Disclosure Risk Evaluation for Fully Synthetic Categorical Data

Jingchen Hu, Jerome P. Reiter, and Quanli Wang[⋆]

Duke University, Durham NC 27708, USA

Abstract. We present an approach for evaluating disclosure risks for fully synthetic categorical data. The basic idea is to compute probability distributions of unknown confidential data values given the synthetic data and assumptions about intruder knowledge. We use a "worst-case" scenario of an intruder knowing all but one of the records in the confidential data. To create the synthetic data, we use a Dirichlet process mixture of products of multinomial distributions, which is a Bayesian version of a latent class model. In addition to generating synthetic data with high utility, the likelihood function admits simple and convenient approximations to the disclosure risk probabilities via importance sampling. We illustrate the disclosure risk computations by synthesizing a subset of data from the American Community Survey.

Keywords: Bayesian, confidentiality, Dirichlet process, disclosure, microdata

1 Introduction

Record-level data, also known as microdata, from the social, behavioral, and economic sciences offer enormous potential benefits to society. When made widely accessible as public use files, these databases facilitate advances in research and policy-making, enable students to develop skills at data analysis, and help ordinary citizens learn about their communities. However, as most stewards of social science data are acutely aware, wide-scale dissemination of microdata can result in unintended disclosures of data subjects' identities and sensitive attributes, thereby violating promises—and in some instances laws—to protect data subjects' privacy and confidentiality.

When microdata are highly sensitive or readily identifiable—as may be the case, for example, for business establishments or in large-scale administrative databases—stewards may not be able to protect confidentiality adequately by suppressing/perturbing only a small fraction of values (which is frequent practice in small-scale probability samples). In such contexts, one approach is to generate and release fully synthetic data (Rubin, 1993; Fienberg, 1994; Reiter, 2002, 2005b, 2009; Raghunathan *et al.*, 2003; Reiter and Raghunathan, 2007). These comprise entirely simulated records generated from statistical models designed to preserve important relationships in the confidential data. A related approach is to release partially synthetic data (Little, 1993; Reiter, 2003, 2004), in which only values deemed sensitive are replaced with simulated values.

The U.S. Census Bureau has adopted synthetic data as a dissemination strategy for several major data products, including the Survey of Income and Program Participation

[⋆] This research was supported by U.S. National Science Foundation grants CNS-10-12141 and SES-11-31897.

(Abowd *et al.*, 2006) and the Longitudinal Business Database (Kinney *et al.*, 2011). In both of these products, all but a handful of variables are replaced with values simulated from models estimated on the confidential data. Other examples of synthetic data applications have appeared in the literature as well (e.g., Kennickell, 1997; Abowd and Woodcock, 2001, 2004; Little *et al.*, 2004; Graham and Penny, 2005; An and Little, 2007; Hawala, 2008; Drechsler *et al.*, 2008a,b; Graham *et al.*, 2009; Machanavajjhala *et al.*, 2008; Drechsler and Reiter, 2010, 2012; Slavkovic and Lee, 2010; Wang and Reiter, 2012; Burgette and Reiter, 2013; Paiva *et al.*, 2014).

With fully synthetic data, disclosure risks generally are considered to be low—it is pointless to match fully synthetic records to records in other databases, since each fully synthetic record does not correspond to any particular individual. However, researchers have identified scenarios where full synthesis carries non-trivial disclosure risks (Abowd and Vilhuber, 2008; Charest, 2010; McClure and Reiter, 2012; Reiter *et al.*, 2014). Typically, these illustrative scenarios involve stylized data (e.g., a 2^4 contingency table) with simple synthesizers (e.g., a Dirichlet-multinomial distribution). To our knowledge, the literature does not include examples of quantified disclosure risks in fully synthetic data in realistic contexts.

In this article, we illustrate disclosure risk evaluations for fully synthetic, categorical data. In particular, we compute Bayesian posterior probabilities that intruders can learn confidential values given the released data and assumptions about their prior knowledge (Duncan and Lambert, 1989; Fienberg *et al.*, 1997; Reiter, 2005a; McClure and Reiter, 2012; Reiter, 2012; Abowd *et al.*, 2013; Reiter *et al.*, 2014). We synthesize a subset of data from the American Community Survey using a Dirichlet process mixture of products of multinomial (DPMPM) distributions. The DPMPM model has been shown in other contexts to be effective at capturing complex dependence structure in contingency tables while requiring little tuning by the data steward (Dunson and Xing, 2009; Si and Reiter, 2013; Manrique-Vallier and Reiter, 2014).

Our goal here is to illustrate the risk evaluations with realistic data. Thus, although we present some evaluations of data utility to assure readers that the DPMPM synthesizer is not generating worthless data, we refrain from making conclusions about the merits of using the DPMPM synthesizer, or fully synthetic data in general as compared to other disclosure protection methods.

2 The DPMPM Synthesizer

Let the confidential data D comprise n individuals measured on p categorical variables. For $i = 1, \dots, n$ and $k = 1, \dots, p$, let x_{ik} denote the value of variable k for individual i, and let $x_i = (x_{i1}, \dots, x_{ip})$. Without loss of generality, assume that each x_{ik} takes on values in $\{1, \dots, d_k\}$, where $d_k \geq 2$ is the total number of categories for variable k. Effectively, the survey variables form a contingency table of $d = d_1 \times d_2 \times \cdots \times d_p$ cells defined by cross-classifications of the p variables. Let X_{ik} and X_i be random variables defined respectively on the sample spaces for x_{ik} and x_i.

We generate synthetic data using a finite number of mixture components in the DPMPM. Paraphrasing from Si and Reiter (2013), the finite DPMPM assumes that each individual i belongs to exactly one of $F < \infty$ latent classes; see Si and Reiter (2013) for advice on determining F. For $i = 1, \ldots, n$, let $\eta_i \in \{1, \ldots, F\}$ indicate the class of individual i, and let $\pi_f = \Pr(\eta_i = f)$. We assume that $\pi = (\pi_1, \ldots, \pi_F)$ is the same for all individuals. Within any class, each of the p variables independently follows a class-specific multinomial distribution, so that individuals in the same latent class have the same cell probabilities. For any value $c \in \{1, \ldots, d_k\}$, let $\phi^{(k)}_{fc} = \Pr(X_{ik} = c \mid \eta_i = f)$ be the probability of $X_{ik} = c$ given that individual i is in class f. Let $\phi = \{\phi^{(k)}_{fc} : c = 1, \ldots, d_k, k = 1, \ldots, p, f = 1, \ldots, F\}$ be the collection of all $\phi^{(k)}_{fc}$. The finite mixture model can be expressed as

$$\begin{aligned} X_{ik} \mid \eta_i, \phi &\overset{ind}{\sim} \text{Multinomial}(\phi^{(k)}_{\eta_i 1}, \ldots, \phi^{(k)}_{\eta_i d_k}) \quad \text{for all } i, k & (1)\\ \eta_i \mid \pi &\sim \text{Multinomial}(\pi_1, \ldots, \pi_F) \quad \text{for all } i, & (2) \end{aligned}$$

where each multinomial distribution has sample size equal to one and the number of levels is implied by the dimension of the corresponding probability vector.

For prior distributions on π and ϕ, we use the truncated stick breaking representation of Sethuraman (1994). We have

$$\begin{aligned} \pi_f &= V_f \prod_{l<f} (1 - V_l) \quad \text{for } f = 1, \ldots, F & (3)\\ V_f &\overset{iid}{\sim} \text{Beta}(1, \alpha) \quad \text{for } f = 1, \ldots, F-1, \quad V_F = 1 & (4)\\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) & (5)\\ \phi^{(k)}_f = (\phi^{(k)}_{f1}, \ldots, \psi^{(k)}_{fd_k}) &\sim \text{Dirichlet}(a_{k1}, \ldots, a_{kd_k}). & (6) \end{aligned}$$

We set $a_{k1} = \cdots = a_{kd_k} = 1$ for all k to correspond to uniform distributions. Following Dunson and Xing (2009) and Si and Reiter (2013), we set $(a_\alpha = .25, b_\alpha = .25)$, which represents a small prior sample size and hence vague specification for the Gamma distribution. In practice, we find these specifications allow the data to dominate the prior distribution. We estimate the posterior distribution of all parameters using a blocked Gibbs sampler (Ishwaran and James, 2001; Si and Reiter, 2013).

We note that this model assumes no structural zeros in the data; that is, all cells in the implied contingency table have non-zero probability. See Manrique-Vallier and Reiter (2014) for variants of DPMPM models that allow structural zeros.

To generate one fully synthetic dataset of size n^*, we first sample a value of (α, π, ϕ) from the posterior distribution. Using the generated π, we sample values of $(\eta_1, \ldots, \eta_{n^*})$ independently from (2). Using the sampled ϕ, for each sampled η_i, where $i = 1, \ldots, n^*$, we then sample the ith synthetic record, $x^*_i = (x^*_{i1}, \ldots, x^*_{ip})$, from independent multinomial distributions with probabilities $\phi^{(k)}_{\eta_i}$ for each k. When n^* is the original sample size n, the synthesis can be conveniently implemented inside the blocked Gibbs sampler—after each Gibbs updating step, we simply sample and save draws of x^*_i for all n records. To create $m > 1$ synthetic datasets, one repeats this process m times, using approximately independent draws of parameters. Approximately independent draws can be obtained by using iterations that are far apart in the estimated MCMC chain.

Let $Z = (Z^{(1)}, \ldots, Z^{(m)})$ be a set of m synthetic categorical datasets under consideration for release by the data steward. In the remainder of the article, we assume that $n^* = n$, although this is not necessary.

3 Disclosure Risk Measure for the DPMPM

With fully synthetic data, disclosure risk metrics based on matching released and external records are generally not applicable, since there is no unique mapping of the rows in Z to the rows in D. Instead, we consider questions of the form: can intruders accurately infer from the synthetic data that some record with particular data values is in the confidential data? When the combination of values is unique in the population (or possibly just the sample), this question essentially asks if intruders can determine whether or not a specific individual is in the confidential data—this may count as a disclosure under some confidentiality protection laws.

3.1 Disclosure Risk Evaluation Strategy

To describe the disclosure risk evaluations, we follow the presentation of Reiter *et al.* (2014). Let x denote an arbitrary realization from the sample space of the contingency table formed by the p categorical variables; x can take on any of d possible values. We suppose that an intruder seeks to learn if a particular x is in D. Let A represent the information known by the intruder about records in D. Let S represent any information known by the intruder about the process of generating Z, for example meta-data indicating the values of F and (a_α, b_α) for the DPMPM synthesizer. Let X be a random variable representing the intruder's uncertain knowledge of whether or not x is in D, where the sample space of X is all possible values of x in the population. Given (Z, A, S), we assume the intruder seeks the Bayesian posterior distribution,

$$p(X = x \mid Z, A, S) = \frac{p(Z \mid X = x, A, S)p(X = x \mid A, S)}{\sum_{x \in \mathcal{U}} p(Z \mid X = x, A, S)p(X = x \mid A, S)} \tag{7}$$

$$\propto p(Z \mid X = x, A, S)p(X = x \mid A, S), \tag{8}$$

where $\mathcal{U}$ represents the universe of all feasible values of x. Here, $p(Z \mid X = x, A, S)$ is the likelihood of generating the particular set of synthetic data given that x is in the confidential data and whatever else is known by the intruder. The $p(X = x \mid A, S)$ can be considered the intruder's prior distribution on X based on (A, S).

Key to the computation of (7) are the assumptions about A and $p(X = x \mid A, S)$. In general, it is not possible for the data steward to know either. We evaluate risks assuming the intruder has very strong prior knowledge in A. In particular, we assume the intruder knows the values of x for all individuals in D except for some record i, also done by Abowd *et al.* (2013). To represent this version of A, we use $D_{-i} = \{x_j : j \neq i\}$. With $A = D_{-i}$, (7) effectively becomes the probability distribution of X_i, i.e., the intruder's distribution of x for the unknown record. For clarity, from now on we write (8) as

$$p(X_i = x \mid Z, D_{-i}, S) \propto p(Z \mid X_i = x, D_{-i}, S)p(X_i = x \mid D_{-i}, S). \tag{9}$$

In many cases, setting $A = D_{-i}$ is conservative, since in contexts involving random sampling from large populations intruders are unlikely to know D_{-i}. Nonetheless, risks deemed acceptable for $A = D_{-i}$ should be acceptable for a weaker A. We note that assuming the intruder knows all records but one is related to, but quite distinct from, the assumptions used in differential privacy (Dwork, 2006).

Intruders can use $p(X_i = x \mid Z, D_{-i}, S)$ to take guesses at the true value x_i. For example, the intruder can find the x that offers the largest probability, and use that as a guess of x_i. Similarly, data stewards can use $p(X_i = x \mid Z, D_{-i}, S)$ in disclosure risk evaluations. For example, for each $x_i \in D$, they can rank each x by its associated value of $p(X_i = x \mid Z, D_{-i}, S)$, and evaluate the rank at the truth, $x = x_i$. When the rank of x_i is high (close to 1, which we define to be the rank associated with the highest probability), the agency may deem that record to be at risk under the strong intruder knowledge scenario. When the rank of x_i is low (far from 1), the agency may deem the risks for that record to be acceptable.

When d is very large, computing the normalizing constant in (7) is impractical. To facilitate computation, we propose to dramatically reduce the support in (7). For any record i, we consider as feasible candidates only those x that differ from x_i in one variable, along with x_i itself; we call this space $\mathcal{R}_i$. Thus, for example, the restricted support of x for a $3 \times 5 \times 2$ table includes only eight possible cases, namely the original x_i and the $2 + 4 + 1$ cases obtained by changing one of the three variables. One can conceive of this support as mimicking an intruder who is knowledgable enough to be searching in neighborhoods near x_i.

When the support is $\mathcal{R}_i$, the resulting values of $p(X_i = x \mid Z, D_{-i}, S)$ for any $x \in \mathcal{R}_i$ are larger than when the support is $\mathcal{U}$. Similarly, when the support is $\mathcal{U}$ the rank of any x is no higher than the corresponding rank when the support is $\mathcal{R}_i$. In this way, restricting support to $\mathcal{R}_i$ results in a conservative ranking of the $x \in \mathcal{R}_i$. Thus, if a data steward determines that the rank of x_i (or any value of x) is acceptably low when using $\mathcal{R}_i$, it also will be acceptably low when using $\mathcal{U}$.

3.2 Computational methods for risk assessment with DPMPM

Let $\Theta = \{\pi, \phi\}$ denote parameters from the DPMPM synthesis model. For Z generated from the DPMPM synthesizer, we can write (9) as

$$\rho_i^x = c \left(\int p(Z \mid X_i = x, D_{-i}, S, \Theta) p(\Theta \mid X_i = x, D_{-i}, S) d\Theta \right) p(X_i = x \mid D_{-i}, S), \tag{10}$$

where c is a normalizing constant. The form of (10) suggests a Monte Carlo approach to estimate ρ_i. First, acting like an intruder, the data steward creates the plausible confidential dataset, $D_i^x = (X_i = x, D_{-i})$. Second, treating D_i^x as if it were the collected data, the data steward samples m values of Θ, i.e., for $l = 1, \ldots, m$, sample a $\Theta^{(l)}$ that could have generated $Z^{(l)}$. Third, for each $(Z^{(l)}, \Theta^{(l)})$, the data steward computes the probability of generating the released $Z^{(l)}$. Fourth, the data steward multiplies the m probabilities; see (12). The value of ρ_i^x is the average of this probability computed over many plausible draws of Θ.

Conceptually, to draw Θ replicates, the data steward could re-estimate the DPMPM model for each D_i^x. However, this would be computationally prohibitive if the data steward intends to examine many x across many records i. Instead, we suggest using the sampled values of Θ from $p(\Theta \mid D)$ as proposals for an importance sampling algorithm. As a brief review of importance sampling, suppose we seek to estimate the expectation of some function $g(\Theta)$, where Θ has density $f(\Theta)$. Further suppose that we have available a sample $(\Theta^{(1)}, \dots, \Theta^{(H)})$ from a convenient distribution $f^*(\Theta)$ that slightly differs from $f(\Theta)$. We can estimate $E_f(g(\Theta))$ using

$$E_f(g(\Theta)) \approx \sum_{j=1}^{H} g(\Theta^{(j)}) \frac{f(\Theta^{(j)})/f^*(\Theta^{(j)})}{\sum_{j=1}^{H} f(\Theta^{(j)})/f^*(\Theta^{(j)})}. \qquad (11)$$

We note that (11) only requires that $f(\Theta)$ and $f^*(\Theta)$ be known up to constants.

We implement importance sampling algorithms to approximate the integral in (10). By construction, we have

$$P(Z \mid D_i^x, S) = \prod_{l=1}^{m} P(Z^{(l)} \mid D_i^x, S), \qquad (12)$$

regardless of the exact values in D_i^x. Thus, for any proposed x, we can use importance sampling to approximate each $P(Z^{(l)} \mid D_i^x, S)$ and substitute the m resulting estimates in the product in (12).

Let $x_i^{*(l)} = (x_{i1}^{*(l)}, \dots, x_{ip}^{*(l)})$ be the ith record's values in synthetic dataset $Z^{(l)}$, where $i = 1, \dots, n^*$ and $l = 1, \dots, m$. For each $Z^{(l)}$ and any proposed x, we define the $g(\Theta)$ in (11) to equal $cP(Z^{(l)} \mid D_i^x, S)$. We approximate the expectation of each $g(\Theta)$ with respect to $f(\Theta) = f(\Theta \mid D_i^x, S)$. In doing so, for any sampled $\Theta^{(j)}$ we use

$$g(\Theta^{(j)}) = P(Z^{(l)} \mid D_i^x, S, \Theta^{(j)}) = \prod_{i=1}^{n} \left(\sum_{f=1}^{F} \pi_f^{(j)} \prod_{k=1}^{p} \phi_{f x_{ik}^{*(l)}}^{(k)(j)} \right). \qquad (13)$$

We set $f^*(\Theta) = f(\Theta \mid D, S)$, so that we can use H draws of Θ from its posterior distribution based on D. Let these H draws be $(\Theta^{(1)}, \dots, \Theta^{(H)})$. We note that one could use any D_i^x to obtain the H draws, so that intruders can use similar importance sampling computations. As evident in (1) and (2), the only differences in the kernels of $f(\Theta)$ and $f^*(\Theta)$ include (i) the components of the likelihood associated with record i and (ii) the normalizing constant for each density. Let $x = (c_1, \dots, c_p)$, where each $c_k \in (1, \dots, d_k)$, be a guess at X_i. After computing the normalized ratio in (11) and canceling common terms from the numerator and denominator, we are left with $P(Z^{(l)} \mid D_i^x, S) = \sum_{j=1}^{H} p_j q_j$ where

$$p_j = \prod_{i=1}^{n} (\sum_{f=1}^{F} \pi_f^{(j)} \prod_{k=1}^{p} \phi_{f x_{ik}^{(*l)}}^{(k)(j)}) \qquad (14)$$

$$q_j = \frac{\sum_{f=1}^{F} \pi_f^{(j)} \prod_{k=1}^{p} \phi_{f c_k}^{(k)(j)} / \sum_{f=1}^{F} \pi_f^{(j)} \prod_{k=1}^{p} \phi_{f x_{ik}}^{(k)(j)}}{\sum_{h=1}^{H} (\sum_{f=1}^{F} \pi_f^{(h)} \prod_{k=1}^{p} \phi_{f c_k}^{(k)(h)} / \sum_{f=1}^{F} \pi_f^{(h)} \prod_{k=1}^{p} \phi_{f x_{ik}}^{(k)(h)})}. \qquad (15)$$

We repeat this computation for each $Z^{(l)}$, plugging the m results into (12).

Finally, to approximate ρ_i^x, we compute (12) for each $x \in \mathcal{R}_i$, multiplying each resulting value by its associated $P(X_i = x \mid D_{-i}, S)$. In what follows, we presume an intruder with a uniform prior distribution over the support $x \in \mathcal{R}_i$. In this case, the prior probabilities cancel from the numerator and denominator of (7), so that risk evaluations are based only on the likelihood function for Z. We discuss evaluation of other prior distributions in the illustrative application, to which we now turn.

4 Illustrative Application

We create and evaluate fully synthetic data for a subset of $n = 10000$ individuals from the 2012 American Community Survey public use microdata sample for the state of North Carolina. The $p = 14$ variables are displayed in Table 1. These 14 variables make a contingency table with $d = 8709120$ cells. The 10000 individuals occupy 3523 of these cells. Of the 3523 observed combinations of x, 2394 appear once, 474 appear twice, and 186 appear three times in the sample. The most frequent combination is repeated 233 times. We note that this table is constructed not to include structural zeros.

Table 1: Variables used in the illustrative application. Data taken from the 2012 American Community Survey public use microdata samples. In the table, PR stands for Puerto Rico.

Variable	Categories
SEX	1 = male, 2 = female
AGEP	age of person: 1 = 18-29, 2 = 30-44, 3 = 45-59, 4 = 60+
RACE1P	1 = White alone, 2 = Black or African American alone, 3 = American Indian alone, 4 = other, 5 = two or more races, 6 = Asian alone
SCHL	1 = less than high school diploma, 2 = high school diploma or GED or alternative credential, 3 = some college, 4 = associate's degree or higher
MAR	1 = married, 2 = widowed, 3 = divorced, 4 = separated, 5 = never married
LANX	1 = speaks another language, 2 = speaks only English
WAOB	born in: 1 = US state, 2 = PR and US island areas, oceania and at sea, 3 = Latin America, 4 = Asia, 5 = Europe, 6 = Africa, 7 = Northern America
MIL	1 = active military duty at some point, 2 = military training for Reserves/National Guard only, 3 = never served in the military
WKL	1 = worked within the past 12 months, 2 = worked 1-5 years ago, 3 = worked over 5 years ago or never worked
DIS	1 = has a disability, 2 = no disability
HICOV	1 = has health insurance coverage, 2 = no coverage
MIG	1 = live in the same house (non movers), 2 = move to outside US and PR, 3 = move to different house in US or PR
SCH	1 = has not attended school in the last 3 months, 2 = in public school or college, 3 = in private school or college or home school
HISP	1 = not Spanish, Hispanic, or Latino, 2 = Spanish, Hispanic, or Latino

4.1 Some evidence of utility of the synthetic data

We generated $m = 5$ synthetic datasets by estimating and sampling from the DPMPM based on the 10000 cases in D. Before illustrating the disclosure risk evaluations, we present evidence that the DPMPM synthesizer generates useful data for this D. Here, we do not intend to offer an exhaustive investigation of data utility; rather, our purpose is to document that the resulting Z are potentially useful for analysis.

Figure 1 and Figure 2 display the joint distributions of (WKL, SCHL) and (HISP, RACE1P), respectively, using the sample percentages in D and the averages of the corresponding percentages in the $m = 5$ synthetic datasets. The DPMPM synthesizer preserves these two joint distributions quite closely. We found similar patterns for the marginal distributions of all variables and for joint distributions involving most other pairs of variables.

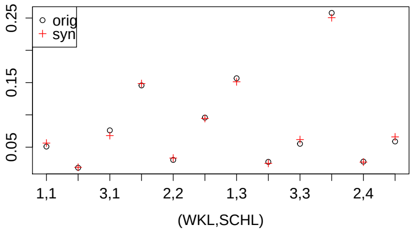

Fig. 1: Estimated joint probabilities for WKL and SCHL across the original and $m = 5$ synthetic datasets.

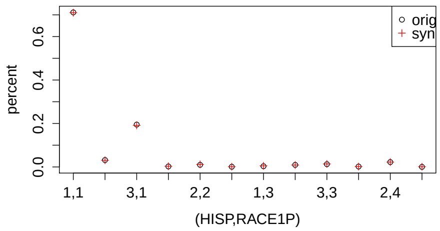

Fig. 2: Estimated joint probabilities for HISP and RACE1P across the original and $m = 5$ synthetic datasets.

Table 2: Point estimates and 95% confidence intervals for coefficients in a logistic regression of disability status on several main effects. Results estimated with the original data and the $m = 5$ generated synthetic datasets.

	Original data		Synthetic (m=5)	
Estimand	Estimate	95% CI	$\bar{q}_5$	95% CI
Intercept	2.212		2.108	[1.523,2.692]
SEX	-0.250		-0.221	[-0.361,-0.081]
MIL	0.239		0.205	[0.1255,0.2854]
MIG	0.049		0.060	[-0.0821,0.2014]
SCH	1.090		0.961	[0.6699,1.2521]
RACE1P	-0.078		-0.065	[-0.1147,-0.0145]
LANX	-1.096		-0.970	[-1.401,-0.539]

Table 2 summarizes the results of logistic regressions of DIS on SEX, MIL, MIG, SCH, RACE1P, and LANX. To estimate the coefficients, we use the maximum likelihood estimates (MLE) from D and the averages of the MLEs from the $m = 5$ synthetic datasets. The 95% confidence interval from the synthetic data derives from Raghunathan *et al.* (2003). Once again, the DPMPM offers reasonable results.

4.2 Disclosure risk assessments

Having demonstrated that Z has some analytic validity, we now turn to illustrating the assessment of disclosure risks. To do so, we drop each record in D one at a time. For each i, we compute the resulting ρ_i^x for all x in a reduced support $\mathcal{R}_i$. Here, each $\mathcal{R}_i$ is defined as the union of the true x_i plus the 34 other combinations of x obtained by changing x_i in one variable. For any two records i and j such that $x_i = x_j$ in D, $\rho_i^x = \rho_j^x$ for any possible x. Thus, we need only do computations for each of the 3523 combinations that appeared in the data. To compute each ρ_i^x, we use a uniform prior distribution over all $x \in \mathcal{R}_i$.

Figure 3 displays the distribution of the rank of the true x_i for each of the 3523 combinations. Here, a rank equal to 1 means the true x_i has the highest probability of being the unknown X_i, whereas a rank of 35 means the true x_i has the lowest probability of being the true X_i. Even armed with D_{-i}, the intruder gives the top rank to the true x_i for only two combinations and gives x_i a ranking in the top three for only 31 combinations; these are displayed in Table 3. We note that 2394 combinations were unique in D, yet evidently the DPMPM synthesizer involves enough smoothing that we do not recover the true x_i in the overwhelming majority of cases.

Figure 4 displays a histogram of the corresponding probabilities associated with the true x_i in each of the 3523 combinations. The largest probability is close to 0.2, and only 16 probabilities exceed 0.08. The majority of probabilities are in the 0.03 range. As we assumed a uniform prior distribution over the 35 possibilities in the support, the ratio of the posterior to prior probability is typically one or less, and only a handful of combinations have ratios exceeding two. Thus, compared to random guesses over a reasonably close neighborhood of the true values, Z typically does not provide much additional information about x_i. We note that high probabilities do not automatically result in top rankings.

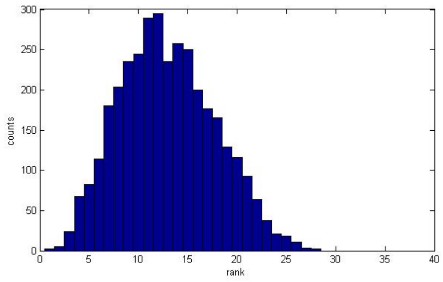

Fig. 3: Histogram of ranks of the probabilities associated with true x_i.

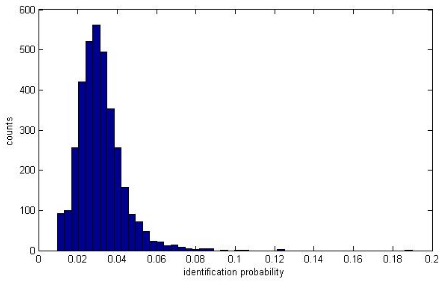

Fig. 4: Histogram of probabilities associated with the true x_i.

Table 3: The 31 combinations in D with the true x_i ranked in the top three highest posterior probabilities. "Reps." is the number of times x_i is repeated in D.

	Value for true x_i		
Combination	Reps.	Rank	Probability
(1,2,2,2,1,1,1,3,1,2,1,1,1,1)	1	1	.051
(2,4,3,2,2,2,1,3,2,2,2,1,1,1)	1	1	.088
(1,1,1,4,5,1,3,3,1,2,2,1,1,2)	1	2	.086
(1,1,5,3,5,2,1,1,2,2,2,3,2,1)	1	2	.070
(1,3,4,3,3,2,1,1,3,2,2,3,1,2)	2	2	.190
(2,1,5,3,5,1,1,3,1,2,1,1,2,1)	1	2	.076
(2,1,6,3,5,2,1,3,3,2,1,1,2,1)	1	2	.087
(1,1,1,3,1,2,1,3,3,2,2,1,1,1)	1	3	.060
(1,1,4,1,5,1,1,3,1,2,2,1,1,2)	1	3	.065
(1,2,1,4,3,2,1,1,3,2,2,3,2,1)	2	3	.125
(1,3,2,1,5,2,5,3,2,1,1,1,1,2)	1	3	.075
(1,3,2,3,3,2,5,1,2,2,2,1,3,1)	1	3	.069
(1,4,1,1,3,2,1,1,3,2,1,1,1,1)	2	3	.047
(1,4,4,4,1,1,3,3,1,2,1,1,1,2)	1	3	.066
(2,1,1,1,5,1,3,3,2,2,1,1,1,2)	1	3	.081
(2,1,1,2,1,2,1,3,3,2,1,1,1,1)	5	3	.046
(2,1,1,2,5,1,5,3,3,2,1,2,3,1)	1	3	.101
(2,1,1,4,3,2,1,2,3,2,1,3,1,1)	1	3	.085
(2,1,3,1,5,2,1,3,2,1,2,1,1,1)	1	3	.055
(2,1,3,3,5,2,1,3,2,2,2,1,1,1)	1	3	.050
(2,1,5,1,1,2,1,3,2,2,2,1,1,1)	1	3	.059
(2,1,5,3,5,1,1,3,1,1,1,3,2,1)	1	3	.066
(2,2,1,4,1,1,5,3,1,2,2,1,1,1)	1	3	.052
(2,2,2,4,5,2,4,3,1,2,1,3,2,1)	1	3	.079
(2,3,1,4,1,1,1,1,1,2,1,3,1,1)	1	3	.051
(2,3,1,4,1,1,5,3,1,2,1,1,1,1)	1	3	.053
(2,3,1,4,1,2,1,1,1,2,1,1,1,2)	1	3	.054
(2,4,1,3,1,1,4,3,2,1,1,1,1,1)	1	3	.121
(2,4,1,3,2,1,4,3,3,2,1,1,1,1)	1	3	.104
(2,4,1,4,2,2,1,3,2,1,2,1,1,1)	1	3	.065
(2,4,6,1,2,1,4,3,3,2,1,1,1,1)	1	3	.080

If desired, data stewards can evaluate risk probabilities for intruders possessing additional information about target records in D. For example, suppose the data steward defines each $x_i = (x_{i(1)}, x_{i(2)})$, where $x_{i(1)}$ is a subset of values known to the intruder (e.g., demographic variables) and $x_{i(2)}$ is the remaining subset of values unknown to the intruder (e.g., health variables). To evaluate risks for intruders seeking to estimate the distribution of $x_{i(2)}$, we define $A = (D_{-i} \cup x_{i(1)})$. Using obvious extensions of notation, we then estimate

$$p(X_{2(i)} = x_{(2)} \mid Z, A, S) \propto p(Z \mid X_{i(2)} = x_{(2)}, A, S)p(X_{i(2)} = x_{(2)} \mid A, S). \tag{16}$$

For some $X_{i(2)}$, the implied support may be small enough that one can evaluate the probability over $\mathcal{U}$ rather than $\mathcal{R}_i$, restricting both sets to cases with $x_{(1)} = x_{i(1)}$.

To illustrate these computations, we suppose that $x_{i(2)}$ includes when an individual last worked (WKL), their disability status (DIS), their health insurance coverage status (HICOV), and their mobility status (MIG); and, $x_{i(1)}$ includes all other variables. The sample space for $x_{i(2)}$ comprises $3{*}2{*}2{*}3 = 36$ possible values. We compute the probabilities in (16) for the particular combination $x_i = (2, 4, 1, 3, 1, 1, 4, 3, \mathbf{2}, \mathbf{1}, \mathbf{1}, \mathbf{1}, 1, 1)$, where the boldface indicates $x_{i(2)}$. This record is somewhat arbitrarily selected for illustration, although it is unique on the entire x_i. The true $x_{i(2)} = (2, 1, 1, 1)$ has probability 0.3695, which ranks first among the 36 cases. Evidently, an intruder armed with this much information can guess the true value for this record. For comparison, we repeated these computations for a person with $x_i = (1, 4, 3, 2, 1, 2, 1, 1, \mathbf{3}, \mathbf{1}, \mathbf{1}, \mathbf{1}, 1, 1)$. Here, the probability is 0.0259, which ranks 16 among the 36 cases.

The data steward also may want to evaluate the marginal distribution of $X_{2(i)}$, namely

$$p(X_{i(2)} = x_2 \mid Z, D_{-i}, S) \propto p(Z \mid X_{i(2)} = x_{(2)}, D_{-i}, S)p(X_{i(2)} = x_{(2)} \mid D_{-i}, S). \tag{17}$$

This allows the data steward to compute, for example, the probability associated with particular combinations of age, race, gender, and education for the ith record under the strong intruder knowledge scenario. To compute this efficiently, one approach is to sum the set of probabilities ρ_i^x where $\{x : x \in \mathcal{R}_i, x_{(2)} = x_{i(2)}\}$. Here, using $\mathcal{R}_i$ may be too restrictive, since by construction many of the x in $\mathcal{R}_i$ have $x_{(2)} = x_{i(2)}$. For this article, we did not investigate other approaches to defining $\mathcal{R}_i$ suitable for estimation of (17); this is a topic for future research.

Finally, data stewards need not use uniform distributions on the support of x for the intruder's prior distribution; the computations apply for any prior distribution. Naturally, the choice of prior distribution affects posterior probabilities, although for large m the posterior probabilities can be practically insensitive to specifications of reasonable prior distributions (Reiter *et al.*, 2014). We suggest that data stewards assess the effects of changing the prior distribution by comparing posterior probabilities for selected pairs of candidate x values under different prior distributions, specifically those for x_i and other candidate values of interest, say $x = b$. When $A = D_{-i}$, the ratio of the posterior probabilities for x_i and b is

$$\frac{p(X_i = x_i \mid Z, D_{-i}, S)}{p(X_i = b \mid Z, D_{-i}, S)} = \frac{p(Z \mid X_i = x_i, D_{-i}, S)}{p(Z \mid X_i = b, D_{-i}, S)} \frac{p(X_i = x_i \mid D_{-i}, S)}{p(X_i = b \mid D_{-i}, S)}. \tag{18}$$

Normalizing constants need not be computed in (18), since they cancel from the numerator and denominator. Hence, once the data steward has computed the likelihoods of Z for x_i and b, it can easily compute the ratio of the posterior probabilities for these two values for arbitrary sets of prior probabilities.

The ratios in (18) allow for convenient investigations of the effects of changing the prior probabilities. For example, suppose that the data steward considers records to be at too high risk if the posterior probability of the true x_i is ranked as the top case (or some other minimum ranking). Suppose that some x_i has the tenth highest posterior probability in $\mathcal{R}_i$ under the uniform prior distribution on $x \in \mathcal{R}_i$. Using the likelihoods from the uniform probability case, the data steward can determine the ratio of the prior probabilities that would change the posterior probability of x_i to become the top ranked. For example, our risk evaluations under the uniform prior assumption reveal that $x_i = (1, 1, 1, 1, 1, 2, 1, 3, 1, 2, 1, 1, 1, 1)$ has the tenth largest ρ_i^x among $x \in \mathcal{R}_i$, with posterior probability equal to 0.0398. The combination with highest probability in $\mathcal{R}_i$ is $b = (1, 1, 1, 1, 1, 2, 1, 3, 1, 2, 2, 1, 1, 1)$, with posterior probability equal to 0.0585. Thus, for an intruder to rank x_i as most likely, the intruder would need to believe *a priori* that $x = x_i$ is 1.46 $(0.0585/0.0398)$ times more likely than $x = b$.

5 Concluding Remarks

When certain x_i have relatively high posterior probabilities, and hence high ranks, data stewards have several options. If the number of cases that are too risky is small, the data steward may decide to release the synthetic data and accept the risks. With this action, the data steward effectively puts low probability on the existence of intruders who know D_{-i} for exactly the records at risk. Alternatively, the data steward could alter the synthesizer, for example by removing risky records from the data used to estimate the synthesis models. It is also prudent for the data steward to examine risks under other assumptions of intruder behavior. For risky cases, data stewards can augment the support of x used to compute posterior probabilities, for example by changing two variables at a time. Implicitly, using a bigger sample space mimics an intruder armed with less precise (but still very substantial) knowledge. The data steward also can use a sensible informative prior distribution to compute risks. For example, the data steward can base prior probabilities for each $x \in \mathcal{R}_i$ on a DPMPM model estimated on D_{-i}. If the resulting posterior probabilities do not differ much from the informative prior probabilities, then arguably releasing Z does not meaningfully increase the disclosure risks for this intruder.

Looking to future research, we see two key next steps in this approach to disclosure risk assessment. First, we would like to develop algorithms for exploring the full support of any X_i. With powerful computing and efficient code—these computations are embarassingly parallel—it may be feasible to compute normalizing constants over much if not all of $\mathcal{U}$. Alternatively, it may be possible to find high probability x via stochastic search algorithms. More complete explorations of $\mathcal{U}$ would allow for more accurate computations of $P(X_{i(2)} \mid Z, D_{-i}, S)$, which helps data stewards assess risks that the intruders learn subsets of key or sensitive variables. Second, we would like to relax the very strong, and perhaps unrealistic, assumption that the intruder knows every case but one. Conceptually, the path to do so is straightforward. If the intruder knows some subset of data D_A, the data steward considers all plausible values in the set $D - D_A$, and identifies sets with high probability of generating Z. Clearly, this approach is computationally very expensive. However, we conjecture that stochastic search algorithms might allow one to identify sets with high probability, and within those sets values of x that appear with regularity.

Bibliography

Abowd, J., Stinson, M., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Tech. rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at `http://www.census.gov/sipp/synth_data.html`.

Abowd, J. and Vilhuber, L. (2008). How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygun, eds., *Privacy in Statistical Databases*, 239–246. New York: Springer-Verlag.

Abowd, J. A., Schneider, M. J., and Vilhuber, L. (2013). Differential privacy applications to Bayesian and linear mixed model estimation. *Journal of Privacy and Confidentiality* **5**, 73–105.

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 215–277. Amsterdam: North-Holland.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 290–297. New York: Springer-Verlag.

An, D. and Little, R. (2007). Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A* **170**, 923–940.

Burgette, L. and Reiter, J. P. (2013). Multiple-shrinkage multinomial probit models with applications to simulating geographies in public use data. *Bayesian Analysis* **8**, 453–478.

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

Drechsler, J., Bender, S., and Rässler, S. (2008a). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy* **1**, 105–130.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008b). A new approach for disclosure control in the IAB Establishment Panel–Multiple imputation for a better data access. *Advances in Statistical Analysis* **92**, 439 – 458.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association* **105**, 1347–1357.

Drechsler, J. and Reiter, J. P. (2012). Combining synthetic data with subsampling to create public use microdata files for large scale surveys. *Survey Methodology* **38**, 73–79.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7**, 207–217.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**, 1042–1051.

Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages, and Programming, part II*, 1–12. Berlin: Springer.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Tech. rep., Department of Statistics, Carnegie-Mellon University.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics* **13**, 75–89.

Graham, P. and Penny, R. (2005). Multiply imputed synthetic data files. Tech. rep., University of Otago, http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm.

Graham, P., Young, J., and Penny, R. (2009). Multiply imputed synthetic data: Evaluation of hierarchical Bayesian imputation models. *Journal of Official Statistics* **25**, 245–268.

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 161–173.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. Washington, D.C.: National Academy Press.

Kinney, S., 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. *International Statistical Review* **79**, 363–384.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Little, R. J. A., Liu, F., and Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 141–152. New York: John Wiley & Sons.

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

Manrique-Vallier, D. and Reiter, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with strutural zeros. *Journal of Computational and Graphical Statistics* to appear.

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

Paiva, T., Chakraborty, A., Reiter, J. P., and Gelfand, A. E. (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine* to appear.

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

Reiter, J. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *International Statistical Review* **77**, 179–195.

Reiter, J. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2005a). Estimating identification risks in microdata. *Journal of the American Statistical Association* **100**, 1103–1113.

Reiter, J. P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2012). Discussion: Bayesian perspectives and disclosure risk assessment. *International Statistical Review* **80**, 373–375.

Reiter, J. P., Wang, Q., and Zhang, B. (2014). Bayesian estimation of disclosure risks in multiply imputed, synthetic data. *Journal of Privacy and Confidentiality* to appear.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Si, Y. and Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics* **38**, 499–521.

Slavkovic, A. B. and Lee, J. (2010). Synthetic two-way contingency tables that preserve conditional frequencies. *Statistical Methodology* **7**, 225–239.

Wang, H. and Reiter, J. P. (2012). Multiple imputation for sharing precise geographies in public use data. *Annals of Applied Statistics* **6**, 229–252.