Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 485–502

Disclosure Limitation Using Perturbation and Related Methods for Categorical Data

Stephen E. Fienberg[1], *Udi E. Makov*[2], *and Russell J. Steele*[1]

During the past twenty-five years, the field of disclosure protection has undergone a "statistical transformation" and has begun to utilize the advances that have occurred within the field of statistics itself as well as in a variety of areas of application. This article reexamines some of the approaches currently employed in statistical disclosure limitation methodology for categorical data, e.g., cell suppression and data swapping, and relates them to the more conventional statistical methods associated with loglinear models and the simulation of exact distributions. It ties this perturbation approach to a general framework for the use of simulated data which we described earlier in Fienberg (1996) and Fienberg, Steele, and Makov (1996).

Key words: Bootstrap; cell suppression; confidentiality; contingency table analysis; data swapping; loglinear models; multiple imputation.

1. Introduction

Disclosure avoidance methodology has developed over the past 20 years as a major area of government statistics research and activity. The advances are impressive (e.g., see the progress chronicled in Subcommittee on Disclosure-Avoidance Techniques, 1994, especially when compared with the methodology described in Subcommittee on Disclosure-Avoidance Techniques, 1978). But all too often these advances appear to be unlinked to the analytical uses to which most census and survey data are put and to the evolving methods of statistical analysis. During this same 20-year period there have also been major advances in statistical methodology and theory. A theme of this article is that many of these statistical tools that come from these latter developments have relevance to the area of disclosure limitation methodology. For a number of reasons situations involving categorical data in the form of a contingency table offer an excellent venue for such consideration.

In this article we:

- Review some current statistical ideas in use for data disclosure avoidance for categorical variables.

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
[2] Department of Statistics, Haifa University, Haifa, Israel.

Acknowledgments: The preparation of this article was supported in part under a contract with Westat and the U.S. Bureau of the Census and by Statistics Netherlands while Fienberg was a visiting researcher there. We are especially appreciative to David Binder, Persi Diaconis, George Duncan, Gary Glonek, Sallie Keller-McNulty, Peter Kooiman, Steffen Lauritzen, Michael Meyer, Peter Müller, Danny Pfefferman, Agostino Nobile, Leon Willenborg and three referees for providing comments, suggestions or assistance in some form during the course of our work and the revisions of this material. None of them, however, bears any responsibility for our use of the suggested materials, or for our admittedly speculative ideas on the applicability of specific statistical methods. Earlier version of material that forms the core of this article appeared in Fienberg (1996) and Fienberg et al. (1996).

- Present a new statistical framework for data release.
- Relate these ideas and approaches to "traditional" statistical methodology associated with loglinear models for cross-classified categorical data and to the simulation of associated exact distributions.

Before doing so, we outline a framework in which the problem of data-disclosure avoidance methodology can be viewed. We consider four different parties: the *agency* or data collector; the *respondents* or data providers; an *intruder* who wants to learn about one or more data providers via the data to be released by the agency; *users* or secondary analysts of the agency data. The question of interest to us is: What data can the agency release for analysis by the users while protecting the respondents from the intruder (i.e., preserving their confidentiality)? The practical way in which this question has been answered is through the application of some disclosure limitation methodology that the agency hopes achieves the desired goals.

For most data releases, especially those from censuses, the U.S. Bureau of the Census has either released data at high levels of aggregation or applied a data disclosure avoidance procedure such as data swapping or cell suppression before preparing micro-data or tables for release.

Consider a sample of n observations on p variables, which may be discrete or continuous. Our general characterization is in terms of the smoothing of a multi-dimensional empirical distribution function (an ordered version of the data), and sampling from it using bootstrap-like selection. Both the smoothing and the sampling introduce alterations to the data and thus a bootstrap sample will not necessarily be the same as the original sample – this works to preserve the confidentiality of individuals providing the original data. Two obvious questions are: How well is confidentiality preserved by such a process? Have the smoothing and sampling disguised fundamental relationships among the p variables of interest to others who will work only with the altered data? In this article we focus primarily on the second of these questions but we do discuss ways to approach answering the first.

In the next section we review some of the specific methods for disclosure avoidance that have been proposed in the literature, and that fit under the broad rubric of "matrix masking." In particular we describe two specific methods for "matrix masking" when all of the variables are categorical – a special case of cell suppression and data swapping. Then, in Section 3, we explain how we view these methods in the context of the users' analytical goals. In Section 4, we suggest a general strategy for disclosure limitation that attends to the proposed goals in a non-standard fashion, and we relate the strategy to some modern approaches from the statistical methodology literature. In Section 5, we describe in further detail our current efforts at implementing a perturbation method related to this general strategy in the context of contingency table problems. We end with an outline of research that would put the general strategy suggested on a firm theoretical foundation.

There are a number of excellent articles that attempt to bridge the gap between the literature on disclosure avoidance and more general statistical methodology, beginning with the pioneering work of Duncan and Lambert (1986) and (1989), and continuing with contributions to the special 1993 issue of the *Journal of Official Statistics* on confidentiality

and data access. This article builds both directly and indirectly on a number of these earlier efforts.

The general strategy proposed here has appeared in other articles in the past; see e.g., Liew, Choi and Liew (1985), Little (1993), Rubin (1993), and Fienberg (1994b). Heer (1993) has suggested a bootstrap method for contingency tables which is related to but different from our proposals for the use of exact distributions in Section 5. Finally, Kennickell (1997) has recently reported on results of a multiple imputation approach to disclosure limitation. To our knowledge, no previous authors have integrated these ideas with both the full literature on loglinear model methods and that on disclosure limitation.

2. Matrix Masking for Micro-data

Duncan and Pearson (1991) give an excellent description of approaches to the masking of microdata. Suppose that $\mathbf{X}$ is an n by p matrix representing the microdata for n individuals or cases on p variables or attributes. Then matrix masking of the microdata file $\mathbf{X}$ provides the user with the transformed file $\mathbf{Z} = \mathbf{AXB} + \mathbf{C}$ in lieu of $\mathbf{X}$. The matrix $\mathbf{A}$ transforms cases, $\mathbf{B}$ transforms variables, and $\mathbf{C}$ blurs the entries of $\mathbf{AXB}$. Cox (1995) explicitly links several of these methods, especially data swapping, to the matrix masking approach, and Fienberg (1994a, 1997) provides a more detailed discussion of the link between matrix masking and a number of proposed disclosure limitation methodologies. Fuller (1993) and Sullivan (1989) provide an informative presentation of the effect of some specific implementations.

A special case involving the deletion of rows is the method of cell suppression. Suppose we are interested in summarizing a set of data in the form of a cross-classification of counts or nonnegative aggregates. Deleting or suppressing a cell value is equivalent to the deletion of those rows of $\mathbf{X}$ for which the entries in columns corresponding to the cross-classifying variables assume the values that specify the cell in question. Cell suppression is widely used for data on establishments because counts of "1" or "2" may uniquely identify a respondent, or one or two establishments dominate an industry, and thus their "share" comprises a large fraction of a weighted total. For simplicity here we focus on the version of cell suppression that weights the respondents equally and thus acts directly on a table of unweighted counts.

In the case of a simple table of counts, current practice at the U.S. Census Bureau and elsewhere would reduce to the suppression of any cell where $k \geq 3$ or fewer respondents make up that cell's value. Such cells are referred to as *primary suppressions*. Typically an agency using such a rule keeps the value of k as well as the method used for selection of cells confidential.

Because reported cross-classifications usually include the corresponding marginal totals, suppressing a single cell produces multiple masks for the same matrix and, taken together, these masks do not disguise the data – the value of a deleted cell in a two-way array can be retrieved from the other entries in the same row or column combined with the corresponding marginal total. Thus methods for cell suppression in cross-classifications also choose other cell values for suppression; these are often referred to as *complementary suppressions*. Determining "desirable" patterns of complementary suppressions is an active area of research, especially for multi-way cross-classifications (see e.g., Cox (1995)).

Table 1. Original $3 \times 2 \times 2$ table with marginals

n_{111}	n_{121}	n_{1+1}	n_{112}	n_{122}	n_{1+2}
n_{211}	n_{221}	n_{2+1}	n_{212}	n_{222}	n_{2+2}
n_{311}	n_{321}	n_{3+1}	n_{312}	n_{322}	n_{3+2}
n_{+11}	n_{+21}	n_{++1}	n_{+12}	n_{+22}	n_{++2}

It is important to note for the present context that the basic approach in cell suppression is one involving margin preservation, i.e., in the 2-way table the method for suppression preserves both sets of one-dimensional marginal totals, $\{n_{i+}\}$ and $\{n_{+j}\}$, by design. In higher dimensions, cell suppression also preserves marginal totals but possibly those of highest order. The principal problem we have with cell suppression as a method is that it intentionally "distorts" the information in the table by purposely selecting cells to suppress. As a consequence, users can be led into misleading and, in particular, biased inferences on the basis of the cell values that are reported.

In 1978, Dalenius and Reiss (1978) proposed a method for "swapping" observations while preserving marginal totals. According to Dalenius and Reiss's definition of a k-order swap, all k-way marginals are preserved. No higher order marginals are guaranteed to be preserved. They present no algorithm for doing these swaps or finding which ones are available. They do, however, present theorems and statements about the probabilities of there being swaps. We can view data swapping as a special case of matrix masking at least in its simplest forms as noted above.

To understand the idea of data swapping, we consider a 3-way contingency table with entries $\{n_{ijl}\}$ as in Table 1. We want to track what happens when we swap the value in the (1,2,1) cell of layer 1 of Table 1 with the (3,1,2) cell in layer 2. Table 2 shows the result. Note that the 2-dimensional total for the first two variables (adding over layers of Table 2) is unchanged, as is the 1-dimensional total for the 3rd variable.

Thus in moving from the original table to the table with the swapped pair of observations we end up by perturbing the data, in a 2×2 subtable using a "local move" of a pair of observations in a way that preserves the two-way totals, $\{n_{ij+}\}$, and the one-way totals, $\{n_{++k}\}$. Data swapping involves the repeated application of such moves of pairs of randomly selected observations.

The U.S. Census Bureau actually used a variant of data swapping for the release of 1990 Census microdata, swapping a somewhat small percentage of records between "nearby" census blocks (see Griffin, Navarro, and Flores-Baez (1989), Navarro, Flores-Baez, and Thompson (1988), as well as Subcommittee on Disclosure Avoidance Techniques, (1994), and Fienberg, Steele, and Makov (1996)). The results were considered to be a success and essentially the same methodology has been proposed for data releases from the 2000 Census. As used in this context, data swapping also distorts the data to some extent

Table 2. Altered $3 \times 2 \times 2$ table with marginals: (1,2,1) cell swapped with (3,1,2) cell

n_{111}	$n_{121} - 1$	$n_{1+1} - 1$	n_{112}	$n_{122} + 1$	$n_{1+2} + 1$
n_{211}	n_{221}	n_{2+1}	n_{212}	n_{222}	n_{2+2}
$n_{311} + 1$	n_{321}	$n_{3+1} + 1$	$n_{312} - 1$	n_{322}	$n_{3+2} - 1$
$n_{+11} + 1$	$n_{+21} - 1$	n_{++1}	$n_{+12} - 1$	$n_{+22} + 1$	n_{++2}

because the number of swaps is not released and the resulting increase in the variability associated with the perturbations cannot easily be incorporated by the user into analyses without full information on the extent of swapping and the margins that are preserved. But if we view data swapping as a first approximation to the method proposed in Section 5 below, then one can show that it is at least consistent. Thus the distortion is only an increase in variance and not the systematic bias that might result from using data to which cell suppression has been applied.

Both the method of cell suppression and the method of data swapping preserve marginal totals in contingency tables. But this is also a property associated with loglinear model methods. What is interesting is that despite the fact that cell suppression and data swapping have been presented in the same sessions in various forums (see e.g., Cox and Sande (1978) and Dalenius and Reiss (1978) and the discussion of the two articles by Zalkind (1978)), previous authors have failed to note this clear relationship between these methods as well as to methods in the contingency table literature. Reviewers and others have questioned whether the preservation of marginal totals is a statistical necessity. In fact, from a modeling perspective one can argue over the desirability of working with fixed margins, but as a practical matter it is consistent with the practice of many statistical agencies, especially when the margins are matched with those from censal records or a baseline survey, through poststratification and/or raking. For us margin preservation in tables is intimately linked to loglinear models, as we have noted, and working with the "exact" distribution for a given loglinear model given its minimal sufficient statistics, as we do in Section 5, is a convenience that matches the practicalities of current agency practices.

More recently, Gouweleeuw et al. (1998) propose a postrandomization method (PRAM) for data perturbation that fits, at least approximately, into the class of matrix masking approaches. PRAM applies a randomization to selected variables in the dataset but in a form that allows the user to draw proper statistical inferences. In PRAM, the equivalent of the matrix $\mathbf{A}$ is stochastic and, instead of an additive matrix $\mathbf{C}$, there is a sampling error associated with each case whose variance depends on the parameters underlying $\mathbf{A}$. Gouweleeuw et al.'s version PRAM acts on individual or blocks of variables independently, and is applied independently to each variable in a microdata file. Duncan and Fienberg (1998) have proposed a generalization of PRAM to allow for multi-way dependencies that preserve specified marginal totals, in a fashion that is closely linked to the methods described in Section 5.

3. Perspective on Data Release and Disclosure Limitation

Typical users of government statistical data are interested in relationships and causal connections for policy choices. They use statistical models to describe such relationships. Often their view of "error" is akin to including an error component in an analytical model (such as a regression error term ϵ in the equation $\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \epsilon$). Otherwise, the user has limited ways to address the multiplicity of information on uncertainty and error coming from the statistical agency that produces the data.

Most users are interested in analytical models, and especially ones with causal implications. Thus we can think of the users' objectives as involving the linking of response variables,

Y, and explanatory variables, **X**, through a statistical model that attempts to represent some underlying substantive phenomenon. Unfortunately we rarely get to observe or measure **Y** and **X** directly. What is produced through a census or a survey questionnaire is often a related but fallible measure of the quantities of real interest. These we label **Y*** and **X***.

We can take as the user's objective the estimation of a multivariate cumulative distribution function (c.d.f.), of the forms $F_{\mathbf{Y}|\mathbf{X}}$ or $F_{\mathbf{Y}|\mathbf{X},\theta}$ for various values of **X**, or at least characteristics of such a multivariate c.d.f. Here the parameter θ might be a population mean or variance, μ or σ^2, or a parameter(s) in a statistical model such as a regression coefficient, β, probably multidimensional in form. In the ensuing discussion we ignore those sources of measurement error in **X** beyond these forms captured in the agency's own evaluation and data preparation activities.

Estimation of a multivariate c.d.f. is a general statistical problem that includes a number of interesting special cases. For example, suppose that all of the variables in the user's model and in the data set are categorical in nature, as is often the case in censal and survey settings. Then the c.d.f. is essentially equivalent to the table of conditional probabilities (for **Y** given **X**) that correspond to the cross-classification of the variables in contingency table form (cf. Bishop, Fienberg, and Holland (1975)). We refer to this special case again in Section 5 and provide an extended set of references and notes on it.

At the risk of oversimplification, we can characterize the standard approach to data collection, processing and release roughly as follows:

- Collect and "clean up" the raw data. This includes editing, matching and all other preliminary processing.
- Protect the data by applying some form of data disclosure avoidance methodology.
- Release the resulting data either as set of marginal tables for some larger cross-classification, or as micro-data files for the variables related to the ones of user interest (**Y***, **X***).
- Estimate θ directly using a sample-based quantity, $\bar{\theta}$.

In effect, the user then follows the agency's lead and estimates the c.d.f. directly from the released data using the "empirical" c.d.f. (suitably weighted to take into account the impact of the survey design), $\bar{F}_{\mathbf{Y}*|\mathbf{X}*}$, or possibly a more elaborate and smoother parametric estimate based on the estimated parameter, i.e. $\bar{F}_{\mathbf{Y}*|\mathbf{X}*,\bar{\theta}}$.

While this approach might make considerable sense for some descriptive statistical problems, the fact is that $\bar{F}_{\mathbf{Y}*|\mathbf{X}*}$ and $\bar{F}_{\mathbf{Y}*|\mathbf{X}*,\bar{\theta}}$ rarely reflect fully aspects of sampling error such as clustering, which many believe to be important, and they almost never reflect the other sources of error listed above that typically dwarf sampling error. Further, given the current state of the art of statistical disclosure limitation methodology, the user may still be able to "identify" individuals in the released data. One way to overcome these shortcomings is to continue to address the various components of error and to separately improve the approach to data disclosure avoidance. Alternatively, we can attempt to reconceptualize the data reporting problem in a new and integrated fashion.

4. Alternative Strategy and Framework

Here we propose an alternative approach to the release of survey data that we described earlier in Fienberg et al. (1996). We begin with the goals of the users and ask how agencies

should organize the data of interest in order to provide data that fit with the user goals. Our approach is cast in terms of the release of a public-use micro-data file that is intended to support analyses for the conditional distribution of $\mathbf{Y}^*$ given $\mathbf{X}^*$. The first step in our prescription is:

1. Combine the census or survey data that the agency would normally have chosen to release, in the form $\bar{F}_{\mathbf{Y}^*|\mathbf{X}^*}$ and $\bar{F}_{\mathbf{Y}^*|\mathbf{X}^*,\hat{\theta}}$, with formal statistical information on error, e.g., form editing, matching, nonresponse, etc. and apply some form of parametric or semi-parametric technique to estimate $F_{\mathbf{Y}|\mathbf{X}}$ and $F_{\mathbf{Y}|\mathbf{X},\theta}$ using all available data by $\hat{F}_{\mathbf{Y}|\mathbf{X}}$ and $\hat{F}_{\mathbf{Y}|\mathbf{X},\hat{\theta}}$ respectively, where $\hat{\theta}$ is a new estimate of θ cast in terms of the distribution of the variables of actual user interest, $\mathbf{Y}$ given $\mathbf{X}$.

For non-parametric estimation of $F_{\mathbf{Y}|\mathbf{X}}$ we can either think in terms of a classical statistical approach using some type of kernel density estimator or a related type of "smooth" estimate, or in terms of a Bayesian approach based on the mixture of Dirichlet processes (see e.g., West, Müller, and Escobar (1994)) or the use of Polya trees (Lavine (1992)). These tools, however, have been used primarily in low-dimensional problems and thus there needs to be additional research to study their adaptation to the high-dimensional censal and survey problems which are the focus of this article. Even if these methods are not especially efficient for statistical estimation purposes, they may serve the needs of data disclosure avoidance which are crucial to the strategy outlined here.

In what ways does this new smoothed estimate of $F_{\mathbf{Y}|\mathbf{X}}$ differ from the one that is explicit or implicit in the current approach? We offer three examples. First, consider the release of decennial census data. In both the U.S. and Canada, there has been extensive documentation of the extent of census undercoverage and how the resulting undercount is distributed across groups in the population and across geographical areas. Failure to correct for such undercoverage in the release of data leads to biased estimates of the true quantity of interest, $F_{\mathbf{Y}|\mathbf{X}}$. Second, by smoothing data to reflect regression-like relationships we can typically achieve improved estimates with much lower variances, although at the price of some potential bias. Finally, by incorporating agency information on components of error (which tends to increase variances) into the statistical estimation process, we produce a new smoothed estimator of $F_{\mathbf{Y}|\mathbf{X}}$.

We hasten to add that this smoothing process should not be viewed simply as standard model selection and fitting, for the goals here are different. The smoothing process possibly involves models but should not be carried out in a way that "oversmooths" the data. Thus the results of smoothing should ideally be compatible with competing models for the data which subsequent analysts could produce by working with the smoothed c.d.f.'s.

The next steps in our prescription are:

2. Instead of releasing the c.d.f. estimated in step 1 above, the agency now "samples" from it to create a "pseudo" micro-data file which we label as $\bar{\bar{F}}_{\mathbf{Y}|\mathbf{X}}$ and $\bar{\bar{F}}_{\mathbf{Y}|\mathbf{X},\hat{\theta}}$. We use the overbar to indicate a sample from the smoothed c.d.f.'s in accord with our earlier notation for the empirical c.d.f., which corresponds to a sample, and the hat to indicate that we are sampling from the smoothed or estimated c.d.f.
3. The agency repeats the process of "sampling" and then releases the resulting replicate "pseudo" micro-data files.

The "pseudo" micro-data files created in the approach outlined above have several interesting features.

First, if we think of them as consisting of a set of released records for individuals, then these "individuals" do not necessarily correspond to any of those in the original sample survey. This fact enhances the public notion of the protection of confidentiality of responses even if an intruder might still be able to indirectly make inferences about individuals in the original sample. This point is especially important from the perspective of data disclosure avoidance. Since the individuals in the pseudo micro-data file are not typically those from the original sample, we have at least in part addressed confidentiality concerns. After all, we no longer even appear to be releasing data for any individual from the original sample. But this discussion of data disclosure avoidance is somewhat illusory. It remains possible that individuals, whose values on $\mathbf{Y}$ and $\mathbf{X}$ are far from those for the rest of the sample, may still in effect be regenerated through this complex statistical estimation process and reemerge virtually intact in the pseudo micro-data file. Thus we would argue that empirical checks on the effectiveness of data disclosure avoidance are still necessary and, in particular, we would advocate examining the issue from the perspective of an intruder (see e.g., Fienberg, Makov, and Sanil (1997), or Lambert (1993)).

Second, there is close connection here with two recently developed statistical methods: (1) the bootstrap (Efron (1979), Efron and Tibshirani (1993)) which is a classical method involving repeated sampling (with replacement) from an empirical distribution function; (2) multiple imputation (Rubin (1987) and (1993)) which is a Bayesian method for generating values that are sampled from a posterior distribution. Our preference is to think about the estimation implicit in the approach outlined here from a Bayesian point of view. Thus, in effect, we are proposing that agencies should first estimate the empirical distribution function, generating the full posterior distribution of $F_{\mathbf{Y}|\mathbf{X}}$ or $F_{\mathbf{Y}|\mathbf{X},\theta}$ and then sample from it using Rubin's multiple imputation approach. From this perspective, the bootstrap can be viewed as a way to sample from something approximately akin to the mean of the posterior distribution.

Third, the sample design for the released records need not be the same as that for the original sample survey. Thus, at least in principle, the agency could use simple random sampling, or even sampling with replacement from $\hat{F}_{\mathbf{Y}|\mathbf{X}}$ or $\hat{F}_{\mathbf{Y}|\mathbf{X},\theta}$. Rubin (1993) emphasizes this point without explaining exactly how to determine what we might call the "equivalent" sample size for the released data files. The heuristic idea is that there is only so much information available in the data and the resampling process cannot increase this. To preserve the appropriate level of accuracy in the data we need to have a bootstrap sample size that at least is conceptually equivalent to the "effective sample size" of the complex sample design, thus reflecting a design effect. This notion is somewhat problematic, however, as the "effective sample size" might well vary from one analytical setting to another!

But perhaps the most important feature of the approach is that users can now analyze pseudo micro-data files to estimate specific quantities of interest, e.g., θ, using standard statistical methodology. In essence the idea is that we can use a standard statistical method such as regression analysis or something more elaborate and thus will produce consistent estimates of the coefficients of interest. What we cannot do, however, is use the usual estimates of standard errors that result from the standard analysis tools.

One of the lessons from both the bootstrap and multiple imputation is that while we can estimate θ using standard statistical methodology applied to the generated bootstrap or multiple imputation sample, we cannot get a proper handle on the variability of our estimates without using replicate versions of the pseudo micro-data file. Generating multiple replicates, however, is a relatively simple task, and estimating variances using the multiple versions of estimated parameters is then straightforward and does not necessarily require special computer programs. This technique of using replicate samples can be exploited for model selection as well, although that process would obviously be superior if done with the full posterior distribution.

5. A Related Approach for the Categorical Data Case

Here we outline the estimation and simulation process of Section 4 for the special case of categorical variables and cross-classification. Our focus is on parametric estimation of the c.d.f., which as we note above is equivalent to estimating the cell probabilities in a contingency table.

The most common class of statistical models used in connection with contingency table data is the loglinear model and for a set of basic sampling schemes (see e.g., Bishop, Fienberg, and Holland (1975) and Whittaker (1990)) there is a direct relationship between a specific hierarchical loglinear model and a set of marginal tables that correspond to the minimal sufficient statistics associated with the model. If we report only those marginal totals appropriate for a loglinear model that fits the data well, then another investigator can, in effect, reconstruct the cell probabilities for the full contingency table (cf., Fienberg (1975)). Further, reporting only a specific set of marginal tables is saying that these are the only totals needed for inference, and this is implicitly suggesting the appropriateness of a specific loglinear model.

As we noted in Section 2, cell suppression and data swapping are in common use as methods for disclosure limitation in categorical variable settings. Unfortunately there seems to be a total disconnect between the literature on disclosure limitation for categorical variables and the now standard literature on loglinear models for categorical data. This is rather unfortunate since, as we noted in Section 2, the notion of margin preservation is fundamental to both cell suppression and data swapping. In the former, cells are suppressed subject to marginal constraints, and in the latter, individuals with one set of margins fixed are swapped between cells, thus preserving other totals. Thus key features of these methods can be embedded in the loglinear model framework, thereby suggesting alternative ways to approach disclosure avoidance. Further results from the loglinear model literature may well be of value in understanding the properties of methods such as cell suppression and data swapping (cf. the discussion in Fienberg (1995) and (1997)), but here we pursue an alternative approach linked to the general strategy described in Section 4.

Our approach needs to be embedded in a model selection and estimation framework where the goal is to develop a replacement table for the original one whose entries are "compatible" with those in the original table, and which, when analyzed would allow the user to draw inferences similar to those drawn from the original table. The first step in such a process is deciding on a model that captures the essential features of the data.

Consistent with the initial smoothing stage of the general strategy proposed in Section 4, we would advocate a model which "overfits" the data, i.e., whose corresponding marginal totals are more extensive than those that might result from a detailed analysis by a specific user. This means that users who analyze the replacement table will be able to search for models and relationships that remain when we preserve the marginal totals and are contained within the model used to generate the replacement table.

Generating the distribution of all cross-classified tables of counts that satisfies a given set of marginal constraints is a problem which has occupied the attention of a substantial number of statisticians in recent years (e.g., see Agresti (1992)). A number of algorithms have been proposed but they have been implemented primarily for two- and three-way cross-classifications. New ideas from the literature on graphical loglinear models suggest that implementation for higher dimensions may at least become feasible (see e.g., Lauritzen (1996), or Whittaker (1990) for details of graphical models). The framework we outline in Section 3 requires us to produce a smooth c.d.f. and then sample from it. In the present context, this seems to suggest, at least heuristically, that we should consider making draws from the exact distribution conditional on a fixed set of marginal totals.

Consider a three-dimensional contingency table with cell counts $\{n_{ij}\}$ and expected cell values $\{m_{ijk}\}$. We can fit loglinear models to the expected cell values such as the model of no 2nd-order interaction.

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} \tag{1}$$

with appropriate side-constraints for identification purposes. The minimal sufficient statistics or "fully efficient statistics" for this model are the margins that correspond to highest order terms: $\{n_{ij+}\}$, $\{n_{+jk}\}$, $\{n_{i+k}\}$.

A special case of Model (1), in which $u_{23(jk)} = 0$ for all j and k, is interpretable as the conditional independence of Variables 2 and 3 given Variable 1. All conditional independence models for a multidimensional contingency table are loglinear models.

In unpublished work, John Darroch and Gary Glonek attempted to construct a Markov chain algorithm for generating draws from the conditional distribution given the margins implied by a loglinear model. The transitions of their Markov chain in effect involved one-step data swaps. Diaconis and Sturmfels (1998) show how to implement a generalization of the Darroch and Glonek approach using the method of Gröbner bases and provide a proof of the convergence of the algorithm through the irreducibility of the Markov chain. The important thing to note for the present circumstances is that simple data swaps are not sufficient to "reach" all possible tables. The method of Gröbner bases gets around this problem by introducing "generalized data swaps" that combine specific pairs, triples, etc. of data swaps in very specific forms. Fienberg et al. (1997) explore this methodology in connection with a number of different non-decomposable graphical loglinear models and we illustrate results from this approach in the example below.

In order to ensure some level of smoothness in the resulting tables associated with random draws from the exact distributions discussed above, we can retain only those draws "compatible" with a more complex loglinear model. Note that the variability of the perturbation methodology used here is directly accessible to the user, since anyone can begin with the reported table and information about the margins that are held fixed, and then run to Diaconis Sturmfels Markov chain algorithm to regenerate the full

distribution of all possible tables with those margins. This then allows the user to make inferences about the added variability in a formal modeling context in a form that is similar to the approach to inference in PRAM by Gouweleeuw et al. (1998). As a consequence, the procedure proposed here, and variants on it, represent a major improvement from the perspective of access to data over cell suppression and data swapping.

An intruder can follow the user in attempting to identify individuals represented in the cross-classification. The principal tool at the intruder's disposal from the released data is the information from the released marginal totals. These can be used to compute upper and lower bounds on the table entries and thus to determine the disclosure exposure of the release (see e.g., Fienberg (1998)). If the upper and lower bounds for some cells are "too close" to one another then the agency must suppress marginal information relevant to the user's needs, thus restricting the utility of the released data. But then the decision to suppress becomes a conscious one and its implications for the subsequent analyses by others can be explored, and perhaps mitigated by instructions or information provided to secondary analysts.

The intruder can alternatively use the information on the released marginals to generate the relevant Gröbner basis and then run the Markov chain procedure to yield all possible tables with the fixed margins. If cells with entries of "1" or "2," for example, are almost always unchanged across tables, this information is akin to that from the bounds, and the response to it must be similar.

An alternative to the procedure outlined in this section would be the generation of a full posterior distribution for the cell probabilities in the table, e.g., using the methods of Epstein and Fienberg (1992) and/or Madigan and York (1995), and then sampling from that posterior distribution as in multiple imputation. We hope to explore this approach in a future article.

6. An Example

The following 3-way table example gives the cross-classification of individuals by race, gender, and income (collapsed into three categories) drawn from the 1990 U.S. Decennial Census Public Use Files (see Table 3).

In Table 4 we present the maximum likelihood estimates for the expected counts corresponding to the entries in Table 3 under the no 2nd-order interaction model with multinomial sampling. We computed these in S-plus. The likelihood ratio chi-squared value for the fit of this model was 2.89 on 4 d.f. This is indicative of a moderately good model fit, although it is actually somewhat difficult to assess the fit given the sparseness of the row in the first layer which has a total count of 1 in it.

Then we generated 1,000,000 tables with the same 2-way margins, using the Diaconis and Sturmfels (1998) algorithm (see the Appendix for details). For each table, we calculated the likelihood ratio chi-squared goodness-of-fit value based on the maximum likelihood estimates for these margins under the no 2nd-order interaction model. In Figure 1, we have plotted the ordered likelihood-ratio chi-squared values against the cumulative values from the corresponding χ^2 distribution with 4 d.f. A good fit to the χ^2 distribution would be represented by a straight line. We can see by the plot that the distribution of our simulated tables does not fit the χ^2 distribution as well as we might have hoped. This is due again to the

Table 3. Three-way cross-classification of gender, race, and income for a census tract. (Source: 1990 Census Public Use Microdata Files)

Gender = Male

Race	Income level			
	< $10,000	> $10,00 and < $25,000	> $25,000	Total
White	96	72	161	329
Black	10	7	6	23
Chinese	1	1	2	4
Total	107	80	169	356

Gender = Female

Race	Income level			
	< $10,000	> $10,00 and < $25,000	> $25,000	Total
White	186	127	51	364
Black	11	7	3	21
Chinese	0	1	0	1
Total	197	135	54	386

sparse nature of the table. Because the maximum likelihood estimates for the sparse cells are not whole numbers, we cannot reach tables that have very low-chi-squared values, because we cannot get close enough to the maximum likelihood estimates. Yet, other than the bias at the low end of the distribution, the distribution seems to be approximately chi-squared.

Clearly the selection of a single table from the distribution of all possible tables, exemplified by the simulation study reported on in this section, poses a problem and we

Table 4. Maximum likelihood estimates for data in Table 3 under the no 2nd-order interaction model

Gender = Male

Race	Income level			
	≤ $10,000	> $10,00 and ≤ $25,000	> $25,000	Total
White	97.09	72.15	159.76	329
Black	9.21	6.41	7.38	23
Chinese	0.70	1.44	1.86	4
Total	107	80	169	356

Gender = Female

Race	Income level			
	≤ $10,000	> $10,00 and ≤ $25,000	> $25,000	Total
White	184.91	126.85	52.24	364
Black	11.79	7.58	1.62	21
Chinese	0.30	0.56	0.14	1
Total	197	135	54	386

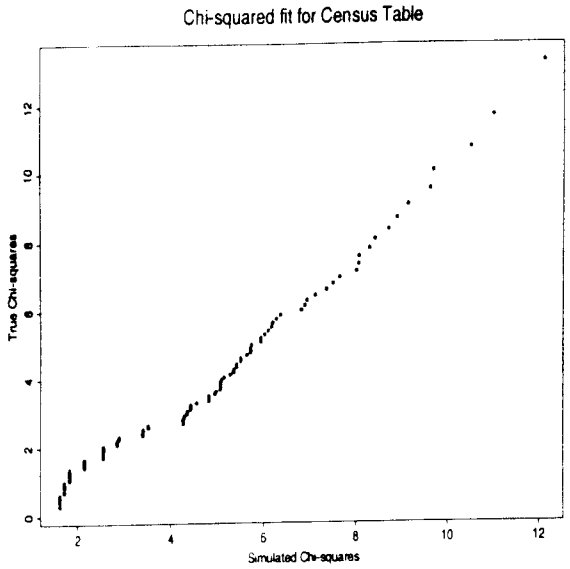

Fig. 1. Chi-squared goodness-of-fit plot for simulated values using the data in Table 3

need to consider "rejecting" those tables far from the original data. We can do this using some type of chi-squared distance measure (see e.g., those described in Bishop et al. (1975) or in Read and Cressie (1988)), although we have not implemented such a rejection process here to select a single table or even a set of tables to report.

The dimensionality of the table in this example has simplified our task of implementing the exact table methodology outlined in Section 5. Implementation for high-dimensional tables would be computationally intensive, with the most difficult task being the generation of the Gröbner basis used in the Markov chain. We continue to explore this methodology for simplifications.

Finally, there is level of disclosure risk associated with this table, because of the two counts of "1" in the marginal totals that are held fixed. Fienberg (1998) computes the upper and lower bounds for the cell entries in this example given the two-way margins, and Fienberg and Makov (1998) discuss the same example in the context of estimating the existence of "population uniques."

7. Discussion and Further Research

It is important to distinguish between the idea of generating public-case micro-data files based on real people and real data through a statistical simulation process, such as we have outlined in this article, and the typical micro-simulation model, which may rely indirectly on data via statistical models but which does not correspond to data on real people. There is a serious difference between "pseudo people" who resemble individuals from whom we have actually collected data of interest, and "imaginary people" for whom we have invented data through a stochastic or nonstochastic modeling process. In this article we propose the former, not the latter. There are, of course alternatives such as

aggregation and collapsing, which fit the matrix masking framework outlined in Section 2, and comparisons with them do need to be made in practice.

There are several virtues of the proposed framework outlined above. First, we believe that it would force agencies to take their own data and their sources of error more seriously, as these are key inputs to the modeling effort outlined in Section 4. Second, we believe that it would solve a large part of the data disclosure avoidance problem. Third, the framework would generate public-use micro-data files of a form that would allow users to apply standard statistical methodology and model search methods (cf., Gouweleeuw et al. (1998).

There are a number of formidable technical details that need to be addressed before an agency could properly implement the proposed framework in a systematic fashion. For the exact distribution method for contingency tables outlined in Section 5, the computational details for high dimensions remain problematic. For the general perturbation approach of Section 4, examples of technical issues include:

- How should an agency combine the multiple sources of error and uncertainty?
- What smoothing methods should be used and how much smoothing is appropriate?
- How do we determine "effective" sample size for pseudo micro-data files? The application of bootstrap ideas relies on certain series expansions (see e.g., Hall (1992)), and these typically require the use of a bootstrap sample of the same size as the original sample. What is the equivalent notion here?
- How many replicates are required for variance estimation? Rubin (1993) suggests the use of four or five replicates in the multiple imputation context. Efron and Tibshirani (1993) use very large numbers of bootstrap replications. Multiple imputation gains its power in this regard from the parametric specification of the full posterior distribution. Will a smaller number of replicates suffice for either approach?

Further, the actual implementation of algorithms of the highly multidimensional situations involved in censal and survey data may require new statistical methods and theory. For example, as we suggested in Section 5, the problem of simulating from distributions for multidimensional contingency tables subject to marginal constraints has been implemented primarily for two- and three-dimensional tables. Implementation for higher dimensions requires new strategies and algorithms. These are at the forefront of current statistical and mathematical research.

Finally, we may need to think about the statistical estimation problems outlined here in a form different from that which we usually find in the methodological literature. It would be wrong, however, to think of the approach suggested here as being rooted solely in bootstrap theory or as relying on Bayesian multiple imputation, as that would in essence be expecting to get usable perturbed data at "no cost" in terms of bias and variability. There is a price to pay for disclosure limitation, and the more restrictions one places on the release of data the bigger the price. Moreover, because of the multiplicity of goals that we are attempting to address, we may need to think in terms of providing the users with data that enable them to approximate the conditional distributions $F_{\mathbf{Y}|\mathbf{X}}$ and $F_{\mathbf{Y}|\mathbf{X},\theta}$ rather than reproduce them in a precise statistical fashion. This relates to Meng's (1994) notion of uncongeniality between an imputer's assessment and those assessments of the users.

In this article, we have tried to suggest that both government agencies and users bear

responsibility when it comes to utilizing census and survey data. It is no longer enough for agencies to prepare public-use files and extensive sets of tabulations as they have in the past. Nor can they continue to ignore the analytical goals of the users of their data. At the same time, the users must learn how various sources of survey error affect their analytical goals, and to build such information into the statistical procedures they use. We have argued that, by looking to and utilizing recent developments in statistical methodology, we may be able to develop an integrated approach to the release and analysis of survey data which will help us all learn to take uncertainty and error seriously. Perhaps the framework proposed in this article will be the first step towards this goal.

Appendix: Algorithm Used to Generate Chi-Squared Values in Example of Section 5

In Section 6, we used the Gröbner bases generated for various examples as input to the generation of values from the exact distribution using the Monte Carlo algorithm proposed in Diaconis and Sturmfels (1998), and described in Fienberg et al. (1997). Sampling from the output generated by the Markov chain involves some care since we need to avoid the dependence associated with persistence in low probability states. We replicated the following detailed approach used by Diaconis and Sturmfels (1998) in their contingency table example.

Read real table, mle table and moves into respective structures
for **i** := 1 to (**number of tables** × 500)
  r1 := *randomly generated number from 1 to number of moves*
  temporarily make move **r1** *to find* **table probability**
  r2 := *randomly generated number from Unif[0,1]*
  if **table probability > r2** and *move creates no negative cells*
    then *make the move permanently*
    else *do not make the move permanently*
  if **i** mod 500 = 0
    then print **current tables' chisquared value**
next **i**

This algorithm is straightforward. In order to keep from oversampling tables that have a low transition probability, the algorithm samples every 500th table, therefore requiring 500 × the number of desired tables iterations to run. The step which avoids the creation of negative cell values was actually coded into the function that changed the table only to make it obvious that one would not want to allow a table with negative cells. Another way to do it would have been to assign zero probability to any table with negative cells. The chi-squared value of the table was based on $G^2 = -2log\lambda$, where λ is the likelihood ratio using the user supplied maximum likelihood estimates (MLEs) for the table cell entries, under the multinomial sampling model. These are functions of the fixed margins, and can be computed using standard algorithms, e.g., we used the routine for MLEs in S-plus.

8. References

Agresti, A. (1992). A Survey of Exact Inference for Contingency Tables (with Discussion). Statistical Science, 7, 131–177.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

Cox, L.H. (1994). Matrix Masking Methods for Disclosure Limitation in Microdata. Survey Methodology, 20, 165–169.

Cox, L. (1995). Network Models for Complementary Cell Suppression. Journal of the American Statistical Association, 90, 1453–1462.

Cox, L. and Sande, G. (1978). Automated Statistical Disclosure Control. American Statistical Association, Proceedings of the Section on Survey Research Methods, 177–182.

Dalenius, T. and Reiss, S.P. (1978). Data-swapping: A Technique for Disclosure Control (extended abstract). American Statistical Association, Proceedings of the Section on Survey Research Methods, 191–194.

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

Diaconis, P. and Sturmfels, B. (1998). Algebraic Algorithms for Sampling from Conditional Distributions. Annals of Statistics, 26, 363–397.

Duncan, G.T. and Fienberg, S.E. (1998). Obtaining Information While Preserving Privacy: A Markov Perturbation Method for Tabular Data. In Statistical Data Protection (SDP'98) Proceedings, IOS Press (1998), Luxembourg, to appear.

Duncan, G.T. and Lambert, D. (1986). Disclosure-limited Data Dissemination (with Discussion). Journal of the American Statistical Association, 81, 10–28.

Duncan, G.T. and Lambert, D. (1989). The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, 7, 207–217.

Duncan, G.T. and Pearson, R.B. (1991). Enhancing Access to Micro-data while Protecting Confidentiality: Prospects for the Future (with Discussion). Statistical Science, 6, 219–239.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7, 1–26.

Efron, B. and Tibshirani, R. (1993). An Introduction to the Bootstrap. New York: Chapman and Hall.

Epstein, A.D. and Fienberg, S.E. (1992). Bayesian Estimation in Multidimensional Contingency Tables. In Proceedings of Indo-U. S. Workshop on Bayesian Analysis in Statistics and Econometrics (P.K. Goel and N.S. Iyengar (eds)). Lecture Notes in Statistics Vol. 75, New York: Springer-Verlag, 37–47.

Fienberg, S.E. (1975). Perspectives Canada as a Social Report. Social Indicators Research, 2, 153–174.

Fienberg, S.E. (1994a). Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality. Journal of Official Statistics, 10, 115–132.

Fienberg, S.E. (1994b). A Radical Proposal for the Provision of Micro-data Samples and the Preservation of Confidentiality. Technical Report No. 611, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, U.S.A.

Fienberg, S.E. (1995). Discussion of Presentations on Statistical Disclosure Methodology. In Seminar on New Directions in Statistical Methodology, Statistical Policy Working Paper No. 23. Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, U.S. Office of Management and Budget, Washington, DC, Part 1, 68–79.

Fienberg, S.E. (1996). Taking Uncertainty and Error in Censuses and Surveys Seriously. Symposium 95: From Data to Information – Methods and Systems. Proceedings, Statistics Canada, Ottawa, 97–105.

Fienberg, S.E. (1997). Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research. Background paper prepared for the Committee on National Statistics.

Fienberg, S.E. (1998). Fréchet and Bonferroni Bounds for Multi-way Tables of Counts with Applications to Disclosure Limitation. Statistical Data Protection (SDP'98) Proceedings, IOS Press, Luxembourg, forthcoming.

Fienberg, S.E. and Makov, U.E. (1998). Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data. Journal of Official Statistics, 14, 385–397.

Fienberg, S.E., Makov, U.E., and Sanil, A.P. (1997). A Bayesian Approach to Data Disclosure. Optimal Intruder Behavior for Continuous Data. Journal of Official Statistics, 13, 75–90.

Fienberg, S.E., Meyer, M.M., Makov, U.E., and Steele, R.J. (1997). Notes on Generating the Exact Distribution for a Contingency Table Given Its Marginal Totals. Unpublished manuscript.

Fienberg, S.E., Steele, R.J., and Makov, U.E. (1996). Statistical Notions of Data Disclosure Avoidance and Their Relationship to Traditional Statistical Methodology: Data Swapping and Log-linear Models. Proceedings of U.S. Bureau of the Census 1996 Annual Research Conference. U.S. Bureau of the Census, Washington, DC, 87–105.

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

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., and de Wolf, P.-P. (1998). Post-randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, 14, 463–478.

Griffin, R., Navarro, A., and Flores-Baez, L. (1989). Disclosure Avoidance for the 1990 Census. Proceedings of the Section on Survey Research, American Statistical Association, 516–521.

Hall, P. (1992). The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag.

Heer, G.R. (1993). A Bootstrap Procedure to Preserve Statistical Confidentiality in Contingency Tables. Proceedings of the International Seminar on Statistical Confidentiality, Dublin, Ireland, September 8–10, 1992, 261–271.

Kennickell, A. (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 SCF. Paper presented at U.S. Census Bureau Workshop on Record Linkage, March, Washington, DC.

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

Lauritzen, S. (1996). Graphical Association Models. New York: Oxford University Press.

Lavine, M. (1992). Some Aspects of Polya Tree Distributions for Statistical Modeling. Annals of Statistics, 20, 1222–1235.

Liew, C., Choi, U., and Liew, C. (1985). A Data Distortion by Probability Distribution. ACM Transactions on Database Systems, 10, 395–411.

Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.

Madigan, D. and York, J. (1995). Bayesian Graphical Models for Discrete Data. International Statistical Review, 63, 215–232.

Meng, X.-L. (1994). Multiple-imputation Inferences with Uncongenial Sources of Inputs, (with Discussion). Statistical Science, 9, 538–573.

Navarro, A., Flores-Baez, L., and Thompson, J. (1988). Results of Data Switching Simulation. Presented at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees.

Read, T.R.C. and Cressie, N.A.C. (1988). Goodness-of-Fit Statistics for Discrete Multivariate Data. New York: Springer-Verlag.

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

Rubin, D.B. (1993). Discussion, Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461–468.

Subcommittee on Disclosure-Avoidance Techniques (1978). 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, DC.

Subcommittee on Disclosure-Avoidance Techniques (1994). 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, DC.

Sullivan, G. (1989). The Use of Added Error to Avoid Disclosure in Microdata Releases. Unpublished Ph.D. dissertation, Department of Statistics, Iowa State University, Ames, IA, U.S.A.

West, M., Müller, P., and Escobar, M. (1994). Hierarchical Priors and Mixture Models, with Application in Regression and Density Estimation. In Aspects of Uncertainty, (P.R. Freeman and A.F.M. Smith (eds)). New York: Wiley, 363–386.

Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. New York: Wiley.

Zalkind, D.L. (1978). Comments on Data-swapping: A Technique for Disclosure Control. American Statistical Association, Proceedings of the Section on Survey Research Methods, 195–196.

Received January 1998
Revised August 1998

Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 503–508

Comment

Peter Kooiman[1]

One person's noise is another person's signal
(Gary S. Brown, 1998)

1. Introduction

The study by Fienberg et al. contains two lines of thought. Sections 2, 5, and 6 deal with data swaps in cross tabulations of categorical variables, keeping certain margins intact. I consider the log linear modeling approach advocated by the authors promising; it could provide a sound statistical underpinning to such data swaps. However, in Sections 3 and 4 the authors extend their approach to a strategy for the release of survey microdata sets broadly. For this type of data release I am quite skeptical about the feasibility of the modeling strategy. Finally I draw a parallel with the *National Accounts process*.

2. Data Swapping in Cross Tabulations

The authors provide an interesting and innovative discussion of data swapping in cross tabulations of categorical variables. Cross tabulations published by statistical agencies typically involve only a few dimensions. Only when very detailed classifications are used, or populations are very skew, disclosure problems may occur in such tables. Then table cells have to be suppressed or data swaps have to be applied, moving table entries from one cell to the other. Hitherto such swapping procedures have been applied rather mechanically or deterministically. In my opinion the main virtue of the study is that it opens up a line of research which could provide sound statistical underpinnings for data swapping methodology. The idea is to first try and reduce the frequency table to be protected by searching for a more parsimonious representation through log-linear modeling. Assuming that a satisfactory model exists which is more economical than the fully saturated one, we can separate off some noise from the signal present in the frequency table. Keeping the signal intact, we can then concentrate our data swaps in the noisy part. From the point of view of subsequent analysis this is harmless, provided we apply the swaps in such a way that no artificial structure emerges where in the original table no structure existed. If the model of the frequency table can be represented as a set of marginal tables these tables contain all useful information there is in the original table, and

[1] Department of Statistical Methods, Statistics Netherlands, Voorburg, The Netherlands <pkmn@cbs.nl>.
Acknowledgments: The views expressed in this comment are those of the author and do not necessarily reflect the policies of Statistics Netherlands. The author thanks Jeroen Pannekoek and Leon Willenborg for stimulating discussions and useful remarks on an earlier version of this comment. They bear no responsibility for any of the views expressed or any of the remaining errors, though.

it is then quite natural to devise procedures which keep these tables intact. As an alternative the agency might conclude that it should revise its set of tables to be published: when all useful information is contained in a subset of marginal tables, why not publish these tables instead of the original higher-dimensional one, contaminated with uninformative noise?

In Section 4 and parts of Section 5 of their study the authors claim that the approach set out above can be extended and developed into a new strategy for the release of survey microdata files. Unfortunately it is not entirely clear to me how the two parts relate. The general strategy is phrased in terms of the conditional distribution of Y given X. Apart from this being very problematical for a statistical agency preparing a data file for general use (almost any variable can act as Y or as X, depending on the research question involved), it is at odds with the log linear modeling of frequency tables which concentrates on the full joint distribution of all table entries, i.e., $F_{Y,X}$ instead of $F_{Y/X}$. Also the general strategy nowhere mentions the problem of simulating from a data model *keeping certain margins intact*, which is at the core of the other part of the study. Indeed almost all of the technical problems arising in the data swapping part of the study are precisely attributable to the fact that we have to simulate conditionally on given margins. The authors implicitly admit the weak relationship between the two parts when they state, a few lines before their Equation (1) in Section 5, that the general strategy applied in the context of log linear modeling of categorical data sets "seems to *suggest*, at least *heuristically*, that we should consider making draws from the exact distribution conditional on a fixed set of margins" (italics mine). This is indeed not a very strong claim.

3. Releasing Microdata

In the remainder of this comment I concentrate on the claim that the general modeling strategy the authors present can provide a basis for the release of survey microdata files. My frame of reference is a statistical agency that purports to provide the academic community with general purpose microdata files for statistical research. The strategy consists of a modeling step, in which the agency develops a data model which is more parsimonious than the data set itself, and a simulation step in which a number of replicate pseudo microdata files are created by drawing from the exact distribution associated with the model. As I understand it, the authors have in mind a situation where a model can be obtained which on the one hand "overfits" the data, so that it does not distort the relevant data patterns, and, on the other hand, is economical enough to leave room for data swaps orthogonal to these data patterns.

To fix ideas let us think of a data set of 10,000 records and 6 categorical variables with 10 categories each. The fully saturated model has 10^6 cells, and clearly represents a considerable overfit. No analyst is likely to be interested in fourth or fifth order interactions; one would not even know how to interpret such effects. In practice almost all analysis concentrates on first order interactions, i.e., second moments of the data, and only incidentally on second order interactions. So, if we represent the data set by a log linear model leaving out all interactions of order three and higher, we will not lose much. This model involves 20 three-way tables with $10 \times 10 \times 10 = 1,000$ cells, accounting

for about 15,000 non-redundant restrictions on the data set. Representing each variable by 10 (0, 1)-dummies the data file contains $6 \times 10,000 = 60,000$ non-zero entries, which we can swap around a bit, provided we do not violate the 15,000 restrictions on the second order interactions. So there is some hope that we have sufficient degrees of freedom to make this a feasible exercise. At the risk of distorting the data for some subsequent analysis one might do a more thorough modeling, and throw out a number of the three-way tables, thereby increasing the degrees of freedom available for data swaps.

Survey data sets associated with the large surveys that statistical agencies conduct are much more detailed than in the example above. A typical data file may contain over 200 variables. These are recorded using very detailed classifications with hundreds or even thousands of categories: location by ZIP-code, industrial activity, profession, educational level, illnesses, causes of death in four of five digits, age in years, and so on. So, as a more typical situation to cope with, we now consider a data file with 50,000 records, and 200 variables with 25 categories per variable. Representing each variable by a set of dummies again, we now have $25 \times 200 = 5,000$ dummies. If we restrict ourselves to first order interactions only we have approximately $0.5 \times 5,000^2 = 1.25 \times 10^7$ cells, representing 1.15×10^7 non-redundant restrictions. There is no hope of keeping all of these intact with only $50,000 \times 200 = 10^7$ non-zero entries to swap around. Things are even worse when we consider a number of very detailed variables. If the file consists of 50,000 records and 10 variables with 500 categories each we have approximately 1.12×10^7 restrictions and 5×10^5 non-zero entries. If we were to include second order interactions, doing justice to the idea of some overfitting of the data, the number of restrictions would explode. With probability close to one, the only data configuration satisfying all these restrictions is the original data set and nothing else. With typical survey data files the number of variables, and the amount of detail about these variables, is such that non-distortive modeling is entirely out of scope.

Researchers are eager to obtain as much detail as they can. They consistently express their discomfort with reductions in detail statistical agencies impose in view of disclosure protection. One of the puzzles here is why researchers want so much detail. Even enormous amounts of records will not provide enough degrees of freedom to support valid statistical inference at the very fine level of detail researchers require. Once they restrict themselves to data patterns that can sensibly be investigated statistically they necessarily resort to far lower dimensional spaces using subsets of variables at far more aggregated levels. This seems to support the modeling approach sketched by the authors. Details beyond a certain level of aggregation will never contribute to valid inference, so what are we going to lose when this is replaced by noise in the sampling process of the pseudo microdata files? The answer is that researchers want to construct their own aggregates, tailor-made for the specific research questions they want to investigate. For certain studies they need age groups from 12–18, for others 17–21 is more appropriate. Having a model based on 5-year classes, 10–15, 16–20, ..., *or a pseudo microdata file representing such a model*, is not helpful to them. Similarly, they want to be able to construct their own derived variables, such as travelling distance between place of living and place of work. When we aggregate such locational variables into relatively crude indicators, researchers can no longer make such derivations. If we want to support all of these research needs, without knowing beforehand the future use of the released microdata, the only solution is to provide

as much detail as possible. I simply do not see how this could ever be accommodated within the framework of the modeling approach advocated by the authors.

It is the task of official statistics to provide society with impartial and trustworthy data reflecting the true state of society as closely as possible. These data constitute the basis for social and scientific debate and subsequent decision making. Survey data collected by statistical agencies constitute an extremely valuable resource for scientific and policy research. The number of questions that can be addressed is enormous. An evolving scientific and policy debate continuously generates new parameters of interest. It is hardly conceivable how such a rich data mass could ever be summarized in a single statistical model in an impartial way. Degrees of freedom considerations necessarily lead to a very restrictive specification. Model selection is an art, and certainly proceeds in crude ways when such masses or variables have to be analysed. Higher order interactions, representing several hundreds or thousands of individual dummy variables, are included or excluded all at once, neglecting underlying subtleties. Detailed classifications can be aggregated in numerous ways, none being uniformly superior to the others. Without a specific research question in mind there is no guidance as to which data patterns are relevant or not. The probability that two equally qualified analysts end up with the same model is close to zero. As long as this is true, a considerable amount of subjectivity cannot be avoided. As a consequence multivariate statistical modeling of large survey data sets cannot provide a foundation for the dissemination of general purpose survey data sets by a statistical agency, *by principle*.

Now, thinking the unthinkable, suppose we have obtained an unambiguously satisfactory model, i.e., one that properly represents all "significant" relationships in the survey data set. When we generate pseudo microdata sets by sampling from this model the information in the samples cannot be more than what was already contained in the model. Otherwise stated: an analyst will at best be able to reconstruct the model underlying the data generation process (or some reduction thereof). If the analyst does not retrieve the true model he or she errs, he or she will end up with invalid conclusions. If the analyst does, he or she might ask why the statistical agency did not simply publish the model instead of disguising it in the form of pseudo microdata files. If the agency does publish the model, or the equivalent set of marginal tables, the knowledgeable analyst will not start analysing the pseudo microdata files at all. It is like cross-word puzzles: nice for entertainment, but not really of interest when the solution is on the back of the envelope. Following this line of thought *ad absurdum* we clearly see the enormous difficulties of the modeling approach: if really successful it would make superfluous *any* subsequent statistical analysis of the pseudo microdata sets. Thus, it necessarily assumes that statistical offices are able and qualified to extract *any useful information there is* from their survey data files. Needless to say, they are not.

The main problem with the approach, which it shares with data swapping, is that it tries to restrict disclosure protection measures to the noise in the data, thereby keeping the signal intact. Swapping noise is harmless for statistical analysis, but can help to protect individual records from re-identification by a data intruder. However, without a specific model noise is hardly defined. Aiming at a general purpose microdata file we must recognize that the only sufficient statistic for all the information that is present in a typical rich survey data set is the data set itself. Adding noise to protect such data against disclosure

necessarily distorts potentially relevant data patterns. For some analyses this may be innocent, since these do not exploit the distorted part of the data patterns. Others are inevitably affected. The alternative approach of Gouweleeuw et al. (1998) recognizes this and therefore no longer tries to keep data patterns intact. Instead it employs the known statistical distribution of the data swaps (i.e., misclassifications) to estimate the latent unperturbed frequency table. Only when we know beforehand which data patterns to concentrate upon, such as when a limited set of low dimensional tables is published from e.g., a census, is it possible to control properly for the distortion due to data swaps. It is for such limited applications that the modeling approach advocated by the authors may be appropriate, especially when it is impractical to publish the set of marginal tables equivalent to the data model employed.

An important remaining question, on which the study touches only briefly, is whether the modeling approach provides sufficient protection against disclosure. The implicit assumption seems to be that the log linear data reduction employed is sufficient to disguise the identities of the subjects underlying the whole exercise. In practice it is difficult to verify such an assumption. Indeed, it is not sufficient to check whether the marginal tables representing the model employed are safe one by one. These tables are linked through their common source, and it is the combination of the tables which matters. Jointly they define a set of admissible solutions for the underlying microdata file. When degrees of freedom are insufficient, as in one of the examples above, this set must degenerate locally (e.g., the General Motors record) or perhaps even globally into a single point, i.e., the original micro data set. So, apart from being a sufficiently rich data representation, we should add the requirement that the log linear model employed entails enough degrees of freedom to support a sufficiently broad set of admissible solutions, especially with respect to all potential identification keys. Verifying this requirement involves very hard combinatorial computations that are unfeasible given the size and the amount of detail of typical survey data sets.

This is further complicated by the release of *replicates* of the data file. By matching replicates an intruder can find clues as to which data fields in which records have been swapped or not, especially when the set of admissible solutions for a specific record is narrow. Using modern matching technology, and modest quantities of noise, almost perfect matches can be obtained, given the large numbers of variables involved (see e.g., Winkler 1998). Perhaps, such matching exercises could be used by the agency to check the safeness of the pseudo micro data files to be released.

4. National Accounts Process

The prescription, by the authors, to include all information the agency has about errors in the data in the modeling exercise, reminds me of the data integration process typically performed by National Accounts people. They try to reconcile conflicting information from several surveys, using their accounting framework as a data model. Correcting for differences in definitions of variables, and supplementing for missing subpopulations, they exploit accounting restrictions, physical demand-supply equalities, and sampling variances to construct a consistent picture of the national or regional economy. Similar accounting systems have been worked out for other phenomena: labour accounts, tourism

accounts, socio-economic and demographic accounts, environmental accounts (see e.g., Van Tuinen 1995). Typically these accounts are both prepared and published in the form of tables at an intermediate level of aggregation. Although in many cases no formal statistical procedures are applied, the resulting figures can nevertheless be conceived of as *full information (gu)estimates* based on all available evidence.

Within the general framework presented by the authors the National Accounts tables can perhaps be identified with the model from which pseudo microdata files could be generated. At Statistics Netherlands a similar idea has been discussed in a quite different context. Due to the corrections made to the primary survey data inputs in the course of the National Accounts process, National Accounts tables are not numerically consistent with tables the agency publishes from the primary survey data sets themselves. To solve this problem it has been contemplated to reweight the surveys *ex post*, taking the National Accounts outcomes as given. The formal underpinning of such a procedure was developed by Renssen and Nieuwenbroek (1997). In following this line of thought we would end up with microdata files consistent with a given set of tables, i.e., the National Accounts tables, or any other applicable accounting framework used to reconcile conflicting survey outcomes. Since we stick to the survey data itself, only adjusting the individual record weights, this obviously would not contribute to the solution of the disclosure protection problem, though.

5. References

Brown, G.S. (1998). Guest Editorial, IEEE Transactions on Antennas and Propagation, 46, 1.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.-P. (1998). Post Randomization for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, 14, 463–478.

Renssen, R.H. and Nieuwenbroek, N.J. (1997). Aligning Estimates for Common Variables in Two or More Sample Surveys. Journal of the American Statistical Association, 92, 368–374.

Van Tuinen, H.K. (1995). Social Indicators, Social Surveys and Integration of Social Statistics. Statistical Journal of the United Nations, 12, 379–394.

Winkler, W.E. (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. Paper presented at the SDP'98 conference, March 25–27, Lisbon.

Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 509–511

Rejoinder

Stephen E. Fienberg[1], *Udi E. Makov*[2], *and Russell J. Steele*[1]

Kooiman offers insightful comments on our article and proposed disclosure limitation methods. Our differences with him are largely a matter of perspective. He is associated with a statistical agency while we are university based and focus on the desires of the statistical users. In the Netherlands there is a tradition of limited releases for research purposes which we contrast with the practice in the United States of the availability of substantial public-use microdata files. In what follows, we attempt to provide answers on four of the issues Kooiman raises.

First, Kooiman asks about the link between what he describes as the two parts of the article, i.e., the "general strategy" and the part based on the exact distribution of a table under a loglinear model conditional on its margins. He describes the relationship as weak; we think of it as strong and reasonably compelling. The interesting thing about the categorical case is that the empirical cumulative distribution function is the contingency table itself. Given the focus by many statistical agencies (e.g., Statistics Canada and the U.S. Bureau of the Census) on fixing selected marginal totals, and on the widespread use of loglinear models for which selected marginal totals are minimal sufficient statistics, then the exact distribution is an estimate for the empirical distribution function in question. Whether it is a good one or not remains to be seen, but we note that many statistical methodologists do recommend inference based on the conditional distribution given the minimal sufficient statistics. How close such an approach is to a fully Bayesian posterior distribution we also do not yet know.

Another reason for thinking about the fixing of marginal totals arises in the context of a sequential query system of the sort described in Keller-McNulty and Unger (1998). Envision a data base consisting of a large contingency table. Queries come in the form of requests for selected marginal tables. Once a marginal table is released by such a system, it remains available to others and so fixing it for all subsequent releases becomes the most reasonable way to proceed.

Second, Kooiman goes on to envision a large example of a 10^6 table. The contingency tables that we encounter in actual surveys have many more variables (as he notes) but each typically has fewer categories, often only two. So the difficulties regarding how to proceed may not quite be as bad as he suggests. Nonetheless, we agree that asking an agency to carry through our prescription with care for every such data set seems unreasonable. But unless it thinks about the underlying phenomena and about models to describe interrelationships, the agency will be totally ad hoc in its functioning and will either release information it should not or severely impair the utility of released data. Thus

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
[2] Department of Statistics, Haifa University, Haifa, Israel.

the agency needs to use some combination of experience and methodological thinking. In this sense, we agree with Kooiman that various forms of aggregation for key variables such as geography and complex classification schemes is a necessity and loglinear models will provide only limited help here. But from this point on, we disagree with his assessment of how to proceed.

Kooiman focuses on the release of only limited amounts of data for restricted purposes, to which he would apply his postrandomization method (PRAM) described in Gouweleeuw et al. (1998). We think PRAM is an innovative technique, but it is very limited, especially when it comes to the preparation of large public-use microdata files. This is because its primary use is for only a small number of key variables, as Kooiman himself notes. In the U.S. at least, such an approach would be unacceptable to the broad group of public data users, and we believe rightfully so. Nonetheless, we recognize and respect the different legal settings and the different expectations of both the public and researchers in other countries around the world. It is for this reason that we hope to see the evolution of a pluralistic approach to disclosure limitation that attempts to take advantage of a range of methodologies, which might include ours, PRAM, Argus, Hundepool et al. (1998a, b), etc.

Third, Kooiman questions the implications and reconciliation of alternative models for our method. He argues that it is impossible to obtain an unambiguously satisfactory model for a survey data set. Since our method depends on this it must be flawed. Perhaps so, but the issue is how badly it is flawed. For complex high-dimensional tables, it is possible to embed multiple user models and questions of interest in the context of some larger statistical model (or at least approximately so). Sampling from the conditional distribution associated with such an enlarged "covering" model is what we propose. If we could achieve this aim only by making choices on aggregation of categories and through other compromises, we believe that this would be far preferable to throwing our hands up in despair or resorting to total ad hockery.

So we come down to the issues of access versus disclosure limitation, noise versus signal, and whether the noise associated with our method overwhelms the signal. Kooiman is correct in noting that for the exact distribution method of Section 5, disclosure is a problem unless there is a sufficiently broad set of admissible solutions. But as long as the counts in margins are sufficiently large, we think that there is promising evidence here that our methods do limit disclosure, and that sufficient signal will remain to make resulting public use data sets of great value to others. Kooiman is skeptical. On disclosure limitation he refers to Winkler (1998), but a close reading of Winkler's results and a replication carried out at Carnegie Mellon suggest that his concerns are generally of limited relevance to the protection of large public use data sets unless there is an intruder with detailed and accurate blocking information and files that allow for a 1–1 match. We suggest, therefore, that the properties of our method are empirical matters worthy of continued investigation.

References

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., and de Wolf, P.-P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, 14, 463–478.

Hundepool, A., Willenborg, L., Wessels, A., Van Gemerden, L., Tiourine, S., and Hurkens, C. (1998a). μ-ARGUS User's Manual. Department of Statistical Methods, Statistics Netherlands.

Hundepool, A., Willenborg, L., Van Gemerden, L., Wessels, A., Fischetti, M., Salazar, J.-J., and Caprara, A. (1998b). τ-ARGUS User's Manual. Department of Statistical Methods, Statistics Netherlands.

Keller-McNulty, S. and Unger, E.A. (1998). A Data System Prototype for Remote Access to Information Based on Confidential Data. Journal of Official Statistics, 14, 347–360.

Winkler (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. Statistical Data Protection (SDP'98) Proceedings, IOS Press, Luxembourg, forthcoming.