APPENDIX A – Technical Notes: Extending Primary Suppression Rules To Other Common Situations

This appendix contains practices the statistical agencies have found useful when applying disclosure limitation to tables in common situations. The primary and complementary suppression procedures for tables of magnitude data discussed in Chapter IV are based on the assumption that the reported data are strictly positive, and that the published number is the simple sum of the data from all respondents. In some situations published data are not simple sums, and it is not clear how to apply primary and complementary suppression methodology. For example, in this appendix we extend primary suppression rules used for tabular data to tables containing imputed data.

Further, the methods discussed in this paper are implicitly to be applied to every published variable. In practice, simplifying assumptions have been made to reduce the workload associated with disclosure limitation and to improve the consistency of published tables over time.

Section 2 presents the disclosure limitation practices that have been used where there may be some question as to how to apply the standard procedures. Section 3 presents the simplifying assumptions that have been found useful by federal statistical agencies. Both sections are intended as a reference for other agencies facing similar situations.

1. Background

The (n, k), pq-ambiguity and p-percent rules described in Chapter IV can all be written in the following form:

$$S(X) = \sum_{i=1}^{n} x_i - c\left(T - \sum_{i=1}^{s} x_i\right)$$

where the values of n, c and s depend on the specific rule and the parameters chosen, T is the total to be published, x_1 is the largest reported value, x_2 is the second largest reported value, and so on. In this framework, the x_i are all nonnegative.

2. Extension of Disclosure Limitation Practices

2.a. Sample Survey Data

The equation above assumes that all data are reported (as in a census). How can this rule be applied to data from a sample survey? One way of handling this is to let the values of the largest respondents, the xi, be specified by the unweighted reported values, but to let T be the weighted total to be published. (Note: this is a consistent way of stating that there is no disclosure with data from a sample survey when no units are selected with certainty and the sampling fractions

are small.)

2.b. Tables Containing Imputed Data

If some data are imputed, disclosure potential depends on the method of imputation.

a) Imputation for a sample survey is done by adjusting weights: In this case, method 2.a applies (the adjusted weights are used to calculate the weighted total, T).

b) Imputed values may be based on other respondent's data, as in "hot decking": In this case, the imputed value should not constitute a disclosure about the nonrespondent, so the imputed value (weighted, if appropriate) is included in the estimated total, T. The imputed value is counted as an individual reported value for purposes of identifying the largest respondents only for the donor respondent.

c) Imputed values may be based on past data from the nonrespondent: If the imputed value were revealed, it could constitute disclosure about the nonrespondent (for example, if the imputed value is based on data submitted by the same respondent in a different time period). The imputed value is included in the estimated total, T, and is also treated as submitted data for purposes of identifying the largest respondents.

2.c. Tables that Report Negative Values

If all reported values are negative, suppression rules can be applied directly by taking the absolute value of the reported data.

2.d. Tables Where Differences Between Positive Values Are Reported

If the published item is the difference between two positive quantities reported for the same time period (e.g. net production equals gross production minus inputs), then apply the primary suppression rule as follows:

a) If the resultant difference is generally positive, apply the suppression procedure to the first item (gross production in the above example).

b) If the resultant difference is generally negative, apply the suppression procedure to the second item (inputs in the above example.)

c) If the resultant difference can be either positive or negative and is not dominated by either, there are two approaches. One method is to set a threshold for the minimum number of respondents in a cell. A very conservative approach is to take the absolute value of the difference before applying the primary suppression rule.

2.e. Tables Reporting Net Changes (that is, Difference Between Values Reported at Different Times)

If either of the values used to calculate net change were suppressed in the original publication, then net change must also be suppressed.

2.f. Tables Reporting Weighted Averages

If a published item is the weighted average of two positive reported quantities, such as volume weighted price, apply the suppression procedure to the weighting variable (volume in this example).

2.g. Output from Statistical Models

Output from statistical models, such as econometric equations estimated using confidential data, may pose a disclosure risk. Often the resulting output from the statistical analyses takes the form of parameter coefficients in various types of regression equations or systems of equations. Since it is only possible to exactly recover input data from a regression equation if the number of coefficients is equal to the number of observations, regression output generally poses no disclose risk. However, sometimes dummy (0,1) variables are used in the model to capture certain effects, and these dummy variables may take on values for only a small number of observations.

One way of handling this situation is provided by the Center for Economic Studies of the Census Bureau. They treat the dummy variables as though they were cells in a table. Using the (n, k) rule, disclosure analysis is performed on the observations for which the dummy variable takes on the value 1.

3. Simplifying Procedures

3.a. Key Item Suppression

In several economic censuses, the Census Bureau employs key item suppression: performing primary disclosure analysis and complementary suppression on certain key data items only, and applying the same suppression pattern to other related items. Under key item suppression, fewer agency resources are devoted to disclosure limitation and data products are more uniform across data items. Key and related items are identified by expert judgment. They should remain stable over time.

3.b. Preliminary and Final Data

For magnitude data released in both preliminary and final form, the suppression pattern identified and used for the preliminary data should be carried forward to the final publication. The final data tables are then subjected to an audit to assure that there are no new disclosures. This conservative approach reduces the risk that a third party will identify a respondent's data from the changes in suppression patterns between preliminary and final publication.

3.c. Time Series Data

For routine monthly or quarterly publications of magnitude data, a standard suppression pattern (primary and complementary) can be developed based on the previous year's monthly data. This suppression pattern, after auditing to assure no new disclosures, would be used in the regular monthly publication.

APPENDIX B – Government References and Websites

1. Report on Statistical and Disclosure-Avoidance Techniques. Statistical Policy Working Paper 2 (May 1978). Washington, DC: U.S. Department of Commerce, Office of Policy and Federal Statistical Standards. This report is available from the National Technical Information Service: NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161; 703-487-4650. The NTIS document number is PB86-211539/AS.

2. Energy Information Administration Standards Manual. (September 2002). Energy Information Administration, U.S. Department of Energy. Washington, DC. http://www.eia.doe.gov/smg/Standard.pdf

3. Federal Statistics: Report of the President's Commission on Federal Statistics, Vol. 1. President's Commission on Federal Statistics. Washington, DC: U.S. Government Printing Office.

4. NASS Policy and Standards Memoranda. National Agricultural Statistics Service, U.S. Department of Agriculture. Washington, DC.

5. NCES Statistical Standards. (June 2003). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/statprog/2002/stdtoc.asp

6. NCES Standard on "Maintaining Confidentiality" National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/statprog/2002/std4_2.asp

7. NCHS Staff Manual on Confidentiality. (September 2004). National Center for Health Statistics, U.S. Department of Health and Human Services. Washington, DC. http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

8. Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies. Publication 1299 (February, 1986). Statistics of Income Division, Internal revenue Service, U.S. Department of Treasury. Washington, DC.

9. SOI Division Operating Manual. (January 1985). Statistics of Income Division, Internal revenue Service, U.S. Department of Treasury. Washington, DC.

WEBSITES FOR ADDITIONAL SOURCES

1) http://www.fcsm.gov/committees/cdac/ Website for the Confidentiality and Data Access Committee. This site provides useful links to resources for disclosure avoidance methodologies and related data access issues.

2) http://www.amstat.org/comm/cmtepc/index.cfm Website for the American Statistical Association's Privacy, Confidentiality, and Data Security. This site provides comprehensive information and references for the methodological, legal, ethical, and technical issues that arise out of protecting and using statistical data

3) www.census.gov/srd/sdc/index.html This site provides links and conventional references for research sponsored by the U.S. Census Bureau in the areas of statistical disclosure control, confidentiality, and disclosure limitation

4) http://aspe.os.dhhs.gov/datacncl/privcmte.htm U.S. Dept. of Health and Human Services Privacy Committee's website.

5) http://neon.vb.cbs.nl/casc/ Website for Computational Aspects of Statistical Confidentiality (CASC) (managed by the Netherlands Statistical Bureau). This site provides links for downloading Mu-Argus and Tau-Argus for applying disclosure avoidance rules to either microdata or tabular data; There are other useful links to books, papers, and presentations.

APPENDIX C – References

The purpose of this listing is to update the references on disclosure limitation methodology that were cited in Statistical Policy Working Paper 2 and the original version of Statistical Policy Working Paper 22. Several papers have been written since both these Statistical Policy Working Papers were published in 1978 and 1994, respectively.

In the Federal statistical system the Census Bureau 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 through their website shown in Appendix B. For these reasons the statistical disclosure limitation research sponsored by the Bureau of the Census is thoroughly and adequately covered in this bibliography. In addition, important papers that 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 are also included.

The "books" listed below in alphabetical order refer to traditional technical books written by a single author or a few co-authors, special collections of papers by many different authors, special issues of journals devoted to disclosure, and various online sources (e.g., references, manuals).

Books

"Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies"; edited by Pat Doyle, Julia I. Lane, Jules J.M. Theeuwes, Laura V. Zayatz. Published in 2001 by Elsevier Science B.V., Amsterdam, The Netherlands.
This volume has sixteen chapters written by leading researchers in a wide variety of disclosure topics. A description and list of articles appears at:

www.elsevier.com/wps/find/bookdescription.cws_home/622129/description#description
Chapter 1 is available online: www.census.gov/srd/sdc/ConfidentialityCH1.pdf

"Elements of Statistical Disclosure Control" by Leon Willenborg and Ton de Waal. Published by Springer in 2001. Lecture Notes in Statistics, volume 155. This volume is more theoretical than the earlier volume by these authors and goes into depth on many important methods. It has chapters on (i) disclosure risk (ii) information loss (iii) non-perturbative techniques (iv) perturbative techniques first for microdata and then for tabular data. There are 119 literature references presented at the end of the volume.

"For the Record, Protecting Electronic Health Information," by the National Academy of Sciences and National Research Council. Published in 1997 by the National Academy Press, Washington, D. C. In 1996, the Computer Science and Telecommunications Board (CSTB) formed a 15 member Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure. The committee addressed threats to healthcare information, adequacy of existing privacy and security measures, and best practices. The results of the committee's work were published in this book.

"Improving Access to and Confidentiality of Research Data", Committee on National Statistics", National Research Council, edited by Christopher Mackie and Norman Bradburn; published by National Research Council, National Academy Press, Washington, D.C., 2000. Summary of a workshop convened by CNSTAT to promote discussion about methods for advancing the often conflicting goals of exploiting the research potential of microdata and maintaining acceptable levels of confidentiality.

"Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics," edited by George T. Duncan, Thomas B. Jabine, Virginia A. de Wolf; published by the Committee on National Statistics and the Social Science Research Council, National Academy Press, Washington, D.C., 1993. This short (23 pages) but important volume consists of the executive summary and recommendations of the Panel on Confidentiality and Data Access. This panel was organized by CNSTAT and the Social Science Research Council to develop recommendations that could aid federal statistical agencies in their stewardship of data for policy decisions and research.

"Record Linkage and Privacy, Issues in Creating New Federal Research and Statistical Information," (GAO-01-126SP). This book provides a summary of various methodologies and matching techniques for matching a microdata file to an outside file. It updates a previous summary of mathematical methods used for matching found in "Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies", Dept of Treasury, IRS, SOI, Publication 1299 (2-86).

"Statistical Disclosure Control in Practice" by Leon Willenborg and Ton de Waal. Published by Springer in 1996. Lecture Notes in Statistics, Volume 111. This book aims to discuss various aspects associated with disseminating personal or business data collected in censuses or surveys or copied from administrative sources. There are two detailed chapters on statistical disclosure control discussing the protection issues for microdata and several techniques that have been developed and used at various agencies. These are similar chapters for tabular data. There are 79 literature references presented at the end of the volume.

Reports Of Conferences and Workshops

Workshop on statistical data confidentiality (Skopje, Macedonia, March 2001). Sponsored by United Nations Economic Commission for Europe (UNECE).
Proceedings are available at http://192.91.247.58/stats/documents/2001.03.confidentiality.htm. This site also provides useful links to the papers and other statistical methodology materials.

"Inference Control in Statistical Databases: From Theory to Practice" (conference in Luxemburg, December 2001). Edited by Josep Domingo-Ferrer. Published by Springer in 2002 in Lecture Notes in Computer Science series, LNCS #2316. The list of articles with brief abstracts are available at: http://www.springerlink.com/app/home/search-articles-results.asp?wasp=5n5d6ynmwn0vwp8d4gfy&referrer=searchmainxml&backto=journal,1,1;linkingpublicationresults,1:105633,1

Workshops sponsored or co-sponsored by Eurostat. "Privacy in Statistical Databases", proceedings of Barcelona, June 2004 conference). Edited by Jose Domingo-Ferrer and Vicenc Torra. Published by Springer in 2004 in Lecture Notes in Computer Science series, #3050. The list of articles with brief abstracts may be found at: http://www.springerlink.com/app/home/search-articles-results.asp?wasp=3l93gmuvtj7yuk32wmf0&referrer=searchmainxml&backto=journal,1,1;linkingpublicationresults,1:105633,1

"Monographs of Official Statistics: Work session on statistical data confidentiality" (Proceedings of Luxembourg conference, April 2003). Published by Eurostat in 2004. The following three online .pdf documents form the entire proceedings.
http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-1/EN/KS-CR-03-004-1-EN.PDF
http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-2/EN/KS-CR-03-004-2-EN.PDF
http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-3/EN/KS-CR-03-004-3-EN.PDF

Special Issues of Journals

Journal of Official Statistics: Special Issue on Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data, Vol. 19, No. 4, December 1998. Edited by Stephen E. Fienberg and Leon C.R.J. Willenborg (This journal is published by Statistics Sweden)
For list of articles see: http://www.jos.nu/Contents/issue.asp?vol=14&no=4

Journal of Official Statistics: Special Issue on Confidentiality and Data Access, Vol.9, No. 2., June 1993. For list of articles see: http://www.jos.nu/Contents/issue.asp?vol=9&no=2

The journal "Of Significance", published by the Association of Public Data Users, had a special issue on Confidentiality in 2000. It is volume 2, number 1 and is available online at: www.apdu.org/resources/docs/OfSignificance_v2n1.pdf

Netherlands Official Statistics: Special issue on Statistical Disclosure Control, vol. 14, Spring 1999. www.cbs.nl/nl/publicaties/publicaties/algemeen/a-125/1999/nos-99-1.pdf

Online References

An annotated list of references in contained in the article by John M. Abowd and Simon D. Woodcock in the volume, "Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies." This list is also available online at http://www.census.gov/srd/sdc/abowd-woodcock2001-appendix-only.pdf

A list of Microdata Confidentiality References compiled by William E. Winkler in March 2004 may also be found at www.census.gov/srd/sdc.

Websites dedicated to disclosure issues and/or references:
www.fcsm.gov/committees/cdac/cdac.html
www.census.gov/srd/sdc.

Manual

Checklist on Disclosure Potential of Proposed Data Releases (prepared by Confidentiality and Data Access Committee (CDAC) of the Federal Committee on Statistical Methodology (FCSM). http://www.fcsm.gov/committees/cdac/cdac.html

Report of the Task Force on Disclosure: GSS Methodology Series, no. 4, Government Statistical Service. Dec 1995, Office of National Statistics, London. This report is available online: http://www.statistics.gov.uk/downloads/theme_other/GSSMethodology_No_04_v2.pdf

National Center for Health Statistics Staff Manual on Confidentiality
http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

Articles

Aboud, J. M. and Lane, J. I., "Synthetic Data and Confidentiality Protection," (September, 2003). Technical Paper No. TP-2003-10, U.S. Census Bureau. The authors describe a method of creating multiple public use files from a single database where the actual values are replaced with scientifically valid estimates. The analytical value of the selected confidential variables is preserved while providing disclosure protection to the file.

Angle, John. (2003). "Imitating the Salamander: Reproduction of the Truncated Right Tail of an Income Distribution." This paper proposes a method to estimate the right tail of an income distribution using knowledge of the left and center portion of the variable's distribution and provides insight in applying top coding to a microdata file.
http://www.fcsm.gov/03papers/Angle_Final.pdf.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," Journal of the American Statistical Association, Vol. 85, p. 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, microaggregation, 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.

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

Cox, L. H., (1980) "Suppression Methodology and Statistical Disclosure Control," Journal of the American Statistical Association, Vol. 75, No. 370, p. 377-385. This article highlights the interrelationships between the processes of disclosure definitions, sub-problem 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. 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 sub-problems. 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. 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, p. 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. and Ernst, L. R. (1982), "Controlled Rounding," INFOR, Canadian Journal of Operation Research and Information Processing, Vol. 20, No. 4, p. 423-432. Reprinted: Some Recent Advances in the Theory, Computation and Application of Network Flow Methods, University of Toronto Press, 1983, p. 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., S.K. McDonald and D.W. Nelson, (1986). "Confidentiality Issues at the U.S. Bureau of the Census," Journal of Official Statistics Vol. 2, No. 2, p. 135 –160. This paper describes the policies and procedures of the U.S. Census Bureau following a major review and research program in data confidentiality protection during the mid-1980's.
http://www.jos.nu/Contents/jos_online.asp

Cox, L. H. (1987), "A Constructive Procedure for Unbiased Controlled Rounding," Journal of the American Statistical Association, Vol. 82, p. 520-524. Unbiased controlled rounding in a table involves rounding to an integer base, preserving additive structure, and assuring that the expected value of the rounded entry equals the original entry. This paper provides an easy-to-implement algorithm for achieving unbiased controlled rounding in a 2-dimensional table. The method also solves the two-way stratification problem in survey sampling and can be used to assure integer sample counts in an unbiased manner following, e.g., iterative proportional fitting (raking).

Cox, L. H. and George, J. A. (1989), "Controlled Rounding for Tables with Subtotals," Annals of Operations Research, 20 (1989) p. 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. (1995), "Network Models for Complementary Cell Suppression," Journal of the American Statistical Association, Vol. 90, No. 432, pp. 1453-1462. 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. (1996), "Protecting Confidentiality in Small Population Health and Environmental Statistics," Statistics in Medicine, Vol. 15, p. 1895-1905. This paper discusses confidentiality problems in small domains and suggests the use of subsampling and supersampling for disclosure limitation in microdata files.

Cox, L. H. (2002), "Bounds on Entries in 3-Dimensional Contingency Tables Subject to Given Marginal Totals," in: Inference Control in Statistical Databases—From Theory to Practice, Lecture Notes in Computer Science 2316 (J. Domingo-Ferrer, ed.), New York: Springer, p. 21-33. This paper examines the problem of determining exact bounds for suppressed entries in 3-dimensional contingency tables given specified marginal totals and flaws in previous approaches, and compares several methods analytically.

Cox, L. H. (2003), "On Properties of Multi-Dimensional Statistical Tables," Journal of Statistical Planning and Inference, Vol. 117, 251-273. This paper examines mathematical properties of multi-dimensional statistical tables, including problems and procedures for assuring the existence of a feasible table given specified marginal tables, failure of linear programming to produce

integer solutions given integer constraints, and conditions under which integral solutions are assured based on network structure and network linear programming.

Cox, L. H. and Dandekar, R. A. (2004), "A New Disclosure Limitation Method for Tabular Data that Preserves Data Accuracy and Ease of Use," Proceedings of the 2002 FCSM Statistical Policy Seminar, Statistical Policy Working Paper 35, Federal Committee on Statistical Methodology, Washington, DC: U.S. Office of Management and Budget, p. 15-30. http://www.fcsm.gov/working-papers/spwp35.html
This paper introduces controlled tabular adjustment to the federal statistical community, focusing on its potential to improve data quality.

Cox, L. H., Kelly J., Patil, R. (2004). "Balancing Quality and Confidentiality for Multi-Variate Tabular Data. This paper proposes the use of certain linear and non-linear models subject to specific constraints that may be used to adjust tabular data in order to preserve additivity, covariance, correlation, and regression coefficients and other data relationships from the original table are preserved.

Cox, L. H., James P. Kelly, and Rahul J. Patil. (2005). "Computational Aspects of Controlled Tabular Adjustment: Algorithm and Analysis" in the book "The Next Wave in Computing, Optimization, and Decision Technologies", ed. B. Golden, S. Raghavan, E. Wasil, published by Springer. This paper presents a cutting plane algorithm for speeding controlled tabular adjustment.

Dandekar, R., Cohen, M., and Kirkendall, N. (2002). "Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique. Lecture Notes in Computer Science, vol. 2316, pp. 117-125, Apr. 2002. ISSN 0302-9743. Vol. Inference Control in Statistical Databases, ed. Josep Domingo-Ferrer, Berlin:Springer-Verlag. This paper discusses a methodology for creating synthetic micro data that can be used in place of actual reported data or to create either additive or multiplicative noise which when merged with the original data can provide disclosure protection while reproducing many of the essential quality of the original micro data file.
Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique
<http://taz/smg/papers/BARCEL.pdf

Dandekar Ramesh A., (2004) "Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems", (2004), p. 428-434, Monographs of Official Statistics, Luxembourg: Eurostat. The paper describes a simplified procedure as an alternative to the linear programming based controlled tabular adjustment (CTA) methodology to generate synthetic tabular data to protect data containing sensitive information. The simplified CTA procedure is a low cost approach that allows statistical agencies to use conventional readily available software tools to generate synthetic tabular data.

Dandekar, Ramesh, (2004). "Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data." Lecture Notes in Computer Science, Springer-Verlag Heidelberg, ISSN: 0302-9743, Vol. 3050 p. 121-135. The paper discusses a simplified version of the CTA and applies it to categorical and magnitude test data. It also

provides a comparative evaluation of this simplified CTA approach and LP-based CTA using magnitude test data. For these test data, the simplified CTA is able to protect the tables with many fewer adjustments to cell values than the LP-based CTA requires.

De Loera, J., Ohn, Shmuel, "All Rational Polytopes Are Transportation Polytopes and All Polytopal Integer Sets Are Contingency Tables." IPCO 2004, LNCS 3064, pp. 338–351. This paper shows that any rational polytope is polynomial-time representable as a "slim" $r \times c \times 3$ three-way line-sum transportation polytope. This universality theorem has important consequences for linear and integer programming and for confidential statistical data disclosure. It provides polynomial-time embedding of arbitrary linear programs and integer programs in such slim transportation programs and in bipartite bi-flow programs. It resolves several standing problems on 3-way transportation polytopes. It also demonstrates that the range of values an entry can attain in any slim 3-way contingency table with specified 2-margins can contain arbitrary gaps, suggesting that disclosure of k-margins of d-tables for $2 \leq k < d$ is confidential. http://www.opt.math.tu-graz.ac.at/IPCO/prog.10

Dobra, Adrian, Fienberg, Stephen E., (2000), "Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs", Proceedings of the National Academy of Sciences Vol. 97 No. 22 p. 1885-1892: Upper and lower bounds on cell counts play an important role in statistical disclosure limitation. This paper provides the theoretical framework and proofs of the exactness of Frechet bounds on decomposable graphical loglinear models. For such models, simple formulae, in lieu of computationally demanding integer programs, yield exact bounds. Some of these models are familiar in statistics, e.g. complete independence models, but overall this entire class of models is relatively small.

Dobra, Adrian, Fienberg, Stephen E. (2001) "Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals with Applications to Disclosure Limitation." Statistical Journal of the United Nations ECE. Vol. 18, p. 363–371. This paper is a more descriptive version of the results presented in Dobra and Fienberg (2000) on computing exact bounds for decomposable graphical models.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001), "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map," Los Alamos National Laboratory Technical Report, LA-UR-01-6428. Methods are discussed for assessing the disclosure risk of a file and trade offs in data utility as the parameters in various disclosure limitation methodologies are changed. The authors describe a method for calculating separate numerical assessments of the disclosure risk and data utility while allowing different values for the disclosure limitation parameters.

Evans, T., Zayatz, L., Slanta, J., (1998). "Using Noise for Disclosure Limitation Establishment Tabular Data," Journal of Official Statistics, Vol. 14, p. 537-551. This paper discusses the disclosure limitation method for protecting establishment magnitude tabular data by adding noise to the underlying microdata prior to tabulation.

Ernst, L., (1989), "Further Applications of Linear Programming to Sampling Problems," Proceedings of the Survey Research Methods Section, American Statistical Association, p. 625-630. In a previous paper, Cox and Ernst (1982), it was demonstrated that a controlled rounding exists for every two-dimensional additive table. In this paper the author establishes by means of

a counter-example that the natural generalization of their result to three dimensions does not hold.

Fienberg, Stephen E. 1997. "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research." Carnegie Mellon Department of Statistics Technical Report, Working Paper No. 668. Carnegie Mellon Department of Statistics. Pittsburgh, Pennsylvania. http://www.stat.cmu.edu/tr/tr668/tr668.html. This paper provides an overview of the statistical issues that are related to the evolving area of statistical disclosure limitation methodology.

Fischetti, M and Salazar, JJ (1999), "Models and Algorithms for the 2-Dimensional Cell Suppression Problem in Statistical Disclosure Control," Mathematical Programming, Vol. 84, 283-312. This paper introduces the Fischetti-Salazar method for solving the decision problem associated with complementary cell suppression. Unlike previous methods, it protects all sensitive cells at once rather than sequentially, and can produce optimal results in medium to large problems.

Fischetti, M. and Salazar, JJ (2000), "Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints," Journal of the American Statistical Association, Vol. 95, p. 916-928. Algorithms for complementary cell suppression for tabular data shown to run to optimality in large, but not enormous, problem settings.

Gomatam, S., Karr, A. F., Sanil, A. P. (2005), "Data swapping as a decision problem," Journal of Official Statistics. This paper discusses risk-utility formulation of data swapping for categorical data.

Gomatam, S., Karr, A. F., Reiter, J. P., Sanil, A. P. (2005), "Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers," Statistical Science, Vol. 20, p. 163 - 177. Remote access analysis servers allow users to submit requests for output from statistical models fit using confidential data. The users are not allowed access to the data themselves. Analysis servers, however, are not free from the risk of disclosure, especially in the face of multiple, interacting queries. In this paper, the authors describe these risks and propose quantifiable measures of risk and data utility that can be used to specify which queries can be answered, and with what output. The risk-utility framework is illustrated for regression models.

Gonzalez, JF and Cox, LH (2005), "Software for Tabular Data Protection," Statistics in Medicine, Vol. 24 (4), p. 659-669. This paper describes software for data protection in two-way tables developed for the National Center for Health Statistics: complementary cell suppression, rounding, perturbation and controlled tabular adjustment. The software is available at no charge.

Greenberg, B. and Zayatz, L. (1992), "Strategies for Measuring Risk in Public Use Microdata Files," Statistica Neerlandica, Vol. 46, No. 1, p. 33-48. Methods of reducing the risk of disclosure for microdata files and factors that 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, p. 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.

Hawala, S., Zayatz, L., Rowland, S., (2004). "American FactFinder: U.S. Bureau of the Census works towards meeting the needs of users while protecting confidentiality," Journal of Official Statistics, Vol. 20, p. 115-124. This paper discusses the special disclosure limitation techniques that are applied to protect the confidentiality of tabulations generated from an online query of microdata files. http://www.jos.nu/Contents/jos_online.asp

Jabine, T. B. (1993a), "Procedures for Restricted Data Access," Journal of Official Statistics, Vol. 9, No. 2, p. 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 that illustrate both successful modes and procedures for providing access, and failures to gain the desired access.
http://www.jos.nu/Contents/jos_online.asp

Jabine, T. B. (1993b), "Statistical Disclosure Limitation Practices of United States Statistical Agencies," Journal of Official Statistics, Vol. 9., No. 2, p. 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.
http://www.jos.nu/Contents/jos_online.asp

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. http://www.census.gov/srd/sdc/Jewett.disc.econ.1992.pdf

Karr, A. F., Lin, X., Reiter, J. P., Sanil, A. P. (2005). Secure regression on distributed databases. Journal of Computational Graphical Statistics Vol. 14 No. (2) p. 263–279. This article presents several methods for performing linear regression on the union of distributed databases that preserve, to varying degrees, confidentiality of those databases. Such methods can be used by federal or state statistical agencies to share information from their individual databases, or to make such information available to others.

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 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.

Kennickell, Arthur B. (1998). "Multiple Imputation in the Survey of Consumer Finances," Proceedings of the Joint Statistical Meetings American Statistical Association 1998. This paper describes the FRITZ system of multiple imputation developed for the Survey of Consumer Finances. In addition to describing the application of the system to ordinary problems of imputation of missing data, the paper presents the results of using the system for a set of experiments in data simulation for disclosure avoidance.
http://www.federalreserve.gov/pubs/oss/oss2/papers/impute98.pdf

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, Proceedings of the Section on Survey Research Methods, p. 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. The author 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. J., and W.E. Winkler (1995). "Masking Microdata Files," American Statistical Association, Proceedings of the Section on Survey Research Methods, p. 114-119. 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.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," Journal of Official Statistics, Vol. 9, No. 2, p. 313-331. The definition of disclosure depends on the context. Sometimes a disclosure is said to occur even though the information revealed is incorrect. A disclosure may violate a respondent's anonymity and sometimes reveal sensitive information. This paper tries 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. http://www.jos.nu/Contents/jos_online.asp

Lee, J., Holloman, C., Karr, A. F. and Sanil, A. P. (2001), "Analysis of Aggregated Data in Survey Sampling with Application to Fertilizer/Pesticide Usage Surveys," Research in Official Statistics, Vol. 4, p. 101-116: This paper proposes a Bayesian simulation approach for analysis of data aggregated to protect disclosure.

Massell, Paul B., (2002). "Optimization Models and Programs for Cell Suppression in Statistical Tables," Proceeding of the Joint Statistical Meetings American Statistical Association 2002. This paper compares the different mathematical approaches to applying cell suppression and evaluates the usefulness of the different programs based on the optimization method as well as other practical considerations. Network based programs and extended network based programs are compared with linear programming, integer based, and hypercube based programs. http://www.census.gov/srd/sdc/Massell.JSM2002.v4.pdf

Massell, Paul B., (2004). "Comparing Statistical Disclosure Control Methods for Tables: Identifying the Key Factors", Proceedings of the Joint Statistical Meetings American Statistical Association 2004. This paper describes the key factors involved in deciding how to select a statistical disclosure method that is suitable for protecting a given set of tables.
http://www.census.gov/srd/sdc/Massell.JSM2004.paper.v3.pdf

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 microdata 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.

Muralidhar, K., Sarathy, R. (May, 2002). "A Data Shuffling Procedure for Masking Data," This paper discusses the methodology and theoretical basis for applying a two-step data swapping procedure for protecting confidential numerical data. Report to the Census Bureau, May, 2002. http://gatton.uky.edu/faculty/muralidhar/maskingpapers.

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

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.

Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. (2003), "Multiple Imputation for Statistical Disclosure Limitation," Journal of Official Statistics, Vol. 19, p. 1-16. This article evaluates the use of the multiple imputation framework to protect the confidentiality of respondents' answers in sample surveys. The basic proposal is to simulate multiple copies of the population from which these respondents have been selected and release a random sample from each of these synthetic populations. Users can analyze the synthetic sample data sets with standard complete-data software for simple random samples, then obtain valid inferences by combining the point and variance estimates using the methods in this article.
http://www.jos.nu/Contents/jos_online.asp

Reiter, J.P. (2002), "Satisfying Disclosure Restrictions with Synthetic Data Sets," Journal of Official Statistics, Vol. 18, No. 4, p. 531-543. To avoid disclosures, Rubin proposed creating multiple, synthetic data sets for public release so that (i) no unit in the released data has sensitive data from an actual unit in the population, and (ii) statistical procedures that are valid for the original data are valid for the released data. This paper discusses through the use of simulation studies that valid inferences can be obtained from synthetic data in a variety of settings, including simple random sampling, probability proportional to size sampling, two-stage cluster sampling, and stratified sampling. http://www.jos.nu/Articles/abstract.asp?article=184531

Reznek, A. P., "Disclosure Risks in Cross-Section Regression Models," (2003). This paper describes the disclosure risks associated with certain types of cross section regression models. In

particular, it shows via examples that models with only fully interacted dummy (0,1) variables on the right-hand side allow recovery of entries from a table of means of the left-hand side variable, broken down by the categories of the dummy variables. Proceedings of the Joint Statistical Meetings American Statistical Association 2003

Reznek, Arnold P. and T. Lynn Riggs (2004). "Disclosure Risks in Regression Models: Some Further Results." Proceedings of the Joint Statistical Meetings American Statistical Association 2004. This paper illustrates that correlation matrices and variance-covariance matrices of variables, as well as variance-covariance matrices of model coefficients, can also allow recovery of table entries if the variables include dummy variables.

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.

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, p. 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.

Sailer, P., Weber, M., and Wong, W., (2001), "Disclosure-Proofing the 1996 Individual Tax Return Public-use File," Proceedings of the American Statistical Association 2001; This paper provides an overview of the disclosure-proofing techniques applied to the Statistics of Income Individual Tax Return Public-Use File (PUF). It also discusses the results of two tests of these procedures: the matching of a publicly available marketing database to the PUF: and the matching of the IRS Individual Master File to the PUF.

Sanil, A. P., Karr, A. F., Lin. X., Reiter, J. P. (2004), "Privacy preserving regression modeling via distributed computation," Proceedings of the Tenth ACM SIGKDD 2004 International Conference on Knowledge Discovery and Data Mining p. 677–682. This paper discusses secure regression for distributed, vertically partitioned data when the response is shared.

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, p. 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.

Steel, Philip M. (2004) "Disclosure Risk Assessment for Microdata," This is an introduction to risk assessment for microdata, for the beginning practioner. It presents some background on legal concepts of identifiablility, discusses risk measurement and its applicability, demonstrates how public data and context can effect risk. There is also an eclectic set of references. http://www.census.gov/srd/sdc/Steel.Disclosure%20Risk%20Assessment%20for%20Microdata.pdf .

Van Den Hout, A., and Van Der Heijden, P. G. M. (2002), "Randomized Response, Statistical Disclosure Control, and Misclassification: A Review." International Statistical Review, Vol. 70 (2), p. 269-288. This paper discusses analysis of categorical data which have been misclassified and where misclassification probabilities are known. Fields where this kind of misclassification occurs are randomized response, statistical disclosure control, and classification with known sensitivity and specificity. Estimates of true frequencies are given, and adjustments to the odds ratio are discussed. Moment estimates and maximum likelihood estimates are compared and it is proved that they are the same in the interior of the parameter space. http://isi.cbs.nl/ISReview/abst01-13.pdf

Winkler, William E. (1998). "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," Research in Official Statistics, Vol. 1, p. 87-114. This paper compares several masking methods in terms of their ability to produce analytically valid, confidential microdata. For a public-use microdata file to be analytically valid, it should, for a very small number of uses, yield analytic results that are approximately the same as the original, confidential file that is not distributed. If a microdata file contains a moderate number of variables and is required to meet a single set of analytic needs, then many more records are likely to be re-identified via modern record linkage methods than via the re-identification methods typically used in the confidentiality literature.

Winkler, William E., (2004). "Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems." This paper provides an overview of various methods applied for masking microdata. It also discusses different measures for estimating disclosure risk for a public-use data file. http://www.census.gov/srd/papers/pdf/rrs2004-03.pdf

Yu, Dunteman, Dai, and Wilson (2004). "Measuring the Performance of MASSC Using NCHS-2000 NHIS Public Use File." The paper discusses the Micro Agglomeration, Substitution, Subsampling, and Calibration disclosure limitation method. Work session on data confidentiality. Conference of European Statisticians 2003. http://www.unece.org/stats/documents/2003.04.confidentiality.htm

Zayatz, L. (1992a). "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 that 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. (1992b), "Linear Programming Methodology for Disclosure Avoidance Purposes at the Census Bureau." Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, p. 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 sub-tables. 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. (2002). "SDC in the 2000 U.S. Decennial Census," Inference Control in Statistical Databases: From Theory to Practice, Springer, p.193-202. This paper describes the statistical disclosure limitation techniques used for all U.S. Census 2000 data products. It includes procedures for short form tables, long form tables, public use microdata files, and an online query system for tables. The procedures that were used include data swapping, rounding, noise addition, collapsing categories, and applying thresholds.

Zayatz, L., Massell, P., and Steel, P. (1999). "Disclosure limitation practices and research at the U. S. Census Bureau" Netherlands Official Statistics, Spring, 1999, Vol. 14, p. 26-29. This paper discusses disclosure limitation practices in effect at the Census Bureau, as well as current Census Bureau research into alternative disclosure limitation procedures and some analysis of these procedures.

APPENDIX D – Confidentiality and Data Access Committee

In 1995, the Interagency Confidentiality and Data Access Group (ICDAG) was formed to (1) promote and implement the goals and recommendations outlined in Chapter 6 of Statistical Policy Working Paper #22 (2) increase cooperation and sharing of statistical disclosure limitation methods among federal agencies and (3) provide a forum for sharing information and ideas on protecting data confidentiality and improving data access. Its members are employees of Executive Branch federal agencies working on data confidentiality and data access issues expressed the need for a forum to share their knowledge and discuss common issues and concerns. Back in 1995, ICDAG was informally affiliated with the Federal Committee on Statistical Methodology (FCSM).

In 1997, the FCSM formally recognized ICDAG as an "Interest Group" to better facilitate communication and cooperation among agencies. In 2000, the name of the group was changed to the Confidentiality and Data Access Committee (CDAC). Since 1997, CDAC has developed several data products to help centralize agency review of disclosure limited data products, share methodology, software, and information across federal agencies on data confidentiality and data access issues and activities. See http://www.fcsm.gov/committees/cdac/ In addition, its members provide presentations on statistical disclosure methodology to various audiences throughout the year to help expand working knowledge in these areas.

Data products that CDAC has developed include:

Checklist on Disclosure Potential of Proposed Data Releases – This document standardizes the review for disclosure risks associated any proposed data release.

Brochure on "Confidentiality and Data Access Issues Among Federal Agencies – This brochure describes some examples of data protections used by federal agencies - legal sanctions, removal of personal identifiers from data sets, the application of statistical procedures to published information, certificates of confidentiality, institutional and disclosure review boards, and restricted data access (research data centers, remote access, special employee status and data licensing).

Restricted Access Procedures - This paper discusses various methods used by five federal agencies for providing access to statistical data while limiting the risk of disclosure of confidential information. The methods include Research Data Centers (RDCs), remote access and on-line query systems, research fellowships and post-doctoral programs, and licensing agreements.

Identifiability in Microdata Files - This document provides an understanding of what variables and types of data might make individual respondents identifiable in a microdata file.

Disclosure Auditing Software – This PC based SAS software identifies the lower and upper bounds on the values of a withheld (suppressed) cell in a tabular statistical table, and provides other useful measures for auditing the suppression pattern in a table.

Reports Available in the Federal Committee on Statistical Methodology's Statistical Policy Working Paper Series

1. ***Report on Statistics for Allocation of Funds***, 1978 (NTIS PB86-211521/AS)
2. ***Report on Statistical Disclosure and Disclosure-Avoidance Techniques***, 1978 (NTIS PB86-211539/AS)
3. ***An Error Profile: Employment as Measured by the Current Population Survey***, 1978 (NTIS PB86-214269/AS)
4. ***Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics***, 1978 (NTIS PB86-211547/AS)
5. ***Report on Exact and Statistical Matching Techniques***, 1980 (NTIS PB86-215829/AS)
6. ***Report on Statistical Uses of Administrative Records***, 1980 (NTIS PB86-214285/AS)
7. ***An Interagency Review of Time-Series Revision Policies***, 1982 (NTIS PB86-232451/AS)
8. ***Statistical Interagency Agreements***, 1982 (NTIS PB86-230570/AS)
9. ***Contracting for Surveys***, 1983 (NTIS PB83-233148)
10. ***Approaches to Developing Questionnaires***, 1983 (NTIS PB84-105055)
11. ***A Review of Industry Coding Systems***, 1984 (NTIS PB84-135276)
12. ***The Role of Telephone Data Collection in Federal Statistics***, 1984 (NTIS PB85-105971)
13. ***Federal Longitudinal Surveys***, 1986 (NTIS PB86-139730)
14. ***Workshop on Statistical Uses of Microcomputers in Federal Agencies***, 1987 (NTIS PB87-166393)
15. ***Quality in Establishment Surveys***, 1988 (NTIS PB88-232921)
16. ***A Comparative Study of Reporting Units in Selected Employer Data Systems***, 1990 (NTIS PB90-205238)
17. ***Survey Coverage***, 1990 (NTIS PB90-205246)
18. ***Data Editing in Federal Statistical Agencies***, 1990 (NTIS PB90-205253)
19. ***Computer Assisted Survey Information Collection***, 1990 (NTIS PB90-205261)
20. ***Seminar on Quality of Federal Data***, 1991 (NTIS PB91-142414)
21. ***Indirect Estimators in Federal Programs***, 1993 (NTIS PB93-209294)
22. ***Report on Statistical Disclosure Limitation Methodology***, Second version 2005
23. ***Seminar on New Directions in Statistical Methodology***, 1995 (NTIS PB95-182978)
24. ***Electronic Dissemination of Statistical Data***, 1995 (NTIS PB96-121629)
25. ***Data Editing Workshop and Exposition***, 1996 (NTIS PB97-104624)
26. ***Seminar on Statistical Methodology in the Public Service***, 1997 (NTIS PB97-162580)
27. ***Training for the Future: Addressing Tomorrow's Survey Tasks***, 1998 (NTIS PB99-102576)
28. ***Seminar on Interagency Coordination and Cooperation***, 1999 (NTIS PB99-132029)
29. ***Federal Committee on Statistical Methodology Research Conference (Conference Papers)***, 1999 (NTIS PB99-166795)
30. ***1999 Federal Committee on Statistical Methodology Research Conference: Complete Proceedings***, 2000 (NTIS PB2000-105886)
31. ***Measuring and Reporting Sources of Error in Surveys***, 2001 (NTIS PB2001-104329)
32. ***Seminar on Integrating Federal Statistical Information and Processes***, 2001 (NTIS PB2001-104626)
33. ***Seminar on the Funding Opportunity in Survey Research***, 2001 (NTIS PB2001-108851)
34. ***Federal Committee on Statistical Methodology Research Conference (Conference Papers)***, 2001 (NTIS PB2002-100103)
35. ***Seminar on Challenges to the Federal Statistical System in Fostering Access to Statistics.*** 2004.
36. ***Seminar on the Funding Opportunity in Survey and Statistical Research.*** 2004.
37. ***Federal Committee on Statistical Methodology Research Conference (Conference Papers)***, 2003.
38. ***Summary Report of the FCSM-GSS Workshop on Web-Based Data Collection.*** 2004.

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161; telephone: 1-800-553-6847. The Statistical Policy Working Paper series is also available electronically from FCSM's web site <*http://www.fcsm.gov*>.