Estimating Risks of Identification Disclosure in Microdata

Jerome P. REITER*

Abstract

When statistical agencies release microdata to the public, malicious users (intruders) may be able to link records in the released data to records in external databases. Releasing data in ways that fail to prevent such identifications may discredit the agency or, for some data, constitute a breach of law. To limit disclosures, agencies often release altered versions of the data; however, there usually remain risks of identifications. This article applies and extends the framework developed by Duncan and Lambert for computing probabilities of identification for sampled units. It describes methods tailored specifically to data altered by recoding and topcoding variables, data swapping, or adding random noise–and combinations of these common data alteration techniques–which agencies can use to assess threats from intruders who possess information on relationships among variables and the methods of data alteration. Using data from the Current Population Survey, the article illustrates a step-by-step process for evaluating identification disclosure risks for competing releases under varying assumptions of intruders' knowledge. Risk measures are presented for individual units and for entire data sets.

KEY WORDS: Confidentiality; Public use data; Record linkage; Survey.

*Jerome Reiter is Assistant Professor of the Practice, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251 (E-mail: jerry@stat.duke.edu). The author thanks Donald Rubin, George Duncan, Diane Lambert, and Eleanor Singer for inspiration and helpful discussions, and the editor, associate editor, and referees for their suggestions and comments. This research was supported by the National Academy of Sciences and was presented at a workshop in October 2003 organized by the Academy's Panel on Confidential Data Access for Research Purposes.

1 INTRODUCTION

When an agency releases microdata (i.e. data on individual units) to the public, it seeks to limit the risk that malicious users are able to identify sampled units in the released data. Such users, hereafter called intruders, can attempt identifications by linking records in the released data to records from external databases. Agencies releasing data in ways that fail to prevent identifications can face serious consequences. Their reputations may be damaged, which in turn can diminish potential respondents' willingness to participate in studies run by the agency. In some cases, unsafe releases break laws that guarantee the confidentiality of respondents' answers (Federal Committee on Statistical Methodology, 1978, 1994; Willenborg and de Waal, 2001; Wallman and Harris-Kojetin, 2004).

To reduce intruders' confidence that attempted links are correct, agencies typically alter microdata before public release. Four of the most commonly employed techniques include: (i) recoding variables into coarse categories, such as releasing only five year intervals for age; (ii) topcoding numerical data, for example reporting all incomes above $100,000 as "$100,000 or more"; (iii) swapping some units' data values with other units' data values; and, (iv) adding random noise to numerical data values. Ideally, the released data maintain a high level of utility; that is, they do not sacrifice too much of the information contained in the collected data. Generally, however, data modifications that increase protection decrease utility (Duncan *et al.*, 2001). Gauging this risk-utility tradeoff necessitates quantitative measures of disclosure risks.

A framework for measuring identification disclosure risks was developed by Duncan and Lambert (1986, 1989) and Lambert (1993). They propose that agencies model the behavior of intruders to obtain probabilities of identification for sampled units, and quantify sources of uncertainty about those probabilities using Bayesian approaches. Unfortunately, the literature on disclosure limitation lacks illustrative applications of the Duncan-Lambert approach on genuine data altered by common disclosure limitation techniques. One exception is Fienberg *et al.* (1997), who describe a Bayesian approach for purely continuous data when adding random noise, although their simulation study uses generated rather than genuine data.

Perhaps because of this dearth of published examples, many agencies do not use formal Duncan-Lambert approaches to assess disclosure risks. Instead, many agencies measure risk as the estimated number of released records that are unique in the population. Some perform reidentification experiments (Federal Committee on Statistical Methodology, 1994, pp. 79 - 80) by using record linkage software to investigate which respondents are most at risk for identification in potential releases. These two approaches are valuable tools for risk analyses; however, as typically implemented, they do not fully mimic the behavior of intruders who know and utilize multivariate relationships in the data and the methods of data alteration, and who quantify uncertainty using Bayesian methods. Hence, they may not fully measure risks from such intruders.

This article applies and extends the Duncan-Lambert framework using data from the Current Population Survey, thereby providing an implementation of this approach on genuine data. It presents methods for obtaining probabilities of identification that utilize multivariate relationships in the data and are tailored specifically to recoding and topcoding, swapping, and adding random noise–as well as combinations of these techniques–that can help agencies assess the threats from sophisticated intruders. The article illustrates a step-by-step process of analyzing the impact on disclosure risk of competing disclosure limitation strategies under differing assumptions of intruders' knowledge and behavior.

The remainder of the paper is organized as follows. Section 2 summarizes approaches to measuring identification disclosure risks in microdata. Section 3 outlines methods that can be used to assess identification probabilities when released data are altered by recoding and topcoding, data swapping, or noise addition. Section 4 illustrates the application of these methods to data from the Current Population Survey, presenting both unit-specific and entire-file measures of identification risk. Section 5 concludes with a general discussion of disclosure risks and limitation methods.

2 MEASURES OF IDENTIFICATION DISCLOSURE RISKS

Strategies for measuring identification risks in microdata can be broadly classified into two categories: 1) estimating the number of records released in the sample whose characteristics are unique in the population, and 2) estimating the probabilities that records possessed by intruders can be identified from the released data.

2.1 Population uniques

Several authors have proposed disclosure risk measures that are some function of 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), and Dale and Elliot (2001). Uniqueness is relevant because population uniques generally have higher risks of identification disclosure than non-uniques. Indeed, it has been suggested that uniqueness is a necessary condition for identification (e.g., Skinner, 1992). Typically, the number of population uniques in the sample is not known and must be estimated by the agency.

While useful, population uniqueness has some limitations as a measure of disclosure risk. First, it does not account for the nature of the information possessed by the intruder. For example, when the intruder knows a particular target is in the sample and knows values of that target's record, the intruder can identify the target when it is a sample unique, even if it is not a population unique. Second, in some data settings there exist a large number of sample and population uniques, especially when the data contain variables that can be treated as continuous. It is not clear that the number of uniques provides much information in these settings. Third, using the number of population uniques may not allow the agency to gauge accurately the effect of some statistical disclosure limitation procedures. For example, suppose an agency releases values of a continuous, key identifier that have been perturbed with Gaussian random noise. The resulting perturbed records may contain just as many estimated population uniques as the original sample contains. Lastly,

estimating the number of population uniques accurately is difficult to do in studies where the sampling fraction is small, so that the measures could be misleading.

2.2 Probabilities of Identification

Other authors have proposed that agencies attempt to link released records with target records, either through direct matching using external databases (Paass, 1988; Blien *et al.*, 1992; Federal Committee on Statistical Methodology, 1994; Yancey *et al.*, 2002) or indirect matching using the existing database (Spruill, 1982; Duncan and Lambert, 1986, 1989; Lambert, 1993; Fienberg *et al.*, 1997; Skinner and Elliot, 2002). In both approaches, the agency essentially mimics the behavior of an intruder trying to match released records to target records.

These approaches address many of the shortcomings of relying on population uniques. They can permit agencies to account for varying degrees of intruder knowledge, are equally appropriate for continuous and categorical data, and can be applied to assess the effects of statistical disclosure limitation techniques. However, they may require strong assumptions about intruder behavior, which if wrong could lead to inaccurate measures of disclosure. They may be expensive to implement, both operationally and computationally. For example, it may be difficult and time-consuming for agencies to obtain external files for use with record linkage techniques like those of Felligi and Sunter (1969). Finally, as mentioned in the introduction, there has been a dearth of evidence and illustrations of these approaches on real-data.

The indirect probabilistic matching approach, which is the Duncan-Lambert approach, avoids the complexity of obtaining external databases, while maintaining the flexibility of modeling intruder behavior. As we shall show, it also easily incorporates information about relationships among variables and the statistical disclosure limitation techniques applied to the data, which can help agencies mimic sophisticated intruders' behavior. We now turn to applying this approach.

3 DESCRIPTION OF DUNCAN-LAMBERT APPROACH AND METHODS

For a collection of n sampled units $\mathcal{S}$, let y_{jk} be the collected data for unit j on variable k, for $k = 0, \ldots, d$ and $j \in \mathcal{S}$. 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 $\mathbf{y}_j = (y_{j1}, \ldots, y_{jd})$ into two sets of variables. Let $\mathbf{y}_j^A$ be the vector of variables available to users from external databases, such as demographic or geographic attributes. And, let $\mathbf{y}_j^U$ be the vector of variables that are unavailable to users except in the released data. The compositions of A and U are determined by the agency for the particular $\mathcal{S}$, based on knowledge of what information exists in external databases. It is assumed that A and U are the same for all units in $\mathcal{S}$.

The agency releases data for $r \leq n$ of the units in $\mathcal{S}$, possibly altered for disclosure limitation. Let z_{jk} be the released value for unit j on variable k. Let $\mathbf{z}_j^A$ and $\mathbf{z}_j^U$ be the released values of the available variables and unavailable variables, respectively. 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 $\mathbf{z}_j^A = (\mathbf{z}_j^{Ap}, \mathbf{z}_j^{Ad})$. The $\mathbf{z}_j^{Ap}$ comprises variables in A whose values are altered from those in $\mathbf{y}_j^A$ by a stochastic perturbation method, such as adding noise or swapping data. The $\mathbf{z}_j^{Ad}$ comprises variables in A whose values do not undergo stochastic perturbation, for example available variables that are globally recoded or available variables for which $z_{jk} = y_{jk}$. Let $\mathbf{z}_j^C = (\mathbf{z}_j^{Ap}, \mathbf{z}_j^U)$, where C indicates the intruder cannot match on these variables with 100% certainty because they have been perturbed or are not in the intruder's target record. Finally, let $\mathbf{Z} = (\mathbf{Z}^A, \mathbf{Z}^U)$ be the matrix of all released data.

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 $\mathbf{z}^{Ad}$ are written as $\mathbf{t}^{Ad}$, and likewise for $\mathbf{t}^{Ap}$. As done by Fienberg *et al.* (1997), we assume that $\mathbf{t} = \mathbf{y}_j^A$ 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.

Let J be a random variable that equals j when $z_{j0} = t_0$ for $j \in \mathbf{Z}$ and equals $r+1$ when $z_{j0} = t_0$ for some $j \notin \mathbf{Z}$. The intruder thus seeks to calculate the $Pr(J = j|\mathbf{t}, \mathbf{Z})$ for $j = 1, \ldots, r+1$. He or she then would decide whether or not any of the identification probabilities for $j = 1, \ldots, r$ are large enough to declare an identification. The $Pr(J = j|\mathbf{t}, \mathbf{Z})$ can be calculated using Bayes rule:

$$Pr(J = j|\mathbf{t}, \mathbf{Z}) = \frac{Pr(\mathbf{Z}^C|J = j, \mathbf{t}, \mathbf{Z}^{Ad})Pr(J = j|\mathbf{t}, \mathbf{Z}^{Ad})}{\sum_{j=1}^{r+1} Pr(\mathbf{Z}^C|J = j, \mathbf{t}, \mathbf{Z}^{Ad})Pr(J = j|\mathbf{t}, \mathbf{Z}^{Ad})} \tag{1}$$

We can split calculations into the available and unavailable components of $\mathbf{z}_j$. To reduce notation, we write this split only for the first term in the numerator of (1):

$$\begin{aligned} Pr(\mathbf{Z}^C|J = j, \mathbf{t}, \mathbf{Z}^{Ad}) &= Pr(\mathbf{z}_1^C, \ldots, \mathbf{z}_{j-1}^C, \mathbf{z}_{j+1}^C, \ldots \mathbf{z}_r^C|\mathbf{z}_j^C, J = j, \mathbf{t}, \mathbf{Z}^{Ad}) \\ &\times Pr(\mathbf{z}_j^U|\mathbf{z}_j^{Ap}, J = j, \mathbf{t}, \mathbf{Z}^{Ad})Pr(\mathbf{z}_j^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad}) \end{aligned} \tag{2}$$

When $j = r+1$, it is convenient not to use (2) and instead calculate directly from $Pr(J = r+1|\mathbf{t}, \mathbf{Z}) = Pr(\mathbf{Z}^C|J = r+1, \mathbf{t}, \mathbf{Z}^{Ad})Pr(J = r+1|\mathbf{t}, \mathbf{Z}^{Ad})$. The components of (1) and (2) can be determined from assumptions about the knowledge and behavior of the intruder, as we now discuss.

3.1 Evaluating $Pr(J = j|\mathbf{t}, \mathbf{Z}^{Ad})$

For any variable k in $\mathbf{z}_j^{Ad}$, when the value of t_k is not consistent with the value of the released z_{jk}, the $Pr(J = j|\mathbf{t}, \mathbf{Z}^{Ad}) = 0$. For example, suppose $\mathbf{t}$ belongs to a 37 year old woman with property taxes of \$10,000. When sexes are not altered in the released data, all males have $Pr(J = j|\mathbf{t}, \mathbf{Z}^{Ad}) = 0$. When age is released in five year intervals rather than exact integers, all people with ages outside 35 to 39 have zero probabilities. When property tax is topcoded at some value $w < 10,000$, all people with property tax less than w 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, the $Pr(J = r + 1|\mathbf{t}, \mathbf{z}^{Ad}) = 0$ and, for $j \leq r$, the $Pr(J = j|\mathbf{t}, \mathbf{Z}^{Ad}) = 1/n_t$, where n_t is the number of units in $\mathbf{Z}$ with $\mathbf{z}_j^{Ad} = \mathbf{t}^{Ad}$. It may be prudent to assume the intruder knows particular target units are in $\mathbf{Z}$, even when $\mathcal{S}$ is not a census. For example, in a survey of households, neighbors may know that an interviewer visited a sampled household. When all records in $\mathcal{S}$ are included in the release, these 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 $Pr(J = r + 1|\mathbf{t}, \mathbf{Z}^{Ad}) = 0$ results in conservative measures of identification disclosure risks.

The calculations are more complicated when $Pr(J = r+1|\mathbf{t}, \mathbf{Z}^{Ad}) \neq 0$. Let N_t be the number of units in the population that would have $\mathbf{z}_j^{Ad} = \mathbf{t}^{Ad}$ if their data were released in $\mathbf{Z}$. Then, $Pr(J = j|\mathbf{t}, \mathbf{Z}^{Ad}) = 1/N_t$ for units whose $\mathbf{z}_j^{Ad}$ are consistent with $\mathbf{t}$, and $Pr(J = r+1|\mathbf{t}, \mathbf{Z}^{Ad}) = (N_t - n_t)/N_t$. For example, suppose $\mathbf{t}$ contains the age, race, sex, and income of an Asian-American man age 57 whose income is \$125,000. Suppose further that, in the population, there are 11,000 Asian-American males age 57, of whom 1,200 have income more than \$100,000 and three have income exactly equal to \$125,000. If age, race, and sex are released without alteration, and income is perturbed stochastically without any restrictions, then N_t=11,000. If age, race, and sex are released without alteration, and income is topcoded at \$100,000 (or is blurred in some way that restricts the released income to be at least \$100,000), then N_t =1,200. If age, race, sex, and income are

released without alteration, then $N_t = 3$.

The agency, and the intruder, may be able to determine N_t from census totals, particularly when $\mathbf{Z}^{Ad}$ 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 $\mathbf{z}_j^{Ad}$ are consistent with $\mathbf{t}$. Although unbiased, the survey-weighted estimate could poorly estimate N_t, especially when units like $\mathbf{t}$ are rare in $\mathcal{S}$ or when $\mathbf{t}$ contains continuous attributes. Alternatively, N_t can be estimated using model-based approaches, such as those used to determine the number of population uniques (see Section 2.1 for references). To avoid making decisions based on overly optimistic estimates of the reidentification probabilities, agencies can adopt a conservative approach and assume the intruder knows the target is in $\mathbf{Z}$.

If $\mathbf{Z}^{Ad}$ contains no variables, for example when all released variables are subject to stochastic dislosure limitation, the $Pr(J = j|\mathbf{t}) = 1/N$ for $j \leq r$, and $Pr(J = r + 1|\mathbf{t}) = (N - r)/N$, where N is the number of units in the population.

3.2 Evaluating $Pr(\mathbf{z}_j^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$

It is reasonable to assume the intruder knows which variables in $\mathbf{Z}^A$ are subject to stochastic perturbation, as well as the general nature of those perturbations. This meta-information might be made available by agencies so that users of the data know the limitations of their analyses. Stochastic perturbations often are done independently on variables. We assume that the intruder believes this to be the case, so that $Pr(\mathbf{z}_j^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad}) = \Pi_k Pr(z_{jk}^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$.

We now describe methods for evaluating probabilities for $j \leq r$ when the $\mathbf{z}_j^{Ap}$ are generated from data swapping or additive Gaussian noise. When $j = r + 1$, the methods described in Section 3.4 are used.

3.2.1 Data Swapping

Values of key identifiers can be swapped to reduce intruders' confidence in the accuracy of the released values, as proposed initially by Dalenius and Reiss (1982). Data swapping may be as simple as choosing two units at random and swapping their values of a variable, or it could involve constraints on which units can be swapped, e.g. only units within similar geographic areas are allowed to be swapped.

Of course, for any variable k subject to swapping, the intruder does not know for any unit j whether $z_{jk} = y_{jk}$. To estimate $Pr(z_{jk}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$ for a swapped variable k, the intruder can simulate the swapping mechanism of the agency on $\mathbf{Z}$, as shall be demonstrated in Section 4. For example, suppose the agency performs an unconstrained, random swap such that all pairs of units (i, j) have probability π_k of having their values of variable k swapped. Further, suppose the intruder can guess the value of π_k reasonably well. The intruder then applies random data swapping with probability π_k to the released data, and calculates the frequencies of the various combinations of the newly swapped values and the original values in $\mathbf{Z}$. This simulation is repeated many times, and the resulting frequencies are averaged across simulations to estimate the $Pr(z_{jk}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$.

Reasonably accurate intruder simulation may not be unrealistic. Intruders know that agencies typically swap only small fractions of values to preserve data utility, and they may possess general descriptions of the swapping methods from agencies' publications. By assuming the intruder can simulate the swapping mechanism precisely, agencies obtain conservative estimates of the identification probabilities. When the agency believes the intruder cannot know the swapping mechanism, or it is too complicated to determine all pairs' swap probabilities, a sensible model for the intruder, and hence agencies' disclosure evaluations, is to assume a constant swap probability for all pairs of units, possibly within geographic areas reflecting any constraints on the swaps. The agency then can assess the sensitivity of disclosure risks to a variety of intruders' choices of π_k.

3.2.2 Noise Addition

For continuous attributes, agencies can add random noise to discourage exact matches. Noise is usually generated from distributions with expectation equal to zero, so as to maintain unbiasedness in the estimators for population means. A common choice for the noise distribution is a Gaussian distribution with mean zero and some variance σ_k^2 specified to provide sufficient protection. When perturbing more than one variable, it is typical to generate uncorrelated noise. This can provide better protection relative to correlated noise (Fuller, 1993).

It is reasonable to assume the intruder knows approximately the distribution used to generate the noise. This might be available directly from the agency itself to allow users to correct for measurement error (Fuller, 1993). When only the distributional family of the noise is released, intruders may be able to estimate parameters of the distribution by comparing the mean and variance of the perturbed data to the mean and variance of external values, or to published summary statistics. To model intruders' behavior for variables perturbed by Gaussian noise, the agency can assume for these variables that $Pr(z_{jk}|J = j, \mathbf{t}, \mathbf{Z}^{A,d}) = N(z_{jk}|t_k, \sigma_k^2)$. When released values are constrained to lie within certain ranges, for example monetary values must be positive, the agency and intruder can use truncated distributions.

3.3 Evaluating $Pr(\mathbf{z}_j^U|\mathbf{z}_j^A, J = j, \mathbf{t}, \mathbf{Z}^{Ad})$

The intruder does not possess exact values of the target's unavailable variables. To evaluate the probability associated with $\mathbf{z}_j^U$, the intruder can use prior beliefs about the values of the $\mathbf{y}_j^U$ and $\mathbf{z}_j^U$. The intruder then averages over the distribution reflecting those beliefs to obtain

$$Pr(\mathbf{z}_j^U|\mathbf{z}_j^A, J = j, \mathbf{t}, \mathbf{Z}^{Ad}) = \int Pr(\mathbf{z}_j^U|\mathbf{y}_j^U, \mathbf{z}_j^A, J = j, \mathbf{t}, \mathbf{Z}^{Ad}) Pr(\mathbf{y}_j^U|\mathbf{z}_j^A, J = j, \mathbf{t}, \mathbf{Z}^{Ad}) d\mathbf{y}_j^U \quad (3)$$

where $Pr(\mathbf{y}_j^U|\mathbf{z}_j^A, J=j, \mathbf{t}, \mathbf{Z}^{Ad})$ is the intruder's distribution on the target's values of the unavailable variables.

Agencies may decide not to alter some of the variables in U. For these variables, the agency should set $Pr(z_{jk}|\mathbf{z}_j^A, J=j, \mathbf{t}, \mathbf{Z}^{Ad}) = 1$. This is because only one possible value can be released as z_{jk}, and it is prudent for agencies to act as if the intruder knows this.

For those variables in U that are altered, the intruder, and the agency seeking to model intruder behavior, specifies $Pr(z_{jk}|y_{jk}, \mathbf{z}_j^A, J=j, \mathbf{t}, \mathbf{Z}^{Ad})$ to reflect the disclosure limitation techniques applied to y_{jk}, as done in Section 3.2. The intruder specifies $Pr(y_{jk}|\mathbf{z}_j^A, J=j, \mathbf{t}, \mathbf{Z}^{Ad})$ based on available information. For example, the intruder might use a regression of Y_k on $\mathbf{Z}^A$, with parameters estimated from external data or even $\mathbf{Z}$. Partial information, for example knowledge of bounds on the target's y_{jk}, also can be incorporated in the probability distribution. When the intruder has no data or beliefs on which to base a distribution, he or she can use a uniform distribution on some reasonable range.

The agency can evaluate the integrals using numerical approximations. An alternative method is to draw many values from $Pr(\mathbf{y}_j^U|\mathbf{z}_j^A, J=j, \mathbf{t}, \mathbf{Z}^{Ad})$, then calculate the $Pr(\mathbf{z}_j^U|\mathbf{y}_j^U, \mathbf{z}_j^A, J=j, \mathbf{t}, \mathbf{Z}^{Ad})$ for those drawn values, and finally average these probabilities across the draws. This simulation method is particularly useful for incorporating partial information about the target's unavailable variables. For example, if the intruder can bound the target's income, the agency can draw repeatedly values of income from its distribution but evaluate the integral using only those draws that lie within the bounded region. It is essential to simulate very large numbers of values, e.g. hundreds of thousands, to obtain accurate estimates, especially for variables with skewed distributions or outliers.

It is, of course, a challenge for the agency to predict exactly what distributions the intruder will use. A prudent strategy is to assume a sophisticated intruder with access to accurate global relationships in the data. Such relationships could exist in external databases, or they may be–and to maintain data utility perhaps should be–preserved in $\mathbf{Z}$. Hence, the agency should predict the unavailable variables as accurately

as possible, using the data in $\mathcal{S}$ to estimate parameters.

Some intruders may decide to forego modeling of the target's unavailable variables, and assume the $Pr(\mathbf{z}_j^U|\mathbf{z}_j^A, J = j, \mathbf{t}, \mathbf{Z}^{Ad}) = 1$. It is therefore wise for agencies to evaluate disclosure risks under this assumption as well, even when the $\mathbf{y}_j^U$ are altered before release.

3.4 Evaluating $Pr(\mathbf{z}_1^C, \ldots, \mathbf{z}_{j-1}^C, \mathbf{z}_{j+1}^C, \ldots \mathbf{z}_r^C|\mathbf{z}_j, J = j, \mathbf{t}, \mathbf{Z}^{Ad})$

As in Section 3.3, the intruder can specify distributions for the $\mathbf{y}_j^C$ and $\mathbf{z}_j^C$ to calculate the final piece of (2). The probability associated with this piece equals:

$$\int Pr(\mathbf{z}_1^C, \ldots, \mathbf{z}_{j-1}^C, \mathbf{z}_{j+1}^C, \ldots \mathbf{z}_r^C|\mathbf{y}^C, \mathbf{z}_j, J = j, \mathbf{t}, \mathbf{Z}^{Ad}) Pr(\mathbf{y}^C|\mathbf{z}_j, J = j, \mathbf{t}, \mathbf{Z}^{Ad}) d\mathbf{y}^C. \tag{4}$$

A simplification arises when the intruder assumes independence in the $(\mathbf{z}_i^C, \mathbf{y}_i^C)$ across units i, so that the probability in (4) can be expressed as the product:

$$\left(\Pi_{i=1}^{j-1} \int Pr(\mathbf{z}_i^C|\mathbf{y}_i^C, \mathbf{z}_i^{Ad}) Pr(\mathbf{y}_i^C|\mathbf{z}_i^{Ad}) d\mathbf{y}_i^C\right) \left(\Pi_{i=j+1}^{r} \int Pr(\mathbf{z}_i^C|\mathbf{y}_i^C, \mathbf{z}_i^{Ad}) Pr(\mathbf{y}_i^C|\mathbf{z}_i^{Ad}) d\mathbf{y}_i^C\right) \tag{5}$$

After substitution of (5) in the numerator and denominator of (1), and subsequent cancellations, we can replace $Pr(\mathbf{z}_1^C, \ldots, \mathbf{z}_{j-1}^C, \mathbf{z}_{j+1}^C, \ldots \mathbf{z}_r^C|\mathbf{z}_j, J = j, \mathbf{t}, \mathbf{z}^{Ad})$ in (1) with $1/\int Pr(\mathbf{z}_j^C|\mathbf{y}_j^C, \mathbf{z}_j^{Ad}) Pr(\mathbf{y}_j^C|\mathbf{z}_j^{Ad}) d\mathbf{y}_j^C$ for $j \leq r$ and with 1 for $j = r + 1$. These probabilities can be determined using simulations or numerical approximations as suggested in Section 3.2 and 3.3. We note that, unlike in Section 3.2, these probabilities are not conditional on J and $\mathbf{t}$. Values of $\mathbf{y}_j^{Ap}$ must be averaged over as well. For those variables in U with $z_{jk} = y_{jk}$ for all j, the agency should assume for all j that $Pr(z_{jk}|\mathbf{z}_j^{Ad}) = 1$.

Some intruders may forego estimating $\int Pr(\mathbf{z}_j^C|\mathbf{y}_j^C, \mathbf{z}_j^{Ad}) Pr(\mathbf{y}_j^C|\mathbf{z}_j^{Ad}) d\mathbf{y}_j^C$. This may be to ease computations or because they do not have strong beliefs in the distributions for the $\mathbf{y}_j^C$. These intruders can act as if only unit j's values have been altered, so that the probabilities in (4) equal one. This gives reasonable

Table 1: Description of variables used in the simulations

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	15 – 90
Child support payments ($)	C	0, 1 – 23,917
Social security payments ($)	S	0, 1 – 50,000
Household property taxes ($)	P	0, 1 – 99,997
Household income ($)	I	-21,011 – 768,742

estimates of identification probabilities when the $Pr(\mathbf{z}_j^C|\mathbf{z}_j^{Ad})$ are roughly equal. It is prudent for agencies to evaluate disclosure risks under this model of intruder behavior, in addition to assuming informative distributions for the $\mathbf{y}_j^C$.

4 SIMULATIONS

To illustrate the methods of Section 3, we use public release data from the March 2000 U.S. Current Population Survey. The data comprise 51,016 heads of households and the nine variables displayed in Table 1. These variables were selected and provided by statisticians at the U.S. Bureau of the Census. There are no geographic identifiers in the data. Survey weights are included on the file. We assume the population comprises 104,781,947 households, which is the sum of the survey weights. These data also were used by Reiter (2003, 2005a,b) to illustrate synthetic data approaches to protecting confidentiality.

Marginally, there are ample numbers of people in each sex, race, marital status, and education category. Many cross-classifications, however, have few or zero people, especially those involving minorities with $M \notin \{1, 7\}$. There are 12,021 people who receive social security payments and 1,677 who receive child support payments, and 33,076 have positive property taxes. There are 132 households with negative income, 582 with zero income, and the remainder with positive income. The negative incomes are legitimate values: some

households actually report paying out more money than they took in over the year. The distributions of positive values for all monetary variables are right-skewed.

4.1 Description of Disclosure Limitation Methods

We presume the agency plans to release all 51,016 records, with some data values possibly altered. The particular disclosure limitation techniques applied here are illustrative and are not claimed to be optimal, and the utility of the resulting released data is not considered here. The techniques include:

- R: Swap randomly 30% of races.
- M: Swap randomly 30% of marital statuses.
- G: Recode age in five year intervals, e.g. $40 \leq G < 45$.
- P: For positive values, add random noise drawn from $N(0, .10^2\sigma_k^2)$, where $\sigma_k^2 = 2907^2$ is the variance of the positive values of sampled property tax values. When altered values of P are negative, re-draw until we get positive values. Zero values are not altered. Topcoding of P using several cutpoints is also employed.
- X, E, C, S, I: Leave sex, education, child support, social security, and income at their original values.

We assume that the user knows these techniques have been applied to the data, although we do not assume the user knows the values of the random noise nor which units were swapped. It is possible to swap randomly identical values of race and marital status, so that for all practical purposes no swap has taken place.

4.2 Identification Disclosure Risks for Individual Units

We calculate $Pr(J|\mathbf{t}, \mathbf{Z})$ for four units in the data set. "Everyman" has values of all variables near their medians. "Unique" is a 39 year old Native American woman whose spouse is not living at home, the only

Table 2: Actual and altered values for units of study

Unit and Data	X	G	R	M	P	I
Everyman						
Original	M	43	1	1	635	40000
Altered	M	40 - 44	1	1	596	40000
Unique						
Original	F	39	3	3	0	12700
Altered	F	35 - 39	3	1	0	12700
Big I						
Original	M	57	1	1	1100	768742
Altered	M	55 - 59	1	1	1210	768742
Big P						
Original	F	79	1	4	99997	94552
Altered	F	75 - 79	1	1	100033	94552

person in the file with that combination of characteristics. "Big I" has the largest income in the data set ($I = 768,742$), which is about \$150,000 larger than the second largest income. "Big P" has the largest property tax value in the data set ($P = 99,997$). Three other people have this amount. The values in the original and altered data are displayed in Table 2.

4.2.1 User Knows $\{X, R, M, G\}$

In this section, we assume that the intruder knows the sex, race, martial status, and exact age of the target $\mathbf{t}$. But, the intruder does not have any knowledge about the values of other variables in the data set. Hence, A contains $\{X, R, M, G\}$, and U contains $\{E, C, S, P, I\}$.

We first consider releasing the data without applying any disclosure limitation, so that $\mathbf{z}_j = \mathbf{y}_j$. It is prudent to assume the intruder knows this, so that we set $Pr(\mathbf{Z}^U | J = j, \mathbf{t}, \mathbf{Z}^A) = 1$ for all j. For any unit j with $\mathbf{z}_j^A \neq \mathbf{t}$, the probability of identification in (1) equals zero. The probability for units whose $\mathbf{z}_j^A = \mathbf{t}$ depends on whether the intruder knows the target is in $\mathcal{S}$. When the intruder does know this, for any unit j whose $\mathbf{z}_j^A = \mathbf{t}$ the $Pr(J = j|\mathbf{t}, \mathbf{Z}) = 1/n_t$, where n_t is the number of units in the sample with $\mathbf{z}_j^A = \mathbf{t}$. When the intruder is not sure the target is in $\mathcal{S}$, we sum the survey weights of these n_t units to estimate

N_t, the number of potential matches in the population. Then, $Pr(J = j|\mathbf{t}, \mathbf{Z}) = 1/N_t$ for $j \leq n$, and $Pr(J = n + 1|\mathbf{t}, \mathbf{Z}) = (N_t - n_t)/N_t$.

The probabilities of identification are displayed in the column of Table 3 labeled "No SDL." When the intruder knows the target is in the released sample, it is not possible to identify with precision any unit but Unique, which can be identified with probability one. The probabilities decline sharply when the intruder does not know the target is in the sample. The estimated probabilities that $J = r + 1$ are at least 0.9995 for all four targets. Because there is only one sampled unit with the characteristics of Unique, her N_t is estimated with high variance, so that $Pr(J = r + 1|\mathbf{t}, \mathbf{Z})$ may be too small when she is the target. It may be prudent for agencies to assume the intruder knows Unique is in the sample.

We next consider recoding age into five year intervals as exemplified in Table 2. Age recoding can increase the number of people who satisfy $\mathbf{z}_j^{Ad} = \mathbf{t}$, which improves protection. The probabilities of identification are displayed in the column of Table 3 labeled "Age recode." When the target is known to be in $\mathcal{S}$, recoding age reduces probabilities by a factor of about five for all targets except Unique. For Unique, no other unit with the same sex, race, and marital status is in the same age interval. We note that changing the age recoding to stretch from 36 to 40 adds a second person to Unique's group, thereby reducing her probability of reidentification to 0.5. When the targets are not known to be in $\mathcal{S}$, all estimated probabilities are small. Here, the $Pr(J = r+1|\mathbf{t}, \mathbf{Z}^{Ad})$ is based on the sum of the weights for units who match $\mathbf{t}$ on race, sex, marital status, and five year age interval.

We next consider swapping some units' race and marital status, but release exact ages. Hence, $\mathbf{Z}^{Ad}$ contains values for sex and age, and $\mathbf{Z}^{Ap}$ contains values for race and marital status. To calculate the $Pr(\mathbf{z}_j^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$, we first simulate the data swapping procedure on the released data $\mathbf{Z}$ to obtain a new data set $\mathbf{Z}^*$. From $\mathbf{Z}^*$ we determine the percentages of white races swapped with other white races, with black races, with Native American races, and with Asian American races. This is repeated for the other three races to obtain a 4×4 matrix of swap percentages, $\mathbf{R}^*$. Similar enumerations are done with marital

Table 3: Probabilities of identification when $A = \{X, R, M, G\}$. In parentheses are the numbers of other units in sample with probability at least as large as the target's probability.

Unit	No SDL	Age recode	RM swaps	Age recode + RM swaps
Intruder knows target in $\mathbf{z}$				
Everyman	.0022 (454)	.00045 (2229)	.0023 (409)	.00045 (2025)
Unique	1 (0)	1 (0)	.022 (9)	.0047 (48)
Big I	.0029 (344)	.00067 (1497)	.0031 (299)	.00069 (1354)
Big P	.0060 (165)	.0013 (775)	.0046 (168)	.00097 (835)
Intruder unsure target in $\mathbf{z}$				
Everyman	.000001	.0000002	.000001	.0000002
Unique	.00032	.00032	.000001	.0000003
Big I	.000002	.0000003	.000002	.0000003
Big P	.000003	.0000006	.000002	.0000005

status to obtain a 7×7 matrix of swap percentages for marital status, $\mathbf{M}^*$. This process of simulation is repeated one hundred times, and the resulting $\mathbf{R}^*$ and $\mathbf{M}^*$ are averaged across simulations to obtain $\bar{\mathbf{R}}$ and $\bar{\mathbf{M}}$. The $Pr(\mathbf{z}_j^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$ is the product of (i) the entry in $\bar{\mathbf{R}}$ corresponding to the race in $\mathbf{t}$ and the released value of race for unit j, and (ii) the entry in $\bar{\mathbf{M}}$ corresponding to the marital status in $\mathbf{t}$ and the released value of marital status for unit j.

We also must calculate the $Pr(\mathbf{z}_j^{Ap}|\mathbf{Z}^{Ad})$, as discussed in Section 3.4. These are used to approximate the $Pr(\mathbf{z}_1^C, \ldots, \mathbf{z}_{j-1}^C, \mathbf{z}_{j+1}^C, \ldots \mathbf{z}_r^C|\mathbf{z}_j^C, J = j, \mathbf{t}, \mathbf{Z}^{Ad})$ in (1). To do so, conceptually we simulate true values of race and marital status for each unit, and then apply the data swapping probabilities in $\bar{\mathbf{R}}$ and $\bar{\mathbf{M}}$ to those simulated values. To ease computing expenses, we approximate this by using the estimated probabilities for each race and marital status combination as follows. Using $\mathbf{Y}^A$, we fit a multinomial regression of the 28 race-marital status categories on age and sex to obtain, for each j, the estimated probability for each of the categories, $\hat{\pi}_{jh}$, where $h = 1, \ldots, 28$. We then compute the dot product of (i) the vector of $\bar{\mathbf{R}}$ corresponding to the unit's released value of race, and (ii) the $\hat{\pi}_{jh}$ for each race. We also compute the dot product of (i) the vector of $\bar{\mathbf{M}}$ corresponding to the unit's released value of marital status, and (ii) the $\hat{\pi}_{jh}$ for each marital

status. The $Pr(\mathbf{z}_j^{Ap}|\mathbf{Z}^{Ad})$ equals the product of the the two dot products.

As an example of the above computations, consider a record j with released race being black. Suppose the predicted probabilities of each race for unit j are .82 for white, .12 for black, .01 for Native American, and .05 for Asian American. These are obtained by summing the predicted probabilities from the multinomial regression across marital statuses for each race. Suppose now that the swap probabilities in $\bar{\mathbf{R}}$ when the released race is black are .05 when the true race is white, .75 when the true race is black, .10 when the true race is Native American, and .10 when the true race is Asian American. We then compute the weighted sum $(.82)(.05)+(.12)(.75)+(.01)(.10)+(.05)(.10)$. A similar process is used for marital status, and the two weighted sums are finally multiplied together.

The estimated probabilities of identification for data swapping alone are displayed in the column "RM swaps" in Table 3. For Everyman and Big I, the swapping does not improve protection relative to releasing actual values. This is because these two units, which have the most common values of race and marital status, do not have their data altered. On the other hand, Unique has its marital status swapped, which dramatically lowers its probability of identification. We also note that, when Unique is the target, other units in the released data have larger identification probabilities, so that intruders matching on the largest probability will obtain a false identification. When the targets are not known to be in $\mathcal{S}$, $Pr(J=r+1|\mathbf{t},\mathbf{Z}^{Ad})$ is based only on the sum of the weights for units matching on sex and exact age.

Some intruders naively might treat the released race and marital status values as real when matching, essentially acting as if race and marital status are components of $\mathbf{Z}^{Ad}$. When not many values are altered by data swapping, intruders who follow this naive strategy can increase the probabilities of matching correctly for records whose values are not altered by swapping, such as Everyman and Big I. But, they decrease the probabilities for records whose values are altered by swapping, such as Unique and Big P. This is illustrated in the CPS data set: the probabilities for the naive strategy–when swapping race and marital status without recoding age and assuming the intruder knows $\mathbf{t}$ is in the sample–are slightly higher for Everyman and Big P

(.0024 and 0.0033, respectively) and equal zero for Unique and Big P. The increases for Everyman and Big P are small because there are many potential matches with $R = M = 1$, and the probabilities of swapping from ones to ones are near .90, so that exact knowledge of race and marital status for these units does not help much. Exact knowledge can increase probabilities substantially for units for whom there are few matches and the chances of true swaps are significantly less than one. For example, for one 34 year old, divorced Native American woman whose race and marital status were not swapped, the probability equals 0.07 using the approach that accounts for swapping and equals one under the naive approach. For both approaches, she has the largest probability of identification and so would be matched correctly.

Lastly, we consider using both the age recoding and the data swapping. In this case, the multinomial regression is fit using the age categories rather than exact ages. The estimated probabilities of reidentification for recoding and data swapping are displayed in the column "Age recode, RM swaps" in Table 3. The combination of swapping and recoding substantially reduces the probability of identification for Unique, perhaps to the point where agencies can feel confident that intruders who know only $\{X, R, M, G\}$ will not be able to identify precisely these targets in the released data.

4.2.2 User Knows $\{X, R, M, G, P\}$

In this section, we assume the intruder also knows the targets' values of property taxes, which are available from most local governments. There are 2,534 distinct values of property taxes, of which 527 have only one household at that value. Out of the 33,076 households with positive property taxes, there are 21,211 households with unique combinations of age, race, sex, marital status, and property tax. Clearly, when the intruder knows certain targets are in the sample, releasing property taxes without some form of alteration could result in easy identifications. Because of the large number of sample uniques when P is known, we adopt a conservative approach to estimating the identification disclosure risk by assuming the intruder knows the targets are in the sample.

Table 4: Probabilities of identification when $A = \{X, R, M, G, P\}$. Intruder knows target in $\mathbf{Z}$. In parentheses are the numbers of other units in sample with probability at least as large as the target's probability.

Unit	No SDL	Age recode + RM swaps	Age recode + RM swaps + P perturbed
Everyman	.5 (1)	.5 (1)	.0016 (134)
Unique	1 (0)	.01 (26)	.01 (26)
Big I	.14 (6)	.05 (20)	.0028 (29)
Big P	1 (0)	1 (0)	1 (0)

The No SDL column of Table 4 displays the probabilities of identifications assuming no disclosure limitation on any variables. The results differ from the corresponding column of Table 3 because P is now assumed known. This knowledge dramatically increases the probabilities of identifications, making Big P unique and Everyman nearly unique. In fact, even knowing just R and P–which occurs, for example, when the intruder knows a certain household head has been sampled, perhaps by direct observation or hearsay, and determines race from characteristics of the neighborhood and property tax from available public records–results in probabilities of 0.2 for Big P and 0.125 for Everyman in the No SDL setting.

We next consider the identification probabilities after recoding age and swapping RM, assuming P is known. To do so, we use the same swapping strategies as previously. The probability calculations proceed as in Section 4.2.1, with the addition of a linear term for P in the multinomial regression for estimating predicted probabilities of the RM categories. The estimated probabilities of identification are shown in the third column of Table 4. It is clear that knowledge of P can increase the identification probabilities substantially. However, when race and sex are swapped, the identification probabilities for all units except Big P are less than .50.

To reduce further the chance of disclosures, the agency can add random noise to the values of P. We simulate this by adding noise drawn from independent $N(0, 290^2)$ to positive property tax values, restricting values to be always positive. Property taxes equal to zero are not altered. We assume the intruder can estimate the noise variance reasonably well, for example by comparing the variance of the released, positive property taxes to published variance estimates or to external databases. Or, the intruder knows the noise

variance when the agency releases information on the noise distribution. We also assume the agency uses age recoding and data swapping of marital status and race, as described in the previous section. Thus, $\mathbf{Z}^{Ad}$ contains values for sex and age, and the $\mathbf{Z}^{Ap}$ contains values for marital status, race, and property taxes.

To determine the additional component of $Pr(\mathbf{z}_j^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$ due to perturbing property taxes, we use the value of the density function for a truncated normal distribution with mean equal to the target's property tax value and variance equal to 290^2. The truncation is at zero because all perturbed taxes must be positive. When property taxes are rounded and released as integers, a slightly more accurate probability is obtained by using the area within ± 0.5 of the released tax value.

To determine the $Pr(\mathbf{z}_j^{Ap}|\mathbf{Z}^{Ad})$, we approximate each integral, $\int Pr(\mathbf{z}_j^C|\mathbf{y}_j^C, \mathbf{z}_j^{Ad})Pr(\mathbf{y}_j^C|\mathbf{z}_j^{Ad})d\mathbf{y}_j^C$. We again assume that $Pr(\mathbf{z}_j^U|\mathbf{Z}^{Ad}) = 1$ for all units. Conceptually, the integrals can be approximated as follows. First, using $\mathbf{Y}^A$, fit a multinomial regression model for the 28 level variable for race and martial status, conditional on sex and age groups. Generate values of race and marital status for each unit based on this fitted multinomial regression. Second, using $\mathbf{Y}^A$, fit a regression of $\log(P)$ on indicator variables for the five-year age categories, marital status, race, and sex. Generate values of the property tax for each unit, based on the fitted regression and the simulated values of race and marital status. Third, for these drawn values $\mathbf{Y}^{Ap*}$, calculate values of the $Pr(\mathbf{z}_j^{Ap}|\mathbf{y}_j^{Ap*}, \mathbf{z}_j^{Ad})$ based on the disclosure limitation procedures. Finally, draw repeatedly values of $\mathbf{Y}^{Ap*}$, and average the resulting $Pr(\mathbf{z}_j^{Ap}|\mathbf{y}_j^{Ap*}, \mathbf{z}_j^{Ad})$ for each unit. This process requires hundreds of thousands of draws of $\mathbf{Y}^{Ap}$ to get draws sufficiently close to the outliers of property taxes. The integral also can be approximated numerically, which is the approach taken here.

The results are displayed in the final column of Table 4. Adding noise to P substantially reduces the probabilities for Big I and Everyman. It does not protect Unique or Big P any further. Unique has a zero property tax, which is not perturbed. Big P is perfectly identified because no other person in the sample in her age group has a property tax within \$60,000 of Big P's value; hence, even when adding noise, she will be identified by an intruder who knows she is in the sample.

Table 5: Probabilities of identification for Big P for differing topcode cutpoints for P, assuming age recoding, swapping of R and M, and no noise added to P. In parentheses are the numbers of other units in sample with probability at least as large as the target's probability.

Cutpoint w	# records with $P > w$	Probability of identification
35,000	28	.500 (1)
10,000	253	.216 (3)
5,000	1,100	.056 (14)
3,000	3,448	.016 (52)

The amount of noise needed to reduce Big P's identification probability sufficiently is so large that it renders property taxes useless for analysis. A potential solution is to use topcoding, releasing values of P greater than some cutpoint w only as "greater than w." Using topcoding, or other deterministic categorizations, without adding noise to P is a form of recoding. Probabilities of identification associated with this strategy are determined using the methods outlined in Section 3.1 and 4.2.1. When swapping race and marital status, the multinomial models used to estimate predicted probabilities of the swapped RM categories include terms for the categories of P.

Probabilities of identification for Big P after topcoding–assuming age recoding, swapping of race and marital status, and not adding noise to P–are presented in Table 5. A topcode of \$35,000 (and at most \$35,999) reduces the probability of identification for Big P to 0.5. Data utility may be minimally affected, since only 28 people fall in this top tier of property taxes. Probabilities are reduced further for smaller values of w. In fact, for $w \leq 10,000$, Big P is not among the highest probability matches. Several records have larger probabilities of identification because, unlike Big P, their released records have $M = 4$, which is Big P's true marital status. Decreasing w coarsens property tax values for increasing numbers of records, which reduces data utility. We note that the probabilities for Everyman, Unique, and Big I do not change appreciably because their property taxes are well below \$3,000.

Topcoding alone may not sufficiently protect those records with $P < w$. The agency could release P as a fully categorical variable, specifying the categories by gauging risk and utility tradeoffs. Determining

probabilities of identification for this strategy is a straightforward application of recoding and is not pursued here. Another approach is to topcode when $P > w$ and add random noise when $P < w$, ensuring that the noise does not push a record with $P < w$ into the topcoded category. For this strategy, agencies can use distributions truncated at w (and zero) when evaluating $Pr(\mathbf{z}_j^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$ for released property tax values less than w. Property taxes exceeding w are part of $\mathbf{Z}^{Ad}$, so that the $Pr(\mathbf{z}_j^{Ap}|J = j, \mathbf{t}, \mathbf{Z}^{Ad})$ involves only probabilities for the swapping of race and marital status. We examine this strategy in the next section.

4.3 Identification Disclosure Risks for Entire Data Set

To assess identificaton disclosure risks for the entire data set, we suppose that the intruder posesses correct records for all units in the released data and seeks to match each of the targets to a record in the released data. This assumes substantial knowledge on the part of the intruder, but it facilitates calculation of identification disclosure risks. The resulting measures likely overstate the global risks when the intruder does not have all targets' exact data or is unsure that certain targets are in the released data.

To simplify calculations, we assume the intruder matches only one target at a time or, equivalently, matches with replacement. The intruder who matches without replacement may be able to increase identification probabilities relative to with-replacement matching, especially when there is a one-to-one correspondence between the released and original data. On the other hand, this intruder could obtain many false matches when mismatches early in the process cause subsequent attempted matches to be incorrect.

The three measures of global identification risk considered here include (i) the number of units with maximum probabilities exceeding an agency-specified threshhold deemed too risky, (ii) the expected number of true matches, and (iii) the number of units with unique true matches. The first measure reflects perceived matching risk rather than true matching risk, since for any particular target the released unit with the largest probability may not be its true match. For illustrative purposes, a probability threshhold of .20 is used. For the second measure, let m_j be the number of units whose probabilities equal the maximum probability for

some target $\mathbf{t}_j$. Let $I_j = 1$ if the correct match is among those m_j units, and let $I_j = 0$ otherwise. The expected number of true matches then equals $\sum_j (1/m_j) I_j$. When the correct match and some $m_j - 1$ other units all have the maximum probability, $1/m_j$ is added to the sum to reflect the intruder randomly guessing a match from the m_j qualifying units. When the correct match's probability is not the maximum, zero is added to the expected match sum. The third measure equals the number of units with $m_j = I_j = 1$.

Using the same data and methods described in Section 4.1 and 4.2, we apply these global risk measures when $A = \{X, R, M, G\}$ and when $A = \{X, R, M, G, P\}$. The results are summarized in Table 6. When P is not known, age recoding is more effective than the swapping of race and marital status. There remain 43 units that can be uniquely and correctly identified even when employing both age recodes and the swaps. This number could be reduced by judicious data swapping, for example making sure to include in swaps the 96 units with unique correct matches after age recoding.

Knowledge of P greatly increases the disclosure risk associated with any strategy. This is because many units have unique combinations when P is included. Here, data swapping provides relatively little extra protection from identity disclosures, even after ages are recoded. Perturbing P helps dramatically because it introduces uncertainty in matches. Nonetheless, intruders can be expected to make a large number of matches.

Of the 1573 correct, unique matches, 655 records have $P > 3,000$. To reduce risks for these records, we consider a topcode of $w = 3,000$ for P in addition to adding constrained random noise when $P \leq 3000$, as described at the end of Section 4.2. We note that topcoding at $w = 3,000$ without adding noise is ineffective, since 11,205 of the 12,739 unique matches after recoding and swapping are for records with $P < 3000$. The topcode effectively reduces risk for the records with $P > 3,000$: of the 914 unique, correct matches, only 44 have property taxes above \$3,000. There are 816 unique, correct matches belonging to records whose property taxes are altered by random noise. To reduce risks for these records, the agency could lower the w, create other categories of property taxes, or increase the amount of random noise. However, these courses

Table 6: Global identification risks for when $A = \{X, R, M, G\}$ or $A = \{X, R, M, G, P\}$. Entries are number of records. The topcode for property tax is set at \$3,000.

	Prob. $> .2$		Expected matches		Unique matches	
Strategy	no P	P known	no P	P known	no P	P known
No SDL	2102	31175	2192	26839	521	21750
Age recode	496	22878	622	18747	96	13746
RM swap	143	31484	1292	24748	264	23891
Age recode + RM swap	8	21905	337	16801	43	12739
Age recode + RM swap + P perturb	8	1164	337	1854	43	1573
Age recode + RM swap + P perturb/topcode	8	267	337	1307	43	914

of action decrease the utility of the data. At some point, further data alteration does not appreciably lower risks and may not be worth the accompanying reductions in data utility.

5 CONCLUDING REMARKS

Identification probabilities have many positive features as a measure of disclosure risk. Identification probabilities allow agencies to assess disclosure risks under different assumptions on the amount of information possessed by intruders. Agencies can determine which units have probabilities large enough to constitute identification disclosure threats, then take action to reduce those probabilities. The reductions in probabilities can be used to gauge the effectiveness of competing disclosure limitation strategies.

The simulations in this article illustrate these features. For example, the results indicate that age recoding improves protection more effectively than random data swapping for these data, and that intruders who know property tax values can achieve substantially higher identification probabilities, including a large number of unique true matches. Perturbation of tax values reduces these probabilities, but targets with unusual tax values remain at risk. These risks can be reduced by topcoding property taxes, along with adding noise to values below the cutpoint.

Identification probabilities also can feed into measures of attribute risk, which is the risk associated

with learning particular sensitive values. Attribute risk can be assessed using decision theoretic approaches (Duncan and Lambert, 1989; Lambert, 1993; Trottini and Fienberg, 2002), which require specifying models for intruders' guesses about sensitive values and loss functions for those guesses. The intruder might estimate some y_k for a particular target $\mathbf{t}$ with a probability-weighted average of the released attribute values, $\sum_j z_{jk} Pr(J = j|\mathbf{t}, \mathbf{Z})$. Alternatively, the intruder who knows that an attribute has been perturbed, and knows the perturbation method, can replace the z_{jk} with estimates derived from measurement error models. To be conservative, it is advisable for the agency to assume the intruder knows any methods of perturbation (although not the information needed to reveal the original values)and can fit as good a predictive model for target values as possible using the original data. The specification of the loss function depends on the context of the data set and disclosure risks. For example, simply knowing an income exceeds a certain amount may constitute an attribute disclosure, in which case a suitable loss function equals one if the intruder's guess exceeds that amount and equals zero otherwise (Lambert, 1993). For some variables, especially categorical ones, agencies can use a loss function equal to one when the intruder's guess equals the target value exactly and equal to zero otherwise. For other variables, such as numerical or ordinal attributes, agencies might use quadratic or absolute loss functions. It is prudent to evaluate more than one loss function under a variety of estimation methods.

Different types of identity and attribute disclosures can have different costs to the agency. For example, identification disclosures with attribute disclosures may be more harmful to the agency than identification disclosures alone. Or, it may be more costly to the agency, and the sampled individuals, if the intruder learns the identity and attributes of a person who is HIV-positive than if the intruder learns about a person who is HIV-negative. Lambert (1993) suggests that agencies attach costs to all types of disclosures to quantify overall harm to the agency. Such overall harm analyses have not appeared in the literature for genuine data sets, although they may be done informally by agencies.

Identification or attribute disclosure risks attached to certain disclosure limitation strategies need to be

weighed against the utility of the released data. Duncan *et al.* (2001) suggest developing formal measures of data utility and plotting the values of risk versus utility for candidate disclosure limitation strategies. Agencies then select the strategy yielding the greatest utility for an acceptable amount of disclosure risk. Utility measures are difficult to construct, and many have been proposed. Some are based on squared or absolute differences in the sample means and covariance matrices between the released and original data (Duncan *et al.*, 2001; Domingo-Ferrer and Torra, 2001; Yancey *et al.*, 2002). Others are based on entropy measures (Willenborg and de Waal, 2001). A drawback of these measures is their failure to account for the accuracy of inferences based on the released data. For example, these measures do not reflect the properties of confidence intervals made from the released data. Further research on measures of data utility is a high priority item on the disclosure limitation research agenda.

Finally, the methods used here assume sophisticated intruders who attempt to quantify all uncertainties about matching. However, intruders using naive matching strategies–for instance, treating swapped data as if they are unaltered values–may be able to match some records more effectively than sophisticated intruders. It is therefore prudent for agencies to evaluate risks under both sophisticated and naive strategies.

References

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

Blien, U., Wirth, H., and Muller, M. (1992). Disclosure Risk for Microdata Stemming from Official Statistics. *Statistica Neerlandica* **46**, 69–82.

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.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A Technique for Disclosure Control. *Journal of Statistical Planning and Inference* **6**, 73–85.

Domingo-Ferrer, J. and Torra, V. (2001). A Quantitative Comparison of Disclosure Control Methods for Microdata. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 111–133. Amsterdam: North-Holland.

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.

Federal Committee on Statistical Methodology (1978). Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques. Subcommittee on Disclosure-Avoidance Techniques. U.S. Department of Commerce, Washington, D.C.

Federal Committee on Statistical Methodology (1994). Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology. Subcommittee on Disclosure Limitation Methodology. Office of Management and Budget, Executive Office of the President, Washington, D.C.

Felligi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association* **64**, 1183–1210.

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.

Fuller, W. A. (1993). Masking Procedures for Microdata Disclosure Limitation. *Journal of Official Statistics* **9**, 383–406.

Greenberg, B. V. and Zayatz, L. V. (1992). Strategies for Measuring Risk in Public Use Microdata Files. *Statistica Neerlandica* **46**, 33–48.

Lambert, D. (1993). Measures of Disclosure Risk and Harm. *Journal of Official Statistics* **9**, 313–331.

Paass, G. (1988). Disclosure Risk and Disclosure Avoidance for Microdata. *Journal of Business and Economic Statistics* **6**, 487–500.

Pannekoek, J. (1999). Statistical Methods for Some Simple Disclosure Limitation Rules. *Statistica Neerlandica* **53**, 55–67.

Reiter, J. P. (2003). Model Diagnostics for Remote Access Servers. *Statistics and Computing* **13**, 371–380.

Reiter, J. P. (2005a). Releasing Multiply-Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2005b). Using CART to Generate Partially Synthetic, Public Use Microdata. *Journal of Official Statistics* forthcoming.

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.

Skinner, C. J. and Elliot, M. J. (2002). A Measure of Disclosure Risk for Microdata. *Journal of the Royal Statistical Society, Series B* **64**, 855–867.

Spruill, N. L. (1982). Measures of Confidentiality. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 260–265.

Trottini, M. and Fienberg, S. E. (2002). Modelling User Incertainty for Disclosure Risk and Data Utility. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems* **10**, 511–527.

Wallman, K. K. and Harris-Kojetin, B. A. (2004). Implementing the Confidentiality Information Protection and Statistical Efficiency Act of 2002. *Chance* **17**, 3, 21–25.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002). Disclosure Risk Assessment in Perturbative Microdata Protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 135–152. Berlin: Springer-Verlag.