RESEARCH REPORT SERIES
(*Statistics #2004-03)*

Re-identification Methods for Masked Microdata

William E. Winkler

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

Report Issued: April 21, 2004

Re-identification Methods for Masked Microdata

William E Winkler[1]

[1] US Bureau of the Census, Washington, DC 20233-9100, USA,
william.e.winkler@census.gov

Abstract. Statistical agencies often mask (or distort) microdata in public-use files so that the confidentiality of information associated with individual entities is preserved. The intent of many of the masking methods is to cause only minor distortions in some of the distributions of the data and possibly no distortion in a few aggregate or marginal statistics In record linkage (as in nearest neighbor methods), metrics are used to determine how close a value of a variable in a record is from the value of the corresponding variable in another record. If a sufficient number of variables in one record have values that are close to values in another record, then the records may be a match and correspond to the same entity. This paper shows that it is possible to create metrics for which re-identification is straightforward in many situations where masking is currently done. We begin by demonstrating how to quickly construct metrics for continuous variables that have been micro-aggregated one at a time using conventional methods. We extend the methods to situations where rank swapping is performed and discuss the situation where several continuous variables are micro-aggregated simultaneously. We close by indicating how metrics might be created for situations of synthetic microdata satisfying several sets of analytic constraints.

1 Introduction

With the advent of readily available computing power and straightforward software packages, many users have requested that significantly more public-use files be made available for analyses. To create the public-use files, statistical agencies mask or distort confidential data with the intent that records associated with individual entities cannot be re-identified using publicly available non-confidential data sources. Agencies adopted many of the masking methods primarily because they could be easily programmed and because other agencies had used the methods.

The primary intent in producing a public-use file is to allow users to reproduce (approximately) certain statistical analyses that might be performed on the original, confidential microdata. To produce such files, agencies typically need to describe what masking method or methods they used, adjustments that users might need during a statistical analysis to get results corresponding to results on the confidential microdata, and limitations of the distributional characteristics of the data. The masked microdata are often intended to reproduce some of the distributional characteristics of individual variables and groups of variables. For instance, if users want to produce a regression

analysis $Y = X \beta$, the agency may use a masking method that allows this type of regression. Palley and Simonoff [26] have observed that, if the user has the independent X variables in a file, then the user may use the beta coefficient β to predict, say, an income variable Y. For a high-income individual with other known characteristics such as age range, race, and sex, the income variable may be sufficient to allow re-identification. The value of the Y variable that can be associated with the individual might be referred to as a *predictive disclosure* of confidential information. Alternatively, if the user (or intruder) has a file with a variable Y' that corresponds to the Y variable and has information about age range, race, and sex of an individual, the intruder may be able to re-identify an individual with the X information in the public-use file.

During the re-identification of (age range, sex, race, Y) with (age range, sex, race, Y'), we use a crude metric that states that the first three variables should agree exactly and the last variables Y and Y' should agree approximately. If Y and Y' are known to be in the tails of a distribution such as a high income or unusual situation, then we may be able to deduce that a range in which we can say Y and Y' are likely to be approximately the same. In some situations, we have a crude functional relationship $f(X) = Y$ that allows us to associate the X variables with a predicted Y variable that may be close to a corresponding Y' variable. In these situations, we can think of the functional relationship $f(X) = Y$ and other knowledge as yielding a metric for the distance between Y and Y'. The variables Y and Y' can be thought of as *weak identifiers* that allow us to associate a record in the public-use file with one or more records in the intruder's file. The intruder's file may contain publicly available information along with an identifier such as name and address.

Statistical agencies have often evaluated the potential confidentiality of a public-use file using elementary methods. For instance, they may have a public-use file with a combination of (X_D, X_C) with discrete variables X_D and continuous variables X_C that they wish to associate with a potential intruder file (X_D', X_C'). In some situations, the intruder file (X_D', X_C') might be the original, unmasked data. To compare records, they may apply a sort/merge utility that does an exact comparison of (X_D, X_C) with (X_D', X_C'). Because continuous variables often show minor variations, the exact comparison will not identify corresponding records. In making the comparison, the agency may be assuming that the intruder only has continuous and other data that do not correspond exactly to the agency's original data or to the masked data. Additionally, the agency may often assume that the intruder has a subset of the data (X_D', X_C') to compare with (X_D, X_C). A complication associated with the assumption that the intruder has a subset of the data (X_D', X_C') is that it often only takes a subset of the data to identify a small proportion of the records. Additionally, as noted above, the intruder may have knowledge of the analyses that can be performed using the data (X_D, X_C) and additional data sources that allow him to create a larger subset of the variables in (X_D', X_C').

More sophisticated record linkage and other methods have been developed in the computer science literature. The newer re-identification methods were developed for linking administrative lists (Winkler [42]) in which names and addresses were not of sufficient quality for accurate linkages. Individuals use other identifiers such as geographic identifiers (when available), numeric data such as income and mortgage pay-

ments that are often available in commonly used public databases, and functional relationships between variables (Scheuren and Winkler [32]). Much more sophisticated link analysis methods McCallum and Wellner [20] and Bilenko et al. [1] are currently being developed for linking information from large numbers of web pages nearly automatically. Further, if users have fairly sophisticated knowledge of how the masking is done and of the analyses that might be done with a file, they can use additional, ad hoc, file- and analysis-specific methods to re-identify (Lambert [19])

This paper demonstrates how to quickly construct new metrics associated with individual variables and groups of variables and add them to re-identification software. The construction of the more sophisticated metrics for linking administrative lists and link analyses has become increasingly more straightforward. Covering the most advanced methods is beyond the scope of this paper. We begin by showing how to construct metrics to re-identify in situations where variables in files are micro-aggregated. It is straightforward to extend to situations where moderately high proportions can be re-identified even when sampling proportions are on the order of 0.1% and upwards of 30% of the variables have errors in them [44], [23].

A fundamental concept is that, if the users of the microdata have accurate information about the underlying distributions of variables in confidential files, then they may have information sufficient for re-identification of some of the records. If the statistical agency needs to evaluate the confidentiality of a potential public-use file, then it needs to use more sophisticated re-identification methods than those that have sometimes been used previously. If a small proportion of records in the potential public-use file can be re-identified, then it may be possible to apply additional masking procedures to *coarsen* the file. The coarsening is intended to reduce or eliminate re-identification while preserving many of the analytic properties of the masked file. Kim and Winkler [17] determined that a small proportion of records in a file masked using additive noise only might be re-identified. To better assure confidentiality of the public-use file, they applied an additional masking procedure in which they swapped information in a set of specified subdomains. After determining that the first and second masking procedures assured disclosure avoidance, they released the public-use file.

In ordinary single-ranking micro-aggregation, we sort each individual variable and aggregate the values of each variable into groups of size k. In each group, we replace the individual values by an aggregate such as a median or the average. Typically, k is taken to be 3 or 4. If k is greater than 4, then varying individuals have shown that basic analytic properties such as regression can be seriously affected. Our key idea is that micro-aggregation almost precisely tells us the underlying distributions of individual variables. It is straightforward to construct highly optimal metrics based on the reported micro-aggregates in the public-use file. If two or three variables are uncorrelated, then the metrics associated with them may allow re-identification. If there is sampling at very low proportions and only a moderate proportion of variables (say 3 of 10) have severe distortion, then the redundancy due to the accurate variables can overcome the inaccuracy of the remaining variables.

In developing additional metrics for other masking situations, we make the key assumption that the public-use file is created in a manner that allows one or two sets of analyses. If the public-use file cannot be demonstrably proven to have analytic prop-

erties, then the distributions of the masked variables might not allow us to construct highly specific metrics that are optimized for re-identification in a specific set of files. For instance, in an extreme situation, we might replace the value of each continuous variable with its average value over the entire file. If the discrete variables associated with a record by themselves do not allow re-identification, then we would not be able to re-identify using the combination of discrete and continuous variables. In this extreme situation, it is unlikely that the analytic properties of the microdata file would be any more useful than the tabulations in the tables in existing publications.

The outline of this paper is as follows. In the second section, we give additional background on the identification methods that are currently being used in computer science. With background in the methods, it becomes straightforward to understand the construction of metrics, the use of redundant information, and the notions of distance between objects (records) taken from a variety of sources. In the general computer science setting of matching administrative and other lists, the interest is on identification rates above 90%. For potential masked, public-use files, we only need the much more attainable 0.1-2% re-identification rates to determine that the masking is not sufficient to release the file to the public. In the third section, we present a summary of how to construct the metrics that can be added to software to enhance re-identification. We also go into some detail about the underlying concepts to provide background on how straightforward it is to create metrics for other situations. In the fourth section, we show how to construct metrics for rank swapping and indicate how they might be constructed in other situations. In the fifth section, we provide ideas related to what might be expected in terms of re-identification with synthetic data. On the surface, it is intuitive that synthetic data is artificial and does not correspond to any individual entity. Fienberg [10], among others, has indicated that if sufficient analytic restraints are placed on the synthetic data, then some (approximate) re-identification might be possible. We provide ideas on how we could associate a small proportion of the records in a synthetic database with specific individuals. Based on the linkages, we show how the information in the synthetic data would allow us to deduce confidential information about an individual entity. In the sixth section, we provide additional caveats by connecting the methods of this paper with some of the ideas and results in the literature. The final section consists of concluding remarks.

2 Background

With increasing demand from users, statistical agencies are creating more and more masked, public-use files that can be analyzed. Agencies seldom demonstrate that the public-use files can be used for analyses that correspond reasonably well to analyses that might be done with the original, non-public data. Users assume that the masked, public-use files can be used for many analyses. They also assume that records in the files cannot be re-identified with individuals using publicly available non-confidential files.

Over the last few years, there has been remarkable progress, primarily in the computer science literature, in linking records associated with individual entities (ei-

ther persons or businesses) from a variety of sources. Some of the ideas originated in record linkage (see e.g. [32]) where many identifiers such as name, address, age, and income can have substantial errors across files. With economic variables such as income from one source file and receipts from another file, Scheuren and Winkler [32] have shown how to significantly increase accuracy in linkages of administrative lists. The correlations between the variables allowed Scheuren and Winkler [32] to create additional weak identifiers called predictors. Straightforward metrics associated with the predictors brought records from one file closer to smaller subsets of other records in another file in comparison to the situations in which the predictors and associated metrics were not used. Using ideas that were independently suggested by economists and record linkage practitioners at Statistics Canada, Winkler [43] has shown how to use auxiliary population files to improve linkage accuracy. For instance, assume that the population file contains a set of variables $Z_1, \ldots, Z_n$ that are either contained in one of the files being matched but not in both simultaneously. If a record from one file is associated with several records in the second file, then the Z-variables may reduce the association to 1 or 0 records in the second file.

Using Probabilistic Relational Models, Getoor et al. [12], Taskar et al. [33], [34], [35], [36], and Koller and Pfeffer [18] have shown how to systematically and iteratively improve linkages in a set of files. In extreme situations, Torra [38], [39] has shown how to create aggregates from quantitative and other data that can be used for linkages. Further, McCallum and Wellner [20] and others have shown how to use Markov Random Fields and Graph Partitioning algorithms to systematically increase the likelihood of a set of linkages between corresponding records in a group of files. The latter methods are often used for extracting and linking information (entities or objects) from a group of web pages.

In this paper we concentrate on the most elementary of the methods that correspond (roughly) to distance between records in a metric space such as might be used in nearest neighbor matching. Record linkage methods use metrics that scale the ranges of variables automatically and partially account for dependencies between variables (Winkler [42]). We show the validity of the re-identification ideas for masking methods that are known to produce analytically valid micro-data in a few situations. Because micro-aggregation methods are often used by statistical agencies, we begin by demonstrating how to compute re-identification metrics for micro-aggregation and show how to create analogous metrics for other situations. If we understand how easily masked, public-use records can be re-identified using the elementary methods that are being increasingly applied, then it may be possible to develop better masking strategies. We do not cover details of the new, more advanced methods. The advanced methods might be applied in situations where the more elementary methods do not allow accurate re-identification. Their application might lead to even better protection strategies.

3 Re-identification for Single-Ranking Micro-aggregation

In this section we provide a summary of the basic re-identification ideas given by Winkler [44]. Muralidhar [23] independently verified the methods. Both Muralidhar and one of his graduate students were each independently able to verify the ease in constructing highly optimized re-identification metrics and their efficacy in re-identification with micro-aggregated data. We provide additional intuition on the underlying concepts so that the extensions and related ideas in subsequent sections can be better understood.

We consider a rectangular data base (table) having fields (variables) X_i $i=1,\ldots,n$, and value states x_{ij}, $j=1,\ldots,n_i$. In many microdata confidentiality experiments, users want 10 or more variables X_i. We assume that each of the variables X_i is continuous, skewed, and not taking zero value states. The second assumption eliminates a few additional technical details. It can easily be eliminated. The third assumption is for convenience. It is not generally needed for the arguments that follow.

We begin our discussion by considering databases with 1000 or more records and situations in which micro-aggregation is on one variable at a time. In this discussion, we demonstrate that micro-aggregation as currently practiced allows almost perfect re-identification with existing record linkage procedures even when k is greater than or equal to 10. We can easily develop nearest-neighbor methods with similar metrics that have almost 100 percent re-identification rates.

We chose any three variables, say X_1, X_2, and X_3 that are pairwise uncorrelated ($R^2 \leq 0.2$). Our procedure is for aggregating variables one at a time. Although we use 3 variables in the following description, re-identification may occur with only two or with four or more variables in analogous other situations. Within each variable, sort the values and aggregate into groups of size 3 or more. Let the new micro-aggregated value-states be denoted by $a_i(x_{ij}) = y_{ij}$, $j = 1, \ldots. k_i$, $i = 1, 2, 3$ where $a_i()$ is the aggregation function. Each (aggregated) value state is assumed three or more times (3 or more records have the same value of the y-variables). Most aggregates will be from three value-states only. In the following $y_{i,ji}$ will denote the j_i value-state of micro-aggregated variable Y_i. The micro-aggregated value $y_{i,ji}$ will be a value such as the average or median. Such a value is in the range of the values being micro-aggregated. We develop new record linkage metrics (or nearest neighbor metrics) as follows. The metrics are for matching a micro-aggregated record R with the original set of data records. Let $R = (y_{1,j1}, y_{2,j2}, y_{3,j3}) = (a_1(x_{1,k1}), a_2(x_{2,k2}), a_3(x_{3,k3}))$ where y_i's are values aggregated by the aggregation operator a_i (.) from original values x_i's. Using the sort ordering for individual variables, for each i, let $p(y_{i,ji})$ be the predecessor of $y_{i,ji}$ and $s(y_{i,ji})$ be the successor of $y_{i,ji}$. In each situation, the predecessor and the successor are distinct from the value $y_{i,ji}$. For $y_{j,ii}$, let the distance be metric $dist(x, y_{i,ji})$ be 1 if x is within distance min ($abs(y_{i,ji} - p(y_{i,ji}))$, $abs(y_{i,ji} - s(y_{i,ji}))/2$ of $y_{i,ji}$; 0, otherwise.

This metric allows us to match the X-variables in the original file with the Y-values in the micro-aggregated file. The metric is highly optimized. It is based on the distribution of the micro-aggregated variables in the public-use file. Each X-variable in the intruder file will be associated with at most one Y-variable in the public-use file. Suitable adjustments should be made for being at the end of the distributions (i.e., one-sided). Let N be the number of records in the original database. Then micro-

aggregated record R has probability close to one of matching with its true corresponding original record. The probability is at least $((N-3)/N)$ on each field X_i that we match with its corresponding micro-aggregated field Y_i. It has probability close to zero of matching with any record other than its original corresponding record on each field. Based on independent empirical work, the possibility of re-identification with single-ranking micro-aggregation has been observed by Domingo et al. [7], [9]. The procedure described above provides a systematic method of re-identifying in all situations where micro-aggregation is used.

The extensions to cover more general (and realistic) situations are straightforward (Winkler [44]). The intruder, who may be using public use data, will often have name, address, and other identifying information that can be associated with individual records. The intruder will only use the quantitative data to associate a record in his files with a corresponding record in the public-use file. If the record of the intruder and the record in the public use file are linked with reasonably high probability, then re-identification occurs. If sampling fractions as low as 0.001 are used, then the metrics can be constructed that still allow us to separate a moderate proportion of records in the public-use sample and associate with it in the intruder database. If a moderate number of variables among a group of ten or more variables have severe error (above thirty percent), then we may still be able to use the remaining, more accurate variables to link records.

We repeat some of the key concepts regarding re-identification to improve the intuitive understanding. If we know that we have two overlapping populations, then it may only take two or three variables to re-identify a proportion of the records. Each variable is a weak identifier that allows us to associate the record in the public-use file with a subset of the records in the intruder data files. Each variable allows us to associate the variable with a different subset of records. Record linkage procedures (or more crudely nearest neighbor procedures) allow us to take an efficient intersection of the records in the second file that might be related to the record in the first file. In a number of situations, this procedure allows us to re-identify a proportion of the records in the masked file with high probability. The record linkage methods are good (efficient) at automatically accounting for the redundancy in a set of agreeing variables (see e.g., [17], [45]).

4 Re-identification for Other Basic Masking Methods

In this section, we describe possible extensions of the metric-construction procedures to rank swapping. We provide issues related to constructing metrics for data in which micro-aggregation is by several variables simultaneously and in which additive noise is used.

Rank swapping (Moore [22]) has somewhat similar characteristics to micro-aggregation. We begin by sorting individual continuous variables in the file, say, in decreasing order. An a priori rank-swapped range p is chosen in which each value of each variable in a record is swapped with the value of a corresponding value of the same variable in another randomly chosen record that is within p% of the ordered

range of the first record. The proportion p is typically between 15 and 20%. If the number of records in the file is even, then each record is swapped once. If the number of records is odd, then one additional record will be swapped twice to assure that the value in the extra record is swapped once. At the very end of the swapping, records (those records with the smallest values) cannot be swapped the full p% of the range.

If the rank-swapping procedure were repeated over different sets of random numbers, then on average the replacement value for a given record would be the average of the records in the p% range of the records. As with micro-aggregation, if the value p is above 1% and the number of records in the p% range is above 10, then the analytic distortions in the resultant data can be very severe. This is particularly true on a subdomain in which the rank-swapping procedure exchanges values of records in the subdomain with values with records in other subdomains.

Even in extreme situations, we will be able to re-identify. If p is equal 1% and the number of records in the p% range is over 100, each value of a variable allows us to construct a metric in which a given record can be associated with at most 1-2% of the other records. This is similar to the micro-aggregation situation. Each value of a variable in a record is a weak identifier that allows us to tentatively link the record to 1-2% and tentatively not link the record to 99-98% of the records in the file. As we accumulate potential linkages over several of the weak-identifiers (variables), we can link a moderate or small proportion of the records with reasonable confidence (probability above 0.5).

If we are matching the masked file with the original unmasked file (i.e., no sampling) and we assume that the original values do not have severe (more than 30% distortion), then, with three uncorrelated variables and the newly constructed metrics, it seems likely that we will be able to re-identify a moderate to high proportion of records. If there is sampling with small proportions and there is a substantial number of variables (say 10) in which a small proportion of variables have severe errors, then it seems likely that we will still be able to re-identify a moderate or small proportion of the records. The re-identification proportion may be less than the corresponding proportions in the micro-aggregation situation because the rank-swapping optimized metrics are used in a much greater range of values (i.e., much larger k) than in the micro-aggregation situation.

Domingo-Ferrer and Mateo-Sanz [6] have shown how to micro-aggregate using two or more variables simultaneously. As they show, analytic properties of the masked data degrade much more rapidly than in single-variable micro-aggregation. The degradation is intuitive because if we simultaneously micro-aggregate on three uncorrelated variables, then the resultant aggregates of the three variables are unlikely to preserve correlations among themselves and with other continuous data in the files. A crude analogy is if we use k-means to cluster data and then micro-aggregate within clusters. If we micro-aggregate using more than three variables simultaneously, then the degradation of analytic properties is likely to be greater than in the three-variable simultaneously situation. In many situations, it seems likely the multi-variable micro-aggregation procedure will preserve confidentiality. The confidentiality is due to k-anonymity [30], [31] because each masked record is likely to be associated with at least k records.

At present, we are uncertain how to compute highly specific re-identification metrics for additive noise (Kim [15], Fuller [11]) or mixtures of additive noise (Yancey et al. [45]). Yancey et al. [45] showed that mixtures of additive noise provided a ten-fold reduction in re-identification rates in comparison to additive noise (Kim and Winkler [17]) while not seriously compromising analytic properties. Brand [2], in a nice tutorial paper, has given more details on how additive noise can compromise the analytic properties of the masked data. Her ideas might be used to determine additional re-identification metrics. Recent work (Kargupta et al. [14]) suggests how better re-identification metrics for additive noise might be constructed. Because Kargupta et al. [14] makes use of ideas from the signal processing and random matrix literature, it may target data situations with considerably less inherent noise and variation than survey data. The Domingo et al. [8] density-estimation procedure is intended to estimate a reconstructed probability distribution of the original, unmasked data. It may suffer from the curse-of-dimensionality problems where the amount of data needed in multivariate situations grows at an exceptionally high exponential rate. It is not clear that the Kargupta et al. [14] or the Domingo et al. [8] procedures can deal with mixtures of additive noise. Even in the situations using ordinary (non-mixture) additive noise, more research is needed to determine whether the methods of Kargupta et al. [14] or the Domingo et al. [8] would yield re-identification rates higher than those obtained by Kim and Winkler [17].

5 Synthetic Data

Palley and Simonoff [26], Fienberg [10], and Reiter [29] have pointed out that it may still be possible that synthetic data may contain some records that allow re-identification of confidential information. Fienberg [10] has given additional methods of re-identification that can be used with either original data that has been masked or synthetic data. Individuals create synthetic data from models that preserve some of the distributional assumptions of the original, confidential data and allow a few analyses that correspond roughly to the analyses that might be performed on the original, unmasked data. An outlier or value in the tail of a distribution in a record in the synthetic data may be much closer to the record in the original data and the corresponding record data available to the intruder than to other information. The intuition is that, when much of the synthetic data are in the interior of a distribution, an intruder can only determine that the synthetic record corresponds to $k \geq 3$ records in the intruder's file. The outlier may allow the intruder to determine that the synthetic record is likely to correspond to at most one record in the intruder's file.

In this section, we describe a situation in which some of the synthetic data might be re-identified (in terms of yielding values in fields or variables) that are reasonably close to confidential values of those fields and can be associated with names and other identifiers in the intruder's file. We make several assumptions that have been made by others who have provided methods for generating analytically valid synthetic microdata. The first is that the original microdata is free of frame errors from undercoverage and duplication and free of edit/imputation errors. The second is that the model

(or set of distributions) that is based on the original microdata allows a reasonable number of analyses in the synthetic data that correspond to analyses on the original microdata. The data producer describes the distributional assumptions and the possible limitations of any analyses so that users of the synthetic data can perform valid analyses within the limitations of the synthetic data.

The re-identification metrics we might use are determined by the set of plausible probability distributions that correspond to the synthetic data. Each distribution can determine one or more metrics. The simplest situation is that described by Fienberg [10] for synthetic data and Lambert [19] for any data. If a value is an outlier in a distribution, then there may only be one plausible value in a record of the intruder that corresponds to that value of the record in the masked file. The identifying information in an intruder's file can be used to compromise even the synthetic data. In this discussion, we use the term *outlier* to represent a value of a variable that is in the tail of a distribution. If the synthetic data allows more and more analyses, it will have corresponding more distributions and metrics that can be used to determine more outliers. Each of the outliers in the distributions of the synthetic data may yield re-identifications. As done by Palley and Simonoff [26] and Fienberg [10], it is possible that information from the non-outliers in the synthetic data and aggregate characteristics of the population such as what types of individual entities in the population may yield information for further improving the re-identification of outliers. A class of examples for which this is true is the class of regression relationships that give good predictive power (i.e., low variance in this situation) of a given variable such as income when the values of other variables and the coefficients of a valid regression relationship are known.

To cast further insight, we provide more detailed examples. The first example is where we produce synthetic data for which only one very simple set of analyses might be performed and for which re-identification is highly unlikely. The example builds intuition on why valid distributional properties in the synthetic data are necessary for building re-identification metrics. The example corresponds roughly to the ideas of Kim and Winkler [17]. We have data (X, S) where X is continuous microdata corresponding to information such as income and mortgage and S is discrete corresponding to information such as age, race, and sex. The potential users of the data specify that they wish to perform regression analyses on the data with the emphasis on the subdomain specified by the S-variables. We obtain the means and covariances of the X variables on each of the subdomains determined by S. We generate synthetic data Y such that the means and covariances of the Y-variables on the subdomains correspond to the means and covariances of the X-variables on the same subdomains. Sample sizes in each subdomain are taken to be at least 500 because covariances of the Y variables do not stabilize to values of covariances of X until sample size is sufficiently large. The slow stabilization is due to the nature of generating multivariate random variables satisfying a number of analytic restraints.

Because there are an exceptionally large number of ways to generate the Y-variables, it is intuitive that in many situations there is no chance that the outliers in the Y distribution will correspond closely with outliers in the X distribution with high probability. The exception is when the X-variables have multivariate normal distribution, the Y-variables have multivariate normal distribution and the number of dimen-

sions (variables) *n* are increased (Paas [25], Mera [21]). We observe that users of the synthetic data will be able to reproduce (approximately) regression analyses on some of the subdomains. The data, however, are virtually useless because they cannot even be used for regression on the entire file (independent of the S variables) and examination of simple statistics such as rank correlations. If we put additional restraints on the generated Y variables such that it preserve regressions on some of the subdomains from aggregating the basic subdomains from the S variables and preserving a few of the rank correlations, then it is likely that we will need to have a considerably larger sample size in each of the subdomains determined by S and that the set of analytic restraints will yield some outliers in the Y data that lead to re-identification.

We need to better understand how valid analytic relationships, including certain aggregates such as regression coefficients and covariances of variables can yield predictive ability with properly constructed metrics that correspond to valid analyses. To do this we need to give an overview about how one creates a model for data. In the following, we will use *model*, *valid parametric form*, and *distribution* to mean the same thing. If we generate synthetic data from the valid model, then the synthetic data will satisfy one or two analytic properties of the original data.

Although there are a number of good examples of the modeling process, we prefer Reiter [29] because it is representative of and builds on a number of good ideas introduced in earlier work. Using the general multiple-imputation framework of Raghunathan et al. [28], Reiter shows how to create a model through a systematic set of regressions (or imputation models) that use subsets of variables to predict other variables. The key component of the modeling procedure is the estimation of the conditional and joint probabilities associated with the data and the proposed analysis. Although Reiter's method targets multiple-imputation, we could also use maximum entropy methods (Polletini [27]), multivariate density estimation (Domingo et al. [8]), or Latin Hypercube methods (Dandekar et al. [4], [5]).

The models have the effect of estimating the probability distribution of a variable conditional on specific values of the (independent) predictor variables. The estimated distributions can serve as predictors of the value of a variable given the values of the variables upon which it is conditioned. The estimated distributions can be used as new metrics for re-identification. As Reiter observes: "When there are parameters with small estimated variances, imputers can check for predictive disclosures and, if necessary, use coarser imputation models." Alternatively, we might state this as "If the model allows predictive values that might lead to re-identification, then we might reduce the analytic effectiveness of the model to reduce predictive disclosure risk."

What we can observe is that the Reiter example (and earlier examples due to Palley and Simonoff [26], Lambert [19], and Fienberg [10]) uses one variable (possibly in the tail of the distribution or in a suitably narrowed range) to obtain predictive disclosure. If we use a substantial number of variables in the model and potential predictive disclosures are possible with different subsets of them, then it is possible that predictive disclosure can increase if suitable metrics are created and placed in record linkage software. This is a research problem.

6 Discussion

In the earlier part of the paper, we demonstrated how to quickly create metrics that allow re-identification with widely used masking procedures such as micro-aggregation. A key feature was that micro-aggregation, as it is typically applied, gives exceptionally good information about the distribution of each individual variable. If the distributional information is used to create a set of metrics for a set of variables, then high rates of re-identification are quite possible. This is particularly true is there are a moderate number of continuous variables that are pairwise uncorrelated or only moderately correlated.

If synthetic data is produced through a valid parametric modeling procedure, then we suspect that re-identification rates using single variables may be in the range 0.001-0.01. Our low estimate of re-identification rates is based on the re-identification rates with mixtures of additive noise (Yancey et al. [45]. With mixtures of additive noise, re-identification at rates of 0.001 to 0.01 occurs primarily with outliers in the tails of distributions. Although determination of general re-identification rates with different types of synthetic data is a research problem, we would expect the re-identification rates to be relatively low. Rates this low may be sufficient to assure confidentiality. Palley and Simonoff [26], Lambert [19], Fienberg [10], and Reiter [29] have all given examples specific to different types of files and types of analyses that make plausible conjectures with respective to predictive disclosure using single variables at a time. If we use a substantial number of variables and have valid information about a model (i.e. probability distribution) representing them, then it seems likely that we can construct metrics that increase re-identification rates. In the simpler situations where re-identification might occur, Reiter [29] suggests coarsening the models to reduce predictive disclosures.

A straightforward procedure may be for the data producer to perform a direct re-identification between the synthetic data and the original data used in the modeling. This can quickly identify potential records in the synthetic data that may lead to predictive disclosure or identify disclosure. Kim and Winkler [17] delineated those records that were most at risk of re-identification when additive noise was used. Their coarsening procedure was to swap information in the at-risk records with the not-at-risk records. As noted by [17], the coarsening procedure had the effect of reducing a number of the analytic properties of the public-use data. Alternatively, Yancey et al. [45] used mixtures of additive noise that reduced disclosure risk by a factor of 10 while not significantly reducing the analytic properties in the masked file.

Dandekar et al. [4], [5], Grim et al. [13], and Thibaudeau and Winkler [37] have all given methods for generating synthetic microdata that do not involve as much detailed modeling effort as those presented by Reiter [29]. The Latin-Hypercube methods of Dandekar et al. [4], [5] may represent a practical alternative. The methods of Grim et al. [13] and Thibaudeau and Winkler [37] may use approximations that compromise many of the analytic properties.

6 Concluding Remarks

This paper provides methods for constructing re-identification metrics that can be used with a few of the masking methods that are commonly used by statistical agencies for producing public-use files. We concentrate on a few masking methods that are known to produce files with one or two analytic properties that correspond to data in unmasked confidential files.

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U. S. Census Bureau. The author thanks Dr. Nancy Gordon, Dr. Cynthia Clark, Dr. Tommy Wright, and two reviewers for comments leading to improved wording and explanation and to several additional references.

References

1. 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
2. Brand, R. Microdata Protection Through Noise Addition, in (J. Domingo-Ferrer, ed.) Inference Control in Statistical Databases, Springer: New York (2002)
3. Dalenius, T. and Reiss, S.P. "Data-swapping: A Technique for Disclosure Control," Journal of Statistical Planning and Inference, 6 (1982) 73-85
4. 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)
5. 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)
6. 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
7. Domingo-Ferrer, J. and Mateo-Sanz, J. M., Organian, A., and Torres, A. On the Security of Microaggregation with Individual Ranking: Analytic Attacks, International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10 (5), (2002) 477-492
8. Domingo-Ferrer, J., Sebe, F., and Castella, J. On the Security of Noise Addition for Privacy in Statistical Databases, in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases 2004, Springer: New York, (2004)
9. 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
10. 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).
11. Fuller, W. A. Masking Procedures for Microdata Disclosure Limitation, Journal of Official Statistics, 9**,** (1993) 383-406
12. Getoor, L., Friedman, N., Koller, D., and Taskar, B. Learning Probabilistic Models for Link Structure, Journal Machine Learning Research, 3, (2003) 679-707

13. Grim, J., Bocek, P., and Pudil, P. Safe Dissemination of Census Results by Means of Interactive Probabilistic Models, Proceedings of 2001 NTTS and ETK, Eurostat: Luxembourg, (2001) 849-856
14. Kargupta, H., Datta, S., Wang, Q., and Ravikumar, K. (2003) Random Data Perturbation Techniques and Privacy Preserving Data Mining, Expanded version of best paper awarded paper from the IEEE International Conference on Data Mining, November, 2003, Orlando, FL
15. 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
16. Kim, J. J. Subdomain Estimation for the Masked Data, American Statistical Association, Proceedings of the Section on Survey Research Methods, (1990) 456-461
17. Kim, J. J. and Winkler, W. E. Masking Microdata Files, American Statistical Association, Proceedings of the Section on Survey Research Methods, (1995) 114-119
18. Koller, D. and Pfeffer, A. Probabilistic Frame-Based Systems, Proceedings of the Fifteenth National Conference on Artificial Intelligience, (1998)
19. Lambert, D. Measures of Disclosure Risk and Harm,Journal of Official Statistics**,** 9**,** (1993) 313-331
20. 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
21. Mera, R. Matrix Masking Methods That Preserve Moments, American Statistical Association, Proceedings of the Section on Survey Research Methods, (1998) 445-450
22. 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) (1996)
23. Muralidhar, K. Verification of Re-identification Rates with Micro-aggregation, private communication (2003)
24. Muralidhar, K. and Sarathy, R. A Theoretical Basis for Perturbation Methods, Statistics and Computing, 13 (4), (2003) 329-335
25. Paas, G. Disclosure Risk and Disclosure Avoidance for Microdata, Journal of Business and Economic Statistics, 6, (1988) 487-500
26. 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
27. Polletini, S. Maximum Entropy Simulation for Microdata Protection, Statistics and Computing, 13 (4), (2003), 307-320
28. Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. Multiple Imputation for Statistical Disclosure Limitation," Journal of Official Statistics, 19, (2003) 1-16
29. Reiter, J.P. Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study, Journal of the Royal Statistical Society, A, (2004)
30. Samarati, P. Protecting Respondents' Identity in Microdata Release, IEEE Transactions on Knowledge and Data Engineering, 13 (6), 2001), 1010-1027
31. Samarati, P. and Sweeney, L. Protecting Privacy when Disclosing Information: k-anonymity and Its Enforcement through Generalization and Cell Suppression, Technical Report, SRI International (1998)
32. Scheuren, F. and Winkler W. Regression Analysis of Data Files that are Computer Matched – Part II, Survey Methodology (1997) 157-165
33. Taskar, B., Abdeel, P., and Koller, D. Discriminative Probabilistic Models for Relational Data, Proceedings of the Conference on Uncertainty in Artificial Intelligence (2002)

34. Taskar, B., Segal, E., and Koller, D. Probabilistic Classification and Clustering in Relational Data, Proceedings of the International Joint Conference on Artificial Intelligence (2001)
35. Taskar, B., Wong, M. F., Abdeel, P., and Koller, D. Link Prediction in Relational Data, Neural Information Processing Systems, (2003)
36. Taskar, B., Wong, M. F., and Koller, D. Learning on Test Data: Leveraging "Unseen" Features, Proceedings of the Twentieth International Conference on Machine Learning, (2003), 744-751
37. 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)
38. Torra, V. Re-Identifying Individuals Using OWA Operators, Proceedings of the Sixth Conference on Soft Computing, Iizuka, Fukuoka, Japan (2000)
39. Torra, V. OWA Operators in Data Modeling and Re-Identification, IEEE Transactions on Fuzzy Systems, to appear
40. Willenborg, L. and De Waal, T. Elements of Statistical Disclosure Control, Vol. 155, Lecture Notes in Statistics, Springer-Verlag, New York (2000)
41. Winkler, W. E. Matching and Record Linkage, in B. G. Cox (ed.) Business Survey Methods, New York: J. Wiley, (1995), 355-384
42. Winkler, W. E. Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, Research in Official Statistics, 1, (1998), 87-104
43. 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
44. 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)
45. 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)