Comparative Study of Differentially Private Data Synthesis Methods  
Claire McKay Bowen* and Fang Liu‡ * 

Abstract

When sharing data among researchers or releasing data for public use, there is a risk of exposing sensitive information of individuals in the data set. Data synthesis (DS) is a statistical disclosure limitation technique for releasing synthetic data sets with pseudo individual records. Traditional DS techniques often rely on strong assumptions of a data intruder’s behaviors and background knowledge to assess disclosure risk. Differential privacy (DP) formulates a theoretical approach for a strong and robust privacy guarantee in data release without having to model intruders’ behaviors. Efforts have been made aiming to incorporate the DP concept in the DS process. In this paper, we examine current Differentially Private Data Synthesis (DIPS) techniques for releasing individual-level surrogate data for the original data, compare the techniques conceptually, and evaluate the statistical utility and inferential properties of the synthetic data via each DIPS technique through extensive simulation studies. Our work sheds light on the practical feasibility and utility of the various DIPS approaches, and suggests future research directions for DIPS.

keywords: differential privacy, DIPS, sufficient statistics, parametric DIPS, non-parametric DIPS, statistical disclosure limitation

* Claire McKay Bowen is a graduate student, and Fang Liu is a Huisking Foundation, Inc. Assistant Professor in the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556 (E-mail: fang.liu.131@nd.edu). Claire McKay Bowen is supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. DGE-1313583. Fang Liu is supported by the NSF Grant 1546373 and University of Notre Dame Faculty Research Support Initiation Grant Program.
AMS 1991 subject classifications: Primary-60-02; secondary-62-99
An earlier version of this paper won the Best Student Paper Competition in the 2017 American Statistical Association Survey Research Methods Section (SRMS), Government Statistics Section (GSS), and Social Statistics Section (SSS).
1 Introduction

When sharing data among collaborators or releasing data publicly, a big concern is the risk of exposing the identification and personal information of the individuals who contribute to the data. Even with key identifiers removed, a data intruder may still identify an individual in a data set via linkage with other public information. Some notable examples on individual identification breach in publicly released or restricted access data include the Netflix prize (Narayanan and Shmatikov, 2008), the genotype and HapMap linkage effort (Homer et al., 2008), the AOL search log release (Götz et al., 2012), and the Washington State health record identification (Sweeney, 2013).

Statistical approaches to protecting data privacy are referred to as statistical disclosure limitation (SDL). SDL techniques aim to provide protection for sensitive information while releasing information and data to the public. Data synthesis (DS) is a SDL technique that focuses on releasing individual-level data synthesized based on the information in the original data (Drechsler, 2011). To propagate the uncertainty arising from the synthesis process, multiple synthetic sets of the identical structure are often released. This procedure is referred to as multiple synthesis (MS). Methods have been developed to combine the results from multiple synthetic data sets to will yield valid statistical inferences (Raghunathan et al., 2003; Reiter, 2002, 2003). However, existing disclosure risk assessment approaches for SDL techniques often depend on the specific values in a given data set as well as various assumptions about the background knowledge and behaviors of data intruders (Reiter, 2005; Hundepool et al., 2012; Manrique-Vallier and Reiter, 2012). In some cases, only heuristic arguments are employed without numerical assessment of disclosure risk.

Differential privacy (DP), a concept popularized in the theoretical computer science community, quantifies privacy risk as a privacy budget and provides strong privacy guarantee in mathematical terms without making assumptions about the background knowledge of data intruders (Dwork et al., 2006b; Dwork, 2008, 2011). This privacy guarantee holds regardless of the amount of background information a data user possesses about the individual. DP has spurred a great amount work in developing differentially private mechanisms in general settings (Dwork et al., 2006b; McSherry and Talwar, 2007; McSherry, 2009; Nissim and Stemmer, 2015) as well as for specific statistical analysis such as data mining (Mohammed et al., 2011), shrinkage regression (Chaudhuri et al., 2011; Kifer et al., 2012), principle component analysis (Chaudhuri et al., 2012), genetic association tests (Yu et al., 2014), Bayesian learning (Wang et al., 2015), location privacy (Xiao and Xiong, 2015), recommender systems (Friedman et al., 2016), deep learning (Abadi et al., 2016), among others. A few research groups created software or web-based interfaces to allow generating or sharing differentially private statistics. At Google, Erlingsson et al. (2014) developed a Randomized Aggregatable Privacy-Preserving Ordinal Response or RAPPOR, an end-user client software on Chrome browser data for differentially private crowdsourcing statistics. Wang et al. (2016)’s RescueDP, an online aggregate monitoring scheme, publishes real-time popu-
lation statistics on spatial-temporal, crowdsourced data from mobile phone users with DP. Gaboardi et al. (2016) created Private data Sharing Interface (PSI) to allow researchers in the social sciences and other fields to share sensitive data while providing a strong privacy guarantee with DP.

DP was originally developed and is the most widely used for releasing aggregate or summary statistics to answering queries submitted to a database. Query-based data release has several shortcomings. The requirement to prespecify the level of privacy budget $\epsilon$ often dictates the number and the types of future queries. The curator of a database will refuse to answer any further queries if the prespecified privacy budget is exhausted from answering all previous queries. Additionally, data users would prefer to directly access the individual-level data to perform statistical analysis on their own.

Efforts have also been made to release differentially private individual-level data, which we will refer to as DIPS (Differentially Private Data Synthesis). Barak et al. (2007) generated synthetic data via the Fourier transformation and linear programming in low-order contingency tables. Blum et al. (2013) discussed differentially private data synthesis from the perspective of the learning theory. Abowd and Vilhuber (2008) proposed an approach to synthesize differentially private tabular data from the predictive posterior distributions of frequencies, which was applied in the simulations studies in Charest (2010) to explore inferences on proportions from synthesized binary data. McClure and Reiter (2012) implemented a similar technique for synthesizing binary data with a different specification of the differentially private prior. Wasserman and Zhou (2010) proposed several paradigms to sample from appropriately differentially private perturbed histograms or empirical distribution functions. They also examined the rate that the probability of empirical distribution of the synthetic data converges to the true distribution of the original data. Zhang et al. (2014) proposed PrivBayes to release high-dimensional data from Bayesian networks with binary nodes and low-order interactions among the nodes. Li et al. (2014) developed DPCopula for synthesizing multivariate data by using Copula functions to take into account the dependency structure. Liu (2016b) proposed a Bayesian technique, model-based DIPS (MODIPS), to release differentially private synthetic data, and explored the inferential properties of the released data. Besides these generic DIPS approaches, there are also DIPS developed for specific type of data such as graphs (Proserpio et al., 2012), edges in social network data (Karwa et al., 2016), mobility data from GPS trajectories (Chen et al., 2013; He et al., 2015).

The goals of this paper are two-fold. First, it introduces the powerful concept of DP to the statistical community and surveys the current development in DIPS. Second, this paper examines and compares some of the generic DIPS approaches based on the statistical and inferential utility of the respective synthesized data; both conceptually and empirically via simulation studies. We aim to, through this comparative examination of different DIPS approaches, demonstrate the useful applications of DP in releasing synthetic data with guaranteed privacy and to provide some guidance on
the feasibility of the DIPS methods for practical use.

The remainder of the paper is organized as follows. Section 2 overviews the basic concepts of DP and some common differentially private mechanisms. Section 3 presents some currently available DIPS approaches. Section 4 compares and examines the utility and inferential properties of the individual-level surrogate data released from the DIPS methods introduced in Section 3 via four simulation studies on different types of data. Concluding remarks are given in Section 5.

2 Concepts

The concepts of DP and the sanitization algorithms were developed originally for releasing results of queries sent to a database. We rephrase the main concepts in DP below in terms of statistics. There is essentially no difference between query results and statistics given that both are functions of data. Denote the target data for protection by \( x = \{x_{ij}\} \) of dimension \( n \times p \) \((i = 1, \ldots, n; j = 1, \ldots, p)\). Each row \( x_i \) represents an individual record with \( p \) variables/attributes.

2.1 differential privacy (DP) and composition properties

Definition 1. Differential Privacy (Dwork et al., 2006b): A sanitization algorithm \( R \) gives \( \epsilon \)-DP if for all data sets \((x, x')\) that is \( d(x, x') = 1 \), and all results \( Q \subseteq \mathcal{T} \)

\[
\left| \log \left( \frac{\Pr(R(s(x)) \in Q)}{\Pr(R(s(x')) \in Q)} \right) \right| \leq \epsilon, \tag{1}
\]

where \( \mathcal{T} \) denotes the output range of the algorithm \( R \), \( \epsilon > 0 \) is the privacy “budget” parameter, and \( s \) denotes the statistics. \( d(x, x') = 1 \) implies that \( x' \) differs from \( x \) by only one individual. Mathematically, Eqn. (1) states that the probability of obtaining the same query result to a query sent to sanitized \( x \) and \( x' \) is roughly the same. In layman’s terms, DP means the chance an individual will be identified based on the sanitized query result is very low since the query results would be about the same with or without the individual in the database. The degree of “about the same” is determined by the value of the privacy budget \( \epsilon \). As \( \epsilon \) becomes smaller, the probabilities of obtaining the same queries from \( x \) and \( x' \) will become more similar. DP provides a robust privacy guarantee, because it does not rely on any background knowledge or behavioral assumptions of a data intruder.

Each new differentially private statistic calculated over the same data set results in a cumulative cost to privacy. Therefore, the data curator must track all queries and analysis conducted on a data set to ensure the overall privacy spending does not exceed the prespecified level. The following composition theorems can be applied when releasing multiple statistics from a data set to maintain \( \epsilon \)-DP.

Theorem 2. Composition Theorems (McSherry, 2009): Suppose a differentially private mechanism \( R_j \) provides \( \epsilon_j \)-DP for \( j = 1, \ldots, r \).
a) **Sequential Composition**: The sequence of $R_j(x)$ executed on the same data set $x$ provides $(\sum_j \epsilon_j)$-DP.

b) **Parallel Composition**: Let $D_j$ be disjoint subsets of the input domain $D$. The sequence of $R_j(x \cap D_j)$ provides $\max(\epsilon_j)$-DP.

For example, assume $r$ queries are sent to the same data set with a total privacy budget of $\epsilon$. The data curator could allocate $\epsilon/r$ privacy budget to each of the $r$ queries per the sequential composition to maintain the privacy budget at $\epsilon$. In the setting explored by the parallel composition, no overlapping information is requested by different queries calculated from disjoint subsets of the data set; thus, the privacy cost does not accumulate. A typical example of the parallel composition being applied is the release of a histogram, where the counts in different bins of the histogram are based on disjoint subsets of data. This results in each bin being perturbed with the full privacy budget $\epsilon$.

### 2.2 differentially private mechanisms

We introduce two commonly used sanitizers to achieve $\epsilon$-DP: the Laplace mechanism and the Exponential mechanism. A key concept in the Laplace mechanism is the global sensitivity (GS) of $s$ (Dwork et al., 2006b), defined as the following: For all $(x, x')$ that is $d(x, x') = 1$, the global sensitivity of statistics $s$ is $\Delta_s = \max_{x,x',d(x,x')=1} ||s(x) - s(x')||_1$.

In layman’s terms, $\Delta_s$ is the maximum change in terms of $l_1$ norm a person would expect in $s$ across all possible configurations of $(x, x')$ and $d(x, x') = 1$. The sensitivity is “global” since it is defined for all possible data sets and all possible ways that two data sets differ by one observation. The higher $\Delta_s$ is the more disclosure risk there will be on the individuals in the data from releasing the original $s$.

**Definition 3. Laplace Mechanism** (Dwork et al., 2006b): The Laplace mechanism of $\epsilon$-DP adds independent noises $e$ from the Laplace distribution with location parameter 0 and scale parameter $\Delta_s \epsilon^{-1}$ to each of the elements of the original result $s$ to generate perturbed result $s^* = s + e$.

By the Laplace distribution, values closer to the raw results $s$ have higher probabilities of being released than those that are further away from $s$. The variance of the Laplace distribution is $2(\Delta_s \epsilon^{-1})^2$, implying the smaller the privacy budget $\epsilon$ and/or the larger the $\Delta_s$, the higher the probability that the perturbed result $s^*$ will be farther way from $s$ when released.

The Laplace mechanism is a quick and simple DP mechanism, but does not apply to all statistics such as statistics that have non-numerical outputs. McSherry and Talwar (2007) introduces a more general DP mechanism, the Exponential mechanism, that applies to all types of queries.

**Definition 4. Exponential Mechanism** (McSherry and Talwar, 2007): In the Ex-
ponential mechanism, a utility function \( u \) assigns a score to each possible output \( s^* \) and releases \( s^* \) with probability

\[
\frac{\exp \left( \frac{u(s^*|x) - u(s^*|x')}{{2\Delta}_u} \right)}{\int \exp \left( \frac{u(s'|x) - u(s^*|x)}{{2\Delta}_u} \right) ds'}
\]

to ensure \( \epsilon \)-DP, where \( {\Delta}_u = \max_{x,x',d(x,x')=1} |u(s^*|x) - u(s^*|x')| \) is the maximum change in score \( u \) with one row change in the data (if \( s^* \) is discrete, the integral in Eqn (2) is replaced with summation).

Per the Exponential mechanism, the probability of returning \( s^* \) increases exponentially with the utility score. For example, if \( s \) is numerical and the goal is to preserve as much original information as possible, metrics measuring the closeness between \( s^* \) and the original \( s \) are good candidates for \( u \) such as the negative \( p \)-norm distance

\[-||s - s^*||_p = -\left( \sum_{j=1}^r |s_j - s^*_j|^p \right)^{1/p}\] (Liu, 2016a). When the \( L_1 \) norm is used, the Exponential mechanism in Definition 4 becomes the Laplace mechanism with halved privacy budget (McSherry and Talwar, 2007; Liu, 2016a). Both the Laplace mechanism and Exponential mechanism are widely applied in developing more complicated mechanisms, such as the multiplicative weight approach of generating synthetic discrete data iteratively (Hardt and Rothblum, 2010) and the median mechanism for efficiently releasing correlated queries (Roth and Roughgarden, 2010).

Besides the Laplace mechanism and the Exponential mechanism, there are other sanitizers for general settings, such as the Gaussian mechanism that adds Gaussian noise to satisfy a softer version of DP (Section 2.3) (Dwork and Roth, 2013; Liu, 2016a) and the generalized Gaussian mechanisms (GGM) that include the Laplace mechanism and the Gaussian mechanism as special cases (Liu, 2016a).

### 2.3 relaxations of \( \epsilon \)-DP

One criticism of the pure \( \epsilon \)-DP in Section 2.1 is the potentially large amount of noise being added to query results to achieve a high level of privacy guarantee. This concern has motivated work on relaxing the pure \( \epsilon \)-DP. We provide a brief overview on three relaxed forms: approximate differential privacy (aDP), probabilistic differential privacy (pDP), and concentrated differential privacy (cDP).

**Definition 5. Approximate Differential Privacy** (Dwork et al., 2006a): A sanitization algorithm \( \mathcal{R} \) gives \((\epsilon, \delta)\)-aDP if for all data sets \((x, x')\) that are \(d(x, x')=1\),

\[
\Pr(\mathcal{R}(s(x)) \in Q) \leq \exp(\epsilon) \Pr(\mathcal{R}(s(x')) \in Q) + \delta,
\]

where \( \delta > 0 \) is typically chosen based on the sample size of the data set \( n \) that satisfies \( \delta(n)/n^r \to 0 \) for all \( r > 0 \). The pure \( \epsilon \)-DP is a special case of aDP when \( \delta = 0 \).
Definition 6. Probabilistic Differential Privacy (Machanavajjhala et al., 2008): A sanitization algorithm \( \mathcal{R} \) gives \((\epsilon, \delta)\)-pDP if
\[
\Pr (\mathcal{R}(s(x)) \in \text{Disc}(x, \epsilon)) \leq \delta
\]
for all data sets \((x, x')\) that are \(d(x, x') = 1\), where \(\text{Disc}(x, \epsilon)\) is the disclosure set \(\mathcal{R}(s(x))\) such that
\[
\left| \ln \left( \frac{\Pr(\mathcal{R}(s(x)) \in Q)}{\Pr(\mathcal{R}(s(x')) \in Q)} \right) \right| > \epsilon.
\]
Eqn (4) can be interpreted as the pure \(\epsilon\)-DP fails with probability \(\delta\).

Definition 7. Concentrated Differential Privacy (Dwork and Rothblum, 2016): For all data sets \((x, x')\) that is \(d(x, x') = 1\), a sanitization algorithm \( \mathcal{R} \) gives \((\mu, \tau)\)-cDP if \(D_{\text{subG}}(\mathcal{R}(x) \| \mathcal{R}(x')) \preceq (\mu, \tau)\), where \(D_{\text{subG}}\) stands for subGaussian divergence, defined as follows: two random variables \(Y\) and \(Z\) are \(D_{\text{subG}}(Y \| Z) \preceq (\mu, \tau)\) if and only if \(\mathbb{E}(L_{Y \| Z}) \leq \mu\) and the centered distribution of \((L_{Y \| Z} - \mathbb{E}(L_{Y \| Z}))\) is defined and \(\tau\)-subgaussian, where \(L_{Y \| Z} \ln (p(Y)/p(Z))\) is the privacy loss random variable.

Both pDP and cDP regard privacy loss as random variables, but cDP has some advantages over pPD. First, cDP has a bounded expected privacy loss whereas pDP has an infinite privacy loss with probability \(\delta\). Second, cDP has better accuracy without compromising the privacy loss from multiple inquiries (Dwork and Rothblum, 2016).

Since most of the methods in Section 3 originally implemented the pure DP, our simulations in Section 4 are based in the setting of the pure DP. Possible future work could explore the trade-off between privacy guarantee and accuracy when applying the various DP relaxations.

3 Differentially private data synthesis (DIPS)

We loosely group the currently available DIPS methods into two categories: the non-parametric approach (NP-DIPS) and the parametric approach (P-DIPS). In the NP-DIPS approach, the synthesizer is constructed based on the empirical distribution of the data. In the P-DIPS approach, the synthesizer is built based on a parametric distribution or an appropriately defined model of the data.

3.1 Non-parametric DIPS (NP-DIPS)

The statistics \(s\) targeted for sanitization are the cell proportions in some types of cross-tabulation for the categorical data in the framework of NP-DIPS. In the case of continuous data, the NP-DIPS techniques can be applied to generate differentially private smoothed density histograms, perturbed density histograms, or to release differentially private empirical distributions via the Exponential mechanism. The summary of the NP-DIPS covered in section is given in Table 1.
Table 1: Summary of Non-parametric DIPS approaches discussed in Section 3.1.

| Method                        | Data Type     | Pros/Cons                                                                 |
|-------------------------------|---------------|---------------------------------------------------------------------------|
| Laplace sanitizer             | counts        | simple and fast / not accurate for large number of queries               |
| Fourier transformation        | contingency   | preserves low-order marginals fairly accurately / computationally expensive as the number of attributes increases |
| multiplicative weight         | contingency   | embarrassingly parallel / less computationally efficient when queries exponentially exceeds data size |
| Exponential mechanism         | contingency   | embryarrassingly parallel / less computationally efficient when queries exponentially exceeds data size |
| histogram with post-hoc sorting | histogram    | more accurate results / possible leak of original information            |
| DPCube                        | histogram     | multidimensional data / inefficiency in constructing accurate high-dimensional histograms |
| NoiseFirst and StructureFirst | histogram     | outperforms several other DP methods/low dimensional histograms          |
| Exponential Fourier perturbation and P-HPartition | histogram | outperforms several other DP methods including NoiseFirst and StructureFirst / depends on histogram compressibility |
| perturbed histogram           | histogram     | computationally efficient / discretization on continuous attributes; doesn’t preserve correlation well |
| smoothed histogram            | histogram     | computationally efficient / discretization on continuous attributes; worse than perturbed histogram in information preservation |
| empirical CDF via Exponential mechanism | histogram | flexible; general / likely to computationally infeasible |

### 3.1.1 Sanitization of categorical data based on contingency tables

In a data set with \( p \) categorical variables, a straightforward approach in generating synthetic data is to add Laplace noise to the cell counts of \( k \)-way cross-tabulation of \( x \), where \( k \leq p \), and then to generate individual level of data from the sanitized counts. If \( k = p \), it is the full cross-tabulation of \( x \), and the individual-level data are straightforward to generate from sanitized counts. If \( k < p \), there are \( \binom{p}{k} \) \( k \)-way contingency tables, and the sanitization process needs to be carefully planned so that all \( k \)-way tables are consistent to yield legitimate marginals and individual-level data.

When \( k = p \), denote the original frequencies of the \( K \) cells formed by the \( p \)-way cross-tabulation of \( x \) by \( \mathbf{n} = (n_1, \ldots, n_K) \). The Laplace sanitizer perturbs the original \( \mathbf{n} \) via \( \mathbf{n}' = \mathbf{n} + \mathbf{e} \), where \( e_j \sim \text{Lap}(0, \Delta_s/\epsilon) \) independently for \( j = 1, \ldots, K \). \( \Delta_s \) is the \( l_1 \) GS from releasing the whole cross-tabulation. \( \Delta_s \) can be set at 2 or 1, depending on
how \( d(\mathbf{x}, \mathbf{x}') = 1 \) is defined (data change in one observation while \( n \) remains the same, thus \( \Delta_s = 2 \); versus removal of one observation, thus \( \Delta_s = 1 \). For the remainder of the paper, we assume \( \Delta_s = 1 \).

When \( k < p \), Barak et al. (2007) conducted early work on constructing \( k \)-way differentially private, consistent, and non-negative contingency tables via a Fourier transformation. The approach identifies the complete set of metrics required to reproduce a contingency table, where each cell is perturbed to achieve the same level of accuracy. However, the algorithm depends on the linear programming and could be a computationally infeasible when \( p \) is large. A less computationally intensive approach to generate individual-level data in the discrete domain is the multiplicative weight Exponential mechanism (MWEM) approach based on linear queries (Hardt et al., 2012). The MWEM approach approximates the original distribution in a differentially private manner through an iterative process. It starts from a uniform distribution over the supports of all the attributes in the original data, and then updates the distribution via multiplicative weighting based on a query sampled via the Exponential mechanism and sanitized via the Laplace mechanism in each iteration. The method is embarrassingly parallel, but is less computationally efficient when the number of queries exponentially exceeds the data set size (Gupta et al., 2012).

\subsection{Sanitization of numerical data based on histograms and empirical distributions}

A straightforward approach for releasing differentially private numerical data is to first generate differentially private histograms, and then synthesize numerical data by drawing a bin according to the relative sanitized frequencies of the histogram bins, and, lastly, sampling data points from the uniform distributions bounded by the sampled bin endpoints in the previous step.

To form histograms on the original numerical data, discretization is necessary. And there could be a large number of data bins/cubes if high-order interactions exist among the data attributes and are taken into account when the histogram is generated. Let \( K \) be the total number of bins (or squares/cubes in the multidimensional case), \( n_k = \sum_{k=1}^{K} I(x_i \in B_k) \) be the number of observations in \( B_k \) for \( k = 1, \ldots, K \), \( \hat{p}_k = n_k/n \), and \( I() \) be the indicator function (\( I(x_i \in B_k) = 1 \) if \( x_i \in B_k \); 0 otherwise), a mean-squared consistent density histogram estimator is \( \hat{f}_K(x) = \sum_{k=1}^{K} \hat{p}_k I(x \in B_k) \) (Scott, 2015). A differentially private perturbed histogram estimator that satisfies \( \epsilon \)-DP is thus

\[
\hat{f}_K^*(x) = \sum_{k=1}^{K} K \hat{p}_k^* I(x \in B_k).
\]

Note that sanitized \( n_k^* \) can be negative since the Laplace noise \( \sim \text{Lap}(0, \Delta_s/\epsilon) \) with \( \Delta_s = 1 \).
replacing negative \( n_k^* \) with 0 (Barak et al., 2007) or using the truncated or boundary inflated truncated Laplace distributions to obtain legitimate data (Liu, 2016c). To incorporate the uncertainty introduced by the sanitization process, releasing multiple sets of \( \tilde{x} \) is suggested, one set per sanitized \( n^* = \{n_k^*\}_{1:K} \).

There are various extensions to the basic perturbed histogram approach with the purposes to improve its accuracy. Hay et al. (2010) suggested boosting the accuracy of differentially private histograms by sorting the bin values before and after sanitation. However, the sorting process itself leaks some original information and the actual privacy cost. Thus, it exceeds the prespecified privacy budget. Xiao et al. (2012) applied a 2-phase partitioning strategy, DPCube, for multidimensional data cubes or histograms. Gardner et al. (2013) implemented DPCube for biomedical data to demonstrate its practical feasibility on real-world data sets, but found that DPCube was still inefficient in constructing accurate high-dimensional histograms. Xu et al. (2013) proposed two mechanisms, NoiseFirst and StructureFirst, that performed well against some DP methods, but only applied to low dimensional histograms. Acs et al. (2012) presented two sanitization techniques, the Exponential Fourier perturbation algorithm and the P-HPartition, that sanitize compressed data to exploit the inherent redundancy of real-life data sets. From the experimental results, the techniques outperformed some DP methods, including NoiseFirst and StructureFirst, but the performance depended on the compressibility of a histogram.

Another method to generate differentially private histograms is the smoothed histogram approach. Wasserman and Zhou (2010) provided the formulation of smoothed histograms of \( \epsilon \)-DP for \( x \in [0,1]^p \), where \( p \) is the number of numerical attributes. It is easy to extend the formulation to the general case when \( x \) is bounded by \( [c_{10}, c_{11}] \times ... \times [c_{p0}, c_{p1}] \). The differentially private smooth histogram is

\[
\hat{f}_K^*(x) = (1 - \lambda) \hat{f}_K(x) + \lambda \Omega, \quad \text{where} \quad \Omega = \left( \prod_{j=1}^{p}(c_{j1} - c_{j0}) \right)^{-1},
\]

and

\[
\lambda \geq \frac{K}{K + n (e^{\epsilon/n} - 1)},
\]

is a constant between 0 and 1 to satisfy \( \epsilon \)-DP. When \( \epsilon \to 0, \lambda \to 1 \), the synthetic data are simulated from a uniform-like \( \hat{f}_K^*(x) \) that is too noisy to be of any use. When \( \epsilon \to \infty, \lambda \to 0, \), the synthetic data would have minimal privacy protection from the DP perspective. Since \( \lambda \) is a constant given \( n, K \) and \( \epsilon \), \( \hat{f}_K(x) \) is not subject to randomness either, it is not necessary to release multiple sets of \( \tilde{x} \) from \( \hat{f}_K(x) \) from an inferential perspective.

In addition to the perturbed histogram and smooth histogram approaches, there is also the approach to generating data from differentially private empirical cumulative density functions (CDF) via the Exponential mechanism (Wasserman and Zhou, 2010).
Specifically, surrogate data \( \tilde{x} \) is simulated from

\[
h(\tilde{x}) = \frac{g_{\tilde{x}}(\tilde{x})}{\int_{c_10, \ldots, c_{p_1}} g_{\tilde{x}}(z) \, dz},
\]

where \( g_{\tilde{x}}(\tilde{x}) = \exp \left( -u(\hat{F}_x, \hat{F}_{\tilde{x}}) \frac{\epsilon}{2\Delta_u} \right) \), \( \Delta_u = \sup_{x, x', \Delta(x, x') = 1} \sup_{\tilde{x}} \left| u(\hat{F}_x, \hat{F}_{\tilde{x}}) - u(\hat{F}_{x'}, \hat{F}_{\tilde{x}}) \right| \),

\( \hat{F}_x \) is the original empirical CDF, \( \hat{F}_{\tilde{x}} \) is the empirical CDF's of the sanitized data, \( u \) is the utility function that denotes a distance measure between the two CDFs, and \( \Delta_u \) is the sensitivity of \( u \). If the Kolmogorov-Smirnov distance is used on \( u \), \( \Delta_u \leq n^{-1} \) (Wasserman and Zhou, 2010). However, releasing \( \tilde{x} \) via the Exponential mechanism defined in Eqn. (8) does not seem to be a viable choice in practice. One difficulty lies in defining the set of all possible candidate CDFs, the size of which increases rapidly with sample size \( n \) and \( p \), making the synthesis process computationally challenging and unrealistic for a large data set. Due to the impracticality of this approach, we did not implement this method in our simulation studies.

3.2 Parametric DIPS (P-DIPS)

The synthesizers in the P-DIPS category are based on an assumed distribution or an appropriately defined model given \( x \). In what follows, we describe the Multinomial-Dirichlet (MD) synthesizer and other methods motivated by the MD synthesizer for categorical data, and the model-based DIPS (MODIPS) approach for general data types based on a Bayesian modeling framework. The summary of the P-DIPS covered in section is given in Table 2. The PrivBayes method based on Bayesian networks and the DPCopula method will not be covered in full details in this section (given that they are not as widely used for routine data analysis).

3.2.1 Multinomial-Dirichlet synthesizer

Abowd and Vilhuber (2008) proposed the Multinomial-Dirichlet (MD) synthesizer to generate differentially private categorical data in the Bayesian framework. The likelihood of proportions \( \pi \) is constructed from \( f(n|\pi) \sim \text{Multinom}(n, \pi) \), where \( n = (n_1, \ldots, n_K) \) contains the original cell counts in \( K \) categories in the original data and \( n = \sum_k n_k \). A Dirichlet prior \( f(\pi) = D(\alpha) \) is imposed on \( \pi \), where each element of \( \alpha \) is set at \( \alpha_k^* = n/(e^\epsilon - 1) \), the minimum value that guarantees \( \epsilon \)-DP, for \( k = 1, \ldots, K \). To generate differentially private surrogate data sets, \( \pi^* \) is first simulated from the posterior distribution \( f(\pi^*|x) = D(\alpha^* + n) \), and then synthetic data is drawn from \( f(\tilde{n}|\pi^*) = \text{Multinom}(n, \pi^*) \). To ensure valid inferences in the synthetic data, multiple sets of \( \tilde{n} \) can be released; one for each differentially private \( \pi^* \). The MD synthesizer reduces to the Binomial-Beta (BB) synthesizer in the binary case. McClure and Reiter (2012) proposed a slightly different approach to synthesizing binary data from \( f(\tilde{n}|n) = \text{Binom} \left( n, \frac{n_1 + n_1}{n + \alpha_1 + \alpha_2} \right) \), where \( \alpha_1 = \alpha_2 = (e^\epsilon - 1)^{-1} \) to satisfy \( \epsilon \)-DP, which
Table 2: Summary of Parametric DIPS approaches discussed in Section 3.2.

| Method                       | Data Type | Pros/Cons                                                                 |
|------------------------------|-----------|---------------------------------------------------------------------------|
| Multinomial-Dirichlet (MD),  | categorical | doesn’t require model specification / performs poorly on sparse data; perturbation amount increases with \( n \); possible biased inferences on proportions |
| Binomial-Beta (BB)           |           |                                                                           |
| BB-MR                        | binary    | simple / perturbation amount increases with \( n \); possible biased inferences on proportions; does not prorogate properly synthesis uncertainty |
| Model-based DIPS (MODIPS)    | any       | general / model dependent and also relies on identification and sanitization of sufficient statistics or likelihood functions |
| PrivBayes                    | binary    | effective in representing correlated high-dimensional data / requires dichotomization on continuous attributes and depends on a quality function that can be computationally inefficient |
| DPCopula                     | any       | general / general limitations for copula models and quadratic time complexity |

we refer to as the BB-MR approach. Different from the BB synthesizer, \( \tilde{\mathbf{n}} \) synthesized via BB-MR has one less layer variability without simulating \( \pi \) from its posterior distribution.

In both the MD/BB and the BB-MR synthesizers, \( \alpha_k^* \) increases with \( n \), implying that when data/observed information increases, the amount of perturbation required to maintain \( \epsilon\)-DP also increases and can be nontrivial for any \( n \). Furthermore, since all \( \alpha_k^* \)'s for \( k = 1, \ldots, K \) are equal, when \( n_k \)'s are not the same across the \( K \) categories, the perturbation will bias the synthetic proportions away from their originals. Charest (2010) modeled explicitly the BB mechanism in a Bayesian framework in the binary data case to reduce the bias of the inferences in the synthetic binary data, which seems to be effective as long as \( \epsilon \) is not too small.

3.2.2 Model-based DIPS (MODIPS)

The MODIPS approach is based in a Bayesian modeling framework and releases \( m \) sets of multiple differentially private surrogate data to the original data to account for the uncertainty of the synthesis model (Liu, 2016b). An illustration of the MODIPS algorithm is given in Figure 1. The MODIPS approach first constructs an appropriate Bayesian model from the original data and identifies the Bayesian sufficient statistics \( s \) associated with the model. The posterior distribution of \( \theta \) can then be represented as \( f(\theta|s) \). The MODIPS then sanitizes \( s \) with privacy budget \( \epsilon/m \). Denote the sanitized \( s \) by \( s^* \). Synthetic data \( \tilde{\mathbf{x}} \) is simulated given \( s^* \) by first drawing \( \theta^* \) from the posterior.
distribution \( f(\theta|s^*) \), and then simulating \( \tilde{x}^* \) from \( f(x|\theta^*) \). The procedure is repeated \( m \) times to generate \( m \) surrogate data sets. The MODIPS does not necessarily need to be based in the Bayesian framework and can draw \( \tilde{x}^* \) directly from the conditional distribution \( f(x|s^*) \). The challenge of this approach lies in obtaining and sampling from \( f(x|s^*) \). Though the MODIPS framework depicted in Figure 1 has more steps, the data augmentation step of drawing \( \theta^* \) first from the posterior distribution \( f(\theta|s^*) \) and then simulating \( \tilde{x}^* \) from \( f(x|\theta^*) \) can be computationally more effective and feasible than calculating and sampling directly from \( f(x|s^*) \).

### 3.3 Inferences from synthetic data via DIPS

Synthetic data generated by DIPS approaches are perturbed through the sanitization process with random noise into the original data. Some P-DIPS approaches (such as the MD synthesizer and MODIPS) also incorporate the uncertainty around the distribution and model assumed on the original data. There are at least two approaches that account for the sanitization/synthesis uncertainty in the inferences based on the synthetic data. The first approach is to model the sanitization/synthesis process directly, and the second approach is to release multiple sets of synthetic data. The second approach can be regarded as a Monte Carlo version of the former. In both Charest (2010), where binary data is synthesized, and Karwa et al. (2016), where the edges of a social network via the exponential random graph models are synthesized, Bayesian modeling and MCMC computations are applied to directly model the sanitization process, which makes the analysis challenging both analytically and computationally on the data users’ end. In the multiple release approach, data users only need to analyze each surrogate data set as if they had the original data set, and then combine the multiple sets of inferences in a legitimate way to yield the final inferences. Suppose the parameter of interest is \( \beta \). Denote the estimate of \( \beta \) in the \( j \)th synthetic data by \( \hat{\beta}_j \) and the associated standard error by \( v_j \). The final point estimate \( \bar{\beta} \) is

\[
\bar{\beta} = m^{-1} \sum_{j=1}^{m} \hat{\beta}_j \tag{9}
\]

with \( \text{Var}(\bar{\beta}) \) estimated by

\[
T = m^{-1} B + W \tag{10}
\]
where $B = \sum_{i=1}^{m} (\hat{\beta}_j - \bar{\beta})^2/(m-1)$ (between-set variability) and $W = m^{-1} \sum_{j=1}^{m} v_j^2$ (average per-set variability); and tests and confidence intervals are based on

$$(\bar{\beta} - \beta)T^{-1/2} \sim t_{\nu=(m-1)(1+mW/B)^2}. \tag{11}$$

The variance combination rule given in Eqn (10) was proposed by Reiter (2003) for dealing with inferences in the context of partial synthesis without DP. Liu (2016b) proved that the combination rule still applies in the case of the MODIPS approach and differs only in what the between-set variability $B$ comprises: $B$ in MODIPS has one additional source of variability from sanitizing $s$ compared to the MS approach without DP. Due to this, inferences from synthesized data via the MODIPS approach will be less precise than those from non-DP MS approaches without DP – a price paid for DP guarantee. Though not formally proved, it is expected Eqns (9) to (11) also apply in the MD synthesizer and other DIPS approaches that use multiple set releases to account for sanitization and synthesis uncertainty, though the sources that compose $B$ might differ.

4 Simulation Studies

We assess the utility and inferential properties of the sanitized data via some of the DIPS approaches presented in Section 3 in four simulation studies. We examine the approaches in the setting of the pure $\epsilon$-DP through the application of the Laplace mechanism, but all the approaches can be executed to satisfy softer versions of DP (such as $(\epsilon, \delta)$-pDP) via the employment of appropriate sanitizers (such as the Gaussian mechanism). The first and second simulation studies focus on categorical data and continuous data generated from the Binomial and Normal distributions, respectively; the third and fourth simulation studies involve a mixture of categorical and continuous variables generated from Gaussian mixture models and a series of logistic regression models, respectively. The results are benchmarked against the statistical inferences based on the original data and the traditional MS technique without DP. In all the synthesis approaches examined, the sample size of each released synthetic set is the same as the original data.

4.1 Simulation study 1: categorical data

The following DIPS methods are compared in this simulation study: the MODIPS synthesizer, the Laplace sanitizer, the BB-MR synthesizer, and the MD synthesizer. Data was simulated from a Bernoulli distribution $f(x_i) = \text{Bern}(\pi)$ for $i = 1, \ldots, n$. We examined 9 simulation scenarios for $n = \{40, 100, 1000\}$ and $\pi = \{0.10, 0.25, 0.50\}$, with 5000 repetitions per scenario. In each DIPS approach, we varied the privacy budget $\epsilon$ from $e^{-10}$ to $e^8$ to examine its effect on the statistical inferences. 5 sets of synthetic data were generated in each of the DIPS approaches, where each synthesis received $\epsilon/5$ privacy budget per the sequential composition.

In the non-DP MS approach, the posterior distribution of $\pi$ given $x$ was $f(\pi|x) =$
\begin{align*}
\text{Beta}(\alpha + n_1, \beta + n - n_1) \text{ with Beta}(\alpha, \beta) \text{ as the prior on } \pi, \text{ where } n_1 = \#\{x_i = 1\},
\text{ and the posterior predictive distribution was } f(\tilde{x}_i|x) = \int f(x_i|\pi)f(\pi|x)d\pi. \text{ We set } \\
\alpha = \beta = 1/3 \text{ (Kerman, 2011). We sampled } \pi \text{ from } f(\pi|x) = \text{Beta}(\alpha + n_1, \beta + n - n_1), \text{ and then } \tilde{x}_i \text{ from } f(\tilde{x}_i|\pi) = \text{Bern}(\pi) \text{ for } i = 1, \ldots, n. \end{align*}

The process was repeated 5 times to obtain 5 sets of synthetic binary data. In the MODIPS approach, we first located the Bayesian sufficient statistics, \(s\), associated with the posterior distribution
\begin{align*}
    f(\pi|x) = \text{Beta}(\alpha + n_1, \beta + n - n_1), \text{ which was } n_1 \text{ with GS } \Delta = 1. \text{ The Laplace mechanism was then employed to obtain } n^*_1 = n_1 + e, \text{ where } e \sim \text{Lap}(0, \epsilon^{-1}). \text{ Finally, we sampled } \\
    \pi^* \text{ from } f(\pi^*|n^*_1) = \text{Beta}(\alpha + n^*_1, \beta + n - n^*_1), \text{ and } \tilde{x}_i \text{ from } f(\tilde{x}_i|\pi^*) = \text{Bern}(\pi^*) \text{ for } i = 1, \ldots, n \text{ to generate one set of synthetic data. The cycle was repeated 5 times (from sanitizing } n_1 \text{ to drawing } \tilde{x}) \text{ to obtain 5 sets of synthetic binary data. Both the BB-MR synthesizer and the MD synthesizer simulated data as in } \tilde{x} \sim \text{Bern}(p^*); \text{ however, } p^* \text{ was fixed at } \frac{n_1 + \alpha^*}{n_1 + \alpha^* + \beta^*}, \text{ with } \alpha^* = \beta^* = (e^{\epsilon/n} - 1)^{-1} \text{ for the BB-MR synthesizer, and was drawn from } \\
    f(p^*|\alpha^*, \beta^*) = \text{Beta}(\alpha^* + n_1, \beta^* + n - n_1) \text{ for the MD synthesizer with } \alpha^* = \beta^* = n/(e^{\epsilon} - 1). \text{ In the Laplace sanitizer, } n^*_1 = n_1 + e, \text{ where } e \sim \text{Lap}(0, \epsilon^{-1}), \text{ and a single data set with } n^*_1 = \#\{\tilde{x}_i = 1\} \text{ was released. Five synthetic data sets were } \\
    \text{generated for the MD and Laplace synthesizers. In the BB-MR synthesizer, a single synthetic data set was generated since no synthesis or sanitization randomness was introduced during the process on top of the sampling error from the binomial data set. In the MODIPS and Laplace sanitizer, sanitized } n^* \text{ could be out of bounds of } [0, 1] \text{ since noise } e \in (-\infty, \infty). \text{ To legitimize } n^*_1, \text{ we applied truncation at the boundaries of } 0 \text{ and } n, \text{ and the boundary inflated truncation (BIT) (setting } n^*_1 \text{ values } < 0 \text{ at } 0 \text{ and those } > n \text{ at } n). \text{ Both post-hoc processing procedures were noninformative and did not leak original data; thus } \epsilon\text{-DP was preserved (Liu, 2016c).}

To obtain inferences on } \pi \text{ from the released data, each of the } 5 \text{ sets was analyzed separately. The point estimate of } \pi \text{ in the } j\text{-th } (j = 1, \ldots, 5) \text{ synthetic data set was the sample proportion } \hat{p}_j, \text{ and its variance was estimated as } v_j = \hat{p}_j(1 - \hat{p}_j)n^{-1}. \text{ Eqns } (9) \text{ to } (11) \text{ were then applied to obtain a final estimate of } \hat{p} \text{ and the associated 95\% confidence interval (CI). Figure 2 depicts the results on the bias, root mean squared error (RMSE), the 95\% CI width, and the coverage probability (CP) of the 95\% CI for } \pi \text{ based on the synthetic data from the DIPS approaches, the non-DP MS approach, and the original data (we present only the results from the BIT post-processing, which was better than the truncation approach in preserving the original information under the same specification of DP).}

Overall, the performances of the MODIPS and Laplace synthesizers were similar while those of the MD and BB-MR synthesizer were similar; and, in general, the inferences in the former seemed to be better than the latter two. In all DIPS approaches, there was noticeable bias, large RMSE, and some undercoverage especially when } \epsilon < 1 \text{ and } n \text{ was small. The inferences improved as } \epsilon \text{ increased (more privacy budget and thus less perturbation), and eventually approached the original or the non-DP MS results (the inferential results from the MODIPS and the Laplace sanitizers approached the non-DP MS results while the MD synthesizer and the BB-MR synthesizer approached}
Figure 2: The bias, root mean square error (RMSE), coverage probability (CP), and 95% confidence interval (CI) width of $\pi$ in simulation study 1. MODIPS represents the model-based differentially private synthesis with boundary inflated truncation (BIT), LAP represents the Laplace sanitizer, MD represents the Multinomial-Dirichlet synthesizer, BB-MR represents BB-MR synthesizer, Ori is the original results without any perturbation, and MS is the traditional multiple synthesis method without DP.
the original results). In the MODIPS and Laplace sanitizer, the noise level remained constant regardless of \( n \). As a result, there was relatively less perturbation as \( n \) increased. In the MD and the BB-MR synthesizers, the perturbation increased monotonically with \( n \). In addition, if the sample proportion was not 0.5, the perturbation would introduce a bias into the released data due to the consistency between the prior information and the data. For the Laplace sanitizer and the MODIPS approach, the inferences were also the best when \( \pi = 0.5 \) since 0.5 was the center point of the support of a proportion, truncating at 0 or 1 did not skew the distribution of \( \pi \) as much as when \( \pi \) was close to 0 or 1. Though The RMSE values from the MODIPS and Laplace sanitizers were slightly larger for very small \( \epsilon \) than those from the BB-MR and MD synthesizers, the former caught up quickly with the latter (around \( \epsilon = e^{-5} \sim e^{-2} \)) when \( \pi \neq 0.5 \); when \( \pi = 0.5 \), the BB-MR and MD synthesizers offered smaller RMSE values for \( \epsilon < 1 \). The CI widths for the Laplace and the MODIPS sanitizers were larger than those from the BB-MR and MD synthesizers until around \( \epsilon = e^{-1} \sim 1 \). The narrow CIs in the latter two led to severe undercoverage when \( \pi \neq 0.5 \) until \( \epsilon \) became very large, while the MODIPS and Laplace sanitizers produced the close to the nominal level coverage (0.95) across all the \( n \), \( \pi \), and \( \epsilon \) values; except for some undercoverage at small \( \epsilon \) and \( n \) due the relatively large bias with the truncation at 0 and 1 for sanitized proportions. Eventually all CPs converged to the nominal level as \( \epsilon \) increased in all the sanitizers except for the BB-MR synthesizer.

4.2 Simulation study 2: continuous data

The following methods are compared in this simulation study: the MODIPS synthesizer, the NP-DIPS synthesizers via the perturbed histogram and the smoothed histogram approaches. Data was simulated from a Normal distribution \( f(x_i) = N(\mu, \sigma^2) \) for \( i = 1, \ldots, n \). We manually truncated the simulated data at \([c_0 = \mu - 3\sigma, c_1 = \mu + 4\sigma] \) to create bounded data. Since there was minimal probability mass (0.0013) outside the \([\mu - 3\sigma, \mu + 4\sigma] \), the normal assumption was hardly affected with the truncation. We examined 9 simulation scenarios for \( n = \{20, 100, 1000\} \) and \( \sigma^2 = \{1, 4, 9\} \), with 5000 repetitions per scenario. Without loss of generality, \( \mu \) was set to 0 in all scenarios without loss of generality. Furthermore, we varied the privacy budget \( \epsilon \) from \( e^{-8} \) to \( e^8 \) to examine its effect on the inferences.

For the non-DP MS method, we assumed prior \( f(\mu, \sigma^2) \propto \sigma^{-2} \). The posterior distributions were \( f(\sigma^2|x) = \text{Inv-Gamma}[\frac{(n-1)/2}{}, (n-1)/2] \) and \( f(\mu|x, \sigma^2) = N(\bar{x}, n^{-1}\sigma^2) \), where \( \bar{x} \) and \( S^2 \) were the sample mean and variance, respectively. The posterior predictive distribution was \( f(\tilde{x}_i|x) = \int f(\tilde{x}_i|\mu, \sigma^2) f(\mu|\sigma^2, x) f(\sigma^2|x) d\mu d\sigma^2 \). Surrogate data were generated by first drawing \( \sigma^2 \) and \( \mu \) from their posterior distributions, and then drawing \( \tilde{x} \) from the normal distribution given the drawn \( \mu \) and \( \sigma^2 \). The process repeated 5 times to generate 5 sets of synthetic data to release.
The Bayesian sufficient statistics from the posterior distributions of $(\mu, \sigma^2)$ was $s = (\bar{x}, S^2)$. The MODIPS procedure started with sanitizing $s$ via the Laplace mechanism to obtain $s^*$. Specifically, the GS was $(c_1 - c_0)n^{-1}$ for $\bar{x}$ and $(c_1 - c_0)^2n^{-1}$ for $S^2$, where $(c_1 - c_0) = 7\sigma$ (Liu, 2016b). Since the range of $\bar{x}$ was $[c_0, c_1]$, and that of $S^2$ was $[0, (c_1 - c_0)^2/4 \cdot n/(n-1)]$ (Macleod and Henderson, 1984), if an element in sanitized $s^*$ was outside its range, it was processed by the BIT procedure. Given the sanitized $s^* = \{\bar{x}^*, S^{2*}\}$, the MODIPS technique drew $\sigma^{2*}$ from Inv-Gamma $[(n-1)/2, (n-1)S^{2*}/2]$ and $\mu^*$ from $N(\bar{x}, n^{-1}\sigma^{2*})$. Finally, $\tilde{x}_i^*$ was simulated from $N(\mu^*, \sigma^{2*})$ for $i = 1, \ldots, n$ to generate one set of surrogate data. The whole procedure was repeated 5 times to generate 5 surrogate data sets. $\epsilon/5$ of the total budget was spent per synthesis. In addition, since there were two statistics, $(\bar{x}, S^2)$, to sanitize over the same set of data, the $\epsilon/5$ budget per synthesis needed to cover the sanitization of both. We applied both the conjoint sanitization and individual sanitization (Liu, 2016b). In the conjoint sanitization, Laplace noise was drawn from $\text{Lap}(\Delta_x(\epsilon/5)^{-1})$, where $\Delta_x = \Delta_x + \Delta_{S^2} = (c_1 - c_0 + (c_1 - c_0)^2)/n$ was the GS of $s = (\bar{x}, S^2)$; in the individual sanitization, $\epsilon/5$ was further split between sanitizing $\bar{x}$ and sanitizing $S^2$, in an allocation ratio of $w_1 : w_2$; that is, noise added onto $\bar{x} \sim \text{Lap}(\{(c_1 - c_0)/n, \Delta_x(w_1\epsilon/5)^{-1}\}$, and those added onto $S^2 \sim \{(c_1 - c_0)^2/n, \Delta_{S^2}(w_2\epsilon/5)^{-1}\}$. In this simulation, we set $w_1 = w_2 = 0.5$. We present only the results from the individual sanitization, which led to better utility of the sanitized data than the conjoint sanitization in terms of the inferences of $\mu$ and $\sigma^2$.

In deciding the number of bins for the histograms in the perturbed and smoothed histogram approaches, we applied the Scott’s Rule after comparing Sturges’ Rule, the Scott’s Rule, and the Freedman-Diaconis rule (Scott, 2015). Specifically, the bin width was $\hat{h} = 3.5Sn^{-1/3}$, where $S$ was the sample standard deviation of $x$ and $n$ was the sample size. The median number of bins was 7, 10, and 21 for $n = 20, 100$, and 1000, respectively, across all simulations (Table 5 in Supplemental Materials). Each bin count was perturbed via the Laplace mechanism with $\Delta_x = 1$ to obtain the perturbed density histogram from Eqn (5). The procedure was repeated 5 times to obtained 5 sets of differentially private $\tilde{p}_j^*$, based on the 5 sets of synthetic data that were simulated. For the smoothed histogram, we first calculated $\lambda$ for a given $\epsilon$ using Eqn (7), and then constructed the smoothed histogram by applying Eqn (6), from which a single set of synthetic data was generated and released.

When obtaining inference on $\mu$ and $\sigma^2$ from the multiple released data sets via the MODIPS, the perturbed histogram sanitizers, and the non-DP MS approach, each of the 5 synthetic sets was analyzed separately to obtain point estimates of $\mu$ and $\sigma^2$, which were $\bar{x}_j$ (the sample mean in the $j^{th}$ synthetic set) and $s_j^2$ (the sample variance in the $j^{th}$ set); the associated within-set variance estimates were $s_j^2/n$ and $(s_j^2)^2(2(n-1)^{-1}+\kappa_jn^{-1})$, respectively, where $\kappa_j$ was the excess kurtosis in the $j^{th}$ set. Eqns (9) to (11) were then applied to obtain the final estimates and 95% CIs.

Figures 3 and 4 depict the bias, RMSE, CP, and CI width of $\mu$ and $\sigma^2$ based
Figure 3: The bias, root mean square error (RMSE), coverage probability (CP), and 95% confidence interval (CI) width of $\mu$ in simulation study 2. MODIPS represents the model-based differentially private synthesis, PERT represents the perturbed histogram method, SMOOTH represents the smoothed histogram method, MS is the traditional multiple synthesis method, and Ori is the original results without any perturbation.
Figure 4: The bias, root mean square error (RMSE), coverage probability (CP), and 95% confidence interval (CI) width of $\sigma^2$ in simulation study 2. MODIPS represents the model-based differentially private synthesis with $\epsilon$ split between $\mu$ and $\sigma^2$, PERT represents the perturbed histogram method, SMOOTH represents the smoothed histogram method, MS is the traditional multiple synthesis method without DP, and Ori is the original results without any perturbation.
on the synthetic data via the 3 DIPS approaches, respectively. For the purposes of comparability across different values of $\sigma^2$, the bias, RMSE, and CI width for $\sigma^2$ were scaled by the true $\sigma^2$, referred to as the relative bias, scaled RMSE and scaled CI width, respectively. Overall, the MODIPS approach outperformed the two histogram-based approaches in the case of $\mu$ with smaller bias and close-to-nominal CP, but it did have larger RMSEs and wider CIs. The MODIPS approach also outperformed the two histogram-based approaches in the case of $\sigma^2$ with much smaller bias and RMSEs and close-to-nominal CP; the CIs were wider in the former. Between the perturbed and smooth histograms, the former offered better inferences.

Specifically, there were some noticeable positive biases and large RMSEs at small $\epsilon$ for both $\mu$ and $\sigma^2$ in all approaches, which improved as $\epsilon$ increased and, eventually, approached the original or the non-DP MS results. For the MODIPS and the perturbed histogram approaches, the amount of injected noise became immaterial as $n$ increased; therefore, inferences improved with $n$. In the the smoothed histogram, $\lambda$ in Eqn. (7) got larger and approached $K/(K + \epsilon)$ as $n$ increased. As a result, increasing $n$ did not seem to help with improving inferences in the smoothed histogram. The observed positive bias in $\sigma^2$ was expected due to the randomness introduced through synthesis and sanitization. The positive bias in $\mu$ can be explained by the asymmetric bounds of data $x [\mu - 3\sigma, \mu + 4\sigma]$ around $\mu$. When the data sanitized $\bar{x}$ and synthesized $\tilde{x}$ were out of bound, they were set at the boundary values. Since the left bound $\mu - 3\sigma$ was closer to $\mu$, there were more values at $\mu - 3\sigma$ than at $\mu + 4\sigma$, resulting in overestimation.

In this simulation, we set asymmetric $[\mu - 3\sigma, \mu + 4\sigma]$ bounds on data, which reflected most of the real life situations, where bounds on data are often dictated with other-than-statistical reasons, and the probability that the chosen bounds would be symmetric about the unknown true parameters is almost 0. As a supplementary analysis, we examined the case when the data bounds were symmetric around the true mean: $[\mu - 4\sigma, \mu + 4\sigma]$. The results are presented in Figures 9 and 10 in the Supplemental Materials. As expected, there were minimal biases on $\mu$ in all the DIPS approaches (there was some fluctuation in MODIPS for small $\epsilon$), and the CP in all approaches were at nominal-level. The histogram-based approaches delivered more precise estimates than MODIPS in the inferences of $\mu$ (smaller RMSE and narrower CIs). However, the histogram-based approaches did not perform as well as MODIPS in the inferences of $\sigma^2$. Both the bias and RMSE were large and there was severe undercoverage at small values of $\epsilon$. 


4.3 Simulation study 3: Gaussian mixture model

In this simulation study, we compared the MODIPS synthesizer and the NP-DIPS synthesizer in data sets with mixed continuous and categorical variables. For NP-DIPS, we applied the Laplace synthesizer on categorical variables + the perturbed histogram for continuous variables. We did not implement the MR synthesis or the smoothed histogram approach given their inferior performances to the the Laplace sanitizer, the perturbed histogram, and the MODIPS in the previous two simulation studies. Data $\mathbf{x}$ was simulated from a Gaussian mixture model, comprising categorical variables $\mathbf{w}$ and continuous variables $\mathbf{z}$, where $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$ with 2, 3, and 4 levels, respectively. Let $n_k$ denote the count in cell $k$ in the full cross-tabulation of $\mathbf{w}$ for $k = 1, \ldots, 24$. First, the counts $\mathbf{n} = \{n_k\}$ in the 24 cells were simulated from a Multinomial distribution with parameter $\mathbf{\pi} = \{\pi_k\}$; and then $\mathbf{z}_{ik} = (z_{ik1}, z_{ik2})'$ was simulated from $N_{(2)}(\mu_k, \Sigma)$ for $i = 1, \ldots, n_k$ and $k = 1, \ldots, 24$, where $\mu_k = (\mu_{k1}, \mu_{k2})'$ was the mean of $\mathbf{z}$ in cell $k$, and $\Sigma$ was the covariance matrix that was set to be the same across all 24 cells. The summary of the parameter values of $\mu_1, \mu_2, \pi$ across the 24 cells are provided in Table 6 in the Supplemental Materials. We set $n = 1000$, the variances of $z_{ik1}$ and $z_{ik2}, \sigma_1^2 = \sigma_2^2 = 1$, and their correlation at $\rho = 0.50$ with 5000 repetitions. $z_{ikj}$ in cell $k$ (where $j = 1, 2$) was truncated at $[c_{0,kj} = \mu_{kj} - 4\sigma_j, c_{1,kj} = \mu_{kj} + 4\sigma_j]$ to generate bounded data. We examined a range of privacy budget $\epsilon$ ranging from $e^{-6}$ to $e^{8}$.

For the non-DP MS approach, we imposed a Dirichlet prior on $\pi$, $f(\pi) = D(\alpha)$, where $\alpha = \{\alpha_1, \ldots, \alpha_K\} = 1/2$, and applied priors $f(\mu, \Sigma) \propto |\Sigma|^{-1}$ to $\mu$ and $\Sigma$. The posterior distributions were $f(\pi | \mathbf{w}) = D(\alpha')$, $f(\Sigma | \mathbf{z}, \mathbf{w}) = \text{Inv-Wishart}(n - K, \mathbf{S})$, and $f(\mu_k | \Sigma, \mathbf{z}, \mathbf{w}) = N_{(2)}(\tilde{\mathbf{z}}_k, n_k^{-1}\Sigma)$, where $\alpha' = \alpha + n$, $\mathbf{S} = n^{-1}\sum_{k=1}^K \sum_{i=1}^{n_k} (\mathbf{z}_{ik} - \bar{\mathbf{z}}_k)(\mathbf{z}_{ik} - \bar{\mathbf{z}}_k)'$, and $\tilde{\mathbf{z}}_k$ contained the sample means of $\mathbf{z}$ in cell $k$. Synthetic data were simulated from the posterior predictive distribution $f(\tilde{\mathbf{z}}_i, \tilde{\mathbf{w}}_i | \mathbf{z}, \mathbf{w}) = \int f(\tilde{\mathbf{z}}_i | \mu, \Sigma, \tilde{\mathbf{w}}_i)f(\tilde{\mathbf{w}}_i | \pi)f(\mu | \Sigma, \mathbf{z}, \mathbf{w})f(\Sigma | \mathbf{z}, \mathbf{w})f(\pi | \mathbf{w}) d\pi d\mu_k d\Sigma_{ij}$ by drawing $\pi$ from $f(\pi | \mathbf{w}) = D(\alpha + n)$, $\Sigma$ from $f(\Sigma | \mathbf{z}, \mathbf{w}) = \text{Inv-Wishart}(n - K, \Sigma)$, $\mu_k$ from $f(\mu_k | \Sigma, \mathbf{z}, \mathbf{w}) = N(\tilde{\mathbf{z}}_k, n_k^{-1}\Sigma)$, $\tilde{\mathbf{w}}_i$ from $f(\tilde{\mathbf{w}}_i | \pi) = \text{Multinom}(n, \pi)$, and $\tilde{\mathbf{z}}_i$ from $f(\mathbf{z}_i | \tilde{\mathbf{w}}_i, \Sigma) = N_{(2)}(\tilde{\mu}_k, \tilde{\Sigma})$ for $i = 1, \ldots, \tilde{n}_k$, where $\tilde{n}_k$ was the count in cell $k$ based on the synthesized $\tilde{\mathbf{w}}$, and $\tilde{k}$ indicates the cell to which the simulated case $i$ belonged given the synthesized $\tilde{\mathbf{w}}_i$. The drawing process repeated 5 times to generate 5 sets of surrogate data.

The Bayesian sufficient statistics from the above Bayesian model was $\mathbf{s} = (\mathbf{n}, \mathbf{S}, \tilde{\mathbf{z}})$, where $\tilde{\mathbf{z}}$ contained the 24 pairs of cell means of $\mathbf{z}_1$ and $\mathbf{z}_2$, among which there was no overlapping information. The MODIPS procedure started with sanitizing $\mathbf{s}$ via the Laplace mechanism to obtain $\mathbf{s}^* = (\mathbf{n}^*, \mathbf{S}^*, \tilde{\mathbf{z}}^*)$ (the $l_1$-GS was 1 for $\mathbf{n}$, $(c_{1,kj} - c_{0,kj})n_k^{-1}$ for $\bar{z}_{kj}$, and $(c_{1,kj} - c_{0,kj})^2(n - 1)(n - K)^{-1}$ for each entry in $\mathbf{S}$ (Liu, 2016c), where $c_{1,kj} - c_{0,kj} = 8\sigma$ for $k = 1, \ldots, 24$ and $j = 1, 2$). Given $\mathbf{s}^*$, the MODIPS method first drew $\pi^*$ from $f(\pi^* | \mathbf{n}^*) = D(\alpha + n^*)$, $\tilde{\mathbf{w}}^*$ from $f(\tilde{\mathbf{w}}^* | \pi^*) = \text{Multinom}(n, \pi^*)$, $\Sigma^*$ from $f(\Sigma^* | \mathbf{S}^*) = \text{Inv-Wishart}(n - K, \mathbf{S}^*)$, $\mu_k^*$ from $f(\mu_k^* | \Sigma^*, \tilde{\mathbf{z}}^*, \mathbf{w}) = N(\tilde{\mathbf{z}}_k^*, n_k^{-1}\Sigma^*)$; and then $\tilde{\mathbf{z}}_i$ was simulated from $f(\mathbf{z}_i | \mu_k^*, \Sigma^*) = N(\mu_k^*, \Sigma^*)$ for $i = 1, \ldots, \tilde{n}_k^*$ to generate one set of surrogate data, where $\tilde{n}_k^*$ was the count in cell $k$ based on the synthesized $\tilde{\mathbf{w}}^*$, and
\( \tilde{k} \) indicates the cell which the simulated case \( i \) belonged to given the synthesized \( \tilde{w}^*_i \). The procedure was repeated 5 times to generate 5 surrogate data sets. Each surrogate data set received 1/5 of the total privacy budget. Since \( s \) contained 6 components: \( \mathbf{n}, \mathbf{z}_1, \mathbf{z}_2 \), two variance terms and one covariance term from \( S \), each received \( \epsilon/30 \) privacy budget (1/6 of \( \epsilon/5 \) budget allocated to each synthesis). Note that though there were 24 elements in \( \mathbf{n} \), these elements were based on disjoints subsets of \( \mathbf{x} \), thus sanitization of each received \( \epsilon/30 \); similarly in the case of \( \mathbf{z}_1 \) and \( \mathbf{z}_2 \).

In the NP-DIPS approach, we applied the Laplace sanitizer to sanitize 24 cell counts \( \mathbf{n} \) formed by the full cross-tabulation of \( \mathbf{w} \), and we employed the perturbed histogram method to sanitize continuous \( \mathbf{z} \) within each of the 24 cells. Since \( \mathbf{z} \) was 2-dimensional, each bin of the histogram of \( \mathbf{z} \) was a square rather than an interval. The number of bins were determined using the Scott’s rule, and the medians ranged from 16 to 49 across the 5000 repeats in the 24 cells (Supplemental Materials Table 7). The process was repeated 5 times to create 5 sets of sanitized \( \tilde{\mathbf{n}} \) and 24 perturbed histograms, from which 5 sets of synthetic data were generated. Each synthesis was allocated 1/5 of the total privacy budget, which was further split between sanitizing the 24 cells formed by \( \mathbf{w} \) and sanitizing the histogram formed by \( \mathbf{z} \) in a 1:1 ratio.

We examined the inferences on \( \mathbf{\mu}_1, \mathbf{\mu}_2, \sigma_1^2, \sigma_2^2, \rho \), and the marginal probabilities of \( \mathbf{w} \) based on the synthetic data sets. In each synthetic set \( l (l = 1, \ldots, 5) \), the marginal probabilities \( \mathbf{\Pi} = \{ \text{Pr}(w_1 = 1), \text{Pr}(w_2 = 1), \text{Pr}(w_2 = 2), \text{Pr}(w_3 = 1), \text{Pr}(w_3 = 2), \text{Pr}(w_3 = 3) \} \) were estimated by the sample marginal probabilities \( \hat{\mathbf{P}}_l; \mathbf{\mu}_1 \) and \( \mathbf{\mu}_2 \) were estimated by the sample cell means \( \bar{z}_{1,l} \) and \( \bar{z}_{2,l} \); and \( \Sigma \) was estimated by the pooled variance-covariance \( \mathbf{S}_l \). The within-set variance was estimated by \( \hat{\mathbf{P}}_l(1 - \hat{\mathbf{P}}_l)n^{-1} \) for \( \mathbf{P}_l \), \( S_{k,l}^2 = \mathbf{n}^{-1} \) for \( \mathbf{z}_k (k = 1, 2), (S_{k,l}^2)^2(2(n - 1)^{-1} + \kappa_l n^{-1}) \) for \( S_{k,l}^2 \), and \( (1 - r_l^2)(n - 2)^{-1} \) for the correlation between \( Z_1 \) and \( Z_2 \), respectively, where \( S_{1,l}^2 \) and \( S_{2,l}^2 \) were the diagonal elements of \( \mathbf{S}_l \). \( \kappa_l \) was the excess kurtosis, and \( r_l \) was derived from \( \mathbf{S}_l \). Eqns (9) to (11) were then applied to obtain the final estimates of the parameters and the 95% CIs.

Figure 5 shows the results on the bias, RMSE, CP, and the 95% CI width of \( \mathbf{\mu}_1 \) and \( \mathbf{\Pi} \). The results on \( \mathbf{\mu}_2, \sigma_1^2, \sigma_2^2, \rho \) are provided in Figures 11 and 12 in the Supplemental Materials. The inferences on \( \mathbf{\mu}_k \) were aggregated via the box plot over the 24 cell means. There was no clear winner in the inferences of \( \mathbf{\Pi} \). The inferences based on the MODIPS were slightly more precise than those based on the NP-DIPS approach. The inferences in the NP-DIPS converged to the original values as \( \epsilon \) increased while those in the MODIPS converged to the non-DP MS values. The inferences of \( \mathbf{\mu}_1 \) and \( \mathbf{\mu}_2 \) based on the synthetic data from the NP-DIPS approach were more precise than those via the MODIPS approach. The MODIPS outperformed the NP-DIPS approach in the inferences on \( \sigma_1^2, \sigma_2^2, \rho \) with much smaller biases and RMSEs for all 3 components when \( \epsilon > e^{-1} \). The NP-DIPS approach experienced severe undercoverage in all 3 components and never reached the nominal level of 95% while the MODIPS approach delivered nominal CP for \( \epsilon > 1 \). The severe undercoverage in the NP-DIPS even when \( \epsilon \) was large is not due to the noise added through the Laplace mechanism, but because
Figure 5: The bias, root mean square error (RMSE), coverage probability (CP), and 95% confidence interval (CI) of $\Pi$ and $\mu_1$ in simulation study 3. In the plot $\Pi$, each line presents a different marginal probability out of 6. MODIPS represents the model-based differentially private data synthesis, NP-DIPS represents the Laplace sanitizer+perturbed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.
of the discretization in forming the histogram bins.

4.4 simulation study 4: logistic models

In this simulation study, we explored a data scenario similar to the simulation in Section 4.3 with a mixture of continuous and categorical variables, but the continuous variables were drawn from a bivariate normal distribution and the categorical variables were generated from a sequence of logistic regression models. We compared the MODIPS synthesizer and the NP-DIPS synthesizer via the Laplace sanitizer + perturbed histogram. Let \( x = (w, z) \), where \( w \) denotes the categorical variables and \( z \) denotes the continuous variables. \( z \) was simulated from the bivariate normal distribution with covariance matrix of \( z \) and \( \epsilon \). We ran 1000 repetitions, and examined the effects of privacy budget \( \epsilon \) ranging from \( e^{-4} \) to \( e^{5} \).

For the traditional non-DP MS approach, we employed priors \( f(\mu, \Sigma, \beta_1, \beta_2, \beta_3, \beta_4) \propto |\Sigma|^{-\frac{1}{2}} \). The posterior distributions of \( \Sigma \) and \( \mu \) were \( f(\Sigma|z) = \text{Inv-Wishart}(n, S) \) and \( f(\mu|z) = N(\bar{z}, n^{-1} \Sigma) \), where \( S = n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})' \) was the sample covariance matrix of \( z \), and \( \bar{z} \) was the sample means of \( z \). With the constant priors on \( (\beta_1, \beta_2, \beta_3, \beta_4) \), their posterior distributions were proportional to the likelihood:

\[
\begin{align*}
\text{a)} f(\beta_1|w_1, z) &\propto \prod_{i=1}^{n} \frac{\text{e}^{w_{1i}(1+z_i-z_{1i})} \beta_1}{1 + \text{e}^{(1+z_i-z_{1i})} \beta_1} = \frac{\text{e}^{\sum_{i=1}^{n} w_{1i}(z_i-z_{1i})} \beta_1}{\prod_{i=1}^{n} (1 + \text{e}^{(1+z_i-z_{1i})} \beta_1)} \\
\text{b)} f(\beta_2|w_2, w_1, z) &\propto \prod_{i=1}^{n} \frac{\text{e}^{w_{2i}(1+z_i-z_{1i})} \beta_2}{1 + \text{e}^{(1+z_i-z_{1i})} \beta_2} = \frac{\text{e}^{\sum_{i=1}^{n} w_{2i}(z_i-z_{1i})} \beta_2}{\prod_{i=1}^{n} (1 + \text{e}^{(1+z_i-z_{1i})} \beta_2)} \\
\text{c)} f(\beta_3, \beta_4|w_3, w_2, w_1, z) &\propto \prod_{i=1}^{n} \left\{ \frac{1}{1 + A_i + B_i} \right\}^{I(w_{3i}=3)} \times \left\{ \frac{\text{e}^{w_{3i}(1+z_i-z_{1i})} \beta_3}{1 + A_i + B_i} \right\}^{I(w_{3i}=2)} \\
&\times \left\{ \frac{\text{e}^{w_{3i}(1+z_i-z_{1i})} \beta_4}{1 + A_i + B_i} \right\}^{I(w_{3i}=3)} \\
&= \frac{\text{e}^{a_i \beta_3 + b_i \beta_4}}{\prod_{i=1}^{n} (1 + A_i + B_i)},
\end{align*}
\]

where \( a_i = (\sum_{i=1}^{n} I(w_{3i}=2), \sum_{i=1}^{n} z_{1i} I(w_{3i}=2), \sum_{i=1}^{n} z_{2i} I(w_{3i}=2), \sum_{i=1}^{n} w_{1i} I(w_{3i}=2), \sum_{i=1}^{n} w_{2i} I(w_{3i}=2), \sum_{i=1}^{n} w_{1i} I(w_{3i}=3), \sum_{i=1}^{n} z_{1i} I(w_{3i}=3), \sum_{i=1}^{n} z_{2i} I(w_{3i}=3), \sum_{i=1}^{n} w_{1i} I(w_{3i}=3), \sum_{i=1}^{n} w_{2i} I(w_{3i}=3)) \), and \( I(\bullet) \) is the indicator function. To
synthesize \( \tilde{z}_i \) for \( i = 1, \ldots, n \) in the traditional non-DP MS approach, we first drew \( \Sigma \) from \( f(\Sigma|z) = \text{Inv-Wishart}(n, S) \), and \( \mu \) from \( f(\mu|\Sigma, z) = \mathcal{N}(\bar{z}, n^{-1}\Sigma) \), and then \( \tilde{z}_i \) from \( f(\tilde{z}_i|\mu, \Sigma) = \mathcal{N}(\mu, \Sigma) \) given the drawn \((\Sigma, \mu)\). To synthesize \( \tilde{w}_i = (\tilde{w}_{i1}, \tilde{w}_{i2}, \tilde{w}_{i3}) \), we applied the Metropolis-Hastings algorithm to sample \( \beta_1, \beta_2, \beta_3, \) and \( \beta_4 \) from their posterior distributions first (2 chains, a burn-in period of 1500, a thinning period of 10, and 6500 iterations to yield a total of \( n \) values), and then simulated \( \tilde{w}_{i1}, \tilde{w}_{i2} \) and \( \tilde{w}_{i3} \) sequentially from \( f(\tilde{w}_{i1}|\beta_1, \Sigma, \tilde{z}) \), \( f(\tilde{w}_{i2}|\beta_2, \tilde{w}_{i1}, \tilde{z}_i) \), and \( f(\tilde{w}_{i3}|\beta_3, \beta_4, \tilde{w}_{i2}, \tilde{w}_{i1}, \tilde{z}_i) \).

For the MODIPS approach, the Bayesian sufficient statistics associated with the posterior distributions of \( \Sigma \) and \( \mu \) were \( S \) and \( \tilde{z} \). The Bayesian sufficient statistics associated with the posterior distribution of \( \beta_1, \beta_2, \beta_3 \) and \( \beta_4 \) in Eqns (12) to (14) were not straightforward since the denominators in these equations involve individual \( z_i \) or \( w_i \), though the statistics needed to be sanitized in the numerators of these equations are easy to identify (e.g., \( \sum_{i=1}^{n} w_{i1}, \sum_{i=1}^{n} z_{1} w_{i1}, \sum_{i=1}^{n} z_{2} w_{i1} \) in Eqn (12)). We tried two alternatives. The first approach was to approximate the denominators as linear functions of \( z \) in Eqns (12) to (14) via the Taylor expansion. However, the approximation error was large at low orders (2 or 3) while higher-order approximations would result in too many sufficient statistics (the high-order polynomial terms) to sanitize. The second approach was to directly sanitize Eqns (12) to (14), where each is a product of \( n \) proportions. We bounded each proportion within \((0, 0.99)\), therefore each equation was bounded within \((0, 0.99^n)\) or \((0, 4.32 \times 10^{-5})\) for \( n = 1000 \) (this simulation). This approach was very easy to implement and also yielded better results, and was adopted in this simulations. All together, the quantities to sanitize included \( S, \tilde{z} \) and the three 3 proportion products in Eqns (12) to (14). Since \( c_{1,j} - c_{0,j} = 8\sigma \) for \( j = 1, 2 \), the \( l_1 \) GS was \( 8\sigma n^{-1} \) for \( \tilde{z}_j \) and \( (8\sigma)^2 n^{-1} \) for each entry in \( S \); the GS for each equation was \( 4.32 \times 10^{-5} \). Once the quantities were sanitized via the Laplace mechanism, the subsequent synthesis steps were similar to the non-DP MS procedure, except that the posterior distributions of the model parameters were based on the sanitized quantities. The whole procedure (from the sanitization of the 5 quantities to the synthesis of \( \tilde{z}^* \) and \( \tilde{w}^* \)) was repeated 5 times to generate 5 surrogate data sets. Each surrogate data set received \( 1/5 \) of the total privacy budget, and was then split equally among the sanitization of the 3 elements \( S \), the 2 elements in \( \tilde{z} \) and the 3 proportion products in Eqns (12) to (14).

In the NP-DIPS approach, no distributions or models were assumed. \( z \) were discretized via the Scott’s Rule and the counts of the 5-way cross-tabulation from \((z_1, z_2, w_1, w_2, w_3)\) were sanitized via the Laplace sanitizer. 5 sets of synthetic data were generated, each with \( 1/5 \) of the total budget \( \epsilon \) (no further split of \( \epsilon/5 \) was needed within each synthesis since there was no overlapping information among the cells of the 5-way cross-tabulation).

We examined the inferences on \( \Sigma, \mu \), and \( \beta_1, \beta_2, \beta_3, \beta_4 \) from the three logistic regression models of \( w \) based on the synthetic data sets. \( \mu_1 \) and \( \mu_2 \) in surrogate data set \( l \) \((l = 1, \ldots, 5)\) were estimated by the sample means \( \bar{z}_{1,l} \) and \( \bar{z}_{2,l} \), and \( \Sigma \) was estimated
by the sample covariance $S_l$. The corresponding within-set variance was estimated by $S_{k,l}^2n^{-1}$ for $\bar{z}_{k,l}$ ($k = 1, 2$), $(S_{k,l}^2)^2(2(n-1)^{-1} + \kappa_l n^{-1})$ for the marginal variances of $z_1$ and $z_2$, and $(1-r_l^2)(n-2)^{-1}$ for the correlation between $Z_1$ and $Z_2$, respectively, where $S_{1,l}^2$, $S_{2,l}^2$ were the diagonal elements of $S_l$, $\kappa_l$ was the excess kurtosis, and $r_l$ was derived from $S_l$. The regression coefficients $\beta$ were estimated using the \texttt{logistf} function with the Firth’s bias reduction method in the R package \texttt{logistf} along with the corresponding estimated variance estimates. Eqns (9) to (11) were then applied to obtain the final estimates of the parameters and the 95% CIs in each DIPS approach.

Due to space limit, we present the results on the bias, log(RMSE, CP), and log(95% CI width) for $\beta_2$ and $\beta_3$ in Figures 13 and 14; the results on $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$, $\rho$, $\beta_1$, and $\beta_4$ are available in the Supplemental Materials. The results on all parameters are summarized as follows. For $\mu_1$ and $\mu_2$, the bias, RMSE, and CI width of the estimates were smaller in the NP-DIPS approach than those in the MODIPS approach at $\epsilon < e$, and were similar for $\epsilon > e$; both provided about 95% CP. For $\Sigma$, the NP-DIPS approach experienced severe undercoverage in all 3 components ($\sigma_1^2$, $\sigma_2^2$, and $\rho$) regardless of $n$ or $\epsilon$ for the same reason as in simulation study 3 (discretization and uniform sampling from each histogram bin). The MODIPS provided nominal CP for $\sigma_1^2$, $\sigma_2^2$, and $\rho$ and had smaller bias and RMSE than the NP-DIPS approach for $\epsilon > e^{-2}$. For the logistic regression coefficients ($\beta_1, \beta_2, \beta_3, \beta_4$), the biases were smaller and converged to the non-DP MS results and zero for the MODIPS whereas the biases from the NP-DIPS approach did not seem converge to zero even when $\epsilon$ was large; however, the bias in the MODIPS was unstable and fluctuate around the zero line when $\epsilon$ was small. The RMSE values from both the NP-DIPS and MODIPS approaches were large especially in the MODIPS. The MODIPS produced coverage at or above the nominal level of 95% for all coefficients with the wide CIs accounting for the large amount of noise being injected at small values of $\epsilon$. The CP results from the NP-DIPS approach varies across parameters: some experienced severe undercoverage across all values of $\epsilon$ or only at small values of $\epsilon$, some had close to 95% coverage across the board. The CI width varied little with $\epsilon$ in the NP-DIPS approach.

Compared to simulation study 3 in Sec 4.3 which also had both continuous and categorical variables and the Gaussian mixture model was applied, the results in simulation study 4 were generally worse for both the NP-DIPS and the MODIPS approaches, but in different ways. In the MODIPS approach, identification of statistics to sanitize was less obvious and the inferences based on the synthetic data were less stable in simulation 4, probably due to the direct sanitization of the likelihood functions. For the NP-DIPS approach, the discretization and sanitization procedure was the same between simulations 3 and 4, but seemed to affect the inferences from logistic regressions more than those from the Gaussian mixture model. The different results from the two simulation studies suggest that even though the NP-DIPS approach was nonparametric, inferences from certain models based on the synthesize data can be more sensitive than others.
Figure 6: The bias, log of the root mean square error (RMSE), 95% coverage probability (CP), and log of the 95% confidence interval (CI) for $\beta_2$ in simulation study 4. MODIPS (with SPLIT and BIT) represents the model-based differentially private synthesis with $\epsilon$ split among the nine sufficient statistics, NP-DIPS represents the Laplace sanitizer + perturbed histogram method, MS is the traditional multiple synthesis method without DP, and Ori is the original results without any perturbation.
Figure 7: The bias, log of the root mean square error (RMSE), 95% coverage probability (CP), and log of the 95% confidence interval (CI) for $\beta_3$ in simulation study 4. MODIPS (with SPLIT and BIT) represents the model-based differentially private synthesis with $\epsilon$ split among the nine sufficient statistics, NP-DIPS represents the Laplace sanitizer + perturbed histogram method, MS is the traditional multiple synthesis method without DP, and Ori is the original results without any perturbation.
5 Discussion

We reviewed the DIPS methods for synthesizing differentially private individual-level data and compared some DIPS methods empirically via extensive simulation studies on the inferential properties of the synthetic data through the methods. To the best of our knowledge, this is the first work that compares the inferential properties of DIPS approaches across various types and sizes of data. From a privacy protection perspective, we recommend the DIPS techniques given the strong guarantee on privacy protection over the traditional non-DP MS approach. As a trade-off, inferences based in the synthetic data from the DIPS approaches are less precise due to the extra layer of noise being injected to ensure DP.

The DIPS methods can be roughly grouped as the nonparametric approaches (NP-DIPS) and the parametric approaches (P-DIPS). NP-DIPS approaches are robust given that they do not need to impose models or distributions on a given data set. However, most NP-DIPS approaches require some degree of discretization on numerical attributes, running the risk of data sparsity in high-order cross-tabulations and sanitization of a large amount of counts. The inferences based on the synthetic data via histogram-based NP-DIPS approaches are also affected by how the histogram bins are formed. P-DIPS approaches, on the other hand, do not have to discretize continuous attributes, but require distributional assumptions and model building. Thus, P-DIPS are subject to model misspecification. In the simulations studies we conducted, the MODIPS approach was the only DIPS approach that provided close-to-nominal coverage regardless of data type, sample size and privacy budget. On the other hand, the MODIPS requires more analytical work and takes longer to produce results if computationally intensive algorithms (e.g., MCMC) are employed to generate data.

Our immediate interest for future work is to improve the precision of the DIPS approaches’ inferences; especially, for the MODIPS methods. The first strategy toward that goal is to take into account the correlations among the statistics during sanitization to ensure the privacy budget is not spent on over overlapping information. In all the simulation studies we conducted, statistics were sanitized independently, implying that redundant noises were introduced on correlated statistics. Accounting for the correlations among the statistics will cut the necessary noises to satisfy DP to the minimum, improving the efficiency of the DIPS procedures. The second strategy is to optimize the privacy budget allocation scheme when the sequential composition is in effect. The third approach is to help reduce the amount of injected noise by applying a relaxed version of DP. Conceptually, all the DIPS methods, regardless of P-DIPS or NP-DIPS, can be implemented in the context of relaxed DP with appropriate sanitizers.

The ultimate goal of developing DIPS approaches is to employ them on real-life data for public release. The most challenging components of the practical application is the large size, high dimension, and complex data structure of real-life data sets with a large number of attributes of various types. Moreover, other issues such as missing data, sparse data, data entry errors, among others add another layer of complexity in
applying DIPS. While we expect our findings to shed light on the data utility aspect and the practical feasibility of applying DIPS approach in data sets of various types and sizes, a regular DIPS approach without any modification might not accommodate to all of the real-life challenges. For example, the OnTheMap project releases commuting patterns of the US population data (Machanavajjhala et al., 2008). The authors proposed combining distance-based coarsening with a probabilistic pruning algorithm and preserving \((4, 10^{-4})\)-pDP, rather than directly applying the MD-synthesizer that had led to poor statistical inferences due to data sparsity.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. arXiv preprint arXiv:1607.00133.

Abowd, J. M. and Villhuber, L. (2008). How protective are synthetic data? In Privacy in Statistical Databases, pages 239–246. Springer.

Acs, G., Castelluccia, C., and Chen, R. (2012). Differentially private histogram publishing through lossy compression. In 2012 IEEE 12th International Conference on Data Mining, pages 1–10. IEEE.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). 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.

Blum, A., Ligett, K., and Roth, A. (2013). A learning theory approach to non-interactive database privacy. Journal of ACM (JACM), 60(2):12.

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

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. JMLR: Workshop and Conference Proceedings, 12:1069–1109.

Chaudhuri, K., Sarwate, A., and Sinha, K. (2012). Near-optimal differentially private principal components. Proc. 26th Annual Conference on Neural Information Processing Systems (NIPS).

Chen, R., Fung, B. C., Mohammed, N., Desai, B. C., and Wang, K. (2013). Privacy-preserving trajectory data publishing by local suppression. Information Sciences, 231:83–97.

Drechsler, J. (2011). Synthetic datasets for Statistical Disclosure Control. Springer, New York.
Dwork, C. (2008). Differential privacy: A survey of results. *Theory and Applications of Models of Computation*, 4978:1–19.

Dwork, C. (2011). Differential privacy. In *Encyclopedia of Cryptography and Security*, pages 338–340. Springer.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer.

Dwork, C. and Roth, A. (2013). The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407.

Dwork, C. and Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

Erlingsson, Ú., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM.

Friedman, A., Berkovsky, S., and Kaafar, M. A. (2016). A differential privacy framework for matrix factorization recommender systems. *User Modeling and User-Adapted Interaction*, pages 1–34.

Gaboardi, M., Honaker, J., King, G., Nissim, K., Ullman, J., and Vadhan, S. (2016). Psi ($\psi$): a private data sharing interface. *arXiv preprint arXiv:1609.04340*.

Gardner, J., Xiong, L., Xiao, Y., Gao, J., Post, A. R., Jiang, X., and Ohno-Machado, L. (2013). Share: system design and case studies for statistical health information release. *Journal of the American Medical Informatics Association*, 20(1):109–116.

Götz, M., Machanavajjhala, A., Wang, G., Xiao, X., and Gehrke, J. (2012). Publishing search logs - a comparative study of privacy guarantees. *IEEE Trans. Knowl. Data Eng.*, 24:5205325.

Gupta, A., Roth, A., and Ullman, J. (2012). Iterative constructions and private data release. In *Theory of Cryptography Conference*, pages 339–356. Springer.

Hardt, M., Ligett, K., and McSherry, F. (2012). A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pages 2339–2347.

Hardt, M. and Rothblum, G. N. (2010). A multiplicative weights mechanism for privacy-preserving data analysis. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 61–70. IEEE.
Hay, M., Rastogi, V., Miklau, G., and Suciu, D. (2010). Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1-2):1021–1032.

He, X., Cormode, G., Machanavajjhala, A., Procopiuc, C. M., and Srivastava, D. (2015). Dpt: Differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165.

Homer, N., Szelinger, S., Redmann, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J., Stephan, D., Nelson, S., and Craig, D. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P.-P. (2012). *Statistical disclosure control*. John Wiley & Sons.

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

Kerman, J. (2011). Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, 5:1450–1470.

Kifer, D., Smith, A., and Thakurta, A. (2012). Private convex empirical risk minimization and high-dimensional regression. *JMLR: Workshop and Conference Proceedings*, 23:25.1–25.40.

Li, H., Xiong, L., and Jiang, X. (2014). Differentially private synthesis of multidimensional data using copula functions. In *Advances in database technology: proceedings. International Conference on Extending Database Technology*, volume 2014, page 475. NIH Public Access.

Liu, F. (2016a). Generalized gaussian mechanism for differential privacy. *arXiv preprint arXiv:1602.06028*.

Liu, F. (2016b). Model-based differentially private data synthesis. *arXiv preprint arXiv:1606.08052*.

Liu, F. (2016c). Statistical properties of sanitized results from differentially private laplace mechanisms with noninformative bounding. *arXiv preprint arXiv:1607.08554*.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. *IEEE ICDE IEEE 24th International Conference*, pages 277 – 286.

Macleod, A. J. and Henderson, G. R. (1984). Bounds for the sample standard deviation. *Teaching Statistics*, 6(3):72–76.
Manrique-Vallier, D. and Reiter, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, 107(500):1385–1394.

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

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

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

Mohammed, N., Chen, R., Fung, B., and Yu, P. S. (2011). Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–501. ACM.

Narayanan, A. and Shmatikov, V. (2008). Robust de anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy*, pages 111–125.

Nissim, K. and Stemmer, U. (2015). On the generalization properties of differential privacy. *CoRR*, abs/1504.05800.

Proserpio, D., Goldberg, S., and McSherry, F. (2012). A workflow for differentially-private graph synthesis. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 13–18. ACM.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531–543.

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

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

Roth, A. and Roughgarden, T. (2010). Interactive privacy via the median mechanism. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 765–774. ACM.
Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Sweeney, L. (2013). Matching known patients to health records in washington state data. *Available at SSRN 2289850*.

Wang, Q., Zhang, Y., Lu, X., Wang, Z., Qin, Z., and Ren, K. (2016). Rescuedp: Real-time spatio-temporal crowd-sourced data publishing with differential privacy. In *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*, pages 1–9. IEEE.

Wang, Y.-X., Fienberg, S. E., and Smola, A. (2015). Privacy for free: Posterior sampling and stochastic gradient monte carlo. *Blei, D., and Bach, F., eds*, 951(15):2493–2502.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.

Xiao, Y., Gardner, J., and Xiong, L. (2012). Dpcube: Releasing differentially private data cubes for health information. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1305–1308. IEEE.

Xiao, Y. and Xiong, L. (2015). Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309. ACM.

Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., and Winslett, M. (2013). Differentially private histogram publication. *The VLDB Journal*, 22(6):797–822.

Yu, F., Fienberg, S. E., Slavkovic, A. B., and Uhler, C. (2014). Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2014). Privbayes: Private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1423–1434. ACM.
Supplemental Materials for “Comparative Study of Differentially Private Data Synthesis Methods”

This file contains the supplementary materials to accompany the paper “Comparative Study of Differentially Private Data Synthesis Methods” with additional results from the four simulation studies.

- Simulation study 1: Tables 3 and 4 show the proportions of usable simulation repeats for all DIPS methods. A usable simulation repeat is defined as a simulated data set that leads to a synthetic data set that contains at least one of each of the two levels of the binary variable. Figure 8 depicts the zoomed-in inferential results when $\pi = 0.50$.

- Simulation study 2: Table 5 shows the summary statistics for the number of histogram bins used in histogram-based DIPS methods; Figures 9 and 10 depict the inferential results of $\mu$ and $\sigma^2$ when the bounds of the data were $[\mu - 4\sigma, \mu + 4\sigma]$ (the main text contains the results on data with asymmetric bounds $[\mu - 3\sigma, \mu + 5\sigma]$).

- Simulation study 3: Table 6 contains the true values of $\mu_1$, $\mu_2$, and $\pi$ for the 24 cells used for simulating the data. Table 7 shows the summary statistics for the number of 2-dimensional histogram bins formed by the continuous variables $(z_1, z_2)$ in each of the 24 cell (needed in the histogram-based DIPS methods); Figures 11 and 12 depict the inferential results on $\mu_2$, $\rho$, $\sigma^2_1$, and $\sigma^2_2$; Table 8 contains the average frequency of empty cells in a synthetic data set.

- Simulation study 4: Table 9 shows the summary statistics for the number of 2-dimensional histogram bins formed by the continuous variables $(z_1, z_2)$ in each of the 12 cells formed by the categorical variables $w$ (needed in the histogram-based DIPS methods); Figures 13 to 16 depict the inferential results of the parameters $\mu_1$, $\mu_2$, $\sigma^2_1$, $\sigma^2_2$, $\rho$, $\beta_1$, and $\beta_4$. 
In Tables 3 and 4, a usable simulation repeat is defined as a simulated data set that leads to a synthetic data set that contains both levels of the binary variable.

Table 3: The proportion of usable simulation repeats (out of 5000) based on the synthetic data via the Laplace sanitizer in Simulation 1. Only the proportions from \( \ln(\epsilon) = -10 \) to 0 are presented since \( \ln(\epsilon) > 0 \) leads to 100% usable repeats.

| \( \ln(\epsilon) \) | -10 | -8 | -6 | -4 | -2 | 0 |
|---------------------|-----|----|----|----|----|---|
| \( n = 40, \pi = 0.15 \) | 0.9374 | 0.9390 | 0.9378 | 0.9532 | 0.9852 | 0.9998 |
| \( n = 100, \pi = 0.15 \) | 0.9392 | 0.9332 | 0.9484 | 0.9684 | 0.9954 | 1.0000 |
| \( n = 1000, \pi = 0.15 \) | 0.9388 | 0.9464 | 0.9752 | 0.9972 | 1.0000 | 1.0000 |
| \( n = 40, \pi = 0.25 \) | 0.9420 | 0.9414 | 0.9394 | 0.9566 | 0.9900 | 1.0000 |
| \( n = 100, \pi = 0.25 \) | 0.9354 | 0.9396 | 0.9436 | 0.9708 | 0.9984 | 1.0000 |
| \( n = 1000, \pi = 0.25 \) | 0.9408 | 0.9486 | 0.9796 | 0.9992 | 1.0000 | 1.0000 |
| \( n = 40, \pi = 0.50 \) | 0.9374 | 0.9384 | 0.9390 | 0.9574 | 0.9952 | 1.0000 |
| \( n = 100, \pi = 0.50 \) | 0.9372 | 0.9338 | 0.9410 | 0.9754 | 0.9998 | 1.0000 |
| \( n = 1000, \pi = 0.50 \) | 0.9374 | 0.9442 | 0.9840 | 1.0000 | 1.0000 | 1.0000 |

Table 4: The proportion of usable simulation repeats (out of 5000) based on the synthetic data via the MODIPS, MD, and BB-MR methods in Simulation 1. Only the proportions from \( \ln(\epsilon) = -9 \) and -10 are presented since \( \ln(\epsilon) > -9 \) leads to 100% usable repeats.

| \( \ln(\epsilon) \) | MODIPS-BIT | MD | BB-MR |
|---------------------|------------|----|-------|
| \( n = 40, \pi = 0.15 \) | 0.9978 | 1.0000 | 1.0000 | 1.0000 |
| \( n = 100, \pi = 0.15 \) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| \( n = 1000, \pi = 0.15 \) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| \( n = 40, \pi = 0.25 \) | 1.0000 | 0.9998 | 1.0000 | 1.0000 |
| \( n = 100, \pi = 0.25 \) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| \( n = 1000, \pi = 0.25 \) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| \( n = 40, \pi = 0.50 \) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| \( n = 100, \pi = 0.50 \) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| \( n = 1000, \pi = 0.50 \) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
Figure 8: Zoomed-in results on the bias, RMSE, and CP of $\pi = 0.50$ in Simulation 1. MODIPS represents the model-based differentially private synthesis, LAP represents the Laplace synthesizer, MD represents the Multinomial-Dirichlet synthesizer, BB-MR represents the BB-MR synthesizer, MS is the traditional non-DP MS, and Ori is the original results without any perturbation.

Table 5: Summary statistics for the number of bins across 5000 simulations repeats in simulation study 2.

| Scenario      | Min | Mean | Median | Max |
|---------------|-----|------|--------|-----|
| $n = 40, \sigma^2 = 1, 4, 9$ | 5   | 7.459| 7      | 12  |
| $n = 100, \sigma^2 = 1, 4, 9$ | 8   | 9.853| 10     | 13  |
| $n = 1000, \sigma^2 = 1, 4, 9$ | 19  | 20.54| 21     | 22  |
Figure 9: The bias, RMSE, CP, and width of the 95% CI of $\mu$ in simulation study 2 when the data bounds were $[\mu - 4\sigma, \mu + 4\sigma]$. MODIPS (with BIT) represents the model-based differentially private data synthesis with $\epsilon$ split 1:1 between $\mu$ and $\sigma^2$, PERT represents the perturbed histogram method, SMOOTH represents the smoothed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.
Figure 10: The relative bias, scaled RMSE, CP, and scaled width of the 95% CI of $\sigma^2$ in simulation study 2 when the bounds were $[\mu - 4\sigma, \mu + 4\sigma]$. MODIPS (with BIT) represents the model-based differentially private data synthesis with $\epsilon$ split 1:1 between $\mu$ and $\sigma^2$, PERT represents the perturbed histogram method, SMOOTH represents the smoothed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.
Table 6: The true values of $\mu_1$, $\mu_2$, and $\pi$ for the 24 cells in simulation study 3

| Cell | $\mu_1$ | $\mu_2$ | $\pi$ |
|------|---------|---------|-------|
| 1    | 1.371   | -0.565  | 0.041 |
| 2    | 0.363   | 0.633   | 0.076 |
| 3    | 0.404   | -0.106  | 0.024 |
| 4    | 1.512   | -0.095  | 0.062 |
| 5    | 2.018   | -0.063  | 0.045 |
| 6    | 1.305   | 2.287   | 0.041 |
| 7    | -1.389  | -0.279  | 0.038 |
| 8    | -0.133  | 0.636   | 0.007 |
| 9    | -0.284  | -2.656  | 0.064 |
| 10   | -2.440  | 1.320   | 0.064 |
| 11   | -0.307  | -1.781  | 0.031 |
| 12   | -0.172  | 1.215   | 0.048 |
| 13   | 1.895   | -0.430  | 0.053 |
| 14   | -0.257  | -1.763  | 0.007 |
| 15   | 0.460   | -0.640  | 0.021 |
| 16   | 0.455   | 0.705   | 0.065 |
| 17   | 1.035   | -0.609  | 0.070 |
| 18   | 0.505   | -1.717  | 0.012 |
| 19   | -0.784  | -0.851  | 0.024 |
| 20   | -2.414  | 0.036   | 0.028 |
| 21   | 0.206   | -0.361  | 0.048 |
| 22   | 0.758   | -0.727  | 0.011 |
| 23   | -1.368  | 0.433   | 0.058 |
| 24   | -0.811  | 1.444   | 0.062 |
Table 7: Summary statistics for the bin number in the 2-dimensional histogram formed by the two continuous variables in the 24 cells formed by the categorical cells (used in the perturbed histogram method) across the 5000 repeats in simulation 3.

| Cell | Min | Mean | Median | Max |
|------|-----|------|--------|-----|
| 1    | 36  | 37.78| 36     | 49  |
| 2    | 49  | 51.09| 49     | 64  |
| 3    | 25  | 34.39| 36     | 36  |
| 4    | 42  | 49.01| 49     | 56  |
| 5    | 36  | 42.06| 42     | 49  |
| 6    | 36  | 37.78| 36     | 49  |
| 7    | 36  | 36.38| 36     | 49  |
| 8    | 16  | 16.04| 16     | 20  |
| 9    | 49  | 49.04| 49     | 56  |
| 10   | 49  | 49.04| 49     | 56  |
| 11   | 36  | 36.00| 36     | 36  |
| 12   | 36  | 45.60| 49     | 49  |
| 13   | 36  | 48.49| 49     | 49  |
| 14   | 16  | 16.04| 16     | 20  |
| 15   | 25  | 28.59| 30     | 36  |
| 16   | 49  | 49.06| 49     | 56  |
| 17   | 49  | 49.37| 49     | 64  |
| 18   | 20  | 25.00| 25     | 25  |
| 19   | 25  | 34.39| 36     | 36  |
| 20   | 25  | 35.97| 36     | 36  |
| 21   | 36  | 45.60| 49     | 49  |
| 22   | 16  | 24.92| 25     | 25  |
| 23   | 36  | 48.97| 49     | 56  |
| 24   | 42  | 49.01| 49     | 56  |
Table 8: The average frequency of empty cells out of 24 per synthetic data set by the MODIPS and the perturbed histogram methods in simulation study 3.

| ln(\(\epsilon\)) | MODIPS | Perturbed Histogram |
|-------------------|--------|---------------------|
| -6                | 4.58   | 23.66               |
| -5                | 4.58   | 23.63               |
| -4                | 4.49   | 23.57               |
| -3                | 4.48   | 23.25               |
| -2                | 4.42   | 21.19               |
| -1                | 4.39   | 7.50                |
| 0                 | 4.12   | 4.05                |
| 1                 | 3.43   | 2.31                |
| 2                 | 2.54   | 1.09                |
| 3                 | 1.31   | 0.46                |
| 4                 | 1.00   | 0.41                |
| 5                 | 0.95   | 0.43                |
| 6                 | 0.95   | 0.40                |
| 7                 | 0.85   | 0.46                |
| 8                 | 0.98   | 0.37                |
| 9                 | 0.88   | 0.46                |
| 10                | 0.88   | 0.43                |
Figure 11: The bias, RMSE, CP, and width of the 95% CI of $\mu_2$ and $\rho$ in simulation study 3. MODIPS represents the model-based differentially private data synthesis with $\epsilon$ split equally among the six sets of sufficient statistics, NP-DIPS represents the Laplace sanitizer + perturbed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.
Figure 12: The bias, RMSE, CP, and width of the 95% CI of $\sigma_1^2$ and $\sigma_2^2$ in simulation study 3. MODIPS represents the model-based differentially private data synthesis with $\epsilon$ split equally among the six sets of sufficient statistics, NP-DIPS represents the Laplace sanitizer + perturbed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.
Table 9: Summary statistics for the number of 2-dimensional histogram bins formed by the two continuous variables $(z_1, z_2)$ in each of the 12 cells by the categorical variables $w$ (needed in the NP-DIPS approach) in simulation study 4.

| Cell | Min | Mean | Median | Max |
|------|-----|------|--------|-----|
| 1    | 49  | 65.23| 64     | 81  |
| 2    | 56  | 64.53| 64     | 81  |
| 3    | 64  | 78.78| 81     | 100 |
| 4    | 36  | 50.90| 49     | 64  |
| 5    | 9   | 19.81| 16     | 30  |
| 6    | 64  | 79.23| 81     | 100 |
| 7    | 36  | 47.67| 49     | 64  |
| 8    | 64  | 77.26| 81     | 90  |
| 9    | 16  | 25.74| 25     | 36  |
| 10   | 0   | 13.89| 16     | 25  |
| 11   | 36  | 47.71| 49     | 64  |
| 12   | 16  | 26.35| 25     | 36  |
Figure 13: Bias, RMSE, CP, and 95% CI width of $\mu_1$ and $\mu_2$ in simulation study 4. MODIPS represents the model-based differentially private data synthesis with $\epsilon$ split equally among the nine sufficient statistics, NP-DIPS represents the Laplace sanitizer + perturbed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.
Figure 14: Bias, RMSE, CP, and 95% CI width of $\sigma_1^2$, $\sigma_2^2$, and $\rho$ in simulation study 4. MODIPS represents the model-based differentially private data synthesis with $\epsilon$ split equally among the nine sufficient statistics, PERT represents the perturbed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.
Figure 15: Bias, log of RMSE, CP, and log of the 95% CI width for $\beta_1$ in simulation study 4. MODIPS represents the model-based differentially private data synthesis with $\epsilon$ split equally among the nine sufficient statistics, PERT represents the perturbed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.
Figure 16: Bias, log of RMSE, CP, and log of the 95% CI width for $\beta_4$ in simulation study 4. MODIPS represents the model-based differentially private data synthesis with $\epsilon$ split equally among the nine sufficient statistics, PERT represents the perturbed histogram method, MS is the traditional MS method without DP, and Ori is the original results without perturbation.