Journal of Privacy and Confidentiality
Vol. 12 (1) 2022

Submitted Dec 2020
Published Jul 2022

A LATENT CLASS MODELING APPROACH FOR GENERATING SYNTHETIC DATA AND MAKING POSTERIOR INFERENCES FROM DIFFERENTIALLY PRIVATE COUNTS

MICHELLE PISTNER NIXON, ANDRÉS F. BARRIENTOS, JEROME P. REITER, AND ALEKSANDRA SLAVKOVIĆ

College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802
e-mail address: pistner@psu.edu

Department of Statistics, Florida State University, Tallahassee, FL 32306
e-mail address: abarrientos@fsu.edu

Department of Statistical Science, Duke University, Durham, NC 27708
e-mail address: jreiter@duke.edu

Department of Statistics, The Pennsylvania State University, University Park, PA 16802
e-mail address: sesa@psu.edu

Abstract. Several algorithms exist for creating differentially private counts from contingency tables, such as two-way or three-way marginal counts. The resulting noisy counts generally do not correspond to a coherent contingency table, so that some post-processing step is needed if one wants the released counts to correspond to a coherent contingency table. We present a latent class modeling approach for post-processing differentially private marginal counts that can be used (i) to create differentially private synthetic data from the set of marginal counts, and (ii) to enable posterior inferences about the confidential counts. We illustrate the approach using a subset of the 2016 American Community Survey Public Use Microdata Sets and the 2004 National Long Term Care Survey.

1. Introduction

National statistical organizations and other data curators, henceforth all called agencies, often seek to share collected information with outside researchers. Doing so requires methods that provide privacy protection to data subjects while maintaining statistical relationships in the data. Many agencies use statistical disclosure control (SDC) methods that blur the private data in a prescribed way, with the aim to provide some level of privacy protection. Traditional SDC methods include data swapping, cell suppression, top- or bottom- coding,

Key words and phrases: Census, disclosure, privacy.
Code accompanying this article is available at Nixon (2022) and https://github.com/journalprivacyconfidentiality/BayesLCM/releases/tag/v20220725.

JPrivacy Confidentiality
www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.768

and the addition of random noise (Hundepool et al., 2012). Generally, traditional SDC techniques have been applied with low intensity, so as not to degrade the quality of the information in the data. However, with the growth in available data and powerful computing, many agencies have become concerned that such SDC methods are not adequately protective.

To address this issue, several agencies have turned to using synthetic data methods (Rubin, 1993; Little, 1993; Reiter, 2005a). The agency releases data that are generated from some statistical model, estimated with the private data. With a full synthesis, agencies can ensure that no sensitive values are released on the file, which can reduce disclosure risks. Further, the agency can offer researchers access to record-level data instead of only summary statistics or other high-level information. Many different methods have been proposed to generate synthetic data (e.g., Reiter, 2005b; Woodcock and Benedetto, 2009; Mateo-Sanz et al., 2004; Slavković and Lee, 2010) and assess the utility and disclosure risk associated with the data release (Drechsler, 2011; Hundepool et al., 2012; Karr et al., 2006; Reiter et al., 2014; Snoke et al., 2018). The U.S. Census Bureau has released synthetic data for the Longitudinal Business Database (Kinney et al., 2011), Survey of Income and Program Participation (Benedetto et al., 2013), and the LEHD Origin-Destination Employment Statistics (with the data product being known as *OnTheMap*) (Machanavajjhala et al., 2008). Affiliated researchers have proposed methodologies for other national surveys (e.g., federal administrative data from the Office of Personnel Management (Barrientos et al., 2018)), and other statistical agencies have developed synthetic data products such as the Scottish Longitudinal Survey (SLS-DSU, 2018) and the German IAB Establishment Data (Drechsler, 2009).

Most SDC methods, including many synthetic data techniques, do not provide formally quantifiable privacy guarantees. Instead, agencies evaluate disclosure risks based on assumptions of intruder knowledge and behavior (Hundepool et al., 2012). An alternative approach is to design and use SDC methods that satisfy differential privacy (DP, (Dwork et al., 2006b)). Differentially private algorithms exist for releasing counts (e.g., Dwork et al., 2006b; Ghosh et al., 2012), k-way marginals (Barak et al., 2007; Yang et al., 2012; Li et al., 2018), regression coefficients (Zhang et al., 2012; Snoke and Slavković, 2018; Awan and Slavkovic, 2021), and many other quantities that can be viewed as answers to user-specified queries. Several statistical agencies are interested in combining the privacy guarantees from differential privacy with the flexibility afforded by synthetic data. For example, the U.S. Census Bureau uses differentially private synthetic data as the backbone of its 2020 population census data releases (Abowd, 2018). However, it can be challenging to generate differentially private synthetic data with low error (Garfinkel et al., 2018).

In this article, we propose methodology to generate differentially private synthetic data sets for multivariate count data, i.e., data from contingency tables. Our approach can be viewed as a post-processing method for generating synthetic data, a strategy also suggested in McKenna et al. (2021). Specifically, we first assume that an agency has selected a set of marginal counts and has used some existing differentially private algorithm to add noise to those counts. For example, the agency could select the margins for which accurate values are especially important. Alternatively, the agency could select a set of margins that covers many types of analyses done by users of the data, e.g., all k-way margins. Second, we specify a Bayesian latent class model for the underlying confidential data (Dunson and Xing, 2009). We write the agency-selected margins as functions of the model parameters. Third, we estimate the parameters in the functions using only the differentially private counts and a composite likelihood-based approach (Lindsay, 1988). Finally, we sample from the estimated model to generate record-level, synthetic data. Agencies can generate multiple copies of the synthetic datasets, without any extra privacy loss, in order to enable secondary data analysts to estimate uncertainty and make inferences (Reiter, 2003). Generating differentially private synthetic data from a user-specified list of summaries can be particularly advantageous for contingency tables with complex structures, e.g., data with structural zeros (Manrique-Vallier and Reiter, 2014; Li et al., 2018) or nested data such as individuals in households (Hu et al., 2018). We illustrate the latter in the online Supplementary Material.

The latent class modeling approach can also be viewed as an engine for approximate posterior inferences for the underlying confidential counts given a collection of differentially private counts, under the assumptions implied by the latent class model. Specifically, analysts can use parameters draws from the Markov chain Monte Carlo (MCMC) sampler to obtain posterior inferences for any functional of the parameters, i.e., any count in the table.

The remainder of the paper is organized as follows. In Section 1.1 we review related literature. In Section 2 we present preliminary information on data privacy and latent class models. In Section 3 we discuss the proposed approach and implementation details. In Section 4, we illustrate the approach by making posterior inferences from differentially private counts with a small set of variables from the 2016 American Community Survey (ACS) Public Use Microdata Sets (PUMS), and compare it to several other approaches from the literature. In Section 5, we illustrate the approach for both posterior inference and synthetic data generation using the 2004 National Long Term Care Survey, which has larger dimensions than the ACS example. For both illustrations, we use the Geometric Mechanism (Ghosh et al., 2012) to ensure differential privacy in the first stage counts. Finally, in Section 6 we end with a discussion.

1.1. **Related Literature.** Protection of counts and contingency tables has been the subject of many papers in the privacy literature. Some of the first examples of differentially private (DP) data were counts and histograms perturbed by the Laplace mechanism (Dwork et al., 2006b). Barak et al. (2007) proposed methodology to produce DP marginal counts by using Fourier bases. Yang et al. (2012) extended this work to multi-dimensional contingency tables and noted some theoretical and practical shortcomings to both approaches. Machanavajjhala et al. (2008) proposed constructing DP contingency tables via a Dirichlet-Multinomial synthesizer, and the Hardt-Ligett-McSherry algorithm (Hardt et al., 2012) used a multiplicative weights approach.

Park and Ghosh (2014) proposed methodology for synthesis of DP contingency tables by disintegrating the data into suitable building blocks, injecting noise to these blocks, and using a Gibbs sampler to draw synthetic samples. Li et al. (2018) proposed to generate synthetic data by constructing an empirical hashed conditional distribution from the whole histogram and applying a Stability Based Algorithm to these empirical distributions. McKenna et al. (2019) proposed constructing DP synthetic data where "suitable building blocks" are a set of selected marginal counts; they add Laplace noise to the selected counts, fit a graphical model to the perturbed counts, and sample synthetic data from this model. Similarly, PrivBayes (Zhang et al., 2017) constructs DP synthetic data sets via Bayesian networks. These authors use networks to approximate the underlying data distribution through lower order marginals, add noise to these marginals, and approximate a full data distribution through the noisy counts and constructed Bayesian network. We note that these two methods placed first (McKenna et al., 2021) and third (Bao et al., 2021), respectively, in the 2018-2019 National Institute of Standards and Technology Public Safety Communications Research division's *Differential Privacy Synthetic Data Challenge* (National Institute of Standards and Technology, 2021); see Bowen and Snoke (2021) for a discussion of the performance of these methods in the context of the challenge. Similarly, the CIPHER method Eugenio and Liu (2021) estimates the joint distribution of a table based on a lower-order set of DP marginals. Their approach can be viewed as a post-processing technique for generating synthetic data.

Hay et al. (2010) showed that accuracy can be greatly improved by accounting for the constraints that a contingency table must satisfy in post-processing steps, although some recent papers argue for post-processing to be part of posterior modeling (e.g., see Seeman et al. (2020), and references therein). Lee et al. (2015) extended this result by also accounting for the noise distribution and developed a fast generic approach for solving the resulting optimization problem.

Bowen and Liu (2020) compared many parametric and nonparametric methods for differentially private data synthesis, beyond those discussed above, and Charest (2012) showed that requiring greater privacy can degrade inferential results. Rinott et al. (2018) discussed several issues related to contingency table release under differential privacy, including the potential effects of rounding and other post-processing steps. They compared several simple mechanisms and offered comments on utility in private data releases. Similarly, Raab (2019) discussed some practical limitations to contingency table protection under formal privacy guarantees and presented a new method to create differentially private contingency tables from a subset of marginals via the Laplace mechanism and iterative proportional fitting.

All of these methods (ours included) are subject to the results of Ullman and Vadhan (2020). They showed that, even in simple cases, it is impossible to construct a polynomial time differentially private algorithm that preserves all two-way marginals.

Accounting for the additional privacy-preserving noise in inference also has been discussed in the literature. Charest (2012) used a simple Bayesian model to account for noise added under the Beta-Binomial synthesizer. Karwa et al. (2015) accounted for noise by treating the original data as missing. As the resulting likelihoods are often intractable, they developed a technique relying on a variational approximation for estimation. Seeman et al. (2020) showed that naive post-processing (such as direct modification of counts to meet constraints) can result in loss of information and proposed a Bayesian sampling scheme to post-process counts based on the noisy counts and outside constraining information. Gong (2019) showed that approximate Bayesian computing can be used to account for DP noise to obtain draws from the correct posterior distribution. Karwa et al. (2017) investigated methods to release DP synthetic networks while also accounting for the additional noise by treating the original private data as missing and using a likelihood-based approach. While not dealing with DP tables, Woo and Slavković (2012) proposed several EM algorithms to obtain correct logistic regression estimates for tables that are protected using the Post Randomization Method (PRAM) which could be extended to DP settings.

Turning to latent class models, Si and Reiter (2013) used latent class specifications for missing data imputation in contingency tables using the model of Dunson and Xing (2009). This latent class model specification belongs to the family of Bayesian nonparametric techniques, which are well-known for being a flexible modeling choice to capture complex data patterns. Manrique-Vallier and Reiter (2014) refined these latent class models to incorporate cases with structural zeros, e.g., a person age five cannot have a college degree. The full table latent class approach was further extended to handle nested data such as people within households (Hu et al., 2018), structural zeros (Akande et al., 2019a), and missing data (Akande et al., 2019b); see the Supplementary Material for a discussion of the model presented in Hu et al. (2018) and how our proposed approach can be augmented for nested data structures.

2. Preliminaries

2.1. **Data Privacy Methods.**

2.1.1. *Synthetic Data.* Let $\mathbf{X} = \{(x_{i1}, \ldots, x_{ip})\}_{i=1}^n$ be an observed data set comprising n individuals measured on p confidential categorical variables. For $i = 1, \ldots, n$ and $j = 1, \ldots, p$, each $x_{ij} \in \{1, \ldots, d_j\}$. We refer to the variables in $\mathbf{X}$ using X_j where $j = 1, \ldots, p$. Synthetic data approaches aim to preserve the joint distribution $f(\mathbf{X})$ of these variables by modeling them simultaneously, e.g., as in Hu et al. (2018) and Akande et al. (2019b), or by using sequential modeling of the form,

$$f(\mathbf{X}) = f(X_1)f(X_2|X_1)\cdots f(X_p|X_1, \ldots, X_{p-1}). \tag{2.1}$$

Different types of statistical models can be used for the conditional distributions in (2.1), including classification and regression trees (CART) (Reiter, 2005b), random forests (Caiola and Reiter, 2010), kernel density estimators (Woodcock and Benedetto, 2009), and combinations thereof (Barrientos et al., 2018). Choice of model depends on both the data structure and desired privacy levels.

2.1.2. *Differential Privacy.* Differential privacy is a formal privacy framework that guarantees privacy protection by bounding the ratio of output densities for all neighboring data sets. Intuitively, it protects against attackers seeking to learn whether or not any specific individual's data were included (or excluded) in the data set over which the output was calculated. That is, the inclusion (or exclusion) of a single person's information does not change the output of the algorithm greatly. In this paper we focus on its strongest form, ϵ-DP, where ϵ is a privacy-loss parameter, with smaller values indicating stronger protection, but our model could be modified for other forms of DP.

Definition 2.1. ϵ-Differential Privacy (Dwork et al., 2006b): A randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if for all data sets $\mathbf{X} = \{(x_{i1}, \ldots, x_{ip})\}_{i=1}^n$ and $\mathbf{X}'$ differing on at most one row, and $\mathcal{S} \subseteq \text{Range}(\mathcal{A})$,

$$\frac{\Pr[\mathcal{A}(\mathbf{X}) \in S]}{\Pr[\mathcal{A}(\mathbf{X}') \in S]} \leq \exp(\epsilon)\,. \tag{2.2}$$

Differential privacy is a property of the algorithm itself. It is achieved by randomness. Many different algorithms have been proposed to release differentially private statistics or data sets, e.g., the Laplace (Dwork et al., 2006b) or Geometric (Ghosh et al., 2012) mechanisms, which are the most relevant to our setting. Several relaxations and extensions of differential privacy have been proposed such as (ϵ, δ)-differential privacy (Dwork et al., 2006a), concentrated differential privacy (Bun and Steinke, 2016), and Rényi differential privacy (Mironov, 2017); for more, see Dwork et al. (2014).

Many DP mechanisms (ours included) rely on the useful properties of post-processing and sequential composition. These properties outline how the DP privacy guarantee can be preserved and how the privacy-loss parameter is propagated through multiple data releases. Theorem 2.2 shows that DP is preserved through post-processing and is key to our proposed approach: the augmented Bayesian latent class model and measurement error model is a post-processing technique for a noisy, privacy-preserving count. Theorem 2.3 shows how the privacy-loss parameter ϵ propagates through multiple data releases on the same individual.

Theorem 2.2. *Post-processing (Dwork et al., 2006b): Let $\mathcal{A}$ be any randomized algorithm such that $\mathcal{A}(\mathbf{X})$ is ϵ-differentially private, and let g be any function. Then, $g(\mathcal{A}(\mathbf{X}))$ also satisfies ϵ-differential privacy.*

Theorem 2.3. *Sequential composition (McSherry, 2009): Let $\mathcal{A}_i$ each provide ϵ_i-differential privacy. The sequence of $\mathcal{A}_i(\mathbf{X})$ provides $(\sum_i \epsilon_i)$-differential privacy.*

2.1.3. *Geometric Mechanism.* The Geometric Mechanism (Ghosh et al., 2012) adds noise from a two-sided geometric distribution to achieve DP. Properties of this distribution, including simulation techniques, are discussed in Inusah and Kozubowski (2006) and its references.

Definition 2.4. A random variable Y distributed as a two-sided geometric distribution has probability mass function

$$P(Y = k) = \frac{1-\alpha}{1+\alpha}\alpha^{|k|}, \tag{2.3}$$

where $0 \le \alpha \le 1$. We refer to this distribution using two-sided-Geom(α).

Theorem 2.5. *Geometric Mechanism (Ghosh et al., 2012): For $f : \mathcal{D} \to \mathbf{R}^d$, the mechanism $\mathcal{A}$ that adds independently drawn noise from a* two-sided-Geom($\exp\{\frac{-\epsilon}{\Delta f}\}$) *distribution to each of the d terms of satisfies ϵ-differential privacy.*

This mechanism has several appealing properties for protecting count data. First, the added noise values are integers, which eliminates the need for a post-processing step to deal with fractional parts. Second, Ghosh et al. (2012) show that this mechanism is optimal for every potential user regardless of the side information that they posses when releasing DP count queries using a Bayesian framework.

2.2. **Bayesian Latent Class Models.** Dunson and Xing (2009) developed a flexible methodology for modeling unordered categorical data using a Bayesian nonparametric mixture model. Their methods are directly applicable to modeling complete contingency tables.

Suppose each observation $i = 1, \dots, n$ belongs to a latent class denoted by $z_i \in \{1, 2, \dots\}$. The probability for each unique combination of variable levels is specified according to these latent classes. For ease of notation, we allow x_{ij} to stand for random variables as well as the observed data. The latent class model is of the form,

$$
\begin{aligned}
x_{ij}|z_i, \{\Psi_h^{(j)}\}_{h=1}^{\infty} &\overset{ind}{\sim} Multinomial\{1, \Psi_{z_i1}^{(j)}, \dots, \Psi_{z_id_j}^{(j)}\}, i = 1, \dots, n,\ j = 1, \dots, p, \\
z_i|\{\pi_h\}_{h=1}^{\infty} &\overset{ind}{\sim} Discrete\{(1, \dots, \infty), (\pi_1, \dots, \pi_\infty)\}, \\
\pi_h &= V_h \prod_{l<h} (1 - V_l), \quad V_h \sim \beta(1, \alpha), \\
\Psi_h^{(j)} &\sim Dirichlet(a_{j1}, \dots, a_{jd_j}),
\end{aligned}
\tag{2.4}
$$

where $\Psi_h^{(j)} = (\Psi_{h1}^{(j)}, \dots, \Psi_{hd_j}^{(j)})$, $\alpha > 0$, and $(a_{j1}, \dots, a_{jdj})$ is a vector with positive components. Under model (2.4), for any feasible set of values $(c_1, \dots, c_p)$ of the categorical variables, one has

$$
Pr(x_{i1} = c_1, \dots, x_{ip} = c_p|\{\Psi_h^{(j)}\}_{h=1}^{\infty}, \{\pi_h\}_{h=1}^{\infty}) = \sum_{h=1}^{\infty} \pi_h \prod_{j=1}^{p} \Psi_{hc_j}^{(j)}, \tag{2.5}
$$

which corresponds to a mixture of product multinomials with a Dirichlet process (Ferguson, 1973) as a mixing distribution.

To facilitate posterior computation, we use a finite dimensional approximation of the model in (2.4), which can be estimated using Gibbs sampling. The approximation relies on the assumption that $z_i \in \{1, \dots, k\}$, where $k < \infty$, which is equivalent to truncating the mixture model (2.5) to k components; see Ishwaran and James (2001) for a discussion of the truncated Dirichlet processes. Similar approximations along with Gibbs sampling have been successfully used for private data synthesis, including in Hu et al. (2018), Manrique-Vallier and Reiter (2014) and Akande et al. (2019b), among others.

2.3. **Composite Likelihood Methods.** Composite likelihood methods allow analysts to circumvent the specification of the full joint likelihood function (Lindsay, 1988; Varin et al., 2011). Instead, analysts can use an approximation to the likelihood function, often with the assumption of some form of independence, to simplify computation.

We use composite likelihood methods to approximate the joint likelihood of the latent class model given an agency-specified set of marginal counts. Let $(M_1, \dots, M_T)$ represent the list of marginal counts of $\mathbf{X}$ with $\theta = \left(\{\Psi_h^{(j)}\}_{h=1,j=1}^{k,p}, \{\pi_h\}_{h=1}^{k}\right)$. Let $\mathcal{L}_t(\theta; M_t)$ be the likelihood function corresponding to M_t induced by Model (2.4). We approximate the joint likelihood of θ given $(M_1, \dots, M_T)$ using

$$
\mathcal{L}_C(\theta; M_1, \dots, M_T) \approx \prod_{t=1}^{T} \mathcal{L}_t(\theta; M_t). \tag{2.6}
$$

The composite likelihood $\mathcal{L}_C$ allows for inference about θ only from the marginal counts (Varin et al., 2011). We note that, while computationally expedient, inference based on only a subset of counts of $\mathbf{X}$ could result in inaccurate estimates of θ (Walker, 2013), particularly for regions of its distribution that describe the counts excluded from the set of selected marginals. Furthermore, not all combinations of marginals and (possibly) conditional distributions fully and uniquely specify the joint distribution; for example, see Fienberg and Slavkovic (2005); Slavković et al. (2015) within the context of confidentiality protection and more general related references therein.

3. Proposed Methods

3.1. **Overview.** We assume the agency adds noise to the true counts for a set of agency-specified marginals, $(M_1, \ldots, M_T)$, computed from $\mathbf{X}$ using the Geometric Mechanism. In our simulation examples, we use the definition of sensitivity based on changing one row; hence, the sensitivity for each M_t, where $t = 1, \ldots, T$, equals two.[1] We assume the agency has determined appropriate values of ϵ to control the overall privacy budget. See Section 4 for a discussion in the context of our simulations.

When estimating the parameters of the Bayesian latent class model, we account for the additional noise due to DP using a measurement error model (Fuller, 2009). Specifically, we treat the true underlying counts as unknown and incorporate the privacy preserving mechanism into the model. Let $\tilde{\mathbf{M}}$ denote the observed noisy marginal counts. We have

$$\begin{aligned} \tilde{\mathbf{M}} &= (M_1 + \varepsilon_1, \ldots, M_T + \varepsilon_T), \\ \varepsilon_t &\overset{ind}{\sim} \text{two-sided-Geom}_{r_t}\left(\exp\left\{\frac{-\epsilon}{\Delta M_t T}\right\}\right), t = 1, \ldots, T, \\ M_t|\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k &\overset{ind}{\sim} \text{Multinomial}_{r_t}(n, P_t(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)),\ t = 1, \ldots, T, \\ \boldsymbol{\pi}_k &= \{\pi_h\}_{h=1}^{k},\ \boldsymbol{\Psi}_k = \{\Psi_h^{(j)}\}_{h=1,j=1}^{k,p}, \end{aligned} \tag{3.1}$$

where each M_t comprises r_t counts, ε_t is a random vector with r_t independent components distributed as two-sided-Geom $\left(\exp\left\{\frac{-\epsilon}{\Delta M_t}\right\}\right)$, and $\Delta M_t = 2$ is the sensitivity of M_t. We complete the model specification by assuming the prior distributions for $\boldsymbol{\pi}_k$ and $\boldsymbol{\Psi}_k$ used in (2.4). Since the summaries M_t are assumed to be marginal counts, the distribution induced by $\mathbf{X} \mapsto M_t$ is multinomial with parameters n and $P_t(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$, where $P_t(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$ represents the probabilities of the counts in M_t and is computed using expression (2.5). We note that the Geometric Mechanism, which generates noise independently for each count, facilitates the assumptions of independence of the marginal counts used in the composite likelihood.

The third line of (3.1) represents our "augmented" step (Lindsay, 1988; Varin et al., 2011): each marginal count is assumed to be independently drawn from multinomial distributions. We rely on this assumption as it is not obvious how to characterize the joint distribution for $(M_1, \ldots, M_T)|\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k$. Although the premise of independence leads to a product of multinomial distributions that places positive probability outside the range of $(M_1, \ldots, M_T)$, we expect such probability to be small since $(P_1(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k), \ldots, P_T(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k))$ is a coherent system of probabilities that accounts for the underlying constraints among $M_1, \ldots, M_T$. Hence, in order to generate synthetic tables that consider such constraints, our strategy is to sample from the distribution of $\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k|\tilde{\boldsymbol{M}}$, and approximate the underlying table cell probabilities by

$$\begin{aligned} &Pr(x_{(n+1)1} = c_1, \ldots, x_{(n+1)p} = c_p|\tilde{\boldsymbol{M}}) = \\ &\qquad \int\int Pr(x_{(n+1)1} = c_1, \ldots, x_{(n+1)p} = c_p|\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k) Pr(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k|\tilde{\boldsymbol{M}}) d\boldsymbol{\pi}_k d\boldsymbol{\Psi}_k. \end{aligned}$$

Here, $Pr(x_{(n+1)1} = c_1, \ldots, x_{(n+1)p} = c_p|\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$ is defined as in (2.5).

Regardless of which marginal counts are used in estimation, probabilities corresponding to any cell of the contingency table or marginal count can be estimated through $(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$. At

[1] Under other definitions of sensitivity, the sensitivity equals 1 since adding or deleting one row changes at most one marginal count.

each iterate, the desired probabilities can be tabulated using the sampled values of $(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$ at that iterate. This results in a posterior distribution for each desired probability. Synthetic data can be created from the full cell counts either by straightforward post-processing (e.g., multiply full table probabilities by the desired table size and expand) or by multinomial sampling of the cells weighted by the full table probabilities.

Any arbitrary set of summary statistics can be fed into the model. However, to increase the statistical usefulness of the estimated table, we suggest using sets of summary statistics that include at least one instance of every variable and its corresponding levels, which we do here. In addition, when privacy budgets are constrained, we recommend using few rather than many marginal counts. Using a smaller set of marginals allocates more of the privacy budget to each marginal table, which can improve accuracy for those counts. We note that the dimension of parameters $(\boldsymbol{\pi}, \boldsymbol{\psi})$ to be estimated does not depend on the number of tables passed to the model.

In general, we expect these results to improve as the accuracy of the formally private counts improves. However, difficulties can arise with the measurement error step of our model if it is hard to specify the DP noise generating process.

This approach to differentially private synthesis does not suffer from privacy loss from running the estimation algorithm. Noise is injected to the marginal counts prior to the MCMC sampling (Chaudhuri et al., 2013). The algorithm uses these noisy marginals and knowledge of the noise distribution (which does not violate differential privacy) to estimate the underlying contingency table through $(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$.

To satisfy DP without using additional budget, one should set the number of latent classes k without using the confidential counts. Further, one should not use a k that exceeds the number of cells in the table; otherwise, the latent class model will have more parameters than observed cells in the data. We suggest starting with a reasonable number like $k = 10$ or $k = 20$, with larger starting sizes based on the size of the table. After fitting the model, examine the estimates of the mixing weights (i.e., the latent class probabilities) and increase k until some are very small. For our illustrative examples, this process resulted in values of k around 10.

3.2. **Illustrative Implementation.** To illustrate implementations, suppose the agency has generated all 40 two-way marginal counts from a 2^5 table. We can write each two-way marginal probability as a sum over all combinations of the remaining three variables, for example,

$$\begin{aligned} Pr(x_{i1} = 0, x_{i2} = 0 | \boldsymbol{\pi}_k, \boldsymbol{\Psi}_k) &= \sum_{j=0}^{1}\sum_{l=0}^{1}\sum_{m=0}^{1} Pr(x_{i1} = 0, x_{i2} = 0, x_{i3} = j, x_{i4} = l, x_{i5} = m | \boldsymbol{\pi}_k, \boldsymbol{\Psi}_k) \\ &= \sum_{h=1}^{k} \pi_h \Psi_{h0}^{(1)} \Psi_{h0}^{(2)} \sum_{j=0}^{1}\sum_{l=0}^{1}\sum_{m=0}^{1} \Psi_{hj}^{(3)} \Psi_{hl}^{(4)} \Psi_{hm}^{(5)} . \end{aligned} \tag{3.2}$$

Similarly, we have

$$
\begin{aligned}
Pr(x_{i1}=0, x_{i2}=1|\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k) &= \sum_{h=1}^{k} \pi_h \Psi_{h0}^{(1)} \Psi_{h1}^{(2)} \sum_{j=0}^{1}\sum_{l=0}^{1}\sum_{m=0}^{1} \Psi_{hj}^{(3)} \Psi_{hl}^{(4)} \Psi_{hm}^{(5)} \\
Pr(x_{i1}=1, x_{i2}=0|\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k) &= \sum_{h=1}^{k} \pi_h \Psi_{h1}^{(1)} \Psi_{h0}^{(2)} \sum_{j=0}^{1}\sum_{l=0}^{1}\sum_{m=0}^{1} \Psi_{hj}^{(3)} \Psi_{hl}^{(4)} \Psi_{hm}^{(5)} \qquad (3.3)\\
Pr(x_{i1}=1, x_{i2}=1|\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k) &= \sum_{h=1}^{k} \pi_h \Psi_{h1}^{(1)} \Psi_{h1}^{(2)} \sum_{j=0}^{1}\sum_{l=0}^{1}\sum_{m=0}^{1} \Psi_{hj}^{(3)} \Psi_{hl}^{(4)} \Psi_{hm}^{(5)}.
\end{aligned}
$$

We denote the probabilities in (3.2) and (3.3) by $P_t(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$ for each corresponding count in the model in (3.1).

After writing all the two-way marginal counts as functions of the latent class parameters, we can estimate the posterior distribution of all parameters using a MCMC sampler; see the Supplementary Material for details. We developed an `R` package to fit the proposed model given an arbitrary set of two-way marginals from a 2^p contingency table of any size[2].

One source of computational overhead for the MCMC sampler is computing all the marginal probabilities from the model parameters, i.e., $P_t(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$. The complexity of this calculation depends on the size of the contingency table, the number of latent classes, and the total number of marginals used as input. In simulations with our `R` package, we find that computing each two-way marginal probability is extremely fast (< 2 milliseconds) when the dimension is ten or less regardless of the number of latent classes. However, run time greatly increases for higher dimensions, leading to slower sampling for the overall model for higher-dimensional tables. Full results are presented in Section 3 of the Supplementary Material.

4. Application to American Community Survey Data

In this section, we illustrate the proposed approach using data from the ACS PUMS and a small number of variables; see Table 1. In this illustration, we use the latent class model as a post-processing algorithm to obtain posterior inferences from a set of released noisy counts.

4.1. Data Description.

For each simulation run, we randomly select a subset of 10,000 individuals from the 2016 one-year ACS PUMS. We use five variables on each individual, namely sex, age, race, citizenship, and income. We recode the latter four variables to binary variables using the rules in Table 1. Table 2 displays all the 2-way marginal tables from one such sample, to give a sense of the typical counts. There are a total of 2^5 table cells corresponding to $\binom{5}{2} \times 4 = 40$ two-way marginal counts.

To demonstrate our proposed approach, we make several simplifications here as well as in the NLTCS illustration in Section 5. First, we ignore that fact that the ACS is a complex survey with weights. We instead analyze the data "as is" without considering the design. Second, we treat all imputations of missing values as reported data and disregard any implications of the imputation procedures for computing the sensitivity of the Geometric

[2]This package, along with all code used in this paper can be found on Github at `https://github.com/michellepistner/BayesLCM`, and archived as Nixon (2022).

Table 1: Data description for the ACS recoded variables. Original data was collected from the 2016 1-year ACS PUMS and recoded according to the guidelines below.

Variable	**Label**	**Levels**
Sex	SEX	0: males, 1: females
Age	AGE	0: under age 18, 1: over age 18
Race	RACE	0: non-white, 1: white
Citizenship	CIT	0: non-U.S. citizen, 1: U.S. citizen
Income	INC	0: income under federal poverty line
		1: income above federal poverty line

Table 2: All two-way marginal tables for one randomly sampled subset of 10,000 observations from the 2016 ACS PUMS.

	Age	
Citizenship	0	1
0	11	596
1	443	8,950

	Race	
Citizenship	0	1
0	299	308
1	1,731	7,662

	Sex	
Citizenship	0	1
0	273	354
1	4,505	4,888

	Income	
Citizenship	0	1
0	294	313
1	2,916	6,477

	Race	
Age	0	1
0	110	344
1	1,920	7,626

	Sex	
Age	0	1
0	239	215
1	4,539	5,007

	Income	
Age	0	1
0	445	9
1	2,765	6,781

	Sex	
Race	0	1
0	945	1,085
1	3,833	4,137

	Income	
Race	0	1
0	827	1,203
1	2,382	5,587

	Income	
Sex	0	1
0	1,281	3,497
1	1,929	3,293

Mechanism. Incorporation of survey weights and imputation are important open problems in formal privacy (Reiter, 2019; Slavkovic and Seeman, 2022).

4.2. **Simulation Details.** We base our example on the presumption that the agency has decided that all two-way marginal tables are useful to the end data user and will create differentially private versions of them. To make the differentially private counts, we add independent noise drawn from the Geometric Mechanism with $\epsilon \in \{.25, .5, 1\}$. We invoke sequential composition theorems to adhere to the specified values of ϵ. In particular, each of the 40 two-way marginal counts belongs to one of $\binom{5}{2} = 10$ two-way tables. Thus, each individual contributes to only one cell per table but contributes across all ten tables. Accordingly, each marginal table should be perturbed with one-tenth of the privacy budget.

After generating the noisy counts in each simulation run, we use a Metropolis-within-Gibbs sampler to iteratively sample the unknown parameters, $(M_1, \dots, M_T, \boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$. Unless otherwise noted, we set the number of latent classes equal to $k = 10$. In fitting, generally at least four classes have non-trivial mass (as determined by $\boldsymbol{\pi_k}$); see the Supplementary Material for model implementation details.

We run the algorithm for 5,000 iterations, discarding the first 2,000 iterations as burn-in. We monitor convergence of model probabilities using trace plots, a standard practice in Bayesian data analysis (Gelman et al., 2013; Roy, 2020).[3] In each iteration, we compute the probabilities for all two-way margins and cells in the full table according to the latent class specification using the sampled parameters $(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$ at that iterate.

We compare outputs from the latent class model to output generated using PrivBayes (Zhang et al., 2017), following the architecture for adapting PrivBayes to private data releases described in Ping et al. (2017). To do so, we adapt their publicly-available Python code, accessible at `https://github.com/DataResponsibly/DataSynthesizer`. We also compare to a graphical model-based approach (McKenna et al., 2019) by adapting their publicly-available Python code, accessible at `https://github.com/ryan112358/private-pgm`. We chose these methods due to their prevalence, performance capabilities, and similarities to our proposed approach in terms of inputs in the sense that all comparison methods are created from marginal counts instead of the full, underlying table.

The codes for these two methods return different outputs. PrivBayes returns a synthetic data set, which we convert into probabilities for the full table. The graphical model-based approach returns a vector of numbers (not necessarily counts) summing to the sample size that corresponds to "counts" for each of the full table cells. We normalize these numbers to create probabilities for the full table. We note that, unlike our proposed latent class approach, neither of these methods generate uncertainty estimates.

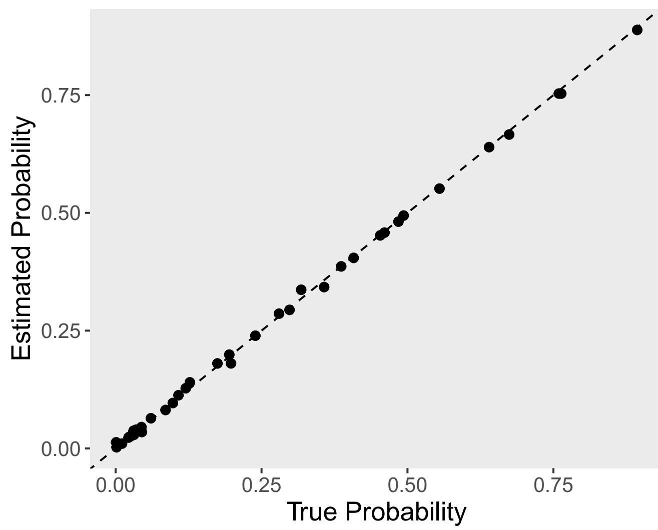
4.3. **Results.** Before turning to the results using differentially private marginals, we first study the usefulness of the composite likelihood approach and the corresponding assumption of independent counts absent privacy concerns, i.e., with $\epsilon = \infty$.

4.3.1. *Performance Absent Privacy Concerns.* We estimate the augmented latent class model for the underlying table without adding privacy-preserving noise to the marginal counts. To do, we fit model (3.1) using the true $(M_1, \ldots, M_T)$ and without the measurement error components in the first two lines of that model.

Figure 1 displays posterior modes from one run of the simulation for the marginal and full table probabilities. Results from additional runs are very similar. Both marginal and full cell probabilities are estimated accurately. The independence assumption appears to have minimal negative effects on estimation of the marginal probabilities, while also returning reasonable estimates for the full table cell probabilities. We note that this latter fact stems from the absence of strong three-way (and higher) interaction effects among the variables. In general, one should not expect accurate estimates for the full table; rather, one should expect accurate estimates of the marginal counts used in the estimation routines.

4.3.2. *Performance with Noisy Counts.* We next generate differentially private counts and corresponding posterior inferences for the data described in Section 4.1, using the algorithm described in Section 3.

[3] Due to potential identifiability issues due to label-switching Gelman et al. (2013), monitoring convergence of $(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$ is infeasible. Instead, we monitor estimated full and marginal probabilities.



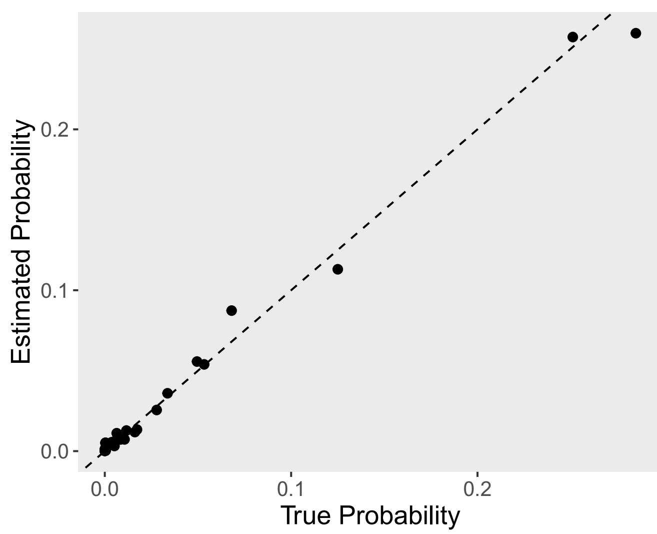


Figure 1: True versus estimated two-way marginal (top) and full table (bottom) probabilities for the ACS simulation with no noise added for privacy. True counts without added noise used as input. The number of latent classes is $k = 10$. As two-way margins were used as inputs, we see more accurate estimation of the marginal probabilities. Accurate estimation of the full cell probabilities is not guaranteed but can occur.

Figure 2 displays posterior means for the estimated marginal and full cell probabilities from one run of the simulation. Results from additional runs are very similar. The results exhibit some expected trends. As the privacy loss budget ϵ increases, the amount of added noise decreases, resulting in better estimates of marginal and full cell probabilities. In the case of marginal probabilities, results for all values of the privacy loss budget return estimates that coincide well with the underlying probabilities, as shown in Figure 2. The results demonstrate this behavior for all values of ϵ. For the full table probabilities, the proposed approach reasonably reconstructs the underlying probabilities even though full

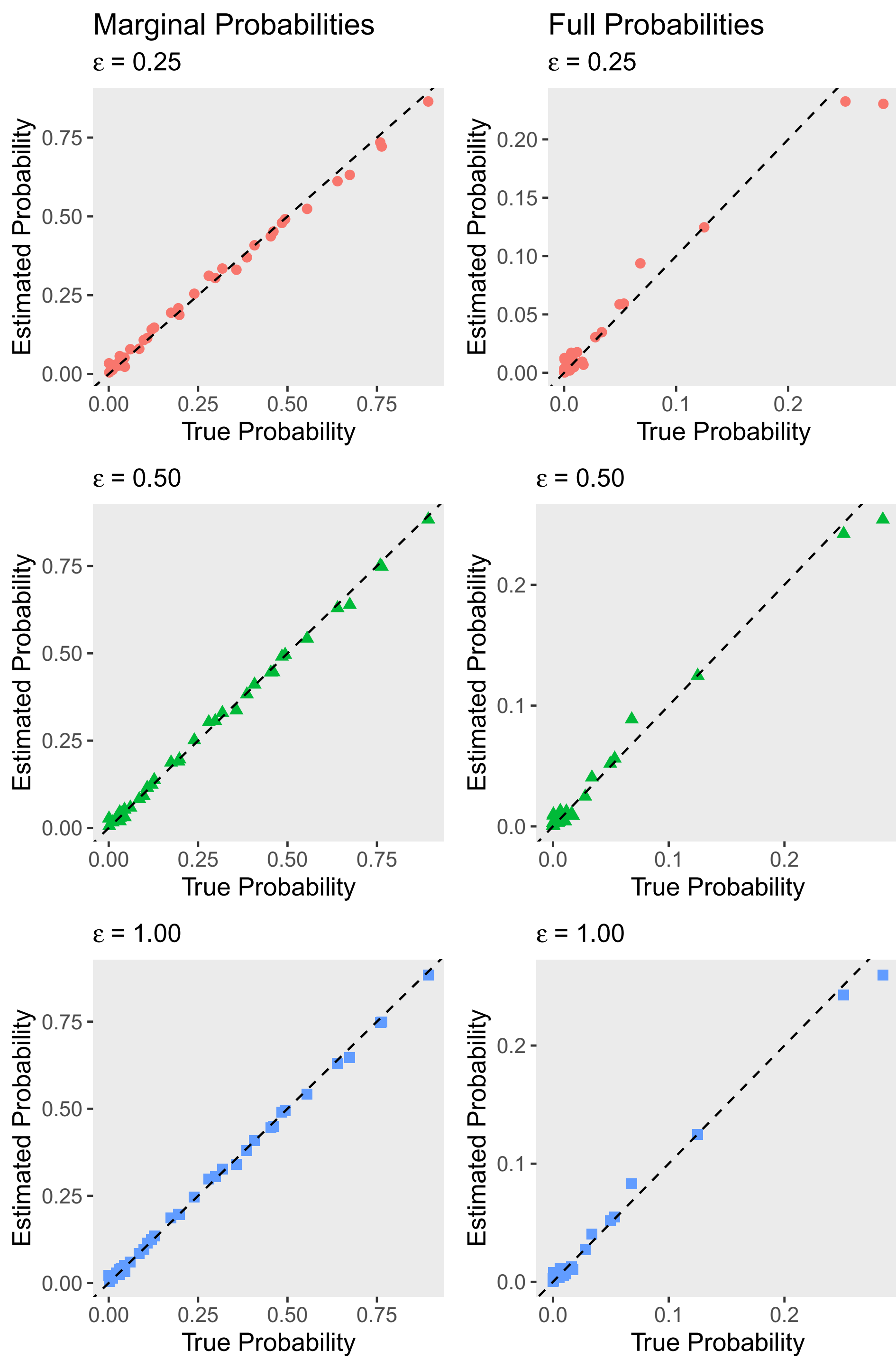


Figure 2: Plots of true marginal probabilities versus estimated marginal probabilities (left) and true full table probabilities versus estimated probabilities (right) for $\epsilon \in \{.25, .5, 1\}$. Estimated marginal probabilities are the posterior means of the corresponding marginal probabilities. As expected, estimation accuracy visually increases as ϵ grows.

Table 3: Average coverage probabilities ("Cov.") and average credible interval length ("Length") for the 40 two-way marginal counts.

	$\epsilon = .25$		$\epsilon = .5$		$\epsilon = 1$		$\epsilon = \infty$	
	Cov.	*Length*	*Cov.*	*Length*	*Cov.*	*Length*	*Cov.*	*Length*
Average:	0.839	0.049	0.896	0.041	0.945	0.038	0.969	0.038

table counts are not used in estimation. In general, there is more variation from the truth when compared to the estimates of the marginal probabilities. This also corresponds with the privacy budget: lower values of ϵ display more disagreement from the true values compared to higher values. This is reasonable to expect as only the marginal counts were used directly in estimation.

We also conducted repeated simulations to assess how well the procedure captures the underlying marginal counts for each value of ϵ. To do so, we repeat the entire procedure 100 times. Each time we sample 10,000 observations from the full original data, compute all two-way marginals, apply the Geometric Mechanism, and implement our method. For each marginal count, we examine the posterior distribution of the corresponding marginal probability and construct 95% credible intervals using its 2.5^{th} and 97.5^{th} quantiles. We record the coverage of the true probability, that is, whether or not the population count is inside the posterior interval.

Table 3 displays the averages of the coverage probabilities and average interval lengths for all 40 two-way probabilities. Results for each of the 40 cells are shown in the Supplementary Material; we summarize the results here. In general, the credible intervals are close to or exceed the nominal coverage rate when $\epsilon = \infty$, confirming the usefulness of the latent class approach absent privacy concerns. The coverage rates stay high for $\epsilon = 1$, with a few coverage rates for small probabilities dipping to around 70%. The coverage rates degrade noticeably when $\epsilon = .25$, especially for the small probabilities. Overall, we notice that coverage probabilities for cells corresponding to statistically insignificant terms in the log-linear model typically perform worse than other counts. This is because the noise has greater impact on small counts, and in a sample of size 10,000 many of these low probability cells have very few sampled cases. The undercoverage also relates, in part, to the sample size in our simulation design. We repeated the simulation using samples of 100,000 records and found coverage rates at or exceeding the nominal rates for all cell probabilities and all four values of ϵ.

4.3.3. *Comparisons to Other Methods.* We next compare the performance of our approach to PrivBayes and the graphical models approach. We ensure that all methods satisfy ϵ-DP with the same ϵ and use consistent definitions of sensitivity. For the graphical models approach, we use all two way marginals as inputs. For PrivBayes, we set the maximum degree of the Bayesian network to two. Results are shown in Figure 3 for one run of each method. Results for other runs are similar. In general, all approaches return reasonable estimates for the marginal probabilities. Again, quality is dependent of the level of privacy. For smaller values of ϵ, variability is most pronounced. While all methods offer accurate point estimates for $\epsilon = 1$, one benefit of the latent class modeling approach is that it generates posterior inferences that account for the measurement error introduced by the privacy protection.

Table 4: Description of NLTCS variables. All variables had two levels: "yes" and "no."

Variable	**Description**	**Counts by Level**
SCN_15_A_Y04	Problem eating by self	"yes": 416; "no": 15,220
SCN_15_B_Y04	Problem getting in/out of bed by self	"yes": 862; "no": 14,774
SCN_15_C_Y04	Problem getting in/out of chairs by self	"yes": 1,054; "no": 14,582
SCN_15_D_Y04	Problem walking inside without help	"yes": 1,763; "no": 13,873
SCN_15_E_Y04	Problem going outside without help	"yes": 2,435; "no": 13,201
SCN_15_F_Y04	Problem dressing without help	"yes": 929; "no": 14,707
SCN_15_G_Y04	Problem bathing without help	"yes": 1,358; "no": 14,278
SCN_15_H_Y04	Problem going to the bathroom without help	"yes": 842; "no": 14,794
SCN_15_I_Y04	Incontinence	"yes": 1,355; "no": 14,281
SCN_17_A_Y04	Prepare meals without help	"yes": 13,670; "no": 1,966
SCN_17_B_Y04	Do laundry without help	"yes": 13,465; "no": 2,171
SCN_17_C_Y04	Light housework without help	"yes": 13,853; "no": 1,783
SCN_17_D_Y04	Shop for groceries without help	"yes": 12,678; "no": 2,958
SCN_17_E_Y04	Manage money without help	"yes": 13,737; "no": 1,899
SCN_17_F_Y04	Take medicine without help	"yes": 14,031; "no": 1,605
SCN_17_G_Y04	Make phone calls without help	"yes": 14,251; "no": 1,385

PrivBayes, the graphical models, and our latent class modeling approach need not be viewed as mutually exclusive competitors; rather, they can be used together. For example, one can use PrivBayes to determine which marginal tables should be released and how to approximate the full table. One then could take the resulting noisy and incoherent marginal tables obtained through PrivBayes and apply the graphical models approach to create a coherent set of these marginal tables. One can use this coherent set as the starting point for $(M_1, \ldots, M_T)$ in the MCMC algorithm for the latent class approach. We illustrate these complementary uses in the next section.

5. Application to the National Long-Term Care Survey

In this section we illustrate the latent class approach on a large-scale dataset from the National Long Term Care Survey (NLTCS).

5.1. **Data Description.**

The NLTCS is a longitudinal survey sponsored by the U.S. National Institute on Aging[4]. First developed in 1982, the survey selects participants from Medicare enrollment files. All participants are over the age of 65, and new participants are added to the survey every five years (National Institute on Aging, 2021). Access to the data is restricted but can be obtained through an approved protocol (Manton, 2010). For this analysis, we restricted ourselves to the 2004 wave of the NLTCS. In addition, we focused on sixteen variables representing markers of daily living as defined in Table 4. Participants with missing or unknown values for any of these variables were dropped, resulting in a 2^{16} contingency table comprised of $n = 15,636$ observations.

[4]The NLTCS (National Long Term Care Study) is sponsored by the National Institute of Aging and was conducted by the Duke University Center for Demographic Studies under Grant No. U01-AG007198.

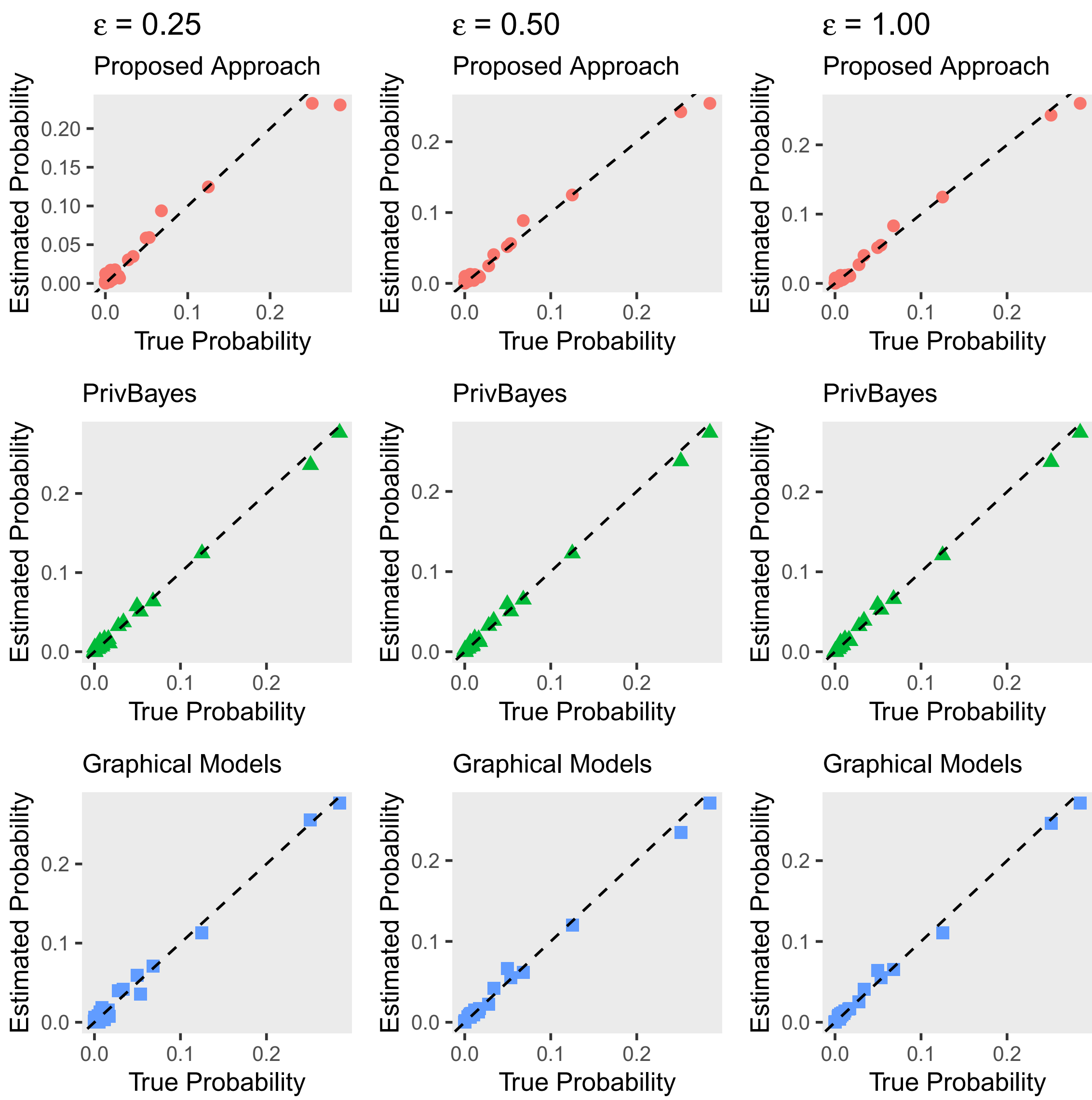


Figure 3: Plots of true full table probabilities versus estimated full table probabilities for $\epsilon \in \{.25, .5, 1\}$. Comparison of latent class modeling approach with a graphical models-based approach and PrivBayes for $\epsilon = .25$ (left), .5 (middle), and 1 (right). For the latent class modeling approach, estimated full cell probabilities denote the mean of the corresponding posterior distribution.

5.2. **Implementation Details.** The 16 binary variables make $\binom{16}{2} = 120$ two-way marginal tables and $120 \times (2 \times 2) = 480$ marginal counts. We selected important margins using the first half of the PrivBayes routine using a quarter of our privacy budget. While this allocates some of our privacy budget, it greatly reduces the number of marginals used in estimating

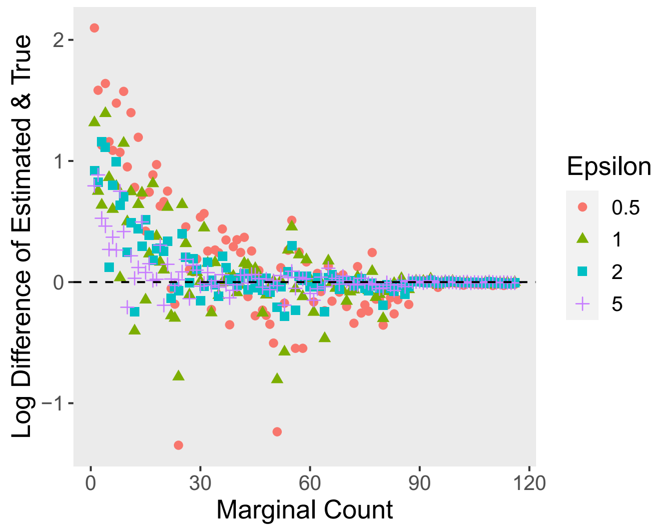


Figure 4: Plot of average log distance by marginal count for $\epsilon \in \{.5, 1, 2, 5\}$. A log distance of 0 indicates that the true and estimated probabilities are identical. Values greater than 0 imply that the estimated probability is larger than the truth. Marginal counts are arranged in ascending order with the smallest counts on the left.

model parameters. This resulted in 29 two-way margins. All counts were used for each pair, corresponding to $29 \times (2 \times 2) = 116$ marginal counts.

As before, we used a Metropolis-within-Gibbs algorithm to sample from the posterior distribution of $\boldsymbol{\pi_k}$, $\boldsymbol{\psi_k}$ and $(M_1, \ldots, M_k)$. To avoid overparametrization of the model, we set the number of latent classes to seven, ensuring that the number of estimated parameters (118) is roughly equal to the number of input marginals (116). We follow a similar implementation strategy as our ACS example; see the Supplementary Material for details. We run the chain for 12,000 iterations with the first 2,000 discarded as burn-in. We use four values of $\epsilon \in \{.5, 1, 2, 5\}$. We estimate the two-way marginal probabilities at each iteration using $\boldsymbol{\pi_k}$ and $\boldsymbol{\psi_k}$. As with the ACS, we ignore the complications from using survey weights and imputations for the NLTCS.

5.3. **Results.** For each of the 116 marginal counts, we compare the estimated and original (non-noisy) two-way probabilities. Figure 4 displays the log ratio (or difference e.g., $\log(P_{est}/P_{true})$) by ϵ. This metric is inspired by compositional data since the marginal probabilities are constrained to sum to one. Equivalence between estimates corresponds to a log ratio of zero. The leftmost estimates correspond to the smallest observed marginal counts whereas the rightmost estimates correspond the the largest observed marginal counts. The average log difference across all counts ranges from .173 for $\epsilon = .5$ to .050 for $\epsilon = 5$. The estimates with $\epsilon = .5$ are improved over using the noisy marginal counts directly: the noisy marginal counts had an average log difference of -.297 for $\epsilon = .5$. However, when $\epsilon = 5$, the results from the noisy counts are better, having an average log difference of .002.

Two observations are readily apparent from Figure 4. First, log differences appear to decrease as a function of ϵ with higher values corresponding to greater similarity between the estimated and true marginal probabilities. Second, log differences appear to be greater for smaller marginal probabilities regardless of ϵ.[5] This result is unsurprising when considering the structure of our model: the variance of the binomial distribution directly corresponds to the probability. In fact, for the sample size of the data, the variance at the minimum marginal probability is 46 times smaller than the corresponding variance at 25%. This effect is not as extreme at larger probabilities as the maximum marginal probability is approximately 92.5%.

Furthermore, the vast majority of the log differences at these lower counts are positive, indicating that the estimated marginal probability is greater than the true marginal probability. We believe that this result is amplified by the inherent non-negativity of the counts and the sum-to-one constraints for some model parameters (e.g., the mixing weights and related latent class probabilities).

The main benefit of the latent class modeling approach is the capability to facilitate posterior inferences and to generate synthetic data. For the former, we compute 95% credible intervals for each marginal probability by selecting the 2.5^{th} and 97.5^{th} quantiles of the posterior draws. To generate synthetic data, we randomly select an iteration of the MCMC chain, and use the set of parameter draws to sample 15,636 synthetic records. We compute various sums of these 15,636 synthetic records to estimate probabilities in the full table. We repeat this process for 20 sufficiently spaced iterations in the chain, resulting in 20 synthetic data sets. We compute the marginal probabilities in each synthetic data set and combine the estimates using the approach in Reiter (2003), resulting in 95% confidence intervals. Table 5 and Table 6 display the credible intervals and synthetic data inferences, respectively, for three randomly selected marginal sub-tables. The posterior intervals and synthetic data intervals are quite similar and typically capture the non-noisy estimate of the marginal probabilities, especially for larger values of ϵ.

6. Concluding Remarks

We present a novel method for posterior inferences and creating differentially private synthetic data for contingency tables based on marginal counts. The simulation results indicate that the latent class modeling approach can offer reasonably accurate estimates of the counts used as inputs to construct the underlying tables (in our example, two-way marginals). While it also did reasonably well in preserving counts from the full table in our illustrative data, analysts generally should not expect this to be the case. Rather than as a mechanism for generating data coherent with the full table probabilities, we view the approach as a means to make posterior inferences and generate synthetic data that are coherent with the marginal counts used as inputs. Thus, we advise agencies to tell data users what counts are used as inputs, so that data users do not expect the synthetic data to be accurate for other counts.

More generally, our approach is to define summary statistics as functions of model parameters, and use composite likelihood approximations to estimate those parameters. This general strategy can be extended to more complex data structures. For example, and as illustrated in the Supplementary Material, the strategy could be used to generate synthetic data for people nested within households, using the latent class model of Hu et al. (2018) as the underlying model. Developing methods for practical implementation in such complex models is an area for future research.

[5] In the true data, the smallest marginal probability is roughly .5%, and about half of all marginal probabilities are smaller than 5%.

Table 5: Posterior intervals for selected marginal counts in the NLTCS example. For each cell, we also present the non-noisy estimate ("Est.") and the noisy estimate obtained after adding geometric noise only; the latter is the input for our latent class modeling approach. Intervals are the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ quantile of the posterior distribution from the latent class modeling approach.

			$\epsilon = .5$		$\epsilon = 1$		$\epsilon = 2$	
Variable	Cell	Est.	*Noisy Est.*	*95% CI*	*Noisy Est.*	*95% CI*	*Noisy Est.*	*95% CI*
SCN_15_E SCN_17_D	00	.041	.033	(.053, .065)	.041	(.044, .052)	.041	(.040, .048)
	01	.114	.112	(.093, .107)	.117	(.106, .113)	.114	(.107, .116)
	10	.769	.776	(.762, .783)	.770	(.764, .774)	.768	(.768, .780)
	11	.075	.076	(.062, .080)	.075	(.070, .077)	.076	(.065, .076)
SCN_17_B SCN_17_C	00	.853	.846	(.839, .851)	.849	(.842, .849)	.854	(.853, .858)
	01	.008	.007	(.016, .033)	.010	(.013, .028)	.008	(.007, .011)
	10	.033	.039	(.032, .059)	.033	(.035, .043)	.033	(.030, .034)
	11	.106	.102	(.067, .102)	.107	(.089, .102)	.107	(.102, .106)
SCN_15_C SCN_17_A	00	.021	.018	(.013, .025)	.021	(.011, .019)	.021	(.015, .021)
	01	.046	.046	(.024, .044)	.036	(.034, .045)	.046	(.045, .050)
	10	.853	.864	(.846, .861)	.854	(.855, .866)	.852	(.853, .859)
	11	.080	.084	(.079, .105)	.085	(.079, .093)	.080	(.076, .081)

Table 6: Interval estimates based on 20 synthetic data sets for selected marginal counts in the NLTCS example. For each cell, we also present the non-noisy estimate ("Est.") and the noisy estimate obtained after adding geometric noise only; the latter is the input for our approach. Point estimates and 95% confidence intervals from the synthetic data computed using the combining rules of Reiter (2003).

			$\epsilon = .5$		$\epsilon = 1$		$\epsilon = 2$	
Variable	Cell	Est.	*Point Est.*	*95% CI*	*Point Est.*	*95% CI*	*Point Est.*	*95% CI*
SCN_15_E SCN_17_D	00	.041	.057	(.053, .061)	.044	(.041, .048)	.044	(.040, .047)
	01	.114	.080	(.076, .085)	.091	(.086, .096)	.107	(.102, .112)
	10	.769	.806	(.798, .815)	.793	(.786, .800)	.779	(.772, .786)
	11	.075	.056	(.050, .061)	.072	(.067, .076)	.070	(.066, .074)
SCN_17_B SCN_17_C	00	.853	.879	(.873, .886)	.869	(.864, .875)	.860	(.854, .865)
	01	.008	.017	(.014, .019)	.012	(.010, .015)	.008	(.007, .010)
	10	.033	.034	(.030, .039)	.030	(.027, .033)	.030	(.028, .033)
	11	.106	.070	(.061, .078)	.088	(.083, .093)	.102	(.097, .106)
SCN_15_C SCN_17_A	00	.021	.016	(.013, .018)	.011	(.008, .013)	.017	(.015, .019)
	01	.046	.023	(.018, .028)	.032	(.029, .036)	.046	(.042, .049)
	10	.853	.886	(.880, .892)	.882	(.876, .887)	.860	(.855, .866)
	11	.080	.076	(.070, .081)	.075	(.071, .080)	.077	(.073, .081)

In the NLTCS example, we show that the latent class model can be estimated from subsets of the full count; not all marginals are required. We leave it to future research to study methods for selecting the marginals, and, similarly, methods for allocating the privacy budget across these marginals. For both of our examples, we equally allocated the privacy budget. However, equal allocation may not be optimal in all problems.

Contingency tables can have very high dimensions and differing numbers of categories across variables. These can lead to potential sparsity issues. The simulation results suggest that, like many formally private algorithms, the latent class method has more difficulty with small counts. While conceptually our approach should work for larger tables, we leave full investigation to future research.

Acknowledgment

This work was supported by the Pennsylvania State University, Duke University, and the National Science Foundation under awards 1443014, 1144860, 1534433, and 1131897.

References

J. M. Abowd. The US Census Bureau adopts differential privacy. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2867–2867. ACM, 2018. https://doi.org/10.1145/3219819.3226070.

O. Akande, A. Barrientos, and J. P. Reiter. Simultaneous edit and imputation for household data with structural zeros. Journal of Survey Statistics and Methodology, 7(4):498–519, 2019a. https://doi.org/10.1093/jssam/smy022.

O. Akande, J. P. Reiter, and A. F. Barrientos. Multiple imputation of missing values in household data with structural zeros. Survey Methodology, 2:498–519, 2019b. https://hdl.handle.net/10161/17537.

J. Awan and A. Slavkovic. Structure and sensitivity in differential privacy: Comparing k-norm mechanisms. Journal of the American Statistical Association, 116(534):935–954, 2021. https://doi.org/10.1080/01621459.2020.1773831.

E. Bao, X. Xiao, J. Zhao, D. Zhang, and B. Ding. Synthetic data generation with differential privacy via Bayesian networks. Journal of Privacy and Confidentiality, 11(3), 2021. https://doi.org/10.29012/jpc.776.

B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In Proceedings of the Twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pages 273–282. ACM, 2007. https://doi.org/10.1145/1265530.1265569.

A. F. Barrientos, A. Bolton, T. Balmat, J. P. Reiter, J. M. de Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, M. DeLong, et al. Providing access to confidential research data through synthesis and verification: An application to data on employees of the US Federal Government. The Annals of Applied Statistics, 12(2):1124–1156, 2018. https://doi.org/10.1214/18-AOAS1194.

G. Benedetto, M. Stinson, and J. M. Abowd. The creation and use of the SIPP Synthetic Beta. Mimeo, U.S. Census Bureau, April 2013. http://hdl.handle.net/1813/43924.

C. M. Bowen and F. Liu. Comparative study of differentially private data synthesis methods. Statistical Science, 35(2):280–307, 2020. https://doi.org/10.1214/19-STS742.

C. M. Bowen and J. Snoke. Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. Journal of Privacy and Confidentiality, 11(1), 2021. http://doi.org/10.29012/jpc.748.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds, 2016.

G. Caiola and J. P. Reiter. Random forests for generating partially synthetic, categorical data. Transactions on Data Privacy, 3:27–42, 2010. http://www.tdp.cat/issues/tdp.a033a09.pdf.

A.-S. Charest. Empirical evaluation of statistical inference from differentially-private contingency tables. In International Conference on Privacy in Statistical Databases, pages 257–272. Springer, 2012. https://doi.org/10.1007/978-3-642-33627-0_20.

K. Chaudhuri, A. D. Sarwate, and K. Sinha. A near-optimal algorithm for differentially-private principal components. Journal of Machine Learning Research, 14:2905–2943, 2013. https://www.jmlr.org/papers/v14/chaudhuri13a.html.

J. Drechsler. Synthetic datasets for the German IAB Establishment Panel. Invited Paper WP.10, Joint UNECE/Eurostat work session on statistical data confidentiality, 2009. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.10.e.pdf.

J. Drechsler. Synthetic Datasets for Statistical Disclosure Control, volume 201 of Lecture Notes in Statistics. New York: Springer Science & Business Media, 2011. ISBN 9781461403258. https://doi.org/10.1007/978-1-4614-0326-5.

D. B. Dunson and C. Xing. Nonparametric Bayes modeling of multivariate categorical data. Journal of the American Statistical Association, 104(487):1042–1051, 2009. https://doi.org/10.1198/jasa.2009.tm08439.

C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In Annual International Conference on the Theory and Applications of Cryptographic Techniques, pages 486–503. Springer, 2006a. https://doi.org/10.1007/11761679_29.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference, pages 265–284. Springer, 2006b. https://link.springer.com/chapter/10.1007/11681878_14.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4):211–407, 2014. https://doi.org/10.1561/0400000042.

E. C. Eugenio and F. Liu. Construction of differentially private empirical distributions from a low-order marginals set through solving linear equations with l_2 regularization. In Intelligent Computing, pages 949–966. Springer, 2021. https://doi.org/10.1007/978-3-030-80129-8_63.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. The Annals of Statistics, pages 209–230, 1973. https://doi.org/10.1214/aos/1176342360.

S. E. Fienberg and A. B. Slavkovic. Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Mining and Knowledge Discovery, 11(2):155–180, 2005. https://doi.org/10.1007/s10618-005-0010-x.

W. A. Fuller. Measurement Error Models. New York: John Wiley & Sons, 2009. https://doi.org/10.1002/9780470316665.

S. L. Garfinkel, J. M. Abowd, and S. Powazek. Issues encountered deploying differential privacy. In Proceedings of the 2018 Workshop on Privacy in the Electronic Society, pages

133–137. ACM, 2018. https://doi.org/10.1145/3267323.3268949.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. Bayesian Data Analysis. Boca Raton, FL: CRC Press, 2013. https://doi.org/10.1201/b16018.

A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. SIAM Journal on Computing, 41(6):1673–1693, 2012. https://doi.org/10.1145/1536414.1536464.

R. Gong. Exact inference with approximate computation for differentially private data via perturbations. arXiv preprint arXiv:1909.12237, 2019. https://arxiv.org/abs/1909.12237.

M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In Advances in Neural Information Processing Systems, pages 2339–2347, 2012. https://proceedings.neurips.cc/paper/2012/file/208e43f0e45c4c78cafadb83d2888cb6-Paper.pdf.

M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. Proceedings of the VLDB Endowment, 3(1-2):1021–1032, 2010. https://doi.org/10.14778/1920841.1920970.

J. Hu, J. P. Reiter, and Q. Wang. Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. Bayesian Analysis, 13(1):183–200, 2018. https://doi.org/10.1214/16-BA1047.

A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf. Statistical Disclosure Control. New York: John Wiley & Sons, 2012. https://doi.org/10.1002/9781118348239.

S. Inusah and T. J. Kozubowski. A discrete analogue of the Laplace distribution. Journal of Statistical Planning and Inference, 136(3):1090–1102, 2006. https://doi.org/10.1016/j.jspi.2004.08.014.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association, 96(453):161–173, 2001. https://doi.org/10.1198/016214501750332758.

A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician, 60(3):224–232, 2006. https://doi.org/10.1198/000313006X124640.

V. Karwa, D. Kifer, and A. B. Slavković. Private posterior distributions from variational approximations. arXiv preprint arXiv:1511.07896, 2015. https://arxiv.org/abs/1511.07896.

V. Karwa, P. N. Krivitsky, and A. B. Slavković. Sharing social network data: Differentially private estimation of exponential family random-graph models. Journal of the Royal Statistical Society Series C, 66(3):481–500, 2017. https://doi.org/10.1111/rssc.12185.

S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. International Statistical Review, 79(3):362–384, 2011. https://doi.org/10.1111/j.1751-5823.2011.00153.x.

J. Lee, Y. Wang, and D. Kifer. Maximum likelihood postprocessing for differential privacy under consistency constraints. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 635–644, 2015. https://doi.org/10.1145/2783258.2783366.

B. Li, V. Karwa, A. Slavković, and R. C. Steorts. A privacy preserving algorithm to release sparse high-dimensional histograms. Journal of Privacy and Confidentiality, 8(1), 2018. https://doi.org/10.29012/jpc.657.

B. G. Lindsay. Composite likelihood methods. Contemporary Mathematics, 80(1):221–239, 1988. https://doi.org/10.1090/conm/080/999014.

R. J. A. Little. Statistical analysis of masked data. Journal of Official Statistics, 9(2):407–426, 1993. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. International Conference on Data Engineering (ICDE), pages 277–286, 2008. https://doi.org/10.1109/ICDE.2008.4497436.

D. Manrique-Vallier and J. P. Reiter. Bayesian estimation of discrete multivariate latent structure models with structural zeros. Journal of Computational and Graphical Statistics, 23(4):1061–1079, 2014. https://doi.org/10.1080/10618600.2013.844700.

K. G. Manton. National Long-Term Care Survey: 1982, 1984, 1989, 1994, 1999, and 2004, 2010. https://doi.org/10.3886/ICPSR09681.v5.

J. M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer. Fast generation of accurate synthetic microdata. In International Workshop on Privacy in Statistical Databases, pages 298–306. Springer, 2004. https://doi.org/10.1007/978-3-540-25955-8_24.

R. McKenna, D. Sheldon, and G. Miklau. Graphical-model based estimation and inference for differential privacy. In International Conference on Machine Learning, pages 4435–4444. PMLR, 2019. https://proceedings.mlr.press/v97/mckenna19a/mckenna19a.pdf.

R. McKenna, G. Miklau, and D. Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. Journal of Privacy and Confidentiality, 11, 2021. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/778.

F. D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, pages 19–30. ACM, 2009. https://doi.org/10.1145/1559845.1559850.

I. Mironov. Rényi differential privacy. In 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pages 263–275. IEEE, 2017. https://doi.org/10.1109/CSF.2017.11.

National Institute of Standards and Technology. 2018 Differential Privacy Synthetic Data Challenge, 2021. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic.

National Institute on Aging. National long term care survey (NLTCS), 2021. https://www.nia.nih.gov/research/resource/national-long-term-care-survey-nltcs.

M. P. Nixon. journalprivacyconfidentiality/bayeslcm: Bayeslcm. Software v20220722, Zenodo, July 2022. https://doi.org/10.5281/zenodo.6903667.

Y. Park and J. Ghosh. PeGS: Perturbed Gibbs samplers that generate privacy-compliant synthetic data. Transactions on Data Privacy, 7(3):253–282, 2014. https://www.tdp.cat/issues11/tdp.a154a13.pdf.

H. Ping, J. Stoyanovich, and B. Howe. DataSynthesizer: Privacy-preserving synthetic datasets. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pages 1–5, 2017. https://doi.org/10.1145/3085504.3091117.

G. Raab. Practical experience with making synthetic data differentially private, 2019. https://simons.berkeley.edu/talks/tba-47.

J. P. Reiter. Inference for partially synthetic, public use microdata sets. Survey Methodology, 29:181–188, 2003. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003002/article/6785-eng.pdf?st=c4i6vBqD.

J. P. Reiter. Estimating risks of identification disclosure in microdata. Journal of the American Statistical Association, 100(472):1103–1112, 2005a. https://doi.org/10.1198/016214505000000619.

J. P. Reiter. Using CART to generate partially synthetic public use microdata. Journal of Official Statistics, 21(3):441–462, 2005b. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/using-cart-to-generate-partially-synthetic-public-use-microdata.pdf.

J. P. Reiter. Differential privacy and federal data releases. Annual Review of Statistics and Its Application, 6:85–101, 2019. https://doi.org/10.1146/annurev-statistics-030718-105142.

J. P. Reiter, Q. Wang, and B. Zhang. Bayesian estimation of disclosure risks for multiply imputed, synthetic data. Journal of Privacy and Confidentiality, 6(1), 2014. https://doi.org/10.29012/jpc.v6i1.635.

Y. Rinott, C. M. O'Keefe, N. Shlomo, and C. Skinner. Confidentiality and differential privacy in the dissemination of frequency tables. Statistical Science, 33(3):358–385, 2018. https://doi.org/10.1214/17-STS641.

V. Roy. Convergence diagnostics for Markov chain Monte Carlo. Annual Review of Statistics and Its Application, 7:387–412, 2020. https://doi.org/10.1146/annurev-statistics-031219-041300.

D. B. Rubin. Discussion: Statistical disclosure limitation. Journal of Official Statistics, 9(2):461–468, 1993. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf.

J. Seeman, A. Slavkovic, and M. Reimherr. Private posterior inference consistent with public information: A case study in small area estimation from synthetic census data. In International Conference on Privacy in Statistical Databases, pages 323–336. Springer, 2020. https://doi.org/10.1007/978-3-030-57521-2_23.

Y. Si and J. P. Reiter. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. Journal of Educational and Behavioral Statistics, 38(5):499–521, 2013. https://doi.org/10.3102/1076998613480394.

A. Slavkovic and J. Seeman. Statistical data privacy: A song of privacy and utility, 2022.

A. Slavković, X. Zhu, and S. Petrović. Fibers of multi-way contingency tables given conditionals: Relation to marginals, cell bounds and markov bases. Annals of the Institute of Statistical Mathematics, 67(4):621–648, 2015. https://doi.org/10.1007/s10463-014-0471-z.

A. B. Slavković and J. Lee. Synthetic two-way contingency tables that preserve conditional frequencies. Statistical Methodology, 7(3):225–239, 2010. https://doi.org/10.1016/j.stamet.2009.11.002.

SLS-DSU. Scottish Longitudinal Study Development & Support Unit, 2018. https://sls.lscs.ac.uk/.

J. Snoke and A. Slavković. pMSE mechanism: Differentially private synthetic data with maximal distributional similarity. In International Conference on Privacy in

Statistical Databases, pages 138–159. Springer, 2018. https://doi.org/10.1007/978-3-319-99771-1_10.

J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic. General and specific utility measures for synthetic data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(3):663–688, 2018. https://doi.org/10.1111/rssa.12358.

J. Ullman and S. Vadhan. PCPs and the hardness of generating synthetic data. Journal of Cryptology, 33(4):2078–2112, 2020. https://doi.org/10.1007/s00145-020-09363-y.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. Statistica Sinica, pages 5–42, 2011. https://www3.stat.sinica.edu.tw/statistica/oldpdf/A21n11.pdf.

S. G. Walker. Bayesian inference with misspecified models. Journal of Statistical Planning and Inference, 143(10):1621–1633, 2013. https://doi.org/10.1016/j.jspi.2013.05.013.

Y. M. J. Woo and A. B. Slavković. Logistic regression with variables subject to post randomization method. In International Conference on Privacy in Statistical Databases, pages 116–130. Springer, 2012. https://doi.org/10.1007/978-3-642-33627-0_10.

S. D. Woodcock and G. Benedetto. Distribution-preserving statistical disclosure limitation. Computational Statistics & Data Analysis, 53(12):4228–4242, 2009. https://doi.org/10.1016/j.csda.2009.05.020.

X. Yang, S. E. Fienberg, and A. Rinaldo. Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. Journal of Privacy and Confidentiality, 4(1), 2012. https://doi.org/10.29012/jpc.v4i1.613.

J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: Regression analysis under differential privacy. Proceedings of the VLDB Endowment, 5(11):1364–1375, 2012. https://doi.org/10.14778/2350229.2350253.

J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. PrivBayes: Private data release via Bayesian networks. ACM Transactions on Database Systems, 42(4):1–41, 2017. https://doi.org/10.1145/3134428.