TRANSACTIONS ON DATA PRIVACY 15 (2022) 141–175

Model-based Differentially Private Data Synthesis and Statistical Inference in Multiple Synthetic Datasets

Fang Liu

Department of Applied and Computational Mathematics and Statistics

University of Notre Dame, Notre Dame, IN 46556, U.S.A.

E-mail: fang.liu.131@nd.edu

Received 26 April 2021; received in revised form 4 January 2022 and 22 April 2022; accepted 29 April 2022

Abstract. We propose the approach of model-based differentially private synthesis (modips) in the Bayesian framework for releasing individual-level surrogate/synthetic datasets with privacy guarantees. The modips technique integrates the concept of differential privacy into model-based data synthesis. We introduce several variants for the modips approach and multiple procedures for obtaining privacy-preserving posterior samples, a key step in the implementation of modips. The uncertainty from the sanitization and synthetic process in modips can be accounted for by releasing multiple synthetic datasets. We propose an inferential combination rule across multiple sets to generate final valid inferences. We run several empirical studies to demonstrate the application of modips and examine the impacts of the number of synthetic sets and the privacy budget allocation schemes on statistical inference based on synthetic data.

Keywords. (Bayesian) sufficient statistics, budget allocation, differentially private posterior sampling, inference, multiple synthesis, sanitization, variance combination

1 Introduction

1.1 Background and Motivation

Data synthesis is a statistical disclosure limitation technique that releases pseudo individual-level data for research and public use. Both parametric and nonparametric Bayesian and frequentist approaches have been proposed for data synthesis [1, 2, 3, 4, 5, 6, 7, 8]. To capture the uncertainty introduced during the synthesis process, multiple sets of synthetic data are often released. Inferential methods that combine information from multiple synthetic datasets to yield valid inference are available [9, 10]. A long-standing research problem in statistical disclosure limitation is the lack of a universally applicable and robust measure of disclosure risk in released data [11, 12]. Many existing disclosure risk assessment approaches rely on strong and ad-hoc assumptions regarding the background knowledge and behaviors of data intruders [13, 14, 15, 16, 17, 18, 19].

*The work was supported by the NSF Grants IIS 1546373 and CNS 1717417

Differential privacy (DP) has gained enormous popularity since its debut in 2006 [20]. DP formalizes privacy in mathematical terms without making assumptions about data intruders and has nice properties such as privacy loss composition and immunity to post processing for the information sanitized through a DP randomized algorithm. DP has spurred a great amount of work in developing differentially private randomized mechanisms to release statistics in general settings as well as for specific types of queries or statistical analyses. The Laplace mechanism [20], the exponential mechanism [21], and the Gaussian mechanism [22, 23] are common differentially private sanitizers for general purposes. Differentially private versions of various statistical analyses are also available, such as point estimators [24, 25], principle components analysis [26], linear and penalized regression [27, 28], model selection [29], release of functions [30], the χ^2 test in genome-wide association studies [31], and deep learning [32, 33, 34], among others.

In addition to sanitization of aggregate statistics for release, there is also differentially private synthesis of individual-level data (dips) that releases surrogate privacy-preserving datasets to original data. An obvious advantage of dips over query-based sanitization and information release is that users of surrogate data can run analyses of their own as if they had the original data, whereas query-based data sharing allows only a certain amount or sometimes only certain types of queries to be answered before a pre-set privacy budget runs outs.

1.2 Related Work

Early dips approaches are often model-free (without assuming data belong to any particular parametric probability distribution during sanitization or synthesis) and focus on categorical data synthesis or require some discretization for numerical attributes. In the framework of categorical data synthesis via model-free approaches, Barak et al. [35] generated synthetic data via the Fourier transformation and linear programming in low-order contingency tables; Blum et al. [36] discussed the possibility of dips from the perspective of the learning theory in a discretized domain; Hardt et al. [37] developed the iterative multiplicative weights exponential mechanism algorithm via "matching" on linear queries; Bowen et al. [38] propose the STEPS procedure that partitions data by attributes according to a practical or statistical importance measure and synthesize the data from a constructed hierarchical attribute tree; Eugenio and Liu [39] propose the CIPHER procedure to construct differentially private empirical distributions from a set of low-order marginals through solving linear equations with l_2 regularization.

For model-based categorical data synthesis, the multinomial-Dirichlet synthesizer [40] designs a prior for cell proportions to achieve DP in the Bayesian framework. The approach was applied to synthesize the US commuting data in Machanavajjhala et al. [41] and its inferential properties were studied in Charest [42]. McClure and Reiter [43] implemented a similar concept but with a different prior to synthesize univariate binary data via the binomial-beta model. Zhang et al. [44] proposed PrivBayes to release high-dimensional categorical data from Bayesian networks.[1]

For numerical data synthesis, Wasserman and Zhou [45] proposed several paradigms to construct differentially private empirical distributions and examined the rate at which the differentially private distributions converge to the true distribution. Hall et al. [30] proposed a differentially private kernel density estimator. In both works, synthetic data can

[1] Bayesian networks are a probabilistic graphical model and does not involve Bayesian modeling or inference; thus PrivBayes conceptually differs from the Bayesian dips framework that we focus on.

be sampled and released from the privacy-preserving distributions; and both are subject to the curse of dimensionality and have trouble of constructing useful differentially private distributions in high dimensional settings.

For general data synthesis, Li et al. [46] proposed DPCopula to generate synthetic data from differentially private copulas for multi-dimensional data. Variational inference (VI) is a state-of-art framework for analytical approximation of hard-to-sample distributions [47]. Some recent work employs neural networks (NN), including generative adversarial networks [48], to release synthetic data from differentially private generative models [49, 50, 51, 52]. One advantage of NN-based approaches is that they rely on machine learning to develop robust generative models and do not make distributional or strong relational assumptions on sample data. On the other hand, NNs are often subject to overfitting and generalization and robustness of a trained NN often need large training data. The recent normalizing flow VI approach [53] is a powerful approach for approximating high-dimensional, irregular, or multi-modal distributions. Work that integrates DP in VI through normalizing flows [54] also exists by achieving DP can be achieved during iterative optimization of the evidence lower bound. The approach at the same size requires large sample sizes to generate useful synthetic samples in a differentially private manner, especially when employed normalizing flows have many parameters.

There also exist dips approaches for specific data types. Quick [55] proposed an approach to generate private synthetic data via the Poisson-gamma model and applied the approach to disease mapping. There are also dips approaches for specific types of data such as graphs [56, 57, 58, 59]; mobility data from GPS trajectories [60, 61], and graph data based on exponential random-graph models in social networks [62, 63]. Dips is also a topic for doctoral dissertation research in recent years [64, 65, 66, 67].

For statistical inference based on differentially private synthetic data, Smith [24] proposed the "subsample-and-aggregate" technique to obtain differentially private α-Winsorized means over subsamples. The approach requires the estimates from subsamples are i.i.d. from an approximately Gaussian distribution with a bounded third moment, for sufficiently large n. Charest [42] explicitly modelled the differentially private mechanism in the Bayesian inference of synthesized univariate binary data; Karwa and Slavković [68] treated the Laplace mechanism as a measurement error on the sufficient statistics of the β-model for random graphs and established the conditions for the existence of the private maximum likelihood estimator for the degree sequence in graphs that achieves the same convergence rate as non-private estimators. Karwa et al. [62] modelled the sanitization mechanism when analyzing synthetic social networks in the framework of exponential family random-graph models. All the work focuses on either just point estimators, a specific type of analysis, or a specific type of data; if uncertainty quantification is involved, then the sanitization mechanisms would be modeled explicitly during the inferential process, meaning that data users would need to be provided with full details of the sanitization mechanism so to model and incorporate the mechanism in their data analysis procedures. This can be challenging especially for users who are not familiar with DP and sanitization mechanisms; even for users who are familiar of DP, incorporating a randomized mechanism in a commonly used analysis procedure may bring analytical and computational challenges.

1.3 Our Contributions

We develop the model-based differential private synthesis (modips) procedure in the Bayesian setting, aiming for population-level information preservation in the synthetic data. Differ-

ent from the existing Bayesian dips approaches such as the multinomial-Dirichlet and the beta-binomial synthesizers, modips does not achieve DP through prior specification, but rather through sanitizing the posterior distribution. modips is a general approach that can handle all data types (numerical, categorical, discrete, graphs) where appropriate Bayesian models can be constructed. Our main contributions are summarized as follows.

- We achieve DP in the modips procedure through sanitizing posterior distributions of model parameters. We propose several procedures for obtaining privacy-preserving posterior samples. The achieved privacy guarantees are preserved in the subsequent synthesis step and in released surrogate datasets with the immunity to post-processing property of DP.
- We propose releasing multiple synthetic datasets to account for uncertainty of sanitization and synthesis in inference from sanitized data. We provide an inferential combination rule across multiple synthetic datasets and examine its asymptotic properties. We run empirical studies to examine the impact of the number of multiple synthetic sets on inference.
- We propose the concepts of communal sanitization and individualized sanitization and study the effect of privacy budget allocation on inference in the individualized sanitization via empirical studies.

The rest of the paper is organized as follows. Sec. 2 overviews the basic concepts of DP, some commonly used differentially private mechanisms, and the traditional non-DP multiple synthesis procedure. Sec. 3 introduces the modips approach and several procedures for obtaining privacy-preserving posterior samples, a key step in modips. Sec. 4 proposes an approach to combine inferential results from multiply synthetic datasets for final inference. Sec. 5 runs simulation studies to illustrate the application of modips, validate the inferential combination rule, examine the effects of the number of released datasets and the impact of privacy budget allocation on inferences; it also summarizes the results from published work that implements the modips approach or uses our inferential combination rule. The paper concludes in Sec. 6 with final remarks and some topics for future work.

2 Preliminaries

In this section, we present some definitions and concepts in differential privacy (DP) (Sec. 2.1) and data synthesis (Sec. 2.2) that are referred to or used in later sections of th paper.

2.1 Differential Privacy

DP is a mathematically formulated formal privacy notion and does not make ad-hoc assumptions on the background knowledge or behaviors of data intruders.

Definition 1. [69, 20] A randomized algorithm $\mathcal{R}$ is ϵ-differentially private if, for all datasets $(\mathbf{x}, \mathbf{x}')$ that differ in one individual and all possible subsets Q to the output range of statistic $\mathbf{s}$ from $\mathcal{R}$, $\left|\log\left(\frac{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x})))\in Q)}{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}')))\in Q)}\right)\right| \leq \epsilon$ for $\epsilon > 0$.

The DP definition implies that the probabilities of obtaining the same statistic from $\mathbf{x}$ and $\mathbf{x}'$ after the sanitization are similar – the ratio between the two probabilities falls within $[e^{-\epsilon}, e^{\epsilon}]$. ϵ is often referred as the privacy loss or budget parameter; the smaller ϵ is, the more privacy protection will be executed on the individuals in the data through $\mathcal{R}$. In layman's

terms, DP implies the chance that an individual in a dataset can be identified based on released sanitized results is low since the results are about the same with or without that individual.

In what follows, we use $d(\mathbf{x}, \mathbf{x}') = 1$ to denote two datasets $\mathbf{x}, \mathbf{x}'$ differing by one individual, which is often defined in two ways. First, $\mathbf{x}, \mathbf{x}'$ have the same sample size, but one and only one record differs in at least one attribute; a substitution would make $\mathbf{x}, \mathbf{x}'$ identical(aka *bounded DP* in Kifer and Machanavajjhala [70]). In the second definition, one dataset has one more record than the other, so the sample sizes differ by 1, and deletion (or insertion) of one record would make $\mathbf{x}, \mathbf{x}'$ identical (aka *unbounded DP*).

In practice, satisfying the pure DP in Definition 1 may lead to much perturbation in released information. To reduce the amount of sanitization, softer versions of DP have been developed, such as the (ϵ, δ)-approximate DP (aDP) [71], the (ϵ, δ)-probabilistic DP (pDP) [41], (ϵ, δ)-random DP (rDP) [72], (ϵ, τ)-concentrated DP (CDP)[73] and zero concentrated DP (zCDP) [74], Rényi DP [75], and Gaussian DP [76]. In some relaxed versions of DP, extra parameters are employed to characterize the amount of relaxation on top of the privacy budget ϵ and include the pure DP as a special case. For example, (ϵ, δ)-aDP and (ϵ, δ)-pDP reduce to ϵ-DP when $\delta = 0$, and (α, ϵ)-Rényi DP reduces to ϵ-DP when $\alpha = \infty$. The relaxed versions are also related; for example, zCDP and Gaussian DP can be converted to aDP, with ϵ as a functional of δ.

Many differentially private mechanisms have been proposed to sanitize statistics, among which the Laplace mechanism, the Gaussian mechanism, and the exponential mechanism are three popular sanitizers for general purposes, the definitions of which are given below.

Definition 2. [20] Let $\mathbf{s} = (s_1, \ldots, s_r)$ be a r-dimensional statistic. The *Laplace mechanism* of ϵ-DP releases sanitized version $\mathbf{s}^*$ of $\mathbf{s}$ via $s_i^* = s_i + e_i$, where $e_i \overset{\text{ind}}{\sim} \text{Laplace}\left(0, \Delta_1 \epsilon^{-1}\right)$ for $i = 1, \ldots, r$ and Δ_1 is the global sensitivity of $\mathbf{s}$.

Δ_1 is a special case of the l_p *global sensitivity* (GS) $\Delta_p = \max_{\mathbf{x},\mathbf{x}', d(\mathbf{x},\mathbf{x}')=1} \|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{x}')\|_p$ when $p = 1$ [23]. The sensitivity is "global" since it is defined for all possible datasets and all pairs of two neighboring datasets. The larger the GS is, the larger the privacy risk is from releasing $\mathbf{s}$, and the more perturbation is needed for $\mathbf{s}$ to offset the privacy risk. This is also reflected in the variance of the Laplace distribution $2\left(\delta_1 \epsilon^{-1}\right)^2$: the larger δ_1 or the smaller ϵ is, the more spread the distribution of $\mathbf{s}^*$ is, and the more likely that extreme $\mathbf{s}^*$ values that are far away from $\mathbf{s}$ will be released.

Definition 3. [23, 77] Let $\mathbf{s} = (s_1, \ldots, s_r)$ be a r-dimensional statistic. The *Gaussian mechanism* sanitizes $\mathbf{s}$ as in $s_i^* = s_i + e_i$, where $e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, r$ with $\sigma \geq c\Delta_2/\epsilon$ (Δ_2 is the l_2 GS of $\mathbf{s}$) for $\epsilon < 1$ and $c^2 > 2\log(1.25/\delta)$ in the case of (ϵ, δ)-aDP and $\sigma \geq (2\epsilon)^{-1}\Delta_2(\sqrt{(\Phi^{-1}(\delta/2))^2 + 2\epsilon} - \Phi^{-1}(\delta/2))$ in the case of (ϵ, δ)-pDP, where Φ^{-1} is the inverse CDF of the standard normal distribution.

Definition 4. [21] Let $\mathbf{s} = (s_1, \ldots, s_r)$ be a r-dimensional statistic and $\mathcal{S}$ be the set containing all possible sanitized outputs. The *exponential mechanism* of ϵ-DP releases $\mathbf{s}^*$ from $p(\mathbf{s}^*) \propto \exp(u(\mathbf{s}^*)\epsilon/(2\Delta_u))$, where $u(\mathbf{s}^*)$ is the utility score of $\mathbf{s}^*$ and $\delta_u = \max_{\mathbf{x},\mathbf{x}', \delta(\mathbf{x},\mathbf{x}')=1} |u(\mathbf{s}^*|\mathbf{x}) - u(\mathbf{s}^*|\mathbf{x}')|$ is the maximum change (sensitivity) in score u for all pairs of neighboring datasets $\mathbf{x}$ and $\mathbf{x}'$.

DP has some nice properties that other privacy notions do not possess, and they are key for successful applications of DP in practical problems. One such property is that released

sanitized results through a DP mechanism are immune to post-processing in the sense that they do not leak more private information if further processed after release (as long as there is no access to the original data from which the results are calculated). Another property is that the total privacy loss from applying multiple differentially private mechanisms to the same dataset is closed[2] (combining the released results from multiple differentially private algorithms is still differentially private). There are two basic composition principles in DP: *parallel composition* and *sequential composition* [21]. If mechanism $\mathcal{R}_j$ is ϵ_j-DP for $j = 1, \ldots, r$ and each is applied on disjoint subsets D_j of a dataset D, then $\prod_j \mathcal{R}_j(\mathbf{y} \cap D_j)$ is $\max(\epsilon_j)$-DP per the parallel composition; if $\mathcal{R}_j$ is applied to the same dataset D, then $\prod_j \mathcal{R}_j(\mathbf{y})$ is $(\sum_j \epsilon_j)$-DP per the sequential composition. Besides the basic compositions, there is also advanced composition [79] in aDP that provides tighter privacy loss bounds. Some relaxed DP notions (e.g., CDP, zCDP, Rényi DP, Gaussian DP) allow exact privacy loss composition or easily track the privacy loss over multiple releases. In summary, immunity to post processing and composition allow users of DP to design sophisticated differentially private algorithms for complicated problems (e.g., releasing intermediate gradients in gradient-based iterative optimization).

2.2 Multiple Synthesis

Each surrogate dataset to an original dataset $\mathbf{x}$ generated through multiple synthesis (MS) have the same structure and format as $\mathbf{x}$[3] but contain pseudo-individuals. Synthetic data can be generated in several approaches. Depending on whether a parametric statistical model for synthesis, a MS approach can be model-free or model-based; depending on the observed data source that synthesis is based on, a MS approach can be population synthesis and sampling [1, 9] or sample synthesis [80]; by the percentage of the synthetic component in a released dataset, data synthesis can be partial synthesis or full synthesis [2, 3]. Cross-labelling of a MS approach is possible, such as "model-based full sample synthesis".

Fig. 1 depicts a Bayesian MS procedure sample full synthesis, the focus of in this paper. $f(\mathbf{x}|\boldsymbol{\theta})$ is the model assumed on the original data $\mathbf{x}$, which also represents the likelihood function, $f(\boldsymbol{\theta})$ is the prior and $f(\boldsymbol{\theta}|\mathbf{x})$ is the posterior distribution of the model parameters $\boldsymbol{\theta}$. The MS procedure outputs $m \geq 1$ sets of synthetic surrogate data from the posterior predictive distribution $f(\tilde{\mathbf{x}}|\mathbf{x})$ by first sampling $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta}|\mathbf{x})$ and then $\tilde{\mathbf{x}}$ from $f(\mathbf{x}|\boldsymbol{\theta})$ given the drawn $\boldsymbol{\theta}$ in the previous step, since $f(\tilde{\mathbf{x}}|\mathbf{x}) = \int f(\tilde{\mathbf{x}}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$. The size of each surrogate data set is kept to be the same as the original sample size n.

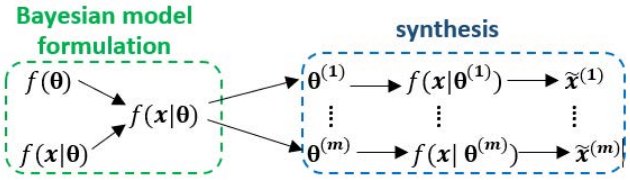


Figure 1: The traditional MS procedure

The MS procedure in Fig. 1 does not use explicit randomized algorithms to sanitize the

[2] See Kifer and Lin [78] for examples on when the property does not hold in some relaxed form of DP.

[3] For example, if $\mathbf{x}$ is tabular data, then $\tilde{\mathbf{x}}$ would have the same dimensionality as $\mathbf{x}$ with the same numbers of samples and attributes; if $\mathbf{x}$ is graph data, the number of nodes in $\tilde{\mathbf{x}}$ is the same as in $\mathbf{x}$.

information in $\mathbf{x}$. The argument for privacy guarantees of this approach is rather heuristic as no individuals in the surrogate data correspond to any real persons. There exists work that connects posterior sampling, which has inherent randomness, with DP. Wang et al. [81] proved that the privacy loss for releasing one posterior sample of $\boldsymbol{\theta}$ given any prior is $4B$, where B is the upper bound of the log-likelihood $\log(l(\boldsymbol{\theta}|\mathbf{x})$. Dimitrakakis et al. [82] show that if the change in the log-likelihood between two neighboring datasets $(\mathbf{x}, \mathbf{x}')$ is bounded by a constant C, releasing one posterior sample of $\boldsymbol{\theta}$ is $2C$-differentially private. For MS, since $m \geq 1$ posterior samples are released, the overall privacy loss becomes $4mB$ and $2mC$, respectively, per the sequential composition. If the bounds B and C are large, the privacy loss that relies on the inherent randomness of posterior sampling can be too large to provide sufficient privacy guarantees.

3 Model-based Differentially Private Data Synthesis (modips)

3.1 The modips Procedure

Fig. 2 illustrates the modips procedure. Compared to the traditional MS procedure in Fig. 1, the modips procedure incorporates the formal DP notion via an additional sanitization step prior to synthesis. Instead of sampling from the posterior distribution of $f(\boldsymbol{\theta}|\mathbf{x})$, the modips samples from its sanitized version $f^*(\boldsymbol{\theta}|\mathbf{x})$, since the information in the original data x is passed to the synthetic data $\tilde{\mathbf{x}}$ through the samples of $\boldsymbol{\theta}$ (refer to Fig. 1), if $\boldsymbol{\theta}$ is privacy-preserving, so are the data $\tilde{\mathbf{x}}$ sampled given $\boldsymbol{\theta}$. We regard the original sample size n as public knowledge and keep the size of each surrogate data set at n in the modips procedure.[4]

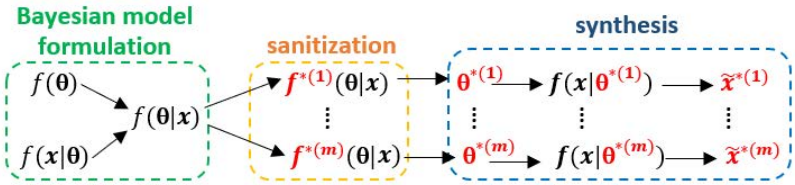


*Figure 2: The modips Procedure (superscript * represents "sanitized")*

Since the modips procedure is model-based, its users need to first specify a model for their data. In the Bayesian framework, this would involve the specification of prior $f(\boldsymbol{\theta})$ and likelihood $f(\mathbf{x}|\boldsymbol{\theta})$. If users have prior or public knowledge on what $f(\mathbf{x}|\boldsymbol{\theta})$ would fit the data $\mathbf{x}$ well, the likelihood can be prespecified without using any specific values in $\mathbf{x}$ and thus costing no privacy; otherwise, the modips procedure starts with a model selection step given $\mathbf{x}$, consuming a portion of privacy. If uses have a candidate set with a finite number of models, they may apply the exponential mechanism (Definition 4) to choose a model with a utility function u that measures the goodness of fit of each model. For example, if all attributes in $\mathbf{x}$ are categorical one may apply a Bayesian network to the data to uncover

[4] A variant to the modips procedure in Fig. 2 is the *nested modips* procedure, the details of which is given in the Appendix. Since the output volume from the nested modips is t folds of that of the standard modips procedure and the analysis of the synthetic data is also more complex with the hierarchical data structure, we suggest not employing the nested modips unless there is an absolute need or interest to separately quantify due to sanitization vs. synthesis (see Sec. 4 for detail).

the dependency structure among the attributes with the Bayesian information criterion as the utility function [83], or using the PrivBayes framework which proposes a novel utility function with low sensitivity [44]. Once a Bayesian network is selected, one may form a corresponding joint distribution of the attributes so to construct the likelihood function for the sanitization and synthesis in subsequent steps. If the attributes are a mixture of numerical and categorical, one may apply sequential regression models to model the dependency structure among the attributes similar to that used in multiple imputation [84]. Users may also use any model selection procedure deemed appropriate for their data, but should always keep in mind this step costs privacy and will need to design a differentially private mechanism to choose a model.

Algorithm 1 presents the steps of the modips procedure, starting with model selection. One can also incorporate the model selection step as part of the "for" loop, so each synthetic dataset is from a different model. There are pros and cons to this approach. On the one hand,

Algorithm 1: The modips Procedure

input : number of surrogate datasets m, privacy budget ϵ, original data $\mathbf{x}$
output: surrogate datasets: $\tilde{\mathbf{x}}^{*(1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}$

1 Select a model with a privacy budget portion ϵ_0 for data $\mathbf{x}$ if no model is prespecified;
2 **for** $i = 1, \ldots, m$ **do**
3 Obtain a posterior sample $\boldsymbol{\theta}^{*(j)}$ from $f(\boldsymbol{\theta}|\mathbf{x})$ corresponding to the selected Bayesian model via a differentially private mechanism with privacy budget $(\epsilon - \epsilon_0)/m$;
4 Draw $\tilde{\mathbf{x}}^{*(j)}$ from $f\left(\mathbf{x}|\boldsymbol{\theta}^{*(j)}\right)$.
5 **end**

statistical inference based on multiple synthetic datasets ought to be more robust as it is implicitly averaged over multiple synthesis models; on the other hand, the inference is subject to more variability with the employment of multiple synthesis models, especially considering that privacy budget ϵ_0 allocated to model selection is further split into m portions, potentially making model selection less meaningful from a data utility perspective.

Each synthetic set is sanitized with $1/m$ of the overall privacy budget per the sequential composition. Since the amount of noise increases with decreasing privacy budget, one set of synthetic data with $m > 1$ is noisier than that with $m = 1$; however, the total amount of information collectively across the $m > 1$ sets may still be compatible with that for $m = 1$. More importantly, releasing multiple sets provides an effective and convenient way to quantify the uncertainty and randomness introduced during the sanitization and synthesis. This enables valid inferences in the released data when no other sources or approaches are available to data users to quantify the uncertainty (see Sec. 4 for details).

Proposition 1. The modips procedure in Algorithm 1 is ϵ-differentially private.

The proof of Proposition 1 is provided in the Appendix. The proof suggests that the step of drawing $\tilde{\mathbf{x}}^*$ does not incur any additional privacy cost as it can be regarded as post-processing of the already-sanitized $\boldsymbol{\theta}^*$.

Algorithm 1 is also applicable to relaxed versions of DP. For example, if (ϵ, δ)-aDP is employed, the data curator would split both ϵ and δ between the model selection and synthesis steps and across the m syntheses; the conclusion in Proposition 1 also applies to (ϵ, δ)-aDP as it is immune to post-processing and closed under composition.

3.2 Differentially Private Posterior Sampling

We present a few approaches to obtaining sanitized samples from $f(\boldsymbol{\theta}|\mathbf{x})$ (line 3 of Algorithm 1), including 1) direct sanitization, 2) sanitization of sufficient statistics, and 3) sanitization of approximate distribution. We introduce each approach in detail below.

1) Direct Sanitization. One may directly sanitize the posterior distribution function $f(\boldsymbol{\theta}|\mathbf{x})$ via a DP mechanism. Though this sounds straightforward conceptually, it can be difficult to implement practically. One reason is that $f(\boldsymbol{\theta}|\mathbf{x})$ is often known up to a constant only in many practical problems; that is, $f(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})$ and the normalizing constant $f(\mathbf{x})$ might not have a close-form expression. This matters in the framework of DP as $f(\mathbf{x})$ is a function of data $\mathbf{x}$ – the target for protection. Even if $f(\boldsymbol{\theta}|\mathbf{x})$ has a closed form, sanitizing $f(\boldsymbol{\theta}|\mathbf{x})$ is not as simple as $f^*(\boldsymbol{\theta}|\mathbf{x}) = f(\boldsymbol{\theta}|\mathbf{x}) + e$, say $e \sim \text{Lap}(0, \Delta_f \epsilon^{-1})$, as $f^*(\boldsymbol{\theta}|\mathbf{x})$ might no longer integrate to 1 or be a proper density function.

2) Sanitization of Sufficient Statistics (SSS). When $f(\boldsymbol{\theta}|\mathbf{x})$ can be reformulated as $f(\boldsymbol{\theta}|\mathbf{s})$, where $\mathbf{s}$ is a scalar or multi-dimensional Bayesian sufficient statistic (that is, the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{x}$ only depends on $\mathbf{s}$), one can sanitize $\mathbf{s}$ to achieve DP guarantees for $f(\boldsymbol{\theta}|\mathbf{s})$ and thus for $f(\boldsymbol{\theta}|\mathbf{x})$ and $f(\tilde{\mathbf{x}}|\boldsymbol{\theta})$. We expect that SSS is easier to implement compared to the direct santization as $\mathbf{s}$ is of finite dimension and there are many existing mechanisms for sanitizing statistics. We refer to this variant of the modips procedure as modips.SSS. The formal privacy guarantee of modips.SSS is provided in Proposition 2 and its proof is given in the Appendix.

Proposition 2. The modips.SSS procedure satisfies ϵ-DP.

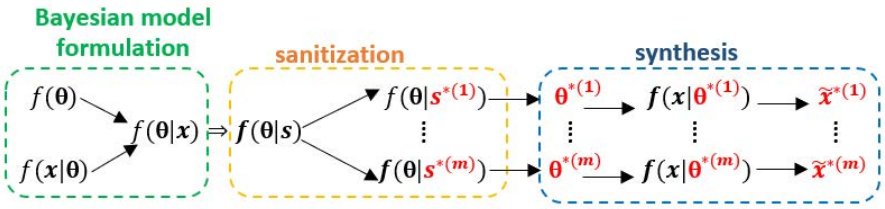


Figure 3: The modips.SSS Procedure

Fig. 3 presents the modips.SSS procedure. The steps for implementing modips.SSS, are similar to Algorithm 1, except for line 3, where $\mathbf{s}$ is sanitized with privacy budget $(\epsilon - \epsilon_0)/m$ (e.g, via the Laplace mechanism) to obtain $\mathbf{s}^{*(j)}$, and then a posterior sample $\boldsymbol{\theta}^{*(j)}$ is drawn from $f(\boldsymbol{\theta}|\mathbf{s}^{*(j)})$. Identification of sufficient statistic $\mathbf{s}$ in a Bayesian model is critical for modips.SSS. Generally speaking, classical sufficiency implies Bayesian sufficiency.[5]

3) Sanitization of Approximate Distribution. The modips.SSS procedure sanitizes sufficient statistics $\mathbf{s}$, the GS of which can be challenging to obtain, not to mention that not all models can be expressed using sufficient statistics only. As an alternative, we may sanitize an approximation to $f(\boldsymbol{\theta}|\mathbf{x})$, say $g(\boldsymbol{\theta}|\mathbf{x})$, such as through the VI approach as introduced in Sec. 1. If analytically tractable distributions $g(\boldsymbol{\theta}|\mathbf{x})$ (e.g., multivariate normal distributions or exponential family distributions) with easily identifiable sufficient statistics, can approximate $f(\boldsymbol{\theta}|\mathbf{x})$ well, we can apply modips.SSS to $g(\boldsymbol{\theta}|\mathbf{x})$ instead. However, in many cases,

[5] There are examples of Bayesian sufficient statistics which are not classically sufficient but those are unusual situations [85, 86, 87].

$f(\boldsymbol{\theta}|\mathbf{x})$ is more complicated than can be well approximated by a "nice" distribution. For this reason, we propose three simple numerical approaches to achieve DP when releasing samples from a posterior distribution, which are Sanitization of Discretized Density function (SiDD), SiDD with Monte Carlo (SiDD.MC), and Sanitization of Posterior Histogram Counts (SiPHiC).

SiDD: Algorithm 2 presents the SiDD procedure. Though $f(\boldsymbol{\theta}|\mathbf{x})$ is listed as the input in Algorithm 2, it does not need to be normalized. The algorithm starts with cutting $f(\boldsymbol{\theta}|\mathbf{x})$ into b bins, which can be challenging in high-dimensional settings. A simpler approach,

Algorithm 2: The SiDD procedure

input : posterior distribution $f(\boldsymbol{\theta}|\mathbf{x})$, number of bins b, privacy budget ϵ.
output: a privacy-preserving posterior sample $\boldsymbol{\theta}^*$

1 Discretize $f(\boldsymbol{\theta}|\mathbf{x})$ into b bins $\{\mathcal{B}_i\}_{i=1,\dots,B}$;
2 Select a bin via the exponential mechanism of ϵ-DP: $\Pr(\mathcal{B}_i) \propto \exp(u_i\epsilon/(2\Delta_u))$, where $u_i = \log\big(\int \mathbb{1}(\theta \in \mathcal{B}_i) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}\big)$; denote the selected bin by $\mathcal{B}_k$ and the index for the marginal bin in the j-th dimension in bin $\mathcal{B}_k$ by $j(k)$ with end points $(c_{j,j(k)-1}, c_{j,j(k)})$;
3 Draw a sample $\boldsymbol{\theta}^*$ from uniform$((c_{1,1(k)-1}, c_{1,1(k)}) \times \cdots \times (c_{p,p(k)-1}, c_{p,p(k)}))$.

though likely less optimal, is to discretize each dimension separately and the total number bins across p dimensions is $B = \prod_{j=1}^{p} B_j$, where B_j is the number of bins in the j^{th} dimension. The step that costs privacy is the bin selection from the discretized distribution (line 2) via the exponential mechanism, with the utility function u being the logarithm of a bin volume. In other words, the probability of a bin $\mathcal{B}_i$ being selected is proportional to the volume of $\mathcal{B}_i$ with a constant exponent $(\epsilon/(2\Delta_u)$ that is calibrated to the privacy budget and the utility sensitivity. Other utility functions can also be used, as long as there is a good rationale. The integral $\int \mathbb{1}(\theta \in \mathcal{B}_i) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$ in u can be evaluated via numerical approaches (e.g. MC approaches). One computationally efficient approach is as follows. We set $\int \mathbb{1}(\theta \in \mathcal{B}_i) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = \bar{f}_i(\boldsymbol{\theta}|\mathbf{x}) V_i$ per the mean value theorem, where V_i is the volume of the hyper-cube defined by the cut points $((c_{1,1(i)-1}, c_{1,1(i)}) \times \cdots \times (c_{p,p(i)-1}, c_{p,p(i)}))$ surrounding $\mathcal{B}_i$ and $\bar{f}_i(\boldsymbol{\theta}|\mathbf{x})$ is the average of the density values evaluated at a large number of uniformly distributed $\boldsymbol{\theta}$ points within $\mathcal{B}_i$ or can be simply set at $f((c_{1,1(i)-1} + c_{1,1(i)})/2, \cdots, (c_{p,p(i)-1} + c_{p,p(i)})/2|\mathbf{x})$. Though the latter may lead to some accuracy loss, since the goal is not to precisely estimate $\int \mathbb{1}(\boldsymbol{\theta} \in \mathcal{B}_i) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$, but rather to define a reasonable utility function u to select a bin via the exponential mechanism, the rough estimate would not cause material harm. Based on $\bar{f}_i(\boldsymbol{\theta}|\mathbf{x}) V_i$, we can calculate the sensitivity of the utility u, the result of which is given in Proposition 3 and its proof is provided in the Appendix.

Proposition 3. Let $A \triangleq \sup_{\mathbf{x},\boldsymbol{\theta}} |\log(f(\boldsymbol{\theta}|\mathbf{x}))|$.[6] The GS of the utility function $u = \log\big(\int \mathbb{1}(\theta \in \mathcal{B}_i) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}\big)$ in Algorithm 2 is $\Delta_u = \max_{\mathcal{B},\boldsymbol{\theta},d(\mathbf{x},\mathbf{x}')=1} |u_i(\mathbf{x}) - u(\mathbf{x}')| = 2A$.

SiDD.MC: The SiDD procedure can also be implemented in an MC manner by sanitizing

[6] A reviewer asked about the rationale behind the usage of "log" and "supreme". The reason why supreme is used is because the upper bound needs to be independent of the data to prevent privacy loss. Supreme here refers to the smallest upper global bound to the log function. Alternatives, such as maximum, cannot be used as the maximum of $|\log(f(\boldsymbol{\theta}|\mathbf{x}))|$ as they depend on the data $\mathbf{x}$ and thus leak privacy; in other words, the bound A would be $A(\mathbf{x})$ if the maximum were used. Regarding "log", since Algorithm 2 uses a log-scale u, its sensitivity relates to the supremum of log-density $\log(f(\boldsymbol{\theta}|\mathbf{x}))$; see the proof in the Appendix for the technical details. If other utility functions are used, the corresponding sensitivities would mostly likely take a different form. The same rationales above apply to Proposition 4.

a histogram constructed from a set of posterior samples from $f(\boldsymbol{\theta}|\mathbf{x})$, referred to as the SiDD.MC procedure. The steps are presented in Algorithm 4. The GS for $\log(N_i^*/N)$ in line 2 of Algorithm 3 is given in Proposition 4; the proof can be found in the Appendix.

Algorithm 3: The SiDD.MC procedure

input : N samples of $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta}|\mathbf{x})$, privacy budget ϵ.
output: a privacy-preserving posterior sample $\boldsymbol{\theta}^*$.
1 Construct a histogram estimator of $f(\boldsymbol{\theta}|\mathbf{x})$ given the N samples of $\boldsymbol{\theta}$; denote the number of histogram bins by B;
2 Obtain bin count N_i for $i = 1, \ldots, B$ and sanitize $\log(N_i/N)$ via the Laplace mechanism: $\log(N_i^*/N) \sim \text{Lap}(\log(N_i/N), \Delta_i/\epsilon)$ to obtain a sanitized histogram;
3 Normalize N_i^* for $i = 1, \ldots, B$ so that $\sum_{i=1}^{B} N_i^* = 1$;
4 Draw a sample $\boldsymbol{\theta}^*$ from the histogram with sanitized counts N_i^* for $i = 1, \ldots, B$.

Proposition 4. Let $A \triangleq \sup_{\mathbf{x},\boldsymbol{\theta}} |\log(f(\boldsymbol{\theta}|\mathbf{x}))|$, B be the total number of bins in the histogram constructed from N posterior samples of $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta}|\mathbf{x})$, and N_i be the count in bin $\mathcal{B}_i$ for $i = 1, \ldots, B$. Then the GS for $\log(N_i^*/N)$ is $\Delta_i = 2A$.

SiDD.MC is asymptotically equivalent to SiDD when N is large and the discretization cut-points are the same between the two. If N is small, constructed histograms in SiDD.MC may deviate significantly from discretized $f(\boldsymbol{\theta}|\mathbf{x})$ in SiDD, leading to a loss of accuracy and supposedly some privacy protection. We present the Sanitization of Posterior Histogram Counts (SiPHiC) procedure in Algorithm 4 that honors the fact that N affects the accuracy of the histogram and privacy guarantees by sanitizing N_i instead of $\log(N_i/N)$. The GS Δ_i in line 2 of Algorithm 4 is given in Proposition 5; the proof is given in the Appendix.

Algorithm 4: The SiPHiC procedure

input : N samples of $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta}|\mathbf{x})$, privacy budget ϵ.
output: a privacy-preserving posterior sample $\boldsymbol{\theta}^*$.
1 Construct a histogram estimator of $f(\boldsymbol{\theta}|\mathbf{x})$ with the N samples of $\boldsymbol{\theta}$;
2 Obtain bin count N_i for $i = 1, \ldots, B$ and sanitize N_i via the Laplace mechanism of ϵ-DP: $N_i^* \sim \text{Lap}(N_i, \Delta_i/\epsilon)$ to obtain a sanitized histogram;
3 Draw a random sample $\boldsymbol{\theta}^*$ from the sanitized histogram with counts N_i^*.

Proposition 5. Let $G \triangleq \sup_{\mathbf{x},\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{x})$, b denote the total number of bins in the histogram constructed from N posterior samples from $f(\boldsymbol{\theta}|\mathbf{x})$, and B_j be the number of bins in the marginal histogram in the j-th dimension for $j = 1 \ldots, p$, and the width of the B_j bins given by $\mathbf{h}_j = (h_{1,j}, \ldots, h_{B_j,j})$. The GS for the bin count N_i in bin $\mathcal{B}_i$ $i = 1, \ldots, B$ is $\Delta_i = N \cdot G \cdot \prod_{j=1}^{p} h_{j(i),j}$, where $j(i) = 1, \ldots, B_j$ is the index of the marginal bin in the j-th dimension for the i-th bin of the p-dimensional histogram.

Proposition 5 suggests the GS of N_i increases linearly with the number of posterior samples N for a fixed $V_i = \prod_{j=1}^{p} h_{j(i),j}$. Existing bin number determination rules can be used to calculate $\mathbf{h}_j$. For example, if the Sturge's rule is used to determine the number of bins separately for each dimension, $B_j \equiv B = \lceil \log_2 N \rceil + 1$ and all elements in $\mathbf{h}_j$ are R_j/B, where R_j is the support range of θ_j; thus $\Delta_i \equiv NG \prod_{j=1}^{p} (R_j/B) = NG\big(\prod_{j=1}^{p} R_j\big)/(\log_2 N + 1)^p$. Given the relationship, we can back calculate N for a desirable value of Δ_i given a fixed B. For example, if $p = 1$, $\Delta_i = NGR/(\log_2 N + 1)$ increases with N at a sub-linear rate. Say $G = 0.1$ and $R = 6$, setting $B = 20$ (thus $h = R/B = 3/10$) and $\Delta_1 = 1$ leads to $N = 1/(Gh) \approx 24$.

The SiDD, SiDD.MC, and SiPHiC procedures can also be used for differentially private posterior sampling from "regular" $f(\boldsymbol{\theta}|\mathbf{x})$ such as multivariate normal (MVN) distributions. Though theoretically the mean and covariance matrix of a MVN can be directly sanitized to achieve DP, their GS, as functions of $\mathbf{x}$, can be difficult to derive analytically, depending on the assumed model on $\mathbf{x}$. In contrast, the application of SiDD only need users to supply an upper bound on $|\log(f(\boldsymbol{\theta}|\mathbf{x}))|$ or $|f(\boldsymbol{\theta}|\mathbf{x})|$, and calculate either the volume of each bin in a discretized MVN distribution in Algorithm 2 or the sensitivity in Algorithm 4, both of which are easy to obtain in the case of MVN. We provide the details how to apply the SiDD, SiDD.MC, and SiPHiC procedures when $f(\boldsymbol{\theta}|\mathbf{x})$ is a MVN in the Appendix.

In summary, all three sanitization of approximate distributions procedures (SiDD, SiDD.MC, and SiPHiC) require discretization of $f(\boldsymbol{\theta}|\mathbf{x})$. The discretization incurs information loss, which, supposedly, also brings in some degree of privacy protection. At the moment, we do not take this into account but rely on explicit randomized mechanisms to achieve DP. The main reason is that the privacy guarantee associated with discretization are not easy to quantify. We will keep this as a topic for future research.

Finally, it should be noted that though there exist approaches to obtaining a privacy-preserving density estimator given a set of data samples [45, 88], they do not directly apply to our setting. The reason is as follows. These approaches deal with the problem that the samples which the density estimate is formed from are the exact data for privacy protection. In our case, the original data $\mathbf{x}$ is subject to privacy protection but its density estimate is not the target for sanitization or release – but rather $f(\boldsymbol{\theta}|\mathbf{x})$; in other words, we aim to protect the privacy information in $\mathbf{x}$ caused by releasing samples of $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta}|\mathbf{x})$.

3.3 Sanitization of Statistics in modips.SSS

The sufficient statistic $\mathbf{s}$ associated with a Bayesian model in the modips.SSS procedure is often multi-dimensional. For a fixed privacy budget, it would be in the best interest of data users to preserve as much original information as possible when sanitizing $\mathbf{s}$. Toward that end, we may first examine whether any elements in $\mathbf{s}$ are calculated based on disjoint subsets of individuals so to leverage the parallel composition principle. For statistics are based on data from at least one shared individual, there are different schemes for budget allocation when it comes to sanitization; we introduce two below. For easy illustration, we present the two definitions in the context of the Laplace mechanism; the definitions are general and apply to other mechanisms such as the Gaussian mechanism and exponential mechanism.

Definition 5. Denote the l_1-GS of a multidimensional $\mathbf{s}$ by $\delta_{\mathbf{s}} = \sum_{i=1}^{r} \Delta_i$, where Δ_i is the l_1-GS of s_i . The *communal sanitization* sanitizes s_i via $s_i^* = s_i + e_i$, where $e_i \overset{\text{ind}}{\sim}$ Laplace $(0, \delta_{\mathbf{s}}\epsilon^{-1})$ for $i = 1, \ldots, r$.

Definition 6. Denote the l_1-GS of s_i in a multidimensional $\mathbf{s}$ by δ_i and by w_i the proportion of ϵ allocated to s_i for $i = 1, \ldots, r$, and $\sum_{i=1}^{r} w_i = 1$. The *individualized sanitization* sanitizes s_i via $s_i^* = s_i + e_i$, where$e_i \overset{\text{ind}}{\sim}$ Laplace$(0, \delta_i(w_i\epsilon)^{-1})$.

In short, all elements in $\mathbf{s}$ are sanitized via the same Laplace mechanism in the communal sanitization, while the sanitation is "individualized" for each element in the individualized sanitization. Remarks 6 compares the communal sanitization and individualized sanitization in two special scenarios. The proof can be found in Appendix I.

Remark 6. (a) The communal sanitization for the Laplace mechanism is a special case of the individualized sanitization when $w_i = \delta_i \left(\sum_{i=1}^r \delta_i\right)^{-1} = \delta_i \delta_{\mathbf{s}}^{-1} \propto \delta_i$ for $i = 1, \ldots, r$. (b) Set $w_i \equiv 1/r$ (equal allocation) in the individualized sanitization. Define the average sensitivity $\bar{\delta}_{\mathbf{s}} \triangleq \delta_{\mathbf{s}}/r$. If $\delta_i < \bar{\delta}_{\mathbf{s}}$, the scale parameter of the Laplace mechanism for the individualized sanitization is smaller that in the communal sanitization; in other words, if the sensitivity of an element s_i in $\mathbf{s}$ is smaller than the average, then allocating the same privacy budget to s_i as to every other statistic in $\mathbf{s}$ in the individualized sanitization leads to less perturbation compared to when the communal sanitization is employed. If the sensitivity of an element s_i in $\mathbf{s}$ is smaller than the average, then allocating the same privacy budget to s_i as to every other statistic in $\mathbf{s}$ in the individualized sanitization leads to more perturbation compared to when the communal sanitization is employed.

The individualized sanitization offers more flexibility as it allows users to specify the privacy budget each s_i receives. There is no restriction on how to specify $w_i > 0$ as along as $\sum_{i=1}^r w_i = 1$ is satisfied. Equal allocation as in Part (b) of Remark 6 may be used; one may define w_i according to how "important" s_i is by some importance metrics.[7] For example, in the context of modips.SSS, if an element in $\mathbf{s}$ is deemed more influential to the quality of synthetic data, then it can be deemed important and receive a big portion of ϵ.

3.4 Model-free dips

Implementation of the modips procedure requires the specification of a Bayesian model given data $\mathbf{x}$ and sanitization of the posterior distribution. If the model does not represent the underlying unknown population distribution well, synthetic data from modips can deviate significantly from the original data, leading to the subsequent invalid inference of population parameters based on the synthetic data. To circumvent this potential problem, we propose a model-free version for modips, as illustrated in Fig. 4(a).

First, an empirical distribution $\hat{f}(\mathbf{x})$ such as histograms is constructed from $\mathbf{x}$ and is then sanitized in a differentially private manner to obtain $f^*(\mathbf{x})$, from which $\tilde{\mathbf{x}}^*$ is sampled. The synthetic data $\tilde{\mathbf{x}}^*$ resembles the original $\mathbf{x}$ except for the variability due to sanitization and the error in constructing $\hat{f}(\mathbf{x})$; however, it ignores the uncertainty of not knowing the underlying population distribution $f(\mathbf{x})$ where the sample data $\mathbf{x}$ come from. This may lead to under-estimated variance for inference based on $\tilde{\mathbf{x}}^*$. One solution to this problem is to incorporate a bootstrap step to propagate the uncertainty from not knowing $f(\mathbf{x})$ in the synthetic data, as demonstrated in Fig. 4(b).

The modips.SSS procedure can also be extended to a model-free setting if sufficient statistics $\mathbf{s}$ is predictive sufficient, that is, $\Pr(\tilde{\mathbf{X}} = \tilde{\mathbf{x}} \mid \mathbf{X} = \mathbf{x}) = \Pr(\tilde{\mathbf{X}} = \tilde{\mathbf{x}} \mid s(\mathbf{X}) = s(\mathbf{x}))$. To implement the model-free modips.SSS, one would sanitize $\mathbf{s}$ to obtain $\mathbf{s}^{(i)*}$, and then draw $\tilde{\mathbf{x}}^{(i)*}$ from $f(\tilde{\mathbf{x}}^{(i)*}|\mathbf{s}^{(i)*})$ if this distribution is easy to compute and sample from.

4 Statistical Inference in Differentially Private Synthetic Data Analysis

To account for the randomness of the synthesis and sanitization process when obtaining the inference from dips data, there are at least two approaches: 1) explicitly modeling the

[7] The definition of "importance" varies from case to case (e.g., statistically v.s. practically important); careful considerations are required when choosing w_i according to importance.

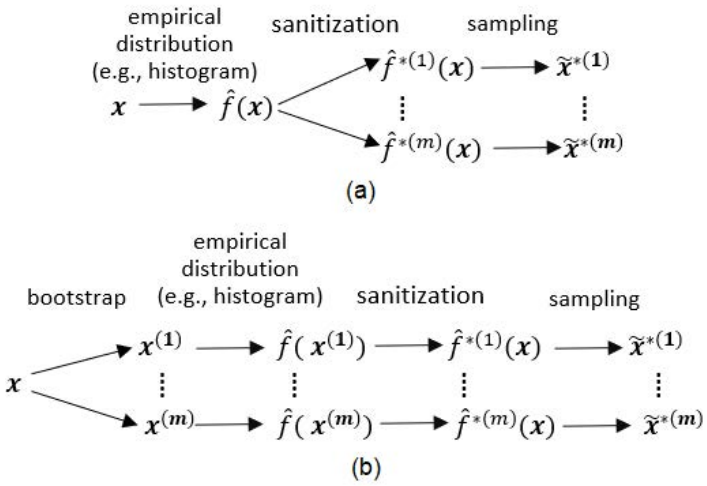


Figure 4: Examples of model-free dips schemes. (a) sample from sanitized empirical distribution of $\mathbf{x}$*; (b) sample from sanitized empirical distribution of bootstrapped samples of* $\mathbf{x}$*.*

sanitization and synthesis mechanisms and estimating variance either analytically or computationally; 2) propagating uncertainty through MS and apply an appropriate variance combination rule. The former does not require the release of multiple synthetic datasets but data users need to be provided with full detail of the sanitization/synthesis mechanisms so to model and incorporate the mechanisms in their data analysis procedures. This can be challenging especially for users who are not familiar with DP and sanitization mechanisms. Even if users are familiar with DP, incorporating a randomized mechanism even in common statistical analysis procedures may be analytically and computationally challenging. In contrast, the MS approach is more user-friendly as there is no need to explicitly model the sanitization mechanism as long as the sanitization/synthesis procedure does not introduce inferential bias and users can analyze each of the synthetic datasets as if they had the original data and then combine the results across multiple sets to generate final inferential results. In what follows, we provide an inferential rule to obtain inferences from multiple dips datasets. The main results are provided in Theorem 7. Before that, we first present Definition 7 on which the main results are based.

Definition 7. Suppose $f^*(\boldsymbol{\theta}|\mathbf{x})$ is a sanitized version of the original posterior distribution $f(\boldsymbol{\theta}|\mathbf{x})$ via a differentially private mechanism with privacy loss ϵ. If $f^*_\epsilon(\boldsymbol{\theta}|\mathbf{x}) \to f(\boldsymbol{\theta}|\mathbf{x})$ as $\epsilon \to \infty\ \forall \boldsymbol{\theta}$ given $\mathbf{x}$, then $f^*_\epsilon(\boldsymbol{\theta}|\mathbf{x})$ is *consistent* for $f(\boldsymbol{\theta}|\mathbf{x})$.

Theorem 7. Assume the model from which the posterior distribution of $\boldsymbol{\theta}$ is obtained in the modips procedure is the same as the one used for synthesis and that $f^*(\boldsymbol{\theta}|\mathbf{x})$ is consistent for $f(\boldsymbol{\theta}|\mathbf{x})$. Let $\tilde{\mathbf{x}}^{*(j)}$ denote the j-th synthetic dataset given sample $\boldsymbol{\theta}^{*(j)}$ from $f^*(\boldsymbol{\theta}|\mathbf{x})$ for $j = 1, \ldots, m$. Denote the parameter of inferential interest by β and assume the statistical procedure to obtain inference for β is the same whether the data is $\tilde{\mathbf{x}}^{*(j)}$ or $\mathbf{x}$. Denote the estimate of β from $\mathbf{x}$ by $\hat{\beta}$ and the corresponding variance estimate by v, those based on $\tilde{\mathbf{x}}^{*(j)}$ by $\hat{\beta}^{*(j)}$ and $\hat{v}^{*(j)}$, respectively. If $\hat{\beta} \xrightarrow{p} \beta, \hat{\beta}^{*(j)} \xrightarrow{p} \beta^*, \mathrm{E}\big(\hat{\beta}^{*(j)}|\mathbf{x}\big) \to \hat{\beta}, \mathrm{E}\big(m^{-1}\sum_{j=1}^m \mathrm{V}(\beta^{*(j)}|\tilde{\mathbf{x}}^{*(j)})|\mathbf{x}\big) \to \mathrm{V}(\beta|\mathbf{x})$, and $\mathrm{E}\big((m\text{–}1)^{-1}\sum_{j=1}^m(\hat{\beta}^{*(j)}\text{–}\bar{\beta}^*)^2|\mathbf{x}\big) \to \mathrm{V}(\beta^{*(j)}|\mathbf{x})$ as $n \to \infty$, where $\bar{\beta}^* = m^{-1}\sum_{j=1}^m \hat{\beta}^{*(j)}$, then

a) $\bar{\beta}^* = m^{-1}\sum_{j=1}^{m}\hat{\beta}^{*(j)}$ is a consistent estimator for β;

b) an asymptotically unbiased estimator for the variance of $\bar{\beta}^*$ is

$$u = \varpi + m^{-1}b, \text{ where}$$
$$\varpi = m^{-1}\sum_{j=1}^{m}\hat{v}^{*(j)} \text{ and } b = (m-1)^{-1}\sum_{j=1}^{m}(\hat{\beta}^{*(j)} - \bar{\beta}^*)^2;$$

ϖ is the averaged within-set variance $\mathrm{E}(\mathrm{V}(\beta|\mathbf{x})|\mathbf{x}^{*1)}, \ldots, \mathbf{x}^{*(m)})$. b is the between-set variance $\mathrm{V}(\mathrm{E}(\beta|\mathbf{x})|\mathbf{x}^{*1)}, \ldots, \mathbf{x}^{*(m)})$ that is comprised of two components b_1 and b_2; b_1 is the variability incurred by sanitization and b_2 is the variability due to synthesis that further comprises b_{21} and b_{22}, corresponding to the posterior variability of $\boldsymbol{\theta}$ and the sampling variability of $\mathbf{x}$ given $\boldsymbol{\theta}$, respectively;

c) the inference of β given $\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}$ is based on $t_\nu(\bar{\beta}^*, m^{-1}b+\varpi)$, where the degree of freedom $\nu = (m-1)(1+m\varpi/b)^2$.

The proof is provided in Appendix H. Though Part b) decomposes the between-set variance component b , there is seldom an interest in quantifying b_1 and b_2 separately. If there is such a need, it can be fulfilled via the nested modips procedure presented in the appendix.

We expect the results in Theorem 7 also apply to general dips approaches including both model-based and model-free approaches that do not have to generate dips data in the Bayesian framework. Depending on the dips procedure, the nature of the between variability b may differ. For example, if we use a model-free dips approach given in Fig. 4(a), then b comprises the sanitization variability b_1 and part of the sanitization variability b_2 (i.e., the sampling variability b_{22} from $\hat{f}(\mathbf{x})$ but not the uncertainty from not knowing the distribution of $\mathbf{x}$). Note that this is not the problem of the variance formula $\varpi + b/m$, but rather because the dips approach in Fig. 4(a) does not take into account the fact that the population distribution $f(\mathbf{x})$ is unknown. If the procedure in Fig. 4(b) is used, then b includes b_1 and both components of b_2 as the uncertainty of $f(\mathbf{x})$ is captured through bootstrap.

The formula $u = \varpi + m^{-1}b$ in Theorem 7 and the formula for variance combination across MS data generated for the partial sample synthesis approach in the non-DP setting [10] [8] are the same, which seems to be a coincidence at first glance. Further analysis, however, suggests that the equivalence between two formulas is not coincidental as the modips procedure can be viewed as a differentially private version of the full sample synthesis with an extra step of explicitly sanitizing the posterior distribution. In other words, we can regard the full sample synthesis as the asymptotic case of modips approach as $\epsilon \to \infty$.

In the non-DP data synthesis setting, the choice of m is mostly driven by computational time and storage considerations; thus small m is preferred as long as it is large enough to capture the between-set variance and deliver valid inference. The empirical studies [10, 9] suggest small m (e.g., $\leq 10 \sim 15$) seems to work. In contrast, in the DP setting, the decision on m is driven by the utility of the sanitized data at a pre-specified privacy budget ϵ and it is not necessarily true that a larger m yields better utility in the synthetic datasets overall. This is because each synthesis receives only $1/m$ of the total budget. While a too small m may not be sufficient to capture the b component, a too large m risks spreading ϵ too thin over m sets and each synthetic set is so"over-perturbed" that aggregating information across m synthetic set cannot remedy the information loss.

We examine the effect of m on the inference in dips data empirically in Sec. 5.1 and plan to

[8] The formula is also applicable to the non-DP partial and full sample synthesis, as it can be viewed as a special case of partial sample synthesis with a 100% synthesis proportion.

provide more theoretical analysis on this problem in the future, which can be a challenging task. We expect that m synthetic sets differ more and more as m increases, leading to an increase in b at least initially. On the other hand, if m is very large, the large amount of perturbation may push each synthetic data to some consistent extremes, causing b to decrease instead. In addition, ϖ may also change with m in a manner that depends on how much the randomized mechanism perturbs the original data in what way. If m does affect ϖ, its effect is expected to be smaller compared to that on b (see the experimental results in Sec. 5.2). The difficulty in the theoretical analysis lies in obtaining a functional form for $u(m) = \varpi(m) + m^{-1}b(m)$, which may vary case by case. If close form for $u(m)$ exists for some problems, the rate of the change of $u(m)$ with m cam be quantified by its first derivative $\varpi'(m) + m^{-1}b'(m) - m^{-2}b(m)$. If $m^2\varpi'(m) + mb'(m) < b(m)$, then $u(m)$ decreases with m; otherwise, $u(m)$ increases with m. Whether there exists an m the leads to a minimum or maximum in u depends on specific problems (see Sec. 5.1); general meaningful theoretical results may not exist.

5 Numerical Examples

This section presents numerical examples to examine the effects of m on statistical inference based on differentially private synthetic data; to demonstrate the validity of the inferential procedure in Sec. 4; and to investigate the impact of budget allocation on inference based on the synthetic data via modips.SSS. We also survey and summarize the published work that has implemented the modips procedure or the inferential combination rule.

Prior to the presentation of the main observations and findings in Sec. 5.1 to 5.4, we first use a simple example to demonstrate the implementation of the modips procedure. This example (Fig. 5) is also part of the simulation studies in Sec. 5.1 and 5.2. The original data $\mathbf{x}$ contains a single binary variable $X \in \{0, 1\}$. Even before seeing the actual values in $\mathbf{x}$, a natural model choice for binary data is $x_i \sim$ Bernoulli(p) for $i = 1, \ldots, n$, where p is the unknown proportion pf $X = 1$. A common prior for p is the conjugate prior to the Bernoulli likelihood $f(\mathbf{x}|p) = p^{n_1}(1-p)^{n_0}$ (assuming the n data points are independent) is $f(p) =$ beta(α, β), where $n_1 = \sum_{i=1}^{n} x_i$, $n_0 = n - n_1$, and α, β are user-specified hyper-parameters. The posterior distribution is $f(p|\mathbf{x}) =$ beta$(n_0 + \alpha, n_1 + \beta)$, with Bayesian sufficient statistic n_1 (or n_0). Since the likelihood is chosen without using any specific value in $\mathbf{x}$, all privacy budget ϵ can be used toward sanitizing n_1 via modips.SSS. Suppose the Laplace mechanism is employed to obtain sanitized $n_1^* = n_1 +$Laplace$(0, m\epsilon^{-1})$ (the GS of n_1 is 1). Plugging in n_1^* in the posterior distribution, we have sanitized $f^*(p|\mathbf{x}) =$ beta $(n_0^* + \alpha, n - n_1^* \beta)$. Finally, we sample p^* from $f^*(p|\mathbf{x})$ and $\tilde{x}_i$ from Bernoulli(p^*) for $i = 1, \ldots, n$ to generate one set of synthetic data $\tilde{\mathbf{x}}^*$. The sanitization of n_1 and sampling of p^* and $\tilde{\mathbf{x}}^*$ are repeated m times to obtain m sets of synthetic binary data.

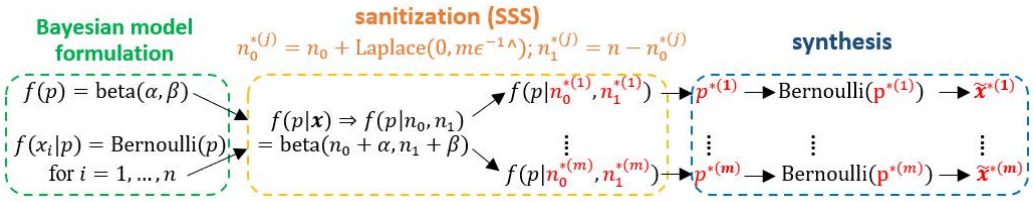


Figure 5: Implementation of modips.SSS in univariate binary data

Upon receiving the m datasets, users would run the analysis of their choice on each set as if they had the original data and then combine the results across the m sets to obtain the final inference using the inferential rule in Sec. 4. For example, if they are interested in estimating p and its 95% CI, they would calculate the sample proportion $\hat{p}^{(j)}$ and the corresponding variance $\varpi^{(j)} = n^{-1}\hat{p}^{(j)}(1-\hat{p}^{(j)})$ from $\tilde{\mathbf{x}}^{*(j)}$ for $j = 1, \ldots, m$. The final p estimate based on the information from the m sets is $\bar{p} = m^{-1}\sum_j \hat{p}^{(j)}$ and $(\hat{p} + t_{0.025,\nu}\sqrt{v}), \hat{p} + t_{0.975,\nu}\sqrt{v})$, where $v = m^{-1}\sum_j \varpi^{(j)} + m^{-1}b$, $b = (m-1)^{-1}\sum_j(\hat{p}^{(j)} - \bar{p})^2$, and $\nu = (m-1)(1+m\varpi/b)^2$.

5.1 Impact of m on Inference

We present two simulation studies to examine the impact of m on statistical inference based on released dips data. The first is the binary data example described above. We simulated binary data for two values on p (0.1, 0.5; $p = 0.1$ represents an unbalanced data scenario). The prior for p is beta$(1, 1)$ and thus $f(p|\mathbf{x}) =$ beta$(n_0 + 1, n_1 + 1)$. The second example concerns Gaussian data and the parameter of interest is mean μ ($\sigma^2 = 1$ and assumed known); we set $\mu = 0$ in the simulation. With prior $f(\mu) \propto$ constant, $f(\mu|\mathbf{x}) = \mathcal{N}(\bar{x}, n^{-1})$, where $\bar{x}$ is the sample mean and also the sufficient statistic in $f(\mu|\mathbf{x})$. The GS of μ is $(c_1 - c_0)/n$, where $[c_0, c_1]$ are the global bounds on data $\mathbf{x}$ [89]. We examined two sets of bounds[9]: $[c_0, c_1] = [-4, 4]$ and $[4, 5]$. Given $\Pr(|X| > 4) = 0.0063\%$, the truncated data can still be well approximated by a Gaussian distribution. The sanitized statistics via the Laplace mechanism can be out of bounds (< 0 or > 1 for sanitized sample proportions, and $< c_0$ or $> c_1$ for sanitized sample means) as the support for the Laplace distribution is the real line. There are two ways to legitimize the out-of-bound values – truncation and boundary inflated truncation (BIT) [89]. The former throws away out-of-bound values and the latter sets the values smaller than the lower bound at the lower bound and those larger than the upper bound at the upper bound.

In both studies, we examine two sample sizes n at 100 and 1000, respectively. We set $m \in [2, 500]$ with the understanding that large m is for investigation purposes and unlikely used in practice. The overall privacy budget is $\epsilon = 1$, split equally across m synthetic datasets. The inferential procedure on p is given above in the binary data case; that on μ in the Gaussian data case is similar with μ estimated by the sample mean in each synthetic set and $\varpi = n^{-1}\sigma^2$. We summarize the bias, standard deviation (SD) estimation, and coverage probability (CP) of the 95% CIs in the inference of p and μ over 5,000 repetitions in Figs 6 and 7, respectively.

The main observations from Fig. 6 are as follows. 1) There is minimal bias in $\bar{p}^*$ (point estimate of p) for all the examined m in all the scenarios except for the slight over-estimation (bias ~ 0.01) for $m > 200$ when $n = 100$ and $p = 0.1$; such a large m is unlikely to be used in practical applications. 2) The SD estimate for $\bar{p}^*$ based on the sanitized data is larger than that based on the original data, which is expected since not only are the synthetic data subject to the sampling error as the original data, but also they have two additional variability sources from sanitization (noise injection to satisfy DP) and synthesis (incorporation of the uncertainty of knowing the underlying population model or parameters) that the original data do not have. For both $n = 100$ and $n = 1,000$, the SD decreases as m increases, gets lose to the original SD around $m = 30 \sim 40$, and then increases with m though minimally at $n = 100$. 3) The CP is near nominal (95%) for almost all m with the slight under-coverage but still $\sim 92.5\%$ at $m = 2$ and severe under-coverage when $m > 200, n = 100, p = 0.1$, a consistent observation with the large bias in the same condition.

[9] Bounds are required for DP purposes; see Liu [89]

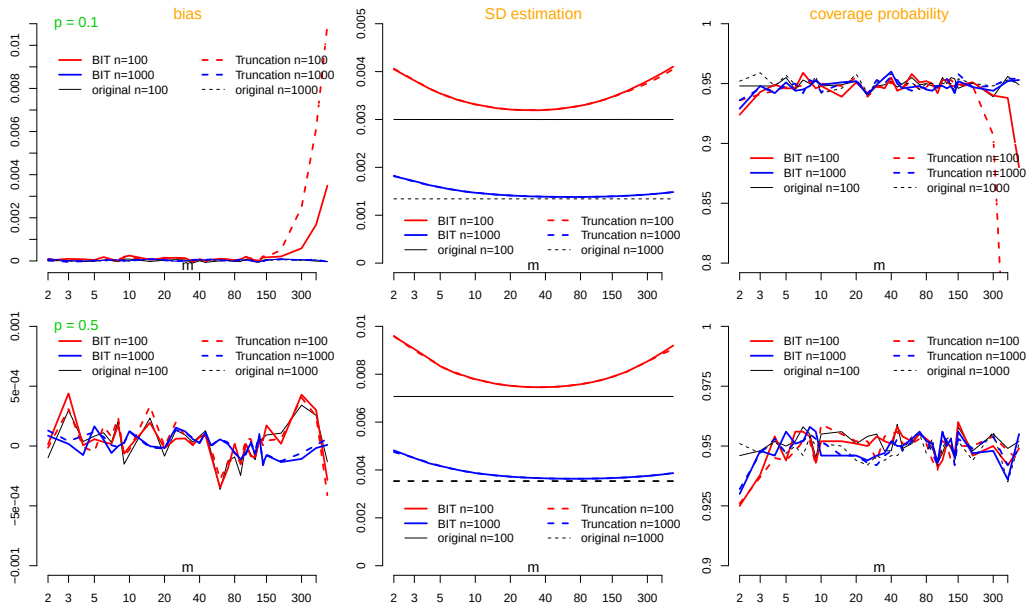


Figure 6: Effect of m (on log scale) on inference based on synthetic binary data via modips.SSS at $\epsilon = 1$

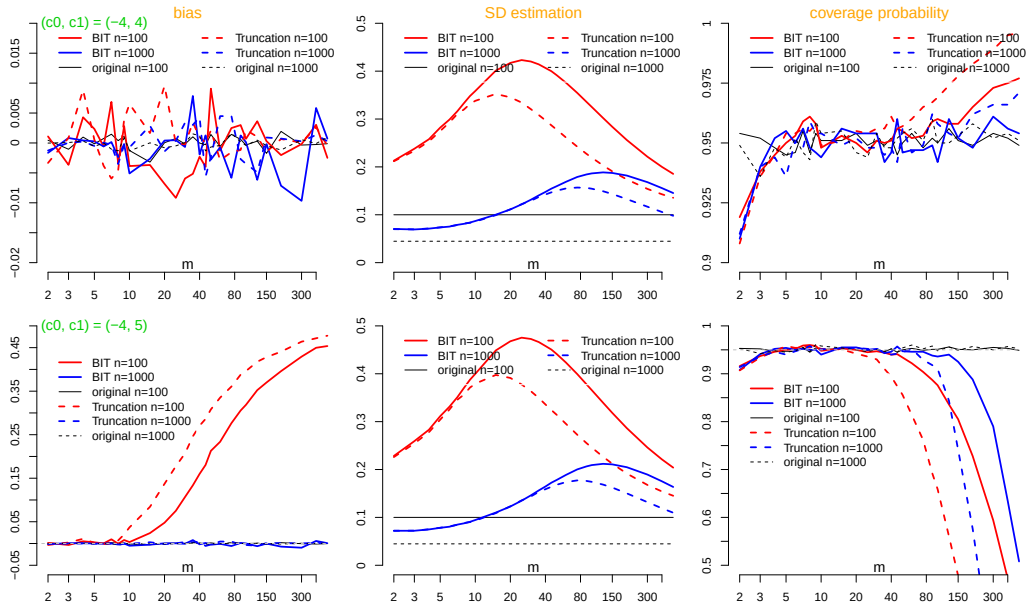


Figure 7: Effect of m (on log scale) inference based on synthetic Gaussian data via modips.SSS at $\epsilon = 1$

For the Gaussian case in Fig. 7, m has a smaller impact on the inference when the bounds are symmetric $(-4, 4)$ around the true mean ($\mu = 0$) than with the asymmetric bounds $(-4, 5)$. Specifically, 1) there is minimal bias in the estimation of μ for all m when the bounds case

are symmetric and there is obvious bias for $m > 10$ in the asymmetric bounds case when $n = 100$. 2) How the variance estimate changes with m is similar between the symmetric and asymmetric bounds cases. When $n = 1,000$, the variance estimate remains roughly constant for $m < 10$ and then increases with m, reaches its peak around $m = 50$ (truncation) to 150 (BIT) before decreasing with m. When $n = 100$, the variance increase with m, reaches its peak around $m = 20$ (truncation) to $m = 40$ (BIT), and then decreases with m. 3) The CP is near nominal (95%) across most m with slight under-coverage ($\sim 91\%$) at $m = 2$ and some over-coverage for $m > 40$ when $n = 100$ when the bounds are symmetric. For the asymmetric bounds case, there is severe under-coverage for $m > 20$ (truncation) and $m > 40$ (BIT) when $n = 100$, and for $m > 80 \sim 150$ when $n = 1,000$.

In summary, Figs 6 and 7 suggest that m affects inference based on sanitized data and in what way depends on true parameter values, global bounds on data, truncation schemes, etc. Variance estimation is the most sensitive to the change in m value, compared to bias and CP. At least in these two examples, $m \in [3, 10]$ seems to be a good choice for satisfactory performance in inference; a large m is unnecessary from a computation and storage perspective and sometimes undesirable from an inferential perspective.

5.2 Variance Combination Rule

We use similar simulation settings to Sec. 5.1 but with smaller n (10 and 100) to compare the variance combination rule $\varpi + m^{-1}b$ in Theorem 7 with three other variance combination rules developed for different but related settings[10]: 1) $(1 + m^{-1})b + \varpi$ that combines inferences from multiply imputed datasets for missing data analysis [90]; 2) $(1+m^{-1})b-\varpi$ for inference in the population full synthesis in the non-DP setting [9]; 3) $(1 + 2/m)\varpi$ for MS in non-DP setting [91]. We used $m = 10$ in all cases. Since the results for asymmetric and symmetric bounds $[c_0, c_1]$ in the Gaussian case and the results from the two bounding schemes (BIT and truncation) are similar, we present those from the symmetric bounds and the BIT scheme only.

The main observations on CP in Table 1 are summarized as follows. 1) Our proposed variance rule $m^{-1}b+\varpi$ provides nominal coverage in all simulation scenarios. 2) $(1+2/m)b$ provides nominal coverage when ϵ is large (≥ 10) and $n = 100$, but leads to severe under-coverage when n or ϵ is small. 3) $(1 + 1/m)b - \varpi$ leads to either under-coverage or over-coverage and hardly produces nominal coverage. (4) $(1+m^{-1})b+\varpi$ is overly conservative and delivers close to 100% coverage in all cases. (5) As expected, the single synthesis leads to severe under-coverage as it does not capture the synthesis and sanitization uncertainty unless users explicitly model the synthesis and sanitization process.

Fig. 8 plots the SD estimates. In the binary case, $(1 + m^{-1})b + \varpi$ produces the largest SD estimate, as expected, followed by $(1+m^{-1})b-\varpi$, $\varpi+b/m$, and $(1+2/m)\varpi$. $\varpi+b/m$ and $(1 + 2/m)\varpi$ are similar when $\epsilon > 1$ for all the examined simulation scenarios. When $\epsilon > 1$, $(1 + m^{-1})b - \varpi$ yields similar results to $\varpi + b/m$ and $(1 + 2/m)\varpi$. In the Gaussian case, $(1+m^{-1})b+\varpi$ and $(1+m^{-1})b-\varpi$ produce very similar results, followed by $\varpi+b/m$ and $(1+2/m)\varpi$; the latter two are similar for $\epsilon > 10$. All rules are similar at $\epsilon = 100$ when $n = 10$ and for $\epsilon > 10$ when $n = 100$. In both the binary and Gaussian cases, the SD estimate is roughly constant across ϵ for $(1+2/m)\varpi$ as the formula ignores b which changes drastically with m in the DP setting, implying the inappropriateness of the formula for DP-based MS. For the other three rules that contain the b component, the SD estimate stabilizes after a

[10] As stated in Sec. 4, the formula in [10] for the partial sample synthesis in the non-DP setting is the same as in Theorem 7 and there is no need to include it as a comparison method.

Table 1: Coverage probability of 95% CI using different variance combination rules ($m = 10$)

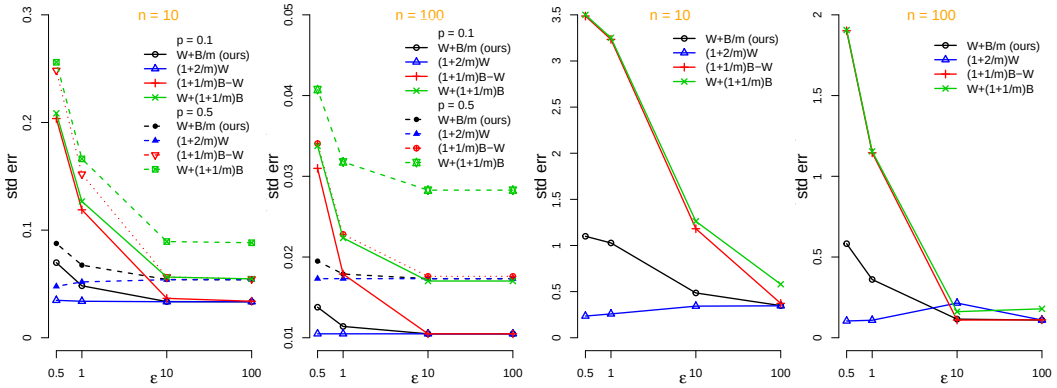
(a) Binary Data

scenario			original	multiple synthesis				single
ϵ	n	p		$B/m+W$ (Thm 7)	$(1+2/m)\varpi$	$(1+1/m)B-\varpi$	$(1+1/m)B+\varpi$	synthesis
100	10	0.5	0.946	0.948	0.948	0.798	0.999	0.736
100	10	0.1	0.935	0.950	0.949	0.793	0.996	0.738
100	100	0.5	0.945	0.952	0.952	0.797	0.999	0.747
100	100	0.1	0.949	0.949	0.950	0.791	0.998	0.743
10	10	0.5	0.942	0.945	0.945	0.814	0.998	0.723
10	10	0.1	0.930	0.946	0.946	0.835	0.997	0.732
10	100	0.5	0.946	0.947	0.947	0.792	1.000	0.738
10	100	0.1	0.952	0.948	0.950	0.795	0.998	0.747
1	10	0.5	0.938	0.947	0.865	0.994	1.000	0.730
1	10	0.1	0.936	0.961	0.840	0.994	0.999	0.726
1	100	0.5	0.948	0.946	0.938	0.898	0.999	0.746
1	100	0.1	0.950	0.952	0.931	0.958	1.000	0.749
0.5	10	0.5	0.942	0.941	0.707	0.999	1.000	0.715
0.5	10	0.1	0.928	0.946	0.532	0.998	0.999	0.686
0.5	100	0.5	0.955	0.953	0.924	0.972	1.000	0.730
0.5	100	0.1	0.949	0.949	0.869	0.992	1.000	0.756

(b) Gaussian Data

scenario		original	multiple synthesis				single
ϵ	n		$B/m+W$ (Thm 7)	$(1+2/m)\varpi$	$(1+1/m)B-\varpi$	$(1+1/m)B+\varpi$	synthesis
100	10	0.951	0.953	0.946	0.825	0.997	0.743
100	100	0.949	0.952	0.952	0.795	0.997	0.741
10	10	0.951	0.952	0.836	0.999	1.000	0.720
10	100	0.948	0.946	0.933	0.924	0.998	0.739
1	10	0.948	0.951	0.392	0.998	0.999	0.450
1	100	0.953	0.956	0.454	1.000	1.000	0.674
0.5	10	0.951	0.954	0.334	0.997	0.998	0.275
0.5	100	0.950	0.951	0.274	1.000	1.000	0.551

certain ϵ, the value of which depends on n and data type, among others.



(a) binary data (b) Gaussian data

Figure 8: Standard error estimate via different variance combination rule

In summary, the variance combination rule $(1 + 2/m)\varpi$ ignores the b component, which matters in the DP setting, leading to under-estimated variance and under-coverage, and is thus invalid. $(1 + 1/m)b + \varpi$ is overly conservative for synthetic data in the DP setting. $(1 + 1/m)b - \varpi$ does not provide the correct combination between b and ϖ either as it focuses wrongly on b when ϖ is the main contributor to the total variance at large ϵ or tends to over-weigh b when it is large at small ϵ. All taken together, none of the three is suitable for inference in the analysis of MS data in the DP setting. In addition, even $(1 + 1/m)b - \varpi$ is not smaller than $m^{-1}B + \varpi$, the former stills leads to under-coverage as the former uses z-statistic for inferences and the latter is based on t-distributions. Finally, $(1 + 1/m)b - \varpi$ is smaller than $(1 + 1/m)b + \varpi$ by 2ϖ; but the difference is not obvious at small ϵ. This is because b is the dominant contributor to the total variance and overshadows the contribution from ϖ. The difference at large ϵ between the two is more obvious in the binary case than in the Gaussian case.

5.3 Impact of Budget Allocation on Inference

We examine how budget allocation schemes in the individualized sanitization (Definitions 5 and 6) affect inference based on synthetic data in this simulation study. We simulated data $\mathbf{x}$ from $\mathcal{N}(\mu = 0, \sigma^2 = 1)$. We examine two sample size cases $n = 100$ and $1,000$, the symmetric and asymmetric bounds scenarios $(c_0, c_1) = (-4, 4)$ and $(-4, 5)$ on $\mathbf{x}$, and the BIT and truncation bounding schemes. The bounds for sanitized means and synthetic data are (c_0, c_1), and for sanitized variances, are $(0, (c_1 - c_0)^2/4)$. We are interested in the inference for both μ and σ^2. With the Gaussian likelihood and prior $f(\mu, \sigma^2) \propto \sigma^{-2}$, the posterior distribution $f(\mu, \sigma^2|\mathbf{x})$ is inverse-gamma$((n-1)/2, (n-1)s^2/2)\mathcal{N}(\bar{x}, \sigma^2/2)$, and the Bayesian sufficient statistics are the sample mean $\bar{x}$ and variance s^2, the l_1 GS of which are $(c_1 - c_0)/n$ and $(c_1 - c_0)^2/n$, respectively [89]. Denote the proportion of the privacy budget ϵ allocated to sanitizing $\bar{x}$ and s^2 by $w \in (0, 1)$ and $1 - w$, respectively. When $w = (c_1 - c_0)/((c_1 - c_0) + (c_1 - c_0)^2) = 1/(1 + c_1 - c_0)$, the individualized sanitization becomes the communal sanitization. We set at $m = 10$ and $\epsilon = 1$ and examine how w affects the inference of μ based on sanitized data. The results on the bias, CP, and half-width of 95% CIs are summarized over 5,000 repeats in Fig. 9.[11]

The findings are summarized as follows. 1) When the bounds (c_0, c_1) are symmetric around μ, w barely affects the accuracy of the point estimate $\bar{\tilde{x}}^*$ except for some numerical fluctuation. When the bounds are asymmetric, there is obvious bias at $n = 100$, which decreases as w increases and remains roughly constant for $w > 0.5$. 2) The CP is around nominal 95% with slight over-coverage for $w \geq 0.5$ for the BIT bounding scheme. For the truncation bounding scheme, there is obvious under-coverage at all w when $n = 100$ and at $w > 0.5$ when $n = 1,000$. 3) It is expected that the half-width of the 95% CI based on the sanitized data is larger than the original CI half-width given the additional variability due to sanitization and synthesis in modips. Specifically, the half-width significantly deviates from the original for all w at $n = 100$ but decreases as w increases; and is close to the original for $w > 0.5$ at $n = 1,000$.

In conclusion, larger w (portion of budgets allocated to sanitizing $\bar{x}$ tends to offer more precise inference for μ with non-inferior accuracy than lower w. The equal allocation scheme, $w = 0.5$ in this case, is a reasonable and convenient choice for this example; the "default"

[11] The inference for μ in Fig. 9 are less accurate and precise than those from the Gaussian example in Sec. 5.1 because the same privacy budget is used to sanitize two statistics (sample mean and variance) instead of just sample in the example in Sec. 5.1.

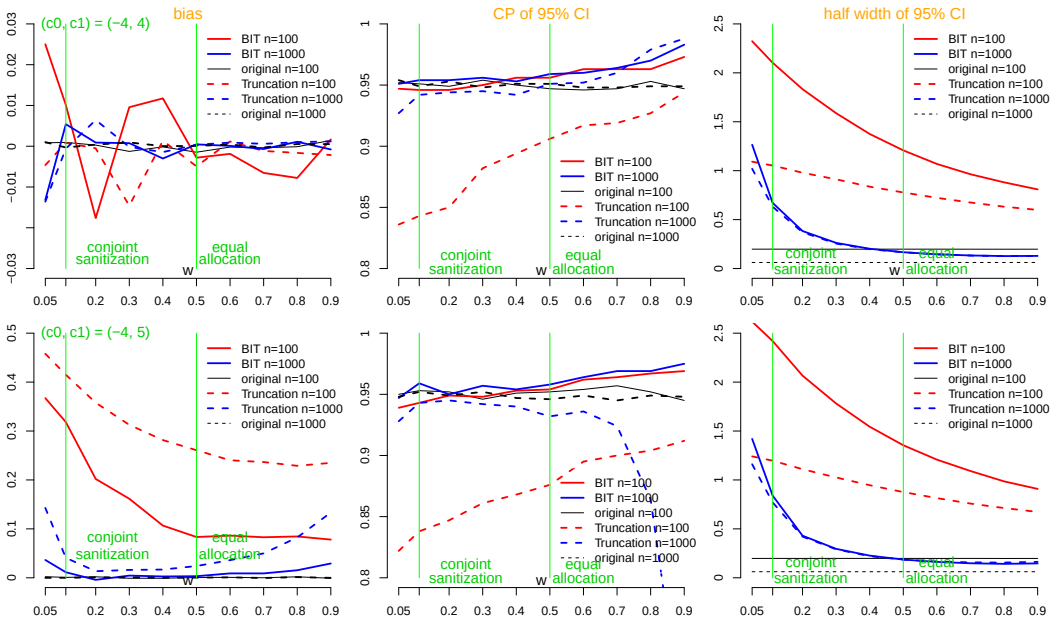


Figure 9: Effect of budget allocation on inference of μ based on sanitized data ($m = 10, \epsilon = 1$)

communal sanitization might not lead to the most efficient or accurate inference for parameters of inferential interest.

5.4 Summary of results in the literature on the application of modips and inferential combination rule

The modips approach and Theorem 7 have been employed in some later work on dips in the literature since an early version of this paper was uploaded onto arXiv.org. We summarize some of those results below.

Chanyaswad et al. [92] proposed the RON-Gauss approach that combines dimensionality reduction via random orthonormal projection and Gaussian generative models to synthesize differentially private data. They run experiments to compare RON-Gauss with 4 other dips methods, including the modips.SSS approach, at the same privacy cost ($\epsilon = 1$) for various data types and learning tasks (image data for grammatical facial expression clustering, mobile-sensing time series data for activity classification, and Twitter data to predict topic popularity) with a large number of attributes p and cases n (p =77, 117, 301 and n =573820, 216752, 27936, respectively). Since RON-Gauss was the newly proposed method, it is not surprising that it is the best performer in utility. Compared to the other 3 methods, the performance of modips varies, depending on the tasks; specifically, it was the second-best in classification and clustering, and was the worst in regression.

Bowen and Liu [93] surveyed various dips techniques including the modips.SSS method, compared them conceptually and empirically, and evaluated the statistical utility and inferential properties of the synthetic data via the techniques through extensive simulation studies. The work employs the inferential rule in Sec. 4 when obtaining inference from multiply synthetic data ($m = 5$). The main conclusions are that with appropriate model

specification, modips.SSS can generate synthetic data with valid statistical inference for a practically reasonably small privacy budget but are often less precise compared to the non-parametric dips that do not consider sampling variability during synthesis.

Liu et al. [63] proposed the DP-ERGM procedure that synthesizes network data via the exponential random graph model (EGRM) in the DP framework. DP-ERGM is a modips procedure and employs Algorithm 4 for deferentially private posterior sampling of ERGM parameters. The work also uses the inferential rule in Sec. 4 with $m = 4$. The experiment results suggest that DP-EGRM preserves the original information significantly better than two competitors in both network summary statistics and statistical inference of some parametric network models.

Bowen et al. [38] and Eugenio and Liu [39] proposed the STEPS and the CIPHER procedures, respectively, to generate differentially private synthetic data to aim for better utility or reduced computational/storage costs. Both methods are model-free in terms of synthesis. Both works obtained the inference from multiply synthetic datasets ($m = 5$) using the inferential rule in Sec. 4 with in their experiments.

6 Discussion

We propose the modips approach for differentially private data synthesis, along with several procedures to obtain differentially private posterior samples for the implementation of modips. In addition, we propose an inferential combination rule to obtain valid inferences based on multiply synthetic datasets. Our empirical studies demonstrate the validity of the combination rule for inference from differentially private synthetic data and provide insights on the impacts of the number of synthetic datasets and privacy budget allocation schemes on statistical inference.

Regarding the choice of m in multiple synthesis, our empirical study suggests that $m \in [3, 10]$ is likely to be a proper range for practical use, consistent with the publish work on dips data, where $m = 3$ to 5 is used [63, 38, 39]. In general, we expect the "optimal" m, in the sense that the original information preservation is maximized with proper uncertainty quantification at a given privacy budget, varies case by case and depends on n, p, sanitization mechanisms, among others. If things are equal, a relatively small m would be preferable so that each synthesis receives a reasonable amount of budget, as long as it is large enough to capture the between-set variability. Small m also helps to save computational/storage costs. We will continue to investigate theoretically and empirically the choice of m for general settings in the future. Note that if a dataset is used mainly for exploratory data analysis or data mining purposes, releasing a single surrogate dataset is workable. If statistical inference or uncertainty quantification are of interest, besides releasing multiple synthetic datasets, direct modelling of sanitization and synthesis mechanisms is another approach but is not as convenient or user-friendly as MS as long as the privatization and synthesis procedure is not biased.

We focus on the modips procedure in the context of the pure ϵ-DP. Extensions of modips to softer versions of DP that are immune to post processing and closed under composition, such as (ϵ, δ)-aDP and Rényi DP, are straightforward. The only modification is to replace the DP mechanism of ϵ-DP with a mechanism that satisfies the softer version of DP.

We presented modips in the context of the full sample synthesis. It may be possible to extend modips to full population synthesis, which will make an interesting topic for future research, but there will be some technical challenges given the missing values in the

unsampled set of a population and the extra sampling step for data release. While it is possible to apply DP in the framework of partial synthesis, we doubt that the robustness and rigor of the privacy guarantees can be retained in the synthetic data, which arguably one of the biggest advantages of DP over other disclosure risk control approaches. This is because partial synthesis assumes that there is minimal privacy risk from retaining and releasing a subset of the original information (a subset of attributes or individuals), the idea of which already contradicts the concepts of DP in some sense.

The modips procedure can be challenging for high-dimensional data with a large number of attributes of various types. The difficulty resides in the construction of a parsimonious but representative Bayesian model; identification, and sanitization of sufficient statistics in the case of modips.SSS; and posterior sampling in the high-dimensional setting. An alternative is to sanitize the likelihood or the posterior distribution density (or their log versions) directly if they are bounded while ensuring the sanitized likelihood and posterior distribution density still lead to proper posterior distributions.

Acknowledgement

We thank two anonymous referees and the editor for their careful reviews and insightful and constructive comments, which helped improve the quality of the manuscript.

Appendix

A Nested modips

The nested modips in Fig. A.1 is a variant of the standard modips (Fig. 3). In brief, for a given $i = 1, \ldots, m$, $t > 1$ sets of $\boldsymbol{\theta}^{*(i,1)}, \ldots, \boldsymbol{\theta}^{*(i,t)}$ are sampled, each of which leads to a synthetic dataset. The released $m \times t$ sets of surrogate data $\tilde{\mathbf{x}}^{*(1,1)}, \ldots, \tilde{\mathbf{x}}^{*(1,t)}, \ldots, \tilde{\mathbf{x}}^{*(m,1)}, \ldots, \tilde{\mathbf{x}}^{*(m,t)}$ takes a 2-layer hierarchical structure. The nested modips is useful when users of the modips procedures are interested in separately quantifying the variability from sanitization and synthesis when running analysis on released synthetic data (see Sec. 4).

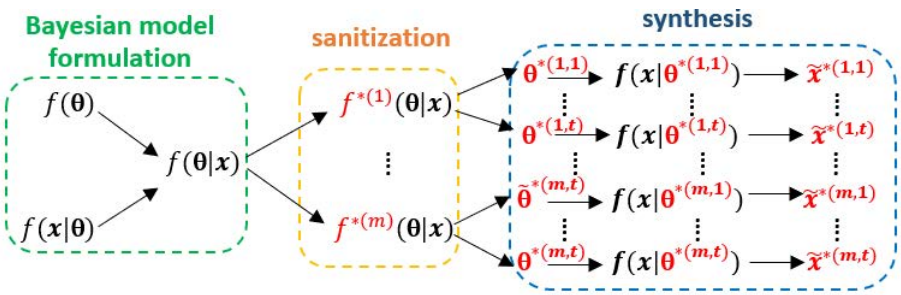


Figure A.1: The nested modips procedure

B Proof of Proposition 1

Proof. To show that the modips procedure in Algorithm 1 satisfied ϵ-DP, we follow the DP definition 1, which states that a randomized algorithm $\mathcal{R}$ is ϵ-differentially private if, for all

datasets $(\mathbf{x}, \mathbf{x}')$ that differ in one individual and all possible subsets Q to the output range of statistics $\mathbf{s}$ from $\mathcal{R}$, $\left|\log\left(\frac{\Pr(\mathcal{R}(\mathbf{s},\mathbf{x}))\in Q)}{\Pr(\mathcal{R}(\mathbf{s},\mathbf{x}'))\in Q)}\right)\right| \leq \epsilon$ for $\epsilon > 0$.

In the case of the modips procedure, the output is synthetic data $\tilde{\mathbf{x}}^{*(j)}$. Denote the total privacy budget by ϵ, the portion allocated to model selection by ϵ_0 and and the number of released dataset is m. Then the budget left to synthesizing each set is $\epsilon' = (\epsilon - \epsilon_0)/m$ DP. Per the DP definition, we aim to establish

$$e^{-\epsilon'} \Pr(\tilde{\mathbf{x}}^{*(j)} \in Q|\mathbf{x}') \leq \Pr(\tilde{\mathbf{x}}^{*(j)} \in Q|\mathbf{x}) \leq e^{\epsilon'} \Pr(\tilde{\mathbf{x}}^{*(j)} \in Q|\mathbf{x}'). \tag{B.1}$$

Since

$$\begin{aligned}
&\Pr(\tilde{\mathbf{x}}^{*(j)} \in Q|\mathbf{x}) = \mathrm{E}_{\tilde{\mathbf{x}}^{*(j)}}(\mathrm{E}_{\boldsymbol{\theta}^{*(j)}}(\mathbb{1}(\tilde{\mathbf{x}}^{*(j)} \in Q)|\boldsymbol{\theta}^{*(j)})|\mathbf{x}) \\
=& \int_{\tilde{\mathbf{x}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\tilde{\mathbf{x}}^{*(j)} \in Q) f(\tilde{\mathbf{x}}^{*(j)}|\boldsymbol{\theta}^{*(j)}) f(\boldsymbol{\theta}^{*(j)}|\mathbf{x}) d\boldsymbol{\theta}^{*(j)} d\tilde{\mathbf{x}}^{*(j)} \\
=& \int_{\tilde{\mathbf{x}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\tilde{\mathbf{x}}^{*(j)} \in Q) f(\tilde{\mathbf{x}}^{*(j)}|\boldsymbol{\theta}^{*(j)}) \frac{f(\boldsymbol{\theta}^{*(j)}|\mathbf{x})}{f(\boldsymbol{\theta}^{*(j)}|\mathbf{x}')} f(\boldsymbol{\theta}^{*(j)}|\mathbf{x}') d\boldsymbol{\theta}^{*(j)} d\tilde{\mathbf{x}}^{*(j)}.
\end{aligned}$$

$\boldsymbol{\theta}^{*(j)}$ is of $\epsilon' = (\epsilon - \epsilon_0)/m$-DP, that is, $e^{-\epsilon'} \leq \frac{f(\boldsymbol{\theta}^{*(j)}|\mathbf{x})}{f(\boldsymbol{\theta}^{*(j)}|\mathbf{x}')} \leq e^{\epsilon'}$, implying that

$$\begin{aligned}
& e^{-\epsilon'} \int_{\tilde{\mathbf{x}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\tilde{\mathbf{x}}^{*(j)} \in Q) f(\tilde{\mathbf{x}}^{*(j)}|\boldsymbol{\theta}^{*(j)}) f(\boldsymbol{\theta}^{*(j)}|\mathbf{x}') d\boldsymbol{\theta}^{*(j)} d\tilde{\mathbf{x}}^{*(j)} \\
\leq& \int_{\tilde{\mathbf{x}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\tilde{\mathbf{x}}^{*(j)} \in Q) \frac{f(\boldsymbol{\theta}^{*(j)}|\mathbf{x})}{f(\boldsymbol{\theta} s^{*(j)}|\mathbf{x}')} f(\tilde{\mathbf{x}}^{*(j)}|\boldsymbol{\theta}^{*(j)}) f(\boldsymbol{\theta}^{*(j)}|\mathbf{x}') d\boldsymbol{\theta}^{*(j)} d\tilde{\mathbf{x}}^{*(j)} \\
\leq& e^{\epsilon'} \int_{\tilde{\mathbf{x}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\tilde{\mathbf{x}}^{*(j)} \in Q) f(\tilde{\mathbf{x}}^{*(j)}|\boldsymbol{\theta}^{*(j)}) f(\boldsymbol{\theta}^{*(j)}|\mathbf{x}') d\boldsymbol{\theta}^{*(j)} d\tilde{\mathbf{x}}^{*(j)}.
\end{aligned}$$

Together with

$$\begin{aligned}
&\int_{\tilde{\mathbf{x}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\tilde{\mathbf{x}}^{*(j)} \in Q) f(\tilde{\mathbf{x}}^{*(j)}|\boldsymbol{\theta}^{*(j)}) f(\boldsymbol{\theta}^{*(j)}|\mathbf{x}') d\boldsymbol{\theta}^{*(j)} d\tilde{\mathbf{x}}^{*(j)} \\
=& \mathrm{E}_{\mathbf{s}^{*(j)}}(\mathrm{E}_{\boldsymbol{\theta}^{*(j)}}(\mathbb{1}(\tilde{\mathbf{x}}^{*(j)} \in Q)|\mathbf{s}^{*(j)})|\mathbf{x}') = \Pr(\tilde{\mathbf{x}}^{*(j)} \in Q|\mathbf{x}'),
\end{aligned}$$

we arrive at Eqn B.1

$$e^{-\epsilon'} \Pr(\tilde{\mathbf{x}}^{*(j)} \in Q|\mathbf{x}') \leq \Pr(\tilde{\mathbf{x}}^{*(j)} \in Q|\mathbf{x}) \leq e^{\epsilon} \Pr(\tilde{\mathbf{x}}^{*(j)} \in Q|\mathbf{x}').$$

■

C Proof of Proposition 2

Proof. In the i-th synthesis for $i = 1, \ldots, m$,

$$\begin{aligned}
&\Pr(\boldsymbol{\theta}^{*(j)} \in Q|\mathbf{x}) = \mathrm{E}_{\mathbf{s}^{*(j)}}(\mathrm{E}_{\boldsymbol{\theta}^{*(j)}}(\mathbb{1}(\boldsymbol{\theta}^{*(j)} \in Q)|\mathbf{s}^{*(j)})|\mathbf{x}) \\
=& \int_{\tilde{\mathbf{s}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\boldsymbol{\theta}^{*(j)} \in Q) f(\boldsymbol{\theta}^{*(j)}|\mathbf{s}^{*(j)}) f(\mathbf{s}^{*(j)}|\mathbf{x}) d\boldsymbol{\theta}^{*(j)} d\mathbf{s}^{*(j)} \\
=& \int_{\tilde{\mathbf{x}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\boldsymbol{\theta}^{*} \in Q) f(\boldsymbol{\theta}^{*}|\mathbf{s}^{*}) \frac{f(\mathbf{s}^{*}|\mathbf{x})}{f(\mathbf{s}^{*(j)}|\mathbf{x}')} f(\mathbf{s}^{*(j)}|\mathbf{x}') d\boldsymbol{\theta}^{*(j)} d\mathbf{s}^{*(j)}.
\end{aligned}$$

Since $\mathbf{s}^{*(j)}$ is obtained through a ϵ'-DP mechanism, $e^{-\epsilon'} \leq \frac{f(\mathbf{s}^{*(j)}|\mathbf{x})}{f(\mathbf{s}^{*(j)}|\mathbf{x}')} \leq e^{\epsilon'}$ and

$$e^{-\epsilon'} \int_{\tilde{\mathbf{s}}^{*(j)}}\int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\boldsymbol{\theta}^{*(j)} \in Q) f(\boldsymbol{\theta}^{*(j)}|\mathbf{s}^{*(j)}) f(\mathbf{s}^{*(j)}|\mathbf{x}') d\boldsymbol{\theta}^{*(j)} d\mathbf{s}^{*(j)}$$

$$\leq \int_{\tilde{\mathbf{s}}^{*(j)}} \int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\boldsymbol{\theta}^{*(j)} \in Q) \frac{f(\mathbf{s}^{*(j)}|\mathbf{x})}{f(\mathbf{s}^{*(j)}|\mathbf{x}')} f(\boldsymbol{\theta}^{*(j)}|\mathbf{s}^{*(j)}) f(\mathbf{s}^{*(j)}|\mathbf{x}) d\boldsymbol{\theta}^{*(j)} d\mathbf{s}^{*(j)}$$

$$\leq e^{\epsilon'} \int_{\tilde{\mathbf{s}}^{*(j)}} \int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\boldsymbol{\theta}^{*(j)} \in Q) f(\boldsymbol{\theta}^{*(j)}|\mathbf{s}^{*(j)}) f(\mathbf{s}^{*(j)}|\mathbf{x}') d\boldsymbol{\theta}^{*(j)} d\mathbf{s}^{*(j)}$$

Together with

$$\int_{\tilde{\mathbf{s}}^{*(j)}} \int_{\boldsymbol{\theta}^{*(j)}} \mathbb{1}(\boldsymbol{\theta}^{*(j)} \in Q) f(\boldsymbol{\theta}^{*(j)}|\mathbf{s}^{*(j)}) f(\mathbf{s}^{*(j)}|\mathbf{x}) d\boldsymbol{\theta}^{*(j)} d\mathbf{s}^{*(j)}$$

$$= \mathrm{E}(_{\mathbf{s}^{*(j)}} E_{\boldsymbol{\theta}^{*(j)}} (\mathbb{1}(\boldsymbol{\theta}^{*(j)} \in Q)|\mathbf{s}^{*(j)})|\mathbf{x}) = \Pr(\boldsymbol{\theta}^{*(j)} \in Q|\mathbf{x}),$$

similarly when $\mathbf{x}$ is replaced by $\mathbf{x}'$, we have

$$e^{-\epsilon} \Pr(\boldsymbol{\theta}^{*(j)} \in Q|\mathbf{x}') \leq \Pr(\boldsymbol{\theta}^{*(j)} \in Q|\mathbf{x}) \leq e^{\epsilon'} \Pr(\boldsymbol{\theta}^{*(j)} \in Q|\mathbf{x}'),$$

$$e^{-\epsilon'} \leq \frac{\Pr(\boldsymbol{\theta}^{*(j)} \in Q|\mathbf{x})}{\Pr(\boldsymbol{\theta}^{*(j)} \in Q|\mathbf{x}')} \leq e^{\epsilon'};$$

that is, releasing a random sample $\boldsymbol{\theta}^{*(j)}$ in the modips.SSS procedure satisfies ϵ'-DP. The rest of the proof is the same as the proof for Proposition 1, leading to the conclusion that sanitized data $\mathbf{x}$ from the modips.SSS satisfy DP. ■

D Proof of Proposition 3

Proof. $u_i = -\log\left(\int \mathbb{1}(\boldsymbol{\theta} \in \mathcal{B}_i) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}\right) = -(\log(\bar{f}_i(\boldsymbol{\theta}|\mathbf{x})) + \log(V_i))$ and $|u_i(\mathbf{x}) - u(\mathbf{x}')| = |\log(\bar{f}_i(\boldsymbol{\theta}|\mathbf{x})) - \log(\bar{f}_i(\boldsymbol{\theta}|\mathbf{x}'))| < 2A$. Therefore, $\Delta_u = \max_{\mathcal{B},\boldsymbol{\theta},d(\mathbf{x},\mathbf{x}')=1} |u_i(\mathbf{x}) - u(\mathbf{x}')| < 2A$. ■

E Proof of Proposition 4

Proof. Draw N samples from $f(\boldsymbol{\theta}|\mathbf{x})$ and $f(\boldsymbol{\theta}|\mathbf{x}')$, where $\mathbf{x}$ and $\mathbf{x}'$ are a pair of neighboring datasets, and form histograms based on the N samples in each case, where the two histograms share the same bin cut points. The number of bins in each histogram is $B = \prod_{j=1}^{p} B_j$, and the number of bins in the j-th marginal is B_j for $j = 1 \ldots, p$ with the widths of the bins $\mathbf{h}_j = (h_{1,j}, \ldots, h_{B_j,j})$.

Denote the counts N_i and N_i' in bin $\mathcal{B}_i$ in the two histograms, respectively. The maximum change in the log(proportion) of bin $\mathcal{B}_i$ given one-individual change from $\mathbf{x}$ to $\mathbf{x}'$ is $\Delta_1 = \max\{|\log(N_i/N) - \log(N_i'/N)|\}$. Since $N_i/N \approx \bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\boldsymbol{\theta}|\mathbf{x}) \prod_{j=1}^{p} h_{j(i),j}$ and $N_i'/N \approx \bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\boldsymbol{\theta}|\mathbf{x}') \prod_{j=1}^{p} h_{j(i),j}$, $\Delta_1 = \max\{|\log(\bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\boldsymbol{\theta}|\mathbf{x}) - \log(\bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\boldsymbol{\theta}|\mathbf{x}'))|\} < 2A$. ■

F Proof of Proposition 5

Proof. The proof is similar to that of Proposition 4 except for the last step. Specifically, let $\prod_{j=1}^{p} h_{j(i),j} = W_i$. Since $N_i \approx N \bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\boldsymbol{\theta}|\mathbf{x}) W_i$ and $N_i' \approx N \bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\boldsymbol{\theta}|\mathbf{x}') W_i$, $\Delta_i = \max\{|N_i - N_i'|\} \approx N W_i \max\{\bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\mathbf{x}) - \bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\mathbf{x}')\} < N V_i \max\{\bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\mathbf{x}), \bar{f}_{\boldsymbol{\theta}\in\mathbf{B_i}}(\mathbf{x}')\} < N W_i G$. ■

G Application of SiDD, SiDD.MC, and SiPHiC when $f(\boldsymbol{\theta}|\mathbf{x})$ is MVN)

The application of SiDD.MC is straightforward and only a upper bound on $|\log(f(\boldsymbol{\theta}|\mathbf{x}))|$.

For SiDD, in addition to an upper bound on $|\log(f(\boldsymbol{\theta}|\mathbf{x}))|$, one would calculate the volume of each bin in the discretized MVN distribution for the exponential mechanism in Algorithm 2, which does not change when the MVN is standardized (mean 0 and marginal variance is 1 for each dimension) as long as the cut points for the bin is also relocated and scaled. After the standardization, the MVN can be expressed as $\mathcal{N}(\mathbf{0}, \mathbf{r})$; the same bound $[-C, C]$ and the number of bins B can be applied to each of the p dimensions of $\boldsymbol{\theta}$. Set $C > 0$ at a large number so that there is ignorable probability mass outside $[-C, C]$; let $h = 2C/B$ and the set of cut points for the bins in each dimension be $\mathbf{C} = [-C, h-C, \cdots, C-h, C]$; and denote the left cut point for bin $\mathcal{B}_i$ in the j-th dimension by $\mathbf{C}[j(i)]$, where $j(i) = 1, \ldots, B+1$. The volume of $\mathcal{B}_i$ is $V_i = \Phi((\mathbf{C}[1(i)+1], \cdots, \mathbf{C}[p(i)]+1); \mathbf{0}, \mathbf{r}) - \Phi((\mathbf{c}[1(i)], \cdots, \mathbf{C}[p(i)]); \mathbf{0}, \mathbf{r})\}$, where Φ is the CDF of $\mathcal{N}(\mathbf{0}, \mathbf{r})$. The posterior correlation matrix $\mathbf{r}$ is a function of $\mathbf{x}$ and needs to be sanitized or specified independently of $\mathbf{x}$ using prior knowledge to save privacy cost. For example, since the larger the elements in $\mathbf{r}$ are, the larger Δ is, we may set all correlations in $\mathbf{r}$ at some rarely large value in practice to be conservative. This approach is employed in Liu et al. [63] to obtain privacy-preserving posterior samples of ERGM parameters fitted on network data.

For SiPHiC, in addition to an upper bound on $f(\boldsymbol{\theta}|\mathbf{x})$, one can easily calculate the sensitivity in Algorithm 4 by $\Delta_i = N \max_i\{V_i\} \leq N(\Phi(\mathbf{h}/2; \mathbf{0}, \mathbf{r}) - \Phi(-\mathbf{h}/2; \mathbf{0}, \mathbf{r}))$, where $\mathbf{h}_{p\times 1} = (h, \ldots, h)^T$. The specification of $\mathbf{r}$ is similar to that for the SiDD procedure.

H Proof of Theorem 7

Part a). The likelihood is $f(\mathbf{x}|\boldsymbol{\theta})$. Synthetic data $\tilde{\mathbf{x}}^*$ via the modips procedure is generated from $f(\mathbf{X}|\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^{*(j)}$ is a random sample from the sanitized posterior distribution $f^*(\boldsymbol{\theta}|\mathbf{x})$. We assume that $f^*(\boldsymbol{\theta}|\mathbf{x})$ is consistent for $f(\boldsymbol{\theta}|\mathbf{x})$. WLOS, suppose θ is a scalar. If the estimator of $\hat{\theta}$ based on original $\mathbf{x}$ is consistent (e.g., MLE, posterior mean) for the parameter of interest θ, then the same estimator $\hat{\theta}^*$ but based on the sanitized $\tilde{\mathbf{x}}^*$ is consistent for θ^* as the distribution that generates $\mathbf{x}$ and $\tilde{\mathbf{x}}^*$ are the same except the underlying parameter values. The mean squared error of $\hat{\theta}^*$ as an estimate for θ is

$$\begin{aligned} &\mathrm{E}_{\mathbf{X}^*}(\hat{\theta}^* - \theta)^2 = \mathrm{E}_{\mathbf{X}^*}(\hat{\theta}^* - \theta^* + \theta^* - \theta)^2 \\ &= \mathrm{E}_{\mathbf{X}^*}(\hat{\theta}^* - \theta^*)^2 + \mathrm{E}_{\mathbf{X}^*}(\theta^* - \theta)^2 + 2\mathrm{E}_{\mathbf{X}^*}[(\hat{\theta}^* - \theta^*)(\theta^* - \theta)] \\ &= \mathrm{E}_{\mathbf{X}^*}(\hat{\theta}^* - \theta^*)^2 + (\theta^* - \theta)^2 + 2(\theta^* - \theta)\mathrm{E}_{\mathbf{X}^*}(\hat{\theta}^* - \theta^*) \qquad \text{(H.1)} \\ &\to (\theta^* - \theta)^2 \text{ as } n \to \infty, \text{ which } \to 0 \text{ as } \epsilon \to \infty. \end{aligned}$$

The first and third terms $\mathrm{E}_{\mathbf{X}^*}(\hat{\theta}^* - \theta^*)^2$ and $\mathrm{E}_{\mathbf{X}^*}(\hat{\theta}^* - \theta^*) \to 0$ as $n \to \infty$ in Eqn (H.1) with the consistency of $\hat{\theta}^*$ for θ^*. θ^* in the second term $(\theta^* - \theta)^2$ is a draw from the sanitized posterior distribution $f^*(\theta|\mathbf{x})$. If there is no sanitization, $f(\theta|\mathbf{x})$ approaches a degenerate distribution at point θ as $n \to \infty$; in other words, $f(\theta|\mathbf{x}) \xrightarrow{d} \theta$ as $n \to \infty$. With sanitization, since $f^*(\theta|\mathbf{x}) \xrightarrow{d} f(\theta|\mathbf{x})$ as $\epsilon \to \infty$ per consistency (definition 7) and $f(\theta|\mathbf{x}) \xrightarrow{d} \theta$ as $n \to \infty$, then $f^*(\theta|\mathbf{x}) \xrightarrow{d} \theta$ as $n \to \infty$ and $\epsilon \to \infty$. Taken together, $(\theta^* - \theta)^2 \to 0$ as $n \to \infty$ and $\epsilon \to \infty$.

In cases where the parameter of inference interest on the users' side is β, rather than θ – the parameter in the sanitization and synthesis on the data curator's side, assume the distribution or model used by the users of the released sanitized data is congenial in a similar sense as in Meng [94], then β should be a function of θ; that is, $\beta = h(\theta)$ and $\beta^* = h(\theta^*)$. By the continuous mapping theorem, since $\hat{\theta}^*$ is consistent for θ as $n \to \infty$

and $\epsilon \to \infty$, so is $\hat{\beta}^*$ for β. With m set of synthetic data, if $\hat{\beta}^{*(j)}$ is consistent for β for $i = 1, \ldots, m$, so is $m^{-1}\sum_{i=1}^{m}\beta^{*(j)}$ per the Slutsky's theorem.

Part b). The proof is based in a similar framework as in Rubin [90] (inferences from multiple imputation) and Reiter [10] (inferences from partial sample synthesis without sanitization), with necessary modifications to allow for the extra variability introduced during the sanitization of the posterior distribution before sampling. We first provide s a Bayesian derivation of the inference and then list the conditions under which these inferences are valid from a frequentist perspective.

In the Bayesian framework, the posterior variance of β given synthetic data $\tilde{\mathbf{x}}^{*(j)}$ for $i = 1, \ldots, m$ is

$$\begin{aligned}&\mathrm{V}(\beta|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}) = \mathrm{V}(\mathrm{E}(\beta|\mathbf{x})|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}) + \mathrm{E}(\mathrm{V}(\beta|\mathbf{x})|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)})\\ =&\mathrm{V}(\hat{\beta}|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}) + \mathrm{E}(\hat{v}|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}),\end{aligned} \tag{H.2}$$

where $\hat{\beta}$ and $\hat{v}$ are the posterior mean and variance of β, respectively, given the original data $\mathbf{x}$; $\hat{\beta}^{*(j)}$ and $\hat{v}^{*(j)}$ are the posterior mean and variance of $\beta^{*(j)}$, respectively, given $\tilde{\mathbf{x}}^{*(j)}$. By the large-sample theory, as $n \to \infty$,

$$\beta|\mathbf{x} \sim N(\hat{\beta}, \hat{v}) \tag{H.3}$$

$$\beta^{*(j)}|\mathbf{x}^{*(j)} \sim N(\hat{\beta}^{*(j)}, \hat{v}^{*(j)}). \tag{H.4}$$

Since $\beta^{*(j)}$ is independent for $i = 1, \ldots, m$ conditional on $\tilde{x}^{*(j)}$, we have, from Eq. (H.4),

$$m^{-1}\textstyle\sum_{i=1}^{m}\beta^{*(j)}|\tilde{\mathbf{x}}^{*(j)} \sim N(m^{-1}\sum_{i=1}^{m}\hat{\beta}^{*(j)}, m^{-2}\sum_{i=1}^{m}\hat{v}^{*(j)}). \tag{H.5}$$

Since $f(\beta^*|\mathbf{x}) \xrightarrow{d} f(\beta|\mathbf{x})$, per the Lyapunov CLT, we have, as $m \to \infty$

$$\beta|\beta^{*1)}, \ldots, \beta^{*(m)} \sim N(m^{-1}\textstyle\sum_{i=1}^{m}\beta^{*(j)}, m^{-2}\sum_{j=1}^{m}v(\beta^{*(j)})) \tag{H.6}$$

Eqs (H.3), (H.5), and (H.6) taken together, it suggests

$$\hat{\beta}|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)} \sim N(m^{-1}\textstyle\sum_{i=1}^{m}\hat{\beta}^{*(j)}, m^{-2}\sum_{i=1}^{m}(v(\beta^{*(j)}) + \hat{v}^{*(j)})). \tag{H.7}$$

In a similar manner, we obtain the conditional distribution of $v(\mathbf{x})$ given $\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}$

$$v(\mathbf{x})|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)} \sim N(m^{-1}\textstyle\sum_{i=1}^{m}\hat{v}^{*(j)}, m^{-2}\sum_{i=1}^{m}(v_v(\beta^{*(j)}) + \hat{v}^{*(j)})). \tag{H.8}$$

Replace the two terms in Eq (H.2) with the conditional variance from Eq (H.7) and mean from Eq (H.8) and denote $m^{-1}\sum_{i=1}^{m}\left(v(\beta^{*(j)}) + \hat{v}^{*(j)}\right)$ by b and $m^{-1}\sum_{i=1}^{m}\hat{v}^{*(j)})$ by ϖ, then

$$\mathrm{V}(\beta|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}) = m^{-1}b + \varpi.$$

For finite m, b is approximated by $(m-1)^{-1}\sum_{i=1}^{m}(\hat{\beta}^{*(j)} - \bar{\beta}^*)^2$ and ϖ by $m^{-1}\sum_{i=1}^{m}\hat{v}^{*(j)}$.

Similar to Reiter [10], the regularity conditions for $m^{-1}b + \varpi$ being an asymptotically unbiased estimator for $\bar{\beta}^*$ in the frequentist framework include 1) $\mathrm{E}\big(\hat{\beta}^{*(j)}|\mathbf{x}\big) \to \hat{\beta}$; 2) $\mathrm{E}\big(m^{-1}\sum_{i=1}^{m}\hat{v}^{*(j)}|\mathbf{x}\big) \to \hat{v}$; and 3) $\mathrm{E}\big((m-1)^{-1}\sum_{i=1}^{m}(\hat{\beta}^{*(j)} - \bar{\beta}^*)^2|\mathbf{x}\big) \to \mathrm{V}(\beta^{*(j)}|\mathbf{x})$.

Part c). The results in parts a) and b) suggest $\beta|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)} \sim N(\bar{\beta}^*, m^{-1}b + \varpi)$. For a finite m, the distribution can be obtained in a similar manner as in Reiter [10], which is $f(\beta|\tilde{\mathbf{x}}^{*1)}, \ldots, \tilde{\mathbf{x}}^{*(m)}) \sim t_\nu(\bar{\beta}^*, m^{-1}b + \varpi)$ with $\nu = (m-1)(1 + m\varpi/b)^2$.

I Proof of Remark 6

Part a). In the communal sanitization, the scale parameter of the Laplace distribution is $\lambda = \delta_{\mathbf{s}}\epsilon^{-1} = r\bar{\delta}_{\mathbf{s}}\epsilon^{-1}$, where $\bar{\delta}_{\mathbf{s}}$ is the average GS. When $w_i \equiv r^{-1}$, every statistic receives the same amount of privacy budget ϵ/r in the individualized sanitization and the scale

parameter of the Laplace distribution for s_i is $\lambda' = \delta_i(\epsilon w_j)^{-1} = r\delta_i\epsilon^{-1}$, which is $< \lambda$ if $\delta_i < \bar{\delta}_{\mathbf{s}}$; and $> \lambda$ otherwise.

Part b). In the individualized sanitization, the scale parameter of the Laplace distribution for s_i is $\delta_i(\epsilon w_i)^{-1} = \epsilon^{-1}\delta_{s_i}(\delta_{s_i})^{-1}\sum_{i=1}^{r}\delta_i = \epsilon^{-1}\sum_{j=1}^{r}\delta_i$, the same as the scale parameter for the Laplace distribution in the communal sanitization.

References

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

[2] F. Liu and RJA Little. Selective multiple imputation of keys for statistical disclosure limitation in microdata. *Proceedings of 2002 American Statistical Association Joint Statistical Meeting*, 2002.

[3] RJA Little, F. Liu, and T. Raghunathan. Statistical disclosure techniques based on multiple imputation. In Andrew Gelman and Xiao-Li Meng, editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An essential journey with Donald Rubin's statistical family*, page Chapter II.13. John Wiley & Sons, 2004.

[4] J.P. Reiter. Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21:441–462, 2005.

[5] Di An and RJA Little. Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):923–940, 2007.

[6] G. Caiola and J. P. Reiter. Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, 3(1):27 – 42, 2010.

[7] J. Drechsler and J. P. Reiter. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic data sets. *Computational Statistics and Data Analysis*, 55(12):461–468, 2011.

[8] Lane F. Burgette and Jerome P. Reiter. Multiple-shrinkage multinomial probit models with applications to simulating geographies in public use data. *Bayesian Analysis*, 8(2): 453–478, 2013.

[9] Trivellore E Raghunathan, J. P. Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of official Statistics*, 19(1):1–16, 2003.

[10] Jerome P Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003.

[11] Josep Domingo-Ferrer and Vicenç Torra. Disclosure risk assessment in statistical data protection. *Journal of Computational and Applied Mathematics*, 164:285–293, 2004.

[12] Chris Skinner. Statistical disclosure risk: Separating potential and harm. *International Statistical Review*, 80(3):349–368, 2012.

[13] Stephen E Fienberg, Udi E Makov, and Ashish P Sanil. A bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics*, 13:75–79, 1997.

[14] J. Domingo-Ferrer and V. Torra. Disclosure control methods and information loss for microdata. In P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L.V. Zayatz, editors, *Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies*, pages 91–110. Elsevier, 2001.

[15] J. Domingo-Ferrer and V. Torra. Disclosure risk assessment in statistical data protection. *Journal of Computational and Appled Mathematics*, 164-165(1):285–293, 2004.

[16] J. P. Reiter. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100(472):1103–1112, 2005.

[17] J. Domingo-Ferrer and Y. Saygin, editors. *Privacy in statistical database*. Springer-Verlag Berlin Heidelberg, 2008.

[18] Daniel Manrique-Vallier and J. P. Reiter. Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, 107(500): 1385–1394, 2012.

[19] 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):2, 2014.

[20] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer, 2006.

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

[22] C. Dwork and A. Roth. *The Algorithmic Foundation of Differential Privacy*. Now Publishes, Inc., 2014.

[23] Fang Liu. Generalized gaussian mechanism for differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):747–756, 2019.

[24] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. *Proceeding of STOC '11 Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822, 2011.

[25] Jing Lei. Differentially private M-estimators. *Proceedings of Advances in Neural Information Processing Systems*, 24, 2011.

[26] Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. A near-optimal differentially private principal components. *The Journal of Machine Learning Research*, 14: 2905–2943, 2013.

[27] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *JMLR: Workshop and Conference Proceedings*, 12:1069–1109, 2011.

[28] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. *JMLR: Workshop and Conference Proceedings*, 23:25.1–25.40, 2012.

[29] Adam Smith and Abhradeep Thakurta. Differentially private model selection via stability arguments and the robustness of the lasso. *JMLR: Workshop and Conference Proceedings*, 30:1–32, 2013.

[30] Rob Hall, Alessandro Rinaldo, and Larry Wasserma. Differential privacy for functions and functional data. *Journal of Machine Learning Research*, 14:703–727, 2013.

[31] Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavkovic, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.

[32] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. *ACM CCS*, pages 1310–1321, 2015.

[33] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[34] Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with gaussian differential privacy. *Harvard data science review*, 23, 2020.

[35] 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*, pages 273–282. ACM, 2007.

[36] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 609–618. ACM, 2008.

[37] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*, 25, 2012.

[38] Claire McKay Bowen, Fang Liu, and Bingyue Su. Differentially private data release via statistical election to partition sequentially. *METRON*, 79(1):1–31, 2021.

[39] Evercita C. Eugenio and Fang Liu. Construction of differentially private empirical distributions from a low-order marginals set through solving linear equations with l_2 regularization. In Kohei Arai, editor, *Intelligent Computing*, pages 949–966. Springer International Publishing, 2021.

[40] John M Abowd and Lars Vilhuber. How protective are synthetic data? In *Privacy in Statistical Databases*, pages 239–246. Springer, 2008.

[41] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. *IEEE ICDE 24th International Conference*, pages 277 – 286, 2008.

[42] A. S. Charest. How can we analyze differentially private synthetic datasets. *Journal of Privacy and Confidentiality*, 2(2):Article 3, 2010.

[43] David McClure and Jerome P Reiter. Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Transactions on Data Privacy*, 5 (3):535–552, 2012.

[44] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1423–1434. ACM, 2014.

[45] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[46] H Li, L Xiong, and X Jiang. Differentially private synthesization of multi-dimensional data using copula functions. *Advances in Database Technology*, pages 475–486, 2014.

[47] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[48] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[49] Gergely Acs, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. Differentially private mixture of generative neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1109–1121, 2018.

[50] Junyoung Kang, Sooyong Jeong, Dowon Hong, and Changho Seo. A study on synthetic data generation based safe differentially private gan. *Journal of the Korea Institute of Information Security & Cryptology*, 30(5):945–956, 2020.

[51] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 510–526. Springer, 2018.

[52] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018.

[53] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

[54] Chris Waites and Rachel Cummings. Differentially private normalizing flows for privacy-preserving density estimation. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2020.

[55] Harrison Quick. Generating poisson-distributed differentially private synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3):1093–1108, 2021.

[56] Davide Proserpio, Sharon Goldberg, and Frank McSherry. A workflow for differentially-private graph synthesis. *Proceedings of the 2012 ACM workshop on online social networks*, pages 13–18, 2012.

[57] Yue Wang and Xintao Wu. Preserving differential privacy in degree-correlation based graph generation. *Transactions on Data Privacy*, 6:127–145, 2013.

[58] Qian Xiao, Rui Chen, and Kian-Lee Tan. Differentially private network data release via structural inference. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 911–920, 2014.

[59] Dai Li, Wei Zhang, and Yunfang Chen. Differentially private network data release via stochastic kronecker graph. In *International Conference on Web Information Systems Engineering*, pages 290–297. Springer, 2016.

[60] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914, 2013.

[61] Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M. Procopiuc, and Divesh Srivastava. Dpt: Differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165, 2015.

[62] Vishesh Karwa, Pavel N. Krivitsky, and Aleksandra B. Slavković. Sharing social network data: differentially private estimation of exponential family random-graph models. *Applied Statistics (JRSS-C)*, page DOI: 10.1111/rssc.12185, 2016.

[63] Fang Liu, Evercita Eugenio, Ick Hoon Jin, and Claire Bowen. Differentially private generation of social networks via exponential random graph models. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1695–1700. IEEE, 2020.

[64] Vincent Bindschadler. *Privacy-preserving seedbased data synthesis*. PhD thesis, University of Illinois at Urbana-Champaign, 2018.

[65] Luca Melis. *Building and evaluating privacy-preserving data processing systems*. PhD thesis, UCL (University College London), 2018.

[66] Claire McKay Bowen. *Data Privacy via Integration of Differential Privacy and Data Synthesis*. PhD thesis, University of Notre Dame, 2018.

[67] Evercita Cuevas Eugenio. *Some Methods for Differentially Private Data Synthesis*. PhD thesis, University of Notre Dame, 2019.

[68] Vishesh Karwa and Aleksandra B. Slavković. Inference using noisy degrees: differentially private β-model and synthetic graphs. *Annals of Statistics*, 44 (1):87–112, 2015.

[69] Cynthia Dwork. Differential privacy. In *Proceedings of the International Colloqium on Automata, Languages and Programming (ICALP)*, pages 1–12. Springer-Verlag ARCoSS, 2006.

[70] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204, 2011.

[71] Cynthia Dwork, K. Kenthapadi, Frank McSherry, I. Mironov, and M. Naor. Our data, ourselves: privacy via distributed noise generation. In *Advances in Cyptology: Proceedings of EUROCRYPT*, pages 485–503. Springer Berlin Heidelberg, 2006.

[72] Rob Hall, Alessandro Rinaldoy, and Larry Wasserman. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2):43–59, 2012.

[73] Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *arXiv:1603.01887v2*, 2016.

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

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

[76] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

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

[78] Daniel Kifer and Bing-Rong Lin. An axiomatic view of statistical privacy and utility. *Journal of Privacy and Confidentiality*, 4(1), 2012.

[79] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

[80] RJA Little. Statistical analysis of masked data. *Journal of the Official Statistics*, 9:407–407, 1993.

[81] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502, 2015.

[82] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 291–305. Springer, 2014.

[83] Zhifa Liu, Brandon Malone, and Changhe Yuan. Empirical evaluation of scoring functions for bayesian network model selection. In *BMC bioinformatics*, volume 13, pages 1–16. Springer, 2012.

[84] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, Peter Solenberger, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.

[85] D. Blackwell and R. V. Ramamoorthi. A bayes but not classically sufficient statistic. *Annals of Statistics*, 10(3):1025–1026, 1982.

[86] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, 1994.

[87] A.G. Nogales, J.A. Oyola, and P. Perez. On conditional independence and the relationship between sufficiency and invariance under the bayesian point of view. *Statistics & Probability Letters*, 46(1):75–84, 2000.

[88] Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *The Journal of Machine Learning Research*, 14(1):703–727, 2013.

[89] Fang Liu. Statistical properties of sanitized results from differentially private laplace mechanism with univariate bounding constraints. *Transactions on Data Privacy*, 12: 169–195, 2019.

[90] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

[91] Gillian M Raab, Beata Nowok, and Chris Dibben. Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7:4, 2017.

[92] T. Chanyaswad, C. Liu, and P. Mittal. Ron-gauss: Enhancing utility in non-interactive private data release. In *Proceedings on Privacy Enhancing Technologies (PETS)*, 2019.

[93] Claire McKay Bowen and Fang Liu. Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2):280–307, 2020.

[94] Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558, 1994.