UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M08/07

Methodology

Estimating Risks Of Identification Disclosure In Partially Synthetic Data

Jerome P. Reiter, Robin Mitra

Abstract

To limit disclosures, statistical agencies and other data disseminators can release partially synthetic, public use microdata sets. These comprise the units originally surveyed, but some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, are replaced with multiple draws from statistical models. Because the original records are on the file, there remain risks of identifications. In this paper, we describe how to evaluate identification disclosure risks in partially synthetic data, accounting for released information from the multiple datasets, the model used to generate synthetic values, and the approach used to select values to synthesize. We illustrate the computations using the Survey of Youths in Custody.

Estimating Risks of Identification Disclosure in Partially Synthetic Data

Jerome P. Reiter and Robin Mitra*

Abstract

To limit disclosures, statistical agencies and other data disseminators can release partially synthetic, public use microdata sets. These comprise the units originally surveyed, but some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, are replaced with multiple draws from statistical models. Because the original records are on the file, there remain risks of identifications. In this paper, we describe how to evaluate identification disclosure risks in partially synthetic data, accounting for released information from the multiple datasets, the model used to generate synthetic values, and the approach used to select values to synthesize. We illustrate the computations using the Survey of Youths in Custody.
KEY WORDS: Confidentiality; Public use data; Record linkage; Survey.

1 INTRODUCTION

To limit the risks of disclosures when releasing data on individual records, statistical agencies can release multiply-imputed, partially synthetic data. These comprise the units originally surveyed with some collected values, e.g. sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. Partially synthetic, public use data sets are in the development stage for the Survey of Income and Program Participation, the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey group quarters data. For other examples and discussions of partially synthetic data, see Little (1993), Kennickell (1997) Abowd and Woodcock (2001, 2004), Reiter (2004b, 2005c), Abowd and Lane (2004), Little *et al.* (2004), and Mitra and Reiter (2006).

To illustrate the general idea of partial synthesis, we adapt the setting of Reiter (2004a). Suppose the agency has collected data on a random sample of people. The data comprise each person's age, race, sex, and income. Some intruder, who knows values of age, race, and sex for individuals in the sample, wants to identify individuals by matching on age, race, and sex. Suppose the agency wants disguise the identities of all people to discourage this linking. To do so, the agency might replace the actual race and sex (and possibly age) values for those people with simulated values. Specifically, the agency estimates the joint distribution of race and sex, conditional on age and income, and samples new values of race and sex for the sampled people. The distribution is estimated using the collected data and other relevant information. The result is one synthetic data set. The agency repeats this process, i.e. draws new values of race and sex, m times to generate m synthetic data sets. These m data sets are then released to the public.

Because the replacement values are simulated from probability models, the relationships among the variables should be preserved on average, provided the models reasonably describe the data. The "on

*Jerome Reiter is an assistant professor and Robin Mitra is a PhD candidate at the Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251 (E-mail: jerry@stat.duke.edu). This research was supported by the National Science Foundation grant, ITR-0427889.

average" caveat is important: parameter estimates from any one simulated data set are unlikely to equal exactly those from the observed data. The synthetic parameter estimates are subject to two sources of variation, namely sampling the collected data and simulating the replacement values. It is not possible to estimate the the latter source of variation from only one released synthetic data set. However, it is possible to do so from multiple synthetic data sets, which explains why multiple synthetic data sets are released. To account for both sources of variability, the user estimates parameters and their variances in each of the synthetic data sets, and combines these results using simple formulas described by Reiter (2003, 2005b). The analyst uses standard methods and software to obtain estimates in each synthetic data set.

The protection afforded by partially synthetic data depends on the nature of the synthesis. Replacing key identifiers with imputations makes it difficult for users to know the values of the original values of those identifiers, which reduces the chance of identifications. Replacing values of sensitive variables makes it difficult for users to learn the exact values of those variables, which can prevent attribute disclosures. Nonetheless, partially synthetic data sets remain susceptible to disclosure risks. The originally sampled units remain in the released files, albeit with some values changed, leaving values that users can utilize for record linkages. Furthermore, the intruder can utilize the multiple copies of the synthetic values, as well as any released meta-data about how they were generated, in disclosure attacks.

This paper describes some approaches that intruders might use to attempt identifications in partially synthetic data. It proposes a general framework for quantifying the identification disclosure risks inherent in releasing partially synthetic data. The framework accounts for (i) the information existing in all the released synthetic data sets, (ii) various assumptions about intruder knowledge and behavior, and (iii) the details released about the synthetic data generation model. The approach is illustrated on a genuine, partially synthesized data set.

2 Risk measure

To describe the framework, we compute probabilities of identification, conditional on the released data. This was first proposed by Duncan and Lambert (1986, 1989) and has been extended by Fienberg *et al.* (1997) and Reiter (2005a).

For a collection of n sampled units, let y_{jk} be the collected data for unit j on variable k, for $k = 0, \ldots, p$ and $j = 1, \ldots, n$. The column $k = 0$ contains unique unit identifiers, such as names or social security numbers, and is never released by the agency. It is convenient to split $\mathbf{y}_j = (y_{j1}, \ldots, y_{jp})$ into two sets of variables. Let $\mathbf{y}_{jA}$ be the vector of variables available to users from external databases, such as demographic or geographic attributes. And, let $\mathbf{y}_{jU}$ be the vector of variables that are available to users only in the released data. The compositions of A and U are determined by the agency based on knowledge of what information exists in external databases. It is assumed that A and U are the same for all units in the sample.

The agency releases $l = 1, \ldots, m$ partially synthetic data sets including all n sampled units. Let $z_{jk}^{(l)}$ be the released value for unit j on variable k in partially synthetic data set l. We assume that the agency synthesizes values only in A to prevent re-identifications. Let $\mathbf{z}_{jA}^{(l)}$ and $\mathbf{z}_{jU} = \mathbf{y}_{jU}$ be the released values of the available variables and unavailable variables, respectively, for unit j in partially synthetic data set l. The sets A and U are the same as those used to partition the $\mathbf{y}_j$. The available variables can be further divided into $\mathbf{z}_{jA}^{(l)} = (\mathbf{z}_{jS}^{(l)}, \mathbf{z}_{jD})$. The $\mathbf{z}_{jS}^{(l)}$ comprises variables in A whose values are replaced with synthetic data drawn from probability distributions. The $\mathbf{z}_{jD}$ comprises variables in A whose values are not synthesized, for example available variables for which $z_{jk} = y_{jk}$ or available variables that are re-coded. Let $\mathbf{z}_{jS} = \left(\mathbf{z}_{jS}^{(1)}, \mathbf{z}_{jS}^{(2)}, \ldots, \mathbf{z}_{jS}^{(m)}\right)$; let $\mathbf{Z}_S$ be the collection of $\mathbf{z}_{jS}$ for all n units in the sample; let $\mathbf{Z}_D$ be the collection of $\mathbf{z}_{jD}$ for all n units in the sample; and, let $\mathbf{Z}$ be all released data. Finally, let $\mathbf{Y}_S$ be all n units' original values of the variables that were synthesized.

The agency generating synthetic data might release meta-data about the synthesis process to help analysts determine if their analyses are reasonably supported in the synthetic data. Let M represent the meta-data released about the models used to generate the synthetic data. The M could include, for example, the code for the models used to generate the synthetic data. Let R represent the meta-data released about why records were selected for synthesis. For example, R could specify that all records that are uniques and duplicates with respect to $\mathbf{y}_A$ undergo synthesis. Either M or R could be empty.

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 $\mathbf{Z}$. The column $k = 0$ in $\mathbf{t}$ contains a unique identifier for that record. The intruder's goal is to match unit j in $\mathbf{Z}$ to the target when $z_{j0} = t_0$, and not to match when $z_{j0} \neq t_0$ for any $j \in \mathbf{Z}$. We assume that $\mathbf{t}$ has some of the same variables as $\mathbf{Z}$—otherwise there is little opportunity for the intruder to match—and we allow $\mathbf{t}$ to include partial information on values. For example, an intruder's $\mathbf{t}$ can include the information that the income for some unit j is above \$100,000, even though the intruder does not know the unit's exact income. The variables of $\mathbf{t}$ that correspond to the variables in $\mathbf{z}_S$ are written as $\mathbf{t}_S$. As done by Fienberg *et al.* (1997), we assume that $\mathbf{t} = \mathbf{y}_{jA}$ for some unit j in the population, although not necessarily for a unit in $\mathbf{Z}$. That is, relative to the sampled values, the intruder's values are not measured with error. This assumption may not be true in practice, but it provides upper limits on the identification probabilities and greatly simplifies calculations. Finally, we assume that users can correctly link the records across synthetic data sets, for example by using record numbers (if released) or by matching on non-synthesized values in each $\mathbf{Z}^{(l)}$.

Let J be a random variable that equals j when $z_{j0} = t_0$ for $j \in \mathbf{Z}$ and equals $n + 1$ when $z_{j0} = t_0$ for some $j \notin \mathbf{Z}$. The intruder thus seeks to calculate the $Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R)$ for $j = 1, \ldots, n+1$. The intruder then decides whether or not the maximum of the identification probabilities for $j = 1, \ldots, n$ is large enough to declare an identification. Because the intruder does not know the actual values in $\mathbf{Y}_S$, the intruder should integrate over its possible values when computing the match probabilities. Hence, we have

$$Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R) = \int Pr(J = j|\mathbf{t}, \mathbf{Z}, \mathbf{Y}_S, M, R) Pr(\mathbf{Y}_S|\mathbf{t}, \mathbf{Z}, M, R) d\mathbf{Y}_S. \quad (1)$$

These probabilities can be determined from assumptions about the knowledge and behavior of the intruder, as we now discuss.

2.1 Evaluating $Pr(J = j|\mathbf{t}, \mathbf{Z}, \mathbf{Y}_S, M, R)$

Given $\mathbf{Y}^s$, the intruder would toss out the synthetic data, $\mathbf{Z}_S$, and use $(\mathbf{Z}_D, \mathbf{Y}_S)$ to attempt re-identifications. We assume that the unavailable variables do not help with re-identifications given $(\mathbf{Z}_D, \mathbf{Y}_S)$. Hence, we have

$$Pr(J = j|\mathbf{t}, \mathbf{Z}, \mathbf{Y}_S, M, R) = Pr(J = j|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S). \quad (2)$$

For any variable k in $\mathbf{z}_{jD}$, when the value of t_k is not consistent with the value of the released z_{jk}, the $Pr(J = j|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = 0$. For example, suppose $\mathbf{t}$ belongs to a 37 year old, married woman. When sex is not altered, all males have $Pr(J = j|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = 0$. When age is released in five year intervals rather than exact integers, all people with ages outside 35 to 39 have zero probabilities.

For variables in $\mathbf{z}_{jS}$, the intruder's actions depend on the nature of the variables. For categorical variables, the intruder treats the $\mathbf{y}_{jS}$ as if it were a part of $\mathbf{z}_{jD}$; that is, he matches directly on $\mathbf{y}_{jS}$. For example, if marital status is synthesized, all women whose marital status in $\mathbf{Y}_S$ differs from married have zero probabilities. For numerical or continuous variables, the intruder also could seek an exact match. However, because the intruder must estimate the values in $\mathbf{Y}_S$, his estimates are very likely to differ from the corresponding values in $\mathbf{t}_S$. This would lead to zero probabilities for most if not all of the records in $\mathbf{Z}$. Alternatively, the intruder can assign zero probabilities to all but a set of plausible matches. For example, among all

candidate records for which the categorical portions of $\mathbf{t}$ and $(\mathbf{z}_{jD}, \mathbf{y}_{jS})$ match exactly, the intruder can define plausible matches as those record(s) whose numerical components of $\mathbf{y}_{jS}$ are within some acceptable Euclidean or Mahalanobis distance from the corresponding $\mathbf{t}_S$. All units not in the set of plausible matches have zero probabilities.

When $\mathbf{t}$ is known to belong to a unit in $\mathbf{Z}$, for example when all records of a census are released or when another version of the data set has been previously released, the $Pr(J = n+1|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = 0$. And, for $j \leq n$, the $Pr(J = j|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = 1/n_t$, where n_t is the number of units in $(\mathbf{Z}_D, \mathbf{Y}_S)$ with $\mathbf{y}_{jA}$ consistent with $\mathbf{t}$, either as exact or plausible matches. When $n_t = 0$ in this setting, which occurs when no values in $(\mathbf{Z}_D, \mathbf{Y}_S)$ match the corresponding values in $\mathbf{t}$, we set $Pr(J = j|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = 1/n_t^*$, where n_t^* is the number of units in $\mathbf{Z}_D$ with $\mathbf{z}_{jD}$ consistent with $\mathbf{t}_D$.

It may be prudent to assume the intruder knows particular target units are in $\mathbf{Z}$, even when the collected data are not a census. For example, in a survey of households, neighbors may know that an interviewer visited a sampled household. Since all records in the sample are included in $\mathbf{Z}$, the neighbors know that household must be in $\mathbf{Z}$. Alternatively, someone with inside information about which units are in the released data may attempt to discredit the agency. Even when knowledge that particular targets are in $\mathbf{Z}$ is difficult to come by, setting $Pr(J = n+1|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = 0$ results in conservative measures of identification disclosure risks.

The calculations are more complicated when $Pr(J = n+1|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) \neq 0$. Let N_t be the number of units in the population that would have $(\mathbf{z}_{jD}, \mathbf{y}_{jS})$ consistent with $\mathbf{t}$ if they were included in $\mathbf{Z}$. Then, $Pr(J = j|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = 1/N_t$ for units whose $(\mathbf{z}_{jD}, \mathbf{Y}_{jS})$ are consistent with $\mathbf{t}$, and $Pr(J = n+1|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = (N_t - n_t)/N_t$. The agency, and the intruder, may be able to determine N_t from census totals, particularly when $\mathbf{Z}_A$ contains only categorical, demographic characteristics. When N_t is not known, it must be estimated from available sources. One approach is to set N_t equal to the sum of the survey weights for all units in $\mathbf{Z}$ whose $(\mathbf{z}_{jD}, \mathbf{y}_{jS})$ are consistent with $\mathbf{t}$. The survey-weighted estimate could poorly estimate N_t, especially when units like $\mathbf{t}$ are rare in the collected data. Alternatively, N_t can be estimated using model-based approaches, such as those used to estimate the number of population uniques. These include, among others, Bethlehem *et al.* (1990), Greenberg and Zayatz (1992), Skinner (1992), Skinner *et al.* (1994), Chen and Keller-McNulty (1998), Fienberg and Makov (1998), Samuels (1998), Pannekoek (1999), Dale and Elliot (2001), and Elamir and Skinner (2006). If $\mathbf{Z}_A$ contains no variables, the $Pr(J = j|\mathbf{t}) = 1/N$ for $j \leq n$, and $Pr(J = n+1|\mathbf{t}) = (N-n)/N$, where N is the number of units in the population.

2.2 Evaluating $Pr(\mathbf{Y}_S|\mathbf{t}, \mathbf{Z}, M, R)$

The construction in (1) suggests a Monte Carlo approach to estimating the $Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R)$. First, we sample a value of $\mathbf{Y}_S$ from $Pr(\mathbf{Y}_S|\mathbf{t}, \mathbf{Z}, M, R)$. Let $\mathbf{Y}^{new}$ represent one set of simulated values. Second, using the values of the N_t and n_t computed from $\mathbf{Y}^{new}$, we compute $Pr(J = j|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S = \mathbf{Y}^{new}, M, R)$ as described in Section 2.1. We iterate this two-step process I times, where ideally h is large, and estimate the quantity in (1) as the average of the resultant h values of $Pr(J = j|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S = \mathbf{Y}^{new}, M, R)$.

The key to this step is the model used to generate the plausible values of $\mathbf{Y}_S$. Here the details in M and R play central roles. We consider three scenarios about M and R that are representative of what might occur in practice. First, the agency releases nothing about the synthetic data generation process, i.e. M and R are empty. Second, the agency releases the exact specification of the models without parameter estimates and releases nothing about why records are selected for synthesis, i.e. M has information and R is empty. Third, the agency releases the exact specification of the models including posterior distributions of the parameter estimates and explains why records are selected for synthesis, i.e. M and R have complete information. There are other possibilities, but these examples illustrate the computations.

When M and R are empty, the intruder's primary source of information about the $\mathbf{Y}_S$ is the $\mathbf{Z}_S$. One approach is to treat the values in $\mathbf{Z}_S$ as equally likely, plausible values of $\mathbf{Y}_S$. That is, the intruder can

assume that, for $l = 1, \ldots, m$, the $Pr(\mathbf{Y}_S = \mathbf{Z}_S^{(l)} | \mathbf{t}, \mathbf{Z}, M, R) = 1/m$. The values of the $\mathbf{Z}_S^{(l)}$ are used to compute each $Pr(J = j | \mathbf{t}, \mathbf{Z}_D, \mathbf{Z}_S^{(l)})$. Alternatively, for any synthesized categorical variable k, the intruder can set y_{jk} equal to the most frequent value in $\mathbf{z}_{jk}$. If several values are tied for the maximum frequency, the intruder can pick one $z_{jk}^{(l)}$ at random from these tied values. In this approach, the intruder only uses the most frequent values in $Pr(J = j | \mathbf{t}, \mathbf{Z}, \mathbf{Y}_S)$, ignoring the uncertainty in the estimate of $\mathbf{Y}_S$.

The intruder might have prior knowledge about the relationships between $\mathbf{Y}_S$ and the other components of $\mathbf{Y}$. Of course, there are infinite numbers of representations of intruder knowledge. To simplify the problem, we adopt the conservative assumption that the intruder with prior knowledge uses the same form of the model for $\mathbf{Y}_S$ as was used to generate $\mathbf{Z}_S$, without knowing the parameter estimates for that model. Equivalently, we assume that scenarios where the intruder has prior knowledge are equivalent to scenarios where the agency releases information about the synthesis model in M, without any parameter estimates. Agencies may release this information in M anyway as meta-data.

When M includes only the specification of the synthesis models, and R is empty, the intruder can fit the models with $\mathbf{Z}$ to estimate the posterior distributions of the model parameters. Using these posterior distributions, the intruder then repeatedly simulates values of $\mathbf{Y}_S$ using predictive simulation. That is, the intruder samples values of the parameters from their posterior distribution, then samples values of $\mathbf{Y}^{new}$ using the sampled parameters and the imputation models described in M. Alternatively, to streamline the computations, the intruder might ignore the uncertainty in the parameter estimates and use their average across the m data sets in the predictive distribution for simulating $\mathbf{Y}^{new}$. For either strategy, the intruder uses the simulated $\mathbf{Y}^{new}$ when computing each $Pr(J = j | \mathbf{t}, \mathbf{Z}, \mathbf{Y}^{new})$. The role of the number of synthetic data sets is apparent in this setting: as m increases, the uncertainty in the parameter estimates decreases, which in turn should decrease variability in the $\mathbf{Y}^{new}$ and improve the matching.

When the agency releases complete information about the synthesis, so that M includes the posterior distributions of the parameters of the synthesis models, the intruder does not need to estimate models with $\mathbf{Z}$. Instead, the predictive simulations are based on draws from the released posterior distributions or, streamlining computations, based on the released posterior modes of these distributions. There is no additional uncertainty due to estimating parameters with finite m. In fact, the magnitude of m is irrelevant when M includes complete information about the synthesis.

The details about R impact the plausibility of certain values of $\mathbf{Y}_S$. The intruder can eliminate values of $\mathbf{Y}_S$ that correspond to values inconsistent with R. This can be done via prior distributions in the predictive simulations. For example, if the agency releases the fact that only minorities' races were synthesized, the intruder can force the prior probability of non-minority race to equal zero when simulating $\mathbf{Y}^{new}$. Or, if the agency reveals that only sample unique records undergo synthesis, the intruder can place zero prior probability on simulating combinations of $(\mathbf{Z}_D, \mathbf{Y}^{new})$ equal to the $\mathbf{Y}_A$ of unaltered records and force all simulated records to be unique.

In some applications, the information in R may not be especially helpful for identifications. For example, if the agency simulates all values of sex and race, the R provides no additional information over what is released in $(\mathbf{Z}, M)$. As another example, suppose the agency simulates all combinations of identifiers that appear no more than five times. With sufficiently large dimensions of $\mathbf{z}_{jS}$, intruders may not gain much by building that prior information into the predictive simulations.

3 Illustrative Example

To illustrate the computations of identification disclosure risk, we synthesize data from the 1987 Survey of Youth in Custody (Lohr, 1999). This survey was used by Mitra and Reiter (2006) to illustrate the role of survey weights in partially synthetic data. 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. For reasons related to data cleaning (Mitra and Reiter, 2006), we deleted all the youths in four facilities, leaving a total of 2562 youths.

We suppose that the set A, i.e. the variables known by the intruder, contains the youth's facility (46 levels), race (five levels), and ethnicity (two levels). We suppose that all other variables are in the set U, i.e. available to users only in the released data. There are 64 youths who have unique combinations of facility, race, and sex.

To reduce the risk of identifications, we synthesize all values of facility and race, without altering other variables. 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 simulated conditional on the values of the synthetic facility indicators.

We suppose that intruders know the values of facility, race, and ethnicity for all units in the survey and would like to identify records from the synthetic data sets. That is, for all targets $\mathbf{t}$ in the sample, we assume the $Pr(J = 2562 + 1|\mathbf{t}, \mathbf{Z}_D, \mathbf{Y}_S) = 0$. For any $\mathbf{t}$ in the sample, we compute $Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R)$ for $j = 1, \ldots, n$. We assume each target's probability is computed independently of other targets' probabilities, i.e. we match with replacement. We determine risk measures for the entire data set using the functions of these match probabilities proposed by Reiter (2005a). The first measure, which we call *perceived match risk*, equals the number of target records for which the highest value of $Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R)$, where $1 \leq j \leq n$, exceeds some threshold deemed too risky. For illustration, we set this threshold to be 0.20. To describe the second and third measures, let c_j be the number of records in the data set 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; and, let $K_j = 1$ when $c_j I_j = 1$ and $K_j = 0$ otherwise. The second risk, which we call *expected match risk*, equals $\sum_j (1/c_j) I_j$. 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 third risk measure, which we call *true match risk*, equals $\sum_j K_j$.

We now evaluate the dependence of the disclosure risk on the number of synthetic data sets—setting m to equal 2, 3, or 10—and the released information about the synthesis models—setting M to be empty, to provide the exact specification of the synthesis models without estimates of the model parameters, or to provide the exact specification models with full information on the posterior distributions of the model parameters. We do not evaluate the dependence of risk on R, since all records for facility and race are synthesized. To mimic intruder behavior, we apply the general disclosure attack strategies outlined in Section 2.1.

When M is empty, we investigate two attack strategies. In the first strategy, which we call the probability-based approach, for each target $\mathbf{t}_j$ we determine the number of records in each $\mathbf{Z}^{(l)}$ with the same values of facility, race, and ethnicity as $\mathbf{t}_j$. We assign equal probability to each of the matches. When there are no exact matches for record j in $\mathbf{Z}^{(l)}$, in that data set we assign equal probability to any records that share the same ethnicity as the target. Once the probabilities associated with each $\mathbf{Z}^{(l)}$ are determined, we average them across synthetic data sets to obtain each target's $Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R)$. In the second strategy, which we call the mode-based approach, for each record j we compute the most frequently occurring values of facility and race in $\mathbf{z}_{jS}$, treating them as the best guess of the true values. For each target, we then determine the number of records with synthetic best guesses and ethnicity that exactly match the target's values. We assign equal probability to any matches.

When M includes details of the synthesis models but not parameter estimates, the attack strategy involves repeated simulation of facility and race, $\mathbf{Y}_S$, using the released synthesis models. We estimate the parameters of the synthesis models by fitting the models on each $\mathbf{Z}^{(l)}$. Specifically, let $\boldsymbol{\beta}^{(l)}$ and $\boldsymbol{\Sigma}^{(l)}$ be respectively the maximum likelihood estimates of the multinomial regression coefficients and their covariance

matrix computed from the lth synthetic data set, for $l = 1, \ldots, m$. Following Reiter (2003, 2005b), let

$$\bar{\boldsymbol{\beta}}_m = \sum_{l=1}^{m} \boldsymbol{\beta}^{(l)}/m \quad (3)$$

$$\bar{\boldsymbol{\Sigma}}_m = \sum_{l=1}^{m} \boldsymbol{\Sigma}^{(l)}/m \quad (4)$$

$$\mathbf{B}_m = \sum_{l=1}^{m} (\boldsymbol{\beta}^{(l)} - \bar{\boldsymbol{\beta}}_m)(\boldsymbol{\beta}^{(l)} - \bar{\boldsymbol{\beta}}_m)'/(m-1) \quad (5)$$

$$\boldsymbol{T}_m = \bar{\boldsymbol{\Sigma}}_m + \mathbf{B}_m/m \quad (6)$$

The point estimate of the regression coefficients equals $\bar{\boldsymbol{\beta}}_m$. The estimate of the covariance associated with these parameter estimates equals $\mathbf{T}_m$.

With these parameter estimates, the intruder can simulate values of $\mathbf{Y}_S$ in two ways. First, he can assume that the true coefficients equal $\bar{\boldsymbol{\beta}}_m$, that is ignore uncertainty in the $\bar{\boldsymbol{\beta}}_m$. Second, he can repeatedly draw values of the coefficients from their posterior distributions. In our setting, we approximate these posterior distributions as normal distributions with mean $\bar{\boldsymbol{\beta}}_m$ and covariance $\boldsymbol{T}_m$. We investigate both approaches, simulating $h = 100$ values of the vector of facilities and races from the synthesis models for each approach. For each target,we average the match probabilities across the 100 simulated data sets to obtain $Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R)$, for $1 \leq j \leq n$.

When M includes details of the synthesis models and the posterior distributions of the model parameters, the intruder uses this distribution to simulate values of $\mathbf{Y}_S$. There is no need to estimate the parameters of the synthesis models from the $\mathbf{Z}$. As before, the intruder can treat the posterior mode as if it equals the true values, or the intruder can repeatedly draw values of the model parameters from their posterior distributions. We investigate both approaches, simulating $h = 100$ new values of the vector of facilities and races. For each target, we average the match probabilities across the 100 simulated data sets to obtain $Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R)$, for $1 \leq j \leq n$.

Table 1 summarizes the risk measures for the different scenarios and attack strategies. Results are based on one simulation run for each scenario. For each level of m, the same synthetic data sets are evaluated across the three specifications of M. We generated another set of synthetic data sets for each scenario, and the general trends in the table hold.

When M is empty, the perceived matching risk decreases with m in the probability-based approach. This is because each $Pr(J = j|\mathbf{t}, \mathbf{Z}, M, R)$ is an average of m probabilities, and the variance of any average decreases as m increases. Put another way, the averages based on small m are noisy estimates of the corresponding averages based on infinite m, and the chances of getting extreme estimated averages (larger than 0.20 in this case) decrease with m. This is not the case for the mode-based approach, since there is no averaging of probabilities. The numbers of expected and true matches for the probability-based approach exceed those from the mode-based approach for all m. Because of this dominance, for the remainder of this section we discuss only the probability-based approach when M is empty.

When M includes the synthesis models (with or without parameter estimates), we typically obtain higher expected and true match rates using point estimates of the parameters than sampling from the posterior distributions of the parameters, although the differences are not dramatic. The point estimates are in fact the maximum likelihood estimates; hence, the observed data values are more likely to be simulated when using the point estimates than when using drawn parameter values, which could be far from the observed values. Thus, for the remainder of this section, we discuss only the point-estimate based methods when M includes the synthesis models.

Table 1 clearly illustrates the impact of releasing additional information about M. The numbers of expected and true matches are lowest when M is empty and largest when M contains everything. When

Information in M	Value of m	Type of Matching Risk Perceived	Expected	True
Empty				
Probability-based	2	143	17.8	3
	3	115	21.9	13
	10	12	24.1	24
Mode-based	2	142	11.8	0
	3	167	11.6	1
	10	142	15.2	2
Synthesis models, no $f(\beta)$				
Fix β at approximate mode	2	18	30.6	30
	3	9	27.7	27
	10	4	31.4	31
Simulate β from approximate posterior	2	3	19.0	19
	3	1	30.0	30
	10	3	29.1	29
Synthesis models and $f(\beta)$				
Fix β at mode of $f(\beta)$	–	10	48.2	48
Simulate β from $f(\beta)$	–	5	39.0	39

Table 1: Summary of risk measures under different scenarios for M and m. The β represents the true values of these parameters in the population, and the $f(\beta)$ represents the released posterior distribution of the parameters of the synthesis model.

M is empty, increasing m increases the numbers of expected and true matches. When M includes details of the synthesis models without parameter estimates, increasing m has unclear impact on risk for these modest values of m. This suggests that, for these data, the intruder does not gain much information for attacking when ten rather than two data sets are released. We expect risks to increase as m gets large, since setting $m = \infty$—which corresponds to releasing everything in M—results in larger risks than setting m to be modest.

Most of the true matches are for records without unique combinations of facility, race, and ethnicity. For example, when using the point estimate approach with M containing everything, only 5 of the 48 true matches are for sample uniques. Not surprisingly, the number of matched uniques is largest when M includes everything and smallest (typically one match) when M is empty. Hence, even if intruders focus solely on targets with unique values of $\mathbf{t}_A$, the identification disclosure risks remain low for this combination of synthesis strategy and knowledge of the intruder.

4 CONCLUDING REMARKS

As these results indicate, agencies and other data disseminators considering the release of partially synthetic data should should account for all released information when assessing identification disclosure risks. When the match probabilities are too large, the data disseminator has several options, including synthesizing more values, reducing m, and restricting the information in M and R. The impacts of these options on identification disclosure risk depends on the particulars of the file to be protected and disseminated. Further

empirical studies of the effectiveness of various options, as well as general guidance on risk reduction strategies as functions of the observed data structure, are important areas of future research.

The specification of $(\mathbf{Z}, M, R)$ also impacts the usefulness of the released data, often called data utility (Duncan *et al.*, 2001; Gomatam *et al.*, 2005; Karr *et al.*, 2006; Woo *et al.*, 2006). Ideally, the data disseminator quantifies the utility associated with any proposed release strategy. The data set with the best balance of risk and utility is ultimately selected for release. For partially synthetic data, typically utility is assessed by comparing inferential quantities—such as confidence intervals for regression coefficients—computed with the synthetic data to the corresponding quantities computed with the observed data. These measures account for the magnitude of m through the estimates of uncertainty in the inferences, but they do not formally incorporate the nature of M and R. Developing quantifiable metrics for the utility of the meta-data in M and R is another area for future research.

References

Abowd, J. M. and Lane, J. I. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 282–289. New York: Springer-Verlag.

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.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association* **85**, 38–45.

Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* **14**, 79–95.

Dale, A. and Elliot, M. (2001). Proposals for 2001 samples of anonymized records: an assessment of disclosure risk. *Journal of the Royal Statistical Society, Series A* **164**, 427–447.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Tech. rep., U.S. National Institute of Statistical Sciences.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association* **81**, 10–28.

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

Elamir, E. and Skinner, C. J. (2006). Record level mieasures of disclosure risk for survey microdata. *Journal of Official Statistics* **22**, 525–539.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics* **14**, 361–372.

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.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. *Statistical Science* **20**, 163–177.

Greenberg, B. V. and Zayatz, L. V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica* **46**, 33–48.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60**, 224–232.

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.

Little, R., 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 Persepectives*, 141–152. New York: John Wiley & Sons.

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

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

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrar, ed., *Privacy in Statistical Databases 2006 (Lecture Notes in Computer Science)*, 177–188. New York: Springer-Verlag.

Pannekoek, J. (1999). Statistical methods for some simple disclosure limitation rules. *Statistica Neerlandica* **53**, 55–67.

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

Reiter, J. P. (2004a). New approaches to data dissemintation: A glimpse into the future (?). *Chance* **17**, 3, 12–16.

Reiter, J. P. (2004b). 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). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.

Reiter, J. P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.

Samuels, S. M. (1998). A Bayesian species-sampling-inspired approach to the uniques problem in microdata. *Journal of Official Statistics* **14**, 373–384.

Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics* **10**, 31–51.

Skinner, C. J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica* **46**, 21–32.

Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2006). Global measures of data utility for microdata masked for disclosure limitation. Tech. rep., National Institute of Statistical Sciences.