RESEARCH REPORT SERIES
(*Statistics #2004-06)*

# Masking and Re-identification Methods for Public-use Microdata: Overview and Research Problems

William E. Winkler

Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: October 21, 2004

# Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems

William E Winkler[1]

[1] U.S. Census Bureau, Washington, DC 20233-9100, USA,
william.e.winkler@census.gov

**Abstract.** This paper provides an overview of methods of masking microdata so that the data can be placed in public-use files. It divides the methods according to whether they have been demonstrated to provide analytic properties or not. For those methods that have been shown to provide one or two sets of analytic properties in the masked data, we indicate where the data may have limitations for most analyses and how re-identification might or can be performed. We cover several methods for producing synthetic data and possible computational extensions for better automating the creation of the underlying statistical models. We finish by providing background on analysis-specific and general information-loss metrics to stimulate research.

## 1 Introduction

This paper presents of an overview of methods for masking microdata. Statistical agencies mask data to create public-use files for analyses that cannot be performed with published tables and related results. In creating the public-use files, the intent is to produce data that might allow individuals to approximately reproduce one or two analyses that might be performed on the original, confidential microdata. Masking methods are often chosen because they are straightforward to implement rather than because they produce analytically valid data.

There are several issues related to the production of microdata. First, if the public-use file is created, then the agency should demonstrate that one or more analyses are possible with the microdata. It may be that the best the agency can do is an ad hoc justification for a particular analysis. This may be sufficient to meet the needs of users. Alternatively, the agency may be able to refer to specific justifications that have been given for similar methods on similar files in previous papers or research reports. If methods such as global recoding, local suppression, and micro-aggregation have never been rigorously justified, then the agency should consider justifying the validity of a method. This is true even if it is in wide-spread use or in readily available generalized software. Second, the public-use file should be demonstrated to be confidential because it does not allow the re-identification of information associated with individual entities.

The paper provides background on the validity of masked microdata files and the possibility or re-identifying information using public-use files and public, non-

confidential microdata. Over the years, considerable research has yielded better methods and models for public-use data that has analytic properties corresponding to the original, confidential microdata and for evaluating risk to avoid disclosure of confidential information. In the second section, we provide an elementary framework in which we can address the issues. We list and describe some of the methods that are in common use for producing confidential microdata. In the third section, we go into detail about some of the analytic properties of various masking methods. A method is *analytically valid* if it can produce masked data that can be used for a few analyses that roughly correspond to analyses that might have been done with the original, confidential microdata. A masking method is *analytically interesting* if it can produce files that have a moderate number of variables (say twelve) and allows two or more types of analyses on a set of subdomains. In the fourth section, we give an overview of re-identification using methods such as record linkage and link analysis that are well-known in the computer science literature. In the fifth section, we provide an overview of research in information-loss metrics model and re-identification methods. Although there are some objective information loss metrics (Domingo-Ferrer and Mateo-Sanz [16], Domingo-Ferrer [15], Duncan et al. [20], Raghunathan et al. [48]), the metrics do not always relate to specific analyses that users may perform on the public-use files. There is substantial need for developing information-loss metrics that can be used in a variety of analytic situations. Key issues with disclosure avoidance are the improved methods of re-identification associated with linking administrative files and the high quality of information in publicly available files. In some situations, the increased amount of publicly available files means that manual methods (Malin et al. [39]) might be used for re-identification. To further improve disclosure-avoidance methods, we need to research some of the key issues in re-identification. The final section consists of concluding remarks.

## 2 Background

This section gives a framework that is often used in disclosure avoidance research and brief descriptions of a variety of methods that are in use for masking microdata. Other specific issues related to some of the masking procedures are covered in subsequent sections.

The framework is as follows. An agency (producer of public-use microdata) may begin with data X consisting of both discrete and continuous variables. The agency applies a masking procedure (some are listed below) to produce data Y. The masking procedure is intended to reduce or eliminate re-identification and provide a number of the analytic properties that users of the data have indicated that they need. The agency might create data Y and evaluate how well it preserves a few analytic properties and then perform a re-identification experiment. A conservative re-identification experiment might match data Y directly with data X. Because the data Y correspond almost precisely (in respects to be made clearer later), some records in X may be re-identified. To avoid disclosure, the agency might apply additional masking procedures to data X to create data Y'. It might also extrapolate or investigate how well a

potential intruder could re-identify using data Y'' that contain a subset of the variables in X and that contain minor or major distortions in some of the variables. After (possibly) several iterations in which the agency determines that disclosure is avoided and some of the analytic properties are preserved, the agency might release the data.

*Global Recoding* and *Local Suppression* are covered by Willenborg and De Waal [66]. Global recoding refers to various global aggregations of identifiers so that re-identification is more difficult. For instance, the geographic identifiers associated with national US data for 50 States might be aggregated into four regions. Local suppression covers the situation when a group of variables might be used in re-identifying a record. The values of one of more of the variables would be blanked or set to defaults so that the combination of variables cannot be used for re-identification. In each situation, the provider of the data might set a default $k$ (say 3 or 4) on the minimum number of records that must agree after global recoding and local suppression.

*Swapping* (Dalenius and Reiss [9], Reiss [49], Schlörer [57]) refers to a method of swapping information from one record to another. In some situations, a subset of the variables is swapped. In variants, information from all records, a purposively chosen subset of records, or a randomly selected subset of records may be swapped. The purposively chosen sample of records may be chosen because they are believed to have a greater risk of re-identification. The advantages of swapping are that it is easily implemented and it is one of the best methods of preserving confidentiality. Its main disadvantage is that, even with a very low swapping rate, it can destroy analytic properties, particularly on subdomains.

*Rank Swapping* (Moore [41]) is another easily implemented masking procedure. With basic single-variable rank swapping, the values of an individual variable are sorted and swapped in a range of $k$-percent of the total range. A randomization determines the specific values of variables that are swapped. Swapping is typically without replacement. The procedure is repeated for each variable until all variables have been rank swapped. If $k$ is relatively small, then analytic distortions on the entire file may be small (Domingo-Ferrer and Mateo-Sanz [16], Moore [41]) for simple regression analyses. If $k$ is relatively large, there is an assumption that the re-identification risk may be reduced.

*Micro-aggregation* (Defays and Answar [12], Domingo-Ferrer and Mateo-Sanz [17]) is a method of aggregating values of variables that is intended to reduce re-identification risk. For single-ranking micro-aggregation in which each variable is aggregated independently of other variables, it is easily implemented. The values of a variable are sorted and divided into groups of size $k$. In practice $k$ is taken to be 3 or 4 to reduce analytic distortions. In each group, the values of the variable are replaced by an aggregate such as the mean or the median. The micro-aggregation is repeated for each of the variables that are considered to be usable for re-identification. Domingo-Ferrer and Mateo-Sanz [17] provided methods for aggregating several variables simultaneously. The methods can be based on multi-variable metrics for clustering variables into the most similar groups. They are not as easily implemented because they can involve sophisticated optimization algorithms. For computational efficiency, the methods are applied to 2-4 variables simultaneously whereas many

public-use files contain 12 or more variables. The advantage of the multi-variable aggregation method is that it provides better protection against re-identification. Its disadvantage is that analytic properties can be severely compromised, particularly if two or three uncorrelated variables are used in the aggregation. The variables that are not micro-aggregated may themselves allow re-identification.

*Additive Noise* was introduced by Kim [32}, [33] and investigated by Fuller [28], Kim and Winkler [34], and Yancey et al. [71]. Let $\mathbf{X}$ be an $n \times k$ data vector consisting of $n$ records with $k$ variables (fields). If we generate random noise $\mathbf{X}_1$ with mean 0 and $\text{cov}(\mathbf{X}_1) = \text{cov}(\mathbf{X})$, then we can replace $\mathbf{X}$ and use $\mathbf{Y} = \mathbf{X} + \varepsilon$ where $\text{cov}(\varepsilon) = c\ \text{cov}(\mathbf{X}_1)$ for $0<c<0.5$. Further, we can do a linear transform of $\mathbf{Y}$ to $\mathbf{Z}$ so that the mean of $\mathbf{Z}$ equals the mean of $\mathbf{X}$, and the correlation of $\mathbf{Z}$ equals correlation of $\mathbf{X}$. Kim [26] also showed that it is theoretically possible to recover means and covariances of $\mathbf{X}$ on arbitrary subdomains. Kim [32], [33] and Fuller [28] both showed that the additive-noise procedures provide good analytic properties such as regression analyses with masked data $\mathbf{Z}$ that closely reproduce regression analyses of the unmasked data $\mathbf{X}$. Because additive noise can yield files with moderate re-identification rates (Fuller [28]], Kim and Winkler [34]), Roque [54] introduced methods that applied mixtures of additive noise that reduced re-identification rates by a factor of 10. Yancey et al. [71] provided much simpler computational procedures that eliminated the nonlinear optimization procedures of Rogue [54]. One criticism has been the need of specialized software for producing and analyzing the data that is (partially) alleviated by high quality software that was developed for Yancey et al. [71].

*Synthetic Microdata from Probabilistic Models* (Fienberg et al. [24], [26], [27], Raghunathan et al. [48], Reiter [52], Little and Liu [37], [38], and Polettini [46]) provide methods for building an accurate statistical model $\mathbf{M}$ of data $\mathbf{X}$. The statistical model $\mathbf{M}$ is generally based on estimates of joint and conditional distributions of the underlying probability densities. The estimation of densities is simplified because the models typically only target one set of analyses. Artificial or synthetic data $\mathbf{Y}$ are created by randomly drawing records from the model $\mathbf{M}$. The intent is that some of the analytic properties of $\mathbf{M}$ are preserved in the artificial data $\mathbf{Y}$. It is generally assumed that re-identification is impossible even when a number of analytic restraints are placed on the data. A clear disadvantage of synthetic data is the amount of modeling expertise that is needed for developing a reasonable model $\mathbf{M}$ of the data with suitable analytic properties. Another disadvantage is that a moderate-to-large amount of high quality data may be needed in the modeling. If there are errors in the original data, then there is a possibility that the errors will be approximately reproduced in the synthetic data. An advantage is that the one or two potential analytic uses of the data can be clearly described.

## 3 Analytic Properties of Microdata

There are several issues with regards to the production of microdata with analytic properties. The first issue is whether the analytic properties of the masked data are justified. Does the masked microdata allow a user to approximately reproduce one or

two analyses that could be performed on the original microdata other than simple tabulations (that are often in published tables)? Is there sufficient detail so that the user can be aware of any potential discrepancies that may occur in an analysis that the producer has stated can be performed with the masked microdata? The second issue is whether the public-use microdata yield analytic results that differ significantly from results that would be obtained using the original microdata. For instance, in an econometric analysis, an economist might not find relationships between certain masked variables where relationships exist between the variables in the original microdata. Additionally, there may be too many or too few individuals in high income categories in comparison with the original microdata.

With the exception of additive noise and synthetic microdata, there has seldom been much work that describes the analytic properties of the masked microdata. Generally, additive noise produces masked microdata that can quite accurately reproduce regression analyses (even on arbitrary subdomains that are reasonably large). If the constant $c$ used with additive noise is close to zero, then additive noise will often not distort the original data much and allow other statistical analyses. A deficiency of additive noise as suggested by Fuller [28] and shown by Kim and Winkler [34] is that small proportions of records (0.5% - 3%) might be re-identified. Kim and Winkler [34] applied additional masking procedures to avoid disclosure and somewhat decrease valid analytic properties of the public-use microdata. We observe that it is possible that re-identification rates go up as more analytic restrictions are placed on the masked data. For instance, many users want to do analyses on subdomains. If the microdata have detailed geographic identifiers (Elliot et al. [21], [22]) or subdomains associated with sparse age-race-sex categories are created, then re-identification rates can increase.

Creation of synthetic data from valid models is appealing because it potentially can preserve some analytic properties. It is assumed that synthetic data cannot be re-identified. An intuitive difficulty of synthetic data, as shown by Reiter [52], [50], is that any analytic properties that are not included in the model will not be in the synthetic data. Reiter [52] has shown that some of the analytic properties that are in the model may not be in the synthetic microdata due to problems with the original microdata. For instance, the nature of the random number generation process and sample sizes, the type of modeling procedures used, outliers, and errors in the original microdata can all affect the quality of the produced microdata. Additionally, it is not possible to have greater accuracy for analytic purposes in the masked microdata than in the original, confidential microdata.

Reiter [51], [52] has used standard modeling methods that are often used in the multiple imputation literature for creating models for the data. Polettini [46] has used maximum entropy methods for creating models of the data. Muralidhar et al. [42], [43] and Sarathy et al. [55] have used copolas to model distributions. Due to the difficulty (both theoretically and computationally) of creating models for data, Thibaudeau and Winkler [62] have suggested using Bayesian Networks because high quality software is available for (semi-)automatically creating models of the data. The Bayesian Network methods can be used only with discrete data or with models in which some of the continuous data can be reasonably approximated by a discrete data representation. Although the models are likely to have lower quality than those of

Reiter [52] and Polettini [46], they are much easier for the statistical agencies to implement. Dandekar et al. [10], [11] use Latin Hypercube methods for creating synthetic data. They have demonstrated that the methods can rapidly produce large synthetic data sets with a considerable number of variables. Further, when the Latin Hypercube methods do not produce synthetic data with the desired accuracy for some analyses, they have iterative refinement methods for improving the analytic properties.

With other masking methods for producing microdata, there is often no justification that the masking methods produce masked microdata with one or two of the analytic properties of the original microdata. The superficial exceptions are when the producers of the masked microdata demonstrate that the masked microdata reproduces several of the tabulations of the original microdata. The producers may also include reproduction of tabulations on a few subdomains. At a minimum, we would hope that the masked microdata allow reproducing tabulations on the entire file and on certain subdomains. Fuller [28] and Lambert [35] have indicated that masked microdata should preserve the first two moments of the original, confidential microdata and at least one other statistical property. Van Den Hout and Van Der Heijden [64] have shown how a few analytic properties can be preserved with the PRAM method (see e.g., Willenborg and De Waal [66]) for producing masked, discrete data. PRAM uses a Markovian strategy for swapping values of a variable across various records. The intent is that certain marginal distributions are approximately preserved. A potential intruder can never be certain what values in a given record have been changed.

## 4 Re-identification of Microdata

The highest standard for estimating the proportion of records that can be re-identified (Lambert [35], Domingo-Ferrer and Torra [19]) is where record linkage is used to match information from public data with the masked microdata. Alternative effective matching methods are the nearest neighbor methods (Domingo-Ferrer and Mateo-Sanz [16]) and the clustering algorithms of Bacher et al. [5]. Record linkage can also be used in the conservative framework described in section 2 in which a masked file of **Y** data is matched directly with the original, confidential file of **X** data that was used in creating it. In some situations, it may be possible to accurately match 0.5-2% of the records in the initial iteration of the potential masked microdata. In this situation, the file is non-confidential and we would apply additional masking procedures to provide disclosure avoidance. The ability to link public information to public-use files has further been compromised by the increased sophistication of record linkage (Yancey et al. [71] and link analysis methods (McCallum and Wellner [40], Bilenko et al. [7]) and the increased availability of high-quality public data. Indeed, due to the quality of public information, Malin et al. [39] are often able to match public information using purely manual methods.

Re-identification of microdata refers to the ability to use publicly available information to attach names, addresses, and other (semi-)unique identifiers to individual records in a public-use file. An identifier is *semi-unique* if it cannot exactly identify a linkage between two records using the identifier by itself but can allow a re-

identification in combination with other variables. For instance, in a file where there are many individuals with the name "John Smith", "John Smith" is a semi-unique identifier. US Postal ZIP code and income may be other semi-unique identifiers for an individual. When the semi-unique identifiers such as name "John Smith", ZIP code, and income are used in combination, they may uniquely determine a linkage between two records. Among the records having name "John Smith" and ZIP code "98100," we may define a metric that allows the income to be matched if the incomes are within 5% of each other. The deviation in the metrics can often be modified according to the types of files being matched, the amount of redundant information that is available for matching, and the stated analytic properties of the public-use microdata. For instance, if we had an additional variable such as profession, it is likely we could re-identify even when the metrics allow significantly larger deviations in income and other characteristics between matching records.

In record linkage (Winkler [67], [68]) and in re-identification experiments (Yancey et al. [71]), the software is designed to deal with both minor and major errors in a subset of the variables used in matching. In most realistic real-world settings, some of the continuous variables are available in external files and can be used in matching. Further, it is straightforward (Scheuren and Winkler [56]) to define new metrics based on relationships between two correlated or related variables to improve matching accuracy significantly. For instance, we might consider the situation of linking two administrative lists of companies in which name and address information are sufficiently poor so most companies in one file might be associated with upwards of fifteen companies in another file. If we also have income on one file, receipts on another file, and a crude function 'f(income) = receipts' that relates income to receipts, then we may be able to significantly increase high quality match rates with the extra, non-name-and-address information. In the situation with a masked file Y, we may have a file Y' that only has a subset of the variables that in the Y file and the subset of Y'-variables is not sufficient for accurate matching. If we can find an extra file Z in which the Z values can be used to predict the Y variables that are missing from the Y' file, then we may have sufficient extra information for significantly improving the matching (and possible re-identification) [69].

Because many individuals are unable to perform sophisticated or elementary record linkage, a number of other re-identification risk measures have been defined. The first measure is the number or proportion of population uniques in a file. The measure appears to be based on elementary ideas in survey sampling for which sort/merge utilities can be used to determine matches on an exact character-by-character basis. A *unique* is a record with identifying information that distinguishes it from other records. If records are tabulated according to their identifying information, then *uniques* are those records that have frequency one. There are elementary models that use the frequency distribution of the identifying characteristics in the sample to obtain estimates of the number of the population uniques in the sample. The intuition is that each population unique might be identified using elementary sort/merge methods. A key observation (Elliot et al. [21], [22]) is that a sizable proportion of population uniques in a file with twelve or more variables may be uniquely determined by three, four, or five variables in the records associated with the uniques. These *special*

*uniques* may be more easily re-identified because the intruder may be more likely to have the smaller number of variables in an external file.

There is a class of risk-estimation measures that are based on statistical models in which the number of uniques in the public-use sample is used to provide an estimate of the number of uniques in the original population file. The methods apply quite sophisticated models that relate the distribution of the sample uniques to the distribution of the population uniques. Early papers with these statistical risk measures are by Bethlehem et al. [6], Fienberg and Makov [25], Benedetti and Franconi [5] and Skinner and Holmes [59]. Later papers with enhanced methods are due to Skinner and Elliot [58], Rinott [53], and Polettini and Stander [47]. The apparent intent is to provide a straightforward, rapid method of estimating the proportion of sample uniques that are also population uniques. We will refer to these methods as *sample-unique-population-unique* (*SUPU*) methods.

There are four obvious criticisms of these *SUPU* methods. The first criticism is that it is straightforward for the statistical agency to determine the risk of disclosure by directly comparing the sample file with the population file from which it is produced. Indeed, determining the number of population uniques that are also in the sample file can be part of the sampling procedure. The second criticism is that a public-use file will typically contain ten or more variables, both discrete and continuous. Even if the continuous variables are broken into a large number of discrete ranges, it is likely that all or almost all of the records in the sample and many or most of the original population records will be unique. The methods that have typically been applied only use a few discrete variables under the assumption that the intruder will only have access to a few discrete variables. This assumption appears to be naïve given the amount of information that is available from Internet and other sources (Sweeney [60], [61]). Because most, if not all, public-use data files contain continuous variables, it is unrealistic not to include them in the statistical models of disclosure risk. The third criticism is that the statistical models only provide an estimate of the proportion of sample uniques that are also population uniques. The methods do not determine which of the sample uniques can be re-identified as would happen in a record linkage experiment (Kim and Winkler [34], Sweeney [60], Winkler [68], Domingo-Ferrer and Torra [19]). The fourth criticism is that the models are often severely biased with the bias varying according to the data source. For instance, if we were to produce a public-use file according to the following two procedures, we would obtain severely biased answers from all of the *SUPU* models. In the first situation, we could sample only from population uniques in producing public-use file $D_1$. In the second situation, we would sample only from population records that are not unique (occurring two or more times according to identifying information) in such a manner that every record in the public-use file $D_2$ is a sample unique. In each of these situations, every *SUPU* model would give very biased estimates of the re-identification risk. In general situations where different sampling procedures might be used, we would still likely have subsets of the public-use file where the biases could approach the two extreme situations previously described.

A second metric associated with re-identification (disclosure) risk for a one-variable situation is the multiplicative inverse of the $Var\ (X_1)$ (Duncan et al. [20]) where the variance is that of an intruder with weak knowledge of the target variable

$X_1$. For instance, if data are micro-aggregated into a group on $n$ items and each item is given the average value $\bar{x}$, the re-identification risk is $n/(n-1)\sigma$ where $\sigma$ is the variance of the original $x$-values. Trottini and Fienberg [63] have extended this metric to a few two-dimensional situations and have noted that the metric is different from many of the other metrics that are commonly in use.

In addition to using methods such as record linkage or nearest-neighbor matching, Lambert [35] and Palley and Simonoff [45] have shown how knowledge of the analytic properties (such as regression parameters) of a masked data file can allow re-identification. De Waal and Willenborg [13], [14] show how knowledge of sampling weights can allow re-identification. If we have detailed information about the survey frame, the sampling design, and the valid uses of the sampled data, then the sampling weights give us useful information about subdomains in which a record may occur. If the sampling weights are combined with the original sampled data or with masked data, we may have sufficient information to allow some re-identifications.

There are several research questions. In many instances, can methods such as record linkage, nearest neighbor, or clustering be used for re-identification? What data will be available on public sources such as the Internet and can be used for re-identification? Can the ideas of constructing functional relationships be substantially extended to allow re-identification in many situations? How can advanced methods such as link analysis be applied for re-identification?

## 5 Discussion of Information Loss Metrics

The best situation for the producer of public-use microdata is when there is one set of users of the microdata with explicitly stated analytic needs. The key point is that, if there is a set of clearly defined users, the data provider can use ad hoc procedures that are specific to a small set of required types of analyses. Generally, individuals have desired more objective information loss metrics that allow comparison of results across several types of analyses or types of files. Before describing some of the attempts at objective information-loss metrics, we describe the situation with clearly user-defined needs.

Kim and Winkler [34] needed to produce masked data that preserved analyses in a clearly defined set of subdomains determined by age, race, sex, and one other variable. In their situations, they, as producers of the data, were able to iteratively negotiate the potential analytic uses. Because some of the initial domains were too small, they were able to get users to agree to some collapsing of ages into age ranges and some of the subdomains determined race-age-sex categories. Based on the consultation with users, they initially created subdomains that had a sufficient number of records. They then applied additive noise with various levels of noise to determine the analytic degradations due to increased noise and determined a noise-level at the user-specified level of accuracy. Finally, they applied a swapping strategy for the most easily re-identified records. The swapping strategy was designed to preserve regression properties in the user-specified subdomains. After completion of the masking procedures, they provided information about analytic degradations in some subdo-

mains that were not considered as important as the main set of subdomains and described specifics of the testing to determine that disclosure was avoided.

Willenborg and De Waal [66] have suggested using entropy with discrete data. The possible intuition is that the masked data may have decreased entropy and be less useful. We observe, however, if $\mathbf{X} = (x_1, \ldots, x_n)$ is discrete data, then a collapsing strategy may produce masked data $\mathbf{Y} = (y_1, \ldots, y_m)$ where the sum of the counts in the y-cells exceeds a lower bound (say 3) and where the sum of the counts agrees the sum of the counts of the x-cells. Each of the y-variables is obtained by aggregating x-variables. Although entropy clearly decreases, it is not clear how a loglinear analysis might be affected. In the best situation, the collapsing may account for the sufficient statistics in original x-variables and the y-variables allow reproduction of an analysis. Domingo-Ferrer and Torra [18] have observed additional difficulties with the use of entropy. In other situations, a loglinear analysis that is possible on the x-variables is impossible on the y-variables. In other words, entropy is unlikely to provide useful information about possible degradations in analyses. Iyengar [30] provides an example of a collapsing strategy that allows a classification problem to be (approximately and accurately) reproduced. Sweeney [60], [61] provides additional details on collapsing methods and strategies for producing k-anonymity. A file is *k-anonymous* if the identifiers agree exactly with the identifiers in at least *k-1* other records.

Willenborg and De Waal [66] have suggested using as an information loss metric a statistic that they refer to as the *variance inflation statistic* $Var(\hat{\theta} \mid D_0) / Var(\hat{\theta} \mid \overline{D})$ where $\hat{\theta}$ is a univariate statistic, $D_0$ is the original data, and $\overline{D}$ is the masked data. We observe that if the data are masked via micro-aggregation, then this statistic increases. If the data are masked by additive noise, then this statistic decreases.

Duncan et al. [20] have suggested using as information loss metric the statistic $1/D_S$ where $D_S$ is the utility of a univariate statistic. For instance, if a file containing one variable $\mathbf{X} = (x_1, \ldots, x_n)$ is masked by replacing the original values $x_i$ in each of the records with the average $\overline{x}$, then the variance $\sigma^2$ of the data is the utility D. If the variance increases, then the utility decreases. The ideas of Duncan et al. [20] are intended to cover both information loss and disclosure risk for a file containing one variable. Trotini and Fienberg [63] have shown how to extend their ideas to a few two-dimensional situations with multivariate normal data.

Agrawal and Aggarwal [2] provide a measure of information loss that may be suitable for further research. Let $\hat{f}_X(x)$ be the density estimated from the masked data *Y*. Let $f_X(x)$ be the density from the original data *X*. Agrawal and Aggarwal [2] estimate $\hat{f}_X(x)$ in the one-dimensional privacy-preserving situation of Agrawal and Srikant [3] in which a simple type additive noise used for masking. In more general situations, the densities could be associated with multivariate data that has been masked via other methods. Then the information loss in the masked data is given by

$$I(f_X, \hat{f}_X) = (1/2)E\left[\int_{\Omega_X} | f_X(x) - \hat{f}_X(x) | dx\right].$$

This metric equals half the expected value of the $L_1$-norm between the original distribution $f_X(x)$ and the estimate $\hat{f}_X(x)$. The information loss $I(f_X, \hat{f}_X)$ takes values between 0 and 1. If $I(f_X, \hat{f}_X)$=0, then the estimate $\hat{f}_X(x)$ perfectly reconstructs $f_X(x)$. An issue with reconstructing the density $f_X(x)$ is the amount of data needed and the accuracy of the estimate $\hat{f}_X(x)$. It may be more difficult to obtain the estimate $\hat{f}_X(x)$ than to do a direct comparison, say, of two regression analyses using the masked and original data. This concern applies to all of the information-loss situations involving synthetic data (given below).

Gomatam and Karr [29] provide a review and empirical comparison of six information-loss metrics for discrete data from the statistical and other literature. In the following *f* and *g* are two discrete distributions where the first might be associated with the unmasked data *X* and the second with the masked data *Y*. They use *Hellinger distance* whose definition is

$$H(f, g) = (1/\sqrt{2})(\sum_x (\sqrt{f(x)} - \sqrt{g(x)})^2)^{1/2}.$$

They use the discrete analog of the distance metric of Agarawal and Aggarwal [2] that they refer to as *total variation*. They use change in entropy that has been used by Willenborg and De Waal [66] and Domingo-Ferrer and Mateo-Sanz [18]. They use Cramer's *V measure of association* on the $\chi^2$ statistic for an $m \times n$ contingency table that is defined as

$$V = (\chi^2 / (N \min(m-1, n-1))^{1/2},$$

where $\chi^2$ is the usual test of independence. They also use *Pearson's contingency coefficient C* that is defined as

$$C = (\chi^2 / (\chi^2 + N))^{1/2}.$$

In the empirical work, Gomatam and Karr [29] apply the information loss metrics to data in which swapping has been performed (Dalenius and Reiss [9]). Willenborg and De Waal [66] define a data swap of 2*k* elements in terms of *k* elementary swaps. In an elementary swap we first make a random selection of two records *i* and *j* from and then interchange of the values of the variables being swapped for these two records. The swap proportion or rate is defined as 2*k*/*N* where *N* is the number of records. Gomatam and Karr observe that each of the metrics generally increases with the increase in the swap rate and that they are quite correlated.

Within the multiple-imputation framework of Raghunathan et al. [48], Little and Liu [37], [38] have also considered an information loss metric for univariate statistics. Before giving the statistic, we need to define some terms. We are interested in the scalar parameter $\phi$. For any completed data set $d = (d = 1, \cdots, D)$ among D copies of the data obtained from randomly drawing from the model, let $\hat{\phi}_d$ denote an estimate of $\phi$ and $V_d$ an estimate of the variance of $\hat{\phi}_d$. The MI estimate of $\phi$ is given by

$$T = W + B / D$$

where $W = \sum_{d=1}^{D} V_d / D$ and $B = \sum_{d=1}^{D} (\hat{\phi}_d - \bar{\phi}) / (D-1)$. Then the information loss associated with the scalar $\phi$ is given by $\gamma = (B / D)T$. Since *B* and *T* are likely to be bounded, the information loss $\gamma$ goes to zero as the number of replicates *D* increases. The metric $\gamma$ is appealing. If the analysis is based on a highly accurate model **M**, then the model **M** provides estimates of scalars for which information loss can go to zero. As MI is a general framework, it would be useful to extend the information loss to multivariate situations and provide a number of examples.

Because of the difficulties in creating detailed models **M**, Little and Liu [38] and Reiter [51] only develop partial models that are applied to a subset of the variables in a data file. Alternatively, Kennickell [31] creates a MI model for data and iteratively blanks and fills values of variables to create a file of synthetic data. The iterative cycling between blanking data and filling in data stops when the data are believed to have converged or sufficient cycles have taken place. The ideas of Kennickell have been adapted and extended by Abowd and Woodcock [1].

The research questions for information loss are particularly difficult? Is it possible to create models that represent data in a form that allows or account for several analyses? With different models, is it possible to have metrics for information loss that relate to several analyses? How does a statistical agency create public-use files that preserve analytic properties and provide statistical justifications of the limitations of the public-use data?

## 6 Concluding Remarks

This paper provides an overview of a number of methods that are in common use for the production of public-use microdata. It covers analytic uses of the microdata and some of the research problems in information-loss metrics. It also covers methods of evaluating re-identification risk.

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U. S.

Census Bureau. The author thanks Nancy Gordon, Cynthia Clark, and two reviewers for comments leading to improved wording and explanation and several additional references.

## References

1. Abowd, J. M. and Woodcock, S. D. Disclosure Limitation in Longitudinal Linked Data, in (P. Doyle et al., eds.) Confidentiality, Disclosure, and Data Access, North Holland: Amsterdam, (2002)
2. Agrawal, D. and Aggarwal, C. C. On the Design on Privacy Preserving Data Mining Algorithms, Proceedings of the ACM SIGPODS, (2001) 247-255
3. Agrawal, R. and Srikant, R. Privacy Preserving Data Mining, Proceedings of the ACM SIGMOD, (2000) 439-450
4. Bacher, J., Brand, R. and Bender, S. Re-identifying Register Data by Survey Data using Cluster Analysis: An Empirical Study, International Journal of Uncertainty, Fuzziness, Knowledge-Based Systems, 10 (5) (2002) 589-608
5. Benedetti, P. and Franconi, L. Statistical and Technological Solutions to the Controlled Data Dissemation. In: Pre-proceedings of New Techniques and Technologies for Statistics, Volume 1, Sorrento (1998) 225-232
6. Bethlehem, J. A., Keller, W. J. and Pannekoek, J. Disclosure Control of Microdata, Journal of the American Statistical Association, 85 (1990) 38-45
7. Bilenko, M., Mooney, R., Cohen, W., Ravikumar P., and Fienberg, S. Adaptive Name Matching in Information Integration, IEEE Intelligent Systems, 18 (5) (2003) 16-23
8. Brand, R. Microdata Protection Through Noise Addition, in (J. Domingo-Ferrer, ed.) Inference Control in Statistical Databases, Springer: New York (2002)
9. Dalenius, T. and Reiss, S.P. Data-swapping: A Technique for Disclosure Control, Journal of Statistical Planning and Inference, 6 (1982) 73-85
10. Dandekar, R. A., Domingo-Ferrer, J., and Sebe, F. LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection, in (J. Domingo-Ferrer, ed.) Inference Control in Statistical Databases, Springer: New York (2002).
11. Dandekar, R., Cohen, M., and Kirkendal, N. Sensitive Microdata Protection Using Latin Hypercube Sampling Technique, in (J. Domingo-Ferrer, ed.) Inference Control in Statistical Databases, Springer: New York (2002)
12. Defays, D. and Anwar, M. N. Masking Microdata Using Micro-aggregation, Journal of Official Statistics, 14, (1998) 449-461
13. De Waal, A. G. and Willenborg, L.C.R.J. Global Recodings and Local Suppressions in Microdata Sets, Proceedings of Statistics Canada Symposium *95*, (1995) 121-132
14. De Waal, A. G. and Willenborg, L.C.R.J. A View of Statistical Disclosure Control for Microdata, Survey Methodology, 22, (1996) 95-103
15. Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases, Springer: New York, (2002)
16. Domingo-Ferrer, J. and Mateo-Sanz, J. M. An Empirical Comparison of SDC Methods for Continuous Microdata in Terms of Information Loss and Re-Identification Risk, presented at the UNECE Workshop On Statistical Data Editing, Skopje, Macedonia, May 2001
17. Domingo-Ferrer, J. and Mateo-Sanz, J. M. Practical Data-Oriented Microaggregation for Statistical Disclosure Control, IEEE Transactions on Knowledge and Data Engineering, 14 (1), (2002) 189-201
18. Domingo-Ferrer, J. and Torra, V. A Quantitative Comparison of Disclosure Control Methods for Microdata, in (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds. Confidentiality,

Disclosure Control and Data Access: Theory and Practical Applications, North Holland, (2001) 111-134
19. Domingo-Ferrer, J. and Torra, V. Statistical Data Protection in Statistical Microdata Protection via Advanced Record Linkage, Statistics and Computing, 13 (4), (2003) 343-354
20. Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map, Los Alamos National Laboratory Technical Report LA-UR-01-6428 (2001)
21. Elliott, M. A., Manning, A. M., and Ford, R. W. A Computational Algorithm for Handling the Special Uniques Problem, International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10 (5), (2002) 493-510
22. Elliott, M. A., Skinner, C. J., and Dale, A. Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk, in Statistical Data Protection '98, Eurostat, Brussels, Belgium, (1998), 261-265, also Research in Official Statistics, 1 (2) 53-68
23. Fellegi, I. P., and Sunter, A. B. A Theory for Record Linkage, Journal of the American Statistical Association, 64, (1969) 1183-1210
24. Fienberg, S. E. Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, commissioned by Committee on National Statistics of the National Academy of Sciences (1997)
25. Fienberg, S. E., and Makov, U. Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data, Journal of Official Statistics, 14 (1998) 385-397
26. Fienberg, S. E., Makov, E. U. and Sanil, A. P., A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data, Journal of Official Statistics, 14, (1997) 75-89
27. Fienberg, S. E., Makov, E. U. and Steel, R. J. Disclosure Limitation using Perturbation and Related Methods for Categorical Data, Journal of Official Statistics, 14, (1998), 485-502
28. Fuller, W. A. Masking Procedures for Microdata Disclosure Limitation, Journal of Official Statistics, 9**,** (1993) 383-406
29. Gomatam, S. V. and Karr, A. On Data Swapping of Categorical Data, American Statistical Association, Proceedings of the Section on Survey Research Methods, (2003), CD-ROM
30. Iyengar, V. Transforming Data to Satisfy Privacy Constraints, Association of Computing Machinery, Special Interest Group on Knowledge Discovery and Datamining '02 (2002)
31. Kennickell, A. B. Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances, in Record Linkage Techniques 1997, Washington, DC: National Academy Press, 248-267 (available at http://www.fcsm.gov ) (1999)
32. Kim, J. J. A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, American Statistical Association, Proceedings of the Section on Survey Research Methods, (1986) 303-308
33. Kim, J. J. Subdomain Estimation for the Masked Data, American Statistical Association, Proceedings of the Section on Survey Research Methods, (1990) 456-461
34. Kim, J. J. and Winkler, W. E. Masking Microdata Files, American Statistical Association, Proceedings of the Section on Survey Research Methods, (1995) 114-119
35. Lambert, D. Measures of Disclosure Risk and Harm,Journal of Official Statistics**,** 9**,** (1993) 313-331
36. Little, R. J. A. Statistical Analysis of Masked Data, Journal of Official Statistics, 9**, (**1993) 407-426
37. Little, R. J. A. and Liu, F. Selective Multiple Imputation of Keys for Statistical Disclosure-Control in Microdata, American Statistical Association, Proceedings of the Section on Survey Research Methods, CD-ROM, (2002)

38. Little, R. J. A. and Liu, F. Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata, American Statistical Association, Proceedings of the Section on Survey Research Methods, (2003)
39. Malin, B. Sweeney, L., and Newton, E. Trail Re-identification: Learning Who You are from Where You have Been, Workshop on Privacy in Data, Carnegie-Mellon University, March 2003.
40. McCallum, A. and Wellner, B. Object Consolidation by Graph Partitioning with a Conditionally-Trained Distance Metric, Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification, Washington DC, August 2003.
41. Moore, R. Controlled Data Swapping Techniques for Masking Public Use Data Sets, U.S. Bureau of the Census, Statistical Research Division Report rr96/04, (available at http://www.census.gov/srd/www/byyear.html) (1995).
42. Muralidhar, K., Parsa, R. and Sarathy, R. A General Additive Data Perturbation Method for Database Security, Management Science, 45 (10), (1999) 1399-1415
43. Muralidhar, K., Sarathy, R. and R. Parsa, R. An Improved Security Requirement for Data Perturbation with Implications for E-Commerce, Decision Sciences, 32 (4), (2001) 683-698
44. Paas, G. Disclosure Risk and Disclosure Avoidance for Microdata, Journal of Business and Economic Statistics, 6, (1988) 487-500
45. Palley, M. A. and Simonoff, J. S. The Use of Regression Methodology for the Compromise of Confidential Information in Statistical Databases, ACM Transactions on Database Systems, 12 (4), (1987) 593-608
46. Polettini, S. Maximum Entropy Simulation for Microdata Protection, Statistics and Computing, 13 (4), (2003) 307-320
47. Polettini, S. and Stander, J. A Bayesian Hierarchical Model Approach to Risk Estimation in Statistical Disclosure Limitation, in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases 2004, Springer: New York, (2004).
48. Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. Multiple Imputation for Statistical Disclosure Limitation, Journal of Official Statistics, 19, (2003) 1-16
49. Reiss, J.P. Practical Data Swapping: The First Steps, ACM Transactions on Database Systems, 9, (1984) 20-37
50. Reiter, J.P. Satisfying Disclosure Restrictions with Synthetic Data Sets, Journal of Official Statistics, 18, (2002) 531-543
51. Reiter, J.P. Inference for Partially Synthetic, Public Use Data Sets, Survey Methodology, (2003)
52. Reiter, J.P. Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study, Journal of the Royal Statistical Society, A, (2004)
53. Rinott, Y. On Models for Statistical Disclosure Risk Estimation, UNECE Work Session on Statistical Data Confidentiality, Luxembourg, April 2003, http://www.unece.org/stats/documents/2003/04/confidentiality/wp.16.e.pdf
54. Roque, G. M.. Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Ph.D. Dissertation, University of California at Riverside, (2000)
55. Sarathy, R., Muralidhar, K., and Parsa, R.. Perturbing Non-Normal Attributes: The Copula Approach, Management Science, 48 (12), (2002) 1613-1627
56. Scheuren, F. and Winkler W. Regression Analysis of Data Files that are Computer Matched – Part II, Survey Methodology (1997) 157-165
57. Schlörer, J. Security of Statistical Databases: Multidimensional Transformation, ACM Transactions on Database Systems, 6, (1981) 91-112
58. Skinner, C. J. and Elliot, M. A. A Measure of Disclosure Risk for Microdata, Journal of the Royal Statistical Society, B 64 (4), (2001), 855-867
59. Skinner, C. J. and Holmes, D. J. Estimating the Re-identification Risk per Record in Microdata, Journal of Official Statistics, 14 (1998) 361-372

60. Sweeney, L. Computational Disclosure Control for Medical Microdata: The Datafly System, in Record Linkage Techniques 1997, Washington, DC: National Academy Press (1999) 442-453
61. Sweeney, L. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10 (5), (2002) 571-588
62. Thibaudeau, Y. and Winkler, W.E. Bayesian Networks Representations, Generalized Imputation, and Synthetic Microdata Satisfying Analytic Restraints, Statistical Research Division report RR 2002/09 at http://www.census.gov/srd/www/byyear.html (2002)
63. Trottini, M. and Fienberg, S. E. Modelling User Uncertainty for Disclosure Risk and Data Utility, International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10 (5), (2002) 511-528
64. Van Den Hout, A. and Van Der Heijden, P. G. M. Randomized Response, Statistical Disclosure Control, and Misclassification: A Review, International Statistical Review, 70 (2) (2002) 269-288
65. Willenborg, L. and De Waal, T. Statistical Disclosure Control in Practice, Vol. 111, Lecture Notes in Statistics, Springer-Verlag, New York (1996)
66. Willenborg, L. and De Waal, T. Elements of Statistical Disclosure Control, Vol. 155, Lecture Notes in Statistics, Springer-Verlag, New York (2000)
67. Winkler, W. E. Matching and Record Linkage, in B. G. Cox (ed.) Business Survey Methods, New York: J. Wiley, (1995) 355-384
68. Winkler, W. E. Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, Research in Official Statistics, 1, (1998) 87-104
69. Winkler, W. E. Issues with Linking Files and Performing Analyses on the Merged Files, Proceedings of the Sections on Government Statistics and Social Statistics, American Statistical Association, (1999) 262-265
70. Winkler, W. E. Single Ranking Micro-aggregation and Re-identification, Statistical Research Division report RR 2002/08 at http://www.census.gov/srd/www/byyear.html (2002)
71. Yancey, W.E., Winkler, W.E., and Creecy, R. H. Disclosure Risk Assessment in Perturbative Microdata Protection, in (J. Domingo-Ferrer, ed.) Inference Control in Statistical Databases, Springer: New York (2002)