1, \mathcal{M}_2)$ satisfies $(\rho + \rho')$-zCDP.*

The Gaussian mechanism (Dwork & Roth, 2014; Bun & Steinke, 2016) with parameter σ^2 takes in a function f, dataset D, and outputs $f(D) + \mathcal{N}(0, \sigma^2)$. The variance of the Gaussian distribution is specified as $\sigma^2 = \frac{\Delta^2}{2\rho}$, given the privacy parameter ρ and sensitivity of the function f, Δ.

3 Synthetic Data Based on Rubin's Approach

Most of the approaches that propose synthetic data as a strategy for facilitating access to sensitive data generate synthetic data by fitting a model to the original data and then drawing new values from this model that are disseminated instead of the original data. The main difference between synthetic data generators proposed in the CS literature and the approach proposed by Rubin is that the letter allows for the direct incorporation of the extra uncertainty introduced through the synthesis process. Rubin takes a Bayesian approach to address this problem, similar to his ideas for multiple imputations for nonresponse. He assumes that the synthetic data are generated by random draws from the posterior distribution of the synthetic data given the observed data. Instead of generating only one synthetic dataset, he suggests generating multiple copies of the synthetic data. The extra uncertainty from the synthesis process can then be quantified easily by looking at the variability of the results obtained from the different datasets.

We will only review the inferential procedures for univariate estimates here, borrowing heavily from Drechsler (2011b) and Drechsler & Haensch (2023). The interested reader is referred to Reiter & Raghunathan (2007), which thoroughly reviews all inferential procedures for synthetic data and the nonresponse context.

3.1 Inferential Methods for Rubin's Approach to Synthetic Data

To understand the procedure of analyzing multiply imputed synthetic datasets, think of an analyst interested in an unknown scalar parameter Q, where Q could be, for example, the mean of a variable, the correlation coefficient between two variables, or a regression coefficient in a linear regression. For simplicity, assume that there are no data with items missing in the observed dataset. Inferences for Q derived from the original dataset usually are based on a point estimate q, an estimate for the variance of q, u, and a normal or Student's t reference distribution. For analysis of the synthetic datasets, let $q^{(i)}$ and $u^{(i)}$ for $i = 1, ..., m$ be the point and variance estimates for each of the m synthetic datasets. The following quantities are needed for inferences for scalar Q:

$$\begin{aligned}
\bar{q}_m &= \sum_{i=1}^{m} q^{(i)}/m, \\
b_m &= \sum_{i=1}^{m} (q^{(i)} - \bar{q}_m)^2/(m-1), \\
\bar{u}_m &= \sum_{i=1}^{m} u^{(i)}/m.
\end{aligned}$$

The analyst can use $\bar{q}_m$ as an unbiased point estimate for Q under the assumption that the synthesis models are correctly specified (that is, they match the true data-generating process) and that q would be an unbiased estimate for Q based on the original data. Its variance can be estimated using.

$$T_f = (1 + m^{-1})b_m - \bar{u}_m,$$

Where b_m is an estimate for the variability of the point estimates between the synthetic datasets and $\bar{u}_m$ is an estimate for the sampling variance. When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_f = (m-1)(1 - \bar{u}_m/((1 + m^{-1})b_m))^2$.

3.2 An Alternative Variance Estimator for Fully Synthetic Data

Raab et al. (2016) proposed an alternative variance estimator extending earlier ideas in Drechsler (2011a) that is more suitable for most of the scenarios considered in this paper as it doesn't require a Bayesian approach for synthesis. Instead, it assumes that maximum likelihood estimates are used directly when generating synthetic data. This variance estimator is given by:

$$T_s = \left(\frac{n_{syn}}{n_{org}} + \frac{1}{m} \right) \bar{u}_m,$$

where n_{syn} is the number of synthetic records and n_{org} is the number of records in the original dataset.

If synthetic data are generated following a Bayesian approach as envisioned by Rubin, the variance estimator becomes

$$T_{PPD} = \left(\frac{n_{syn}}{n_{org}} + \frac{1 + n_{syn}/n_{org}}{m} \right) \bar{u}_m,$$

The subscript PPD indicates that the synthetic data are obtained by taking draws from the posterior predictive distribution.

Note that these variance estimators do not rely on the between imputation variance b_m. This offers three crucial advantages compared to T_f, the variance estimator for fully synthetic data discussed above: (i) the estimators can never be negative, (ii) they have less variability than T_f, and (iii) valid variance estimates can be obtained from a single synthetic dataset. See Drechsler (2018) for further discussion of the advantages and disadvantages of the different synthesis strategies and which variance estimator is appropriate in which scenario.

4 On the equivalence of Gaussian Synthetic Data and the Gaussian Mechanism

We start with the following setting: A data curator (e.g., a researcher or a statistical agency) collects data and wishes to release a synthetic data set that preserves some information about the original data. For simplicity, we assume the data is univariate and follows a normal distribution with parameters μ and σ^2. The normality assumption is only required to obtain valid inferences based on the procedures discussed in Section 3. The privacy guarantees of the synthesizer discussed below still hold, even if the normality assumption is violated.

The data curator observes a sample of n observations from that distribution, where each observation is drawn according to

$$X_i \sim \mathcal{N}(\mu, \sigma^2).$$

They then set up a parametric model to provide synthetic data to end users by drawing from

$$X_i^{syn} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2).$$

Depending on the different settings discussed below, the parameters $\hat{\mu}$ and $\hat{\sigma}^2$ are either treated as known, directly estimated from the data (for the plug-in approach), or based on posterior draws (for the Bayesian approach).

We want to know to what extent such a synthetic data generator offers formal privacy guarantees.

4.1 Plug-in Synthesis with Known Variance

We start by analyzing a simplified setting, which assumes that the population variance σ^2 can be treated as public knowledge. From a privacy perspective, this information does not need to be protected. Still, it also means the data curator using the plug-in approach can set up the following synthetic data model:

$$X_i^{syn} \sim \mathcal{N}(\bar{X}, \sigma^2), \tag{1}$$

where $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ denotes the sample mean.

We can compare this model to the ρ- zero concentrated differentially private estimate of the mean using the Gaussian mechanism(Bun & Steinke, 2016):

$$\tilde{X} = \bar{X} + \mathcal{N}(0, \frac{\Delta^2}{2\rho}), \tag{2}$$

where Δ denotes the sensitivity of the mean, i.e., the maximum possible difference between the mean $\bar{X}$ and the mean on a neighboring data set X'. If we assume bounded DP[1] and data that is bounded on the interval $[-d, d]$ (i.e., any value outside this interval is clamped to the closest boundary point of the interval), $\Delta = \frac{|2d|}{n}$[2] Now, if we compare Equation (1) with Equation (2) and solve for ρ, we see that generating a single synthetic record based on Equation (1) offers ρ- zero concentrated DP with $\rho = \frac{4d^2}{2n^2\sigma^2}$. Using the composition property of ρ- zero concentrated DP, this implies that releasing m copies of synthetic data[3], each containing n_{syn} records based on this model will imply a privacy loss of

$$\rho = \frac{mn_{syn}4d^2}{2n^2\sigma^2}.$$

[1] Bounded DP assumes that the neighboring datasets X and X' are obtained by changing one record in the data but keeping the size of the data fixed.

[2] This worst case is obtained if a row of X changes from $-d$ to d.

[3] The idea of releasing m synthetic copies of a single variable is contrived as releasing m copies of synthetic data with n_{syn} records is conceptually the same as releasing a single dataset containing mn_{syn} synthetic records. It only serves to illustrate the impacts of releasing multiple copies of synthetic data.

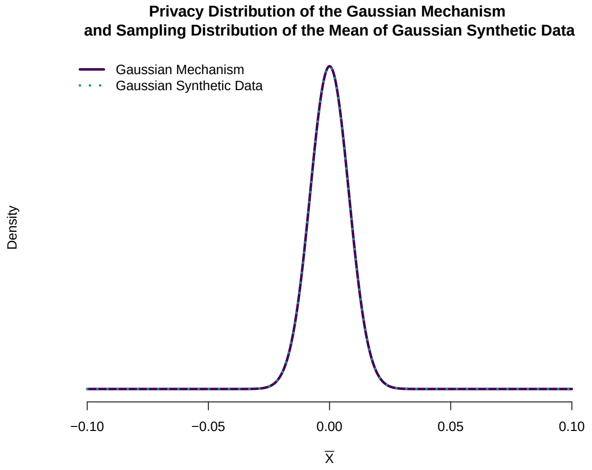

Figure 1: Simulations assume the original data are drawn from a standard normal, with $n = 1,000$. The purple line shows the privacy distribution of the Gaussian mechanism with $\rho = 0.5$ and $\Delta = \frac{8}{1000}$ (i.e., clamping to the interval [-4,4]), assuming the sample mean exactly matches the true mean, $\bar{X} = 0$. The dotted teal line shows the sampling distribution of the mean from synthetic data with $\bar{X} = 0$, $\sigma^2 = 1$, and $n_{syn} = 15625$.

Equivalently, we can determine the number of synthetic samples we can release for any desired level of privacy ρ as

$$n_{syn} = \frac{2n^2\rho\sigma^2}{4d^2m}.$$

In Figure 1 we show that when the number of synthetic samples n_{syn} is chosen according to this equivalence, and the synthetic data are used to estimate μ, The distribution of the estimated mean of the synthetic data over repeated simulations is the same as the distribution if the Gaussian Mechanism is used to compute a DP mean from the original data.

This simple equivalence is helpful to formalize some intuitions about the privacy guarantees of synthetic data. First, we cannot use synthetic data to get meaningful privacy guarantees if the original data has no variance. This makes intuitive sense, as any **one** synthetic sample released from such a constant data set would release **all** the information about the original data. Hence, $\lim_{\sigma^2 \to 0} \rho = \infty$. This, in turn, also means that the release of such parametric synthetic data is not formally private in all cases. In contrast, the Gaussian mechanism could still release formally private statistics about such a constant data set. However, it is reasonable to assume that $\sigma > 0$ for most useful social science applications. Second, releasing more synthetic samples using such a parametric model leads to less privacy and vice versa. Intuitively, if we release no synthetic data, we would perfectly protect the privacy of everyone in the original sample ($\rho = 0$). The more synthetic data we release, the more precise information we release regarding the sample mean. Third, to get any meaningful formal privacy guarantees when releasing synthetic data, we need to limit the influence that any single observation can have on the parametric model's sufficient statistic(s). If the influence of any single observation is unbounded, the only way to preserve privacy is by not releasing any synthetic data.

4.2 Bayesian Synthesis with Known Variance

Assuming a known fixed variance, Bayesian synthesis of a normally distributed variable would proceed in two steps. In the first step, a new value for the mean would be drawn from its posterior distribution:

$$\mu^*|X, \sigma^2 \sim N(\bar{X}, \sigma^2/n).$$

In the second step, synthetic data will be generated by repeatedly drawing from

$$X_i^{syn}|X, \mu^*, \sigma^2 \sim \mathcal{N}(\mu^*, \sigma^2).$$

Given that both steps require drawing from a normal distribution, the synthesis can be conducted in one step by integrating over μ^* and drawing new synthetic records from

$$X_i^{syn}|X, \sigma^2 \sim \mathcal{N}(\bar{X}, \sigma^2(1 + 1/n)).$$

Thus, the privacy loss of the Bayesian synthesis approach is

$$\rho = \frac{mn_{syn}4d^2}{2n^2\sigma^2(1 + 1/n)}.$$

Equivalently, the number of synthetic samples that can be released for a fixed privacy level ρ is given by

$$n_{syn} = \frac{2n^2\rho\sigma^2(1 + 1/n)}{4d^2m}.$$

Interestingly, the Bayesian approach allows for an alternative interpretation, in which the draw of μ^* is interpreted as the Gaussian mechanism, and the synthetic data generation is treated as post-processing. Under this setting, the privacy loss would be

$$\rho = \frac{4d^2n}{2\sigma^2}.$$

Note that since the actual synthesis step is considered post-processing under this setting, the privacy loss no longer depends on the number of synthetic records generated. In most cases, the privacy loss under this setting will be larger than the loss that accounts for both steps of the synthesis. The approach relying on post-processing will only lead to less privacy loss in cases, in which the number of synthetic records to be released is substantially larger than the number of records in the original data, specifically if $n^2(n+1) < mn_{syn}$

4.3 Plug-in Synthesis with Unknown Variance

Moving to the more realistic case of releasing synthetic data with unknown variance complicates matters, as the equivalence of the Gaussian mechanism and the Gaussian synthetic data generator only holds when the variance can be treated as public knowledge. Yet, the intuition that limiting the number of synthetic samples should provide privacy guarantees also holds for this case.

Under this scenario, the data curator using the plug-in approach will generate synthetic data based on the following model:

$$X_i^{syn} \sim \mathcal{N}(\bar{X}, s^2),$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ denotes the estimated variance based on the sample.

To analyze the privacy guarantees of synthetic data generated based on this model, it is important to note that the mean and the variance are sufficient statistics for the normal distribution. This implies that all that can be learned from the synthetic data are the sample mean and sample variance of the original data. Thus, it should be sufficient to limit the privacy leakage about these two statistics to quantify the privacy guarantees.

For the sample mean, the results from Section 4.1 still hold, with the slight modification that σ^2 is replaced by s^2. This also has the significant drawback that ρ is now data dependent and thus can no longer be released to the public without leaking information about the original data. Therefore, although formal privacy can still be achieved under this setting, the fundamental DP principle of full transparency about the protection measures is violated.

We compute the privacy loss random variable for the sampling variance to obtain bounds on the sampling variance and identify suitable bounds using $(\varepsilon, \delta) - DP$. The privacy loss random variable is the logged ratio of the densities of the output of interest computed over two neighboring datasets X and X', i.e.,

$$\mathcal{L} = \log \frac{P(M(X) = \tau)}{P(M(X') = \tau)},$$

where M is some randomized mechanism generating the output.

In our case, the output of interest is the sampling variance estimated from the synthetic data, and M is the mechanism that first generates synthetic data according to the model defined above and then estimates the sample variance based on the generated data.

Since the synthetic data are generated from a normal distribution, the sampling distribution of the estimated variance based on the synthetic data follows a Gamma distribution $\mathcal{G}(\alpha, \beta)$ with shape parameter $\alpha = \frac{n_{syn}-1}{2}$ and rate parameter $\beta = \frac{n_{syn}-1}{2s^2}$, where both $\alpha > 0$ and $\beta > 0$.

We need to analyze the privacy loss random variable to compute the privacy guarantees by looking at the maximum possible difference between $M(X)$ and $M(X')$. Given that the sensitivity of the sample variance is $\Delta_S = \frac{|2d|^2}{n}$, we know that the maximum difference is obtained if X has estimated sample variance s^2. The estimated sample variance of X' is $s^2 + \frac{|2d|^2}{n}$. Under this scenario, the privacy loss random variable is given as:

$$
\begin{aligned}
&\ln\left(\frac{\frac{{\frac{n_{syn}-1}{2s^2}}^{\frac{n_{syn}-1}{2}}}{\Gamma(\frac{n_{syn}-1}{2})} x^{\frac{n_{syn}-1}{2}-1} e^{-\frac{n_{syn}-1}{2s^2}x}}{\frac{{\frac{n_{syn}-1}{2s^2+\frac{|2d|^2}{n}}}^{\frac{n_{syn}-1}{2}}}{\Gamma(\frac{n_{syn}-1}{2})} x^{\frac{n_{syn}-1}{2}-1} e^{-\frac{n_{syn}-1}{2s^2+\frac{|2d|^2}{n}}x}}\right) = \\
&= \ln\left(\frac{{\frac{n_{syn}-1}{2s^2}}^{\frac{n_{syn}-1}{2}} e^{-\frac{n_{syn}-1}{2s^2}x}}{{\frac{n_{syn}-1}{2s^2+\frac{|2d|^2}{n}}}^{\frac{n_{syn}-1}{2}} e^{-\frac{n_{syn}-1}{2s^2+\frac{|2d|^2}{n}}x}}\right) = \\
&= \ln\left({\frac{n_{syn}-1}{2s^2}}^{\frac{n_{syn}-1}{2}} e^{-\frac{n_{syn}-1}{2s^2}x}\right) - \ln\left({\frac{n_{syn}-1}{2s^2+\frac{|2d|^2}{n}}}^{\frac{n_{syn}-1}{2}} e^{-\frac{n_{syn}-1}{2s^2+\frac{|2d|^2}{n}}x}\right) = \\
&= \frac{n_{syn}-1}{2}\ln\left(\frac{n_{syn}-1}{2s^2}\right) - \frac{n_{syn}-1}{2s^2}x - \frac{n_{syn}-1}{2}\ln\left(\frac{n_{syn}-1}{2s^2+\frac{|2d|^2}{n}}\right) + \frac{n_{syn}-1}{2s^2+\frac{|2d|^2}{n}}x = \\
&= \frac{n_{syn}-1}{2}\ln\left(\frac{2s^2+\frac{|2d|^2}{n}}{2s^2}\right) - \frac{(n_{syn}-1)\frac{|2d|^2}{n}}{2s^2(2s^2+\frac{|2d|^2}{n})}x
\end{aligned}
$$

To bound this ratio, we consider its absolute value:

$$
|\frac{n_{syn}-1}{2}\ln\left(\frac{2s^2+\frac{|2d|^2}{n}}{2s^2}\right) - \frac{(n_{syn}-1)\frac{|2d|^2}{n}}{2s^2(2s^2+\frac{|2d|^2}{n})}x|.
$$

Since the Gamma distribution is not symmetric, we need to look at both neighboring cases, i.e., the case where an observation changing from the minimum $-d$ to the maximum d increases the sample variance of X' by $\frac{|2d|^2}{n}$, and the case where the change decreases the variance by $\frac{|2d|^2}{n}$, separately.

Now to bound the privacy loss, we need to choose a sensible value for $\delta$and look up the $\frac{\delta}{2}$ and $1-\frac{\delta}{2}$ percentiles of the Gamma distribution $\mathcal{G}(\frac{n_{syn}-1}{2}, \frac{n_{syn}-1}{2s^2})$. The maximum then gives

$$\begin{aligned}&\epsilon \geq \\ &\max(|\frac{n_{syn}-1}{2}\ln\left(\frac{2s^2+\frac{|2d|^2}{n}}{2s^2}\right) - \frac{(n_{syn}-1)\frac{|2d|^2}{n}}{2s^2(2s^2+\frac{|2d|^2}{n})}x|, \\ &|\frac{n_{syn}-1}{2}\ln\left(\frac{2s^2-\frac{|2d|^2}{n}}{2s^2}\right) - \frac{(n_{syn}-1)\frac{|2d|^2}{n}}{2s^2(2s^2-\frac{|2d|^2}{n})}x|),\end{aligned}$$

evaluated at the values of the percentiles for the chosen δ of the Gamma distribution.

Getting the overall privacy guarantee of the synthesis with unknown variance requires to compose the guarantees for the mean and the variance. For example, by translating the guarantee for the mean to an $(\epsilon, \delta)-$DP guarantee and sequentially composing the two guarantees. That means, for two mechanisms satisfying $(\epsilon_1, \delta_1)-$DP and $(\epsilon_2, \delta_2)-$DP, the overall guarantee is $(\epsilon_1+\epsilon_2, \delta_1+\delta_2)$.

5 Discussion

In this paper, we showed that simple parametric synthetic data models can satisfy formal privacy guarantees. However, we also show that moving towards more realistic synthetic data models is complicated. We plan to extend this work by analyzing multivariate synthetic data models. For a multivariate normal synthetic data model with a known variance-covariance matrix, we expect similar results as described for the plug-in synthesis with known variance in section 4.1. We expect that moving to the multivariate normal case with unknown variance-covariance matrix is significantly harder than the univariate case. An open research question is whether and to what extent non-parametric synthetic data models can satisfy formal privacy guarantees.

References

Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '07, pp. 273–282, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936851. doi: 10.1145/1265530.1265569.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In Cynthia Dwork (ed.), *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pp. 609–618. ACM, 2008.

Alexander Theodorus Petrus Boudewijn, Andrea Filippo Ferraris, Daniele Panfilo, Vanessa Cocca, Sabrina Zinutti, Karel De Schepper, and Carlo Rossi Chauvenet. Privacy measurements in tabular synthetic data: State of the art and future research directions. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

Anne-Sophie Charest. How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Confidentiality*, 2(2), 2011.

Joerg Drechsler and Anna-Carolina Haensch. 30 years of synthetic data. *arXiv preprint arXiv:2304.02107*, 2023.

Jörg Drechsler. Improved variance estimation for fully synthetic datasets. *Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*, 2011a.

Jörg Drechsler. *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer Science & Business Media, 2011b.

Jörg Drechsler. Some clarifications regarding fully synthetic data. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings*, pp. 109–121. Springer, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4): 211–407, 2014.

Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pp. 265–284, 2006b.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006c.

Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Zhiwei Steven Wu. Dual query: Practical private query release for high dimensional data. In *Proceedings of the 31th International Conference on Machine Learning, {ICML} 2014, Beijing, China, 21-26 June 2014*, volume 32 of *{JMLR} Workshop and Conference Proceedings*, pp. 1170–1178. JMLR.org, 2014. URL http://proceedings.mlr.press/v32/gaboardi14.html.

Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st annual symposium on foundations of computer science*, pp. 61–70. IEEE, 2010.

Terrance Liu, Jingwu Tang, Giuseppe Vietri, and Steven Wu. Generating private synthetic data with genetic algorithms. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22009–22027. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23ag.html.

Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the {nist} contest: {A} scalable and general approach to differentially private synthetic data. *J. Priv. Confidentiality*, 11(3), 2021. doi: 10.29012/jpc.778. URL https://doi.org/10.29012/jpc.778.

Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. {AIM:} an adaptive and iterative mechanism for differentially private synthetic data. *Proc. {VLDB} Endow.*, 15(11):2599–2612, 2022. doi: 10.14778/3551793.3551817. URL https://www.vldb.org/pvldb/vol15/p2599-mckenna.pdf.

Marcel Neunhoeffer, Steven Wu, and Cynthia Dwork. Private post-gan boosting. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6isfR3JCbi.

Gillian M Raab, Beata Nowok, and Chris Dibben. Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3):67–97, 2016.

Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1):1, 2003.

Jerome P Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531, 2002.

Jerome P Reiter and Trivellore E Raghunathan. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471, 2007.

Jerome P Reiter, Quanli Wang, and Biyuan Zhang. Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality*, 6(1), 2014.

Donald B Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pp. 20–34. American Statistical Association Alexandria, VA, USA, 1978.

Donald B Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 1987.

Donald B Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

Jennifer Taub and Mark Elliot. The synthetic data challenge. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, The Netherlands*, 2019.

Jonathan Ullman and Salil Vadhan. Pcps and the hardness of generating private synthetic data. In Yuval Ishai (ed.), *Theory of Cryptography*, pp. 400–416, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19571-6.

Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. *arXiv preprint arXiv:2302.12580*, 2023.

Giuseppe Vietri, Grace Tian, Mark Bun, Thomas Steinke, and Zhiwei Steven Wu. New oracle-efficient algorithms for private synthetic data release. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9765–9774. PMLR, 2020. URL `http://proceedings.mlr.press/v119/vietri20b.html`.

Hao Wang, Shivchander Sudalairaj, John Henning, Kristjan Greenewald, and Akash Srivastava. Post-processing private synthetic data for improving utility on selected measures. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=neu9JlNweE`.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), oct 2017. ISSN 0362-5915. doi: 10.1145/3134428. URL `https://doi.org/10.1145/3134428`.