Appendix to Disclosure Limitation in Longitudinal Linked Data

John M. Abowd
Cornell University, U.S. Census Bureau, CREST, and NBER

Simon D. Woodcock
Cornell University

August 20, 2001

Acknowledgement

Extracted from "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds.), (Amsterdam: North Holland, 2001), forthcoming. Reproduced with permission.

The research reported in this paper was partially sponsored by the U.S. Census Bureau, the National Science Foundation (SES-9978093), and the French Institut National de la Statistique et des Etudes Economiques (INSEE) in association with the Cornell Restricted Access Data Center. The views expressed in the paper are those of the authors and not of any of the sponsoring agencies. The data used in this paper are confidential but the authors' access is not exclusive. No public use data sets were released as a part of this research. Restricted access to the French data was provided to Abowd by INSEE through an agreement with Cornell University. The authors thank Benoit Dostie, Sam Hawala, Janet Heslop, Paul Massell, Carol Murphree, Philip Steel, Lars Vilhuber, Marty Wells, Bill Winkler, and Laura Zayatz for helpful comments on earlier versions of this research.

1 Appendix: Recent Research on Disclosure Limitation

In recent years, statistical agencies have seen an increasing demand for the data they collect, coupled with increasing concerns about confidentiality. This presents new challenges for statistical agencies, who must balance these concerns. Doing so requires techniques to allow dissemination of data that is both analytically useful and preserves the confidentiality of respondents.

This appendix presents recent research on disclosure limitation methods and concepts. Although our primary interest is in methods appropriate to longitudinal linked data, this topic has not been well-addressed in the literature. Hence, we review disclosure limitation methods and concepts appropriate to microdata in general. Since there are a number of reviews which provide a good summary of early research, (e.g., Subcommittee on Disclosure Avoidance Techniques (1978a), Subcommittee on Disclosure Avoidance Techniques (1994b), and Jabine (1993)), we concentrate on recent research only.

The review is divided into four parts. The first presents general research on the disclosure limitation problem. The second presents research on measures of disclosure risk and harm. Part three discusses recent research into disclosure limitation methods for microdata, and the final part discusses the analysis of disclosure-proofed data.

1.1 General Research and Survey Articles

Evans et al. (1996) This paper summarizes recent applications of a variety of disclosure limitation techniques to Census Bureau data, and outlines current research efforts to develop new techniques. The paper briefly discusses methods for microdata (including the two-stage additive noise and data-swapping technique of Kim and Winkler (1997), the rank-based proximity swapping method of Moore (1996b), and developing synthetic data based on log-linear models), methods under consideration for the 2000 Decennial Census, and methods for establishment tabular data (see the more detailed discussion of Evans et al. (1998) below).

Fienberg (1997) This paper presents an excellent review of the disclosure limitation problem and recent research to address it. The paper is organized in eight sections. The first section is introductory. In the second, the author defines notions of confidentiality and disclosure, presents the debate between limited access versus limited data, and describes the role of the intruder in defining notions of disclosure and methods of disclosure limitation. The third section presents two detailed examples to illustrate the issues, namely issues surrounding the Decennial Census and the Welfare Reform Act. The fourth section classifies various disclosure limitation methodologies that have been proposed, and illustrates them in some detail. The fifth section considers notions of uniqueness in the sample and uniqueness in the population, and their role in defining notions of disclosure risk (see the discussion of Fienberg and Makov (1998), Boudreau (1995), and Franconi (1999), below). The sixth section presents two integrated proposals for disclosure limitation: the ARGUS project (see the discussion of Hundepool and Willenborg (1999) and Nordholt (1999) below) and proposals for the release of simulated data (see Section 1.3.2 below). The seventh section presents a brief discussion of issues pertaining to longitudinal data, and section 8 concludes.

Winkler (1997) This paper briefly reviews modern record-linkage techniques, and describes their application in re-identification experiments. Such experiments can be used to determine the level of confidentiality protection afforded by disclosure limitation methods. The author stresses the power of such techniques to match records from disclosure-proofed data to other data sources. Emerging record-linkage techniques will allow re-identification in many existing public-use files, even though these files were produced by conscientious individuals who believed they were using effective disclosure limitation tools.

National Research Council (2000) This book describes the proceedings of a workshop convened by the Committee on National Statistics (CNSTAT) to identify ways of advancing the often conflicting goals of exploiting the research potential of microdata and preserving confidentiality. The emphasis of the workshop was on linked longitudinal data – particularly longitudinal data linked to administrative data. The workshop addressed four key issues. These are: (1) the trade-off between increasing data access on the one hand, and improving data security on the other; (2) the ethical and legal requirements associated with data dissemination; (3) alternative approaches for limiting disclosure risks and facilitating data access – primarily the debate between restricting access and altering data; and (4) a review of current agency and organization practices. Some interesting observations in the report include:

- Researchers participating in the workshop indicated a preference for restricted access to unaltered data over broader access to altered data. However, these researchers also recognized the research costs associated with the former option.
- Although linking databases can generate new disclosure risks, it does not necessarily do so. In particular, many native databases are already sensitive, and hence require confidentiality protection. Linking may create a combined data set that increases disclosure risk, however a disclosure incident

occurring from a linked data source is not necessarily caused by the linking. The breach of disclosure may have occurred in the native data as well.

- An appropriate measure of disclosure risk is a measure of the *marginal* risk. In other words, rather than comparing risks under various schemes with disclosure probability zero, one might consider the change in probability of disclosure as a result of a specific data release or linkage, or for adding or masking fields in a data set – the marginal risk associated with an action.
- In defense of data alteration methods, Fienberg noted that all data sets are approximations of the real data for a group of individuals. Samples are rarely representative of the group about which a researcher is attempting to draw inferences, rather it represents those for whom information is available. Even population data are imperfect, due to coding and keying errors, missing data, and the like. Hence, Fienberg finds the argument that perturbed data is not useful for intricate analysis not altogether compelling.

1.2 Measures of Disclosure Risk and Harm

Lambert (1993) This paper considers various definitions of disclosure, disclosure risk, and disclosure harm. The author stresses that disclosure is in large part a matter of perception – specifically, what an intruder *believes* has been disclosed, even if it is false, is key. The result of a false disclosure may be just as harmful (if not worse) than the result of a true disclosure. Having distinguished between disclosure risk and disclosure harm, the author develops general measures of these.

The author defines two major types of disclosure. In an *identity disclosure* (or *identification*, or *re-identification*), a respondent is linked to a particular record in a released data file. Even if the intruder learns no sensitive information from the identification, it may nevertheless compromise the security of the data file, and damage the reputation of the releasing agency. To distinguish between true identification and an intruder's beliefs about identification, the author defines *perceived identification,* which occurs when an intruder believes a record has been correctly identified, whether or not this is the case. An *attribute disclosure* occurs when an intruder believes new information has been learned about the respondent. This may occur with or without identification. The *risk of disclosure* is defined as the risk of identification of a released record, and the *harm from disclosure* depends on what is learned from the identification.

Suppose the agency holds N records in a data file Z, and releases a random sample $\mathsf{X} = (\mathsf{x}_1, ..., \mathsf{x}_n)$ of n masked records on p variables. Lambert (1993) defines several measures of perceived disclosure risk. A "pessimistic" risk of disclosure is given by:

$$\begin{aligned} D(\mathsf{X}) &= \max_{1\le j\le N}\max_{1\le i\le n} \Pr\left[i^{th}\text{released record is } j^{th} \text{ respondent's record}|\mathsf{X}\right] \qquad (1)\\ &= \max_{1\le j\le N}\max_{1\le i\le n} \Pr\left[\mathsf{x}_i \text{ is } j^{th} \text{ respondent's record}|\mathsf{X}\right]. \end{aligned}$$

Minimizing the measure in (1) protects against an intruder looking for the easiest record to identify. Alternate measures of disclosure risk can also be defined on the basis of (1) , for example:

$$D_{average}(\mathsf{X}) = \frac{1}{N}\sum_{j=1}^{N}\max_{1\le i\le n} \Pr\left[\mathsf{x}_i \text{ is } j^{th} \text{ respondent's record}|\mathsf{X}\right] \qquad (2)$$

$$D_{total}(\mathsf{X}) = ND_{average}(\mathsf{X}) \ . \qquad (3)$$

Equation (2) is a measure of data vulnerability based on the average risk of perceived disclosure, whereas (3) is a measure of the cumulative risk. An alternate measure of the total risk of perceived identification can be defined as the number of cases for which the risk of perceived disclosure exceeds a threshold τ :

$$D_\tau(\mathsf{X}) = \#\left\{ j : \max_{1 \le i \le n} \Pr\left[\mathsf{x}_i \text{ is } j^{th} \text{ respondent's record} | \mathsf{X} \right] \ge \tau \right\} .$$

The author proceeds to develop several detailed examples, and provide a general measure of disclosure harm, which is not presented here.

Fienberg and Makov (1998) This paper reviews several concepts, namely uniqueness in sample, uniqueness in the population, and some notions of disclosure. The main contribution is a proposed approach for assessing disclosure potential as a result of sample uniqueness, based on log-linear models. A detailed description of this method follows.

Suppose a population is cross-classified by some set of categorical variables. If the cross-classification yields a cell with an entry of "1" then the individual associated with this entry is defined as a *population unique*. Population uniqueness poses a disclosure risk, since an intruder with matching data has the potential to match his or her records against those of the population unique. This creates the possibility of both re-identification and attribute disclosure.

A *sample unique* is defined similarly – an individual associated with a cell count of "1" in the cross-classification of the sample data. Population uniques are also sample uniques if they are selected into the sample, but being a sample unique does not necessarily imply being a population unique. The focus of the Fienberg and Makov (1998) approach for assessing disclosure potential is to use uniqueness in the sample to determine the probability of uniqueness in the population. Note that sample uniqueness is not necessarily required for such an endeavor – "small" cell counts may also pose a disclosure risk. For example, a count of "2" may allow an individual with almost unique characteristics to identify the only other individual in the sample with those characteristics. If the intruder did not also possess these characteristics, then a cell count of "2" could allow the individuals to be linked to the intruder's data with probability 1/2. The extension to larger yet still "small" cell counts is obvious.

Let N denote the population size, n the size of the released sample, and K the maximum number of "types" of individuals in the data, as defined by the cross-classifying variables (*i.e.*, the total number of cells). Let F_i and f_i, $i = 1, ..., K$, denote the counts in the cells of the multiway table summarizing the entire population and sample, respectively. Then a crucial measure of the vulnerability of the data is given by:

$$\sum_{i=1}^{K} \Pr\left(F_i = 1 | \, f_i = 1\right) \tag{4}$$

Most prior attempts to estimate (4) assumed distributions for F_i and f_i (e.g., Bethlehem et al. (1990) and Skinner and Holmes (1993)). The Fienberg and Makov (1998) approach differs by assuming the released sample is drawn from a population with cell probabilities $\left\{\pi_i^{(N)}\right\}$ which follow a log linear model (including terms such as main effects interactions), of the form

$$\log(\pi_i^{(N)}) = g_N(\boldsymbol{\theta}_i) \tag{5}$$

where $\boldsymbol{\theta}_i$ are parameters. The authors propose to fit (5) to the observed counts $\{f_i\}$. Denote the estimated cell probabilities $\left\{\hat{\pi}_i^{(n)}\right\}$. Ideally, one would like to develop analytical formulae for $\Pr(F_i = 1 | f_i = 1)$, but this will frequently be infeasible since many of the log-linear models that result from the estimation process will not have a closed-form representation in terms of the minimal marginal sufficient statistics. Instead, the authors propose the following simulation approach. First, use the records on the n individuals in the sample $(x_1, ..., x_n)$ to generate $(N - n) \times H$ records from $\left\{\hat{\pi}_i^{(n)}\right\}$. This results in H populations of

size N, each containing $(N-n)$ "new" records obtained by some form of imputation (*e.g.*, see Little and Rubin (1987)), or multiple imputations from some posterior distribution (*e.g.*, see Rubin (1987)). Next, let $\bar{F}_i(j) = \bar{F}_i(x_1, ..., x_N, j)$ be the count in cell i of the jth imputed population. Similarly, let $\bar{f}_i = \bar{f}_i(x_1, ..., x_N)$ be the count in cell i of the released data (the sample). Clearly, $\bar{f}_i \neq 1 \Longrightarrow \bar{F}_i(j) \neq 1$. We can estimate (4) by:

$$\sum_{i=1}^{K} \widehat{\Pr}(F_i = 1 | f_i = 1) = \sum_{i=1}^{K} \sum_{j=1}^{H} \frac{\mathbf{1}\left[\left(\bar{F}_i(j) = 1\right) \cap \left(\bar{f}_i = 1\right)\right]}{H} \tag{6}$$

where the function $\mathbf{1}[A] = 1$ if A is true, and zero otherwise. Equation (6) can be used to assess the disclosure risk of the released data for a given release size n. Since (6) is likely to decrease as $(N-n)$ increases, the statistical agency is motivated to reduce n to the point that (6) indicates disclosure is infeasible. Note that if we remove the summation over i in (6), then we can obtain a cell-specific measure of disclosure risk.

Fienberg and Makov (1998) do not address the sample error of the estimate in (6). They also do not address the inherent trade-off that an agency faces when choosing n based on (6) between reduced disclosure risk and increased uncertainty in the released data.

Boudreau (1995) This paper presents another measure of disclosure risk based on the probability of population uniqueness given sample uniqueness. For the case of microdata containing discrete key variables, the author determines the exact relationship between unique elements in the sample and those in the population. The author also gives an unbiased estimator of the number of population uniques, based on sample data. Since this estimator exhibits great sampling variability for small sampling fractions, the author models this relationship. After observing this conditional probability for a number of real populations, the author provides a parametric formulation of it. This formulation is empirical only – it has not theoretical justification. However, the empirical formulation is much more flexible than earlier measures of disclosure risk based on uniqueness which required distributional assumptions (e.g. the Poisson-Gamma model of Bethlehem et al. (1990) or the Poisson-Lognormal model of Skinner and Holmes (1993)).

Willenborg and Kardaun (1999) This paper presents an alternate measure of disclosure risk appropriate to microdata sets for research (as opposed to public use files). In such files there is generally no requirement that all records be absolutely safe, since their use is usually covered by a contractual agreement which includes a non-record-matching obligation. The approach is to define a measure of the "degree of uniqueness" of an observation, called a *fingerprint*. A fingerprint is a combination of values of identifying (key) variables that are unique in the data set at hand, and contain no proper subset with this property (so it is a minimum set with the uniqueness property). The authors contend that records with "many" "short" fingerprints (i.e., fingerprints comprised of a small number of variables) are "risky", and should not be released. Appropriate definitions of "many," "short" and "risky" are at the discretion of the data collection/dissemination agency. In this way, defining disclosure risk in terms of fingerprints is very flexible. The authors propose that agencies use the fingerprinting criterion to identify risky records, and then apply disclosure-limitation measures to these records. The paper contains a discussion of some design criteria for an implementation of the fingerprinting criterion, and stipulates some useful heuristics for algorithm design.

Franconi (1999) This paper reviews recent developments in measures and definitions of disclosure risk. The author stresses differences between methods appropriate to social data and to business data. These differences are due to differences in the underlying data. Namely, social data are generally from large populations, have an inherent dependent structure (*i.e.*, groups such as families or households exist in the data), and are characterized by key variables of a categorical nature. These characteristics allow one to tackle the disclosure limitation problem via concepts of uniqueness. Business data, on the other hand are generally from small populations, with a skewed distribution, and have key variables which are primarily

continuous. Uniqueness concepts are generally not useful here, since nearly all cases would be considered unique. In both cases, the author stresses the need to take account of hierarchies in the data, such as the grouping of cases into families and households. These hierarchies provide additional information to an intruder attempting to identify records, hence they should be incorporated into measures of disclosure risk.

1.3 Disclosure Limitation Methods for Microdata

1.3.1 Additive Noise Methods

Fuller (1993) This paper considers a variety of masking methods in which error is added to data elements prior to release. These fall generally within the class of measurement error methods. The author stresses that to obtain consistent estimates of higher-order moments of the masked data and functions of these moments such as regression coefficients, measurement error methods and specialized software are required. Other techniques, such as data switching and imputation, can produce biased estimates of some sample covariances and other higher-order moments. The approach is related to that of Kim and Winkler (1997), but applicable to data which is not necessarily multivariate normal.

Kim and Winkler (1997) This paper presents a two-stage disclosure limitation strategy, applied to matched CPS-IRS data. The disclosure concern in this data arises from the match: the CPS data are already masked, but the IRS tax data is not. The IRS data need to be sufficiently well-masked so they cannot easily be used in re-identifications, either alone or in conjunction with unmasked key variables from the CPS. The procedure is as follows.

The data in question are known to be approximately multivariate normal. Hence, in the first stage noise from a multivariate normal distribution with mean zero and the same correlation structure as the unmasked data is added to the IRS income variables. As discussed in Little (1993) and Fuller (1993), such an approach is currently the only method that preserves correlations. Following the addition of noise to the data, the authors determine the re-identification risk associated with the data by matching the raw linked data to the masked file. In cases where the re-identification risk was deemed too great, the authors randomly swap quantitative data within collapsed (age $\times$ race $\times$ sex) cells. This approach preserves means and correlations in the subdomains on which the swap was done, and in unions of these subdomains. However, the swapping algorithm may severely distort means and correlations on arbitrary subdomains. Finally, the authors assess both the confidentiality protection offered by their method and the analytic usefulness of the resulting files, and conclude that both are good.

Moore (1996a) This paper provides a critical examination of the degree of confidentiality protection and analytic usefulness provided by the Kim and Winkler (1997) method. The author concludes that the method is both useful and highly feasible. The author also considers some particular aspects of the algorithm, such as optimal parameter values which generate "sufficient" masking with minimal distortion to second moments. Finally, the author considers how much masking is "sufficient," given reasonable assumptions on intruder knowledge, tools, and objectives.

Winkler (1998) This paper compares the effectiveness of a number of competing disclosure limitation methodologies to preserve both confidentiality and analytic usefulness. The methods considered include the additive-noise and swapping techniques of Kim and Winkler (1997), the additive-noise approach of Fuller (1993), and μ-ARGUS suppression as described in Hundepool and Willenborg (1999) and Nordholt (1999) in Section 1.3.3. The author arrives at several conclusions. First, the Fuller (1993) additive-noise method may not provide as much protection as that author had originally suggested. In particular, sophisticated matching techniques may allow for a significantly higher re-identification rate than previously thought. Second, a naive application of μ-ARGUS to the linked CPS-IRS data described in Kim and Winkler (1997) did little

to preserve either confidentiality or analytic usefulness. More sophisticated methods, including a variant of the Kim and Winkler (1997) method that included a μ-ARGUS pass on the masked data, were much more successful. The authors conclude that additive-noise methods can produce masked files that allow some analyses to approximately reproduce the results obtained with unmasked data. When additional masking procedures are applied such as limited swapping or probability adjustment (Fuller (1993)), then disclosure risk is significantly reduced, though analytic properties are somewhat compromised.

Duncan and Mukherjee (1998) This paper derives an optimal disclosure limitation strategy for statistical databases – *i.e.*, micro-databases which respond to queries with aggregate statistics. As in all disclosure limitation problems, the aim is to maximize legitimate data access while keeping disclosure risk below an acceptable level. The particular confidentiality breach considered is called a *tracker attack*: a well known intruder method in databases with query set size (QSR) control. QSR control is a query restriction technique where a query is disallowed if the number of records satisfying the query is too small (or too large, by inference from the complementary query). A tracker attack is a finite sequence of legitimate queries that yields the same information as a query precluded under QSR. The authors show that the optimal method for thwarting tracker attacks is a combination of query restriction and data masking based on additive noise. The authors also derive conditions under which autocorrelated noise is preferable to independent noise or "permanent" data perturbation.

Evans et al. (1998) This paper presents an additive-noise method for disclosure limitation which is appropriate to establishment tabular data. The authors propose adding noise to the underlying microdata prior to tabulation. Under their approach, "more sensitive" cells receive more noise than less sensitive cells. There is no attempt to preserve marginal totals. This proposal has numerous advantages over the cell-suppression approach which is usually applied to such data. In particular, it is far simpler and less time-consuming than cell-suppression techniques. It also eliminates the need to coordinate cell suppressions between tables, and eliminates the need for secondary suppressions, which can seriously reduce the amount of information in tabular releases. The authors also contend that an additive noise approach may offer more protection than cell-suppression, although suppression may give the appearance of offering more protection.

Pursey (1999) This paper discusses the disclosure control methods developed and implemented by Statistics Canada to release a Public Use Microdata File (PUMF) of financial data from small businesses. This is a fairly unique enterprise – in most cases, statistical agencies deem it too difficult to release public use microdata on businesses that preserve confidentiality. The paper discusses the five steps taken to create the PUMF: (1) make assumptions about an intruder's motivation, information, and tools; (2) make disclosure control goals based on these assumptions; (3) translate these goals into mathematical rules; (4) implement these rules to create the PUMF; and (4) measure the data quality of the PUMF. These are discussed briefly below.

It is assumed that an intruder seeks to identify any record in the PUMF, and has access to the population data file from which the PUMF records are drawn. It is assumed that identification is achieved via nearest-neighbor matching to the population file. Given these assumptions, the following disclosure control goals were set:

- Ensure a low probability that a business from the population appears in the PUMF (less than $r\%$), and that an intruder cannot determine that a particular business is in the PUMF.
- Ensure that each continuous variable is perturbed and that an intruder cannot undo the perturbation.
- Ensure a low probability that a PUMF record can be correctly linked to itself in the population file (less than $p\%$), and that an intruder cannot determine whether a link has been correctly or incorrectly made.

- Remove unique records.

Continuous variables were perturbed according to methods similar to those of Kim and Winkler (1997). First, independent random noise was added to each datum, subject to the constraints that the minimum and maximum proportion of random noise is constant for each datum, and that within a record the perturbations are either always positive or always negative. Next, the three highest data values of each variable in each cell were replaced with their average. Finally, all data values were rounded to the nearest \$1000. Since a less than $p\%$ linkage rate was deemed necessary, in industry cells with a correct linkage rate greater than $p\%$, the data was further perturbed by data swapping with the second-nearest neighbor until a $p\%$ linkage rate was achieved.

After implementing the above disclosure control methods, the resulting data quality was analyzed. The general measure used was one of relative distance: $Rd = (x_a - x_b)/(x_a + x_b)$, where x_a is data or a sample statistic after disclosure control, and x_b is the same data or sample statistic before disclosure control. All variables in the PUMF and a variety of sample statistics were analyzed according to this distance measure. The results indicated that the resulting data quality was good to fair for unincorporated businesses, fair to poor for incorporated businesses.

1.3.2 Multiple Imputation and Related Methods

Rubin (1993) Rubin (1993) is the first paper to suggest the use of multiple imputation techniques for disclosure limitation for microdata analyses. His radical suggestion – to release only synthetic data generated from actual data by multiple imputation – is motivated by the forces outlined at the outset of this review. Namely, an increase in the demand for public use microdata, and increasing concern about the confidentiality of such data.

Rubin's (1993) approach has a number of advantages over competing proposals for disclosure limitation, such as microdata masking. For example, valid statistical analyses of masked microdata generally require "not only knowledge of which masking techniques were used, but also special-purpose statistical software tuned to those masking techniques" (Rubin (1993, p. 461)). In contrast, analysis of multiply-imputed synthetic data can be validly undertaken using standard statistical software simply by using repeated applications of complete-data methods. Furthermore, an estimate of the degree to which the disclosure proofing techniques influence estimated model parameters can be obtained from between-imputation variability. Finally, since the released data is synthetic, *i.e.*, contains no data on actual units, it poses no disclosure risk.

The details of Rubin's (1993) proposal are as follows. Consider an actual microdata sample of size n drawn using design D from a much larger population of N units. Let X represent background variables (observed, in principle, for all N units), Z represent outcome variables with no confidentiality concerns, and Y represent outcome variables with some confidentiality concerns. Note that Z and Y are only observed for the n sampled units, and missing for the $N - n$ unsampled units. A multiply-imputed population consists of the actual X data for all N units, the actual $\left[\begin{array}{cc} Z & Y \end{array}\right]$ data for the n units in the sample, and M matrices of $\left[\begin{array}{cc} Z & Y \end{array}\right]$ data for the $N-n$ unsampled units, where M is the number of multiple imputations. The multiply-imputed values of $\left[\begin{array}{cc} Z & Y \end{array}\right]$ are obtained from some model with predictors X. Given such a multiply-imputed population and a new survey design D^* for the microdata to be released (possibly the same as D), the statistical agency can draw a sample of $n^* \ll N$ units from the multiply-imputed population which is structurally like an actual microdata sample of size n^* drawn from the actual population using design D^*. This can be done M times to create M replicates of the $\left[\begin{array}{cc} Z & Y \end{array}\right]$ values. To ensure that no actual data are released, the statistical agency could draw the samples from the multiply-imputed population excluding the n actual units.

Rubin (1993) recognizes the information loss inherent in the multiple-imputation technique. However, some aspect of this information loss are subtle, and he presents these as the following two facts. First,

although the actual $\left[\begin{array}{cc} Z & Y \end{array}\right]$ and the population values of X contain more information than the multiply-imputed population, if the imputation model is correct, then as M increases, the information in the latter is essentially the same as in the former. Second, the information in the original microdata sample of size n may be greater than, less than, or equal to the information in the multiply-imputed sample of size n^*; the relationship will depend on the estimand under investigation, the relative sizes of n and n^*, the magnitude of M, the designs D and D^*, and the ability of X to predict $\left[\begin{array}{cc} Z & Y \end{array}\right]$.

Fienberg (1994) Fienberg (1994) proposes a method of confidentiality protection in the spirit of Rubin (1993). Whereas Rubin (1993) suggests generating synthetic microdata sets by multiple imputation, Fienberg (1994) suggests generating synthetic microdata by bootstrap methods. This method retains many of the desirable properties of Rubin's (1993) proposal – namely disclosure risk is reduced because only synthetic data are released, and the resultant microdata can be analyzed using standard statistical methods.

To discuss the details of his proposal, let us restate the statistical agency's problem. As before, suppose the agency collects data on a random sample of size n from a population of size N (ignore aspects of the sample design). Let F be the true p-dimensional c.d.f. of the data in the population, and let $\bar{F}$ be the empirical c.d.f. based on the sample of size n. The disclosure problem arises because researchers request, in essence, access to the full empirical p-dimensional c.d.f., $\bar{F}$. Because of guarantees of confidentiality, the agency believes it cannot release $\bar{F}$ since an intruder may be able to identify one or more individuals in the data.

Fienberg's (1994) proposal is as follows. Suppose the statistical agency has a "smoothed" estimate of the c.d.f., $\hat{F}$, derived from the original sample c.d.f. $\bar{F}$. Rather than releasing either $\bar{F}$ or $\hat{F}$, the agency could sample from $\hat{F}$ and generate a synthetic bootstrap-like sample of size n. Denote the empirical c.d.f. of the synthetic microdata file as $\bar{G}$. Fienberg (1994) notes some technical details surrounding $\bar{G}$ which have yet to be addressed. Namely, under what conditions would replicates of $\bar{G}$, say $\bar{G}_i$ for $i = 1, ..., B$, be such that as $B \rightarrow \infty$, $\frac{1}{B}\sum_{i=1}^{B} \bar{G}_i \rightarrow \hat{F}$? Is a single replicate sufficient, or would multiple replicates be required for valid analyses, or possibly the average of multiple replicates? Bootstrap theory may provide some insight into these issues.

Fienberg et al. (1998) The authors reiterate Feinberg's (1994) proposal for generating synthetic data via bootstrap methods, and present a related application to the case of categorical data. Categorical data can be represented by a contingency table, for which there is a direct relationship between a specific hierarchical loglinear model and a set of marginal tables that represent the minimal sufficient statistics of the model. The authors present an example of a three-way table, for which they obtain maximum likelihood estimates of the expected cell values under a loglinear model. The suggestion is to release the MLEs as a public use product, rather than the actual data. They then generate 1,000,000 tables with the same two-way margins, and perform a goodness-of-fit test based on the MLEs. They find that the sparseness of the table in their example presents some problems for accurate loglinear modeling.

In a comment to this article, Kooiman (1998) expresses doubt as to the feasibility of the Fienberg (1994) and Fienberg et al. (1998) proposal for generating synthetic data. He makes a connection between the proposed method and a data-swapping exercise subject to fixed margins. Kooiman (1998) shows that for large data sets with many categorical variables and many categories, such an exercise is likely impossible. He also finds the relationship between the synthetic data proposal and the categorical data example tenuous, at best.

Kennickell (1991), Kennickell (1997), Kennickell (1998), Kennickell (2000) In a series of articles, Kennickell (1991), Kennickell (1997), Kennickell (1998), Kennickell (2000), describes the Federal Reserve Imputation Technique Zeta (FRITZ), used for both missing value imputation and disclosure limitation in the Survey of Consumer Finances (SCF). The SCF is a triennial survey administered by the Federal Reserve

Board to collect detailed information on all household assets and liabilities. Because holdings of many types of assets are highly concentrated in a relatively small fraction of the population, the SCF heavily oversamples wealthy households. Since such households are likely to be well-known, at least in their localities, the data collection process presents a considerable disclosure risk. As a first step towards implementing the proposal of Rubin (1993), the SCF simulates data for a subset of sample cases, using the FRITZ multiple imputation algorithm. This approach is highly relevant for our current research, and hence we discuss it in some detail here.

Using The FRITZ Algorithm for Missing Data Imputation As mentioned above, the FRITZ algorithm is used both for missing value imputation and disclosure limitation in the SCF. The algorithm is most easily understood in the context of missing data imputation. We return to the issue of its application to disclosure limitation below.

The FRITZ model is sequential in the sense that it follows a predetermined path through the survey variables, imputing missing values one (occasionally two) at a time. The model is also iterative in that it proceeds by filling in all missing values in the survey data set, using that information as a basis for imputing the following round, and continuing the process until key estimates are stable. Five imputations are made for every missing value, hence the method is in the spirit of Rubin's (1993) proposal. The following describes the FRITZ technique for imputing missing continuous variables.

For convenience, suppose the iterative process has completed $\ell-1$ rounds, and we are currently somewhere in round ℓ, with a data structure as given below: (reproduced from Kennickell (1998))

$$\left[\begin{array}{cccc} & \text{Iteration } \ell-1 & & \\ y_1 & \chi_{11}^{\ell-1} & x_{12} & x_{13} \\ \Psi_2^{\ell-1} & \chi_{21}^{\ell-1} & x_{22} & \chi_{23}^{\ell-1} \\ y_3 & x_{31} & x_{32} & x_{33} \\ \ldots & & & \\ y_{n-2} & x_{n-2,1} & x_{n-2,2} & x_{n-2,3} \\ y_{n-1} & x_{n-1,1} & \chi_{n-1,2}^{\ell-1} & x_{n-1,3} \\ \Psi_n^{\ell-1} & \chi_{n1}^{\ell-1} & \chi_{n2}^{\ell-1} & x_{n3} \end{array}\right] \quad \left[\begin{array}{cccc} & \text{Iteration } \ell & & \\ y_1 & \cdot & x_{12} & x_{13} \\ \cdot & r_{21} & x_{22} & \cdot \\ y_3 & x_{31} & x_{32} & x_{33} \\ \ldots & & & \\ r_{n-2} & x_{n-2,1} & x_{n-2,2} & x_{n-2,3} \\ y_{n-1} & x_{n-1,1} & \chi_{n-1,2}^{\ell} & x_{n-1,3} \\ \cdot & r_{n1} & \chi_{n2}^{\ell} & x_{n3} \end{array}\right]$$

Here, y indicates complete (non-missing) reports for the variable currently the subject of imputation; Ψ^p represents round p imputations of missing values of y; x represents complete reports of the of the set of variables available to condition the imputation; χ^p represents completed imputations of x from iteration p. Variables that were originally reported as a range but are not currently imputed are represented by r, and $\cdot$ represents values that are completely missing that are not yet imputed. Every x variable becomes a y variable at its place in the sequence of imputations within each iteration. Note that no missing values remain in the stylized $\ell - 1$ data set.

Ideally, one would like to condition every imputation on as many variables as possible, as well as on interactions and higher powers of those terms. Of course there are always practical limits to such a strategy due to degrees of freedom constraints, and some judgement must be applied in selecting a "maximal" set of conditioning variables, X. Of that maximal set, not every element may be non-missing at a given stage of imputation. For each variable to be imputed, the FRITZ algorithm determines the set of non-missing variables among the maximal set of of conditioning variables for each observation, denoted $X_{(i)}$ for observation i. Given the set of available conditioning variables $X_{(i)}$, the model essentially regresses the target imputation variable on the subset of conditioning variables using values *from the previous iteration of the model*. This process is made more efficient by estimating a maximal normalized cross-product matrix for each variable to be imputed, denoted $\sum (X, Y)_{\ell-1}$, and then subsetting the rows and columns corresponding to the non-missing conditioning variables for a given observation, denoted $\sum \left(X_{(i)}, Y\right)_{\ell-1}$. The imputation

for observation i in iteration ℓ is thus given by:

$$\Psi_{i\ell} = \beta_{(i)\ell} X_{(i)i\ell} + e_{i\ell} \tag{7}$$

where $X_{(i)i\ell}$ is the rows of $X_{(i)\ell}$ corresponding to i; $X_{(i)\ell}$ is the subset of X that is available for i in iteration ℓ; $\beta_{(i)\ell} = \sum \left(X_{(i)} X_{(i)}\right)^{-1}_{\ell-1} \sum \left(X_{(i)} Y\right)_{\ell-1}$, and $e_{i\ell}$ is a random error term. Once a value is imputed, its imputed value is used (along with reported values) in conditioning later imputations.

The choice of error term $e_{i\ell}$ has been the subject of several experiments (see Kennickell (1998)). In early releases of the SCF, $e_{i\ell}$ was taken to be a draw from a truncated normal distribution. The draw was restricted to the central 95 percent of the distribution, with occasional supplementary constraints imposed by the structure of the data or respondent-provided ranges for the variable under imputation. More recently, $e_{i\ell}$ has been drawn from an empirical distribution.

The FRITZ algorithm for imputing multinomial and binary variables works similarly, with an appropriate "regression" substituted for (7) .

Using The FRITZ Algorithm for Disclosure Limitation The FRITZ algorithm is applied to the confidentiality protection problem in a straightforward manner. In the 1995 SCF, all dollar values for selected cases were simulated. The procedure is as follows. First, a set of cases which present excessive disclosure risk are selected (see Kennickell (1997)). These are selected on the basis of having unusual levels of wealth or income given other characteristics, or other unusual combinations of responses. Second, a random set of cases is selected to reduce the ability of an intruder to determine even the set of cases determined to present an excessive disclosure risk. Then, a new data set is created for all the selected cases, and shadow variables (which detail the "type" of response given for a particular case-variable pair, e.g., a complete report, a range report, or non-response) are set so that the FRITZ model interprets the responses as range responses. The type of range mimics one where the respondent volunteered a dollar range – a dollar amount of $\pm p$ percent (where p is and undisclosed number between 10 and 20 percent) is stored in a data set normally used to contain unique range reports. Finally, the actual dollar values are set to missing, and the FRITZ algorithm is applied to the selected cases, using the simulated range reports to constrain the imputed values. Subsequent evaluation of the 1995 SCF (Fries et al. (1997)) indicates that while the imputations substantially masked individual cases, the effect on important distributional characteristics was minimal.

1.3.3 Other Methods

Moore (1996b) This paper presents a brief overview of data-swapping techniques for disclosure limitation, and presents a more sophisticated technique than found elsewhere in the literature. The author presents an algorithm for a controlled data swap based on the rank-based proximity swap of Greenberg (1987). The contribution in this paper is to provide a technique which preserves univariate and bivariate relationships in the data. Based on a simulation using the 1993 Annual Housing Survey Public Use File, the author concludes that the algorithm preserves the desired moments to an acceptable degree (and hence retains some degree of analytic usefulness), while providing a level of confidentiality protection comparable to simple additive-noise methods.

Moore (1996c) This paper suggests modifications to the Confidentiality Edit, the data-swapping procedure used for disclosure limitation in the 1990 Decennial Census. The suggested improvements are based on the ARGUS system for determining high-risk cases (see Hundepool and Willenborg (1999) and Nordholt (1999) below), and the German SAFE system for perturbing data. The author also presents two measures of the degree of distortion induced by the swap, and an algorithm to minimize this distortion.

Mayda et al. (1997) This paper examines the relationship between variance estimation and confidentiality protection in surveys with complex designs. In particular, the authors consider the case of the Canadian National Population Health Survey (NPHS), a longitudinal survey with a multi-stage clustered design. To prepare a public use file, it was deemed necessary to remove specific design information such as stratum and cluster identifiers due to the extremely detailed level of geography they represented. Furthermore, providing cluster information could allow users to reconstitute households, increasing the probability of identifying individuals. However, specific design information is necessary to correctly compute variances using jackknife or other methods. This highlights yet another aspect of the conflict between providing high quality data and protecting confidentiality. The authors describe the approach taken to resolve this conflict. Specifically, strata and clusters are collapsed to form "super-strata" and "super-clusters" in the public use file, which protect confidentiality while providing enough information for researchers to obtain unbiased variance estimates under certain conditions. The drawback of this approach is that it does not generate the exact variance corresponding to the original design, and that collapsing reduces degrees of freedom and hence the precision of variance estimates.

Nadeau et al. (1999) This paper presents a discussion of confidentiality issues surrounding Statistics Canada's Survey of Labour and Income Dynamics (SLID), and presents the release strategy for microdata on individual and family income. SLID is a longitudinal survey designed to support studies of economic well-being of individuals and families, and of their determinants over time. With the demise of the Canadian Survey of Consumer Finances (SCF) in 1998, SLID became the official source of information for both longitudinal *and* cross-sectional income data on individuals and families. This presented some rather unique issues for disclosure limitation.

Prior to integrating SLID and SCF, Statistics Canada did not release sufficient information in the SLID Public Use Microdata Files (PUMFs) to allow household reconstitution. It was considered too difficult to protect confidentiality at the household level in a longitudinal microdata file. However, since integrating SLID and SCF, it has become a priority to release cross-sectional PUMFs that meet needs of former SCF users. In particular, the cross-sectional PUMFs now contain household and family identifiers, which allow household and family reconstitution. This compromises the release of longitudinal PUMFs. Instead, Statistics Canada has opted to explore other options for the release of longitudinal data – namely release of synthetic files, and creation of research data centres. In the meantime, a number of disclosure limitation methods have been explored for the cross-sectional PUMFs to limit the ability of intruders to link records dynamically (constructing their own longitudinal file, considered "too risky" for re-identification) and/or re-identify records by linking to the Income Tax Data File (ITDF).

The disclosure control methods applied in the cross-sectional PUMFs include both data reduction and data modification methods. The data reduction methods include dropping direct identifiers, aggregating geographic variables, and categorical grouping for some occupational variables. Data modification methods are applied to numeric variables. In particular, year of birth is perturbed with additive noise; income variables are both bottom- and top-coded, and the remaining values are perturbed with a combined random-rounding and additive noise method.

Finally, the authors assess how successful these measures are at protecting confidentiality and maintaining analytical usefulness. To address the former, they consider both linking consecutive cross-sectional PUMFs and linking to the ITDF. In both cases, they consider both direct matches and nearest-neighbor matches. They find that the ability of an intruder to match records in either consecutive PUMFs or to the ITDF is severely limited by the disclosure control measures. As for the usefulness of the data, they find little difference in the marginal distribution of most variables at highly aggregated levels (*i.e.*, the national level), but more significant differences at lower levels of aggregation (i.e., the province×sex level).

Hundepool and Willenborg (1999), Nordholt (1999) These papers describe the τ-ARGUS and μ-ARGUS software packages developed by Statistics Netherlands for disclosure limitation. Nordholt (1999) describes their specific application to the Annual Survey on Employment and Earnings (ASEE). The τ-ARGUS software tackles the problem of disclosure limitation in tabular data. It automatically applies a series of primary and secondary suppressions to tabular data on the basis of a dominance rule: a cell is considered unsafe if the n major contributors to that cell are responsible for at least p percent of the total cell value. The μ-ARGUS software is used to create a public use microdata file from the ASEE. The public use microdata has to satisfy two criteria, which are implemented with μ-ARGUS: first, every category of an identifying variable must occur "frequently enough" (200,000 times is the default for ASEE); second, every bivariate combination of values must occur "frequently enough" (1,000 times is the default for ASEE). These objectives are achieved via global recoding and local suppression.

1.4 Analysis of Disclosure-Proofed Data

Little (1993) Little (1993) develops a model-based likelihood theory for the analysis of masked data. His approach is to formally model the mechanism whereby case-variable pairs are selected for masking, the masking method, and derive an appropriate model for analysis of the resulting data. His method is sufficiently general to allow for a variety of masking selection mechanisms, and such diverse masking methods as deletion, coarsening, imputation, and aggregation. The formal theory follows.

Let $\mathsf{X} = \{x_{ij}\}$ denote an $(n \times p)$ unmasked data matrix of n observations on p variables. Let $\mathsf{M} = \{m_{ij}\}$ denote the masking indicator matrix, where $m_{ij} = 1$ if x_{ij} is masked, and $m_{ij} = 0$ otherwise. Let $\mathsf{Z} = \{z_{ij}\}$ denote the masked data, i.e., z_{ij} is the masked value of x_{ij} if $m_{ij} = 1$, and $z_{ij} = x_{ij}$ if $m_{ij} = 0$. Model the joint distribution of X, Z, and M with the density function:

$$f\left(\mathsf{X}, \mathsf{Z}, \mathsf{M}|\boldsymbol{\theta}\right) = f_X\left(\mathsf{X}|\boldsymbol{\theta}\right) f_Z\left(\mathsf{Z}|\mathsf{X}\right) f_M\left(\mathsf{M}|\mathsf{X}, \mathsf{Z}\right). \tag{8}$$

Here $f_X\left(\mathsf{X}|\boldsymbol{\theta}\right)$ is the density of the unmasked data given unknown parameters $\boldsymbol{\theta}$, which would be the basis for analysis in the absence of masking; $f_Z\left(\mathsf{Z}|\mathsf{X}\right)$ formalizes the masking treatment; and $f_M\left(\mathsf{M}|\mathsf{X}, \mathsf{Z}\right)$ formalizes the masking selection mechanism. If the analyst knows which values are masked and the method masking, then the analyst knows M, as well as the distributions of M and Z. If not, then M is unknown. A more general specification would also index the distributions of M and/or Z by unknown parameters, and a full likelihood analysis would then involve both $\boldsymbol{\theta}$ and these unknown masking parameters.

Let $\mathsf{X} = (\mathsf{X}_{obs}, \mathsf{X}_{mis})$ and $\mathsf{Z} = (\mathsf{Z}_{obs}, \mathsf{Z}_{mis})$ where obs denotes observed components, and mis denotes missing components of each matrix. Analysis of the masked data is based on the likelihood for $\boldsymbol{\theta}$ given the data $\mathsf{M}, \mathsf{X}_{obs}$, and Z_{obs}. This is obtained formally by integrating the joint density in (8) over the missing values X_{mis} and Z_{mis} :

$$L\left(\boldsymbol{\theta}|\,\mathsf{M}, \mathsf{X}_{obs}, \mathsf{Z}_{obs}\right) = \int f_X\left(\mathsf{X}|\boldsymbol{\theta}\right) f_Z\left(\mathsf{Z}|\mathsf{X}\right) f_M\left(\mathsf{M}|\mathsf{X}, \mathsf{Z}\right) d\mathsf{X}_{mis} d\mathsf{Z}_{mis}. \tag{9}$$

Since the distribution of M in (9) may depend on X and Z_{obs}, but should not depend on Z_{mis}, we can write $f_M\left(\mathsf{M}|\mathsf{X}, \mathsf{Z}\right) = f_M\left(\mathsf{M}|\mathsf{X}, \mathsf{Z}_{obs}\right)$. Thus we can rewrite (9) as:

$$L\left(\boldsymbol{\theta}|\,\mathsf{M}, \mathsf{X}_{obs}, \mathsf{Z}_{obs}\right) = \int f_X\left(\mathsf{X}|\boldsymbol{\theta}\right) f_Z^*\left(\mathsf{Z}_{obs}|\mathsf{X}\right) f_M\left(\mathsf{M}|\mathsf{X}, \mathsf{Z}_{obs}\right) d\mathsf{X}_{mis} \tag{10}$$

where $f_Z^*\left(\mathsf{Z}_{obs}|\mathsf{X}\right) = \int f_Z\left(\mathsf{Z}|\mathsf{X}\right) d\mathsf{Z}_{mis}$.

The author notes that the likelihood in (10) can be simplified if the masking selection and treatment mechanisms satisfy certain ignorability conditions, in the sense of Rubin (1976) and Rubin (1978). Specifically, if the masking selection mechanism is ignorable, then $f_M\left(\mathsf{M}|\mathsf{X}, \mathsf{Z}\right) = f_M\left(\mathsf{M}|\mathsf{X}_{obs}, \mathsf{Z}_{obs}\right)$ for all X_{mis}, Z_{mis}. In this case, the density of M can be omitted from (10). Similarly, the masking treatment mechanism

is ignorable if $f_Z^*(\mathsf{Z}_{obs}|\mathsf{X}) = f_Z^*(\mathsf{Z}_{obs}|\mathsf{X}_{obs})$ for all X_{mis}. In this case, the density of Z_{obs} can be omitted from (10). Finally, if both mechanisms are ignorable, then the likelihood reduces to:

$$L(\boldsymbol{\theta}|\mathsf{M}, \mathsf{X}_{obs}, \mathsf{Z}_{obs}) = \int f_X(\mathsf{X}|\boldsymbol{\theta})\, d\mathsf{X}_{mis}$$

which is proportional to the marginal density of X_{obs}.

References

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

Boudreau, J.-R. (1995, November). Assessment and reduction of disclosure risk in microdata files containing discrete data. Presented at Statistics Canada Symposium 95.

Duncan, G. T. and S. Mukherjee (1998, June). Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. Heinz School of Public Policy and Management Working Paper No. 1998-15.

Evans, B. T., R. Moore, and L. Zayatz (1996). New directions in disclosure limitation at the Census Bureau. U.S. Census Bureau Research Report No. LVZ96/01.

Evans, T., L. Zayatz, and J. Slanta (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics 14*(4), 537–551.

Fienberg, S. E. (1994, December). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Carnegie Mellon University Department of Statistics Technical Report No. 611.

Fienberg, S. E. (1997, September). Confidentiality and disclosure limitation methodology: Challenges for national statistics and statistical research. Presented at the Committee on National Statistics 25th Anniversary Meeting. Carnegie Mellon Department of Statistics Technical Report, Working Paper No. 668.

Fienberg, S. E. and U. E. Makov (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics 14*(4), 385–397.

Fienberg, S. E., U. E. Makov, and R. J. Steele (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics 14*(4), 485–502.

Franconi, L. (1999, March). Level of safety in microdata: Comparisons between different definitions of disclosure risk and estimation models. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 4.

Fries, G., B. W. Johnson, and R. L. Woodburn (1997, September). Analyzing disclosure review procedures for the Survey of Consumer Finances. SCF Working Paper, presented at the 1997 Joint Statistical Meetings, Anaheim, CA.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics 9*(2), 383–406.

Greenberg, B. (1987). Rank swapping for masking ordinal microdata. U.S. Census Bureau, unpublished manuscript.

Hundepool, A. and L. Willenborg (1999, March). ARGUS: Software from the SDC project. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 7.

Jabine, T. B. (1993). Statistical disclosure limitation practices of the United States statistical agencies. *Journal of Official Statistics 9*(2), 427–454.

Kennickell, A. B. (1991, October). Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. SCF Working Paper, prepared for the Annual Meetings of the Amercian Statistical Association, Atlanta, Georgia, August 1991.

Kennickell, A. B. (1997, November). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. SCF Working Paper.

Kennickell, A. B. (1998, September). Multiple imputation in the Survey of Consumer Finances. SCF Working Paper, prepared for the August 1998 Joint Statistical Meetings, Dallas, TX.

Kennickell, A. B. (2000, May). Wealth measurement in the Survey of Consumer Finances: Methodology and directions for future research. SCF Working Paper, prepared for the May 2000 annual meetings of the American Association for Public Opinion Research, Portland, Oregon.

Kim, J. J. and W. E. Winkler (1997). Masking microdata files. U.S. Census Bureau Research Report No. RR97/03.

Kooiman, P. (1998). Comment on disclosure limitation for categorical data. *Journal of Official Statistics 14*(4), 503–508.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics 9*(2), 313–331.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*(2), 407–426.

Little, R. J. and D. B. Rubin (1987). *Statistical Analysis of Missing Data*. New York: Wiley.

Mayda, J., C. Mohl, and J. Tambay (1997). Variance estimation and confidentiality: They are related! Unpublished Manuscript, Statistics Canada.

Moore, Jr, R. A. (1996a). Analysis of the Kim-Winkler algorithm for masking microdata files - how much masking is necessary and sufficient? Conjectures for the development of a controllable algorithm. U.S. Census Bureau Research Report No. RR96/05.

Moore, Jr, R. A. (1996b). Controlled data-swapping techniques for masking public use microdata sets. U.S. Census Bureau Research Report No. RR96/04.

Moore, Jr, R. A. (1996c). Preliminary recommendations for disclosure limitation for the 2000 Census: Improving the 1990 confidentiality edit procedure. U.S. Census Bureau Statstical Research Report Series, No. RR96/06.

Nadeau, C., E. Gagnon, and M. Latouche (1999). Disclosure control strategy for the release of microdata in the Canadian Survey of Labour and Income Dynamics. Presented at the 1999 Joint Statistical Meetings, Baltimore, MD.

National Research Council (2000). Improving access to and confidentiality of research data: Report of a workshop. Committee on National Statistics, Christopher Mackie and Norman Bradburn, Eds. Commission on Behavioral and Social Sciences and Education, National Academy Press, Washington, D.C.

Nordholt, E. S. (1999, March). Statistical disclosure control of the Statistics Netherlands employment and earnings data. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 2.

Pursey, S. (1999, March). Disclosure control methods in the public release of a microdata file of small businesses. Presented at the Joint ECE/Eurostat Work Session on Statisical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 5.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics 6*, 34–58.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1993). Discussion of statistical disclosure limitation. *Journal of Official Statistics 9*(2), 461–468.

Skinner, C. and D. Holmes (1993). Modelling population uniqueness. In *Proceedings of International Seminar on Statistical Confidentiality*, Luxembourg, pp. 175–199. EUROSTAT.

Subcommittee on Disclosure Avoidance Techniques (1978a). Statistical policy working paper no. 2: Report on statistical disclosure and disclosure avoidance techniques. Federal Committee on Statistical Methodology, Office of Federal Policy and Standards, U.S. Department of Commerce, Washington, D.C.

Subcommittee on Disclosure Avoidance Techniques (1994b). Statistical policy working paper no. 22: Report on statistical disclosure limitation methodology. Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, U.S. Office of Management and Budget, Washington, D.C.

Willenborg, L. and J. Kardaun (1999, March). Fingerprints in microdata sets. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 10.

Winkler, W. E. (1997). Views on the production and use of confidential microdata. U.S. Census Bureau Research Report No. RR97/01.

Winkler, W. E. (1998). Producing public-use files that are analytically valid and confidential. U.S. Census Bureau Research Report No. RR98/02.