Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data

Jörg Drechsler[1] and Jerome P. Reiter[2]

[1] Institute for Employment Research, 90478 Nuremberg Germany
[2] Duke University, Durham NC 27708, USA

Abstract. Partially synthetic data comprise the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple draws from statistical models. Because the original records remain on the file, intruders may be able to link those records to external databases, even though values are synthesized. We illustrate how statistical agencies can evaluate the risks of identification disclosures before releasing such data. We compute risk measures when intruders know who is in the sample and when the intruders do not know who is in the sample. We use classification and regression trees to synthesize data from the U.S. Current Population Survey.

Keywords: CART, Disclosure, Risk, Synthetic data.

1 Introduction

Several national statistical agencies disseminate multiply-imputed, partially synthetic data to the public. These comprise the units originally surveyed with only some collected values replaced with multiple imputations [1,2]. For example, in the Survey of Consumer Finances, the U.S. Federal Reserve Board replaces monetary values at high disclosure risk with multiple imputations, releasing a mixture of imputed values and the not replaced, collected values [3]. The U.S. Bureau of the Census protects data in the Survey of Income and Program Participation [4] and in longitudinal business databases [5,6] by replacing all values of sensitive variables with multiple imputations, leaving non-sensitive variables at their actual values. They also have created synthesized origin-destination matrices, i.e. where people live and work, available to the public as maps via the web (On The Map, http://lehdmap.did.census.gov/). They plan to protect the identities of people in group quarters (e.g., prisons, shelters) in the American Communities Survey by replacing quasi-identifiers for records at high disclosure risk with imputations. Partially synthetic, public use data are being developed for the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey veterans and full sample data. Other examples of partially synthetic data are in [7,8,9,10].

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 227–238, 2008.

Because the original records remain on the file, intruders may be able to link those records to external databases, even though values are synthesized. It is prudent for agencies to assess the risks of such identification disclosures before releasing the file. When they are too high, additional synthesis or some other action is needed before release. In this article, we illustrate how to compute risks of identification disclosure for partially synthetic data using a subset of the U. S. Current Population Survey. We show how to incorporate intruders' uncertainty about which records are in the sample and how to assess different synthesis strategies. We also illustrate an application of classification and regression tree methodology for generating partially synthetic data.

2 Review of Partially Synthetic Data

The agency constructs partially synthetic datasets based on the s records in the observed data, D_{obs}, in a two-part process. First, the agency selects the values from the observed data that will be replaced with imputations. Second, the agency imputes new values to replace those selected values. Let $Y_{\text{rep},i}$ be all the imputed (replaced) values in the ith synthetic dataset, and let Y_{nrep} be all unchanged (not replaced) values. The values in Y_{nrep} are the same in all synthetic datasets. Each synthetic dataset, D_i, is then comprised of $(Y_{\text{rep},i}, Y_{\text{nrep}})$. Imputations are made independently for $i = 1, \ldots, m$ times to yield m different synthetic datasets. These synthetic datasets are released to the public.

When using parametric imputation models, the $Y_{\text{rep},i}$ should be generated from the Bayesian posterior predictive distribution of $(Y_{\text{rep},i}|D_{\text{obs}})$, or some approximation to it such as the sequential regression imputation methods [11]. In this article, we generate the $Y_{\text{rep},i}$ from a series of regression tree (CART) models. These models are described in Section 4.1.

Inferences about some scalar estimand, say Q, are obtained by combining results from the D_i. Specifically, suppose that the data analyst estimates Q with some point estimator q and estimates the variance of q with some estimator v. For $i = 1, \ldots, m$, let q_i and v_i be respectively the values of q and v in D_i. It is assumed that the analyst determines q_i and v_i as if D_i was in fact a random sample collected with the original sampling design. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{2}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i/m \ . \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_{\text{p}} = b_m/m + \bar{v}_m \tag{4}$$

to estimate the variance of $\bar{q}_m$. When s is large, inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_p = (m-1)(1+r_m^{-1})^2$, where $r_m = (m^{-1}b_m/\bar{v}_m)$. Derivations of these methods are presented in [2]. Extensions for multivariate Q are presented in [12].

3 Identification Disclosure Risk Measures for Partial Synthesis

To evaluate disclosure risks, we compute probabilities of identification by following the approach in [13]. Related approaches for non-synthetic data are in [14,15,16,17]. Roughly, in this approach we mimic the behavior of an ill-intentioned user of the released data who possesses the true values of the quasi-identifiers for selected target records (or even the entire population). The intruder has a vector of information, $\mathbf{t}$, on a particular target unit in the population which may or may not correspond to a unit in the m partially synthetic datasets, $\mathbf{D} = \{D_1, \ldots, D_m\}$. Let t_0 be the unique identifier (e.g., full name and address of a survey respondent) of the target, and let d_{j0} be the (not released) unique identifier for record j in $\mathbf{D}$, where $j = 1, \ldots, s$. Let M be any information released about the simulation models.

The intruder's goal is to match unit j in $\mathbf{D}$ to the target when $d_{j0} = t_0$, and not to match when $d_{j0} \neq t_0$ for any $j \in \mathbf{D}$. Let J be a random variable that equals j when $d_{j0} = t_0$ for $j \in \mathbf{D}$ and equals $s+1$ when $d_{j0} = t_0$ for some $j \notin \mathbf{D}$. The intruder thus seeks to calculate the $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$ for $j = 1, \ldots, s+1$. He or she then would decide whether or not any of the identification probabilities for $j = 1, \ldots, s$ are large enough to declare an identification. Let Y_{rep} be all original values of the variables that were synthesized. Because the intruder does not know the actual values in Y_{rep}, he or she should integrate over its possible values when computing the match probabilities. Hence, for each record in $\mathbf{D}$, we compute

$$Pr(J = j|\mathbf{t}, \mathbf{D}, M) = \int Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{\text{rep}}, M) Pr(Y_{\text{rep}}|\mathbf{t}, \mathbf{D}, M) dY_{\text{rep}} \ . \quad (5)$$

This construction suggests a Monte Carlo approach to estimating each $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$. First, sample a value of Y_{rep} from $Pr(Y_{\text{rep}}|\mathbf{t}, \mathbf{D}, M)$. Let Y^{new} represent one set of simulated values. Second, compute $Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{\text{rep}} = Y^{\text{new}}, M)$ using exact or, for continuous synthesized variables, distance-based matching assuming Y^{new} are collected values. This two-step process is iterated R times, where ideally R is large, and (5) is estimated as the average of the resultant R values of $Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{\text{rep}} = Y^{\text{new}}, M)$. When M has no information, the intruder can treat the simulated values in each $Y_{\text{rep},i}$ as plausible draws of Y_{rep}.

To illustrate, suppose that age, race, and sex are the only quasi-identifiers in a survey of households. The agency releases $m > 1$ partially synthetic datasets with all values of race and age synthesized and sex not changed. We suppose that the agency does not release any information about the imputation model

but does reveal which variables are synthesized. Suppose that an intruder seeks to identify a white male aged 45, *and he knows that this target is in the sample.* In each D_i, the intruder would search for all records matching the target on age, race, and sex. Let $N_{\mathbf{t},i}$ be the number of matching records in $\mathbf{D}_i$, where $i = 1, \ldots, m$. When no one with all of those characteristics is in $\mathbf{D}_i$, set $N_{\mathbf{t},i}$ equal to the number of males in D_{obs}, i.e., match on all non-simulated quasi-identifiers. For $j = 1, \ldots, s$,

$$Pr(J = j|\mathbf{t}, \mathbf{D}, M) = (1/m) \sum_i (1/N_{\mathbf{t},i})(Y_{ij}^{\text{new}} = \mathbf{t}) \ , \tag{6}$$

where $(Y_{ij}^{\text{new}} = \mathbf{t}) = 1$ when record j is among the $N_{\mathbf{t},i}$ matches in D_i and equals zero otherwise. We note that $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) = 0$ because the intruder knows this target is in the sample.

Now suppose that the intruder *does not know that this target is in the sample.* For $j = 1, \ldots, s$, we have to replace $N_{\mathbf{t},i}$ in (6) with $F_{\mathbf{t}}$, the number of records in the population that match the target on age, race, and sex. When the intruder and the agency do not know $F_{\mathbf{t}}$, it can be estimated using the approach in [17], which assumes that the population counts follow an all-two-way-interactions log-linear model. The agency can determine the estimated counts, $\hat{F}_{\mathbf{t}}$, by fitting this log-linear model with D_{obs}. Alternatively, since D_{obs} is in general not available to intruders, the agency can fit a log-linear model with each D_i, resulting in the estimates $\hat{F}_{\mathbf{t},i}$ for $i = 1, \ldots, m$. We note that $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) = 1 - \sum_{j=1}^{s} Pr(J = j|\mathbf{t}, \mathbf{D}, M)$.

For some target records, the value of $N_{\mathbf{t},i}$ might exceed $F_{\mathbf{t}}$ (or $\hat{F}_{\mathbf{t}}$ if it is used). It should not exceed $\hat{F}_{\mathbf{t},i}$, since $\hat{F}_{\mathbf{t},i}$ is required to be at least as large as $N_{\mathbf{t},i}$. For such cases, we presume that the intruder sets $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) = 0$ and picks one of the matching records at random. To account for this case, we can re-write (6) for $j = 1, \ldots, s$ as

$$Pr(J = j|\mathbf{t}, \mathbf{D}, M) = (1/m) \sum_i \min\left(1/F_{\mathbf{t}}, 1/N_{\mathbf{t},i}\right) (Y_{ij}^{\text{new}} = \mathbf{t}) \ . \tag{7}$$

As suggested in [16], we quantify disclosure risks with summaries of the identification probabilities in (6) and (7). It is reasonable to assume that the intruder selects as a match for $\mathbf{t}$ the record j with the highest value of $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$, if a unique maximum exists. We consider three disclosure risk measures. To calculate these measures, we need some further definitions. Let $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_{|\mathbf{T}|}\}$ be the set of the intruder's targets. Let c_j be the number of records in the released data with the highest match probability for the target $\mathbf{t}_j$; let $I_j = 1$ if the true match is among the c_j units and $I_j = 0$ otherwise. Let $K_j = 1$ when $c_j I_j = 1$ and $K_j = 0$ otherwise. The *expected match risk* is defined as $\sum_{j \in \mathbf{T}} (1/c_j) I_j$. When $I_j = 1$ and $c_j > 1$, the contribution of unit j to the expected match risk reflects the intruder randomly guessing at the correct match from the c_j candidates. The *true match risk* equals $\sum_{j \in \mathbf{T}} K_j$. Finally, we introduce the *true match rate* equal to $\sum_{j \in \mathbf{T}} K_j / \sum_{j \in \mathbf{T}} (c_j = 1)$, which is the percentage of true matches for the targets that have a unique match in $\mathbf{D}$.

Table 1. Description of variables used in the empirical studies

Variable	Label	Range
Sex	X	male, female
Race	R	white, black, American Indian, Asian
Marital status	M	7 categories, coded 1–7
Highest attained education level	E	16 categories, coded 31–46
Age (years)	G	0 – 90
Child support payments ($)	C	0, 1 – 23,917
Social security payments ($)	S	0, 1 – 50,000
Household alimony payments ($)	A	0, 1 – 54,008
Household property taxes ($)	P	0, 1 – 99,997
Household income ($)	I	-21,011 – 768,742

4 Empirical Evaluation

We simulate partial synthesis for a subset of public release data from the March 2000 U.S. Current Population Survey. The data comprise ten variables measured on $N = 51,016$ heads of households. The variables, displayed in Table 1, were selected and provided by statisticians at the U.S. Bureau of the Census. Similar data are used in [18] to illustrate and evaluate releasing fully synthetic data.

Marginally, there are ample numbers of people in each sex, race, marital status, and education category. Many cross-classifications have few people, especially those involving minorities with $M \notin \{1, 7\}$. There are 521 records with unique combinations of age, race, marital status, and sex. There are 284 combinations of the four variables that have only two records in the dataset. There are 2064 empty cells in the four-way contingency table.

We treat the N records as a population and take a random sample of $n = 10,000$ for D_{obs}. We consider age, race, marital status, and sex to be quasi-identifiers that intruders may know precisely. Cross-classification of these four variables in the sample yields 473 sample uniques, 241 duplicates and 2909 empty cells in the four-way contingency table. Intruders might have access to other variables on the file, such as property taxes. Thus, the computations in this section serve to illustrate our suggested disclosure risk measures rather than to evaluate the actual disclosure risks for this specific dataset (which is already in the public domain).

We generate synthetic datasets for each of two scenarios: replace all values of age, marital status, and race without changing sex; and, replace all values of marital status and race without changing age and sex. The synthetic data are generated using regression trees, as we now describe.

4.1 CART Synthesis Models

CART models are a flexible tool for estimating the conditional distribution of a univariate outcome given multivariate predictors. Essentially, the CART model

partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes. The partitions are found by recursive binary splits of the predictors. The series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units.

CART models also can be used to generate partially synthetic data [19]. To synthesize all values of age, marital status, and race, we proceed as follows. First, using D_{obs} we fit the tree of age on all other variables except race and marital status. Label this tree $\mathcal{Y}_{(G)}$. We require a minimum of five records in each leaf of the tree and do not prune it; see [19] for discussion of pruning and minimum leaf size. Let L_{Gw} be the wth leaf in $\mathcal{Y}_{(G)}$, and let $Y_{(G)}^{L_{Gw}}$ be the $n_{L_{Gw}}$ values of $Y_{(G)}$ in leaf L_{Gw}. In each L_{Gw} in the tree, we generate a new set of values by drawing from $Y_{(G)}^{L_{Gw}}$ using the Bayesian bootstrap [20]. These sampled values are the replacement imputations for the $n_{L_{Gw}}$ units that belong to L_{Gw}. Repeating the Bayesian bootstrap in each leaf of the age tree results in the ith set of synthetic ages, $Y_{(G)\text{rep},i}$.

To avoid releasing only values of the observed ages in each leaf, we could take an additional step suggested in [19]. In each leaf, we could estimate the density of the bootstrapped values using a Gaussian kernel density estimator with support over the smallest to the largest value of $Y_{(G)}$. Then, for each unit, we would sample randomly from the estimated density in that unit's leaf using an inverse-cdf method. The sampled values rounded to the nearest integer would be the $Y_{(G)rep,i}$. We do not take this extra step here.

Imputations are next made for marital status. Using D_{obs}, we fit the tree, $\mathcal{Y}_{(M)}$, with all variables except race as predictors. To maintain consistency with $Y_{(G)\text{rep},i}$, units' leaves in $\mathcal{Y}_{(M)}$ are located using $Y_{(G)\text{rep},i}$. Occasionally, some units may have combinations of values that do not belong to one of the leaves of $\mathcal{Y}_{(M)}$. For these units, we search up the tree until we find a node that contains the combination, then treat that node as if it were the unit's leaf. Once each unit's leaf is located, values of $Y_{(M)\text{rep},i}$ are generated using the Bayesian bootstrap. Imputing races follows the same process: we fit the tree $\mathcal{Y}_{(R)}$ using all variables as predictors, place each unit in the leaves of $\mathcal{Y}_{(R)}$ based on their synthesized values of age and marital status, and sample new races using the Bayesian bootstrap.

The process is repeated independently $m = 5$ times. These m datasets would be released to the public. All CART models are fit in S-Plus using the "tree" function. It takes about five minutes to generate five synthetic datasets with all three variables. The sequential order of imputation is $G - M - R$; see [19] for a discussion of the ordering of the trees. The synthesis of only marital status and race is similar except that the process begins with marital status. Although we use the CART method only to generate categorical data, it is straightforward to apply the method to generate continuous variables [19].

4.2 Data Utility

Evaluating disclosure risk is, of course, only part of the story. We could create completely worthless data and have very low disclosure risks. Hence, it is important to examine data usefulness when evaluating disclosure risks.

Table 2. Point estimates and standard errors for observed data, synthetic data with age not replaced, and synthetic data with age replaced

Estimand	Observed Data q_{obs} (SE)	True Age $\bar{q}_5$ ($\sqrt{T_{\text{p}}}$)	Synth. Age $\bar{q}_5$ ($\sqrt{T_{\text{p}}}$)
Avg. education for married black females	39.5 (.21)	39.6 (.21)	39.7 (.20)
Coefficient in regression of $\sqrt{C}$ on:			
Intercept	-94.5 (27)	-94.5 (27)	-95.3 (27)
Female	12.5 (5.4)	12.4 (5.4)	12.2 (5.4)
Non-white	-1.72 (4.7)	-0.34 (4.9)	-0.53 (4.8)
Education	3.44 (0.6)	3.44 (.60)	3.46 (0.6)
Number of youths in house	1.33 (1.6)	1.34 (1.6)	1.37 (1.6)
Coefficient in regression of $\sqrt{S}$ on:			
Intercept	81.0 (4.5)	78.1 (4.6)	79.4 (4.9)
Female	-11.1 (1.1)	-11.1 (1.1)	-10.6 (1.1)
Black	-7.0 (1.6)	-6.3 (1.9)	-5.3 (1.8)
American Indian	-8.2 (4.7)	-8.9 (7.1)	-10.8 (5.5)
Asian	0.1 (3.3)	-3.1 (3.8)	2.3 (3.7)
Widowed	5.0 (1.2)	4.7 (1.2)	4.3 (1.2)
Divorced	-3.0 (1.7)	-0.3 (1.8)	0.3 (1.8)
Single	-1.4 (2.1)	2.0 (2.1)	3.5 (2.2)
High school	3.6 (1.1)	3.8 (1.1)	3.8 (1.1)
Some college	5.2 (1.3)	5.1 (1.3)	5.7 (1.3)
College degree	8.3 (1.7)	8.3 (1.7)	8.1 (1.7)
Advanced degree	10.1 (2.1)	9.8 (2.2)	9.8 (2.2)
Age	0.22 (.06)	0.25 (.06)	0.23 (.07)
Coefficient in regression of $\log(I)$ on			
Intercept	4.80 (.10)	4.78 (.10)	4.82 (.15)
Black	-0.14 (.03)	-0.16 (.03)	-0.12 (.03)
American Indian	-0.20 (.07)	-0.21 (.09)	-0.12 (.08)
Asian	-0.01 (.05)	0.04 (.06)	0.01 (.05)
Female	0.02 (.02)	0.01 (.03)	-0.002 (.03)
Married in armed forces	-0.04 (.10)	-0.30 (.15)	-0.19 (.11)
Widowed	-0.07 (.06)	-0.17 (.07)	-0.30 (.08)
Divorced	-0.11 (.04)	-0.14 (.05)	-0.13 (.04)
Separated	-0.28 (.09)	-0.13 (.11)	-0.24 (.10)
Single	-0.15 (.04)	-0.11 (.04)	-0.12 (.04)
Education	0.113 (.003)	0.113 (.003)	0.114 (.003)
Household size > 1	0.54 (.03)	0.54 (.03)	0.52 (.03)
Females married in armed forces	-0.49 (.14)	-0.22 (.16)	-0.39 (.14)
Widowed females	-0.27 (.07)	-0.15 (.07)	-.07 (.08)
Divorced females	-0.34 (.05)	-0.31 (.06)	-0.33 (.06)
Separated females	-0.45 (.11)	-0.48 (.13)	-0.41 (.12)
Single females	-0.35 (.05)	-0.37 (.05)	-0.33 (.05)
Age	0.043 (.003)	0.043 (.003)	.041 (.003)
Age^2 $\times 1000$	-0.42 (.03)	-0.42 (.03)	-0.41 (.03)
Property tax $\times 10000$	0.27 (.03)	0.29 (.03)	0.29 (.03)

Child support regression fit using records with $C > 0$. Social security regression fit using records with $S > 0$ and $G > 54$. Income regression fit using records with $I > 0$.

Table 2 provides some evidence of the usefulness of the five synthetic datasets. It displays the point estimates and standard errors for several quantities based on the observed and partially synthetic data. Synthetic estimates are computed from the $m = 5$ datasets using the methods described in Section 2. The synthetic data point estimates are generally within two standard errors of the observed data point estimates. The biggest differences are for quantities associated with small sub-groups, such as married in the armed forces. We believe that the results in Table 2 are evidence of good quality, especially since the regressions involved subsets of data, transformations of variables, and interaction effects. We note that these results were obtained without any tuning other than to decide on the minimum number of records for each leaf and the order of synthesis. We also note that the results for synthesizing or not synthesizing age are similar.

4.3 Disclosure Risk

We consider four scenarios with different assumptions about the information available to the intruder. Across all scenarios, we assume the intruder knows the sex, age, race and marital status of some target records, for example from an external database.

- Scenario I: the intruder knows the identifiers for 10,000 randomly specified units in the population but does not know who is in the survey.
- Scenario II: the intruder knows the identifiers for 10,000 randomly specified units in the population and knows who is in the survey.
- Scenario III: the intruder knows the identifiers for all $N = 51,016$ units in the population but does not know who is in the survey.
- Scenario IV: the intruder knows the identifiers for all $N = 51,016$ units in the population and knows who is in the survey.

For Scenarios I and II, 1,968 of the intruder's target records are included in D_{obs}. For Scenario I, we estimate each $\hat{F}_{\mathbf{t},i}$ by fitting the all-two-way-interactions log-linear model on each D_i. An intruder might do this if he is unsure whether or not his $10,000$ records are representative of the population. It is prudent for the agency to assess the disclosure risk using estimated counts based on D_{obs} as well. For Scenario III, the intruder presumably would use the known values of $F_{\mathbf{t}}$. For interest, we report the results for the first and third scenarios using both estimated and true population counts.

For Scenarios I and III, we consider three intruder strategies. The first is that the intruder matches to the released data no matter what the value of $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$. That is, the intruder ignores the chance that a record is not in the sample. The second is that the intruder matches to the released data only when $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) < \gamma$, where $0 < \gamma < 1$. The third is that the intruder does not match whenever $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$ is the maximum probability for the target.

We compare the risks when only race and marital status are synthesized to the risks when age, race, and marital status are synthesized.

Table 3. Disclosure risks when only marital status and race are synthesized and intruder matches regardless of the value of $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$

	Scen. I		Scen. II	Scen. III		Scen. IV
	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$		$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$	
Expected match risk	72.3	71.6	74.7	367.8	361.1	365.1
True match risk	26	40	37	131	201	172
Number of single matches	1,942	3,445	593	9,769	17,555	2,905
True match rate (%)	1.34	1.16	6.24	1.34	1.14	5.92

Synthesis of Race and Marital Status Only. Table 3 displays the risk measures when age is left unchanged and the intruder matches regardless of the value of $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$. In all scenarios, the great majority of declared matches are incorrect, as evident by the low true match rates. True match rates are highest when the intruder knows who is in the sample, as might be expected. Given $\mathbf{T}$, the expected match risk measures are very similar for an intruder with response knowledge and an intruder not knowing who participated in the survey. The true match risk measures are higher when using the $\hat{F}_{\mathbf{t},i}$ instead of $F_{\mathbf{t}}$. This is because the number of matches with $c_j = 1$ is higher when matching with $\hat{F}_{\mathbf{t},i}$ instead of $F_{\mathbf{t}}$, as evident in the third row of the table.

Naturally, the numbers of expected and true matches increase when the intruder has information for the whole population rather than only for a sample. Quite simply, there are more targets to match. The expected and true risk measures when only around 2000 records are in $\mathbf{T} \cap D_{\text{obs}}$ are roughly 1/5 the magnitudes when all 10000 records in D are in $\mathbf{T} \cap D_{\text{obs}}$.

The results in Table 3 presume that the intruder always considers the record j with maximum $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$, where $j = 1, \ldots, s$ a match no matter how small this maximum is. With this strategy, the number of true matches is swamped by the number of false matches. For targets with $J = s + 1$ as the maximum match probability, the intruder might not match if he deems $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M)$ to be too high, say exceeding a threshold γ. Large values of γ result in a higher number of true and false matches. Small values of γ reduce the chance of false matches but miss out on some true matches. Table 4 presents the risk measures for Scenario I and III using $\gamma = 0.5$. As expected, there is a reduction in both the number of true matches and the total number of single matches. In fact, in Scenario I the intruder detects very few correct matches. However, in both scenarios the true match rate increases from around 1% to at least 8%.

The intruder also might choose not to match for targets with $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) \geq Pr(J = j|\mathbf{t}, \mathbf{D}, M)$ for $j = 1, \ldots, s$. Applying this strategy, the intruder obtains 2 true matches (with a match rate of 50%) in Scenario I and 6 true matches (with a match rate of 20%) in Scenario III.

Synthesis of Age, Race, and Marital Status. The agency may decide that the disclosure risks are too high when synthesizing only race and marital status.

Table 4. Disclosure risks for Scenario I and III when only marital status and race are synthesized and the intruder matches if $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) \leq 0.5$

	Scen. I		Scen. III	
	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$
Expected match risk	3	1	9.5	6
True match risk	3	1	9	6
Number of single matches	17	11	102	64
True match rate (%)	17.65	9.09	8.82	9.37

Table 5. Disclosure risks when age, marital status, and race are synthesized and intruder matches regardless of the value of $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$

	Scen. I		Scen. II	Scen. III		Scen. IV
	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$		$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$	
Expected match risk	3.5	3.0	4.2	14.0	14.7	16.0
True match risk	2	3	3	4	12	12
Number of single matches	2,651	6,879	1,252	13,641	34,972	6,359
True match rate (%)	0.075	0.044	0.240	0.029	0.034	0.189

Table 6. Disclosure risks for Scenario I and III when age, marital status, and race are synthesized and the intruder matches if $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) \leq 0.5$

	Scen. I		Scen. III	
	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$
Expected match risk	0	0	0	0
True match risk	0	0	0	0
Number of single matches	6	6	48	41
True match rate (%)	0	0	0	0

Table 5 displays the results if age is also synthesized, assuming that the intruder matches no matter what. The risks decrease significantly. The true match rate drops well below 1% for all scenarios. Table 6 displays the risks when the intruder matches only if $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) < 0.5$. The intruder cannot detect any correct matches.

Synthesizing age appears to reduce the disclosure risks substantially for this dataset. Given the similarity in the data utility of the two approaches, we suspect that many agencies would opt to synthesize age.

5 Concluding Remarks

The simulation results suggest several conclusions about disclosure risks in partially synthetic data. These include:

1. Knowing which targets are in the sample increases the true match rate compared to not knowing which targets are in the sample, so that disclosure risks increase.
2. Intruders who match to the synthetic data regardless of the value of $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$ can find more true matches at the expense of a higher false match rate than intruders who would not match when $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M)$ is large.
3. There are differences in the risk measures when using estimated population counts versus true population counts. However, they tend to be small and arguably not worth worrying about.
4. Synthesizing variables that are primary contributors to the disclosure risks, in particular age, can reduce disclosure risks substantially.

In general, it is difficult for the agency to know what information is owned by intruders. We recommend that the agency evaluate disclosure risks under conservative but realistic assumptions of intruder knowledge. For example, to begin, the agency can assume that intruders know exactly who is in the sample and have correct values of all quasi-identifiers. The agency then can back off these assumptions, for example assuming that intruders do not know who is in the sample or that intruders do now know some quasi-identifiers. By computing risk and utility under a variety of assumptions, the agency can decide if the disclosure risks are adequately low for the proposed microdata release.

Acknowledgments. This research was supported by grants from the U.S. National Science Foundation (NSF-ITR-0427889 and NSF-MMS-0751671) and the IAB in Germany.

References

1. Little, R.J.A.: Statistical analysis of masked data. J. Off. Stat. 9, 407–426 (1993)
2. Reiter, J.P.: Inference for partially synthetic, public use microdata sets. Surv. Methodol. 29, 181–189 (2003)
3. Kennickell, A.B.: Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In: Record Linkage Techniques, pp. 248–267. National Academy Press, Washington (1997)
4. Abowd, J.M., Stinson, M., Benedetto, G.: Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program (2006)
5. Abowd, J.M., Woodcock, S.D.: Disclosure limitation in longitudinal linked data. In: Doyle, P., Lane, J., Zayatz, L., Theeuwes, J. (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 215–277. North-Holland, Amsterdam (2001)
6. Abowd, J.M., Woodcock, S.D.: Multiply-imputing confidential characteristics and file links in longitudinal linked data. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 290–297. Springer, Heidelberg (2004)
7. Reiter, J.P.: Simultaneous use of multiple imputation for missing data and disclosure limitation. Surv. Methodol. 30, 235–242 (2004)

8. Little, R.J.A., Liu, F., Raghunathan, T.E.: Statistical disclosure techniques based on multiple imputation. In: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, pp. 141–152. John Wiley & Sons, New York (2004)
9. Mitra, R., Reiter, J.P.: Adjusting survey weights when altering identifying design variables via synthetic data. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 177–188. Springer, Heidelberg (2006)
10. Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. Joint Eurostat UNECE Worksession on Statistical Data Confidentiality, Manchester, WP. 11 (2007)
11. Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a series of regression models. Surv. Methodol. 27, 85–96 (2001)
12. Reiter, J.P.: Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. J. Stat. Plan. Inf. 131, 365–377 (2005)
13. Reiter, J.P., Mitra, R.: Estimating risks of identification disclosure in partially synthetic data. J. Priv. Conf. (to appear)
14. Duncan, G.T., Lambert, D.: The Risk of disclosure for microdata. Journal of Business and Economic Statistics 7, 207–217 (1989)
15. Fienberg, S.E., Makov, U.E., Sanil, A.P.: A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. J. Off. Stat. 13, 75–89 (1997)
16. Reiter, J.P.: Estimating identification risks in microdata. J. Amer. Stat. Assoc. 100, 1103–1113 (2005)
17. Elamir, E.A.H., Skinner, C.J.: Record level measures of disclosure risk for survey microdata. J. Off. Stat. 22, 525–529 (2006)
18. Reiter, J.P.: Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. J. Roy. Stat. Soc. A 168, 531–544 (2005)
19. Reiter, J.P.: Using CART to generate partially synthetic, public use microdata. J. Off. Stat. 21, 441–462 (2005)
20. Rubin, D.B.: The Bayesian bootstrap. Ann. Stat. 9, 130–134 (1981)