Journal of Privacy and Confidentiality (2014) **6**, Number 1, 17–33

Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data

Jerome Reiter*, Quanli Wang†, and Biyuan E. Zhang‡

1 Introduction

Many national statistical agencies, survey and research organizations, and businesses—henceforth all called agencies—collect data that they intend to share with others. These agencies strive to release data that (i) protect the confidentiality of data subjects' identities and sensitive attributes, (ii) are informative for a wide range of analyses, and (iii) are relatively straightforward for secondary data analysts to use. Most strategies for meeting these three criteria involve altering data values, such as suppressing values, aggregating variables, swapping values across records [10], and adding random noise to data values [20]. An alternative to releasing datasets is to release perturbed results of user-specified queries [16]; we consider only dataset releases here.

As the threats to confidentiality grow, aggregation and perturbation techniques may have to be applied with high intensity to ensure adequate protection. However, applying these methods with high intensity can have serious consequences for secondary statistical analyses. For example, aggregation of geography to high levels disables small area estimation and hides spatial variation; top-coding (reporting all values exceeding a threshold c as "value exceeds c") eliminates learning about tails of distributions—which are often most interesting—and degrades analyses reliant on entire distributions [25]; swapping at high rates destroys correlations among swapped and not swapped variables [14, 50]; and, adding random noise introduces measurement error that distorts distributions and attenuates correlations [20]. In fact, Elliott and Purdam [17] use the public use files from the U. K. census to show empirically that the quality of statistical analyses can be degraded even when using recoding, swapping, or stochastic perturbation at modest intensity levels. These problems only would get worse with high intensity applications.

The Census Bureau's Longitudinal Business Database (LBD), which contains annual total payroll and employee size since 1975 for every U. S. business establishment with paid employees, is an informative case study on the challenges of data dissemination [27]. Because the LBD is subject to Title 13 and Title 26 of the U. S. code, no actual values for individual establishments in the LBD can be released to the public; even the fact that an establishment filed taxes—and hence is in the dataset—is protected. Thus, top-coding cannot be used on monetary data as a large fraction of exact values would be released.

*Department of Statistical Science, Box 90251, Duke University, Durham, NC, mailto:jerry@stat.duke.edu.

†Department of Statistical Science, Box 90251, Duke University, Durham, NC, mailto:quanli@stat.duke.edu.

‡Department of Economics, Duke University, Durham, NC, mailto:bz30.duke@gmail.com.

This also suggests that swapping would have to be done at an extremely high rate, in which case the released data would be useless for any analysis involving relationships with swapped variables. Furthermore, the core variables of interest to researchers and policy makers, number of employees and total payroll, have highly skewed distributions even within industry classifications. The amount of added noise necessary to disguise these observations would have to be very large, resulting in data of limited usefulness.

When large fractions of values must be altered to protect confidentiality, as is the case in the LBD, agencies instead can replace sensitive values with multiple draws from statistical models designed to preserve important relationships in the confidential data. This approach is known in the statistical community as multiply-imputed, synthetic data [46, 18, 28, 35, 36, 37, 39, 40, 41]. Synthetic data come in two flavors: fully and partially synthetic data. In fully synthetic data, every value on the file is replaced with draws from the synthesis model. In partially synthetic data, only collected values deemed sensitive are replaced with simulated values. With either flavor, analysts can obtain valid inferences for wide classes of estimands by combining standard likelihood-based or survey-weighted estimates with simple formulas; analysts need not learn new statistical methods or software to adjust for the effects of the disclosure limitation treatments [34, 36, 37, 45]. This is true even for high fractions of replacement, whereas swapping high percentages of values or adding noise with large variance can destroy much of the data utility. The released data can include simulated values in the tails of distributions (no top-coding) and avoid category collapsing. Because many quasi-identifiers can be simulated, finer details of geography can be released [49, 8]. Finally, the method is flexible: synthesis can be targeted to particular values for at-risk records [14], to entire variables, or to the entire dataset. Because of these potential benefits, 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 [1], the American Community Survey group quarters data [23], the OnTheMap origin-destination data [31], and the LBD [27]. Other examples of synthetic data applications have appeared in the literature [e.g., 26, 3, 4, 29, 21, 5, 11, 12, 22, 48].

Despite these applications, there has been little research on methods for assessing the disclosure risks inherent in releasing model-based, synthetic data. For partially synthetic data, Reiter and Mitra [44] and Drechsler and Reiter [13] present risk measures for an intruder who knows the collected values of a single target record and searches the released data to identify that record. These approaches do not apply to fully synthetic data, nor do they account for intruders with knowledge about multiple records. An alternative approach that is feasible in some settings is to generate synthetic data to satisfy, at least approximately, some variant of differential privacy [e.g., 6, 7, 31, 2, 9, 24].

In this article, we present a generic framework for estimating disclosure risks in model-based, synthetic data, fleshing out an approach outlined by Reiter [42] and Wang and Reiter [49]. The basic approach is as follows. Motivated by—although quite distinct from—risk assessments via differential privacy, we create risk measures for an intruder who knows the true values of sensitive data for all records in the original database except for one. The intruder evaluates the posterior distribution of possible original values for the one unknown record, given the released synthetic data and information

about the data generation mechanism. The intruder uses values with high probability as reasonable guesses at the unknown true values. We illustrate such computations for fully synthetic data using simulations based on a 2^4 contingency table (Section 3) and for partially synthetic data using the Survey of Youth in Custody (Section 4).

2 Disclosure Risk Measures

Let (x_i, y_i) be a p-dimensional vector of numeric values for some record i, where x_i includes non-sensitive values (if any) that the agency releases without alteration and y_i includes confidential values subject to synthesis. Let $D = \{(x_i, y_i) : i = 1, \ldots, n\}$ be the $n \times p$ matrix comprising the database of interest, with direct identifiers (e.g., name, address, tax number) removed. The goal is to release a version of D that has acceptable disclosure risks and supports a wide range of valid analyses. Let $Z^{(l)}$ be one synthetic dataset, constructed so that all confidential values are replaced with simulations from models estimated with D; see Section 3 and Section 4 for examples of data synthesizers. We assume that the data owner disseminates $m > 1$ synthetic replicates, $Z = (Z^{(1)}, \ldots, Z^{(m)})$, so as to enable the user to estimate uncertainty appropriately [45].

We suppose that an intruder seeks to learn the value of y_i for some record i 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, code for the synthesizer or descriptions of the synthesis models. Let Y_i denote the random variable representing the intruder's uncertain knowledge of y_i, where the sample space of Y_i is all possible values of y in the population. Given (Z, X, A, S)—note that intruders who see Z can determine $X = \{x_i : i = 1, \ldots, n\}$ when $m > 1$—we assume the intruder seeks the Bayesian posterior distribution for Y_i [15, 19, 38, 32, 42], namely

$$p(Y_i \mid Z, X, A, S) \propto p(Z \mid Y_i, X, A, S) p(Y_i \mid X, A, S). \tag{1}$$

Here, $p(Y_i \mid X, A, S)$ is the intruder's prior distribution on y_i based on (X, A, S), and Z serves to sharpen the intruder's prior beliefs about Y_i. Essentially, the intruder guesses at the true y_i according to the prior beliefs. Guesses that result in relatively low probability of generating Z (given X, A, and S) are downweighted compared to guesses that result in relatively high probability of generating Z.

Of course, agencies cannot know any particular intruder's prior beliefs. Instead, agencies can adopt the recommendation of Skinner [47] and evaluate risks under reasonable prior distributions. For example, the agency can use a uniform distribution over the sample space of y_i to reflect vague prior knowledge. Alternatively, the agency can use a sensible predictive model, for example the one used in S. We discuss the impact of prior distributions further in Section 3 and Section 4.

Similarly, it is impossible for the agency to know the auxiliary information possessed by intruders. One approach, which we adopt here, is to evaluate risks under a "strong intruder knowledge" scenario by assuming that the intruder knows y_i for all individuals

except one. We call this set $D_{-i} = \{\{(x_j, y_j), \text{for } j \neq i\} \cup x_i\}$. In many contexts, 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 weaker A.

The agency can compute many disclosure risk measures based on $p(Y_i \mid Z, X, D_{-i}, S)$. For example, for $i = 1, \ldots, n$ and discrete Y_i, the agency can compute

$$R_i = I(argmax_y p(Y_i = y|Z, X, D_{-i}, S) = y_i) \tag{2}$$

and decide if the percentage of correct guesses, $R = \sum_{i=1}^{n} R_i/n$, is acceptably low. The agency also could examine $p(Y_i = y_i|Z, X, D_{-i}, S)/p(Y_i = y_i|X, D_{-i}, S)$ to examine the multiplicative increase in the disclosure probability attached to the true value (perhaps restricted to cases where $R_i = 1$). For (approximately) continuous Y_i, the agency could compute the expected difference,

$$E_i = \int yp(y|Z, X, D_{-i}, S)dy - y_i \tag{3}$$

for all i and decide if the distances are sufficiently large. In what follows, we focus on illustrating R_i for discrete data.

As noted by a reviewer, some agencies may view setting $A = D_{-i}$ as overly conservative, believing it unrealistic that intruders would have this amount of knowledge. In such cases, high values of risk measures like R should not automatically deter agencies from releasing the synthetic data. Nonetheless, the measures still offer a type of "upper bound" on the disclosure risks, both for individual records and the entire file, given the posited assumptions about intruder behavior.

3 Simulations with Fully Synthetic Data

We use a simple simulation scenario that illustrates many of the main issues: protecting a 2^4 binary table with fully synthetic data. For $i = 1, \ldots, 1000 = n$, let $y_i = (y_{1i}, y_{2i}, y_{3i}, y_{4i})$ comprise four binary variables. Let each of the $K = 16$ possible combinations be denoted c_k, where $k = 1, \ldots, 16$. Let $c_{16} = (0, 0, 0, 0)$, and let $C_{-16} = (c_1, \ldots, c_{15})$. We generate an observed dataset D as follows. For $i = 1, \ldots, n-1 = 999$, sample y_i from a multinomial distribution such that $p(y_i = c_k) = 1/15$ for all $c_k \in C_{-16}$. Set $y_{1000} = c_{16}$. Since we do full synthesis, $X = \emptyset$.

With this design, we create a record that is guaranteed to be unique in the sample. Intuitively, we expect such records potentially to face higher risks, since they can offer information to the synthesis model that is not available from other records. Whether or not this is true depends on the nature of the synthesizer; to illustrate this, we examine results for different types of synthesizers, which we now describe. We emphasize that we select the three synthesizers only to illustrate how the risk measure in (1) depends on the synthesizer; we do not intend to study which synthesizer is optimal, which is clearly a data-specific question.

3.1 Three Synthesizers

We implement three approaches to generating fully synthetic datasets. The first is a Dirichlet-multinomial synthesizer. The second and third involve smoothing assumptions in the form of sequences of logistic regressions.

Dirichlet-Multinomial Synthesizer

With low-dimensional tables, it is feasible to estimate and generate data from multinomial distributions,

$$y_i \mid \theta \sim \text{Multinomial}(1, \theta_1, \ldots, \theta_K). \tag{4}$$

Here $K = 16$, $\theta_k = p(Y = c_k \mid \theta)$, and $\theta = (\theta_1, \ldots, \theta_K)$. In fully synthetic data contexts [34], we assume that we do not know θ (e.g., D is a sample from a population) and so we use the posterior predictive distribution of Y to generate data. This requires specification of a prior distribution for θ. We consider two prior distributions, $\theta \sim \text{Dirichlet}(a, \ldots, a)$, where $a \in \{.0001, 1\}$. Note that this is not the intruder's prior distribution for unknown Y_i; rather, it is set by the agency to facilitate the synthesis.

Formally, we draw synthetic values using a two part process. First, we sample a value of θ from $p(\theta|D) \sim \text{Dirichlet}(n_1 + a, ..., n_{16} + a)$, where each n_k denotes the number of incidences of c_k in D. Second, using the sampled value of θ, we sample synthetic values $(y_1^*, \ldots, y_n^*)$ from (4) where each record y_i^* is sampled independently. We repeat these two steps $m = 5$ times to release five synthetic datasets.

Logistic Regression Synthesizers

The Dirichlet-Multinomial synthesizer includes one parameter for each of the sixteen cells in the implied contingency table. Alternatively, we might synthesize from models with fewer parameters. This can be essential in settings with many variables, as the number of parameters can become impractically large. A convenient way to implement such smoothing is to use a sequence of conditional models; that is, we write the joint distribution as

$$p(y_{1i}, y_{2i}, y_{3i}, y_{4i}) = p(y_{1i})p(y_{2i} \mid y_{1i})p(y_{3i} \mid y_{1i}, y_{2i})p(y_{4i} \mid y_{1i}, y_{2i}, y_{3i}). \tag{5}$$

We impose parameter restrictions on the conditional models to effect smoothing.

We implement a smoothed synthesizer by using logistic regressions as the conditional models and setting high-order interaction terms in the models to zero. Specifically, we generate synthetic data as follows:

1. For $i = 1, \ldots, n$, draw synthetic y_{1i}^* using a Dirichlet-Binomial synthesizer. This is the same process as the Dirichlet-Multinomial synthesizer with $K = 2$ outcomes.
2. Using all n records in D, estimate the logistic regression, $logit(p(y_{2i})) = (1, y_{1i})\beta_2$, where β_2 is a 2×1 vector of coefficients. Draw a value of $\beta_2|D \sim N_2(\hat{\beta}_2, \hat{V}(\hat{\beta}_2))$,

where $\hat{\beta}_2$ is the maximum likelihood estimate (MLE) of β_2 and $\hat{V}(\hat{\beta}_2)$ is MLE of its covariance-variance matrix. We use the drawn value of β_2 and the previously generated y^*_{1i} to compute predicted probabilities based on the logistic regression equation, which we then use in Bernoulli draws to get synthetic y^*_{2i}.

3. Using all n records in D, estimate the logistic regression, $logit(p(y_{3i})) = (1, y_{1i}, y_{2i}, y_{1i}y_{2i})\beta_3$, where β_3 is a 4×1 vector of coefficients. Draw a value of $\beta_3|D \sim N_4(\hat{\beta}_3, \hat{V}(\hat{\beta}_3))$, where $\hat{\beta}_3$ is MLE of β_3 and $\hat{V}(\hat{\beta}_3)$ is the MLE of its covariance-variance matrix. We use the drawn value of β_3 and the previously generated (y^*_{1i}, y^*_{2i}) to compute predicted probabilities based on the logistic regression equation, which we then use in Bernoulli draws to get synthetic y^*_{3i}.

4. Using all n records in D, estimate the logistic regression, $logit(p(y_{4i})) = (1, y_{1i}, y_{2i}, y_{3i}, y_{1i}y_{2i}, y_{1i}y_{3i}, y_{2i}y_{3i})\beta_4$, where β_4 is a 7×1 vector of coefficients. Draw a value of $\beta_4|D \sim N_7(\hat{\beta}_4, \hat{V}(\hat{\beta}_4))$, where $\hat{\beta}_4$ is MLE of β_4 and $\hat{V}(\hat{\beta}_4)$ is the MLE of its covariance-variance matrix. We use the drawn value of β_4 and the previously generated $(y^*_{1i}, y^*_{2i}, y^*_{3i})$ to compute predicted probabilities based on the logistic regression equation, which we then use in Bernoulli draws to get synthetic y^*_{4i}.

We call this the no-three-way interactions synthesizer (N3WI), since none of the logistic regressions include three-way interaction effects. We note that the DM synthesizer can be expressed as the sequence of logistic regressions in steps (1-(3) plus a logistic regression in step (4) that additionally includes the three-way interaction $y_{1i}y_{2i}y_{3i}$ as a predictor; thus, the N3WI synthesizer represents only a modest amount of smoothing compared to the DM synthesizer.

We also consider more substantial smoothing in the form of the no-two-way interactions synthesizer (N2WI). This synthesizer uses the same steps (1) and (2), but replaces (3) and (4) with main effects only logistic regression models. Specifically, we use

(3') Using all n records in D, estimate the logistic regression, $logit(p(y_{3i})) = (1, y_{1i}, y_{2i})$ β_3. Draw a value of $\beta_3|D \sim N_3(\hat{\beta}_3, \hat{V}(\hat{\beta}_3))$, where $\hat{\beta}_3$ is MLE of β_3 and $\hat{V}(\hat{\beta}_3)$ is the MLE of its covariance-variance matrix. We use the drawn value of β_3 and the previously generated (y^*_{1i}, y^*_{2i}) to compute predicted probabilities based on the logistic regression equation, which we then use in Bernoulli draws to get synthetic y^*_{3i}.

(4') Using all n records in D, estimate the logistic regression, $logit(p(y_{4i})) = (1, y_{1i}, y_{2i}, y_{3i})\beta_4$. Draw a value of $\beta_4|D \sim N_4(\hat{\beta}_4, \hat{V}(\hat{\beta}_4))$, where $\hat{\beta}_4$ is MLE of β_4 and $\hat{V}(\hat{\beta}_4)$ is the MLE of its covariance-variance matrix. We use the drawn value of β_4 and the previously generated $(y^*_{1i}, y^*_{2i}, y^*_{3i})$ to compute predicted probabilities based on the logistic regression equation, which we then use in Bernoulli draws to get synthetic y^*_{4i}.

When implementing the N3WI or N2WI synthesizers, we repeat the four steps independently for $m = 5$ times to generate five synthetic datasets for public release.

3.2 Disclosure Risk Measurement

We now explicate the disclosure risk measures for this simulation design. For purposes of illustration, we assume that the intruder's target is y_{1000}, which is unique in the sample. Following (1), a key probability of interest is

$$p\left(Y_{1000}=(0,0,0,0)|Z,D_{-i},S\right)=\frac{p(Z|D_{-i},Y_i=c_{16},S)p(Y_i=c_{16}|D_{-i},S)}{\sum_{k=1}^{16}p(Z|D_{-i},Y_i=c_k,S)p(Y_i=c_k|D_{-i},S)}. \quad (6)$$

For this illustration, we assume $p(Y_i=y|D_{-i})=1/16$ for all y in the support. Thus, the prior probabilities cancel from the numerator and denominator. Using a uniform prior distribution is effectively equivalent to mimicking an intruder who searches over all possible values of y to find the one that results in the highest probability of $p(Z \mid D_{-i},Y_i=y,S)$.

By construction, we have

$$p(Z \mid D_{-i},Y_i,S)=\prod_{l=1}^{m}p(Z^{(l)} \mid D_{-i},Y_i,S). \quad (7)$$

Thus, we need to compute the posterior predictive distributions for each $Z^{(l)}$ under all possible values of Y_i.

For the DM synthesizer, we proceed as follows. Let $n_i^{(*k)}$ be the count of c_k in $D_i^{(*k)}=\{D_{-i},Y_i=c_k\}$. Let $n_k^{(l)}$ be the count of c_k in synthetic dataset $Z^{(l)}$. We draw many (say $H=1000$) values of θ from Dirichlet$(n_i^{(*1)}+a,...,n_i^{(*16)}+a)$. For any drawn $\theta^{(h)}$, we compute the multinomial probability,

$$p(Z^{(l)} \mid \theta^{(h)},D_{-i},Y_i=c_k,S)=\binom{n}{n_1^{(l)},\ldots,n_{16}^{(l)}}(\theta_1^{(h)})^{n_1^{(l)}}\cdots(\theta_{16}^{(h)})^{n_{16}^{(l)}}. \quad (8)$$

Finally, we approximate the posterior predictive probability as

$$p(Z^{(l)} \mid D_{-i},Y_i=c_k,S)=(1/H)\sum_h\binom{n}{n_1^{(l)},\ldots,n_{16}^{(l)}}(\theta_1^{(h)})^{n_1^{(l)}}\cdots(\theta_{16}^{(h)})^{n_{16}^{(l)}}. \quad (9)$$

For the N3WI and N2WI synthesizers, we follow a similar logic using the logistic regressions rather than Dirichlet-multinomial distributions. Let $\beta_1=Pr(Y_1=1)$. We seek to compute

$$\begin{aligned} p(Z^{(l)}|D_{-i},Y_i=c_k,S) &= \int\prod_{j=1}^{4}p(Z_j^{(l)} \mid \beta_j,D_{-i},Y_i=c_k,S) \\ &\quad p(\beta_j|D_{-i},Y_i=c_k,S)d\beta_j. \end{aligned} \quad (10)$$

To do so, we sample H values of $\beta=(\beta_1,\beta_2,\beta_3,\beta_4)$ following the method used in Section 3.1.2. However, we base the MLEs and estimated variance matrices on $D_i^{(*k)}$ rather than

D. Given a draw of $\beta^{(h)}$, we compute the probabilities for all sixteen possible c_k based on the logistic regressions and the expression in (5). We use these probabilities to form a multinomial distribution for use in (8) and (9).

3.3 Simulation Results

After generating an observed dataset, we generate a set of $m = 5$ synthetic datasets using each of the three synthesizers. In each set of synthetic datasets, we compute the posterior probability for y_{1000} with the different methods. We use $H = 1000$ draws of parameters.

Table 1 displays the estimated posterior probabilities for the different synthesizers. The DM synthesizer with $a = .0001$ offers no protection to the target $(0, 0, 0, 0)$. Basically, with this prior distribution (for the synthesis process) the synthesizer cannot generate a synthetic person at $(0, 0, 0, 0)$ unless there is someone in the data with that value. On the other hand, when we make $a = 1$ the risk to the target drops dramatically. This is because the prior distribution introduces a substantial probability of generating a $(0, 0, 0, 0)$ even without the target in the database. In fact, the intruder's posterior probabilities for $(1, 0, 1, 1)$ and $(1, 0, 0, 1)$ are very close to the target's probability, so that the intruder does not have strong evidence for selecting the right guess (although $(0, 0, 0, 0)$ remains the highest posterior probability guess). Once we use smoothed synthesizers N3WI and N2WI, $(0, 0, 0, 0)$ is no longer the highest probability guess. This indicates that the smoothing, which is typical in most applications of synthetic data, offers additional protection.

Although not a focus here, the choice of synthesizer also impacts the quality of the synthetic data. In particular, the quality of Z degrades as we go from left to right in Table 1, which also corresponds to increased use of smoothing in the synthesizer. Of note, using a DM synthesizer with $\alpha = 1$ nearly doubles the expected number of synthetic records with $(0, 0, 0, 0)$ compared to using a DM synthesizer with $\alpha = .0001$, resulting in biased estimates of $Pr(Y = c_{16})$. These types of biases engendered by smoothing are inherent in trading off reductions in data utility for reductions in disclosure risk.

By using a uniform distribution for the intruder's prior distribution for Y_{1000}, we actually over-weight the prior probability that $Y_{1000} = c_{16}$ compared to the frequency of c_{16} in the data. If instead we had based the intruder's prior distribution on the observed frequencies in D_{-1000}, we would expect the posterior probabilities from (6) to decrease compared to the reported values in Table 1. In fact, with no instances of c_{16} in D_{-1000}, sensible prior distributions constructed from D_{-1000} would put almost no prior mass on $Y_{1000} = c_{16}$. This generally would drive the probability in (6) to zero (except possibly for DM with $a = .0001$).

We examined other simulation designs that do not use uniform distributions for generating counts for C_{-16}. We found similar trends in the posterior probabilities when assuming the unique case is in fact the target.

Table 1: Posterior probabilities of guessing unknown record correctly in the contingency table simulation. Results based on one observed dataset and $m = 5$ synthetic datasets. True value of unknown record is $(0, 0, 0, 0)$.

Combination				DM ($a = .0001$)	DM ($a = 1$)	N3WI	N2WI
0	0	0	0	1	.077	.064	.070
0	0	0	1	0	.056	.084	077
0	0	1	0	0	.059	.062	.051
0	0	1	1	0	.059	.061	.068
0	1	0	0	0	.059	.051	.059
0	1	0	1	0	.058	.065	.056
0	1	1	0	0	.065	.059	.050
0	1	1	1	0	.062	.066	.062
1	0	0	0	0	.058	.071	.055
1	0	0	1	0	.073	.059	.062
1	0	1	0	0	.050	.070	.055
1	0	1	1	0	.074	.067	.050
1	1	0	0	0	.057	.078	.069
1	1	0	1	0	.058	.054	.061
1	1	1	0	0	.065	.067	.067
1	1	1	1	0	.060	.055	.063

4 Simulations with Partially Synthetic Data

The computations of the risk measures for partially synthetic data are similar to those for fully synthetic data with one key difference. For partial synthesis, we have to condition on any unchanged values when computing each $p\left(Y_i = y \mid Z, X, D_{-i}, S\right)$. For example, in the setting of Section 3, suppose that we replace (y_{3i}, y_{4i}) for each individual but leave each $x_i = (y_{1i}, y_{2i})$ at their collected values. Suppose we use Step 3 and Step 4 of the N3WI synthesizer (or Step 3' and Step 4' of the N2WI synthesizer) to generate the replacement values. To estimate the posterior probabilities, we would proceed as in Section 3.2 but replace (10) with

$$\begin{aligned} p(Z^{(l)}|D_{-i}, Y_i = c_k, x_i, S) &= \int \prod_{j=3}^{4} p(Z_j^{(l)} \mid \beta_j, D_{-i}, Y_i = c_k, x_i, S) \\ &\quad p(\beta_j|D_{-i}, Y_i = c_k, x_i, S) d\beta_j. \end{aligned} \tag{11}$$

Here, each (y_{1i}, y_{2i}) is not considered a random variable when evaluating (11), since these values are not changed in the synthetic data.

To illustrate the computations for partially synthetic data, we synthesize data from the 1987 Survey of Youth in Custody [30]. The survey interviewed youths in juvenile institutions about their family background, previous criminal history, and drug and alcohol use. The survey contains 2621 youths in 50 facilities. There are 23 variables on the file, including facility and race (measured in five categories). For reasons related to

data cleaning [33], we deleted all the youths in four facilities, leaving a total of $n = 2562$ youths.

We replace all values of facility and race (y variables), without altering other variables (x variables), using models developed by Mitra and Reiter [33] and Reiter and Mitra [44]. We first synthesize facility using multinomial regressions that include all other variables as predictors, except race and some variables that cause multi-collinearity. We then synthesize race using multinomial regressions that include all other variables plus indicator variables for facilities as predictors, except those that cause multicollinearity. The new values of race are sampled conditional on the values of the synthetic facility indicators. Similar models were shown previously to produce synthetic data with high analytic utility [33]. When generating the synthetic data, we use the MLEs directly from the multinomial regressions without drawing parameter values, as the theory of partial synthesis does not require drawing parameter values [43].

Nominally, there are $K = 230$ possible combinations of facility and race for each youth. However, facilities are grouped into two strata based on population size. We generate synthetic data separately for each stratum so that youths from stratum 1 (or 2) are forced to be in stratum 1 (or 2) in the synthetic data. While not strictly necessary, this improves the quality of the analyses of the synthetic data [33]. Thus, when computing risks we restrict the possible combinations to the $K_1 = 175$ observed combinations of facility and race for youths in stratum 1 and the $K_2 = 55$ observed combinations of facility and race for youths in stratum 2.

We assume that an intruder seeks to learn the facility and race of some individual i using the synthetic data. We again make the conservative assumption that the intruder knows the true facility and race values of all records except the target. Let y_i be the 2×1 vector comprising the actual facility number and race of youth i, and let x_i comprise all other variables. Let Y_i be the random variable corresponding to the intruder's guess at y_i. Using notation from Section 2, for all possible combinations y we compute

$$p\left(Y_i = y \mid Z, X, D_{-i}, S\right) = \frac{p(Z \mid X, D_{-i}, Y_i = y, S)p(Y_i = y \mid X, D_{-i}, S)}{\sum_y p(Z|X, D_{-i}, Y_i = y, S)p(Y_i = y|X, D_{-i})}. \quad (12)$$

We evaluate (12) under two prior distributions. The first represents an intruder with vague knowledge: we set $p(Y_i = y|X, D_{-i}, S)$ to be a discrete uniform distribution over the space of all possible combinations of facility and race in the stratum of record i. The second represents an intruder who uses D_{-i} to form prior beliefs. For each i, we estimate the MLEs of the coefficients in the multinomial regressions for facility and race using only D_{-i}. Using x_i and these MLEs (without sampling parameter values), we calculate the predicted probabilities for all possible combinations of y within the stratum. This set of probabilities is the prior distribution for record i. We repeat this process for each i, so that each target has its own prior distribution. We also compute a "prior risk" measure by counting the percentage of the n individuals whose race and facility combination is correctly predicted from the informative prior distribution; that

Table 2: Prior risk (PR) and posterior risk (R) measures for $m = 5$ partially synthetic datasets for the Survey of Youth in Custody. R values reported for all 5 synthetic datasets and for each individually.

Stratum	Synthetic Data	PR	R
1	All $m = 5$	.1402	.1471
1	$Z^{(1)}$		.1454
1	$Z^{(2)}$		.1454
1	$Z^{(3)}$		.1552
1	$Z^{(4)}$		.1506
1	$Z^{(5)}$		.1494
2	All $m = 5$	.2445	.2616
2	$Z^{(1)}$		.2409
2	$Z^{(2)}$		.2457
2	$Z^{(3)}$		.2372
2	$Z^{(4)}$		.2470
2	$Z^{(5)}$		.2530

is, analogous to (2), we compute

$$PR = (1/n)\sum_{i=1}^{n} I(argmax_y p(Y_i = y|X, D_{-i}, S) = y_i). \tag{13}$$

The computation of the posterior probabilities in (12) proceeds as follows. First, we set y to one of the possible combinations in the stratum, thus creating a plausible true dataset $D_i^{(*y)} = (D_{-i}, Y_i = y)$. Second, we estimate the MLEs of the multinomial regressions for facility and race from S using $D_i^{(*y)}$. Using the estimated coefficients, for $j = 1, \ldots, n$ we compute the multinomial probabilities of all K possible race-facility combinations from the multinomial logit equations, $\hat{\pi}_{ij}^{(*y)} = (\hat{\pi}_{ij}^{(*y1)}, \ldots, \hat{\pi}_{ij}^{(*yK)})$. Let $\hat{\pi}_i^{(*y)} = (\hat{\pi}_{i1}^{(*y)}, \ldots, \hat{\pi}_{in}^{(*y)})$. We compute

$$p(Z^{(l)} \mid X, D_{-i}, Y_i = y, S, \hat{\pi}_i^{(*y)}) = \prod_{j=1}^{n}\prod_{k=1}^{K}(\hat{\pi}_{ij}^{(*yk)})^{I(z_j^{(l)}=k)} \tag{14}$$

$$p(Z \mid D_{-i}, Y_i = y, X, S, \hat{\pi}_i^{(*y)}) = \prod_{l=1}^{m} p(Z^{(l)} \mid D_{-i}, Y_i = y, X, S, \hat{\pi}_i^{(*y)}). \tag{15}$$

Table 2 displays the prior (PR) and posterior risk measures (R) using all $m = 5$ synthetic datasets, as well as the risk measures for each individually, when the intruder uses the informative prior distribution. Both PR and R are higher for stratum 2, which primarily stems from the fact that stratum 2 has fewer possible combinations of facility and race (55 as compared to 175 from stratum 1). The values of R are not much

Table 3: Evaluation of how much the synthetic data add to the disclosure risks using the informative prior in the partial synthesis of the Survey of Youth in Custody.

		Posterior Prediction	
Prior Prediction	Total	Correct	Incorrect
Stratum 1			
Correct	244	219	25
Incorrect	1496	37	1459
Stratum 2			
Correct	201	184	17
Incorrect	621	31	590

Table 4: Posterior risk (R) associated with release of multiple synthetic datasets for stratum 2 in Survey of Youth in Custody.

m	1	2	3	4	5	6	7	8	9	10
R	.241	.247	.243	.247	.262	.258	.263	.268	.263	.258

different than those of PR, indicating that the release of $m = 5$ synthetic datasets (or single datasets) does not add much information to the informative prior distribution. In this case, intruders already can predict y_i accurately from D_{-i} for many records. As evident in Table 3, when the intruder's prediction based on the informative prior distribution is incorrect, the synthetic data allow the intruder to find the match for roughly an additional 2% to 5% of cases. However, roughly 10% of this intruder's prior correct guesses turn incorrect a posteriori.

When the intruder uses a uniform prior distribution, the values of R based on all $m = 5$ datasets are very similar: 13.9% for stratum 1 and 26.0% for stratum 2. However, when $m = 1$, the uniform prior distribution generates noticeably lower values of R: around 10% for stratum 1 and 18% for stratum. Evidently, the posterior probabilities in R are dominated by the likelihood function for Z when $m = 5$ (or larger, presumably), at least for these two prior distributions.

We also explored the extra risk associated with the release of increasing numbers of synthetic datasets. To do so, we generated $m = 10$ datasets and computed R after releasing them sequentially. As reported for stratum 2 in Table 4, the overall trend follows that observed by Reiter and Mitra [44]: disclosure risks tend to increase with m, with relatively large increases at small values of m. In these data, setting $m > 5$ appears not to carry appreciably higher disclosure risks than setting $m = 5$.

5 Concluding Remarks

In the simulations used here, we easily could search over the sample space of any Y_i, which allows us to compute the normalizing constants for the posterior probabilities. In

problems with large sample spaces, for example a full synthesis of all variables in the Survey of Youth in Custody, an exhaustive search would be computationally infeasible. For many records, however, an exhaustive search is likely unnecessary. For each record i, the agency can compute just the kernel of (1) for $Y_i = y_i$ and for a sequence of similar y, e.g., y values resulting from plausible changes in one field of y_i. When a sufficient number of these y values yield higher posterior probability than setting $Y_i = y_i$, the agency can declare that record sufficiently protected and stop computations with that i. Since the computations are trivially run in parallel across i, this strategy could be used as a screening device to identify records potentially at risk and deserving of in depth investigation. For such records, there may be approximations that can cut down the computation, for example, restricting the support of y to the $\approx n$ combinations of y observed in the collected data. For some synthesis models, it may be possible to draw (full synthesis) or maximize (partial synthesis) the parameters in the synthesis models without re-fitting the models from scratch, e.g, using importance sampling or delete-one-case formulas used in regression case influence diagnostics. Finally, for some synthesis models it may be possible to determine analytically the maximum a posteriori estimate of y_i.

Acknowledgments

This research was supported by National Science Foundation grant CNS-10-12141 and SES-11-31897.

References

[1] 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. report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at `http://www.census.gov/sipp/synth_data.html`.

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

[3] 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*. Amsterdam: North-Holland. 215–277.

[4] 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*. New York: Springer-Verlag. 290–297.

[5] 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.

[6] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the 27^{th} International Conference on Management of Data/Principles of Database Systems (SIGMOD/PODS '07)*. ACM Press. 273–282.

[7] Blum, A., Ligett, K., and Roth, A. (2008). A learning theory approach to non-interactive database privacy. In *Proceedings of the 40^{th} Symposium on Thoery of Computing (STOC '08)*. ACM Press. 609–618.

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

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

[10] Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6:73–85.

[11] 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.

[12] 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.

[13] Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin (eds), *Privacy in Statistical Databases*, vol. 5262 of *LNCS*. New York: Springer-Verlag. 227–238.

[14] 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.

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

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

[17] Elliott, M. and Purdam, K. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. *Environment and Planning A*, 39:1101–1118.

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

[19] 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.

[20] Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9:383–406.

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

[22] 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.

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

[24] Karwa, V. and Slavkovic, A. S. (2012). Differentially private graph degree sequences and synthetic graphs. In J. Domingo-Ferrer and I. Tinnirello (eds), *Privacy in Statistical Databases*, vol. 7556 of *LNCS*. Berlin: Springer-Verlag. 273–285.

[25] Kennickell, A. and Lane, J. (2006). Measuring the impact of data protection techniques on data utility: Evidence from the Survey of Consumer Finances. In J. Domingo-Ferrar (ed), *Privacy in Statistical Databases*, vol. 4302 of *LNCS*. New York: Springer-Verlag. 291–303.

[26] 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*. Washington, D.C.: National Academy Press. 248–267.

[27] Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review*, 79:363–384.

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

[29] 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*. New York: John Wiley & Sons. 141–152.

[30] Lohr, S. L. (1999). *Sampling: Design and Analysis*. New York: Duxbury Press.

[31] 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*. IEEE Computing Society. 277–286.

[32] 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.

[33] Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrer and L. Franconi (eds), *Privacy in Statistical Databases*. New York: Springer-Verlag. 177-188.

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

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

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

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

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

[39] 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.

[40] Reiter, J. P. (2005c). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131: 365–377.

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

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

[43] Reiter, J. P. and Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, 28:583–590.

[44] Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1:99–110.

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

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

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

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

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

[50] Winkler, W. E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Tech. report, U.S. Census Bureau Research Report Series, No. 2007-21.