Statistical Science
2020, Vol. 35, No. 2, 280–307
https://doi.org/10.1214/19-STS742

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 is a statistical disclosure limitation technique for releasing synthetic data sets with pseudo individual records. Traditional data synthesis 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 data synthesis 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.

Key words and phrases: Differential privacy, DIPS, sufficient statistics, parametric DIPS, nonparametric DIPS, statistical disclosure limitation.

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. These techniques aim to provide protection for sensitive information while releasing information and data to the public. Data synthesis is a statistical disclosure limitation technique that focuses on releasing individual-level data synthesized based on the information in the original data (Rubin, 1993, Little, 1993, Liu and Little, 2003, Raghunathan, Reiter and Rubin, 2003, Reiter, 2003, 2009, Little, Liu and Raghunathan, 2004, Drechsler, 2011). Multiple synthetic sets of the identical structure are often released as a way to propagate the uncertainty arising from the synthesis process, a procedure referred to as multiple synthesis (MS). Methods have been developed to combine the results from multiple synthetic data sets to yield valid statistical inferences (Raghunathan, Reiter and Rubin, 2003, Reiter, 2002, 2003). However, existing disclosure risk assessment approaches for statistical disclosure limitation 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, provides strong privacy guarantee in mathematical terms without making assumptions about the background knowledge of data intruders (Dwork et al., 2006a, Dwork, 2008, 2011). In brief, if a statistic is released via a ϵ-differentially private mechanism, then when the statistic is calculated from

Claire McKay Bowen is Lead Data Scientist, Urban Institute, 500 L'Enfant Plaza SW, Washington, DC 20024, USA (e-mail: cbowen@urban.org). Fang Liu is Associate Professor, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, 201B Crowley Hall, Notre Dame, Indiana 46556, USA (e-mail: fang.liu.131@nd.edu).

two neighboring data sets that differ by one record, the log-difference on the probability to obtain a specific value of that statistic is bounded between $(-\epsilon, \epsilon)$. In layman's terms, DP means the chance an individual will be identified based on the sanitized statistic is low (the smaller ϵ is, the lower the probability is) since the statistic would be about the same with or without the individual in the database.

DP has spurred a great amount work in developing differentially private mechanisms in general settings (Dwork et al., 2006a, 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, Monteleoni and Sarwate, 2011, Kifer, Smith and Thakurta, 2012), principle component analysis (Chaudhuri, Sarwate and Sinha, 2012), genetic association tests (Yu et al., 2014), Bayesian learning (Wang, Fienberg and Smola, 2015), location privacy (Xiao and Xiong, 2015), recommender systems (Friedman, Berkovsky and Kaafar, 2016), deep learning (Abadi et al., 2016), among others. Software or web-based interfaces to generate differentially private statistics are also in development, such as RescueDP (Wang et al., 2016), an online aggregate monitoring scheme that publishes real-time population statistics on spatial-temporal, crowd-sourced data from mobile phone users with DP and Private data Sharing Interface (Gaboardi et al., 2016) that aims to allow data sharing among researchers in the social sciences and other fields while satisfying DP.

DP was originally developed and is widely used for releasing aggregate or summary statistics to answering queries submitted to a database. However, query-based data release has several shortcomings. The requirement to prespecify the level of privacy budget ϵ 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, Ligett and Roth (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 explored 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. (2017) created PrivBayes to release high-dimensional data from Bayesian networks with binary nodes and low-order interactions among the nodes. Li, Xiong and Jiang (2014a) developed DPCopula for synthesizing multivariate data by using Copula functions to take into account the dependency structure. Liu (2016) 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, Goldberg and McSherry, 2012), and 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, it examines and compares some of the general DIPS approaches based on the statistical and inferential utility of the respective synthesized data; both conceptually and empirically via simulation studies and a real-life case study. 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 some of the DIPS methods introduced in Section 3 via four simulation studies on different types of data. Section 5 evaluates the practical feasibility of DIPS on real-world data. Concluding remarks are given in Section 6.

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 $\mathbf{x} = \{x_{ij}\}$ of dimension $n \times p$ $(i = 1, \ldots, n;\ j = 1, \ldots, p)$. Each row $\mathbf{x}_i$ represents an individual record with p variables/attributes.

2.1 Differential Privacy (DP) and Composition Properties

DEFINITION 2.1 (Differential privacy (Dwork et al., 2006a)). A sanitization algorithm $\mathcal{R}$ gives ϵ-DP if for all data sets $(\mathbf{x}, \mathbf{x}')$ that is $d(\mathbf{x}, \mathbf{x}') = 1$ and all subsets $Q \subseteq \mathcal{T}$,

$$\left|\log\left(\frac{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x})) \in Q)}{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}')) \in Q)}\right)\right| \leq \epsilon, \tag{2.1}$$

where $\mathcal{T}$ denotes the output range of the algorithm $\mathcal{R}$, $\epsilon > 0$ is the privacy budget (or loss) and $\mathbf{s}$ denotes the statistics. $d(\mathbf{x}, \mathbf{x}') = 1$ implies that $\mathbf{x}'$ differs from $\mathbf{x}$ by only one individual. Mathematically, equation (2.1) states that the log-difference on the probability of obtaining a specific value of $\mathbf{s}$ via $\mathcal{R}$ is bounded by $(-\epsilon, \epsilon)$ when it is calculated from two neighboring data sets that differ by one record. If ϵ is small, then the chance an individual will be identified based on the sanitized query result is low since the query result would be about the same with or without the individual in the database. Inversely, if ϵ is larger, then the more differentiable the sanitized results are when an individual is absent or present in the data set.

Regarding what value of ϵ is considered to be appropriate or acceptable for practical use, Dwork (2008) discussed the choice of ϵ is a social question (and "beyond the scope of" her paper), but suggests $0.01 \sim \ln(3)$, or even up to 3 in some cases, as possible ϵ values. Lee and Clifton (2011) stated that ϵ does not easily relate to practically relevant measures of privacy and suggest a formula to calculate ϵ if the goal is to hide any individual's presence (or absence) in the database. The formula relies on some assumptions, like query-dependency and also requires knowing the data universe as well as the subset of that universe to be queried. Abowd and Schmutte (2015) examined the question from the economic perspective by accounting for the public-good properties of privacy loss and data utility, and quantify the optimal choice of ϵ by formulating a social planner's problem and incorporating (ϵ, δ)-DP (another relaxation of the pure DP; see Definition 2.5) and (α, δ)-accuracy (the l_1 error in released statistics is bounded by α with probability $1 - \beta$) to release normalized histograms via the private multiplicative weights method. In two applications, they have examined the optimal ϵ is 0.067 and 0.044 for $\alpha = 0.096$ and 0.073 (resp.) when $\beta = 0.01$ and $\delta = 0.9/N$, where N is the population size for some specific settings on the population size and query set size. If deemed valid, the ϵ values suggested in the (ϵ, δ)-DP setting can also be considered for ϵ-DP as the latter implies stricter privacy protection at the same ϵ value. However, the caveat of possible worse information preservative compared to its relaxed counterpart. Other ϵ values also came up in the literature. For example, Machanavajjhala et al. (2008) applied DP in the OnTheMap data (commuting patterns of the US population) and used $(\epsilon = 8.6, \delta = 10^{-5})$-probabilistic DP (a relaxation of the pure DP in equation (2.1); see Definition 2.6) to synthesize commuter data. Karwa, Krivitsky and Slavković (2017) use $\epsilon = 3$ and $\epsilon = 6$ when synthesizing edges in social networks via a randomized response mechanism with ϵ-edge DP. Ding et al. (2011) and Li, Xiong and Jiang (2014b) use $\epsilon = 1$ in their experiments.

All of the work above suggests there are many factors that affect the choice of ϵ, including the type of information released to the public, social perception of privacy protection, statistical accuracy of the release data, among others. Also, that it remains an open question that warrants more research and further investigation. The smaller ϵ is, the less the privacy loss, but the less accurate the released information. Choosing an "appropriate" ϵ is essentially finding a good trade-off between privacy loss and released information accuracy. We will provide more discussion regarding the choice of ϵ in Section 6, reviewing what we have learned from the literature and the simulation/case studies.

Often in practice a data set is queried for multiple statistics especially when the data is high-dimensional. Every time the data set is queried, there is a privacy cost (loss) as information is being asked about the same set of individuals. 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; say, r queries are sent to the same data set with a total privacy budget of ϵ. The data curator could allocate ϵ/r privacy budget to each of the r queries to maintain the total privacy cost at ϵ. On the other hand, if each query is sent to a disjoint set of data such that each set has no overlapping individuals, then the privacy cost does not accumulate. A typical example is the release of a histogram, where the counts in different bins of the histogram are based on disjoint subsets of data, and each bin is perturbed with the full privacy budget ϵ. These principles are presented in the sequential composition and parallel composition theorems below.

THEOREM 2.2 (Composition theorems (McSherry, 2009)). *Suppose a differentially private mechanism $\mathcal{R}_j$ provides ϵ_j-DP for $j = 1, \ldots, r$.*

(a) Sequential composition: *The sequence of $\mathcal{R}_j(\mathbf{x})$ executed on the same data set $\mathbf{x}$ provides $(\sum_j \epsilon_j)$-DP.*

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

2.2 Relationship Between DP and Disclosure Risk Assessment in the Traditional Statistical Disclosure Limitation Setting

The concept of DP is different from the traditional disclosure risk assessment in statistical disclosure limitation. The former does not rely on any background knowledge

or behavioral assumptions of a data intruder while the latter often models what the data intruder knows and how the disclosure risk is formulated or calculated and could vary significantly, depending on the data and the approaches for assessing disclosure risk, lacking a unified principle. Provided below is a concrete example that illustrates the differences between DP and the traditional disclosure risk assessment.

Suppose a data set contains 11 attributes, one out of the 11 is a sensitive variable, such as HIV status, and the other 10 are pseudo-identifiers such as age, gender, etc. In a typical disclosure risk assessment, the data curator would first make an assumption about what the intruder knows and what the intruder will do to obtain the information she/he is interested in. Therefore, the curator will likely assume in this case that: (1) the intruder A wants the information on the sensitive variable on individual B, and A knows that B is in the data set; (2) A knows all 10 pseudo-identifiers of B; (3) A fits a logistic model to calculate the probability of having HIV with the released data set. Suppose the true HIV status of B is T and estimated $\Pr(\text{HIV} = \text{T}|\text{the 10 attributes})$ is 5% from the logistic model based on a synthetic copy of the original data; then from the perspective of the traditional disclosure risk assessment, we would consider B is at a lower risk of getting his/her personal information disclosed. However, how confident are we with this 5%? What if the data intruder has more information in addition to the released data? What if the data intruder has a more efficient method than the logistic regression to predict the HIV status with high accurate? In other words, the single value 5% with all the above assumptions could be far from being optimal in reflecting the true disclosure risk.

If the surrogate data set is synthesized via a technique based on the DP framework, then it is guaranteed that any individual (including B) from the original data has little impact on any statistics calculated from the synthetic data, and the impact is quantified by the probabilities of obtaining the same statistic with versus without any single individual, the ratio between which is bounded by $(e^{-\epsilon}, e^{\epsilon})$. In this example, the statistic s is $\Pr(\text{HIV}|\text{the 10 attributes})$, and the the ratio $\frac{\Pr(s^*|\mathbf{x}^*)}{\Pr(s^*|\mathbf{x}'^*)} \in (e^{-\epsilon}, e^{\epsilon})$, where data $\mathbf{x}^*$ and $\mathbf{x}'^*$ differ by one individual and s^* is the sanitized version of the observed original s based on the synthetic data. Using a small ϵ leads to a tight neighborhood $(e^{-\epsilon}, e^{\epsilon})$ around 1, and a small privacy loss.

A reviewer asks if DP can be used as an upper bound for disclosure risk assessment. The above example suggests the way the DP bounds the absolute log-ratio of two distribution functions on the sanitized version of s obtained from two neighboring data sets ($\mathbf{x}$ and $\mathbf{x}'$) by ϵ rather than providing a direct measure on the probability that an individual would be identified or have his/her true value on a sensitive variable disclosed. Lee and Clifton (2011) calculated an upper bound for the posterior probability of a correct guess from an adversary on whether an individual is in a data set given discrete query results sanitized via the Laplace mechanism under some assumptions. However, the bound is not tight.

In summary, the link between DP and the traditional disclosure risk assessment is an interesting topic and open question. One thing for certain is that DP integrates out all the unknowns (e.g., whether and how data intruder would use that data set, whether an individual in a particular data set or will participate in any future studies, etc.) and covers the worst-case scenario, whether the data curator can think of or not, from the perspective of protecting every individual.

2.3 Empirical DP and Local DP

Classical DP has inspired other privacy concepts such as the empirical DP (Abowd, Schneider and Vilhuber, 2013) and local DP (Duchi, Jordan and Wainwright, 2013), both of which look for bounding some type of "privacy" using a single parameter ϵ. We will not examine the two concepts further in this discussion for the reasons given below.

Empirical DP was first proposed for privacy protection in Bayesian mixed-effects modeling. In empirical DP, a prior distribution is designed to guarantee that the log difference on the posterior distribution of a parameter with versus without each of the individuals in the original data is bounded by $(-\epsilon, \epsilon)$. Charest and Hou (2017) showed that empirical DP is more of an empirical measurement of sensitivity, and relates to the so-called "local sensitivity" (Nissim, Raskhodnikova and Smith, 2007) rather than a guarantee or an empirical estimate of DP. In addition, empirical DP is computationally sensitive to how many posterior samples are drawn and how they are binned in its numerical calculation as the analytical form of the posterior distribution is often not available.

For local DP, although its mathematical formulation seems similar to the classical DP, the two are conceptually different. In local DP, the true response of an individual goes through a locally ϵ-differentially private randomization mechanism that generates a perturbed response, which is recorded and released. Different from the traditional DP (where the privacy budget ϵ is possessed by a whole data set), each individual receives a privacy budget ϵ in local DP, and the log-difference in the probability of generating the same perturbed response from two different individual responses is bounded by $(-\epsilon, \epsilon)$. Local DP has been applied in practice to collect users' data (Erlingsson, Pihur and Korolova, 2014, Fanti, Pihur and Erlingsson, 2016, Tang et al., 2017); but given its conceptual difference from the classical DP, we leave the in-depth investigation and exploration of local DP for future research (more discussion on the local DP is provided in Section 6).

2.4 Differentially Private Mechanisms

We introduce two commonly used sanitizers to achieve ϵ-DP: the Laplace mechanism and the exponential mechanism. A key concept in the Laplace mechanism is the global sensitivity of $\mathbf{s}$ (Dwork et al., 2006a), defined as the following: For all $(\mathbf{x}, \mathbf{x}')$ that is $d(\mathbf{x}, \mathbf{x}') = 1$, the global sensitivity of statistics $\mathbf{s}$ is $\Delta_{\mathbf{s}} = \max_{\mathbf{x},\mathbf{x}', d(\mathbf{x},\mathbf{x}')=1} \|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{x}')\|_1$. In layman's terms, $\Delta_{\mathbf{s}}$ is the maximum change in terms of l_1 norm a person would expect in $\mathbf{s}$ across all possible configurations of $(\mathbf{x}, \mathbf{x}')$ and $d(\mathbf{x}, \mathbf{x}') = 1$. 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 $\Delta_{\mathbf{s}}$ is the more disclosure risk there will be on the individuals in the data from releasing the original $\mathbf{s}$.

DEFINITION 2.3 (Laplace mechanism (Dwork et al., 2006a)). The Laplace mechanism of ϵ-DP adds independent noises $\mathbf{e}$ sampled from the Laplace distribution with location parameter 0 and scale parameter $\Delta_{\mathbf{s}}\epsilon^{-1}$ to each of the elements of the original result $\mathbf{s}$ to generate perturbed $\mathbf{s}^* = \mathbf{s} + \mathbf{e}$.

Per the Laplace distribution, values closer to the raw results $\mathbf{s}$ have higher probabilities of being released than those that are further away from $\mathbf{s}$. The variance of the Laplace distribution is $2(\Delta_{\mathbf{s}}\epsilon^{-1})^2$, implying the smaller the privacy budget ϵ and/or the larger the $\Delta_{\mathbf{s}}$, the higher the probability that the perturbed result $\mathbf{s}^*$ will be farther way from $\mathbf{s}$ when released. The Laplace mechanism is a quick and simple DP mechanism, but does not apply to all statistics such as statistics that have nonnumerical outputs. McSherry and Talwar (2007) introduces a more general DP mechanism, the exponential mechanism, that applies to all types of queries.

DEFINITION 2.4 (Exponential mechanism (McSherry and Talwar, 2007)). In the exponential mechanism, a utility function u assigns a score to each possible output $\mathbf{s}^*$ and releases $\mathbf{s}^*$ with probability

$$\frac{\exp(u(\mathbf{s}^*|\mathbf{x})\frac{\epsilon}{2\Delta_u})}{\int \exp(u(\mathbf{s}^*|\mathbf{x})\frac{\epsilon}{2\Delta_u})\, d\mathbf{s}^*} \tag{2.2}$$

to ensure ϵ-DP, where $\Delta_u = \max_{\mathbf{x},\mathbf{x}', d(\mathbf{x},\mathbf{x}')=1} |u(\mathbf{s}^*|\mathbf{x}) - u(\mathbf{s}^*|\mathbf{x}')|$ is the maximum change in score u with one row change in the data (if $\mathbf{s}^*$ is discrete, the integral in equation (2.2) is replaced with summation).

Per the exponential mechanism, the probability of returning $\mathbf{s}^*$ increases exponentially with the utility score. For example, if $\mathbf{s}$ is numerical and the goal is to preserve as much original information as possible, metrics measuring the closeness between $\mathbf{s}^*$ and the original $\mathbf{s}$ are good candidates for u such as the negative p-norm distance $-\|\mathbf{s} - \mathbf{s}^*\|_p = -(\sum_{j=1}^r |s_j - s_j^*|^p)^{1/p}$ (Liu, 2019a). When the l_1 norm is used, the exponential mechanism in Definition 2.4 becomes the Laplace mechanism with halved privacy budget (McSherry and Talwar, 2007, Liu, 2019a). 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 2.5) (Dwork and Roth, 2014, Liu, 2019a) and the generalized Gaussian mechanisms that include the Laplace mechanism and the Gaussian mechanism as special cases (Liu, 2019a).

2.5 Relaxations of Pure ϵ-DP

The pure ϵ-DP in Section 2.1 can lead to potentially large amounts of noise injected to query results to achieve a high level of privacy guarantee. This concern has motivated work on relaxing the pure ϵ-DP. We briefly overview three relaxations: approximate differential privacy (aDP), probabilistic differential privacy (pDP) and concentrated differential privacy (cDP).

DEFINITION 2.5 (Approximate differential privacy (Dwork et al., 2006b)). A sanitization algorithm $\mathcal{R}$ gives (ϵ, δ)-aDP if for all data sets $(\mathbf{x}, \mathbf{x}')$ that are $d(\mathbf{x}, \mathbf{x}') = 1$,

$$\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x})) \in Q) \le \exp(\epsilon) \Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}')) \in Q) + \delta, \tag{2.3}$$

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 ϵ-DP is a special case of aDP when $\delta = 0$.

DEFINITION 2.6 (Probabilistic differential privacy (Machanavajjhala et al., 2008)). A sanitization algorithm $\mathcal{R}$ gives (ϵ, δ)-pDP if

$$\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x})) \in \text{Disc}(\mathbf{x}, \epsilon)) \le \delta \tag{2.4}$$

for all data sets $(\mathbf{x}, \mathbf{x}')$ that are $d(\mathbf{x}, \mathbf{x}') = 1$, where $\text{Disc}(\mathbf{x}, \epsilon)$ is the disclosure set $\mathcal{R}(\mathbf{s}(x))$ such that $|\ln(\frac{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}))\in Q)}{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}'))\in Q)})| > \epsilon$. Equation (2.4) can be interpreted as the pure ϵ-DP fails with probability δ.

DEFINITION 2.7 (Concentrated differential privacy (Dwork and Rothblum, 2016)). For all data sets $(\mathbf{x}, \mathbf{x}')$, that is $d(\mathbf{x}, \mathbf{x}') = 1$, a sanitization algorithm $\mathcal{R}$ gives (μ, τ)-cDP if $D_{\text{subG}}(\mathcal{R}(\mathbf{x})\|\mathcal{R}(\mathbf{x}')) \preceq (\mu, \tau)$, where D_{subG} stands for *sub-Gaussian divergence*, defined as follows: two random variables Y and Z are $D_{\text{subG}}(Y\|Z) \preceq (\mu, \tau)$ if and only if $\mathbb{E}(L_{Y\|Z}) \le \mu$ and the centered distribution of $(L_{Y\|Z} - \mathbb{E}(L_{Y\|Z})$ is defined and τ-sub-Gaussian, 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 pDP. First, cDP has a bounded expected privacy loss whereas pDP has an infinite privacy loss with probability δ. Second, cDP has better accuracy without compromising the privacy loss from multiple inquiries (Dwork and Rothblum, 2016).

3. DIFFERENTIALLY PRIVATE DATA SYNTHESIS (DIPS)

We loosely group the currently available DIPS methods into two categories: the nonparametric 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, while in the P-DIPS approach it is constructed based on a parametric distribution or an appropriately defined model for the original data.

3.1 Nonparametric DIPS (NP-DIPS)

When the original data is categorical, the statistics $\mathbf{s}$ targeted for differentially private sanitization are the cell counts or proportions in some types of cross-tabulation in NP-DIPS, from which the synthetic data will be from generated. In the case of continuous data, the NP-DIPS techniques can be applied to generate differentially private histograms, kernel density estimators or empirical distributions. The list of the NP-DIPS covered in this section is given in Table 1. Our goal is not to discuss every NP-DIPS method out there in the literature, which would be impossible to achieve in one paper. The list is not exhaustive, but should provide the readers an idea on how DIPS works in the nonparametric setting.

3.1.1 *Synthesis of categorical data.* 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 $\mathbf{x}$, 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 $\mathbf{x}$, 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 $\mathbf{x}$ by $\boldsymbol{n} = (n_1, \ldots, n_K)$. The Laplace sanitizer perturbs the original $\mathbf{n}$ via $n_j^* = n_j + e_j$, where $e_j \sim \text{Lap}(0, \Delta_{\mathbf{s}}/\epsilon)$ independently for $j = 1, \ldots, K$. $\Delta_{\mathbf{s}}$ is the l_1 global sensitivity from releasing the whole cross-tabulation. $\Delta_{\mathbf{s}}$ can be set at 2 or 1, depending on how $d(\mathbf{x}, \mathbf{x}') = 1$ is defined. Specifically, if the change in one individual refers to the case that n remains the same, but the data in exactly one individual change, then $\Delta_{\mathbf{s}} = 2$. If the change in one individual refers to removal of one individual from the data, then $\Delta_{\mathbf{s}} = 1$. When n is relatively large, say > 30, the difference in the standard deviations of the Laplace noises $\sqrt{2}n^{-1}$ between

TABLE 1
Summary of NP-DIPS approaches discussed in Section 3.1

Sec	Method	Pros	Cons
3.1.1	Laplace sanitizer	simple; fast	not accurate for large number of queries
3.1.1	Fourier transformation	preserves low-order marginals accurately	computationally expensive as the number of attributes increases
3.1.1	multiplicative weights Exponential mechanism	adaptive, preserves consistency of marginals across tables	difficulty of choosing an appropriate iteration number; inaccuracy
3.1.2	perturbed histogram	simple; fast	discretization of continuous attributes; doesn't preserve correlation well
3.1.2	smoothed histogram	simple; fast	discretization of continuous attributes; worse than perturbed histogram in accuracy
3.1.2	empirical cumulative density function via Exponential mechanism	flexible; general	computational infeasibility
3.1.2	kernel density estimator with Gaussian process noise	general	works for (ϵ, δ)-aDP; curse of dimensionality
3.1.3	histogram with constrained inferences	better accuracy than perturbed histogram	constrains are publicly known or inherent
3.1.3	universal histogram	accuracy for low-order counts	less accurate for high-order counts
3.1.3	DPCube	multidimensional data	inefficiency in constructing accurate high-dimensional histograms; performs worse than the Laplace sanitizer
3.1.3	NoiseFirst and StructureFirst	outperforms several other DP methods	low dimensional histograms; nonconsistency as $\epsilon \to \infty$
3.1.3	Exponential Fourier perturbation and P-H Partition	better than NoiseFirst and StructureFirst	depends on histogram compressibility

the two versions is $O(10^{-}2)$. There is no practical difference on which one to use. In the simulation studies and the case study presented later, we used $\Delta_{\mathbf{s}} = 1$. Given the smallest n examined was 40, we expect the results to remain roughly the same if we had used $\Delta_s = 2$.

When $k < p$, Barak et al. (2007) conducted early work on constructing k-way differentially private, consistent, and nonnegative 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. The Fourier transformation based algorithm depends on the linear programming and could be computationally infeasible when p is large. For this reason, we do not evaluate this method in Section 4.

Another approach that can be used to generate individual-level data in the discrete domain is the multiplicative weights Exponential mechanism given a set of linear queries (Hardt, Ligett and McSherry, 2012). The multiplicative weights exponential mechanism yields an differentially private empirical distribution that approximates the original empirical distribution in terms of the input linear queries through an iterative process. It often 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. Since every iteration accesses the original data, the total privacy needs to be divided by the number of iterations. It can be difficult to choose an optimal iteration number especially when p is large. A small number of iterations would not be sufficient to capture the information in the original queries, leading to biased results, while a large number of iterations will introduce too much noise during the data generation process, rendering the synthetic data useless. The inaccuracy of the multiplicative weights exponential mechanism is documented in Li et al. (2016), Vadhan (2017), Kowalczyk et al. (2017) and is also confirmed by the simulation studies we have conducted. For these reasons, we do not evaluate the multiplicative weights exponential mechanism in Section 4.

3.1.2 *Synthesis of numerical data*. 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, sample data 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. In addition, 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), $n_k = \sum_{k=1}^{K} I(\mathbf{x}_i \in B_k)$ 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 ($I(\mathbf{x}_i \in B_k) = 1$ if $\mathbf{x}_i \in B_k$; 0 otherwise), a mean-squared consistent density histogram estimator is $\hat{f}_K(\mathbf{x}) = \sum_{k=1}^{K} K \hat{p}_k I(\mathbf{x} \in B_k)$ (Scott, 2015). A differentially private perturbed histogram is a direct application of the Laplace mechanism. The sanitized bin counts and proportions with ϵ-DP are given by $n_k^* = n_k + e_k$ and $\hat{p}_k^* = n_k^* / \sum_k n_k^*$, respectively, where $e_k \overset{\text{iid}}{\sim} \text{Lap}(0, \Delta_{\mathbf{s}}/\epsilon)$ with $\Delta_{\mathbf{s}} = 1$. The density histogram estimator that satisfies ϵ-DP is thus

$$\hat{f}_K^*(\mathbf{x}) = \sum_{k=1}^{K} K \hat{p}_k^* I(\mathbf{x} \in B_k). \tag{3.1}$$

Note that sanitized n_k^* can be negative since the Laplace noise $\in (-\infty, \infty)$, especially when n_k is small or ϵ is small. Commonly used post-hoc processing approaches include 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, 2019b). To incorporate the uncertainty introduced by the sanitization process, releasing multiple sets of $\tilde{\mathbf{x}}$ is suggested, one set per sanitized $\mathbf{n}^* = \{n_k^*\}_{1:K}$.

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

$$\hat{f}_K^*(\mathbf{x}) = (1 - \lambda) \hat{f}_K(\mathbf{x}) + \lambda \Omega$$

$$\text{where } \Omega = \left(\prod_{j=1}^{p} (c_{j1} - c_{j0}) \right)^{-1} \tag{3.2}$$

$$\text{and } \lambda \geq \frac{K}{K + n(e^{\epsilon/n} - 1)} \tag{3.3}$$

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

(Wasserman and Zhou, 2010). Specifically, surrogate data $\tilde{\mathbf{x}}$ is simulated from

$$h(\tilde{\mathbf{x}}) = \frac{g_{\mathbf{x}}(\tilde{\mathbf{x}})}{\int_{[c_{10},c_{11}]\times\cdots\times[c_{p0},c_{p1}]} g_{\mathbf{x}}(\mathbf{z})\,d\mathbf{z}},$$

$$\text{where } g_{\mathbf{x}}(\tilde{\mathbf{x}}) = \exp\left(-u(\hat{F}_{\mathbf{x}}, \hat{F}_{\tilde{\mathbf{x}}})\frac{\epsilon}{2\Delta_u}\right), \tag{3.4}$$

$$\Delta_u = \sup_{\mathbf{x},\mathbf{x}',\Delta(\mathbf{x},\mathbf{x}')=1} \sup_{\tilde{\mathbf{x}}} |u(\hat{F}_{\mathbf{x}}, \hat{F}_{\tilde{\mathbf{x}}}) - u(\hat{F}_{\mathbf{x}'}, \hat{F}_{\tilde{\mathbf{x}}})|,$$

$\hat{F}_{\mathbf{x}}$ is the original empirical cumulative density function, $\hat{F}_{\tilde{\mathbf{x}}}$ is the empirical cumulative density function's of the sanitized data, u is the utility function that denotes a distance measure between the two cumulative density functions, and Δ_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 $\tilde{\mathbf{x}}$ via the Exponential mechanism defined in equation (3.4) does not seem to be a viable choice in practice. One difficulty lies in defining the set of all possible candidate cumulative density functions, 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.

Hall, Rinaldo and Wasserman (2013) proposed sanitizing kernel density estimator by adding noise from a Gaussian process to yield DP, from which synthetic data can be generated. If a Gaussian kernel is used, they show there is no loss of accuracy in the differentially private kernel density estimator to the original one with the optimal bandwidth that minimizes the integrated mean squared error. However, the method is currently only available for (ϵ, δ)-aDP for $\delta > 0$, and suffers the same curse of dimensionality for large p (Scott, 2015).

3.1.3 *Other NP-DIPS methods*. There are also various extensions to the basic Laplace sanitizer and the perturbed histogram approach with the purposes to improve their accuracy. Hay et al. (2010) suggested boosting the accuracy of differentially private histograms by sorting the bin values after sanitation if the order of the bin size is known to the public. They also developed a universal histogram approach by exploring the inherent consistency in a hierarchical histogram, and proved that the accuracy of lower-order contingency tables/marginals is improved, but at the sacrifice of high-order contingency tables (Qardaji, Yang and Li, 2013, Hay et al., 2016). Xiao, Gardner and Xiong (2012) applied a 2-phase partitioning strategy, DPCube, to multidimensional data cubes. Gardner et al. (2013) implemented DPCube in 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. Additionally, Hay et al. (2016) showed that DPCube performed worse than the Laplace sanitizer. Xu et al. (2013) proposed two mechanisms, NoiseFirst and StructureFirst, that performed well against some DP methods, but only applied to low-dimensional histograms due to long running time. Moreover, StructureFirst is inconsistent; where the error of the statistics does not tend to 0 as ϵ increases to infinity (Qardaji, Yang and Li, 2013, Hay et al., 2016). Acs, Castelluccia and Chen (2012) presented two sanitization techniques, the exponential Fourier perturbation algorithm and the P-H Partition, 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.

In summary, the accuracy improvements, if any, of the above methods over the basic Laplace sanitizer or the perturbed histogram either utilize some constraints that only exist in certain types of histograms/data, or only benefit low dimensional histograms. For these reasons, we do not explore these extensions in the simulations studies in Section 4. That being said, it is of our interest to further explore these extended methods that provide better accuracy in low-dimensional histograms in the future, by coupling them with efficient and accurate dimensional reduction techniques.

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 the original data. In what follows, we describe the Multinomial-Dirichlet synthesizer and other methods motivated by the Multinomial-Dirichlet synthesizer for categorical data, the model-based DIPS (MODIPS) approach for general data types based on a Bayesian modeling framework and sequential regression synthesizers. The list of the P-DIPS covered in this section is given in Table 2. The list is not meant to be exhaustive nor does it list the methods that deal with a specific type of data, but it gives readers an idea on how DIPS works in the parametric setting. The PrivBayes and the DPCopula methods, though listed in Table 2, 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 synthesizer to generate differentially private categorical data in the Bayesian framework. The likelihood of proportions $\boldsymbol{\pi}$ is constructed from $f(\mathbf{n}|\boldsymbol{\pi}) \sim \text{Multinom}(n, \boldsymbol{\pi})$, where $\mathbf{n} = (n_1, \ldots, n_K)$ contains the original cell counts in K categories in the original data and $n = \sum_k n_k$. A Dirichlet prior $f(\boldsymbol{\pi}) = \text{D}(\boldsymbol{\alpha})$ is imposed on $\boldsymbol{\pi}$, where each element of $\boldsymbol{\alpha}$ is set at $\alpha_k^* = n/(e^\epsilon - 1)$, the minimum value that guarantees ϵ-DP, for $k = 1, \ldots, K$. To generate differentially private surrogate data sets, $\boldsymbol{\pi}^*$ is first simulated from the posterior distribution $f(\boldsymbol{\pi}^*|\mathbf{x}) = D(\boldsymbol{\alpha}^* +$

TABLE 2
Summary of Parametric DIPS approaches discussed in Section 3.2

Sec	Method	Data	Pros	Cons
3.2.1	Multinomial-Dirichlet, Binomial-Beta	categorical	straightforward; easy to implement	performs poorly on sparse data; perturbation amount increases with n; possible biased inferences on proportions
3.2.2	Model-based DIPS (MODIPS)	any	general; the DP version of the model-based multiple synthesis without DP	model-dependent; relies on identification and sanitization of sufficient statistics or likelihood functions
3.2.3	Sequential Regression Modeling Synthesizers	any	general; models the correlations among all variables	large amount noises for large p
	PrivBayes	categorical	models dependency among variable; has an inherent model selection component	requires dichotomization on continuous attributes; depends on a quality function that can be computationally inefficient
	DPCopula	any	general	same limitations for copula models in general and quadratic time complexity

$\boldsymbol{n}$), and then synthetic data is drawn from $f(\tilde{\mathbf{n}}|\boldsymbol{\pi}^*) = \text{Multinom}(n, \boldsymbol{\pi}^*)$. To ensure valid inferences in the synthetic data, multiple sets of $\tilde{\mathbf{n}}$ can be released; one for each differentially private $\boldsymbol{\pi}^*$. The Multinomial-Dirichlet synthesizer reduces to the Binomial-Beta synthesizer in the binary case. McClure and Reiter (2012) proposed a slightly different approach to synthesizing binary data from $f(\tilde{\mathbf{n}}|\mathbf{n}) = \text{Binom}(n, \frac{n_1+\alpha_1}{n+\alpha_1+\alpha_2})$, where $\alpha_1 = \alpha_2 = (e^{\epsilon/n} - 1)^{-1}$ to satisfy ϵ-DP, which we refer to as the Binomial-Beta McClure–Reiter approach. The Binomial-Beta McClure–Reiter differs from the Binomial-Beta synthesizer not only in how the prior on π differ, but also that it does not simulate $\boldsymbol{\pi}$ from its posterior distribution thus $\tilde{\mathbf{n}}$ synthesized via the Binomial-Beta McClure–Reiter has one less layer variability.

In both the Multinomial-Dirichlet/Binomial-Beta and the Binomial-Beta McClure–Reiter synthesizers, α_k^* increases with n, implying that when data/observed information increases, the amount of perturbation required to maintain ϵ-DP also increases and can be nontrivial for any n. Furthermore, since all α_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 Binomial-Beta 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 ϵ 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, 2016). 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 $\mathbf{s}$ associated with the model. The posterior distribution of $\boldsymbol{\theta}$ can then be represented as $f(\boldsymbol{\theta}|\mathbf{s})$. The MODIPS then sanitizes $\mathbf{s}$ with privacy budget ϵ/m. Denote the sanitized $\mathbf{s}$ by $\mathbf{s}^*$. Synthetic data $\tilde{\mathbf{x}}$ is simulated given $\mathbf{s}^*$ by first drawing $\boldsymbol{\theta}^*$ from the posterior distribution $f(\boldsymbol{\theta}|\mathbf{s}^*)$, and then simulating $\tilde{\mathbf{x}}^*$ from $f(\mathbf{x}|\boldsymbol{\theta}^*)$. The procedure is repeated m times to generate m surrogate data sets.

Since the MODIPS approach is model-dependent, the identification and validation of an appropriate model for data $\mathbf{x}$ is critical, and model misspecification will generate biased samples. If the identification of a suitable model is based on previous knowledge and common practice, then no privacy will need to be spend; however, if the model selection procedure is based on the the data

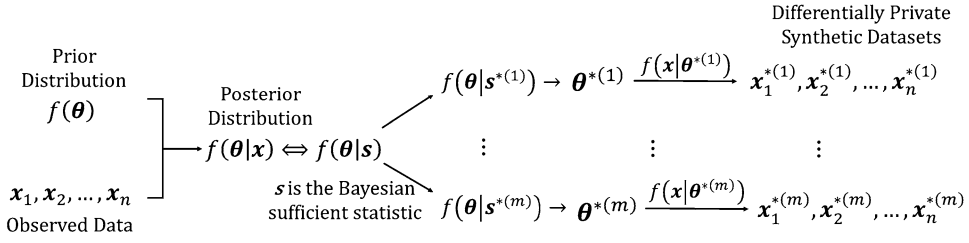

FIG. 1. *The MODIPS algorithm.*

to be released, then the data curator will have to a certain portion of the total privacy budget to the model selection procedure. Differentially private model selection is a separate research topic that is beyond the scope of this paper. When there are several plausible models, the model averaging idea can be incorporated into the synthetic data generation and serves as a mitigation measure to model misspecification and the dependency of the synthetic data on a single synthesis model. The model averaging can be implemented using the Bayesian model averaging method (Hoeting et al., 1999); but it can be analytically and computationally demanding. An alternative approach, less formal but practically more straightforward manner, can also be applied. Say there are 3 reasonable models—M1, M2 and M3—for the original data, with the "plausibility" weights 0.4, 0.3 and 0.3 for each model (e.g., per Bayes factors). Suppose 10 sets of synthetic data are to be released; we could then generate 4 sets of synthetic data from M1, 3 sets from M2 and 3 sets from M3, leading to 10 sets synthesized by 3 different models. The inferences based on the 10 sets (combined using the method given in Section 3.3) will implicitly integrate out the model uncertainty, and are more robust and less sensitive to the model selection and specification than in the case where all 10 sets are generated from a single model. The downside of the model averaging idea is that it will result in more uncertainty in the synthetic data, a price paid for more robustness. In addition, the weights associated with the set of the models can be subjective, even via the Bayes factor approach, which is known for its dependence on the priors.

3.2.3 *Sequential regression modeling synthesizers.* Another method to generate DIPS data is through a sequential regression modeling synthesizer approach, which broadly speaking, can also be regarded as the MODIPS approach. Specifically, suppose the variables from the data are $X_1, X_2, \ldots, X_p$. The joint distribution of $f(X_1, \ldots, X_p)$ can be decomposed as $f(X_1)f(X_2|X_1)\cdots f(X_p|X_1, \ldots, X_{p-1})$, suggesting data can be generated sequentially by first synthesizing X_1 from $f(X_1)$, then X_2 from the conditional model of X_2 given X_1, and so on. Each of the p regression models needs to differently private and can borrow the existing framework on differentially private empirical risk minimization (ERM) or differentially private regression models. For instance, Chaudhuri and Monteleoni (2009) proposed directly perturbing the minimizers or perturbing the empirical risk to obtain differentially private minimizers in l_2 regularized logistic regression, which is extended to differentially private empirical risk minimization with differentiable and strongly convex regularizers in Chaudhuri, Monteleoni and Sarwate (2011) and with a nondifferentiable regularizer in Kifer, Smith and Thakurta (2012). Zhang et al. (2012) proposed the functional mechanism that adds noise to the objective function using the Laplace mechanism and estimates the global sensitivity through a polynomial representation. The functional mechanism applies to both linear and logistic regression, but the latter is based on the approximation through the Taylor expansion and is susceptible to a large amount of noise (Zhang et al., 2013). Sheffet (2017) examined differentially private inferences (hypothesis testing and confidence interval construction) for ordinary least squares and ridge estimator in linear regression.

The sequential regression modeling synthesizer accounts for the correlations among the various variables. In addition, the synthesizer can be implemented in most data types as long as the differentially private versions for the commonly used regression model types are available. The sequential modeling framework has been successfully implemented in practice for imputing missing data (Raghunathan et al., 2001, Raghunathan, Solenberger and Hoewyk, 2017) and data synthesis (Kinney et al., 2011)). Used as a DIPS methods, there are a few potential drawbacks. First, the total privacy budget needs to be divided into p portions due to the DP sequential composition, and each regression model receives only a single portion. If p is large, this approach could perform poorly in terms of statistical utility of the synthetic data due to the lack of privacy budget per regression. Most of the DP regression techniques mentioned above often output a single point estimate for the parameters involved in each regression models, which can be plugged in to generate synthetic data. To properly propagate the uncertainty around the parameters, we either have to model the synthesis process analytically or release multiple synthetic data sets by drawing and plugging in multiple sets of parameters as in the MODIPS approach. Third, efficient DP regression models are not available for all model types (e.g., the Cox regression for survival data). One possible solution that circumvents regression modeling is the STEPS approach (Bowen and Liu, 2018), a nonparametric synthesizer that is also based sequentially "modeling" of the data.

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 Multinomial-Dirichlet 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, such as in Charest (2010) for synthesizing binary data and in Karwa, Krivitsky and Slavković (2017), where the edges of a social network are synthesized via a randomized response mechanism under ϵ-edge DP in the

exponential random graph models and then likelihood-based inference for missing data and Markov chain Monte Carlo techniques (more specifically, Metropolis–Hastings algorithms) are applied to model the synthesis process. This approach can be demanding for data users both analytically and computationally. The second approach is to release multiple sets of synthetic data, which can be regarded as a Monte Carlo version of the former. 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 β. Denote the estimate of β in the lth synthetic data by $\hat{\beta}^{(l)}$ and the associated standard error by $v^{(l)}$. The final point estimate β is

$$\bar{\beta} = m^{-1} \sum_{l=1}^{m} \hat{\beta}^{(l)} \tag{3.5}$$

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

$$T = m^{-1} B + W, \tag{3.6}$$

where $B = \sum_{l=1}^{m} (\hat{\beta}^{(l)} - \bar{\beta})^2/(m-1)$ (between-set variability) and $W = m^{-1} \sum_{j=1}^{m} v^{2^{(l)}}$ (average per-set variability); and tests and confidence intervals are based on

$$(\bar{\beta} - \beta) T^{-1/2} \sim t_{\nu=(m-1)(1+mW/B)^2}. \tag{3.7}$$

The formal proof of the variance combination rule for the MODIPS approach can be found in Liu (2016).

The estimator in equation (3.6) is the same as the variance combination rule in Reiter (2003) for obtaining inferences from multiply synthetic sets in the context of non-DP setting for partial synthesis. The equivalence between the two is not just a random coincidence but likely due to that the synthesis processes are similar between the two, with the only difference in the extra variability brought into the synthetic data via a differentially private mechanism in the DP setting, which is nicely captured by the between-set variability component B and does not affect how B and W are combined. Given this extra variability, inferences from the differentially privately synthesized data will be less precise than those from non-DP MS approaches—a price paid for the DP guarantee. Although not formally proved, it is expected equation (3.5) to (3.7) also apply in the Multinomial-Dirichlet synthesizer and other DIPS approaches that use multiple set releases to account for sanitization and synthesis uncertainty, though the sources that compose B would differ. Reiter and Kinney (2012) and Raab, Nowok and Dibben (2017) suggested that the estimator $W + B/m$, although derived for partial synthesis, also works for complete (full) synthesis and does not require the synthetic data to be generated from the posterior predictive distribution. These arguments further connects $W + B/m$ with the DIPS methods and support its potential as the variance estimator for the DIPS methods in general. First, DIPS falls under the complete synthesis scenario, but without generating a synthetic population from which a synthetic sample data set is drawn or having any known population auxiliary variables X as referred to by Raghunathan, Reiter and Rubin (2003). Second, many DIPS methods do not synthesize data from posterior predictive distributions (e.g., the Laplace sanitizer), which is a case that $W + B/m$ can accommodate.

The estimator for the DIPS presented in equation (3.6) is the first proposed variance estimator in the DP setting. Although there exist several variance estimators in the non-DP setting (see Raab, Nowok and Dibben (2017), it provides an overview of estimators and recommendations on which one to use in different scenarios. We conducted simulation studies to examine how these variance estimators would work in the DIPS setting. While the simulations studies are neither comprehensive nor confirmatory, they provide some interesting findings: (1) $(1 + m^{-1})B - W$ (Raghunathan, Reiter and Rubin, 2003) can lead to an underestimation of the variance, an undercoverage of CIs when ϵ is small and an overcoverage when ϵ is large; (2) $W(1 + 2/m)$ (Raab, Nowok and Dibben, 2017) can lead to a severe undercoverage when ϵ is small and deliver nominal coverage when ϵ is relatively large; (3) $W + (1 + 1/m)B$, the very first combination rule for inferences in multiple imputation in the missing data setting leads to an overcoverage; (4) $W + B/m$ in equation (3.6) delivers the nominal coverage in all the examined simulation scenarios. In summary, $W + B/m$ seems to work the best based on the theoretical and empirical evidence collected so far; but both aspects are somewhat limited in depth and scope, calling for more research on the development and validation of the variance estimator in the setting of DIPS.

4. SIMULATION STUDIES

We assess the utility and inferential properties of the sanitized data via some of the DIPS approaches presented in Section 3 through four simulation studies. We examine the approaches in the setting of the pure ϵ-DP through the application of the Laplace mechanism, but all the DIPS approaches can be applied under softer versions of DP (e.g., (ϵ, δ)-pDP) via the employment of appropriate sanitizers (such as the Gaussian mechanism). The first and second simulation studies focus on univariate categorical data and univariate continuous data, respectively; the third and fourth simulation studies involve a mixture of categorical and continuous variables, but data are generated from different models.

Despite the simplicity of the first and second simulation studies, the results on the impacts of DP on the statistical inferences are rather insightful, especially considering there is very little research in comparing the utility

of various DIPS approaches in terms of statistical inferences. These simulations also provide justifications for the choices of the DIPS approaches used in the third and fourth simulation studies. In the fourth simulation, we examine the effect of misspecification on the synthesis model for the MODIPS approach, and investigate the importance of selection and validation of an appropriate synthesis model. We did not implement the model selection/validation procedures in simulation studies 1 to 2 due to the simplicity of the data and thus the obvious choice of an appropriate model. We also did not conduct the model selection/validation in simulation study 3 given that its similarity in the data structure with simulation 4. There is no difference whether simulation studies 3 or 4 is used for the purposes of illustrating the model selection/validation procedure in the MODIPS approach.

We varied the privacy budget ϵ from e^{-4} to e^4 in each simulation to examine its effect on the statistical inferences. While ϵ as large as e^4 might not be used in practice due to privacy considerations, it is a useful theoretical exploration on the amount of privacy sacrifice in order to have inferences based on the synthetic data to be close to the original; likewise, ϵ as small as e^{-4} helps us to understand what level of privacy would ruin the inferences to an unacceptable degree based on the synthetic data. In all the examined DIPS approaches, the sample size of each released synthetic set is the same as the original data, and 5 sets of synthetic data are generated in DIPS approaches except for the smoothed histogram and the Binomial-Beta McClure–Reiter approaches for reasons stated in Section 3. For the DIPS approaches that generate 5 synthetic data sets, each synthesis receives $1/5$ of the total privacy budget ϵ per the sequential composition principle. The inferences based on the DIPS synthetic data are benchmarked against those based on the original data and the traditional non-DP MS technique.

4.1 Simulation Study 1: Categorical Data

The following DIPS methods are compared in this simulation study: the MODIPS synthesizer, the Laplace sanitizer, the Binomial-Beta McClure–Reiter synthesizer and the Multinomial-Dirichlet 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 $n \in \{40, 100, 1000\}$ and $\pi \in \{0.10, 0.25, 0.50\}$, with 5000 repetitions per scenario.

The non-DP MS and the MODIPS approaches are model-based, and usually model selection and validation should be applied to select an appropriate synthesis model. However, we did not perform model and selection and validation given obvious choice of the likelihood with the simplicity of the data in this simulation. With the binomial likelihood and prior $\text{Beta}(\alpha, \beta)$ on π, the posterior distribution of π given $\mathbf{x}$ is $f(\pi|\mathbf{x}) = \text{Beta}(\alpha + n_1, \beta + n - n_1)$ where $n_1 = \#\{x_i = 1\}$. We set $\alpha = \beta = 1/3$ (Kerman, 2011). In the MODIPS approach, we first located the Bayesian sufficient statistics $\mathbf{s}$ associated with the posterior distribution $f(\pi|\mathbf{x})$, which is n_1 with global sensitivity being 1. The Laplace mechanism was then employed to obtain $n_1^* = n_1 + e$, where $e \sim \text{Lap}(0, \epsilon^{-1})$. Finally, we sampled π^* from $f(\pi^*|n_1^*) = \text{Beta}(\alpha + n_1^*, \beta + n - n_1^*)$, and $\tilde{x}_i$ from $f(\tilde{x}_i|\pi^*) = \text{Bern}(\pi^*)$ for $i = 1, \ldots, n$ to generate one set of synthetic data. The cycle was repeated 5 times (from sanitizing n_1 to drawing $\tilde{\mathbf{x}}$) to obtain 5 sets of synthetic binary data. The non-DP MS approach generated synthetic data in a similar manner to the MODIPS approach except that there was no perturbation of n_1 and π was sampled directly from $f(\pi|\mathbf{x}) = \text{Beta}(\alpha + n_1, \beta + n - n_1)$, and then $\tilde{x}_i$ was sampled from $f(\tilde{x}_i|\pi) = \text{Bern}(\pi)$. In the Laplace sanitizer, five sets of sanitized binary data were directly generated per $n_1^* = n_1 + e$ without any distributional assumption or model fitting.

In both the Laplace sanitizer and the MODIPS, the sanitized n^* could be out of bounds $[0, n]$ as the Laplace noise is drawn from the real line. To legitimize n_1^*, we applied truncation (out-of-bounds n_1^* is thrown away and only in-bounds values are kept), and the boundary inflated truncation (setting $n_1^* < 0$ values at 0 and those $> n$ at n). Neither post-hoc processing procedures compromise DP as no new information is leaked from the original data (sample size n is assumed to be insensitive information and can be released) (Liu, 2019b). Bounding n^* at $[0, n]$ in the MODIPS implies that $\alpha^* = \alpha + n_1^* \geq 0$ and $\beta^* = \beta + n - n_1^* \geq \beta$ in $f(\pi^*|n_1^*) = \text{Beta}(\alpha + n_1^*, \beta + n - n_1^*)$. A reviewer suggested bounding n^* at $[-\alpha, \beta + n]$, thus $\alpha^* \geq 0$ and $\beta^* \geq 0$ and a wider range of π^* can be sampled. We compared this truncation scheme with the above two in this simulation and no significant differences were found.

Both the Binomial-Beta McClure–Reiter synthesizer and the Multinomial-Dirichlet synthesizer simulated data $\tilde{\mathbf{x}}$ from $\text{Bern}(p^*)$; however, p^* was fixed at $\frac{n_1 + \alpha^*}{n + \alpha^* + \beta^*}$ with $\alpha^* = \beta^* = (e^{\epsilon/n} - 1)^{-1}$ for the Binomial-Beta McClure–Reiter synthesizer, and was drawn from $f(p^*|\alpha^*, \beta^*) = \text{Beta}(\alpha^* + n_1, \beta^* + n - n_1)$ for the Multinomial-Dirichlet synthesizer with $\alpha^* = \beta^* = n/(e^\epsilon - 1)$. In the Binomial-Beta McClure–Reiter sanitizer, a single synthetic set was released. In the Multinomial-Dirichlet synthesizer, five synthetic sets were generated, one per each sanitized p^*.

To obtain inferences on π from the released data, each of the 5 sets was analyzed separately in all the above synthesis approach except for the Binomial-Beta McClure–Reiter approach. The point estimate of π in the lth ($l = 1, \ldots, 5$) synthetic data set was the sample proportion $\hat{p}^{(l)}$, and its variance was estimated as $v^{(l)} = \hat{p}^{(l)}(1 - \hat{p}^{(l)})n^{-1}$. Equations (3.5) to (3.7) were then applied to

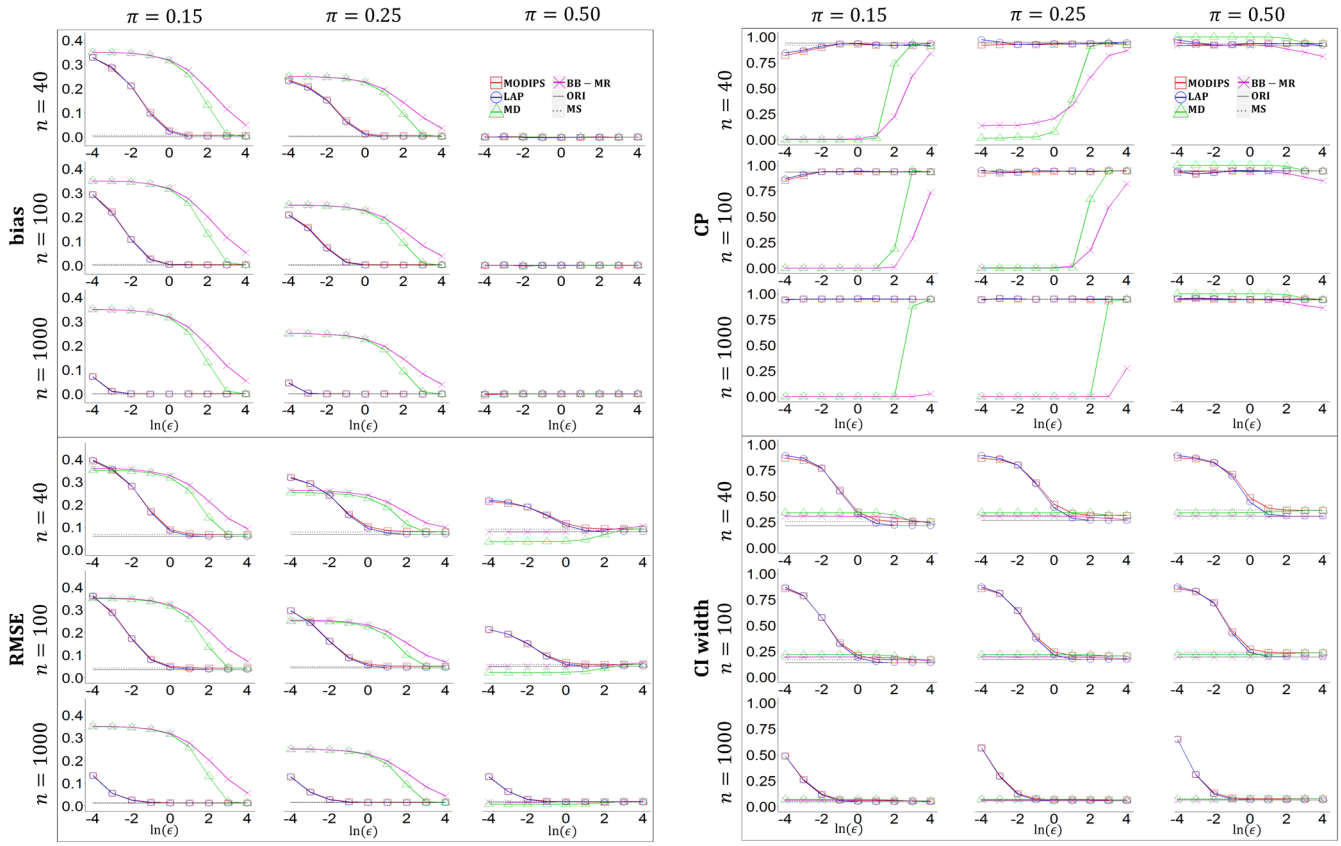

FIG. 2. *The bias, RMSE, 95% coverage probability (CP) and 95% confidence interval (CI) width of π in simulation study* 1. *MODIPS represents the model-based differentially private synthesis, LAP represents the Laplace sanitizer, MD represents the Multinomial-Dirichlet synthesizer, BB-MR represents Binomial-Beta McClure–Reiter synthesizer, Ori is the original results without any perturbation and MS is the traditional multiple synthesis method without DP.*

obtain a final estimate of $\hat{p}$ and the associated 95% confidence interval (CI). Figure 2 depicts the results on the bias and root mean squared error (RMSE), CI width and the coverage probabilities (CPs) of the 95% CIs for π from each DIPS approach, the non-DP MS approach and the original data (we present only the results from the boundary inflated truncation post-processing, which were better than the results from the truncation approach).

The results are summarized as follows. (1) The overall performances of the MODIPS and the Laplace synthesizer are similar while those of the Multinomial-Dirichlet and Binomial-Beta McClure–Reiter synthesizers are similar; in general, the inferences in the former two are better than the latter two. (2) There is noticeable bias, large RMSE and some undercoverage especially when $\epsilon < 1$ and n is small across all DIPS approaches. The inferences improve as ϵ increases (more privacy budget, and thus less perturbation), eventually approaching the original results for the Multinomial-Dirichlet and Binomial-Beta McClure–Reiter synthesizers, and approaching the non-DP MS results for the MODIPS and the Laplace sanitizer. (3) In the MODIPS and Laplace sanitizer, the amount of noise remains constant regardless of n; in other words, the noise becomes less significant for larger n. In the Multinomial-Dirichlet and the Binomial-Beta McClure–Reiter synthesizers, the perturbation introduced through the prior information increases monotonically with n. As a result, there is little improvement in inferences with larger n, which is a significant drawback for the Multinomial-Dirichlet and the Binomial-Beta McClure–Reiter synthesizers. (4) Since the prior mean of π is 0.5 for the Multinomial-Dirichlet and the Binomial-Beta McClure–Reiter synthesizers, when the sample proportion is not 0.5, the two sanitizers introduce bias into the released data. Therefore, the inferences are the best when $\pi = 0.5$ for the Multinomial-Dirichlet and Binomial-Beta McClure–Reiter synthesizers given the consistency between the prior information and the data. (5) For the Laplace sanitizer and the MODIPS approach, the inferences are also the best when $\pi = 0.5$ since 0.5 is the midpoint for the range of a proportion, truncating at 0 or 1 does not skew the distribution of π as much as when π is close to 0 or 1. (6) The RMSE values from the MODIPS and Laplace sanitizers are much smaller than those from the Binomial-Beta McClure–Reiter and Multinomial-Dirichlet synthesizes for most ϵ values when $\pi \neq 0.5$; when $\pi = 0.5$, the Binomial-Beta McClure–Reiter and Multinomial-Dirichlet synthesizers offer smaller RMSE

values for small ϵ; actually, the values are even smaller than the original RMSE values for small ϵ values and decrease when there is more perturbation (ϵ shrinks). Again, this is due to the consistency of the prior information and the data when $\pi = 0.5$. As ϵ decreases, the prior in the Multinomial-Dirichlet and Binomial-Beta McClure–Reiter priors become more "informative," and inject more "useful" prior information about π that is consistent with the data, leading to smaller RMSE. (7) The MODIPS and Laplace sanitizers produce close-to-nominal coverage (0.95) across all the n and π values, except for some undercoverage at small ϵ and n due the relatively large bias with the truncation at 0 and 1 for sanitized proportions. Eventually all CPs converge to the nominal level as ϵ increases in all the sanitizers except for the Binomial-Beta McClure–Reiter synthesizer. (8) On the other hand, the CIs for the Laplace and the MODIPS sanitizers are much wider when $\epsilon < e^{-1}$ than those from the Binomial-Beta McClure–Reiter and Multinomial-Dirichlet synthesizes.

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 $\mathrm{N}(\mu, \sigma^2)$. We manually truncated the simulated data at bounds $[c_0 = \mu - 3\sigma, c_1 = \mu + 4\sigma]$ around μ to generate bounded data so that global sensitivity for the sample mean and variances are finite and calculable. Since there is minimal probability mass (0.0013) outside the $[\mu - 3\sigma, \mu + 4\sigma]$, the normal assumption is hardly affected with the truncation (note that the bounds we used are asymmetric around the true μ, which is more representative of real life data than symmetric bounds). We also examined symmetric bounds, but present the results in the Supplementary Material (Bowen and Liu, 2020). 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, μ was set to 0 in all scenarios.

With the obvious choice of the likelihood given the simplicity of the data, we did not perform model selection and validation in this simulation. Given prior $f(\mu, \sigma^2) \propto \sigma^{-2}$, the posterior distributions are $f(\sigma^2|\mathbf{x}) = \text{Inv-Gamma}[(n-1)/2, (n-1)S^2/2]$ and $f(\mu|\mathbf{x}, \sigma^2) = \mathrm{N}(\bar{x}, n^{-1}\sigma^2)$, where $\bar{x}$ and S^2 are the sample mean and variance, respectively. In the non-DP MS, a synthetic set was generated by first drawing σ^2 and μ from their posterior distributions, and then drawing $\tilde{\mathbf{x}}$ from the normal distribution given the drawn μ and σ^2. The process was repeated 5 times to generate 5 sets of synthetic data to release.

The MODIPS procedure started with sanitizing sufficient statistics $\mathbf{s}$ via the Laplace mechanism, which are, in the posterior distribution $f(\mu, \sigma^2|\mathbf{x})$, $\mathbf{s} = (\bar{x}, S^2)$. To calculate the global sensitivity for $\bar{x}$ and S^2, we needed the global bounds of X. We assumed the bounds of the data were publicly known knowledge, which is a realistic assumption in general as it is very likely an attribute in a data set was never studied previously; there the global bounds on the attribute values are known (e.g., human height, income or published biomarkers, etc). A reviewer questioned the possible conservativeness of the bounds. If the bounds are conservative given the local data, then it is not a concern as DP protects against the worst case scenario and the global bounds are what is needed instead of the local data. If the bounds are conservative at the global level, this implies there is insufficient information on the attribute. In this case, it would be better to be conservative rather than not from a privacy protection perspective though it means more than necessary noise is injected. Future studies are expected to help gain more understanding on the attribute and tighten the bounds. Note that using the local bounds directly would violate privacy even if one is willing to spend some privacy budget to perturb the bounds before using them. However, how to perturb the minimum and maximum can be difficult without knowing the global bounds in the first place.

The global sensitivity is $(c_1 - c_0)n^{-1}$ for $\bar{x}$ and $(c_1 - c_0)^2 n^{-1}$ for S^2, where $(c_1 - c_0) = 7\sigma$ (Liu, 2016). Since the data are bounded, so are $\bar{x}$ and S^2. Specifically, the bounds for $\bar{x}$ are $[c_0, c_1]$, and those of S^2 are $[0, (c_1 - c_0)^2/4 \cdot n/(n-1)]$ (Macleod and Henderson, 1984). If a sanitized statistic was outside its range, it was post-processed by the boundary inflated truncation procedure. Given the sanitized $\mathbf{s}^* = \{\bar{x}^*, S^{2*}\}$, the MODIPS technique drew σ^{2*} from Inv-Gamma$[(n-1)/2, (n-1)S^{2*}/2]$ and μ^* from $\mathrm{N}(\bar{x}^*, n^{-1}\sigma^{2*})$. Finally, $\tilde{x}_i^*$ was simulated from $\mathrm{N}(\mu^*, \sigma^{2*})$ for $i = 1, \ldots, n$ to generate one synthetic set. 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 are two statistics, $(\bar{x}, S^2)$, to sanitize over the same set of data, the $\epsilon/5$ budget per synthesis was further split in half between the sanitization of $\bar{x}$ and S^2.

In deciding the number of bins for the histograms for the perturbed and smoothed histogram approaches, we applied Scott's rule after comparing it with the Sturges' rule and the Freedman–Diaconis rule (Scott, 2015). Specifically, the bin width is set at $\hat{h} = 3.5Sn^{-1/3}$, where S is the sample standard deviation of $\mathbf{x}$ and n is the sample size. The median number of bins is 7, 10 and 21 for $n = 20$, 100 and 1000, respectively, across all simulations (Table 1 in the Supplementary Material). In the perturbed histogram, all bin counts were perturbed via the Laplace mechanism with $\Delta_{\mathbf{s}} = 1$ to obtain the perturbed density histogram (equation (3.1)). The procedure was repeated 5 times to obtain 5 sets of differentially private $\hat{\mathbf{p}}^{(l)}$ (the perturbed bin counts), based on the 5 sets of synthetic data

that were simulated. For the smoothed histogram, we first calculated λ for a given ϵ using equation (3.3), and then constructed the smoothed histogram by applying equation (3.2), from which a single set of synthetic data was generated and released.

To obtain the inference on μ and σ^2 from the multiple released data sets via the MODIPS, the perturbed histogram sanitizers, and the non-DP MS approach, each synthetic set l was analyzed to obtain point estimates of μ and σ^2, which were $\bar{x}^{(l)}$ (the sample mean) and $s^{2(l)}$ (the sample variance), respectively; the associated within-set variance estimates were $s^{2(l)}/n$ and $(s^{2(l)})^2(2(n-1)^{-1} + \kappa^{(l)}n^{-1})$, respectively, where $\kappa^{(l)}$ was the excess kurtosis in the lth set. Equations (3.5) to (3.7) were then applied to obtain the final estimates and 95% CIs.

Figures 3 and 4 depict the bias, RMSE, 95% CI width and the CP of the 95% CI for μ and σ^2 based on the synthetic data of μ and σ^2 based on the synthetic data via the 3 DIPS approaches, respectively. For the purposes of comparability across different values of σ^2, the bias, RMSE and CI width for σ^2 are scaled by the true σ^2, referred to as the relative bias, scaled RMSE and scaled CI width, respectively.

The results are summarized as follows. (1) For all approaches, there are some noticeable biases and large RMSE at small ϵ for both μ and σ^2, which shrink as ϵ increases and eventually approach the original or the non-DP MS results. Overall, the perturbed histogram seems to offer the best trade-off between bias and variance for the inferences based on the synthetic data. (2) For the MODIPS and the perturbed histogram approaches, the amount of injected noise becomes immaterial and the inferences improve as n increases. In the smoothed histogram, λ in equation (3.3) gets larger and approaches $K/(K+\epsilon)$ as n increases. As a result, increasing n does not help the inferences in the smoothed histogram. (3) The positive bias in μ can be explained by the asymmetric bounds $[\mu - 3\sigma, \mu + 4\sigma]$ of data $\mathbf{x}$ around μ. When sanitized $\bar{x}^*$ or synthesized data are out of bound, they are set at the boundary values per the boundary inflated truncation procedure. Since the left bound $\mu - 3\sigma$ is closer to μ, there are more values at $\mu - 3\sigma$ than at $\mu + 4\sigma$, resulting in overestimation. The observed positive bias in σ^2 is expected due to the randomness introduced via synthesis and sanitization. (4) In terms of RMSE, the histogram-based approaches produce smaller RMSE for μ than the MODIPS for most of n and $\epsilon < 1$, but the situation changes for σ^2 with the smallest RMSE coming from the MODIPS. (5) In terms of CP, the MODIPS produces close-to-nominal level coverage in all examined scenarios for both μ and σ^2 at the cost of wide CIs for $\epsilon < 1$; the perturbed histogram has moderate to mild undercoverage

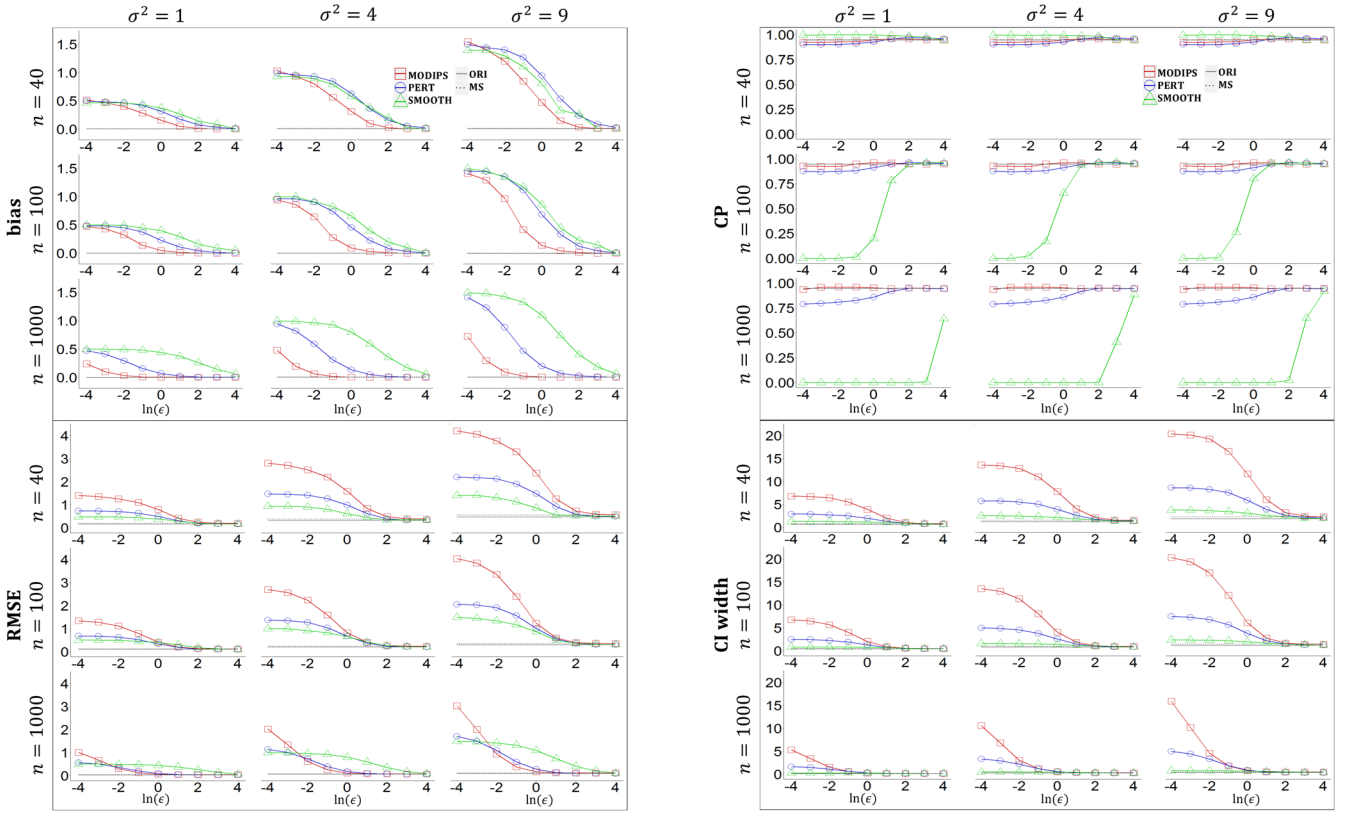

FIG. 3. *The bias, RMSE, 95% CP and 95% CI width of μ in simulation study 2. MODIPS represents the model-based differentially data private synthesis, PERT represents the perturbed histogram method, SMOOTH represents the smoothed histogram method, MS is the traditional multiple synthesis method without DP and Ori is the original results without any perturbation.*

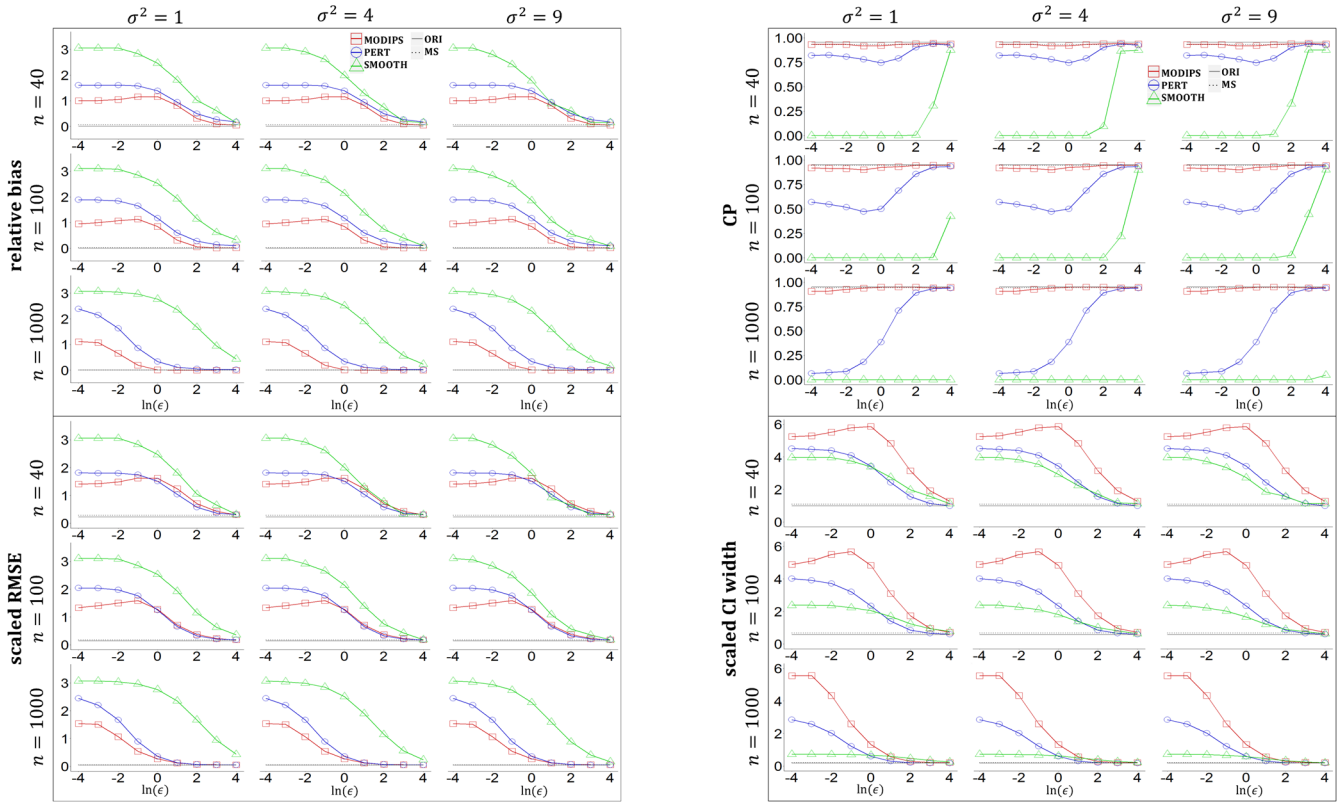

FIG. 4. *The bias, RMSE, 95% CP and 95% CI width of* σ^2 *in simulation study* 2. *MODIPS represents the model-based differentially data private synthesis, PERT represents the perturbed histogram method, SMOOTH represents the smoothed histogram method, MS is the traditional multiple synthesis method without DP and Ori is the original results without any perturbation.*

with much narrower CIs; and the smooth histogram has unacceptable severe undercoverage at small ϵ for large n.

The results when the data bounds $[\mu - 4\sigma, \mu + 4\sigma]$ are symmetric around the true mean are presented in Figures 2 to 5 in the Supplementary Material. As expected, there are minimal biases on μ in all the DIPS approaches (there was some fluctuation in MODIPS for small ϵ), and the CPs in all approaches are at nominal-level. The histogram-based approaches deliver more precise estimates than MODIPS in the inferences of μ (smaller RMSE and narrower CIs). However, the histogram-based approaches do not perform as well as MODIPS in the inferences of σ^2. Both the bias and RMSE are large and there is severe undercoverage at small values of ϵ.

4.3 Simulation Studies 3 and 4: Mixture of Continuous and Categorical Data

In simulation studies 3 and 4, we compare the MODIPS synthesizer and the NP-DIPS synthesizer in data with mixed Gaussian variables $\mathbf{x}$ and categorical variables $\mathbf{w}$. The data were generated from the GLOM (*G*eneral *LO*cation *M*odel) based on $f(\mathbf{x}|\mathbf{w})f(\mathbf{w})$ in simulation 3, and from the SLOMAG model (*S*equential *LO*gistic regression with *MA*rginal *G*aussian distribution) $f(\mathbf{w}|\mathbf{x})f(\mathbf{x})$ in simulation 4. We also investigate the impact of misspecification of the synthesis models. For NP-DIPS in both simulations, we applied the Laplace synthesizer on $\mathbf{w}$ coupled with the perturbed histogram for $\mathbf{x}$. We did not implement the Multinomial-Dirichlet synthesis or the smoothed histogram approach given their inferior performances to the the Laplace sanitizer, the perturbed histogram and the MODIPS based on the results from simulation studies 1 and 2.

4.3.1 *Simulation study* 3: *GLOM model*. Data $\mathbf{x}$ comprise three categorical variables $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$ with 2, 3 and 4 levels, respectively, and continuous variables $\mathbf{z}$. Let n_k denote the count in cell k in the full cross-tabulation of $\mathbf{w}$ for $k = 1, \ldots, 24$. The counts $\mathbf{n} = \{n_k\}$ in the 24 cells were first simulated from a multinomial distribution with parameter $\boldsymbol{\pi} = \{\pi_k\}$ (which are summarized in Table 3 from the Supplementary Material); $\mathbf{z}_{ik} = (z_{ik1}, z_{ik2})'$ was then simulated from $\mathrm{N}_{(2)}(\boldsymbol{\mu}_k, \Sigma)$ for $i = 1, \ldots, n_k$ and $k = 1, \ldots, 24$, where $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})'$ is the mean of $\mathbf{z}$ in cell k, and Σ is the covariance matrix that is the same across all 24 cells. The summary of the parameter values of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\pi}$ across the 24 cells are provided in Table 2 in the Supplementary Material. 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.

Additionally, in the Supplementary Material, Table 3 depicts the summary statistics for the number of observations in the 24 cells across the 5000 repetitions.

For the non-DP MS approach, Given priors $f(\boldsymbol{\pi}) = \text{Dirichlet}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_{24}\} = 1/2$, and $f(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{24}, \Sigma) \propto |\Sigma^{-1}|$. The posterior distributions are $f(\boldsymbol{\pi}|\mathbf{w}) = \text{D}(\boldsymbol{\alpha}')$, $f(\Sigma|\mathbf{z}, \mathbf{w}) = \text{Inv-Wishart}(n - K, \mathbf{S})$ and $f(\boldsymbol{\mu}_k|\Sigma, \mathbf{z}, \mathbf{w}) = \text{N}_{(2)}(\bar{\mathbf{z}}_k, n_k^{-1}\Sigma)$, where $\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \boldsymbol{n}$, $\mathbf{S} = n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathbf{z}_{ik} - \bar{\mathbf{z}}_k)(\mathbf{z}_{ik} - \bar{\mathbf{z}}_k)'$, and $\bar{\mathbf{z}}_k$ contains the sample means of $\mathbf{z}$ in cell k. Synthetic data were simulated from the posterior predictive distribution $f(\tilde{\mathbf{z}}_i, \tilde{\mathbf{w}}_i|\mathbf{z}, \mathbf{w})$ by first drawing $\boldsymbol{\pi} \sim f(\boldsymbol{\pi}|\mathbf{w}) = D(\boldsymbol{\alpha} + \boldsymbol{n})$, Σ from $f(\Sigma|\mathbf{z}, \mathbf{w}) = \text{Inv-Wishart}(n - K, \hat{\Sigma})$ and $\boldsymbol{\mu}_k$ from $f(\boldsymbol{\mu}_k|\Sigma, \mathbf{z}, \mathbf{w}) = \text{N}_{(2)}(\bar{\mathbf{z}}_k, n_k^{-1}\Sigma)$; then sampling $\tilde{\mathbf{w}}$ from $f(\tilde{\mathbf{w}}|\boldsymbol{\pi}) = \text{Multinom}(n, \boldsymbol{\pi})$, and $\tilde{\mathbf{z}}_i$ from $f(\mathbf{z}_i|\tilde{\mathbf{w}}_i, \Sigma) = \text{N}_{(2)}(\boldsymbol{\mu}_k, \Sigma)$ for $i = 1, \ldots, \tilde{n}_k$, where $\tilde{n}_k$ is the count in cell k based on the synthesized $\tilde{\mathbf{w}}$. The drawing process was repeated 5 times to generate 5 synthetic sets.

The Bayesian sufficient statistics from the above Bayesian model are $\mathbf{s} = (\mathbf{n}, S, \bar{\mathbf{z}})$; $\bar{\mathbf{z}}$ contain the 24 pairs of cell means of $\mathbf{z}_1$ and $\mathbf{z}_2$. The MODIPS procedure started with sanitizing $\mathbf{s}$ via the Laplace mechanism to obtain $\mathbf{s}^* = (\mathbf{n}^*, S^*, \bar{\mathbf{z}}^*)$ (the l_1 global sensitivity was 1 for $\mathbf{n}$, $(c_{1,kj} - c_{0,kj})n_k^{-1}$ for $\bar{z}_{kj}$ and $(c_{1,kj} - c_{0,kj})^2(n - 1)(n(n - K))^{-1}$ for each entry in S (Liu, 2019b), 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 $\boldsymbol{\pi}^*$ from $f(\boldsymbol{\pi}^*|\mathbf{n}^*) = D(\boldsymbol{\alpha} + \mathbf{n}^*)$, $\tilde{\mathbf{w}}^*$ from $f(\tilde{\mathbf{w}}^*|\boldsymbol{\pi}^*) = \text{Multinom}(n, \boldsymbol{\pi}^*)$, Σ^* from $f(\Sigma^*|\mathbf{S}^*) = \text{Inv-Wishart}(n - K, \mathbf{S}^*)$, $\boldsymbol{\mu}_k^*$ from $f(\boldsymbol{\mu}_k^*|\Sigma^*, \bar{\mathbf{z}}^*, \mathbf{w}) = \text{N}(\bar{\mathbf{z}}_k^*, n_k^{-1}\Sigma^*)$; and then $\tilde{\mathbf{z}}_i$ was simulated from $f(\mathbf{z}_i|\boldsymbol{\mu}_{\tilde{k}}^*, \Sigma^*) = \text{N}(\boldsymbol{\mu}_{\tilde{k}}^*, \Sigma^*)$ for $i = 1, \ldots, \tilde{n}_k^*$ to generate one set of surrogate data, where $\tilde{n}_k^*$ is 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 synthetic sets with $\epsilon/5$ privacy budget each. Since $\mathbf{s}$ contains 6 components: $\mathbf{n}$, $\bar{\mathbf{z}}_1$, $\bar{\mathbf{z}}_2$, two variance terms and one covariance term from S, each received $1/6$ of $\epsilon/5$ budget allocated to each synthesis (there is no need to split $\epsilon/30$ further among the 24 elements in $\mathbf{n}$ per the parallel composition as they are calculated over nonoverlapping subsets; similar for $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$, respectively).

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 the perturbed histogram method to sanitize continuous $\mathbf{z}$ within each of the 24 cells. Since $\mathbf{z}$ is 2-dimensional, each bin of the histogram of $\mathbf{z}$ is a square rather than an interval. The number of bins were determined using the Scott's rule, and the medians range from 16 to 49 across the 5000 repeats in the 24 cells (Supplementary Material Table 4). 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 examine the inferences on $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, σ_1^2, σ_2^2, ρ and probabilities $\boldsymbol{\Pi} = \{\Pr(w_1 = 1), \Pr(w_2 = 1), \Pr(w_2 = 2), \Pr(w_3 = 1), \Pr(w_3 = 2), \Pr(w_3 = 3)\}$ based on the synthetic data sets. In each synthetic set l $(l = 1, \ldots, 5)$, $\boldsymbol{\Pi}$ was estimated by the corresponding sample marginal probability $\hat{\mathbf{P}}^{(l)}$; $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ were estimated by the sample cell means $\bar{\mathbf{z}}_{1,l}$ and $\bar{\mathbf{z}}_{2,l}$; and Σ was estimated by the pooled variance-covariance $S^{(l)}$. The within-set variance was estimated by $\hat{\mathbf{P}}^{(l)}(1 - \hat{\mathbf{P}}^{(l)})n^{-1}$ for $\hat{\mathbf{P}}^{(l)}$, $S_{j,l}^2 n^{-1}$ for $\bar{\mathbf{z}}_j$ $(j = 1, 2)$, $(S_{j,l}^2)^2(2(n - 1)^{-1} + \kappa^{(l)}n^{-1})$ for $S_{k,l}^2$ and $(1 - r^{2(l)})(n - 2)^{-1}$ for the correlation between Z_1 and Z_2, respectively, where $S_{1,l}^2$ and $S_{2,l}^2$ are the diagonal elements of $S^{(l)}$, $\kappa^{(l)}$ is the excess kurtosis and $r^{(l)}$ is derived from $S^{(l)}$. Equations (3.5) to (3.7) were 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 $\boldsymbol{\mu}_1$ and $\boldsymbol{\Pi}$. The results on $\boldsymbol{\mu}_2$, σ_1^2, σ_2^2, ρ are provided in Figures 6 and 7 in the Supplementary Material. The results are summarized as follows. (1) NP-DIPS performs better than MODIPS for the inferences of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Pi}$ with smaller bias, similar or smaller RMSE, closer-to-nominal coverage and slightly narrower CIs for $\epsilon > e^{-2}$. (2) On the other hand, the MODIPS outperforms the NP-DIPS approach for the inferences on σ_1^2, σ_2^2, ρ with much smaller bias and RMSE for $\epsilon > e^{-1}$ and delivers nominal CP with reasonable CI widths for $\epsilon > 1$; the NP-DIPS approach experiences severe undercoverage in all 3 variance/covariance components and never reaches the nominal level of 95% at all level of ϵ. The severe undercoverage in the NP-DIPS at large ϵ (where the injected noise is supposed to be small) is due to the discretization in forming the histogram bins. The performance of the MODIPS is based on the correct specification of the synthesis model (the GLOM). Misspecification of the synthesis model is expected to lead to worse results, which we will explore in simulation study 4.

4.3.2 *Simulation study* 4: *SLOMAG model*. In this simulation, we first simulated $\mathbf{z}$ from the bivariate normal distribution $f(\mathbf{Z}) = \text{N}_{(2)}(\boldsymbol{\mu}, \Sigma)$ and then generated the categorical variables $\mathbf{w}$ from a sequence of logistic regression models. We set $\boldsymbol{\mu} = (\mu_1, \mu_2)' = \mathbf{0}$, the variances of Z_1 and Z_2 at $\sigma_1^2 = \sigma_2^2 = 1$ and their correlation at $\rho = 0.50$. Z_j $(j = 1, 2)$ was truncated at $[c_{0j} = \mu_j - 4\sigma_j, c_{1j} = \mu_j + 4\sigma_j]$ to generate bounded data. $\mathbf{w}$ contains 3 categorical variables W_1, W_2, W_3 with 2, 2 and 3 levels, respectively, and was generated from $W_1|Z_1, Z_2 \sim \text{Bern}(\pi_1)$ with $\pi_1 = e^{(1,Z_1,Z_2)\boldsymbol{\beta}_1}(1 + e^{(1,Z_1,Z_2)\boldsymbol{\beta}_1})^{-1}$, $W_2|Z_1, Z_2, W_1 \sim \text{Bern}(\pi_2)$ with $\pi_2 = e^{(1,Z_1,Z_2,W_1)\boldsymbol{\beta}_2}(1 + e^{(1,Z_1,Z_2,W_1)\boldsymbol{\beta}_2})^{-1}$ and $W_3|Z_1, Z_2, W_1, W_2 \sim \text{Multinom}(1, (\pi_{31}, \pi_{32}, \pi_{33}))$, where $\pi_{31} =$

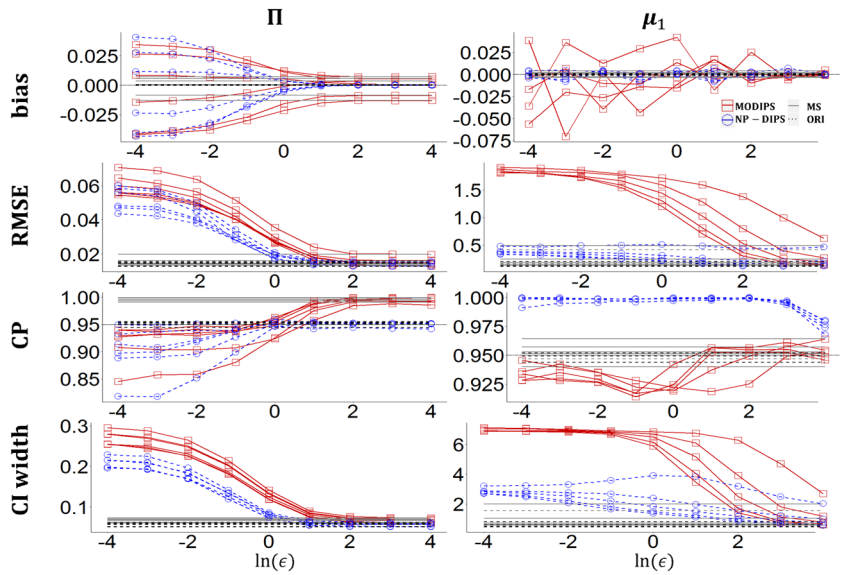

FIG. 5. *The bias, RMSE, CP and 95% CI width of* $\mathbf{\Pi}$ *and* $\boldsymbol{\mu}_1$ *in simulation study* 3. *In the plot of* $\mathbf{\Pi}$, *each line presents a different marginal probability. In the plot of* $\boldsymbol{\mu}_1$, *the lines represent the min, med, max, Q*1, *Q*3 *of the true* 24 *cell means, respectively. 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.*

$(1+A+B)^{-1}$, $\pi_{32} = A(1+A+B)^{-1}$, $\pi_{33} = B(1+A+B)^{-1}$, $A = e^{(1,Z_1,Z_2,W_1,W_2)\boldsymbol{\beta}_3}$, $B = e^{(1,Z_1,Z_2,W_1,W_2)\boldsymbol{\beta}_4}$; and $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11}, \beta_{21})' = (-1, 0.5, -1)'$, $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12}, \beta_{22}, \beta_{32})' = (-2, -1, 1.5, 0.5)'$, $\boldsymbol{\beta}_3 = (\beta_{03}, \beta_{13}, \beta_{23}, \beta_{33}, \beta_{43})' = (0, -2.5, 1, 0.5, 0.4)'$ and $\boldsymbol{\beta}_4 = (\beta_{04}, \beta_{14}, \beta_{24}, \beta_{34}, \beta_{44})' = (0.1, -1, -0.5, 0, 1.5)'$. We ran 1000 repetitions, each sized at $n = 1000$.

The implementation of the NP-DIPS approach is straightforward. $\mathbf{z}$ was first discretized via the Scott's rule to form a 2-way histogram (Table 6 presents the number of histogram bins in the Supplementary Material), which was then combined with $(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$ to form a 5-way cross-tabulation. The counts from which were sanitized via the Laplace sanitizer with global sensitivity is 1, from which 5 sets of synthetic data were generated.

The non-DP MS and the MODIPS methods rely on specifying a synthesis model. In the case of the non-DP MS, an appropriate model can be identified without having to worry about privacy costs. For MODIPS, if the identification of a suitable model is based on previous knowledge and common practice, then no privacy is needed to be spent; however, if the model selection procedure is based on the the data to be released, then the data curator will have to allocate a certain portion of the total privacy budget to the model selection procedure. Differentially private model selection is a separate research topic that is beyond the scope of this paper. For simplicity, we assume the correct SLOMAG model is identified beforehand without using the current data set.

In the SLOMAG model, we employed prior $f(\boldsymbol{\mu}, \Sigma) \propto |\Sigma^{-1}|$ and assumed $f(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4) = f(\boldsymbol{\beta}_1) f(\boldsymbol{\beta}_2) \times f(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$. The joint posterior distribution of the parameters can be factorized as $f(\Sigma|\mathbf{z}) f(\boldsymbol{\mu}|\Sigma, \mathbf{z}) f(\boldsymbol{\beta}_1|\mathbf{z}, \mathbf{w}_1) \times f(\boldsymbol{\beta}_2|\mathbf{z}, \mathbf{w}_1, \mathbf{w}_2) f(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4|\mathbf{z}, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$, where $f(\Sigma|\mathbf{z}) = \text{Inv-Wishart}(n, \boldsymbol{S})$, $f(\boldsymbol{\mu}|\Sigma, \mathbf{z}) = \text{N}(\bar{\mathbf{z}}, n^{-1}\Sigma)$ ($\bar{\mathbf{z}}$ contains the sample means of $\mathbf{z}$ and $\boldsymbol{S} = n^{-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})'$ is the sample covariance matrix of $\mathbf{z}$) and

$$f(\boldsymbol{\beta}_1|\mathbf{w}_1, \mathbf{z}) \propto f(\boldsymbol{\beta}_1) \prod_{i=1}^{n} \frac{e^{w_{i1}(1,z_{i1},z_{i2})\boldsymbol{\beta}_1}}{1 + e^{(1,z_{i1},z_{i2})\boldsymbol{\beta}_1}}, \tag{4.1}$$

$$f(\boldsymbol{\beta}_2|\mathbf{w}_2, \mathbf{w}_1, \mathbf{z}) \propto f(\boldsymbol{\beta}_2) \prod_{i=1}^{n} \frac{e^{w_{i2}(1,z_{i1},z_{i2},w_{i1})\boldsymbol{\beta}_2}}{1 + e^{(1,z_{i1},z_{i2},w_{i1})\boldsymbol{\beta}_2}}, \tag{4.2}$$

$$\begin{aligned} &f(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4|\mathbf{w}_3, \mathbf{w}_2, \mathbf{w}_1, \mathbf{z}) \\ &\quad \propto \prod_{i=1}^{n} \left\{ \left(\frac{1}{1 + A_i + B_i} \right)^{I(w_{i3}=1)} \right. \\ &\qquad \times \left(\frac{e^{w_{i3}(1,z_{i1},z_{i2},w_{i1},w_{i2})\boldsymbol{\beta}_3}}{1 + A_i + B_i} \right)^{I(w_{i3}=2)} \\ &\qquad \left. \times \left(\frac{e^{w_{i3}(1,z_{i1},z_{i2},w_{i3},w_{i2})\boldsymbol{\beta}_4}}{1 + A_i + B_i} \right)^{I(w_{i3}=3)} \right\} \\ &\qquad \times f(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4) \\ &\quad = \frac{e^{a_i\boldsymbol{\beta}_3 + b_i\boldsymbol{\beta}_4}}{\prod_{i=1}^{n}(1 + A_i + B_i)} f(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4), \end{aligned} \tag{4.3}$$

where $a_i = (\sum_{i=1}^{n} I(w_{i3} = 2), \sum_{i=1}^{n} z_{i1} I(w_{i3} = 2), \sum_{i=1}^{n} z_{i2} I(w_{i3} = 2), \sum_{i=1}^{n} w_{i1} I(w_{i3} = 2), \sum_{i=1}^{n} w_{i2} \times I(w_{i3} = 2))$, $b_i = (\sum_{i=1}^{n} I(w_{i3} = 3), \sum_{i=1}^{n} z_{i1} \times I(w_{i3} = 3), \sum_{i=1}^{n} z_{i2} I(w_{i3} = 3), \sum_{i=1}^{n} w_{i1} I(w_{i3} = 3), \sum_{i=1}^{n} w_{i2} I(w_{i3} = 3))$, $A_i = e^{(1,z_{i1},z_{i2},w_{i1},w_{i2})\boldsymbol{\beta}_3}$ and $B_i = e^{(1,z_{i1},z_{i2},w_{i1},w_{i2})\boldsymbol{\beta}_4}$.

To synthesize $\tilde{\mathbf{z}}_i$ for $i = 1, \ldots, n$ in the traditional non-DP MS approach via the SLOMAG model, we first drew Σ from $f(\Sigma|\mathbf{z}) = \text{Inv-Wishart}(n, \boldsymbol{S})$, and $\boldsymbol{\mu}$ from $f(\boldsymbol{\mu}|\Sigma, \mathbf{z}) = \text{N}(\bar{\mathbf{z}}, n^{-1}\Sigma)$ and then simulated $\tilde{\mathbf{z}}_i$ from $f(\tilde{\mathbf{z}}_i|\boldsymbol{\mu}, \Sigma) = \text{N}(\boldsymbol{\mu}, \Sigma)$ given the drawn $(\Sigma, \boldsymbol{\mu})$. To synthesize $\tilde{\mathbf{w}}_i = (\tilde{w}_{i1}, \tilde{w}_{i2}, \tilde{w}_{i3})$, we assumed $f(\boldsymbol{\beta}_1)f(\boldsymbol{\beta}_1) \times f(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4) \propto$ constant and applied the Metropolis algorithm to sample $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_4$ from their posterior distributions and, after checking on the convergence of the MCMC chains (2 chains, a burn-in period of 1500, a thinning period of 10 and 10,000 iterations to yield a total of 7650 samples), simulated $\tilde{w}_{i1}$, $\tilde{w}_{i2}$ and $\tilde{w}_{i3}$ from $f(\tilde{w}_{i1}|\boldsymbol{\beta}_1, \Sigma, \tilde{\mathbf{z}})$, $f(\tilde{w}_{i2}|\boldsymbol{\beta}_2, \tilde{w}_{1i}, \tilde{\mathbf{z}}_i)$ and $f(\tilde{w}_{i3}|\boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \tilde{w}_{i2}, \tilde{w}_{i1}, \tilde{\mathbf{z}}_i)$, respectively. We calculated the potential scale reduction factor (psrf) using the `R` package `coda` to check on the convergence of the MCMC chains. In the Supplementary Material, we provide the MCMC trace plots from a random sample out of the 1000 repeats on $\boldsymbol{\beta}_1$ as an example.

For the MODIPS approach, there are total 8 sets of quantities to be sanitized: $\bar{\mathbf{z}}$, $\mathbf{S}$ and 3 sets of estimated regression coefficients $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$ and $(\hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\beta}}_4)$, implying the total privacy budget per synthesis $(\epsilon/5)$ should be divided by 8. $\bar{\mathbf{z}}$ (2 components) and $\mathbf{S}$ (3 components) are the Bayesian sufficient statistics associated with the posterior distributions of Σ and $\boldsymbol{\mu}$. Since $c_{1,j} - c_{0,j} = 8\sigma$ for $j = 1, 2$, the l_1 global sensitivity is $8\sigma n^{-1}$ for $\bar{z}_j$ and $(8\sigma)^2 n^{-1}$ for each entry in $\mathbf{S}$. To obtain differentially private samples from the posterior distributions of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ in equation (4.1) to (4.3), we implemented Algorithm 1 in Chaudhuri and Monteleoni (2009). Specifically, denote by $\hat{\boldsymbol{\beta}}$ the optimizer of the loss function from a logistic regression model with l_2 regularization (tuning parameter λ) and normalized predictors $\mathbf{x}$ for all $i = 1, \ldots, n$ (per Euclidean norm $\|\mathbf{x}\|_i \leq 1$), then the differentially private coefficient estimates are given by $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} + \mathbf{e}$, where the distribution of the noises $\mathbf{e}$ is $f(\mathbf{e}) \propto \exp(-n\lambda\epsilon\|\mathbf{e}\|/2)$. In this simulation study, we added the noise $\mathbf{e}$ to a random draw from the posterior distribution of the $\boldsymbol{\beta}$ instead of to the optimizer (posterior mode in the Bayesian context) as originally targeted in Chaudhuri and Monteleoni (2009) because it can be easily shown that the same global sensitivity $2/(n\lambda)$ is applicable for other values of $\boldsymbol{\beta}$ in addition to the optimizer. λ is often chosen by cross-validation if there is no privacy concern, but costs privacy otherwise. Chaudhuri, Monteleoni and Sarwate (2011) suggested two ways of selecting λ; using a separate public data set, which does not cost privacy budget; or subsetting the data and then apply the Exponential mechanism with the prediction accuracy as the scoring function to choose λ (Algorithm 4 in Chaudhuri, Monteleoni and Sarwate (2011)). Here, we assumed there exists a public data set (which was simulated from the same joint distribution of $\mathbf{X}$ and $\mathbf{W}$ and attempted five different λ values $(0.01, 0.05, 0.1, 0.5, 1)$ in each regression on this public data set. We found λ_1, λ_2, λ_3 around 0.5 performed the best in terms of prediction accuracy, which was used as the final λ's in MODIPS approach.

In summary, the steps for generating the differentially private data from the SLOMAG model are as follow. (1) sanitize $\bar{\mathbf{z}}$ and $\mathbf{S}$, draw $\boldsymbol{\mu}$ and Σ from their posterior distributions with the sanitized $\bar{z}^*$ and $\mathbf{S}^*$, and simulate $\tilde{\mathbf{z}}^*$ from its posterior predictive distribution given $\boldsymbol{\mu}^*$ and Σ^*. (2) Fit the l_2 regularized logistic regression on $\mathbf{w}_1$ in the Bayesian framework with the normalized predictor $\mathbf{z}'$ and prior $f(\boldsymbol{\beta}_1) \overset{\text{ind}}{\sim} \text{N}(0, \lambda_1^{-1})$; draw $\boldsymbol{\beta}_1$ from its posterior distribution and sanitize it as outlined above; simulate $\tilde{\mathbf{w}}_1^*$ given the sanitized $\boldsymbol{\beta}_1^*$ and the normalized sanitized $\tilde{\mathbf{z}}^{*'}$ from the first step. (3) Fit the l_2 regularized logistic regression on $\mathbf{w}_2$ in the Bayesian framework with the normalized predictor $(\mathbf{z}', \mathbf{w}_1')$ and prior $f(\boldsymbol{\beta}_2) \overset{\text{ind}}{\sim} \text{N}(0, \lambda_2^{-1})$; draw $\boldsymbol{\beta}_2$ from its posterior distribution and sanitize it as outlined above; simulate $\tilde{\mathbf{w}}_2^*$ given the sanitized $\boldsymbol{\beta}_2^*$ and the normalized $(\tilde{\mathbf{z}}^{*'}, \tilde{\mathbf{w}}_1^{*'})$ from the first two steps. (4) Fit the l_2 regularized multinomial logistic regression on $\mathbf{w}_3$ in the Bayesian framework with the normalized predictor $(\mathbf{z}', \mathbf{w}_1', \mathbf{w}_2')$ and prior $f(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4) \overset{\text{ind}}{\sim} \text{N}(0, \lambda_3^{-1})$; draw $(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$ from their posterior distributions and sanitize them as outlined above; simulate $\tilde{\mathbf{w}}_3^*$ given the sanitized $(\boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*)$ and the normalized $(\tilde{\mathbf{w}}_1^*, \tilde{\mathbf{w}}_2^*, \tilde{\mathbf{z}}^*)$ from the previous three steps. Similar to the non-DP MS case, we calculated the psrf to check on the convergence of the MCMC chains. In the Supplementary Material, we provide the MCMC trace plots from a random sample out of the 1000 repeats on $\boldsymbol{\beta}_1$ as an example.

To check on the impact of the misspecification of synthesis models in MODIPS on the inferences from the synthetic data, we also synthesized data from a misspecificated model. There are many possible models for a mixture of categorical and continuous variables, for example, assuming independence among $(\mathbf{x}, \mathbf{w})$, dropping a predictor in one of the logistic regression equations above, switching the order of the logistic regression sequence, or applying the GLOM, which are all regarded as mis-specifications in this simulation. For the purposes of checking on the impact of the mis-specification of synthesis models, it makes no essential difference which misspecified model to be compared with the correct specification. Therefore, we used the GLOM as a representation for the model misspecification. The Bayesian GLOM in this case is similar to that in Section 4.3.1, with the multinomial distribution on (W_1, W_2, W_3) (12 cells) and the bivariate Gaussian assumption on (Z_1, Z_2) in each of the 12 cells, to generate 5 synthetic data sets. To examine how bad the misspecification could be in terms of inferences, the Supplementary Material present a posterior predictive

check from fitting the SLOMAG and the GLOM models to the original data. Although misspecification leads to biased synthetic data, it does bring some potential side benefit in terms of privacy protection. Specifically, misspecification could offer some additional privacy guarantee as it provides another type of noise that deviates the synthetic information from the original. Therefore, the comparison in statistical utility below from the synthetic data might be unfair to the GLOM model as the actual privacy it provides might be more than the nominal ϵ-DP, which the other synthesizer guarantees. On the other hand, the additional privacy protection, if there is any, can be difficult to trace or quantify; or it may be unnecessary given that the pre-specified privacy is already guaranteed through DP.

We examine the inferences on Σ, $\boldsymbol{\mu}$ and $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_4$. μ_1 and μ_2 in synthetic data set l ($l = 1, \ldots, 5$) were estimated by the sample means $\bar{z}_1^{(l)}$ and $\bar{z}_2^{(l)}$, and Σ was estimated by the sample covariance $\boldsymbol{S}^{(l)}$. The corresponding within-set variance was estimated by $(S_k^2)^{(l)} n^{-1}$ for $\bar{z}_k^{(l)}$ ($k = 1, 2$), $((S_k^2)^{(l)})^2 (2(n-1)^{-1} + \kappa^{(l)} n^{-1})$ for the marginal variances of $\mathbf{z}_1$ and $\mathbf{z}_2$, and $(1 - (r^{(l)})^2)(n-2)^{-1}$ for the correlation between Z_1 and Z_2, respectively, where $(S_1^2)^{(l)}$, $(S_2^2)^{(l)}$ are the diagonal elements of $\boldsymbol{S}^{(l)}$, $\kappa^{(l)}$ is the excess kurtosis and $r^{(l)}$ is derived from $\mathbf{S}^{(l)}$. The regression coefficients $\boldsymbol{\beta}$ were estimated using the `logistf` function with the Firth's bias reduction method in the `R` package `logistf` along with the corresponding estimated variance estimates. Equations (3.5) to (3.7) 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 $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ in Figures 6 and 7; the results on μ_1, μ_2, σ_1^2, σ_2^2, ρ, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_4$ are available in the Supplementary Material. For very small values of $\epsilon = e^{-4}$ to e^{-2}, the logistic regression based on the synthetic data from the MODIPS approach failed to converge thus the results were not available for plotting. The results are summarized as follows. (1) First, as expected, MODIPS-Wrong fails to capture the original information due to the model misspecification (large bias and undercover coverage). (2) Overall, the biases for MODIPS-Correct get smaller and are close zero for $\epsilon > e^{-1} \approx 0.368$, whereas the biases from the NP-DIPS approach do not seem to diminish even at large ϵ. However, the bias in the MODIPS-Correct method is unstable and larger than the other methods when ϵ is small (not plotted). (3) The RMSE values in general are large compared to the original RMSE values across all DIPS methods. (4) The MODIPS-Correct approach produces coverage at or above the nominal level of 95% for $\epsilon > e^{-1}$ at the cost of wide CIs. The CP results from the NP-DIPS approach vary across parameters: some experience severe undercoverage across all values of ϵ or only at small values of ϵ, some have close to 95% coverage across the board. The CI width varies little with ϵ in the NP-DIPS approach. (5) For μ_1 and μ_2, the bias, RMSE and CI width of the estimates are smaller in the NP-DIPS approach than those in the MODIPS-Correct approach for $\epsilon < e$ and are similar for $\epsilon > e$; and both provide about 95% CP. Although the RMSE and CI width decrease as ϵ increases for MODIPS-Wrong, the bias and CP deviate significantly from the original values. (6) For Σ,

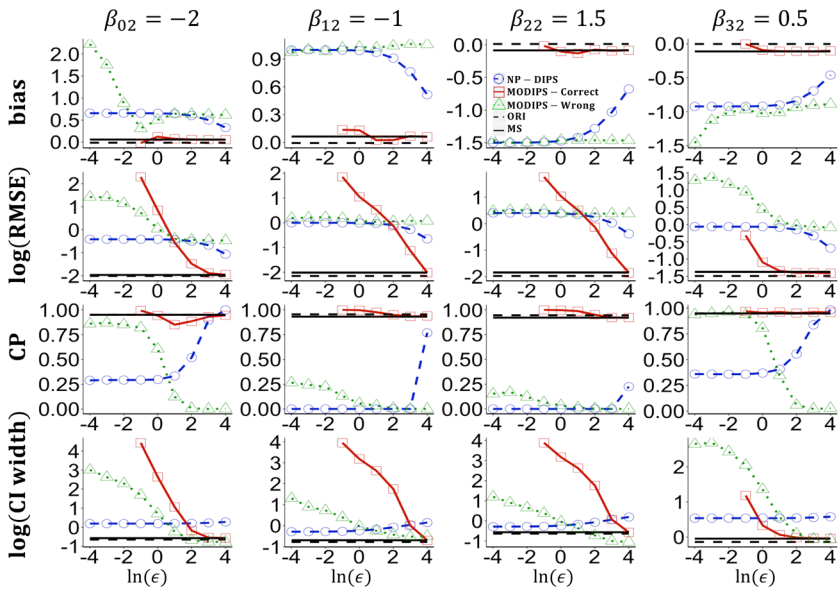

FIG. 6. *The bias, log(RMSE), 95% CP and log(95% CI width) for $\boldsymbol{\beta}_2$ in simulation study 4. MODIPS represents the model-based differentially private synthesis, NP-DIPS represents the Laplace sanitizer + perturbed histogram method, MS is the traditional multiple synthesis method without DP and Ori is the original results without any perturbation.*

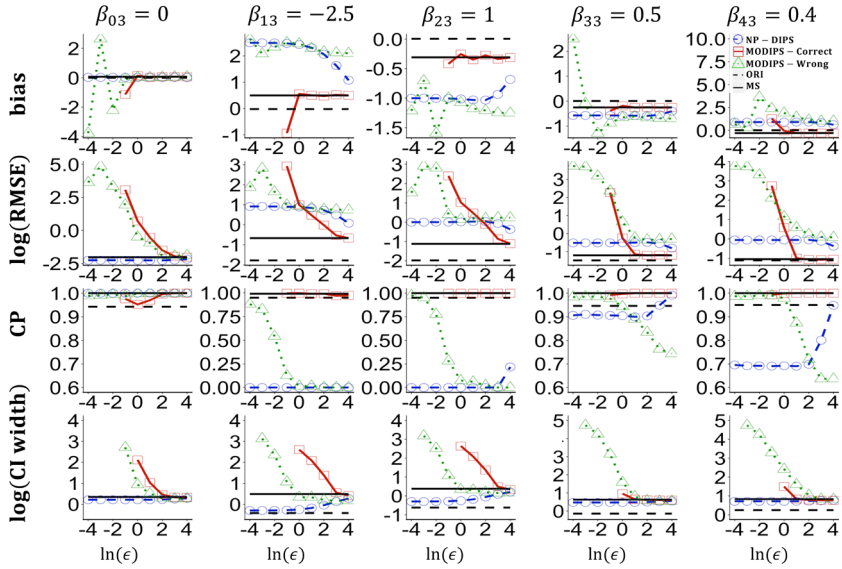

FIG. 7. *The bias, log(RMSE), 95% CP and log(95% CI width) for* $\boldsymbol{\beta}_3$ *in simulation study* 4. *MODIPS represents the model-based differentially private synthesis, NP-DIPS represents the Laplace sanitizer + perturbed histogram method, MS is the traditional multiple synthesis method without DP and Ori is the original results without any perturbation.*

the NP-DIPS and MODIPS–Wrong approaches experience severe undercoverage in all 3 components (σ_1^2, σ_2^2 and ρ) regardless of ϵ. The former sufferes for the same reason given in simulation study 3 (discretization and uniform sampling from each histogram bin). The MODIPS–Correct method provides nominal CP for σ_1^2, σ_2^2 and ρ and has smaller bias and RMSE than the NP-DIPS approach for $\epsilon > e^{-2}$, at the cost of wide CIs for $\epsilon < e$.

Compared to simulation study 3 which also examines a mixture of continuous and categorical variables, the results in simulation study 4 are generally worse for both the NP-DIPS and the MODIPS approaches, but in different ways. The identification of statistics to sanitize with the SLOMAG model is less obvious and the inferences based on the synthetic data are less stable in simulation 4 for the MODIPS approach, probably due to the direct sanitization of the likelihood functions. For the NP-DIPS approach, the discretization and sanitization procedure is the same between simulations 3 and 4, but seems to affect the inferences from the SLOMAG model more than those from the GLOM. The different results from the two simulation studies suggest that even though the NP-DIPS approach is nonparametric, inferences based on the synthesized data in certain models can be more sensitive than others.

5. CASE STUDY

We applies several DIPS approaches to a real-life data set to assess the feasibility these approaches in generating useful synthetic data sets in practice. We used the fertility data set from Gil et al. (2012) in a study of 100 student volunteers at the University of Alicante. Each participant provided a semen sample after 3 to 6 days of sexual abstinence, and answered a questionnaire about their life habits and health status. The attributes in the data are summarized in Table 3 (there are originally 35 variables in Gil et al. (2012), but only 10 variables are publicly available on the UCI Machine Learning Repository). The data set is useful for studying risk factors possibly associated with the fertility rate. On the other hand, sharing the data set publically could cause privacy concerns as some of the variables such as "diagnosis of seminal quality" are generally regarded as sensitive information.

The main goal of Gil et al. (2012) is to compare the performance of three machine learning techniques (decision trees, Multilayer Perception and Support Vector Machine/SVM) in predicting the seminal quality given various predictors. The authors found that Multilayer Perception and SVM outperformed decision trees with SVM

TABLE 3
Variables from the fertility data in Gil et al. (2012)

Variable	Values
Season of the analysis	Winter, Spring, Summer, Fall
Age at the time of analysis (years)	$18 \sim 36$
Childish diseases	Yes, No
Accident or serious trauma	Yes, No
Surgical intervention	Yes, No
High fevers in the last year	<3 months ago, >3 months ago, no
Frequency of alcohol consumption	several times a day, every day, several times a week, once a week, hardly ever or never
Smoking habit	never, occasional, daily
Number of hours sitting per day	$1 \sim 16$
Diagnosis of seminal quality	normal, altered

slightly more accurate. Therefore, we only employed the SVM in this case study. Specifically, we first randomly split the original data into a training set of 80 subjects and a test set of 20 subjects. The same test set was then used to evaluate the predictive power of the SVMs constructed from the synthetic data via different DIPS approaches (so to avoid testing the SVM on a test data that was synthesized via the same DIPS approach for generating the training data). Since this analysis did not involve statistical inferences, we generated a single synthetic data set with $\epsilon = e^1 = 2.72$, a practically small and reasonable privacy budget.

We employed the Laplace sanitizer (ND-DIPS) and the MODIPS (P-DIPS) approach. In the Laplace sanitizer, we first discretized the two continuous variables (age and hours sitting) into a 2D histogram, then sanitized the cell counts $\mathbf{n}$ from the full cross tabulation of the 10 variables with the additive noise from $\text{Lap}(0, \epsilon^{-1})$. For the MODIPS approach, the first step was to select an appropriate synthesis model. There are 8 categorical variables with some of them having sparse cell counts in their marginal distribution (e.g., in alcohol consumption, there is only 1 person in the categories of several times a day and every day, resp.); both of the two continuous variables (age and hours sitting) deviate from Gaussian distributions. Given the small sample size ($n = 80$ in the training set), the GLOM is not expected to work well as there would be too many empty or sparse cells from the full cross-tabulation of the categorical variables; the SLOMAG model could generate noisy synthetic data based on its performance in the simulation study 4 where the sample size ($n = 1000$) is much higher and has a much smaller cross-tabulation of the categorical variables than this case study. We also tried the second-order mixed graphical model approach on the data, and the prediction was not good even without DP perturbation. All taken together, we discretized the two continuous variables and then fitted a saturated log-linear model. The Bayesian sufficient statistics is $\mathbf{n}$. We first sanitized $\mathbf{n}$ via the Laplace mechanism with scale ϵ^{-1} to obtain $\mathbf{n}^*$. Given $\mathbf{s}^*$, we drew $\boldsymbol{\pi}^*$ from $f(\boldsymbol{\pi}^*|\mathbf{n}^*) = D(\boldsymbol{\alpha} + \mathbf{n}^*)$, then $\tilde{\mathbf{x}}$ from $f(\tilde{\mathbf{x}}|\boldsymbol{\pi}^*) = \text{Multinom}(n, \boldsymbol{\pi}^*)$. We ran 100 repetitions.

As a benchmark, we run the PrivateSVM (Algorithm 2 in Rubinstein et al. (2009)), an approach designed specifically for releasing differentially private SVM results. PrivateSVM first applies the SVM to the original data and then returns the noisy weight vector via the Laplace mechanism. Calculating the global sensitivity of the weight vector for PrivateSVM is nontrivial. We employed the linear kernel for PrivateSVM, the global sensitivity based which is $4LC\kappa\sqrt{F}/n$ (Rubinstein et al., 2009), where $L = 1$ is the linear kernel, $C = 1$ is cost of constraints violation (the C-constant of the regularization term in the Lagrange formulation in SVM), κ is the upper bound for the linear kernel (9 with the normalized data in this case), F is the number of features (9 in this case) and n is the number of observations in the training sample (80 in this case). Thus, the total global sensitivity is 1.35.

If the DIPS approaches perform similarly to PrivateSVM in classification accuracy, DIPS would be preferable as data users will have the individual-level synthetic data and can perform their own analyses whereas PrivateSVM only provides a differentially private SVM. When constructing the SVM on the synthetic data, we employed the `svm` command in R package `e1071` with `kernel="linear"` and a 5-fold cross validation.

Table 4 shows the averaged confusion matrices and the classification accuracy on the 20 testing cases over the 100 repeats by the SVMs constructed from PrivateSVM as well as the synthetic data from the Laplace sanitizer and MODIPS approaches. As expected, the prediction accuracy of the SVMs constructed on the synthetic data via the Laplace sanitizer and the MODIPS approach is not as good (64.7% and 50.0%, respectively) as the original SVM (85%) at the privacy budget of $\epsilon = 2.72$, a cost we have to pay to achieve some level of privacy. The Laplace sanitizer is no worse than PrivateSVM (64.9%). The significant decreases (20%) in predictive accuracy from the original results in all 3 differentially private approaches might have something to do with the small sample size and the unbalancedness between the two categories of the

TABLE 4

Accuracy of SVMs constructed from PrivateSVM approach and synthetic data via the Laplace Sanitizer and MODIPS approaches when $\epsilon = e$

Observed	Predicted based on							
	Original data		PrivateSVM		Laplace sanitizer		MODIPS	
	+	−	+	−	+	−	+	−
+	0	3	0.84	2.16	1.04	1.96	1.50	1.50
−	0	17	4.87	12.13	5.11	11.89	8.51	8.49
CR†	**17/20 = 85%**		**12.97/20 = 64.9%**		**12.93/20 = 64.7%**		**9.99/20 = 50.0%**	

†CR: consistency rate

TABLE 5
Some summary statistics on the synthetic data

Statistic	Attribute	Original data	Laplace santizer	MODIPS
Mean (SD)	age	0.67 (0.12)	0.66 (0.14)	0.66 (0.14)
	hours	0.41 (0.19)	0.44 (0.21)	0.50 (0.29)
Total variation distance (TVD)	1-way table	–	0.228	0.250
	2-way table	–	0.353	0.379
	3-way table	–	0.311	0.330
	8-way table	–	0.451	0.483

†TVD $= t^{-1} \sum_{j=1}^{t} |\mathbf{p}_j^* - \mathbf{p}_j|_1$, where t is the number of tables, $\mathbf{p}_j$ and $\mathbf{p}_j^*$ represent the vector of cell probabilities in table j constructed from the original and synthetic data, respectively.

outcome (88:12 normal vs. altered). On the other hand, there might exist more efficient DIPS methods that can better preserve the original info while satisfying DP, a topic we will continue to work on.

We also examine summary statistics to further assess the synthetic data quality. Table 5 summarizes the continuous variables (*Age at the time of analysis* and*Number of hours sitting per day*, which are referred to as *Age* and *Hours*, resp.) by the mean and SD, and the categories variables by the averaged total variation distance in the 1-, 2-way, 3-way and 8-way (full) cross-tabulations, respectively, constructed based on the synthetic versus the original data. In summary, for both the Laplace sanitizer and the MODIPS approach, the mean and SD of the two continuous variables are close to the original data's values; and the averaged total variation distances in all the cross-tabulations are consistently smaller with the Laplace sanitizer than the MODIPS approach.

6. DISCUSSION

We have reviewed various DIPS methods for synthesizing differentially private individual-level data and compared some DIPS methods empirically through simulation studies and a real-life case study on the utility and inferential properties of the synthetic data generated by the DIPS 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.

The NP-DIPS approaches are robust given that they do not impose model or distribution assumptions on a given data set. However, most NP-DIPS approaches require some degree of discretization on numerical attributes. When the number of attributes p is large, an important question to ask is whether there exists a "consistent" high-dimensional histogram density estimator f_n for the underlying true density f for a given sample size n, even before the employment of a DP technique. It is known that the number of bins of a high-dimensional histogram grows exponentially with dimension p, and the rate of decrease of the mean integrated squared error $\mathrm{E}\|(f_n - f)\|_2^2$ degrades rapidly as p increases compared to the ideal parametric rate $O(n)$ (Scott, 2015). In addition, when p increase, most of the hypercube bins in the high-dimensional histogram become empty and the histogram will be rough and provides reasonable estimates only near the mode and away from the tails. When there are correlations among the variables, smaller bin widths are required to "track" the correlations (implying more bins), and the asymptotic mean integrated squared error is always larger than the independent case. In summary, in high-dimensional histograms, the variance and bias tradeoff is not favorable unless n is large. If the original high-dimensional histogram is not already a good estimator for the distribution of the data, it is not meaningful to further sanitize it. Additionally, inferences based on the synthetic data via histogram-based NP-DIPS approaches are affected by how the histogram bins are formed. There exists theoretical work in the computer science community that examines the relationship between the sample size n and the accuracy of p binary proportion. This accuracy is defined as how close the sanitized histogram is to the original and does not involve drawing inferences about population parameters. For example, the average l_1 error of answering p 1-way binary proportions has a lower bound of $\Omega(p/(n\epsilon))$ (Hardt and Talwar, 2010); and the maximum l_1 error from answering p binary proportion given n has a upper bound of $\Omega(p\log(p)/(n\epsilon))$ (Steinke and Ullman, 2017).

The P-DIPS approaches, on the other hand, often require distributional assumptions and model building, and thus are subject to appropriate model misspecification. None of the P-DIPS procedures we have examined (except for PrivBayes in Zhang et al. (2017)) have the inherent model-selection component, implying they are applied after a suitable model is identified. Broadly speaking, there are two model selection scenarios—one costs privacy budget and the other does not. Specifically, if the model is chosen not using the knowledge in the current data, but based on previous studies and common practice, then no privacy needs to be spent. If the synthesis

model is selected via a selection procedure using the data to be released, then we will need to split the privacy budget between model selection and data synthesis. The current research on differentially private model selection focuses on feature selection in the setting of a certain model type, such as Kifer, Smith and Thakurta (2012), Smith and Thakurta (2013), Lei et al. (2018) for linear regression; and Zhang et al. (2017) for Bayesian networks. More research will be needed in differentially privately selecting among models that do not have to be of the same type, maybe by perturbing model selection criteria such as AIC or BIC. Meanwhile, to mitigate the concern on model specification or when there are several plausible models, we incorporated the model averaging idea into the synthetic data generation, which also helps loosening the restriction the dependency of the synthetic data on a single synthesis model.

An obvious drawback for all DIPS approaches is that the data user will not know how much the results based on the synthetic data deviate from those if they had access to the original data. If the differentially private synthetic data contain too much noise, the decisions made based on the analysis of the synthetic might be improper or wrong. Barrientos et al. (2019) proposed a differentially privately mechanism to release the test statistic and p-value from testing a regression coefficient against 0 from a linear regression model. Their numerical results suggest the sign of the test statistic and the conclusion of the hypothesis test have a high probability of being consistent with the original results. The authors also proposed that the approach can be used to validate the linear regression analysis based on synthetic data from a DIPS method. The validation system hinted in Barrientos et al. (2019) is developed by Barrientos et al. (2018) in a more comprehensive and integrated fashion using the U.S. federal government employee longitudinal data as an example. Specifically, the system has three components: (1) release synthetic data generated from a joint distribution of the data; (2) verif/validate the statistical utility of a certain analysis (query) by comparing the results based on the synthetic data with the query result released by a differentially private mechanism and (3) provide the raw/confidential data to approved data users via secure remote access. In the examples given in both papers, all the given privacy budget is spent on testing or verifying a single query result, while the reality is that a data user is often interested in estimating more than parameters. From a DP perspective, the total privacy budget will have to be split among all queries to be verified, leading to potentially a large amount noised injected per query and diminishing the value of a validation system. In addition, the synthesis in the validation system mentioned in Barrientos et al. (2018) does not have to be differentially private (and is not in the example given in the paper). The validation system needs another disclosure risk assessment step on the released synthetic data, which relies on strong assumption and can be ad-hoc as compared to the robustness of the DP concept. In addition, since DP aims to cover the worst case scenario, this means the statistical utility of the query result obtained via a differentially private mechanism can be further away from the original than the inferences based on the synthetic data without DP. In other words, a significant discrepancy between the two as suggested in Barrientos et al. (2018) does not necessarily invalidate one or the other.

An alternative to enhancing data users' confidence in synthetic data is to develop more efficient DP mechanisms at the same privacy cost, but with less noise injected, such as taking into account the correlations among the statistics during sanitization so that the privacy budget is not spent on overlapping information, or optimizing the privacy budget allocation scheme when the sequential composition is in effect. 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, improving the efficiency of the DIPS procedures. In addition, we could always employ a relaxed version of DP (such as aDP or pDP) to generate synthetic data as long as there is consensus the relaxed DP still provide satisfactory privacy protection. Conceptually, all the DIPS methods introduced and examined in the paper can be implemented with relaxed DP, assuming the appropriate sanitizer is employed.

Some future work could also involve developing a system that compares the various DP definitions, mechanisms and algorithms, and recommends DP mechanisms/algorithms to users. Given the wealth of DP methods, a data user might face difficulty in selecting the most well-suited DP approach for his/her data, including considerations on the practicality and computational limitations of those DP methods. Hay et al. (2016) attempted to address the issue proposing DPBench as approach for standardized evaluation of privacy algorithm, as well as valuable observations and findings after comparing various data-dependent and data-independent DP methods. However, their work is limited to 1- and 2-dimensional range queries. Motivated by DPBench, Kotsogiannis et al. (2017) developed Pythia, a meta-algorithm that measures the input features to select a particular DP method. Similarly, Pythia is limited to releasing certain queries such as histograms, range queries and Naive Bayes classifiers.

The choice of ϵ (and the parameter that quantifies the relaxation of the strict ϵ-DP if a relaxed version of DP is used) remains an open question. The concept of the ϵ-DP is abstract and does not easily relate to practically relevant measures of privacy, making the justification of a socially acceptable of ϵ difficult. Based on the literature we have surveyed as well as the observations on the statistical

utility from the simulation studies and the case study we conducted, ϵ in the neighborhood 1 (which is neither too small nor too large) seems to produce synthetic data of acceptable statistical utility. Additionally, $\epsilon = 1$ has been explored frequently in experiments run in literature. We believe more research and further investigation on this issue will help narrow down ϵ to a generally acceptable set of values.

The ultimate goal of developing DIPS approaches is to employ them for public data release in practice. The US Census Bureau aims to employ DP in major data products like the 2020 Census of Population and House, the Economic Census and the annual American Community Surveys (Abowd et al., 2017). On the other hand, real-life data can be large in size, complex in structure and have a large number of attributes of various types. In addition, issues such as missing data, sparse data, data entry errors, among others further complicate the application of DIPS. There is still a huge gap from the research work on DP to the wide practical application of DP. The status quo is that a large body of DP literature focus only on categorical/binary attributes and ignore missing data or data entry errors. Machanavajjhala et al. (2008) demonstrated that the Multinomial-Dirichlet synthesizer led to poor inferences due to data sparsity when releasing the commuting patterns of the US population data and proposed combining distance-based coarsening with a probabilistic pruning algorithm and preserving $(\epsilon = 8.6, \delta = 10^{-5})$-pDP. The relatively low classification accuracy based on synthetic data in our case study in Section 5 also suggests that direct application of a DIPS approach without any modification might not accommodate real-life situation well enough. On the other hand, local DP has been employed by big tech companies (e.g., Google and Apple) to collect users data. Though these applications seem to be successful, multiple sources suggest the privacy budget ϵ employed by Apple to collect users data on mobile devices is too high to be acceptable for privacy protection (Tang et al., 2017, Orr, 2017). Although Apple has provided some information about their DP approach, the information is vague. Tang et al. (2017) attempted to replicate the method without success and stated that "*We applaud Apple's deployment of DP for its bold demonstration of feasibility of innovation while guaranteeing rigorous privacy. However, we argue that in order to claim the full benefits of differentially private data collection, Apple must give full transparency of its implementation and privacy loss choices, enable user choice in areas related to privacy loss and set meaningful defaults on the daily and device lifetime privacy loss permitted.*"

ACKNOWLEDGMENTS

Claire McKay Bowen was supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. DGE-1313583 during part of the development of this paper. Fang Liu is supported by NSF Grants #1546373, #1717417 and the University of Notre Dame Faculty Research Support Initiation Grant Program. 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). The publication has been assigned the Los Alamos National Laboratory identifier LA-UR-18-31132.

We thank the Editor, the Associate Editor and the two referees for their valuable comments and suggestions that improved the quality of the manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "Comparative Study of Differentially Private Data Synthesis Methods" (DOI: 10.1214/19-STS742SUPP; .pdf). This file contains the Supplementary Material to accompany the paper "Comparative Study of Differentially Private Data Synthesis Methods" with additional results from the four simulation studies.

REFERENCES

ABADI, M., CHU, A., GOODFELLOW, I., MCMAHAN, H. B., MIRONOV, I., TALWAR, K. and ZHANG, L. (2016). Deep learning with differential privacy. In *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security* 308–318. ACM, New York.

ABOWD, J. M. and SCHMUTTE, I. M. (2015). Revisiting the economics of privacy: Population statistics and confidentiality protection as public goods. Cornell University ILR School. Available at https://digitalcommons.ilr.cornell.edu/ldi/22/.

ABOWD, J. M., SCHNEIDER, M. J. and VILHUBER, L. (2013). Differential privacy applications to Bayesian and linear mixed model estimation. *J. Priv. Confid.* **5** 4.

ABOWD, J. M. and VILHUBER, L. (2008). How protective are synthetic data? In *Privacy in Statistical Databases* 239–246. Springer, Berlin.

ABOWD, J., ALVISI, L., DWORK, C., KANNAN, S., MACHANAVAJJHALA, A. and REITER, J. (2017). Privacy-preserving data analysis for the Federal Statistical Agencies. Preprint. Available at arXiv:1701.00752.

ACS, G., CASTELLUCCIA, C. and CHEN, R. (2012). Differentially private histogram publishing through lossy compression. In 2012 *IEEE* 12*th International Conference on Data Mining* 1–10. IEEE, Washington, DC.

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* 273–282. ACM, New York.

BARRIENTOS, A. F., BOLTON, A., BALMAT, T., REITER, J. P., DE FIGUEIREDO, J. M., MACHANAVAJJHALA, A., CHEN, Y., KNEIFEL, C. and DELONG, M. (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *Ann. Appl. Stat.* **12** 1124–1156. MR3834297 https://doi.org/10.1214/18-AOAS1194

BARRIENTOS, A. F., REITER, J. P., MACHANAVAJJHALA, A. and CHEN, Y. (2019). Differentially private significance tests for regression coefficients. *J. Comput. Graph. Statist.* **28** 440–453. MR3974892 https://doi.org/10.1080/10618600.2018.1538881

BLUM, A., LIGETT, K. and ROTH, A. (2013). A learning theory approach to noninteractive database privacy. *J. ACM* **60** Art. 12. MR3060810 https://doi.org/10.1145/2450142.2450148

BOWEN, C. M. K. and LIU, F. (2018). Differentially private release and analysis of youth voter registration data via statistical election to partition sequentially. Preprint. Available at arXiv:1803.06763.

BOWEN, C. M. and LIU, F. (2020). Supplement to "Comparative study of differentially private data synthesis methods." https://doi.org/10.1214/19-STS742SUPP.

CHAREST, A. S. (2010). How can we analyze differentially private synthetic datasets. *J. Priv. Confid.* **2** 3.

CHAREST, A.-S. and HOU, Y. (2017). On the meaning and limits of empirical differential privacy. *J. Priv. Confid.* **7** 3.

CHAUDHURI, K. and MONTELEONI, C. (2009). Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems* 289–296.

CHAUDHURI, K., MONTELEONI, C. and SARWATE, A. D. (2011). Differentially private empirical risk minimization. *J. Mach. Learn. Res.* **12** 1069–1109. MR2786918

CHAUDHURI, K., SARWATE, A. and SINHA, K. (2012). Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems* 989–997.

CHEN, R., FUNG, B. C., MOHAMMED, N., DESAI, B. C. and WANG, K. (2013). Privacy-preserving trajectory data publishing by local suppression. *Inform. Sci.* **231** 83–97.

DING, B., WINSLETT, M., HAN, J. and LI, Z. (2011). Differentially private data cubes: Optimizing noise sources and consistency. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (*SIGMOD* 2011) 217–228. ACM, New York.

DRECHSLER, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*: *Theory and Implementation*. *Lecture Notes in Statistics* **201**. Springer, New York. MR2809912 https://doi.org/10.1007/978-1-4614-0326-5

DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2013). Local privacy and statistical minimax rates. In 2013 *IEEE* 54*th Annual Symposium on Foundations of Computer Science—FOCS* 2013 429–438. IEEE Computer Soc., Los Alamitos, CA. MR3246246 https://doi.org/10.1109/FOCS.2013.53

DWORK, C. (2008). Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*. *Lecture Notes in Computer Science* **4978** 1–19. Springer, Berlin. MR2472670 https://doi.org/10.1007/978-3-540-79228-4_1

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

DWORK, C. and ROTH, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9** 211–407. MR3254020 https://doi.org/10.1561/0400000042

DWORK, C. and ROTHBLUM, G. N. (2016). Concentrated differential priacy. Preprint. Available at arXiv:1603.01887.

DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006a). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*. *Lecture Notes in Computer Science* **3876** 265–284. Springer, Berlin. MR2241676 https://doi.org/10.1007/11681878_14

DWORK, C., KENTHAPADI, K., MCSHERRY, F., MIRONOV, I. and NAOR, M. (2006b). Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology—EUROCRYPT* 2006. *Lecture Notes in Computer Science* **4004** 486–503. Springer, Berlin. MR2423560 https://doi.org/10.1007/11761679_29

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* 1054–1067. ACM, New York.

FANTI, G., PIHUR, V. and ERLINGSSON, Ú. (2016). Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proc. Priv. Enhanc. Technol.* **2016** 41–61.

FRIEDMAN, A., BERKOVSKY, S. and KAAFAR, M. A. (2016). A differential privacy framework for matrix factorization recommender systems. *User Model. User-Adapt. Interact.* **26** 425–458.

GABOARDI, M., HONAKER, J., KING, G., NISSIM, K., ULLMAN, J. and VADHAN, S. (2016). PSI (Ψ): A private data sharing interface. Preprint. Available at 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. *J. Am. Med. Inform. Assoc.* **20** 109–116.

GIL, D., GIRELA, J. L., DE JUAN, J., GOMEZ-TORRES, M. J. and JOHNSSON, M. (2012). Predicting seminal quality with artificial intelligence methods. *Expert Syst. Appl.* **39** 12564–12573.

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** 520–532.

HALL, R., RINALDO, A. and WASSERMAN, L. (2013). Differential privacy for functions and functional data. *J. Mach. Learn. Res.* **14** 703–727. MR3033345

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* 2339–2347.

HARDT, M. and ROTHBLUM, G. N. (2010). A multiplicative weights mechanism for privacy-preserving data analysis. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 61–70. IEEE Computer Soc., Los Alamitos, CA. MR3024776

HARDT, M. and TALWAR, K. (2010). On the geometry of differential privacy. In *STOC*'10*—Proceedings of the* 2010 *ACM International Symposium on Theory of Computing* 705–714. ACM, New York. MR2743320

HAY, M., RASTOGI, V., MIKLAU, G. and SUCIU, D. (2010). Boosting the accuracy of differentially private histograms through consistency. *Proc. VLDB Endow.* **3** 1021–1032.

HAY, M., MACHANAVAJJHALA, A., MIKLAU, G., CHEN, Y. and ZHANG, D. (2016). Principled evaluation of differentially private algorithms using dpbench. In *Proceedings of the* 2016 *International Conference on Management of Data* 139–154. ACM, New York.

HE, X., CORMODE, G., MACHANAVAJJHALA, A., PROCOPIUC, C. M. and SRIVASTAVA, D. (2015). DPT: Differentially private trajectory synthesis using hierarchical reference systems. *Proc. VLDB Endow.* **8** 1154–1165.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. MR1765176 https://doi.org/10.1214/ss/1009212519

HOMER, N., SZELINGER, S., REDMANN, M., DUGGAN, D., TEMBE, W., MUEHLING, J., PEARSON, J. V., STEPHAN, D. A., NELSON, S. F. et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4** e1000167.

HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K. and DE WOLF, P.-P. (2012). *Statistical Disclosure Control*. *Wiley Series in Survey Methodology*. Wiley, Chichester. MR3026260 https://doi.org/10.1002/9781118348239

KARWA, V., KRIVITSKY, P. N. and SLAVKOVIĆ, A. B. (2017). Sharing social network data: Differentially private estimation of exponential family random-graph models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 481–500. MR3632338 https://doi.org/10.1111/rssc.12185

KERMAN, J. (2011). Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electron. J. Stat.* **5** 1450–1470. MR2851686 https://doi.org/10.1214/11-EJS648

KIFER, D., SMITH, A. and THAKURTA, A. (2012). Private convex empirical risk minimization and high-dimensional regression. *J. Mach. Learn. Res. Workshop Conf. Proc.* **23** 1–40.

KINNEY, S. K., REITER, J. P., REZNEK, A. P., MIRANDA, J., JARMIN, R. S. and ABOWD, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *Int. Stat. Rev.* **79** 362–384.

KOTSOGIANNIS, I., MACHANAVAJJHALA, A., HAY, M. and MIKLAU, G. (2017). Pythia: Data dependent differentially private algorithm selection. In *Proceedings of the* 2017 *ACM International Conference on Management of Data* 1323–1337. ACM, New York.

KOWALCZYK, L., MALKIN, T., ULLMAN, J. and WICHS, D. (2017). Hardness of non-interactive differential privacy from one-way functions. *Semantic Scholar*. Available at https://pdfs.semanticscholar.org/5a40/c47078efdd3d6b6d15f323d6c3bc0d709ea8.pdf.

LEE, J. and CLIFTON, C. (2011). How much is enough? Choosing for differential privacy. In *Information Security*: *Proceedings of the* 14*th*, *International Conference*, *ISC* 2011 (X. Lai, J. Zhou and H. Li, eds.) 325–340. ACM, New York.

LEI, J., CHAREST, A.-S., SLAVKOVIC, A., SMITH, A. and FIENBERG, S. (2018). Differentially private model selection with penalized and constrained likelihood. *J. Roy. Statist. Soc. Ser. A* **181** 609–633. MR3807500 https://doi.org/10.1111/rssa.12324

LI, H., XIONG, L. and JIANG, X. (2014a). Differentially private synthesization of multi-dimensional data using copula functions. In *Advances in Database Technology*: *Proceedings. International Conference on Extending Database Technology* **2014** 475. NIH Public Access.

LI, H., XIONG, L. and JIANG, X. (2014b). Differentially private synthesization of multi-dimensional data using copula functions. In *Proc.* 17*th International Conference on Extending Database Technology* (*EDBT*) 475–486.

LI, N., LYU, M., SU, D. and YANG, W. (2016). Differential privacy: From theory to practice. *Synth. Lect. Inf. Secur. Priv. Trust* **8** 1–138.

LITTLE, R. J. A. (1993). Statistical analysis of masked data. *J. Off. Stat.* **9** 407.

LITTLE, R. J. A., LIU, F. and RAGHUNATHAN, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives Wiley Ser. Probab. Stat.* 141–152. Wiley, Chichester. MR2138251 https://doi.org/10.1002/0470090456.ch13

LIU, F. (2016). Model-based differentially private data synthesis. Preprint. Available at arXiv:1606.08052.

LIU, F. (2019a). Generalized Gaussian mechanism for differential privacy. *IEEE Trans. Knowl. Data Eng.* **31** 747–756.

LIU, F. (2019b). Statistical properties of sanitized results from differentially private Laplace mechanisms with noninformative bounding. *Trans. Data Priv.* **12** 169–195.

LIU, F. and LITTLE, R. J. A. (2003). SMIKe vs. data swapping and PRAM for statistical disclosure limitation in microdata: A simulation study. In *Proceedings of* 2003 *American Statistical Association Joint Statistical Meeting*.

MACHANAVAJJHALA, A., KIFER, D., ABOWD, J., GEHRKE, J. and VILHUBER, L. (2008). Privacy: Theory meets practice on the map. In *IEEE ICDE IEEE* 24*th International Conference* 277–286.

MACLEOD, A. J. and HENDERSON, G. R. (1984). Bounds for the sample standard deviation. *Teach. Stat.* **6** 72–76.

MANRIQUE-VALLIER, D. and REITER, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *J. Amer. Statist. Assoc.* **107** 1385–1394. MR3036402 https://doi.org/10.1080/01621459.2012.710508

MCCLURE, D. and REITER, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Priv.* **5** 535–552. MR3018910

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* 19–30. ACM, New York.

MCSHERRY, F. and TALWAR, K. (2007). Mechanism design via differential privacy. In *Foundations of Computer Science*, 2007. *FOCS'*07. 48*th Annual IEEE Symposium* 94–103. IEEE, Washington, DC.

MOHAMMED, N., CHEN, R., FUNG, B. and YU, P. S. (2011). Differentially private data release for data mining. In *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 493–501. ACM, New York.

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

NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2007). Smooth sensitivity and sampling in private data analysis. In *STOC'*07*—Proceedings of the* 39*th Annual ACM Symposium on Theory of Computing* 75–84. ACM, New York. MR2402430 https://doi.org/10.1145/1250790.1250803

NISSIM, K. and STEMMER, U. (2015). On the generalization properties of differential privacy. *CoRR*, Abs/1504.05800.

ORR, A. (2017). Google's differential privacy may be better than Apple's. https://www.macobserver.com/analysis/google-apple-differential-privacy/.

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* 13–18. ACM, New York.

QARDAJI, W., YANG, W. and LI, N. (2013). Understanding hierarchical methods for differentially private histograms. *Proc. VLDB Endow.* **6** 1954–1965.

RAAB, G. M., NOWOK, B. and DIBBEN, C. (2017). Practical data synthesis for large samples. *J. Priv. Confid.* **7** 67–97.

RAGHUNATHAN, T. E., REITER, J. P. and RUBIN, D. B. (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* **19** 1–16.

RAGHUNATHAN, T. E., SOLENBERGER, P. W. and HOEWYK, J. V. (2017). IVEware: Imputation and Variance Estimation software for SRMI. Available at http://www.isr.umich.edu/src/smp/ive (accessed December 20, 2018).

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VANHOEWYK, J. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* **27** 85–95.

REITER, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *J. Off. Stat.* **18** 531–543.

REITER, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Surv. Methodol.* **29** 181–188.

REITER, J. P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100** 1103–1112. MR2236926 https://doi.org/10.1198/016214505000000619

REITER, J. P. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *Int. Stat. Rev.* **77** 179–195.

REITER, J. P. and KINNEY, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *J. Off. Stat.* **28** 583–590.

ROTH, A. and ROUGHGARDEN, T. (2010). Interactive privacy via the median mechanism. In *STOC'*10*—Proceedings of the* 2010 *ACM International Symposium on Theory of Computing* 765–774. ACM, New York. MR2743326

RUBIN, D. B. (1993). Discussion statistical disclosure limitation. *J. Off. Stat.* **9** 461–468.

RUBINSTEIN, B. I., BARTLETT, P. L., HUANG, L. and TAFT, N. (2009). Learning in a large function space: Privacy-preserving mechanisms for SVM learning. Preprint. Available at arXiv:0911.5708.

SCOTT, D. W. (2015). *Multivariate Density Estimation*: *Theory*, *Practice*, *and Visualization*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3329609

SHEFFET, O. (2017). Differentially private ordinary least squares. In *Proceedings of the* 34*th International Conference on Machine Learning—Vol*. 70 3105–3114. JMLR.org.

SMITH, A. and THAKURTA, A. (2013). Differentially private model selection via stability arguments and the robustness of the Lasso. *J. Mach. Learn. Res. Workshop Conf. Proc.* **30** 1–31.

STEINKE, T. and ULLMAN, J. (2017). Between pure and approximate differential privacy. *J. Priv. Confid.* **7** 2.

SWEENEY, L. (2013). Matching known patients to health records in Washington state data. *Social Science Research Network*. id=2289850.

TANG, J., KOROLOVA, A., BAI, X., WANG, X. and WANG, X. (2017). Privacy loss in Apple's implementation of differential privacy on macOS 10.12. Preprint. Available at arXiv:1709.02753.

VADHAN, S. (2017). The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*. *Inf. Secur. Cryptography* 347–450. Springer, Cham. MR3837668

WANG, Y.-X., FIENBERG, S. and SMOLA, A. (2015). Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *International Conference on Machine Learning* 2493–2502.

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* 35*th Annual IEEE International Conference on* 1–9. IEEE, Washington, DC.

WASSERMAN, L. and ZHOU, S. (2010). A statistical framework for differential privacy. *J. Amer. Statist. Assoc.* **105** 375–389. MR2656057 https://doi.org/10.1198/jasa.2009.tm08651

XIAO, Y., GARDNER, J. and XIONG, L. (2012). Dpcube: Releasing differentially private data cubes for health information. In 2012 *IEEE* 28*th International Conference on Data Engineering* 1305–1308. IEEE, Washington, DC.

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

XU, J., ZHANG, Z., XIAO, X., YANG, Y., YU, G. and WINSLETT, M. (2013). Differentially private histogram publication. *VLDB J.* **22** 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. *J. Biomed. Inform.* **50** 133–141.

ZHANG, J., ZHANG, Z., XIAO, X., YANG, Y. and WINSLETT, M. (2012). Functional mechanism: Regression analysis under differential privacy. *Proc. VLDB Endow.* **5** 1364–1375.

ZHANG, J., XIAO, X., YANG, Y., ZHANG, Z. and WINSLETT, M. (2013). PrivGene: Differentially private model fitting using genetic algorithms. In *Proceedings of the* 2013 *ACM SIGMOD International Conference on Management of Data* 665–676. ACM, New York.

ZHANG, J., CORMODE, G., PROCOPIUC, C. M., SRIVASTAVA, D. and XIAO, X. (2017). PrivBayes: Private data release via Bayesian networks. *ACM Trans. Database Syst.* **42** Art. 25. MR3730676 https://doi.org/10.1145/3134428