Obtaining Information while Preserving Privacy: A Markov Perturbation Method for Tabular Data[1]

George T. Duncan[2] and Stephen E. Fienberg[3]
Carnegie Mellon University, Pittsburgh, Pennsylvania USA

Abstract. Preserving privacy appears to conflict with providing information. Statistical information can, however, be provided while preserving a specified level of confidentiality protection. The general approach is to provide disclosure-limited data that maximizes its statistical utility subject to confidentiality constraints. Disclosure limitation based on Markov chain methods that respect the underlying uncertainty in real data is examined. For use with categorical data tables a method called Markov perturbation is proposed as an extension of the PRAM method of Kooiman, Willenborg, and Gouweleeuw (1997). Markov perturbation allows cross-classified marginal totals to be maintained and promises to provide more information than the commonly used cell suppression technique.

Key Words: Confidentiality; Data Access; Data Security; Hierarchical Models; Markov Chains; Perturbation Methods; Simulated Data.

1. Introduction

Information organizations must resolve the tension between demands by data users for access and demands by data subjects and providers for privacy and confidentiality (Duncan, Jabine, and de Wolf 1993, Fienberg 1994). In doing so the information organizations have two fundamental tools: (1) *restricting access*, i.e., implementing administrative policies to limit access to the data and (2) *restricting data*, i.e., providing access to data that have been transformed to reduce the risk of disclosure of individual attributes of data subjects (Duncan 1995). This paper deals exclusively with the second set of tools, disclosure limitation methods. We develop disclosure-limitation procedures to maximize the extent of statistical information subject to confidentiality constraints. This constrained optimization perspective is consistent with the mission of information organizations, which must rest solidly on the twin pillars of data access and confidentiality protection. Our perspective is depicted in Figure 1. The

---

[1] Presented at Statistical Data Protection '98, Eurostat, 27 March 1998, Lisbon, Portugal. A preliminary version of this paper was presented at the Annual Meeting of the American Statistical Association, Anaheim, CA, 1997 August 11. The authors thank two anonymous reviewers for comments, and Larry Cox, John Engberg, and Gordon Sande for comments and suggestions.
[2] Professor of Statistics, Heinz School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA 15213-3890, 412-268-2172, gd17+@andrew.cmu.edu.
[3] Maurice Falk University Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, 412-268-2723, fienberg@stat.cmu.edu.

knowledge of a data user is assessed on two dimensions: (1) the horizontal axis is the level of knowledge about the legitimate object of empirical inquiry; (2) the vertical axis is the level of knowledge about the confidential item.

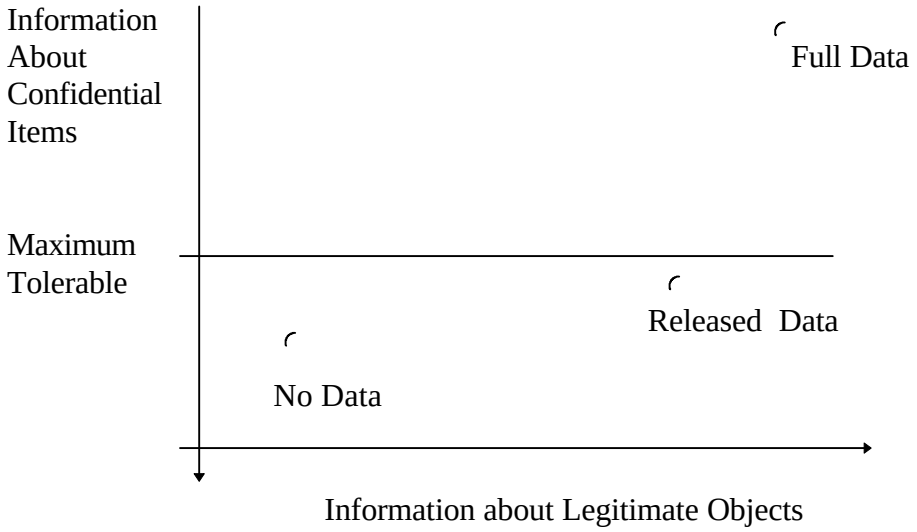


Figure 1. Characterization of Data User's Knowledge

Data release entails disclosure risk. We release disclosure-limited data so this disclosure risk is below some maximum tolerable level. With reference to Figure 1, our goal is to move the information about confidential items in the released data below the maximum tolerable level while losing little information about legitimate objects.

To illustrate our thinking, we focus on the most common data product of statistical agencies, the table. It is typically a cross-classification with the cell entries being frequency counts or other aggregate statistics. Since data users find microdata more useful than aggregations for many purposes, we note that our results have implications beyond tables to categorical microdata. The tabular structure emphasizes cross-classifications in the data. Hence this structure highlights the value of having disclosure-limitation procedures that respect certain marginal summaries, such as row and column totals.

2. Disclosure Limitation Through Cell Suppression

In practice disclosure risk is lowered in the release of categorical data through cell suppression (e.g., Carvalho, Delbert, and Osorio 1994; Cox 1980, 1995; Kelly, Golden, and Assad 1992). A *primary suppression* is made if a cell count is too low or if a few data subjects represent too much of a cell value aggregate. To avoid undue information loss, the information organization releases marginal totals even for attribute combinations in which cells have been suppressed. They must then make *complementary* or *secondary suppressions* so that the primary cell suppression values cannot be recovered from the margins and the other released values. Consider, for example, the table for employment in the SIC code 2052, Cookies and Crackers, as published in the Bureau of Labor Statistics publication, *Employment and Wages Annual Averages, 1995* (Bureau of Labor Statistics 1996). Suppressed are employment figures for 31 states, including Colorado, Louisiana, Maryland, and Massachusetts. We argue that cell suppression suffers from three fundamental problems: (1) it loses too much information, (2) it is analytically difficult, and (3) it can lead to misleading inferences.

Consider first the problem of information loss. Bureau of Labor Statistics (1996) gives U.S. total figures for "cookies and crackers" employment as 52,543 employees. Data for so many states are suppressed, however, that it is impossible to calculate total employment for the six-state Census region of New England. Every state value is suppressed (except for Connecticut's which is listed as 0)!

To further illustrate this difficulty with cell suppression we examine an aggregated version of Table 1 from Fienberg (1997a).

Table 1: Two-way cross-classification of race and income for a selected U.S. census tract. (Source: Fienberg 1997a from 1990 census public use microdata files)

Race/Income Level	**≤$10,000**	**>$10,000 and ≤25,000**	**>$25,000**	**Total**
White	282	199	212	**693**
Black	21	14	9	**44**
Chinese	1	2	2	**5**
Total	**304**	**215**	**223**	**742**

According to common cell suppression criteria, all three cell entries for Chinese would be primary cell suppressions. Thus if these data were in fact a cross-tabulation of the population of the census tract, this data release would be a confidentiality violation. In fact, it is a sample from this population. Since the data contains only a small fraction of the total population, the chance of a specific individual being included in the data is equally small. For purposes of our exposition we will treat the data of Table 1 as if they were population data.

In data dissemination with the cells for Chinese as primary suppressions, the three cells for Black would also need to be suppressed. Because their values could be recovered from the marginal values, they would be complementary cell suppressions. Thus using cell suppressions (denoted by n) Table 1 would be replaced by Table 2 in dissemination to data users, with substantial loss of statistical information. All income distribution data are lost for both the small Chinese population and the not so small Black population.

Table 2: Two-way cross-classification of race and income for a selected U.S. census tract. Cell suppressions. (Source: Fienberg 1997a from 1990 census public use microdata files)

Race/Income Level	**≤$10,000**	**>$10,000 and ≤$25,000**	**>$25,000**	**Total**
White	282	199	212	**693**
Black	n	n	n	**44**
Chinese	n	n	n	**5**
Total	**304**	**215**	**223**	**742**

Some obvious alternatives to cell suppression also have deficiencies because of equally serious loss of statistical information. If the table was collapsed to combine Black and Chinese (global recoding in the terminology of ARGUS (Willenborg and de Waal 1996)), we lose the marginal information on the relative proportions of Black and Chinese. If we eliminate the Chinese category entirely, we lose the marginal information on the total number of Chinese in the census tract.

The second problem is the analytical difficulty of determining an appropriate set of secondary suppressions. For tables of dimension frequently encountered in practice this determination is a computationally daunting problem of combinatorial optimization (Cox

1995), although Fishchetti and Salazar (1996) have made the problem seem tractable at least for relatively low-dimensional tables. Computational approaches for multi-way tables that exploit relevant statistical theory as well as modern operations research methodology are still noticeable by their absence.

The third problem is that cell suppression can lead to misleading inferences. Table 2 can be misleading to the data user about the income distribution of Blacks and Chinese. A relatively naïve data user might take a cognitive clue from the array of numbers for the White population and think that Blacks and Chinese follow roughly the same pattern. Thus the naïve data user might fill in across the income cells the values of (18, 13, 13) for Blacks and (2, 1, 2) for Chinese. These values fail to reflect the reality of fewer Blacks in the highest income category and more Chinese. A somewhat more sophisticated data user might use both the row totals and the column totals to obtain "guesses" of the suppressed cell values. So, for the "Black-Under $10,000" cell, the estimate could be (44/49)(304-282) = 19.8. Rounding and extending to the other missing cells, we get (20, 14, 10) for Blacks and (2, 2, 1) for Chinese. These "guesses" move in the right direction, that is, toward the true values but not as far as they should. More generally, an iterative proportional fitting algorithm or raking (as it is known in some circles) could be used to obtain cell entries that are consistent with given marginal totals (Bishop, Fienberg, and Holland 1975), but since the primary suppressions are purposely chosen, the iterative proportional fitting algorithm may be far from optimal.

3. Two Alternatives to Cell-Suppression: Hierarchical Models and Data-Conditioned Models

Cell suppression suffers from information loss, computational difficulty, and user imputation through misleading patterns in the released data. We seek better alternatives according to a different paradigm. Our objective is to provide additional information to data users without compromising confidentiality protection. In doing this we work within a paradigm that has been labeled according to emphasis as perturbation methods (Fienberg 1997a), data swapping (Dalenius and Reiss 1982, Zayatz 1997), or simulated data methods (Rubin 1993). Two basic approaches are possible: one is a hierarchical approach and the other is a data-conditioned approach.

In the hierarchical approach the original data is viewed as a realization of a probabilistic process from some population. The mechanism for this view is apparent in the case where the data are a sample. If instead the data form a population, as in a complete census, this view suggests a super-population. Statistical inference then informs us about the parameters of this higher-level population. The hierarchical approach is conceptually a comfortable one for statistical inference, but it requires careful model specification. The structures of all probability distributions need to be specified. In this context a natural and flexible approach would be to use a log-linear model.

The data-conditioned approach is simpler and is the focus of this present work. This approach takes the original data, itself, as the focus of statistical attention. Certainly this is the case for a statistical analysis whose basic purpose is descriptive rather than inferential. Further it avoids having to make shaky model specifications. It has a further advantage. The disseminating information organization need not pre-specify the nature of statistical analyses that data users may choose to perform--at least not to the extent that they would in the hierarchical model approach. When a table cannot be released because of disclosure risk, we propose that the information organization release instead one or more tables that are stochastically perturbed from the original table. All tables whose release would be problematical because of confidentiality concerns would be replaced by tables randomly generated from a probability distribution.

Our approach is analogous to, but operationally distinct from, adding noise to a continuous microdata value, say X. In many applications it is reasonable to take X as the sum of a number of (independent or dependent) random variables. The release of $X+\varepsilon$ where ε is

mean zero noise just continues this process by adding another increment to an already noisy X. Noise addition perturbs X further in an unbiased way. A table count entry Y can be viewed similarly. Take it as the sum of indicator variables, each flagging the classification of an entity into that cell cross-classification. In typical large-scale data there are misclassifications, that is an entity ought to be identified in one cell and an error is inadvertently made by classifying that entity into another cell. Our approach is to stochastically move cell entry values in the table. Such moves introduce perturbations in the form of misclassifications. This is an error mechanism for which statistical analysis methods exist. We focus on perturbations that leave the expected value of each cell entry unchanged. This stationarity is, in purpose, akin to adding noise with mean zero to the continuous microdata value X. We will make use of *local moves*—data swapping through sequences of moves of one observation from one cell to another (Glonek 1987). To see most easily the structure of our approach we begin with an idea called PRAM.

4. PRAM

PRAM stands for Post Randomization Method. With PRAM, Kooiman, Willenborg, and Gouweleeuw (1997) (also see Gouweleeuw, Kooiman, Willenborg, and de Wolf (1997)), followed a line of suggestions for the use of randomized response techniques for confidentiality protection going back to Warner (1965) (e.g., see Särndal, Swensson, and Wretman (1992)). In PRAM, instead of randomizing the presentation of a *question*, the *responses* (so post-questioning) are randomized. Applying this idea to categorical variables, they move entities independently among categories. This is similar to the method of data swapping (Dalenius and Reiss 1982), but in PRAM the moves are according to a Markov chain. The idea is in the spirit of fixed-data perturbation as proposed by Adam and Wortman (1989). This probabilistic perturbing of the original responses makes it hard for a data snooper to have confidence in making inferences about confidential information. The important advantage of PRAM over the usual implementations of data swapping is that it has a sound statistical basis. The data user is assured of an appropriate mode of analysis.

We initiate our extension of PRAM with an elementary dichotomy, say employed or not. From this starting point we show that we can deal systematically with a variety of constrained structures for the release of tabular data. For an elementary dichotomy, take the fractions employed and not in the original data as **p** = (r, 1-r). The data distribution is invariant if the 2×2 transition probability matrix **P** is chosen so that **pP = p**. It is easy to show that the class of Markov operators that satisfy this fixed-point equation has the structure,

$$\mathbf{P} = \begin{bmatrix} 1-q & q \\ \frac{1-r}{r}q & 1-\frac{r}{1-r}q \end{bmatrix}, \text{ where } 0 \le q \le \min\left[\tfrac{1-r}{r},\ 1\right].$$

This form is consistent with the invariant formulation suggested by Gouweleeuw, Kooiman, Willenborg, and de Wolf (1997), but in this 2×2 case the formulation fully characterizes all solutions **P** to **pP = p**. Note that θ = 0 corresponds to an identity transformation, the obvious solution to the equation and one that trivially gives no perturbation. The parameter θ can be adjusted upward to give a higher probability of perturbation while maintaining stationarity. This approach can be extended to perturb an arbitrary m-state classification by applying it to each of the $\binom{m}{2}$ pairs of variables in some sequence.

This structure has an appealing interpretation in terms of misclassification error. Take the data to comprise a total of n individuals. Let each individual move independently according to the Markov matrix **P**, with nr individuals beginning in the employed state and n(1-r) individuals beginning in the unemployed state. After one stage of moves by the n individuals, some of them may be misclassified with respect to the classification of the original data

(whether or not that original classification is in fact “correct”). Chen (1979) reviews randomized response schemes as purposive misclassifications. He examines how log-linear models can be used to estimate the underlying true values, and this work is directly relevant to estimation for the type of perturbations explored in this paper.

As Kooiman, Willenborg, and Gouweleeuw (1997) demonstrate, PRAM can be applied to multi-way tables, say one that is I×J×K. In its simplest form this just involves stringing the IJK = m, say, elements in a stacked vector and proceeding as before with an m-state classification. It involves nothing new. Tables do however have a structure, most especially their marginal distributions. The information organization may wish not to perturb certain such marginal distributions. As Denise Lievesley has suggested, one reason for this desire is that aggregated statistics of this form are often routinely released prior to consideration of dissemination of more refined data. Statistical agencies want to maintain consistency in their data releases. Maintenance of marginal totals, again as Kooiman, Willenborg, and Gouweleeuw (1997) show, is not difficult if the set of variables is classified into two categories, one set to be perturbed and one set not. For example in an I⊗J table with constraints just on the row totals, we can use I separate PRAM processes each on the J states in their respective rows. They do not, however, treat the more interesting and difficult case of a genuine cross-classification where marginal distributions are to be maintained. It is to this problem that we turn our attention.

5. Maintaining Margins in Cross-Classified Tables

Consider a two-way cross-classification in which we permit only perturbations that leave both row and column totals fixed. Our interpretation of the basic approach involves the movement of an entity from one cell to another. Think about a movement between rows in one column. This has no effect on column totals but will change row totals unless there is an equal and opposite movement in another column. Thus in the case of cross-classified constraints, moves must be *coupled*. This coupling is consistent with the more general Gröbner basis structure laid out in Diaconis and Sturmfels (1998):

$$\begin{matrix} + & - \\ - & + \end{matrix} \quad \text{or} \quad \begin{matrix} - & + \\ + & - \end{matrix}$$

In a graphical representation it is consistent with data flows corresponding to an *alternating cycle*, as discussed by Cox (1987), with the following two possible structures:

Note that each vertex has in-degree 2 or out-degree 2. There are two possible states; the switch is shown by reversing the direction of each arc.

This idea motivates expanding our elementary dichotomy to an *elementary data square*, which is a 2×2 matrix. For example, in Table 1 one elementary data square is the upper-left block: $\begin{matrix} 282 & 199 \\ 11 & 14 \end{matrix}$

The potential movers are of two-classes: those we label M_1 from the low-income White group and those we label M_2 from the low-income Black group. The entries in the elementary data square are then classified into those that *must* stay and those that *may* move. Including the coupled moves, we can display this as:

$$\begin{bmatrix} 268, 14M_1 & 178, 21M_2 \\ 21M_2 & 14M_1 \end{bmatrix}$$

After one Markov move, the number of individuals in the upper left cell is a random variable X with the following form:

$$X = 268 + Binomial(14, 1-a) + Binomial(21, \tfrac{r}{1-r}a), \text{ where } r = \tfrac{282}{303}.$$

The number Y in the lower left cell is a random variable Y = 303 - X. To extend to a whole two-way, say I×J, table we apply the method to a random sequence of the $\binom{I-1}{2}\binom{J-1}{2}$ possible elementary data squares.

6. Data Squares are Limiting

A special case exists when one more of the cell entries is a zero. For example, if there is a zero in the (2, 1)-cell, the stationary probability vector is p = (1, 0) which is required for unbiasedness. The transition probability matrix **P** must then be the identity, so there is no perturbation. This discussion assumes that zero cell entries should not be perturbed. Note that with a cell zero in a 2×2 table, row and column totals determine all cell entries. Since movements just among data squares get blocked by cell entries of zero, it is worthwhile to explore other perturbation patterns. This leads to more general alternating cycles (see Cox 1987). For example, consider the following 3x3 table:

0	2	1	**3**
3	0	1	**4**
4	2	0	**6**
7	**4**	**2**	**13**

The alternating cycle prompted by a movement from cell (1, 2) to cell (1, 3) looks like this:

0	2	1	**3**
3	0	1	**4**
4	2	0	**6**
7	**4**	**2**	**13**

This forms an alternating cycle of length 6. Again, there are two possible transition patterns; reversing the direction of any one arc reverses them all. Thus, as with the basic data square, moves can be modeled with a two-state Markov chain.

One might wonder what graphical form would be generated by a diagonal move, say from the (1, 2)-cell to the (2, 1)-cell. This is not a row-column (or column-row) move. Substantively, it would typically be a less plausible misclassification than a move along a single row or single column. The generated graph has the following form:

0 2 1 **3**

3	0	1	**4**
4	2	0	**6**
7	**4**	**2**	**13**

Interestingly, this graph results in exactly the same cell-entry changes as the original alternating cycle. It can be shown that this holds in general, that is any alternating cycle that involves diagonal moves can be replaced with an equivalent alternating cycle involving only row-column (column-row) moves.

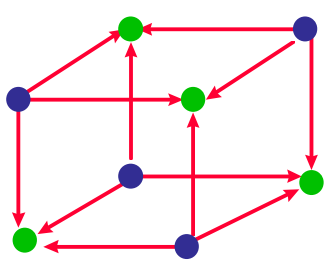
7. Three-Dimensional Tables

Now let's take up Markov perturbation for a three-way, say I×J×K, table. Consider the case where all two-way marginal totals are fixed. Rees and Duke-Williams (1997) note the problem that arises in directly extending the Cox (1987) alternating cycle notion from two to three dimensions. A single strand path with alternating signs through a three-way table need not have the appropriate compensation on the third dimension.

For three-way tables, the coupling is among three moves, rather than among two moves as it was for two-way tables. We make use of an *elementary data cube* for which the coupled moves have the structure

+	−	−	+
−	+	+	−

over the two layers. This is the twinning of alternating cycle paths that is suggested by Rees and Oliver-Williams (1997; Figure 18). It is equivalent to a particular graph over this cube:



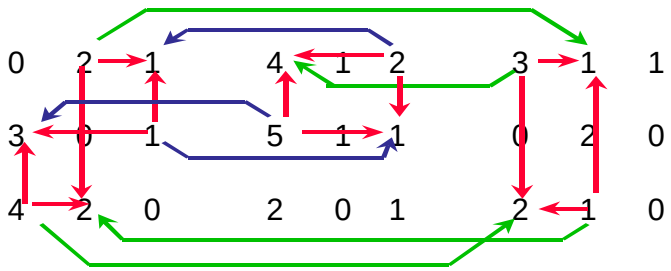
In this graph, each vertex has in-degree 3 or out-degree 3. Again, there are two possible states. The second state occurs with a reversal of all edge directions.

Generalizing alternating cycles in this way to the 3-dimensional case, we see that the graph must be an alternating cycle (possibly degenerate) in each plane; just to have an alternating cycle in 3 dimensions is inadequate. This can be hard to achieve with various patterns of zeroes in the table. There are times, however, when perturbation that respects both zeroes and marginal totals can be achieved. Consider the following non-trivial 3×3×3 example:

0	2	1	4	1	2	3	1	1
3	0	1	5	1	1	0	2	0

4 2 0 2 0 1 2 1 0

A 3-dimensional graph that would allow perturbation has the following structure:



In an M-way table with M>2, maintaining all M-1 dimensional projections (subtotals) can place substantial restrictions on the cell entries. Indeed such constraints may uniquely specify these cell entries (Willenborg and de Waal 1996, p. 130). In such a case, to allow perturbation it is necessary to relax the constraints. This can be done by specifying that each of the marginal totals must lie within some specified range about the actual data totals. Stochastic perturbation is then possible by first choosing a feasible set of marginal totals and then proceeding as outlined previously. It is possible in this context that the use of Gröbner bases lets us finesse some of these problems, but it also leads to computational complexities of an order of magnitude as great as for cell suppression.

8. Discussion and Conclusions

In this paper we have focused largely on attempts to preserve the "unbiasedness" of the post randomization perturbations, especially with regard to zero cell entries. As we indicated in the previous section the presence of zeros then forces zero entries to remain zero. For the large sparse tables that occur in practice this may be unwise. Consider for example an M-way table where M=100 or even 200. Even if we have data based on a survey of thousands of households or establishments, the full cross-classification will be full of cells with zeros and ones in them. Preserving the zero entries in such circumstances will often lead to the non-existence of a Markov-like perturbation of the sort we explore in this paper. A far more sensible way to view such tables is to consider "exchanging" the zeros and ones using similar Markov like procedures, subject to marginal constraints. The unbiasedness will then come for the lower-order margins.

The Markov perturbation method gives workable disclosure limitation for categorical data. Importantly, it applies to data where structure as cross-classified marginal totals is to be maintained. It provides more information to the legitimate data user than cell suppression. Like all disclosure limitation methods, it (1) requires time and attention to implement and perform the technique, and (2) raises questions of consistency when related tables are released, perhaps at different times. The noise deliberately introduced through the Markov perturbation method is comfortably interpretable as misclassification error. This is error that experienced data analysts know how to handle. The stationarity of the method eases statistical analysis.

References

[1] Adam, N. R. and Wortman, J. C. (1989) Security-control methods for statistical databases: a comparative study. ACM Computing Surveys, Volume 21, No. 4, 555-556.

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

[3] Bureau of Labor Statistics (1996) *Employment and Wages, Annual Averages, 1995,* U. S. Department of Labor, Washington, D. C.

[4] Carvalho, F. D., Dellaert, N., and Osorio, M. S. (1994) Statistical disclosure in two-dimensional tables: Positive tables. *Journal of the American Statistical Association*, ***89***, 1547-1557.

[5] Chen, T. Timothy (1979) Analysis of randomized response as purposively misclassified data. *Proceedings of the Section on Survey Research Methods.* American Statistical Association 158-163.

[6] Cox, Lawrence H. (1980) Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, **75**, 377-385.

[7] Cox, Lawrence H. (1987) A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, **82**, 520-524.

[8] Cox, Lawrence H. (1995) Network models for complementary cell suppression, *Journal of the American Statistical Association*, **90** 1453-1462.

[9] Dalenius, T. and Reiss, S. P. (1982) Data-swapping: a technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73-85.

[10] Diaconis, Persi and Sturmfels, Bernd (1998) Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, **26**, 363-397.

[11] Diaconis, Persi and Gangolli, A. (1995) Rectangular arrays with fixed margins. In D. Aldous, P. Diaconis, J. Spencer, and J. M. Steele, eds. *Discrete Probability Algorithms.* Springer-Verlag, New York, 15-41.

[12] Duncan, George T. and Lambert, Diane (1986) Disclosure-limited data dissemination. *Journal of the American Statistical Association*, **81**, 10-28.

[13] Duncan, George T., Jabine, Thomas B., and de Wolf, Virginia A. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Panel on Confidentiality and Data Access, Committee on National Statistics, National Academy Press, Washington, DC.

[14] Duncan, George T. (1995) Restricted data versus restricted access: A perspective from *Private Lives and Public Policies*. In *Seminar on New Directions in Statistical Methodology*, Statistical Working Paper No. 23, Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC, Part 1, pp 43-56.

[15] Fienberg, Stephen E. (1994) Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, **10**, 115-132.

[16] Fienberg, Stephen E. (1997a) Confidentiality and disclosure limitation methodology: challenges for national statistics and statistical research. Paper commissioned by the Committee on National Statistics for presentation at its 25th anniversary meeting on 1997 February 21.

[17] Fienberg, Stephen E. (1997b) Notes on generating the exact distribution for a contingency table given its marginal totals. Unpublished manuscript.

[18] 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 Bureau of the Census 1996 Annual Research Conference.* U.S. Bureau of the Census, Washington, D.C. 87-105.

[19] Fischetti, M. and Salazar, J. J. (1996) Models and algorithms for the cell suppression problem. *Proceedings of Third International Seminar on Confidentiality,* Bled, Slovenia, 114-122.

[20] Glonek, G. (1987) *Some Aspects of Log Linear Models* Ph.D. Thesis, School of Mathematical Sciences, Flinders University of South Australia.

[21] Gouweleeuw, J. M., Kooiman, P., Willenborg, L.C.R.J., and de Wolf, P.-P. (1997) Post randomization for statistical disclosure control: theory and implementation, Research Paper No. 9731, Statistics Netherlands, Division Research and Development, Department of Statistical Methods.

[22] Kelly, T., Golden, S., and Assad, P. (1990) Cell suppression disclosure protection for sensitive tabular data, *Networks* **22** 397-417.

[23] Kooiman, Peter, Willenborg, Leon, and Gouweleeuw, Jose (1997) PRAM: a method for disclosure limitation of microdata. Statistics Netherlands. Division Research and Development, Department of Statistical Methods.

[24] Patefield, W. M. (1981) An efficient method of generating random R⟐C tables with given row and column totals. *Applied Statistics*, **30,** 91-95.

[25] Rees, P. H. and Duke-Williams, O. (1997) Methods for estimating missing data on migrants in the 1991 British census. *International Journal of Population Geography*, **3**, 323-368.

[26] Rubin, Donald B. (1996) Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. *Journal of Official Statistics*, **9**, 461-468.

[27] Särndal, C. E., Swensson, B., and Wretman, J. (1992) *Model Assisted Survey Sampling.* Springer Verlag, New York.

[28] Warner, S. L. (1965) Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.

[29] Willenborg, Leon and de Waal, Ton (1996) *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics 111. Springer Verlag, New York.

[30] Zayatz, Laura V. (1997) Disclosure limitation for the 2000 census of housing and population. Unpublished manuscript, U.S. Bureau of the Census, Washington, DC.