Bibliography

The purpose of this bibliography is to update the references on disclosure limitation methodology that were cited in **Working Paper 2**. Much has been written since **Working Paper 2** was published in 1978. Subcommittee members reviewed papers dealing with methodological issues which appeared after 1980 and prepared the abstracts in this bibliography. **An asterisk (*) indicates that the document has been specifically referenced in this report.**

In the Federal statistical system the Bureau of Census has been the leading agency for conducting research into statistical disclosure limitation methods. The Census Bureau staff has been very active in publishing the results of their research and has been well represented on the Subcommittee. For these reasons the statistical disclosure limitation research that has been sponsored by the Bureau of the Census is thoroughly and adequately covered in this bibliography. In addition the Subcommittee tried to include important papers which either describe new methodology or summarize important research questions in the areas of disclosure limitation for tables of magnitude data, tables of frequency data and microdata.

Within the past two years statistical disclosure limitation research has been highlighted in publications from Western Europe. An international Seminar on Statistical Confidentiality was held in September, 1992, in Dublin, Ireland. The seminar was organized by Eurostat (Statistical Office of the European Community) and ISI (International Statistical Institute). The papers were published in Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute. In addition, a special issue of Statistica Neerlandica, Vol. 46, No. 1, 1992 was dedicated to disclosure limitation. Selected papers from these sources are included in the annotated bibliography.

In 1993, a special issue of the Journal of Official Statistics, Vol. 9, No. 2 was dedicated to disclosure limitation. That issue contains the papers which were presented at a workshop sponsored by the Panel on Confidentiality and Data Access, of the Committee on National Statistics. The panel report was published later in 1993. It is entitled Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, by Duncan et. al., and was published by the Committee on National Statistics and the Social Science Research Council, National Academy Press, Washington, DC. The panel report and selected papers from the special issue of the Journal of Official Statistics are included in the bibliography.

Areas of potential applicability which are not covered in this bibliography include mathematical methods of matching a microdata file to an outside file. A good summary of the state-of-the-art in exact matching as of 1985 can be found in "Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies", Dept of Treasury, IRS, SOI, Publication

1299 (2-86). A more recent reference is a special issue of Survey Methodology, Volume 19, Number 1, 1993.

The Subcommittee on Disclosure Limitation Methodology would like to thank the following people who contributed to the annotated bibliography: Robert Burton, National Center for Education Statistics; Russell Hudson, Social Security Administration; Dorothy Wellington, Environmental Protection Agency.

Bibliography on Methodology for Disclosure Limitation

*Adam, N. R. and Wortmann, J. C., "Security-control Methods for Statistical Databases: A Comparative Study," ACM Computing Surveys, Vol 21, No. 4, pp. 515-556 (Dec. 1989).

The authors carefully define and explain the various problems inherent in disclosure control of on-line systems. Proposed solutions, along with their strengths and weaknesses are discussed. This paper is written on a tutorial level, the purpose being to educate the reader in the current methods. Security control methods are classified into four approaches: conceptual, query restriction, data perturbation and output perturbation. Methods based on these approaches are compared. Promising methods for protecting dynamic-online statistical databases are presented.

*Alexander, L. B., and Jabine, T. B. (1978), "Access to Social Security Microdata Files for Research and Statistical Purposes," Social Security Bulletin, Vol. 41, No. 8, pp. 3-17.

This article focuses on the characteristics of SSA microdata files and on the development of a disclosure policy aimed at serving the public interest while protecting the privacy of individuals and the confidentiality of research and statistical information. Several dimensions of the disclosure question are explored. The factors controlling the decision whether or not to release microdata are also discussed. Some particular practices are described to illustrate application of present policy principles.

*Barabba, V. P. and Kaplan, D. L. (1975), "U. S. Census Bureau Statistical Techniques to Prevent Disclosure -- The Right to Privacy vs. the Need to Know," paper read at the 40th session of the International Statistical Institute, Warsaw.

*Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," Journal of the American Statistical Association, Vol. 85, pp. 38-45.

A general overview of disclosure risk in the release of microdata is presented. Topics discussed are population uniqueness, sample uniqueness, subpopulation uniqueness and disclosure protection procedures such as adding noise, data swapping, micro aggregation, rounding and collapsing. One conclusion reached by the authors is that it is very difficult to protect a data set from disclosure because of the possible use of matching procedures. Their view is that the data should be released to users with legal restrictions which preclude the use of matching.

Blien, U., Wirth, H., and Muller, M. (1992), "Disclosure Risk for Microdata Stemming from Official Statistics," Statistica Neerlandica, Vol. 46, No. 1, pp. 69-82.

Empirical results from matching a portion of a microdata file on individuals against a reference file are presented. Two matching strategies are considered: a simple (or exact) procedure, and a procedure developed by Paass and based on the use of discriminant function analysis. It is concluded that exact matching is a poor strategy and would not be

used by an astute data snooper. It is further concluded that Paass' procedure is also deficient, on the grounds that, among the "identifications" that it yields, only a small proportion are true identifications, and the snooper does not know which matches are correct. One criticism of this is that the authors do not consider the possibility that a "small proportion" of correct matches may be too many. This weakens their overall conclusion that disclosure in microdata files may be a less serious problem than it is typically thought to be.

Bowden, R. J. and Sim, A. B. (1992), "The Privacy Bootstrap," Journal of Business and Economic Statistics, Vol. 10, No. 3, pp. 337-345.

The authors describe a method of masking microdata by adding noise. The noise is generated by bootstrapping from the original empirical distribution of the data. The technique is analyzed in terms of protection gained and statistical efficiency lost, and it is compared with the technique of adding random noise.

Burton, R. (1990), "Strategies to Ensure the Confidentiality of Data Collected by the National Center for Education Statistics," unpublished manuscript presented to the Washington Statistical Society.

The author discusses the confidentiality problems that arise with microdata files containing information from complex national surveys, the steps taken at NCES to minimize disclosure risk from these files, and the testing procedures that are used to assess risk. Since education data bases are typically hierarchical, disclosure risk centers on identification of schools, rather than on direct identification of individuals. It is therefore necessary to recode school data, which, at NCES, takes the form of converting continuous data to categorical.

The recoded files are tested by matching against publicly available reference files, using the "nearest neighbor" concept, which is defined in terms of Euclidean distance. The author discusses the difficulty of developing a strict criterion of "sufficiently safe" in this context, and presents the criteria that NCES has used.

Caudill, C. (1990), "A Federal Agency Looks for Answers to Data Sharing/Confidentiality Issues," presented at the annual meeting of the American Agricultural Economics Association, Vancouver, British Columbia.

NASS has a clear set of publication rules. No reports can be released which may reveal information about individual operations. NASS will only issue information which is based on reports of three or more operations. Also, data will not be released if one operation accounts for 60 percent or more of a total. These publication rules often mean that geographic subdivisions must be combined to avoid revealing information about individual operations. Data for many counties cannot be published for some crop and livestock items and State level data must be suppressed in other situations.

If only a few operations are present in a particular universe, such as hatcheries in a State or firms holding cold storage supplies of specific commodities, it may not be possible to publish totals at all. NASS can publish data in such cases only if a signed waiver is received under which an operation accounting for more than 60 percent of a total agrees to allow data to be published. If waivers cannot be obtained, data are not published. If waivers are obtained, the waivers are reviewed and signed periodically to be sure that cooperation has not changed.

Causey, B., Cox, L. H., and Ernst, L. R. (1985), "Application of Transportation Theory to Statistical Problems," Journal of the American Statistical Association, 80, 392, pp. 903-909.

This paper demonstrates that the transportation theory that solves the two-dimensional (zero-restricted) controlled rounding problem, Cox and Ernst (1982), can be used for other statistical problems: 1) general statistical problems which involve replacing nonintegers by integers in tabular arrays (eg. iterative proportional fitting or raking); 2) controlled selection for a sample; and 3) sample selection to maximize the overlap between old and new primary sampling units after a sample redesign. The paper mentions that the controlled rounding of a two-way table can be used to prevent statistical disclosure in a microdata release (by replacing the true value by an appropriately rounded value). The paper also provides a simple example that shows that the three-way controlled rounding problem does not always have a solution.

*Cecil, J. S. (1993), "Confidentiality Legislation and the United States Federal Statistical System," Journal of Official Statistics, Vol. 9, No. 2, pp. 519-535.

Access to records, both statistical and administrative, maintained by federal agencies in the United States is governed by a complex web of federal statutes. The author provides some detail concerning the Privacy Act of 1974, which applies to all agencies, and the laws which apply specifically to the U. S. Bureau of Census, the National Center for Education Statistics and the National Center for Health Statistics. The author also describes ways these agencies have made data available to researchers.

Cigrang, M. and Rainwater, L. (1990), "Balancing Data Access and Data Protection: the Luxembourg Income Study Experience" Proceedings of the Statistical Computing Section, American Statistical Association, Alexandria, VA, pp. 24-26.

Details of a computer system allowing access to multi-national microdata files are presented. Data protection is achieved through use of a security system based on user identification, passwords, output control, operating system safeguards, and review of both job requests and output. The authors do not address analytical issues.

Cox, L. H. (1979), "Confidentiality Problems in Microdata Release," Proceedings of the Third Annual Symposium on Computer Applications in Medical Care, IEEE Computer Society, pp. 397-402.

The examples of disclosure avoidance techniques given and views expressed were drawn from Statistical Policy Working Paper 2: Report on Disclosure-Avoidance and Disclosure Avoidance Techniques. This was followed by Cox's observations. He points out that there are no generally accepted and quantifiable notions of the degree of disclosure or the degree of protection. Thus, there is no concept of sensitivity of microdata upon which the necessary protective techniques must be defined. It is difficult to accurately measure the degree to which a technique reduces the sensitivity of a microdata set without first quantifying the notion of sensitive data. He suggested empirical research into quantifying the concept of sensitivity, simulating likely privacy invasion tactics and engaging in related cost-benefit analyses. For theoretical research he suggested casting the problem in terms of data base theory, which he claims includes data base security, multidimensional transformation and data swapping.

One interesting thing about this paper is that although some of the research projects have been tried, the same questions remain and many of the same protection techniques are still used.

Cox, L. H. (1980), "Suppression Methodology and Statistical Disclosure Control," Journal of the American Statistical Association, Vol. 75, pp. 377-385.

This article highlights the interrelationships between the processes of disclosure definitions, subproblem construction, complementary cell suppression, and validation of the results. It introduces the application of linear programming (transportation theory) to complementary suppression analysis and validation. It presents a mathematical algorithm for minimizing the total number of complementary suppressions along rows and columns in two dimensional statistical tables. This method formed the basis of an automated system for disclosure control used by the Census Bureau in the 1977 and 1982 Economic Censuses.

In a census or major survey, the typically large number of tabulation cells and linear relations between them necessitate partitioning a single disclosure problem into a well-defined sequence of inter-related subproblems. Over suppression can be minimized and processing efficiency maintained if the cell suppression and validation processes are first performed on the highest level aggregations and successively on the lower level aggregates. This approach was implemented in a data base environment in an automated disclosure control system for the 1977 U.S. Economic Censuses.

The paper gives an example of a table with 2 or more suppressed cells in each row and column, where the value of the sensitive cell can be determined exactly, as an example of the need for validation.

*Cox, L. H. (1981), "Linear Sensitivity Measures in Statistical Disclosure Control," Journal of Statistical Planning and Inference, Vol.5, pp. 153-164.

Through analysis of important sensitivity criteria such as concentration rules, linear sensitivity measures are seen to arise naturally from practical definitions of statistical disclosure. This paper provides a quantitative condition for determining whether a particular linear sensitivity measure is subadditive. This is a basis on which to accept or reject proposed disclosure definitions. Restricting attention to subadditive linear sensitivity measures leads to well-defined techniques of complementary suppression.

This paper presents the mathematical basis for claiming that any linear suppression rule used for disclose rule must be "subadditive". It gives as examples the n-k rule, the pq rule, and the p percent rule and discusses the question of sensitivity of cell unions. It provides bounding arguments for evaluating (in special cases) whether a candidate complementary cell might protect a sensitive cell.

Cox, L. H. (1983), "Some Mathematical Problems Arising from Confidentiality Concerns," Essays in Honour of Tore E. Dalenius Statistical Review, Vol. 21, Number 5, Statistics Sweden, pp. 179-189.

This is a nice summary of disclosure problems in tables, both of frequency data and magnitude data. Included are discussions of the definition of a sensitive cell and the mathematics involved in selection of cells for complementary suppression and controlled rounding.

Cox, L. H. (1984), "Disclosure Control Methods for Frequency Count Data," presented at the Census Advisory Committee Meeting of the American Statistical Association, Bureau of the Census.

Four methods for controlling statistical disclosure in frequency count data are discussed along with their pros and cons: cell suppression, random perturbation, random rounding and controlled rounding. Each method is viable for single 2-way tables. With some effort, cell suppression can be applied consistently between sets of tables but creates problems for data use. Because the other methods distort every value somewhat, they avoid abbreviating detail and do not produce seemingly arbitrary and possible cumbersome patterns of data suppression. Among these other methods, only controlled rounding meets all of the following objectives: additivity, unbiasedness and reducing data distortion. The paper recommends research concerning the extent to which various methods, particularly controlled rounding can be applied consistently between tables.

This paper defines disclosure in frequency count data to occur when one can infer with certainty that the number of respondents is less than a predetermined threshold. Most other references say that disclosure occurs when the number of respondents is less than the predetermined threshold.

Cox, L. H. (1987a), "New Results in Disclosure Avoidance for Tabulations," International Statistical Institute-Proceedings of the 46th Session: Contributed Papers, Tokyo, pp. 83-84.

For two-way tables, this paper considers the three standard disclosure avoidance procedures, suppression, perturbation and rounding in a single mathematical framework. The two unifying formulations mentioned are the use of alternating cycles and network optimization models. Alternating cycles are described in more detail in Cox (1987b). Network optimization models are described in more detail in Cox (1992).

Cox, L. H. (1987b), "A Constructive Procedure for Unbiased Controlled Rounding," Journal of the American Statistical Association, Vol. 82, June 1987, pp. 520-524.

A constructive algorithm for achieving zero-restricted unbiased controlled rounding, simple enough to be implemented by hand is presented. The procedure is based on adjustments in alternating cycles of cells in an array. Gives a counterexample to the existence of unbiased controlled rounding in three dimensional tables. Cox's solution also allows one to perform random data perturbation in a way that assures additivity.

Cox, L. H. (1991), "Comment," a comment on Duncan, G. and Pearson, R., "Enhancing Access to Microdata while protecting Confidentiality: Prospects for the Future," Statistical Science, No. 6, pp. 232-234.

This is an initial formulation of matrix masking, which is described more completely in Cox (1993b).

Cox, L. H. (1992), "Solving Confidentiality Protection Problems in Tabulations Using Network Optimization: A Network Model for Cell Suppression in U.S. Economic Censuses." Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, Dublin, pp. 229-245.

Mathematical issues in confidentiality protection for aggregate economic statistics are discussed. A network model of the minimum-cell cell suppression algorithm of Cox (1980) is presented and the development of network models that combine minimum-cell and minimum-value-suppressed criteria are discussed.

Cox, L. H. (1993a), "Network Models for Complementary Cell Suppression", unpublished manuscript.

Complementary cell suppression is a method for protecting data pertaining to individual respondents from statistical disclosure when the data are presented in statistical tables. Several mathematical methods to perform complementary cell suppression have been proposed in the statistical literature, some of which have been implemented in large-scale statistical data processing environments. Each proposed method has limitations either theoretically or computationally. This paper presents solutions to the complementary cell suppression problem based on linear optimization over a mathematical network. these

methods are shown to be optimal for certain problems and to offer several theoretical and practical advantages, including tractability and computational efficiency.

Cox, L. H. (1993b), "Matrix Masking methods for Disclosure Limitation in Microdata," unpublished manuscript.

The statistical literature contains many methods for disclosure limitation in microdata. However, their use and understanding of their properties and effects has been limited. For purposes of furthering education, research, and use of these methods, and facilitating their evaluation, comparison, implementation and quality assurance, it would be desirable to formulate them within a single framework. A framework called "matrix masking"-- based on ordinary matrix arithmetic--is presented, and explicit matrix mask formulations are given for the principal microdata disclosure limitation methods in current use. This enables improved understanding and implementation of these methods by statistical agencies and other practitioners.

Cox, L. H. (1994), "Protecting Confidentiality in Establishment Surveys," in Survey Methods for Businesses, Farms and Institutions, Brenda Cox (ed.), John Wiley and Sons, NY.

This paper focuses on the issue of disclosure limitation in tables of establishment data, namely cell suppression and some of the mathematical issues associated with determining "optimal" patterns of complementary cell suppressions. The methods used by the U. S. Census Bureau and Statistics Canada, current research by the Census Bureau, and problems associated with microdata files for establishment data are described.

Cox, L. H. and Ernst, L. R. (1982), "Controlled Rounding," INFOR, Vol 20, No. 4, pp. 423-432. Reprinted: Some Recent Advances in the Theory, Computation and Application of Network Flow Methods, University of Toronto Press, 1983, pp. 139-148.)

This paper demonstrates that a solution to the (zero-restricted) controlled rounding problem in two-way tables always exists. The solution is based on a capacitated transportation problem.

Cox, L. H., Fagan, J. T., Greenberg, B., and Hemmig, R. (1986), "Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data, "Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 388-393.

Results obtained by the Census Bureau's confidentiality staff in its research in disclosure avoidance methods for publicly released tabular data are described. The paper reports new procedures (based on network theory) developed for rounding, perturbation, and cell suppression in two-dimensional tables, with a focus on their common underlying structure. The goal is to develop unbiased procedures which maintain additivity and alter marginals as infrequently as feasible.

The common underlying structure considered in this paper is using circuits in a graph, referred to as "alternating cycles" in Cox (1987b). This paper describes the approach and illustrates its use for unbiased controlled rounding, unbiased controlled perturbation, unbiased restricted controlled perturbation, auditing protection in a suppression scheme, and selecting cells for complementary suppression.

Cox, L. H. and George, J. A. (1989), "Controlled Rounding for Tables with Subtotals," Annals of Operations Research, 20 (1989) pp. 141-157.

Controlled rounding in two-way tables, Cox and Ernst (1982), is extended to two-way tables with subtotal constraints. The paper notes that these methods can be viewed as providing unbiased solutions. The method used is a capacitated network (transshipment) formulation. The solution is exact with row or column subtotals. It is demonstrated that the network solution with both row and column subtotal constraints is additive, but that it may fail zero-restricted constraints and may leave grand-totals of the subtables uncontrolled for the adjacency condition. An example is given of a table for which no zero-restricted controlled rounding exists.

*Cox, L. H., Johnson, B., McDonald, S., Nelson, D. and Vazquez, V. (1985), "Confidentiality Issues at the Census Bureau," Proceeding of the Bureau of the Census First Annual Research Conference, Bureau of the Census, Washington D. C., pp. 199-218. (Revised: Cox, L. H., McDonald, S. K. and Nelson, D. (1986), "Confidentiality Issues of the U.S. Bureau of the Census," Journal of Official Statistics, 2, 2, pp. 135-160.)

This paper summarizes confidentiality issues and presents a fair amount of detail in selected areas such as methods applied to tables of frequency counts. U. S. Census Bureau studies in the early 1980's pointed out the need for an integrated program of research and development work to contribute to key decisions on important policy issues. This paper presents the major ideas that were raised in these studies and merit further attention as opportunities for research, and highlights research in progress.

*Cox, L. H. and Kim, J. (1991), "Concept Paper: Thwarting unique identification in Microdata Files. A Proposal for Research," unpublished manuscript.

This paper proposes a disturbance technique involving super and subsampling. The original data are sampled with replacement to create a file larger than the intended microdata file. Differential probabilities of selection are used for the unique records in the original data set, and record weights are adjusted. This larger file is then subsampled to create the final microdata file.

Cox, L. H. and Zayatz, L. (1993), "Setting an Agenda for Research in the Federal Statistical System: Needs for Statistical Disclosure Limitation Procedures," Proceedings of the Section on Government Statistics, American Statistical Association.

The authors describe the confidentiality protection problem for different types of data, summarize confidentiality protection techniques in current use, and present an agenda for future research in confidentiality protection.

Dalenius, T. (1981), "A Simple Procedure for Controlled Rounding," Statistisk Tidskrift, No. 3, pp. 202-208.

Disclosure control in frequency tables is discussed and available methods are summarized. Dalenius proposes a method of rounding only part of a table, which assures that the rounded table preserve the marginal totals. The cells to be rounded include the sensitive cells and selected other cells to assure that the marginal totals are unchanged. Selecting the additional cells for the "rounding" may be akin to selecting cells for complementary suppression.

Dalenius, T. (1982), "Disclosure Control of Magnitude Data," Statistisk Tidskrift, No. 3, pp. 173-175.

Discusses disclosure control in tables of magnitudes, where cells are determined to be sensitive either because there are too few respondents, or because they fail the (n,k) rule. The approach is similar to that described in Dalenius (1981) in that for one sensitive cell, three additional cells are selected to complete a rectangle. Then random rounding is applied to the counts in four cells, and the magnitude to be published is calculated based on the adjusted number of respondents and the assumption that each respondent has the average volume. The new aggregates are unbiased estimates for the originals.

Dalenius, T. (1986), "Finding a Needle in a Haystack or Identifying Anonymous Census Records," Journal of Official Statistics, Vol. 2, pp. 329-336.

The author presents two variants of a simple method for identifying the unique records in microdata. The first variant involves three different approaches to sorting the data to identify unique records. In the second variant he considers two types of transformation of the data and shows how sorting can identify the unique records under either transformation. A cost function to determine which variant to use is based on number of variables with data in the public domain and computer memory necessary to perform that variant. The identified data can be protected by destroying the unique records in the public domain or by blocking out some but not all of the data, by data perturbation (replacing original data with different values for one or more variables), or by encryption (a reversible transformation of the data.)

Dalenius, T. (1988), Controlling Invasion of Privacy in Surveys, Department of Development and Research, Statistics Sweden.

This book discusses many problems associated with protecting against invasion of privacy in surveys. It was intended as a text for a course on the subject, and includes many examples from Europe and the U.S. Included are chapters on the basic concept and legal framework, safeguards provided as a result of agency codes, professional codes and informed consent. Other chapters discuss safeguards provided by sampling, measurement methods (including randomized response), and transformations. These are followed by a discussion of safeguards taken during the data processing and safeguards for use in the release of statistics (publication of tables and microdata). These chapters on release are the most applicable for this bibliography.

Dalenius defines disclosure and provides examples. He describes disclosure control for tables of counts to involve cell suppression, changing the classification scheme, perturbation and rounding. For tables of magnitude data disclosure control may be based on cell suppression, changing the classification scheme or perturbation. He discuses release of low order moments, release through a data-base and queries of the data base. Finally he describes release of microdata. Protective measures for microdata include deidentification, sampling, placing a restriction on population size, reduction in detail, adding noise to the data, removing well known data subjects, suppression, data-swapping, and transformations.

The book concludes with a discussion of the safeguards involved in the closing operations of a survey, including deidentification of records, file-splitting, taking action on unique vectors, and encryption. The epilogue summarizes what is ahead and includes a discussion of research ideas, most of which do not deal with the methodological issues of release.

Dalenius, T. (1993), "Safeguarding Privacy in Surveys at the Turn of the Century," unpublished manuscript.

This memo discusses the change in public perception concerning invasion of privacy with surveys and considers what survey statisticians can do to counteract concerns by survey respondents. The assumption is made that the sources of public concern in the past may be present in the next few years, but with different forces. This assumption makes it necessary to identify the key sources in the past likely to have a force in the future that cannot be neglected.

Dalenius, T. (1993), "Disclosure Control of Microdata using Data Shifting," unpublished manuscript.

This memo proposes "data shifting" as a way to limit disclosure in microdata. Data shifting is related to, but not the same as, "data swapping".

Dalenius, T. and Denning, D. E. (1982), "A Hybrid Scheme for Release of Statistics," Statistisk Tidskrift, Vol 2, pp. 97-102.

For population survey data this paper proposes for the release of data a scheme that is a hybrid of microstatistics and macrostatistics (tables and summary statistics). A specified set of low-order finite moments of the variables are computed and released, allowing users to compute the low-order statistics corresponding to their needs. They consider the computational feasibility of doing this and discuss the protection implied. They also observe that users would not be able to calculate even simple moments for subgroups of the respondents (eg. all females.) The authors balance the feasibility of giving higher order moments against the increased amount of computation needed as well as the increased risk of disclosure.

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

The data-swapping technique proposed by the authors can be used on categorical data to produce microdata and to release statistical tabulations while protecting confidentiality. The raw data matrix is converted to a new matrix by rearranging entries in such a way that the marginals up to a specified order of cross-tabulation are unaffected and the desired order of statistics is preserved. The authors illustrate mathematically how the data base presented only in terms of t-order statistics is unlikely to be compromised. An appended comment points out that this approach is applicable to data from individual respondents with relatively few categorical responses for each data item. This technique can be used to both produce microdata and release statistical tabulations so that confidentiality is not violated. This is the technique which has been used by the U. S. Census Bureau as part of the Confidentiality Edit. The Confidentiality Edit was used to protect data tables from the 1990 census.

*Denning, D. E. (1982), Cryptography and Data Security, Addison-Wesley, Reading, MA.

This is **the** standard book addressing computer security issues. The book is quite rigorous and very thorough in coverage, including cryptography and transmission issues as well as data-base security. Statistical databases are covered in depth, incorporating much of the author's previous work. The topics covered still provide the basis for understanding more recent work in this area.

DeSilets, L., Golden, B., Kumar, R., and Wang, Q. (1992), "A Neural Network Model for Cell Suppression of Tabular Data," College of Business and Management, University of Maryland, College Park, MD.

For three-dimensional tables, the objective is to select cells for complementary suppression which minimize the total value suppressed but assure that the sensitive cells are protected to within pre-specified tolerance levels. A neural network is trained on the solutions from the heuristic, network based model described in Kelly et al. (1992). Thus, the neural

network can be used on a new problem to quickly identify a good starting solution for the more general optimization method. This paper provides detail on the neural network design and training. Results are promising. Run time of the network is minimal once the network is trained. The trained neural network was able to match about 80% of the cell suppression solution for a new table.

Duncan, G. T. (1990), "Inferential Disclosure-Limited Microdata Dissemination," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 440-445.

Duncan discusses the various forms of disclosure such as complete identification of a record (identity disclosure), obtaining a good approximation of confidential data (attribute disclosure) and the more general inferential disclosure in which the release of the microdata data allows a user to make more accurate estimates of confidential information. This paper also presents a method to measure the risk of disclosure in terms of unauthorized information gained when microdata are released. This method can then be used to measure the effectiveness of data masking techniques.

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

The authors briefly summarize the legal aspects of maintaining the confidentiality of records, in particular they site various United States laws. The most important part of this paper deals with a general disclosure limiting approach that quantifies the extent of statistical disclosure by means of an uncertainty function applied to predictive distributions.

Duncan, G. T. and Lambert, D. (1987), "The Risk of Disclosure for Microdata," Proceedings of the Bureau of the Census Third Annual Research Conference, Bureau of the Census, Washington, DC.

Various types of disclosure are discussed, including identity and attribute disclosure. The authors then present a model to estimate the risk of disclosure that can take into account the user's prior knowledge and also the type of masking technique that has been used. The model presented uses predictive distributions and loss functions. Using this model they show that sampling and the including of simulated artificial records can reduce the disclosure risk.

Duncan, G. T., Jabine, T. B. and de Wolf, V. A., eds. (1993), Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, Committee on national Statistics and the Social Science Research Council, National Academy Press, Washington, DC.

This is the final report of the Panel on Confidentiality and Data Access, which was jointly sponsored by the Committee on National Statistics of the National Research Council and the Social Science Research Council. The Panel's charge was to develop recommendations that could help federal statistical agencies to protect the confidentiality of data subjects and,

at the same time, facilitate responsible dissemination of data to users. Chapter 6, "Technical and Administrative Procedures," covers statistical disclosure limitation methodology and administrative procedures for restricting access to data. The chapter includes recommendations on both topics.

Duncan, G. T. and Pearson, R. W. (1991), "Enhancing Access to the Microdata While Protecting Confidentiality: Prospects for the Future" (with comment), Statistical Science, Vol. 6, No. 3, pp. 219-239.

Methods on increasing data access while assuring an acceptable level of protection are discussed, including statistical masking, electronic gatekeepers, licensing contracts, punitive damages for improper use of data and researchers code of ethics. The authors also suggest that respondents to data collection procedures should be informed that there is a remote risk of re-identification of their responses.

Ernst, L., (1989), "Further Applications of Linear Programming to Sampling Problems," Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA, pp. 625-630.

Cox and Ernst (1982) demonstrated that a controlled rounding exists for every two-dimensional additive table. Here, the author establishes by means of a counter-example, that the natural generalization of their result to three dimensions does not hold. However, a rounding does always exist under less restrictive conditions.

Fagan, J. T., Greenberg, B. V. and Hemmig, R., (1988), "Controlled Rounding of Three Dimensional Tables," Bureau of the Census, SRD Research Report No: Census/SRD/RR-88/02

A heuristic procedure for finding controlled roundings of three dimensional tables is presented. The three-dimensional controlled rounding problem is much more difficult than its two-dimensional counterpart. The solution to the two dimensional problem involves representing the table as a system of linear equations, formulating a network flow problem, modeling the system of equations, finding a saturated flow through the network and interpreting the flow as a controlled rounding of the original table. In three dimensions, the system of linear equations cannot be represented as a single network. The heuristic model discussed in this paper employs a sequence of network flow problems; the solution of each reduces the size of the table to be rounded. The sequence of solutions is then used to attempt to extract a controlled rounding of the original table, if one exists. An alternative approach to the problem is in Kelly, Golden, Assad and Baker (1988).

Fienberg, S. (1993), "Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality," Technical Report #577, Department of Statistics, Carnegie Mellon University. An earlier version of this paper was published in Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, 1992.

This paper examines the conflicts between the two perspectives of data access and confidentiality protection and briefly outlines some of the issues involved from the perspectives of governments, statistical agencies, other large-scale gatherers of data, and individual researchers.

Fuller, W.A. (1991), "Masking Procedures for Disclosure Limitation," Journal of Official Statistics, Vol. 9, No. 2, pp. 383-406.

In this study Fuller focuses on masking through the addition of noise to the characteristics, or to a subset of the characteristics, by data switching or imputation methods. He illustrates how an "intruder" can estimate the characteristics of a particular individual, and by what probability, under different characteristics of the microdata and different masking procedures. Fuller maintains that masking error can be treated as measurement error. He outlines a method Sullivan (1989) developed for adding measurement error to the variables of a data set in which the masking procedure is designed to maintain the marginal distribution functions and the covariance structure of the original data. Fuller then applies measurement error procedures to the masked data to construct consistent estimators of regression parameters and other higher order statistics.

*Gates, G. W. (1988), "Census Bureau Microdata: Providing Useful Research Data while Protecting the Anonymity of Respondents," Presented at the annual meeting of the American Statistical Association, New Orleans, LA.

Gates describes some solutions used by the Census Bureau to provide microdata for public use while controlling disclosure risk. These include: thorough review of the files to evaluate risk of individual identification; research on the methodological evaluation of various masking techniques; microaggregations; and remote access whereby users submit computer programs to be run by authorized staff. He also lists administrative solutions such as surveys that are reimbursable rather than protective, special sworn employees, and programs for which all research must be done on site at the Census Bureau. Gates also lists various legal options for dealing with the problem.

*George, J. A. and Penny, R. N. (1987), "Initial Experience in Implementing Controlled Rounding for Confidentiality Control, "Proceedings of the Bureau of the Census Third Annual Research Conference, Bureau of the Census, Washington DC., pp. 253-262.

The New Zealand Bureau of Statistics has been using random rounding. This paper documents a study of controlled rounding to offset the disadvantages of random rounding: (1) that the published values in the rows and columns of the table to not necessarily add to the published marginal totals, and (2) the procedure would not result in consistent

random roundings if applied to the same table at different times. They use the methodology for controlled rounding in two dimensional tables with subtotal constraints described in the paper by Cox and George (1989).

This paper describes the capacitated transshipment formulation for controlled rounding in terms of a network formed from nodes and arcs, with flows created by the nodes and relayed along the arcs. The network is defined to be capacitated when there are non-zero upper and/or lower limits on the flow along some or all of the arcs. The search for a controlled rounding solution becomes the search for a solution to a network that has an integer flow in every arc. The authors describe their implementation of this search using the SAS/OR Network module, but state that most commercial Mathematical Programming systems will solve the problem. The effect of different types of tables are considered as well as the difficulties encountered the implementation.

The authors point to the need for further theoretical work on controlled rounding for multi-dimensional tables and tables with other complex structures, given the advantage of controlled over random rounding in preserving the additivity of table totals.

Govoni, J. P. and Waite, P. J. (1985), "Development of a Public Use File for Manufacturing," Proceedings of the Section on Business and Economic Statistics, American Statistical Association, Alexandria, VA, pp. 300-302.

A procedure for producing a public use data product for the Longitudinal Establishment Data file (LED) is described. The procedure involves sorting on value of shipments within 4-digit SIC code, and then aggregating 3 or more establishments at a time to form pseudo-establishments. The exact extent of aggregation depends upon the (n,k) rules that would be used in publishing tabular data for the same data set.

Testing led to the conclusion that the resulting file was disclosure-free. There is, however, no description of the testing method, other than the statement that testing involved matching to the original data file. In terms of utility of the public use file, the authors noted that correlations were increased by aggregation, but that relative relationships seemed to be preserved.

Greenberg, B. (1985), "Notes on Confidentiality Issues when Releasing Survey or Census Data," presented at the Conference on Access to Public Data sponsored by the Social Science Research Council.

The author notes that currently there is no measure of disclosure risk for microdata files and explains the need for such a measure. Techniques for reducing the disclosure risk of a microdata file such as top-coding, coding into ranges, and limiting geographic detail are discussed. The majority of the paper describes the role and the review process of the Microdata Review Panel.

*Greenberg, B. (1986), "Designing a Disclosure Avoidance Methodology for the 1990 Decennial Censuses," presented at the 1990 Census Data Products Fall Conference, Arlington, VA.

The Census Bureau's objective data release strategy is to maximize the level of user statistical information provided subject to the condition that pledges of confidentiality are not violated. A Confidentiality Staff has been established at the Census Bureau to develop disclosure avoidance methods for use in Census products, most prominently the 1990 Decennial Censuses data products. This paper describes procedures developed, their impact, and how they relate of the Bureau's goals. The two types of procedures described in the paper for reducing disclosure risk in the release of tabular data are suppression (primary and complementary) and noise introduction (controlled rounding and controlled perturbation). The paper concludes that controlled rounding appears to be the preferred method for use on the 1990 Decennial Census data products. However, it was not used (see Griffin, et. al.).

Greenberg, B. (1988a), "An Alternative Formulation of Controlled Rounding," Statistical Research Division Report Series, Census/SRD/RR-88/01, Bureau of the Census, Washington, DC.

The standard definition of controlled rounding is extended to allow a non-zero multiple of the base to decrease as well as increase. Greenberg compares the formulations of the standard and of this extended version of controlled rounding. He shows that, in their respective solutions, the underlying networks differ only with respect to arcs and to costs. The paper gives step-by-step examples of each procedure and contrasts their performances. The procedures developed by the author minimize a measure of closeness-of-fit to provide solutions to either of the controlled rounding definitions. Greenberg asserts that the new definition and its solution process also can be applied to tables of more than two dimensions and refers to Fagan, Greenberg, and Hemmig (1988).

Greenberg, B. (1988b), "Disclosure Avoidance Research at the Census Bureau," Presented at the Joint Advisory Committee Meeting, April 13-14, Oxon Hill, MD.

Greenberg discusses research in a) improving complementary cell suppression procedures for economic tabular data, b) assessing risk inherent in public use demographic microdata, c) design of data release and masking strategies for demographic public use microdata, and d) design of data release and masking for economic microdata.

Greenberg, B. (1988c), "Disclosure Avoidance Research for Economic Data," Presented at the Joint Advisory Committee Meeting, October 13-14, Oxon Hill, MD.

The primary method of releasing economic data by the Census Bureau is through cross-classified tables of aggregate amounts. This report discusses the basic disclosure avoidance methodology employed for tabular data. Even though economic microdata is not systematically released by the Census Bureau, they also report on research into methods for the design of surrogate economic microdata files. This report outlines the

complementary suppression problem, with examples and briefly discusses the use of network theory.

Greenberg, B. (1990a), "Disclosure Avoidance Research at the Census Bureau," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 144-166.

The confidentiality staff at Census is involved in diverse projects, including; (a) improving cell suppression procedures for economic tabular data, (b) assessing risk inherent in public use demographic microdata, (c) development and implementation of data masking schemes for demographic public use microdata, and (d) design of data release and masking strategies for economic microdata. The author discusses these projects focusing on objectives, progress to date, current activities, and future work.

Greenberg, B. (1990b), "Disclosure Avoidance Practices at the Census Bureau," presented at the Seminar on Quality of Federal Statistics sponsored by the Council of Professional Associations on Federal Statistics, Washington, DC.

A data collection agency has the obligation to release as much information to the public as possible while adhering to pledges of confidentiality given to respondents. The author discusses the trade-offs between the completeness and the accuracy of microdata and tabular data. For microdata this reduces to releasing fewer variables and collapsing categories versus adding noise to the data or to trading completeness for one data attribute at the expense of completeness for another. For tabular data one either suppresses information and collapses categories or introduces noise. Both these actions can be thought of as data masking. The first option reduces completeness, while the second option reduces accuracy.

Greenberg, B. and Voshell, L. (1990a), "Relating Risk of Disclosure for Microdata and Geographic Area Size," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 450-455.

This paper examines the percent of records on a microdata file that represent individuals or households with a unique combination of characteristics variables. In particular, the authors describe the relationship between the percent of population uniques on a file from a specific geographic region and the size of that region.

Greenberg, B. and Voshell, L. (1990b), "The Geographic Component of Disclosure Risk for Microdata," Statistical Research Division Report Series, Census/SRD/RR-90/12, Bureau of the Census, Statistical Research Division, Washington, DC.

The relationship between the percent of population uniques on a microdata file from a specific geographic region and the size of that region is described. The authors also introduce the idea of using random subsets of microdata records to simulate geographic subsets of microdata records.

Greenberg, B. and Zayatz, L. (1992), "Strategies for Measuring Risk in Public Use Microdata Files," Statistica Neerlandica, Vol. 46, No. 1, pp. 33-48.

Methods of reducing the risk of disclosure for microdata files and factors which diminish the ability to link files and to obtain correct matches are described. Two methods of estimating the percent of population uniques on a microdata file are explained. A measure of relative risk for a microdata file based on the notion of entropy is introduced.

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

This paper presents the 1990 Census disclosure avoidance procedures for 100 percent and sample data and the effects on the data. The Census Bureau's objective is to maximize the level of useful statistical information provided subject to the condition that confidentiality is not violated. Three types of procedures for 100 percent data have been investigated: suppression, controlled rounding, and confidentiality edit. Advantages and disadvantages of each are discussed. Confidentiality Edit is based on selecting a small sample of census households from the internal census data files and interchanging their data with other households that have identical characteristics on a set of selected key variables. For the census sample data, the sampling provides adequate protection except in small blocks. A blanking and imputation-based methodology is proposed to reduce the risk of disclosure in small blocks.

*Jabine, T. B. (1993a), "Procedures for Restricted Data Access," Journal of Official Statistics, Vol. 9, No. 2, pp. 537-589.

Statistical agencies have two main options for protecting the confidentiality of the data they release. One is to restrict the data through the use of statistical disclosure limitation procedures. The other is to impose conditions on who may have access, for what purpose, at what locations, and so forth. For the second option, the term **restricted access** is used. This paper is a summary of restricted access procedures that U. S. statistical agencies use to make data available to other statistical agencies and to other organizations and individuals. Included are many examples which illustrate both successful modes and procedures for providing access, and failures to gain the desired access.

Jabine, T. B. (1993b), "Statistical Disclosure Limitation Practices of United States Statistical Agencies," Journal of Official Statistics, Vol 9., No. 2, pp. 427-454.

One of the topics examined by the Panel on Confidentiality and Data Access of the Committee on National Statistics of the National Academy of Sciences was the use of statistical disclosure limitation procedures to limit the risk of disclosure of individual information when data are released by Federal statistical agencies in tabular or microdata formats. To assist the Panel in its review, the author prepared a summary of the disclosure

limitation procedures that were being used by the agencies in early 1991. This paper is an updated version of that summary.

Jewett, R. (1993), "Disclosure Analysis for the 1992 Economic Census," unpublished manuscript, Economic Programming Division, Bureau of Census, Washington, DC.

The author describes in detail the network flow methodology used for cell suppression for the 1992 Economic Censuses. The programs used in the disclosure system and their inputs and outputs are also described.

Jones, D. H. and Adam, N. R. (1989), "Disclosure Avoidance Using the Bootstrap and Other Re-sampling Schemes, "Proceedings of the Bureau of the Census Fifth Annual Research Conference," Bureau of the Census, Washington, DC., pp. 446-455.

Methods to protect confidentiality from cleverly designed complex sequences of queries are classified under four general approaches: conceptual modeling, query restriction, data perturbation, and output perturbation. The authors present data coding as the basis of a new perturbation method, and also propose an output perturbation approach based on the bootstrap.

Keller, W. J. and Bethlehem, J. G. (1992), "Disclosure Protection of Microdata: Problems and Solutions," Statistica Neerlandica, Vol. 46, No. 1, pp. 5-19.

Identification and disclosure problems related to the release of microdata in the Netherlands are discussed. The authors discuss both population and sample uniqueness. An argument is presented that disclosure avoidance should be achieved by legal provisions and not by reducing the amount or quality of data releases.

Keller-McNulty, S., McNulty, M. S., and Unger, E. A. (1989), "The Protection of Confidential Data," Proceeding of the 21st Symposium on the Interface, American Statistical Association, Alexandria, VA, pp. 215-219.

A broad overview of analytic methods that have been or might be used to protect confidentiality is provided for both microdata files and for tabular releases. Some little-known methods that might be used with microdata, e.g.. "blurring," "slicing," are described. The authors also discuss the need for a standard measure of "control" or protection.

Keller-McNulty, S. and Unger, E., (1993), "Database Systems: Inferential Security," Journal of Official Statistics, Vol. 9, No. 2, pp. 475-499.

The problems of data security and confidentiality have been studied by computer scientists and statisticians. The areas of emphasis within these disciplines on data security are different but not disjoint. One of the main differences is how one views data release. Statisticians have focused on aggregate data release and on single static files of microdata

records. Computer scientists have focused on data release through sequential queries to a database. An initial integrating factor of the two fields is the concept of information stored as a federated database. This paper synthesizes the research done in both of these disciplines and provides an extensive review of the literature. Some basic definitions integrating the two fields are given and data security and confidentiality methodologies studied in both disciplines is discussed.

Kelly, J. P. (1990), "Confidentiality Protection in Two- and Three-Dimensional Tables," Ph.D. Dissertation, University of Maryland, College Park, MD.

Contains proof that the integer programming problem of finding an optimal set of complementary suppressions is "NP-hard"; that is, the number of computations increases (roughly) exponentially with the number of primary suppressions.

Kelly, J. P., Assad, A. A. and Golden, B. L. (1990), "The controlled Rounding Problem: Relaxations and Complexity Issues," OR Spektrum, Springer-Verlag, 12, pp. 129-138.

The three-dimensional controlled rounding problem is described and proved to be NP-complete. For tables where a solution does not exist, a series of relaxations of the zero-restricted problem is described that can lead to solutions. Examples of tables that need various orders of relaxation are given.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1990a), "Cell Suppression: Disclosure Protection for Sensitive Tabular Data," Working Paper MS/S 90-001, College of Business and Management, University of Maryland, College Park, MD.

This paper formulates and develops solution techniques for the problem of selecting cells for complementary suppression in two dimensional tables. (Sensitive cells must not be able to be estimated to within a specified tolerance interval). The authors present a network flow-based heuristic procedure for the complementary suppression problem. The objective function is the minimization of the total of the suppressed values. The authors use the network flow based heuristic procedure currently used by the Census bureau (a sequence of network flow models) to find a feasible solution, then implement a "clean-up" procedure to improve the solution. The paper also develops a lower bounding procedure, which can be used to estimate the quality of the heuristic solution. It can also generate a starting point for the heuristic procedure. Extensive computational results based on real-world and randomly generated tables demonstrate the effectiveness of the heuristic.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1990b), "Cell Suppression Using Sliding Protection Ranges," Working Paper Series MS/S 90-007, College of Business and Management, University of Maryland, College Park, MD.

Complementary suppression cells must be selected to protect those cells identified as primary suppressions. Traditionally, the protection required is defined by an interval centered around the value of each primary suppression. This paper formulates and

develops solution techniques for the problem where sliding protection ranges are allowed; this represents a relaxation of the traditional problem. In this problem, the protection ranges have fixed widths but are free to slide; the only restriction is that they must contain the values of the primary suppressions.

The authors present a network flow-based heuristic for this modified cell suppression problem and use a lower-bounding procedure to evaluate the performance of the heuristic. Extensive computational results based on real-world and randomly generated tables demonstrate that sliding protection ranges can significantly reduce the total amount of suppressed data, as compared to the traditional suppression scheme.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1990c), "A Review of the Controlled Rounding Problem," Proceedings of the 22nd Symposium on the Interface, Interface Foundation of North America, Springer-Verlag, pp. 387-391.

A review of the state of the art in controlled rounding. Notes that three dimensional controlled rounding does not lend itself to a network representation on which to base an effective solution. They quote previous work which demonstrates that the zero restricted controlled rounding problem is NP-Complete. It has also been shown that not every three-way table has a zero-restricted controlled rounding solution. They relax the zero-restricted requirement and discuss their linear programming solution to the relaxed problem (discussed in detail in one of their 1990 papers.) Their procedure (ROUND and BACK) has an advantage in that it either finds a solution or proves that none exists. They also discuss pre-processor heuristics to speed convergence. These are called "Round-Round and Back,""Quick-Round and Back" and "Anneal-Round and Back". The latter appears to be quickest and most successful to date.

Kelly, J. P., Golden, B. L., Assad, A. A. and Baker, E. K. (1988), "Controlled Rounding of Tabular Data," Working Paper MS/S 88-013, College of Business and Management, University of Maryland, College park, MD. (also published in Operations Research, Vol. 38, No. 5, pp. 760-772.)

The authors describe the use of a binary tree search algorithm based on linear programming techniques for solving three-dimensional controlled rounding problems. The algorithm determines whether or not a solution exists, and effectively finds a solution if one does exist. Computational results are presented. A technique for decreasing the running time of the algorithm is also described.

Kim, J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 370-374.

Although noise addition is effective in reducing disclosure risk, it has an adverse affect on any data analysis. If one knows how the data are to be used, transformations of the data before and after the addition of noise can maintain the usefulness of the data. Kim

recommends using linear transformations subject to the constraints that the first and second moments of the new variable are identical to those of the original. He presents the properties of the transformed variable when the variance is known, and when it is estimated. He sets forth the impacts of masking on the regression parameter estimates under different conditions of preserving the first and second moments of the original data.

Kim, J. (1990a), "Masking Microdata for National Opinion Research Center," Final Project Report, Bureau of the Census.

No single masking scheme so far meets the needs of all data users. This article describes the masking scheme used for a specific case of providing microdata to two users that took into account their analytic needs. Since it was done before Kim (1990b), each group was masked separately. In this example the user planned to construct multiple regression models, with the dependent variable of two types - proportions transformed into logits, and medians. Kim discusses 1) whether to add the noise before or after transformation, 2) what distribution of the noise to use, and 3) whether to add correlated or uncorrelated noise. He presents in clear detail the masking process, the statistical properties of the masked variables, and how they satisfied these users' needs. Excellent results were obtained for estimates of the mean and variance/covariance, except when considerable censoring accompanied the logit transformation of the proportions.

Kim, J. (1990b), "Subpopulation Estimation for the Masked Data," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 456-461.

Kim derives estimators for the mean and variance/covariance parameters of a subgroup under different uncorrelated and correlated data conditions when the data set as a whole is masked by additive noise or by additive noise plus transformation. He illustrates the good results using a mixed population data base. He concludes that it is safe to mask the whole data set once and to let the users estimate the mean and variance/covariance of subpopulations using his formulas.

Kumar, F., Golden, B. L. and Assad, A. A. (1992), "Cell Suppression Strategies for Three-Dimensional Tabular Data," Proceedings of the Bureau of the Census 1992 Annual Research Conference, Bureau of the Census, Washington, D. C.

The authors present a cell suppression strategy for three-dimensional tabular data which involves linear programming techniques and heuristic search. The linear programming techniques find a sufficient set of complementary suppressions and a lower bound on the total value that must be suppressed to obtain a sufficient set. If the set obtained has a much higher total value than the lower bound, a heuristic search attempts to find a better solution. An analysis of results is given.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm,," Journal of Official Statistics, Vol. 9, No. 2, pp. 313-331.

The definition of disclosure depends on the context. Sometimes it is enough to violate anonymity. Sometimes sensitive information has to be revealed. Sometimes a disclosure is said to occur even though the information revealed is incorrect. This paper tires to untangle disclosure issues by differentiating between linking a respondent to a record and learning sensitive information from the linking. The extent to which a released record can be linked to a respondent determines disclosure risk; the information revealed when a respondent is linked to a released record determines disclosure harm. There can be harm even if the wrong record is identified or an incorrect sensitive value inferred. In this paper, measures of disclosure risk and harm that reflect what is learned about a respondent are studied, and some implications for data release policies are given.

Little, R. J. A. (1993), "Statistical Analysis of Masked Data," Journal of Official Statistics, Vol. 9, No. 2, pp. 407-426.

A model-based likelihood theory is presented for the analysis of data masked for confidentiality purposes. The theory builds on frameworks for missing data and treatment assignment, and a theory for coarsened data. It distinguishes a model for the masking selection mechanism, which determines which data values are masked, and the masking treatment mechanism, which specifies how the masking is carried out. The framework is applied.

Lougee-Heimer, R. (1989), "Guarantying Confidentiality: The Protection of Tabular Data," Master's Degree Thesis, Department of Mathematical Sciences, Clemson University.

Looks at selection of complimentary cells for three-way tables of magnitude data; describes the Census Bureau's current two-way procedure and demonstrates a three-way procedure based on finding linear dependent sets of vectors in a system of linear equations. (Procedure sequentially fixes complementary suppression cells in stages by solving linear programming subproblems.) The suppression method quoted is to protect each primary cell to within a "tolerance". That is upper and lower bounds are specified (in an undisclosed way) for each primary cell. The problem is to select complimentary cells so that it is impossible to estimate the value of the primary cells more accurately than their tolerance regions.

Lunt, T. F. (1990), "Using Statistics to Track Intruders," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC.

The author presents a detailed account of a real-time intrusion-detection expert system that identifies computer users who exhibit unexpected or suspicious behavior. Security systems of this type, while certainly applicable to disclosure avoidance in microdata files, do not fall into the general class of analytic approaches to confidentiality protection.

McGuckin, R. H., (1992), "Analytic Use of Economic Microdata: A Model for Researcher Access With Confidentiality Protection,", Center for Economic Studies Discussion paper CES 92-8, Bureau of the Census, Washington D. C.

This paper describes the benefits of analytic research with economic microdata and describes the administrative arrangements that have been developed by the Census Bureau to provide access to microdata files by selected researchers who are appointed as special sworn employees of the Census Bureau and work on site at the Center for Economic Studies. The author proposes expansion of facilities for user access, including provision of access at regional centers located in universities or Census Bureau regional offices. He also recommends that the Census Bureau use somewhat broader criteria to decide which research projects are relevant to Census Bureau program needs and therefore meet statutory requirements for this mode of access to the Census Bureau's economic microdata.

McGuckin, R. H. and Nguyen, S. V. (1988), "Use of 'Surrogate Files' to Conduct Economic Studies with Longitudinal Microdata," Proceedings of the Fourth Annual Research Conference, Bureau of the Census, Washington, DC., pp. 193-209.

Essentially same paper as one below.

McGuckin, R. H. and Nguyen, S. V. (1990), "Public Use Microdata: Disclosure and Usefulness," Journal of Economic and Social Measurement, Vol. 16, pp. 19-39.

The authors discuss and compare methods for masking economic microdata for public use data files, given the economic data characteristics of uniqueness of particular information and skewed size distribution of business units. They examine summary statistics methods such as data grouping under the (n,k) rule, and providing the variances, covariances and means of the original data. They also discuss using surrogate files involving stochastic and deterministic data transformations.

The authors address theoretical aspects of various transformations that might be applied to longitudinal datasets in order to protect confidentiality. The focus is on the ability of the transformed dataset to yield unbiased estimates of parameters for economic models. Both stochastic and deterministic transformations are considered, all are found to be flawed in one way or another. The authors conclude that it may be more useful to release variance-covariance matrices than to develop transformed microdata files.

*Michalewicz, Zbigniew (1991), "Security of a Statistical Database," in Statistical and Scientific Data-bases, ed., Ellis Horwood, Ltd.

This article discusses statistical database security, also known as inference control or disclosure control. It is assumed that all data is available in an on-line, as in a micro-data file. A critique of current methods, both query restriction and perturbation, is included using an abstract model of a statistical database. **Tracker** type attacks are extensively discussed. The balance between security and usability is developed, with usability for

query restriction methods being dependent upon the number and ranges of restricted data intervals. Methods of determining these intervals are compared.

Mokken, R. J., Kooiman, P., Pannekoek, J. and Willenborg, L. C. R. J. (1992), "Assessing Disclosure Risks for Microdata," Statistica Neerlandica, Vol. 46, No. 1, pp. 49-67.

The authors provided methods to estimate the probability that the release of a microdata set allows users to obtain confidential data for population unique individuals that are known by the user to exist. This paper covers the situation were there are multiple users of a geographically stratified micro data release that contain multiple identification variables.

Mokken, R. J., Pannekoek, J., and Willenborg, L. C. R. J. (1992), "Microdata and Disclosure Risks," Statistica Neerlandica, Vol 46, No 1.

The authors provide a method to calculate the risk that an investigator is able to re-identify at least one individual in an microdata set. This risk is shown to depend on some variables that are readily controlled by the releasing agency such as the coarseness of the key variables and the size of the subsample that is released.

Mugge, R. H. (1983a), "Issues in Protecting Confidentiality in National Health Statistics," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 592-594.

Statistical programs must be administered in such a way as to bring the best scientific results but also to protect any participating subjects. This paper discusses four particular policy issues as they relate to NCHS.

Public use microdata files are protected by, first, suppressing all direct identifiers, and then, suppressing or categorizing other variables that might lead to identification. The author notes that most NCHS releases cannot be linked to known comparison files, so that it is impossible to test the risk of disclosure. He also notes the small sampling fractions that are used in most NCHS surveys, which he believes to be an additional safeguard against disclosure.

Mugge, R. H. (1983b), "The Microdata Release Program of the National Center for Health Statistics," Statistical Policy Working paper 20, pp. 367-376.

The author presents an overview of the NCHS microdata release procedures, which include suppression of some data elements, but not transformation. He notes that NCHS data typically include enough noise so that purposeful addition of more noise would probably be redundant. No testing program exists at the agency, so there are no measures of the level at which confidentiality is protected. However, as far as is known, there has never been a breech of confidentiality.

*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.

This paper documents some of the findings from a simulation of a data switching procedure. This procedure is one part of a disclosure limitation technique termed the "Confidentiality Edit" that will be used on data from the 1990 Decennial Census prior to forming demographic tables of frequency counts. From this simulation, it was discovered that the data from small blocks needed additional protection. The success of the procedure and its effect on the statistical properties of the data are described.

Paass, G. (1988), "Disclosure Risk and Disclosure Avoidance for Microdata," Journal of Business and Economic Statistics, Vol. 6, pp. 487-500.

Paass gives estimates for the fraction of identifiable records when specific types of outside information may be available to the investigator, this fraction being dependent primarily on the number of variables in common, and the frequency and distribution of the values of these variables. He also discusses the costs involved. Paass then evaluates the performance of disclosure-avoidance measures such as slicing, microaggregations, and recombinations. In an appendix, he presents the technical details of the proposed methods.

Paass, G. (1989), "Stochastic Generation of a Synthetic Sample from Marginal Information," Proceedings of the Bureau of the Census Fifth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 431-445.

In this paper the author describes a stochastic modification algorithm (SMA) used to construct a synthetic sample X from different input sources, the sources being independent samples or summary statistics from an underlying population. The first step in the process is to construct an X as a best fit to the data by a maximum likelihood or minimum cost criterion, and the second step is to generate a sample with a cost value near the minimum which also has maximum entropy. Paass tests his method on income tax data for the German Treasury.

*Qian, X., Stickel, M., Karp, P., Lunt, T. and Garvey, T., "Detection and Elimination of Inference Channels in Multilevel Relational Database Systems," IEEE Symposium on Research in Security and Privacy, Oakland, CA, May 24-26, 1993.

This paper addresses the problem where information from one table may be used to **infer** information contained in another table. it assumes an on-line, relational database system of several tables. The implied solution to the problem is to classify (and thus to deny access to) appropriate data. The advantage of this approach is that such discoveries are made at the **design** time, not execution time. The disadvantage is that the technique only addresses those situations where inferences always hold, not those cases where the inference is dependant upon specific values of data. The technique needs to be investigated for applicability to the disclosure limitation problem.

*Robertson, D. A. (1993), "Cell Suppression at Statistics Canada," Proceedings of the Bureau of the 1993 Census Annual Research Conference, Bureau of the Census, Washington, DC, pp. 107-131.

Statistics Canada has developed Computer software (CONFID) to ensure respondent confidentiality via cell suppression. It assembles tabulation cells from microdata and identifies confidential cells and then selects complementary suppressions. This paper discusses the design and algorithms used and its performance in the 1991 Canadian Census of Agriculture.

Rowe, E. (1991), "Some Considerations in the Use of Linear Networks to Suppress Tabular Data," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 357-362.

The paper discusses network flow theory and its application to finding optimal complementary suppressions in two dimensional tables. The network flow methodology uses closed paths in the table. This analysis considers defining costs to try to assure that the selected path both includes all primary suppressions and minimizes the sum of the suppressed cells (total cost). The paper points out the problems associated with treating one primary cell at a time in terms of finding the "least cost" path.

*Rubin, D. (1993), "Discussion, Statistical Disclosure Limitation," Journal of Official Statistics, Vol. 9, No. 2, pp. 461-468.

Rubin proposes that the government should release only "synthetic data" rather than actual micro-data. The synthetic data would be generated using multiple imputation. They would look like individual reported data and would have the same multivariate statistical properties. However, with this scheme there would be no possibility of disclosure, as no individual data would be released.

Saalfeld, A., Zayatz, L. and Hoel, E. (1992), "Contextual Variables via Geographic Sorting: A Moving Averages Approach," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 691-696.

Social scientists would like to perform spatial analysis on microdata. They want to know relative geographic information about each record such as average income of neighboring individuals. Variables providing this type of information are called "contextual variables." This paper introduces a technique which could generate contextual variables which do not comprise the exact location of respondents. The technique is based on taking moving averages of a sorted data set.

Sande, G. (1984), "Automated Cell Suppression to Preserve Confidentiality of Business Statistics, "Statistical Journal of the United Nations, ECE 2, pp. 33-41.

Discusses in general terms the application of linear programming to complementary suppression. Also outlines the CONFID program developed by Sande at Statistics Canada.

*Singer, E. and Miller, E. (1993), "Recent Research on Confidentiality Issues at the Census Bureau," Proceedings of the Bureau of the Census 1993 Annual Research Conference, Bureau of the Census, Washington, DC, pp. 99-106.

The Census Bureau conducted focus group discussions concerning participants' reactions to the use of administrative records for the Year 2000 Census, their fears concerning confidentiality breaches, their reactions to a set of motivational statements, and ways of reassuring them about the confidentiality of their data. This paper highlights results of these discussions and relates findings from other research in this area.

Skinner, C. J. (1992), "On Identification Disclosure and Prediction Disclosure for Microdata," Statistica Neerlandica, Vol 46, No. 1, pp. 21-32.

Skinner discusses how to estimate the probability of disclosure for two types of disclosure (identification and prediction.) In particular he demonstrates how a Poisson-gamma model can be used to estimate the number of population unique records.

Skinner, C. J. and Holmes, D. J. (1992), "Modelling Population Uniqueness," Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, pp. 175-199.

Authors present various statistical models to be used to estimate the number of population unique records using data collected from a sample from the population. In particular there are examples that demonstrate the effectiveness of a Poisson-lognormal model.

Skinner, C. J., Marsh, C., Openshaw, S., and Wymer, C. (1990), "Disclosure Avoidance for Census Microdata in Great Britain," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC, pp.131-143.

The authors lay out in detail the structural logic of estimating the risk of disclosure and its dependence on factors which can be controlled by those masking methods that preserve the integrity of the data (not by contamination methods). They characterize the type of investigator, the degree of identification, the scenario by which the investigator attempts identification, and the two steps that must be achieved for identification, i.e. i) locate a record that matches the key individual on all the variables common to the microdata and to the additional information file, and ii) infer with some degree of confidence that this record does belong to the target individual.

The authors then summarize the four conditions under which identification is possible as 1) the target individual does appear in the microdata; 2) the common variable values of the target individual are recorded identically in the additional information and the microdata; 3) the combination of common variable values for the target individual is unique in the population; and 4) the investigator infers with some degree of confidence that the combination of common variable values is unique in the population.

The assessment of the probability of each of the four conditions in the context of census microdata is then set forth in some detail, and their product is proposed as the estimate of the risk of disclosure.

Spruill, N. L. (1982), "Measure of Confidentiality," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 260-265.

Spruill uses a monte-carlo procedure to measure the effectiveness of five disclosure avoidance procedures: adding random noise, multiplying by random noise, aggregation, random rounding and data swapping.

Spruill, N. L. (1983), "The Confidentiality and Analytic Usefulness of Masked Business Microdata,"Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 602-607.

This paper presents empirical results on the trade-off between confidentiality protection and data usefulness. Both simulated and real data are used, and several masking procedures are used with each dataset: additive random error, multiplicative random error, grouping, rounding, and data swapping. Confidentiality protection is measured in terms of the proportion of entities that can be correctly linked to a public file.

The results indicate that, when the number of matching variables is small (4 to 6) all masking procedures can be used successfully. When this number is high (20 to 32), masking is much more problematic, although grouping becomes a more attractive procedure. It is noted that the proportion of zeroes in the data set can be an important consideration.

Strudler, M., Oh, H. L. and Scheuren, F. (1986), "Protection of Taxpayer Confidentiality with Respect to the Tax Model," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 375-381.

The Tax Model, a microdata file of a sample of individual taxpayer' returns, is made available to the public. This paper describes the methods used to minimize disclosure risk from the Tax model, and the results of empirical testing of the resulting file. Disclosure risk is reduced by rounding, by "blurring" independently across sensitive variables, and by lowering subsampling rates in the high-income strata. Testing was based on the "Spruill method" of finding best matches by minimizing the sum of absolute deviations,

taken across variables for which it is believed that true data may be known with certainty. Testing indicated that a useful file could be released to the public.

Sullivan, C. M. (1992a), "An Overview of Disclosure Practices," Statistical Research Division Research Report Series, Census/SRD/RR-92/09, Bureau of the Census, Washington, DC.

This paper gives a general definition of the problem of disclosure avoidance for tabular data. The author describes sensitive data, cell suppression, primary suppression rules, cost functions, and feasible ranges for suppressed values.

Sullivan, C. M. (1992), "The Fundamental Principles of a Network Flow Disclosure Avoidance System," Statistical Research Division Research Report Series, Census/SRD/RR-92/10, Bureau of the Census, Washington, DC.

This paper provides a very clear explanation of how to translate the problem of cell suppression in a table into a problem that can be solved with network flow methodology. In particular it explains how to use a network to describe tables which are related additively in one dimension and additive with a hierarchial structure in the other dimension.

Sullivan, C. M. (1993a), "A Comparison of Cell Suppression Methods," ESMD-9301, Economic Statistical Methods Division, Bureau of the Census, Washington, DC.

Two techniques for removing superfluous suppressions when network flow methodology is used to apply complementary suppression are described. Also discussed are problems encountered when applying the techniques to actual economic census data.

Sullivan, C. M. (1993b), "Adjustment Techniques to Supplement a Network Flow Disclosure Avoidance System," Proceedings of the International Conference on Establishment Surveys.

This is an extended version of Sullivan (1993a).

Sullivan, C. and Rowe, E. (1992), "A Data Structure Technique to Facilitate Cell Suppression Strategies," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 685-690.

The authors describe the network flow methodology currently used at the Census Bureau for identifying complementary suppressions for a primary suppression. They then introduce an integer programming technique which may be used after the network flow technique has identified complementary suppressions for all primary suppressions to release superfluous complementary suppressions.

Sullivan, C. and Zayatz, L. (1991), "A Network Flow Disclosure Avoidance System Applied to the Census of Agriculture," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 363-368.

Since rounding and perturbation are unsatisfactory for aggregate magnitude data, the Economic and Agriculture Divisions have always chosen a cell suppression technique to protect aggregate data. The objective in applying complementary suppressions is to ensure the protection of the sensitive data value at minimum cost. Commonly, the original data value that would have appeared in the publication is assigned as the cost. Minimizing the cost incurred through complementary suppression produces a publishable table with maximum data utility; that is, the greatest amount of usable data is provided. (Note: The solutions obtained are not optimal.)

For the 1992 Census of Agriculture, research was conducted on the cell suppression technique using the network flow system of applying complementary suppressions. However, the existing network flow system was not optimal for agricultural data because of the complexity of the data structure and its use of systems of three dimensional tables. Therefore the network flow methodology required customizing. This paper discusses the formulation of the customized network methodology and the limitation encountered with the customized version when applied to agricultural data.

The network-flow methodology for agricultural data was successfully adapted to some extent. However, in its present form, the authors feel it is still unsuitable.

Sullivan, G. R. (1989), "The Use of Added Error to Avoid Disclosure in Microdata Releases," Unpublished Ph.D. Dissertation, Iowa State University, Ames, Iowa.

This paper discusses methods of adding error to observations by means of a masking algorithm that creates a data set that is statistically representative of the original data records in three ways. First, the masked data set should have the same first and second moments as the original data set. Second, the correlation structure of the original and masked data sets should be nearly identical. Third, the univariate distribution functions of the original and masked data should also be the same. The paper also investigates the statistical usefulness of the masked data set by comparing statistical analyses performed on the original and masked data and it evaluates the effectiveness of the mask with regard to disclosure avoidance.

Sullivan, G. R. and Fuller, W. A. (1989), "The Use of Measurement Error to Avoid Disclosure," Proceedings of the Section on Survey Research, American Statistical Association, Alexandria, VA, pp. 802-807.

The authors describe the general technique of adding a random error vector to mask each data vector. On the basis of an approach an intruder would use to construct predictions of confidential variables from conditional probabilities derived from data already known, they illustrate how to select an error covariance matrix for masking normally distributed

data. To find a balance between an error variance large enough to sufficiently lower the probability of matching a record but not to severely distort the data, they illustrate the efficacy of adding vectors of error that have a covariance matrix equal to a multiple of the covariance matrix of the original unmasked data vectors.

Sullivan, G. R. and Fuller, W. A. (1990), "Construction of Masking Error for Categorical Variables," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 435-439.

The authors present a method for masking categorical variables to reduce the risk of attribute disclosure, in which each classification variable is transformed into a Bernoulli variable, and then further transformed into a standard normal variate using the sample univariate distribution functions. The masking is completed by adding a normally distributed error vector to each transformed vector of normalized data. They illustrate how to back-transform the data to the original scale, and then to convert the Bernoulli variables back to their categorical values, and provide an example in terms of its correlation structure.

Tendrick, P. and Matloff, N. S. (1987), "Recent Results on the Noise Addition Method for Database Security," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 406-409.

Adding random noise to sensitive variables maintains expected mean values, but adds bias to other common estimators. The authors investigate bias in percentile ranks and in regression coefficients, and, for the case of regression coefficients, develop methods for eliminating this bias. They also discuss the advantages of adding multivariate noise (when several sensitive variables are involved) that has the same covariance structure as the original data.

Vemulapalli, K. C. and Unger, E. A. (1991), "Output Perturbation Techniques for the Security of Statistical Databases," Proceedings of the 14th National Computer Security Conference, Washington, DC.

This paper addresses the technique of adding "noise" or perturbations to query answers in order to prevent disclosure. The study analyses not only the amount of protection, but also the amount of bias introduced. In particular, the analysis is done for **sum** queries. Ensuring that identical query sets return identical answers is proposed as a solution to compromise by averaging. The favored solutions offer high security while requiring relatively little extra processing time, and so are suitable for on-line systems.

Vogel, F. A. (1992), "Data Sharing: Can We Afford It?," Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, Dublin.

The author discusses several issues that have been raised concerning disclosure, confidentiality, and the sharing of data collected for the National Agricultural Statistics

Service estimating program. There is general agreement within NASS that public benefits can be derived from further analysis of survey data. However, under no circumstances will the preservation of the confidentiality pledge be violated. Special tabulations will be done and on-site analysis can occur only if all confidentiality provisions have been met. Full data sharing does not exist because of NASS supported confidentiality standards.

Voshell, L. (1990), "Constrained Noise for Masking Microdata Records," Statistical Research Division Report Series, Census/SRD/RR-90/04, Statistical Research Division, Bureau of the Census, Washington, DC.

The author presents two algorithms which transform data generated by a random number generator into data satisfying certain constraints on means and variance-covariance structure. Data sets such as these may be beneficial when used for introducing noise in order to mask microdata as a disclosure avoidance technique.

Wang, Q., Sun, X., and Golden, B. L. (1993), "Neural Networks as Optimizers: A Success Story," College of Business and Management, University of Maryland, College Park, MD.

The authors apply a modified, continuous Hopfield neural network to attack the problem of cell suppression. They design an energy function and a learning algorithm to solve two-dimensional suppression problems. The approach is shown to perform well.

Wester, W. C. and Hemmig, R. (1984), "Disclosure Analysis for the Economic Censuses," Proceedings of the Business and Economic Statistics Section, American Statistical Association, Alexandria, VA, pp. 406-409.

Discusses the practical implementation of a complementary disclosure scheme for the complex set of tables published in the Economic Censuses.

Willenborg, L. C. R. J. (1992a), "Disclosure Risk for Microdata Sets: Stratified Populations and Multiple Investigators," Statistica Neerlandica, Vol 46, No 1.

Willenborg discusses the estimation of the risk of disclosure from the release of a geographically stratified microdata set to multiple investigators. The risk of disclosure is high when data is released for small geographic areas because an investigator is very likely to be aware of population uniques for small geographic areas .

Willenborg, L. C. R. J. (1992b), "Remarks on Disclosure Control of Microdata," Statistica Neerlandica, Vol. 46, No. 1.

Willenborg discusses what conditions need to hold so that a computationally easy function can be used to estimate disclosure risk. He also discusses use of subsampling and redefinition of key variables to reduce the risk of disclosure. In addition it is shown how contamination of the original data by adding noise to the key values can reduce the disclosure risk.

Willenborg, L. C. R. J., Mokken, R. J., and Pannekoek, J. (1990), "Microdata and Disclosure Risks," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 167-180.

The authors introduce a measure of the disclosure risk of a microdata file. The measure involves the probabilities that a respondent is in the sample, that the intruder knows information about the respondent, that the information known by the intruder identifies the respondent to be unique in the population, and that the intruder knows that the respondent is unique and finds and recognizes the respondent in the microdata file. A Poisson-Gamma model which can be used to predict uniqueness in the population is described.

Wolf, M. K. (1988), "Microaggregation and Disclosure Avoidance for Economic Establishment Data," Proceedings of the Business and Economic Statistics Section, American Statistical Association, Alexandria, VA.

This paper describes the development of a microaggregate data file. The author evaluates the degree to which microaggregation preserves information contained in the original data. The effect of microaggregation on disclosure risk of a data file is also discussed.

Wright, D. and Ahmed, S. (1990), "Implementing NCES' New Confidentiality Protections," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 446-449.

Public Law 100-297 (Stafford Hawkins Act) imposed new and more stringent confidentiality requirements on NCES. The authors provide an overview of the methods that have been developed and implemented to meet the new requirements, and of the process that led to these methods. With regard to public use tapes, the paper discusses the data masking and testing procedures that were used for various surveys, focusing on the identification of publicly available reference files, on the use of a Euclidean distance measure for matching sample schools to reference schools, and on the problems that arise when school coordinators know the identities of teachers who were sampled in their schools.

Zayatz, L. (1991a), "Estimation of the Number of Unique Population Elements Using a Sample," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 369-373.

The author introduces two methods of estimating the percent of unique population elements in a sample microdata file. Examples of performance are given.

Zayatz, L. (1991b), "Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample," Statistical Research Division Report Series, Census/SRD/RR-91/08, Bureau of the Census, Statistical Research Division, Washington, DC.

The author introduces two methods of estimating the percent of unique population elements in a sample microdata file. A third method is evaluated. Examples of performance of all three methods are given.

Zayatz, L. (1991c), "Estimation of the Percent of Records on a Death Information Microdata File that Could be Correctly Matched to CPS Microdata," Statistical Research Division Technical Note Series, No. RR-91/02, Bureau of the Census, Statistical Research Division, Washington, DC.

This paper describes the results of using two techniques to estimate the percent of records on one version of a microdata file (the Death Information file) that could be linked to a Current Population Survey microdata file. The Microdata Review Panel considered these results, made some changes to the file, and then approved the release of the file.

Zayatz, L. (1992a), "The Effect of Geographic Detail on the Disclosure Risk of Microdata from the Survey of Income and Program Participation," Statistical Research Division Report Series, Bureau of the Census, Statistical Research Division, Washington, DC, No. CCRR-92/03.

The author describes the relationship between the percent of population uniques and the geographic detail on a Survey of Income and Program Participation microdata file. The relationship between the percent of sample uniques and the geographic detail on such a file is also examined. The objective is to relate the consequences in terms of disclosure risk of lowering the required minimum number of persons in the sampled population per identified geographic region on SIPP microdata files.

Zayatz, L. (1992b), "Using Linear Programming Methodology for Disclosure Avoidance Purposes," Statistical Research Division Report Series, Census/SRD/RR-92/02, Bureau of the Census, Statistical Research Division, Washington, DC.

This paper presents a linear-programming scheme for finding complementary suppressions for a primary suppression which is applicable to two or three dimensional tables. The method yields good but not optimal results. The paper discusses three ways of improving results: 1) sorting the primary suppressions by the protection they need and finding complementary cells for each primary cell sequentially beginning with the largest; 2) adding an additional run through the linear program with an adjusted cost function to eliminate unnecessary complementary suppressions identified in the first run; and 3) using different cost functions. A general comparison with network flow methodology is also given. The paper also provides an example using the commercially available linear programming package, LINDO.

Zayatz, L. V. (1992c), "Linear Programming Methodology for Disclosure Avoidance Purposes at the Census Bureau," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 679-684.

This paper recommends specific approaches for finding complementary suppressions for two dimensional tables, small three dimensional tables and large three dimensional tables. Network flow procedures are recommended for two dimensional tables. Linear programming methods are recommended (and described) for small three dimensional tables. In the case of large three dimensional tables, the recommended procedure is a sequence of network flow algorithms applied to the two-dimensional subtables. The resultant system of suppressions must then be audited to assure that the sensitive cells are protected. A linear programming algorithm for validating a pattern of suppressions is described.

Zayatz, L. V. (1992d), "Using Linear Programming Methodology for Disclosure Avoidance Purposes," Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, pp. 341-351.

This paper is based on Zayatz (1992b). It describes the implementation of linear-programming to find complementary suppressions for a single primary suppression. The method identifies the complete set of complementary suppressions by considering the primary suppressions sequentially. The procedure is applicable to two or three dimensional tables. The three ways of improving results, listed above under Zayatz (1992b), are discussed.