Journal of Official Statistics, Vol. 11, No. 2, 1995, pp. 205–220

An Agenda for Research in Statistical Disclosure Limitation[1]

Lawrence H. Cox[2] and Laura V. Zayatz[3]

Methods for statistical disclosure limitation have appeared in the published literature for over twenty years and in many cases have been incorporated into the practices of national statistical organizations. This paper presents the authors' views on establishing a research agenda that combines agency needs and potential advantages to respondents and data users. Separately for tabular data and microdata, the disclosure problem is discussed, methods and current practice are reviewed, and an agenda for research is presented. Issues regarding the effects of disclosure limitation on data quality and usefulness are examined. Disclosure limitation methods regarded as superior and suitable for general use are identified and the development and dissemination of corresponding statistical software is recommended.

Key words: Confidentiality; microdata; tabular data; disclosure limitation.

1. Introduction

1.1. Confidentiality protection

Most data collected by government statistical agencies are obtained by direct data collection from the respondent or through access to administrative records. Direct data collection includes face-to-face and telephone interviewing and mail-back response through completion and return of a questionnaire. Administrative records, which in the United States include Internal Revenue Service (tax), Social Security (earnings/disability/retirement), and building permits and other municipal records, are used in surveys to develop sampling frames, reduce respondent burden, improve data accuracy and completeness, and minimize data collection costs. Whether collected directly from respondents or from administrative records, much of the data

[1] This paper was stimulated by discussions stemming from work of the Subcommittee on Disclosure Limitation Methodology, U.S. Federal Committee on Statistical Methodology (FCSM). However, the views expressed are those of the authors and are not intended to reflect the views, policies or practices of the FCSM or of the U.S. Environmental Protection Agency or the U.S. Bureau of the Census. The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency. It has been subjected to Agency review and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

[2] U.S. Environmental Protection Agency, Atmospheric Research and Exposure Assessment Laboratory, Research Triangle Park, NC 27711, U.S.A.

[3] U.S. Bureau of the Census, Statistical Research Division, Washington, DC 20233, U.S.A.

collected by statistical agencies pertains to aspects of personal or business life that the respondent would regard as *confidential.*

Confidentiality protection is a recognized responsibility of statistical agencies internationally and is assured by most agencies and in most surveys (for the U.S. situation, see Federal Committee on Statistical Methodology 1994; EUROSTAT 1993 provides selected international approaches). This is done for three reasons. First, confidentiality protection is regarded as ethical behavior in the statistics profession, and is among the ethical standards for statisticians adopted internationally (International Statistical Institute 1986) and in the United States (American Statistical Association 1989). Second, confidentiality protection is often required by law or government regulation (EUROSTAT 1993; National Research Council 1993; Federal Committee on Statistical Methodology 1994). Third, practitioners believe that most respondents would not respond, or not respond truthfully, without a guarantee that confidential data will not be divulged directly or indirectly to a third party (Singer 1993). This amounts to a social contract between the respondent and the statistical agency under which the respondent provides confidential data in return for assurance that the agency will preserve the confidentiality of these data.

1.2. Confidentiality protection policies

The goals of confidentiality protection described above are shared by most national statistical agencies. However, specific requirements vary between agencies due to differing laws, policies and experiences governing agency practices. For example, the U.S. Bureau of the Census collects most of its information under Title 13, U.S. Code, which prohibits the Bureau of the Census from releasing "any publication whereby the data furnished by a particular establishment or individual under this title can be identified." A section of the U.S. Internal Revenue Code dealing with "Statistical Publications and Studies" places the same restrictions on publications of the Internal Revenue Service. The U.S. Social Security Administration's Regulation Number 1, Section 1106 of the Social Security Act, contains similar confidentiality requirements. The U.S. Bureau of Labor Statistics is required to protect data confidentiality under Commissioner's Order No. 2-80. Other U.S. statistical agencies may not be required by law to maintain confidentiality, but it is their policy to do so (Federal Committee on Statistical Methodology 1994). Statistical agencies in Western countries including Canada, the Netherlands, and the United Kingdom, and in Australia and New Zealand, follow similar laws and policies. Different agency policies may dictate different technical approaches as illustrated later.

1.3. Purposes of this paper

The purposes of this paper are to (1) describe the confidentiality protection problem for tabular data and microdata, (2) summarize confidentiality protection techniques in current use, and (3) present an agenda for future research in confidentiality protection.

This research was stimulated by the authors' participation on the Subcommittee on Disclosure Limitation Methodology of the U.S. Federal Committee on Statistical Methodology, the final report of which appears as Federal Committee on Statistical

Methodology (1994). This paper examines some of the technical issues raised in the report, focusing on research opportunities rather than on current practices. The report is an excellent reference for the reader unfamiliar with the problems and issues discussed here.

2. The Confidentiality Protection Problem

Statistical agencies collect respondent-level data in order to produce data products. Data products are designed to inform public policy, research, or public information. The *confidentiality protection problem* is to thwart the association of a respondent with his/her confidential data. *Disclosure* occurs when confidential respondent data are divulged directly or indirectly from a data product. Another term for confidentiality protection is *disclosure limitation.* The technical details and methodological approaches to confidentiality protection differ according to the type of data product.

A key problem, largely unexamined, is to balance the need for confidentiality protection with legitimate needs of data users (Section 5.2). Every disclosure limitation method in some way affects or modifies true data values and relationships. Ideally, these effects can be quantified so that the selection and use of disclosure limitation methods can be guided by their anticipated effects on data quality, completeness, utility and ease of use.

2.1. Tabular data

The responding unit in some surveys and censuses is an individual or a household. The data collected and published from such surveys and censuses are demographic data. Typical demographic variables include Age, Race, and Sex of the responding individual or Rent, Total Income, and Number in Household for the responding household. Demographic data are typically released in tabular form as sets of frequency count tables. Respondents can be identified, and disclosure may result, in frequency tabulations by examination of small counts or complements of small counts, or, conversely, if nonzero counts are concentrated in too few tabular cells.

The responding unit in many other surveys and censuses is an establishment or business. The data collected and published from such surveys and censuses are establishment data. Typical economic variables include Standard Industrial Classification code, Employment Size, and Value of Shipments of the responding establishment. Establishment data are typically released as tables of aggregates, ratios or averages.

In the presence of detailed geographic information, or simply on the basis of individual characteristics, many establishments are directly identifiable as contributors to particular table cells. In general, statistical agencies assume that establishments can be identified from general knowledge or publicly available sources. Therefore, agencies place primary emphasis on protecting confidential establishment data, and not on reducing the identifiability of establishments. For demographic data, additional effort is made to thwart the identification of individual respondents (i.e., the frequency cell to which the respondent belongs), as frequency cells are defined by characteristics that are often regarded as confidential.

2.2. *Microdata*

Microdata are unit-record data. Often, microdata records pertain to individual respondents and therefore pose serious risk of disclosure. Direct disclosure can occur if identifying information, such as name and address is present on the microdata file, if too many variables are presented with great precision, or if certain respondents are salient in the population. Typically, disclosure can occur from uncensored release of the microdata file, so, at a minimum, original microdata must be abbreviated and data categories broadened before a microdata file is considered disclosure-limited (Cox, McDonald, and Nelson 1986).

Statistical agencies have released disclosure-limited microdata files from demographic surveys and censuses for *public use* or *restricted use* since the 1960s. Due to typically highly skewed distributions for establishment data, microdata are rarely released from establishment censuses and surveys. A notable exception is the U.S. Census of Agriculture, for which microdata at high levels of geographic aggregation and data categorization have been released since the 1987 U.S. Census of Agriculture.

2.3. *Other forms of data*

Data products cannot meet exactly the needs of all users. New or expanded avenues and formats for providing statistical data are constantly being proposed. One avenue, attractive from a data use perspective, is *on-line data query systems* whereby the user requests any and all tabulations and other statistical analyses of interest (McNulty and Unger 1989). Unfortunately, this scenario creates disclosure problems that remain unresolved. Other proposed formats for releasing information include releasing regression coefficients, variance/covariance matrices, low order finite moments (Dalenius and Denning 1982), and simulated data with the statistical properties of real data (Rubin 1993). The disclosure problems presented by these and other proposed data release avenues and formats will not be examined in detail here.

3. Disclosure Limitation in Tabular Data: Status and Research Agenda

3.1. *The disclosure problem*

In frequency count tables, disclosure occurs when small counts are released or may be inferred. For example, if a table shows only one Black male in a census block and shows his marital status as "Divorced," then disclosure has occurred. Or, if a table shows only two physicians in a town and also provides income data for them, then each can infer the other's income. Disclosure can also occur if distributions are spiked. For example, if there are 20 physicians in a town and the data show that 20 physicians earn between $180,000 and $200,000 per year, then confidential income information for each physician has been divulged. Operationally, disclosure is defined in frequency count tables as counts or complements of counts that are small. The precise values defining "small" depend on circumstances.

In tables of aggregate magnitude data such as total sales for all retail establishments in a town, it is often easy to associate establishments with particular cells. The disclosure limitation problem is then to ensure that aggregate cell values do not closely approximate data for any one respondent in the cell and, moreover, that one respondent or a coalition of respondents cannot subtract their contribution from the cell value to achieve a close estimate of the contribution of another respondent. The operational definition of disclosure is then a cell for which a predetermined linear combination of respondent values within the cell is positive. The linear combination is called a *linear sensitivity measure* (Cox 1981) which quantifies the agency's notion of disclosure. For example, to prevent any respondent in the cell or third party from estimating the contribution of any other respondent to the cell to within p percent, the linear sensitivity measure $S(X) = (p/100)x_1 - x_{2+}$ would be used, where x_1 is the contribution of the largest respondent and x_{2+} is the sum of the contributions of the third, fourth, etc., respondents.

3.2. Disclosure limitation techniques

There are four principal methods for disclosure limitation in tabular data: data rounding, data perturbation, cell suppression, and modification of the underlying microdata.

If frequency tabulations are rounded to an appropriate base (viz., a base larger than the definition of "small" counts), then disclosure has been limited. Similarly, perturbation of cell values by adding randomly selected small quantities makes the precise determination of small counts impossible. A difficulty with both methods is that modifying table counts independently can destroy additivity to totals (i.e., the rounding or perturbation is not *controlled*).

Cox and Ernst (1982) solve the controlled rounding problem for two dimensional tables. Causey, Cox, and Ernst (1985) and Cox (1987a) present procedures for performing two dimensional controlled rounding in an unbiased manner. Ernst (1989) and Causey, Cox, and Ernst (1985) demonstrate, respectively, that these results cannot be extended to three dimensions. Cox and George (1989) investigate the extent to which controlled rounding can be performed in two-way tables subject to row and column subtotal constraints. Cox, Fagan, Greenberg, and Hemmig (1986) demonstrate that the rounding and perturbation problems in two dimensions are identical mathematically, thereby extending these results to tables of perturbed counts.

Controlled rounding has been used by Statistics Canada and Statistics New Zealand. Data perturbation has been used by the U.K. Office of Population Censuses and Surveys.

Cell suppression is a disclosure limitation method that removes from publication all *disclosure cells* – those cells for which $S(X) > 0$ – plus sufficiently many additional cells – called *complementary suppressions* – to ensure that the values of the disclosure cells cannot be narrowly estimated through manipulation of additive relationships between cell values and subtotals, subtotals and totals, etc. Cell suppression strategies aim to minimize data loss as measured by the number of complementary suppressions, their total value, or a combination of these measures. In general, it is computationally infeasible to solve these problems optimally, so heuristic methods are often used.

Confidentiality protection in the 1977 and 1982 U.S. Economic Censuses was accomplished by means of a cell suppression algorithm called INTRA and a data management strategy based on mathematical lattices. INTRA produced optimal solutions to the problem of minimizing the number of complementary suppressions taken in a single two-way table in order to provide disclosure protection along individual rows and columns of the table. The INTRA algorithm was augmented by a heuristic procedure to deal with disclosure resulting from combination of row and column equations. The result was a disclosure pattern that provided sufficient protection, but in some cases involved more than the theoretical minimum number of complementary suppressions needed for full two dimensional protection. Three dimensional tables were handled by a heuristic procedure based on a well-defined procedure for stacking constituent two dimensional components of the table. The data management system proceeded from higher to lower levels of aggregation to create the two and three dimensional tables, apply INTRA, and keep track of suppressions. See Cox (1980) for details.

During this period, Statistics Canada developed a cell suppression algorithm and computer system called CONFID based on general linear programming (Robertson 1993). In lieu of creating tables, CONFID analyzes the full set of linear equations corresponding to the tabular structure from cells to subtotals, subtotals to totals, etc. Linear programming is used to select complementary suppressions that protect single disclosure cells one at a time, beginning with the cells requiring the most protection. Using an objective function that is approximately linear in the logarithm of the cell value, suppression patterns representing a compromise between minimum number of complementary suppressions and minimum total value suppressed are produced. Whereas INTRA protected an entire table in one operation, CONFID protects one cell at a time. CONFID has been used in Canadian economic and agriculture censuses and surveys since the 1980s.

In the 1987 and 1992 U.S. Economic Censuses, cell suppression was accomplished using a specialized linear programming structure called a *mathematical network* (Cox 1993a; Zayatz 1993). The change from INTRA was motivated by two criteria: to include the ability to protect against disclosure between row and column equations directly into a single algorithm, and to attempt to minimize total value suppressed, rather than minimizing number of complementary suppressions. Networks were selected because they meet the first criterion and, through careful selection of objective function, could be made to simulate the second in some cases (Cox 1987b). Networks offer an additional advantage: they run extremely fast computationally. The system design remained table based. Disclosure protection is provided one disclosure cell at a time. The design was generalized for the 1992 Economic Censuses to incorporate a collection of tables related hierarchically along one dimension into a single network. Cox (1995a) provides a detailed summary of disclosure limitation in U.S. economic censuses.

Methods based on mathematical graph theory have been proposed to *audit* two dimensional tables to identify disclosure and for complementary cell suppression (Gusfield 1988). These methods have not gained acceptance within statistical agencies, with the exception of the Australian Bureau of Statistics which uses them for table auditing.

The fourth confidentiality protection method for tabular data is to modify the underlying microdata prior to tabulation. One such approach, the *confidentiality edit*, was developed by the U.S. Bureau of the Census for the 1990 Decennial Census of Population and Housing. By this method, a sample of individual census records is selected, matched on key variables, and *switched* between small geographic areas. Official census counts and tabulations are based on the switched data file. Switching was selected in favor of rounding or suppression in order to avoid changing or suppressing small area counts on which reapportionment of legislative districts and decisions of national and regional importance are based (Griffin, Navarro, and Flores-Baez 1989).

3.3. What needs to be done

3.3.1. Standardized software

Research over the past 20 years into disclosure limitation methodology for tabular data has led to a body of proved disclosure limitation methods and operational computer systems. Unfortunately, this expertise and these resources are localized within a few organizations. Architectures for large scale computer systems are complex and most likely specialized within organizations and surveys, and therefore not transportable to other organizations. However, properly designed, transportable software implementing selected disclosure limitation methods could be developed without major incremental effort. We recommend that the following standardized software be developed and made broadly available within the U.S. and international community of statistical agencies:

* data rounding and perturbation in two dimensional frequency count tables
  - using mathematical networks (Cox and Ernst 1982; Cox, Fagan, Greenberg, and Hemmig 1986; Cox 1987a)
  - for tables with subtotals (Cox and George 1989)

* complementary cell suppression
  - network implementation (Sullivan 1992; Cox 1993a, 1995b)
  - network extended to tables related along one dimension (Cox and George 1989; Sullivan 1992)
  - CONFID (Robertson 1993)
  - network software to audit tables with suppressions (Cox 1995b).

Accompanying documentation should describe the statistical method, the operation of the software by a general user, and the organization and functional characteristics of the underlying computer code. This recommendation is consistent with that of Cox (1993b) and the major recommendations of the Federal Committee on Statistical Methodology (1994).

3.3.2. Research agenda

The development and dissemination of standardized software will greatly improve the state of practice in disclosure limitation for tabular data. Still, there remain interesting and important research problems. We focus on those problems listed below that promise to improve the efficiency of disclosure limitation, both in terms of the effort required and its effects on data quality, completeness and use. The data completeness problem is of particular importance in large tabulation structures such as censuses because cell suppressions tend to cascade from higher to lower levels of aggregation, with each suppression potentially necessitating additional complementary suppressions. The following research is recommended:

* determine which method, or a combination, among CONFID, network optimization, INTRA, and graph theoretic methods will produce the most efficient results across a representative set and number of cases;
* identify/develop the most efficient network, mathematical graph and linear programming algorithms for disclosure limitation;
* determine the extent to which networks can be extended to three dimensions and complex tabular structures;
* develop network based methods for protecting all/multiple disclosure cells in a two dimensional table simultaneously (Cox 1995b);
* embed networks in linear programming formulations, such as with a parallel computing architecture (Cox 1993a);
* embed INTRA in network formulations (Cox 1995b);
* develop methods to reduce oversuppression (Sullivan 1992; Carvalho, Dellaert, and Osorio 1994);
* develop linear programming methods for three dimensional tables (Zayatz 1993);
* investigate the computational complexity of cell suppression problems (Kelly, Golden, and Assad 1992);
* investigate publishing intervals in place of suppressed cells (Cox 1993b).

4. Disclosure Limitation in Microdata: Status and Research Agenda

The risk of disclosure for microdata can be significant (Cox, McDonald, and Nelson 1986; Bethlehem, Keller, and Pannekoek 1990). If an outside data user can in some way correctly link a respondent to a record on a microdata file, the statistical agency releasing that file has violated its data confidentiality pledge.

4.1. The disclosure problem

There are two main sources of the disclosure risk of a microdata file. One source of risk is the existence of high visibility records. Some records on the file may represent respondents with unique characteristics such as very unusual jobs (movie star, federal judge) or very large incomes (over one million dollars). An agency must do something to decrease the visibility of such records.

The second source of disclosure risk for microdata files is potential external matching files. There may be individuals or establishments in the population which possess a unique combination of the characteristic variables on the file. If some of those individuals or establishments happen to be chosen in the sample of the population that is represented on that file, there is a disclosure risk. Intruders potentially could use outside files which possess the same characteristic variables and identifiers to link these unique respondents to their records on the microdata file.

4.2. Disclosure limitation techniques

Obviously, a microdata file must be purged of all direct personal and institutional identifiers such as name, address, Social Security number, and Employer Identification number. Disclosure risk can be limited either by restricting the amount of published information or by disturbing the data.

Agencies releasing microdata files often set *geographic cut-offs* which are lower bounds on the size of the sampled population of each geographic region identified on microdata files (Greenberg and Voshell 1990). For example, the U.S. Bureau of the Census does not identify on its microdata files geographic regions containing less than 100,000 persons in the sampling frame. Statistical agencies often attempt to identify outside files which are potentially matchable to the microdata file in question. Comparability of all such files with the file in question must be examined.

Microrecords can be reidentified because of their high visibility or as the result of a concerted effort to match the microdata file to an outside file. In these cases, appropriate *topcodes*, bottomcodes, or recodes may be applied to continuous high visibility or high risk variables on a microdata file. For example, rather than publishing a record showing an income value of $2,000,000, the record may only show a representative value for the upper tail of the distribution, such as the cut-off value for the tail or the median value for the tail.

Some agencies have review panels which recommend and facilitate the use of disclosure limitation techniques. For example, the U.S. Bureau of the Census established a Microdata Review Panel in 1981 (Cox, McDonald, and Nelson 1986).

4.3. What needs to be done

4.3.1. Standard software

Standard software for disclosure analysis and disclosure limitation in microdata would be of considerable benefit within and across agencies. The software would ensure that disclosure limitation techniques were carried out correctly and would serve as a data analysis and quality assurance tool for review panels. An integrated disclosure limitation setting involving software that could perform the following tasks would be useful:

* disclosure analysis
  - produce distributions of variables
  - perform cross tabulations of sets of variables
  - estimate the number of population uniques in a sample data set
  - characterize sample uniques under alternative data release options
  - tabulate various risk measures under alternative data release options

* disclosure limitation
  - data disturbance techniques
    - · noise inoculation (Skinner 1991; McGuckin and Nguyen 1988; Fuller 1993; Kim 1986, 1990; Voshell 1990)
    - · blurring (Strudler, Oh, and Scheuren 1986)
    - · blanking and imputation (Griffin et al. 1989)
    - · swapping (or switching) (Dalenius 1988; Griffin et al. 1989)
    - · microaggregation (Govoni and Waite 1985)
  - other disclosure limitation techniques
    - · topcoding
    - · subsampling
    - · rounding
    - · categorization of variables subject to a minimum category size.

One proposed way of representing and implementing these techniques is *matrix masking* – the representation of disclosure limitation methods in terms of matrix algebra to facilitate their implementation and analysis (Duncan and Pearson 1991; Cox 1991, 1994). Matrix masking should be explored as a format for representing, implementing and comparing microdata disclosure limitation methods.

4.3.2. Research agenda

Successful research in the following areas should increase the ability of statistical agencies to release microdata subject to confidentiality constraints.

* probability based measures of disclosure risk for microdata
* assessing disclosure risk and usefulness of expanded microdata release formats
  - establishment based microdata
  - longitudinal microdata
  - appending contextual variable data to microdata (Saalfeld, Zayatz, and Hoel 1992)
  - synthetic microdata (Rubin 1993)
  - hierarchical data files
* encrypted data bases
* reidentification studies.

Defining and assessing disclosure in microdata needs to be put on a sound statistical footing. Probability theory provides an intuitively appealing framework for defining disclosure in microdata in which disclosure is related to probability of reidentification (Skinner, Marsh, Openshaw, and Wymer 1990; Duncan and Lambert 1989; Mokken, Kooiman, Pannekoek, and Willenborg 1991; Paass 1988; Cox and Kim 1991; Spruill 1982). On this basis, disclosure limitation could be measured quantitatively and extensions, such as analysis based on prior knowledge, could be incorporated. Part of this research involves developing a method of estimating the percentage of records on a microdata file which represent unique persons or establishments in the population (Greenberg and Zayatz 1991; Skinner and Holmes 1993; Bethlehem, Keller, and Pannekoek 1990). Another part involves developing a measure of marginal risk for each variable on a microdata file and analyzing changes in risk that result from changes in detail of variables.

Disclosure risk and limitation of some particular types of microdata need to be carefully examined. The feasibility of releasing establishment based microdata such as from economic and agriculture censuses and surveys should be investigated and models for release or administrative alternatives proposed. The additional risk in releasing longitudinal microdata files where data from different time periods can be linked for each respondent needs to be quantified. A study is needed which will identify an affordable method of generating microdata files with contextual variables which will not lead to identification of small areas (Saalfeld, Zayatz, and Hoel 1992). The quality and usefulness of synthetic data with the statistical properties of real data need to be examined. Research is needed to quantify the additional risk in releasing hierarchical data files (e.g., children within schools).

A study is needed to analyze the feasibility of storing and allowing access to microdata in an encrypted data base with security controls and inferential disclosure limiting techniques (McNulty and Unger 1989; Lunt 1993). Users of the data could not access individual records in the data base but could obtain results of statistical analyses of the data.

The principal risk in releasing microdata is that a data spy will be able to match microrecords to another file containing identifiable information with reasonable accuracy. It would be beneficial to attempt to link two files that overlap in terms of respondents represented and analyze the outcome (Blien, Wirth, and Muller 1991). One way would be to focus on one-to-one exact record matches between the two files based on a set of overlapping categorical variables. A controversial proposal involves hiring an "intruder" to attempt to link respondents to their microrecords. It would be useful to see how an intruder might approach the problem, whether or not any correct matches could be made, and if correct matches were made, how long did the process take and how much work was required. Research on the kinds of external files available for matching would also be useful.

5. Other Issues

5.1. Multiple modes of data release

When microdata are released from a census or survey, tabulations from the same data are often released. The resulting problem is that of ensuring confidentiality protection when both microdata and tabulations are made (publicly) available. This problem has several facets.

If sample weights are recorded on the microdata, and if these weights have not been changed during disclosure limitation, then discrepancies between released tabulations and tabulations computed from the microdata can potentially lead to disclosure. If released tabulations are computed from the microdata file, then, for the most part, this problem disappears. However, to do so is likely to introduce unnecessary distortion into the released tabular data. From a data quality and usefulness standpoint, it is desirable that released tabulations and those computed from the microdata conform as closely as possible to one another. This might be accomplished using the controlled rounding and perturbation methodologies discussed earlier. However, that leads potentially to the problem of having provided the data user with the means to reconstruct suppressed entries in the released tabulations. These problems remain largely unexamined. Research that simulates these problems in realistic scenarios is needed to assess the scope of these problems.

5.2. *Balancing confidentiality protection with data quality and usefulness*

Meaningful, broadly useful analyses on disclosure-limited data need to be performed and documented to assess the quality, completeness, usefulness and ease of use of disclosure-limited data products. The first step is to develop quantitative measures of data quality and utility. Such measures should quantify the amount of data reduction (e.g., suppression, geographic cutoffs, topcoding), the amount of data distortion (e.g., rounding, perturbation, switching), and effects of disclosure limitation on relationships, or independence, among data items.

The second step is to apply these measures to data before and after disclosure limitation to assess the degradation in these factors caused by disclosure limitation. The final step is to develop new or refined disclosure limitation methods whose effects are acceptable and, if not minimal, at least balanced with regard to confidentiality concerns.

Related measures need to be developed for the effects of disclosure limitation methods on confidentiality protection. Ideally, such measures would be based on a quantitative model of *disclosure risk.* Data disturbance techniques such as data perturbation or swapping have been used for microdata. Research is needed to investigate the protection provided by data disturbance techniques and the usefulness of the resulting microdata. Data users should be consulted when deciding upon a protection method for frequency data. Users also should be consulted as to the benefit of publishing ranges or median values for suppressed cells in tables of magnitude data. Agencies expend considerable human and other resources in collecting and publishing data. Agencies should work with data users when designing data release strategies to ensure that as much of this valuable information is released as possible without disclosing the identities of respondents. Alternative forms of data release (e.g., releasing ranges in lieu of blanking suppressed data; releasing moment matrices or multiply imputed simulated microdata in lieu of high-risk microdata) should be investigated.

5.3. *Uniform theoretical framework for disclosure*

A final issue is whether a uniform theoretical framework is needed for disclosure limitation. Theory for disclosure limitation in tabulations is based primarily on linear algebra and extensions. While there are notable exceptions (e.g., Duncan and Lambert 1986, 1989; Greenberg and Zayatz 1991; and Skinner and Holmes 1993), a general theoretical framework for microdata disclosure risk and disclosure limitation is lacking. It is conceivable that a unified theoretical framework for both (and perhaps other) data types would provide a solution to the combined release problem. While it is not clear that such a framework would enhance research and methods in tabular disclosure limitation, its pursuit is of interest and would likely strengthen the theoretical underpinnings of microdata disclosure limitation.

6. References

American Statistical Association, Committee on Professional Ethics (1989). Ethical Guidelines for Statistical Practice. American Statistical Association, Alexandria, VA.

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

Blien, U., Wirth, H., and Muller, M. (1991). Disclosure Risk for Microdata Stemming From Official Statistics. Statistica Neerlandica, 46, 69–82.

Carvalho, F. de, Dellaert, N., and Osorio, M. de Sanches (1994). Statistical Disclosure in Two-Dimensional Tables: General Tables. Journal of the American Statistical Association, 89, 1547–1557.

Causey, B., Cox, L., and Ernst, L. (1985). Applications of Transportation Theory to Statistical Problems. Journal of the American Statistical Association, 80, 903–909.

Cox, L. (1980). Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association, 75, 377–385.

Cox, L. (1981). Linear Sensitivity Measures and Statistical Disclosure Control. Journal of Statistical Planning and Inference, 5, 153–164.

Cox, L. (1987a). A Constructive Procedure for Unbiased Controlled Rounding. Journal of the American Statistical Association, 82, 520–524.

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

Cox, L. (1991). Comment (on Duncan, G. and Pearson, R., "Enhancing Access to the Microdata While Protecting Confidentiality: Prospects for the Future"). Statistical Science, 6, 232–234.

Cox, L. (1993a). Solving Confidentiality 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, EUROSTAT, 229–245.

Cox, L. (1993b). Discussion. Proceedings of the Annual Research Conference, U.S. Bureau of the Census, Washington, DC: U.S. Department of Commerce, 132–135.

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

Cox, L. (1995a). Protecting Confidentiality in Establishment Surveys. Survey Methods for Businesses, Farms and Institutions. New York: John Wiley & Sons, 443–473.

Cox, L. (1995b). Network Models for Complementary Cell Suppression. Journal of the American Statistical Association, to appear.

Cox, L. and Ernst, L. (1982). Controlled Rounding. INFOR: Canadian Journal of Operation Research and Information Processing, 20, 423–432.

Cox, L., Fagan, J., 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, 388–393.

Cox, L. and George, J. (1989). Controlled Rounding for Tables with Subtotals. Annals of Operations Research, 20, 141–157.

Cox, L. and Kim, J. (1991). Thwarting Unique Identification in Microdata Files: A Proposal for Research. Statistical Research Division, U.S. Bureau of the Census (unpublished).

Cox, L., McDonald, S.-K., and Nelson, D. (1986). Confidentiality Issues at the United States Bureau of the Census. Journal of Official Statistics, 2, 135–160.

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

Dalenius, T. and Denning, D. (1982). A Hybrid Scheme for Release of Statistics. Statistisk tidskrift, 20, 97–102.

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

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

Duncan, G. and Pearson, R. (1991). Enhancing Access to the Microdata While Protecting Confidentiality: Prospects for the Future (with comment). Statistical Science, 6, 219–239.

Ernst, L. (1989). Further Applications of Linear Programming to Sampling Problems. Proceedings of the Section on Survey Research Methods, American Statistical Association, 625–630.

EUROSTAT (1993). Proceedings of the International Seminar on Statistical Confidentiality, Luxemburg.

Federal Committee on Statistical Methodology (1994). Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology. Washington, DC: U.S. Office of Management and Budget.

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

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

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

Greenberg, B. and Zayatz, L. (1991). Strategies for Measuring Risk in Public Use Microdata Files. Statistica Neerlandica, 46, 33–48.

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

Gusfield, D. (1988). A Graph Theoretic Approach to Statistical Data Security. SIAM Journal on Computing, 17, 552–571.

International Statistical Institute (1986). Declaration of Professional Ethics. International Statistical Review, 54, 227–242.

Kelly, J., Golden, B., and Assad, A. (1992). Cell Suppression: Disclosure Protection for Sensitive Tabular Data. NETWORKS, 22, 397–417.

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, 370–374.

Kim, J. (1990). Masking Microdata for National Opinion Research Center. Final Project Report, U.S. Bureau of the Census, Washington, DC.

Lunt, T. (1993). Discussion: Computer Security. Journal of Official Statistics, 9, 507–510.

McGuckin, R. and Nguyen, S. (1988). Use of "Surrogate Files" to Conduct Economic Studies with Longitudinal Microdata. Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census, Washington, DC: U.S. Department of Commerce, 193–209.

McNulty, S. and Unger, E. (1989). The Protection of Confidential Data. Computer Science and Statistics: Proceedings of the 21st Symposium on the Interface, American Statistical Association, 215–219.

Mokken, R., Kooiman, P., Pannekoek, J., and Willenborg, L. (1991). Microdata and Disclosure Risks. Statistica Neerlandica, 46, 49–67.

National Research Council (1993). Private Lives and Public Policies – Confidentiality and Accessibility of Government Statistics. Washington, DC: National Academy Press.

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

Robertson, D. (1993). Cell Suppression at Statistics Canada. Proceedings of the Annual Research Conference, U.S. Bureau of the Census, Washington, DC: U.S. Department of Commerce, 107–131.

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

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, 691–696.

Singer, E. (1993). Recent Research on Confidentiality Issues at the Census Bureau. Proceedings of the Annual Research Conference, U.S. Bureau of the Census, Washington, DC: U.S. Department of Commerce, 97–106.

Skinner, C. (1991). Statistical Disclosure Issues for Census Microdata. Statistica Neerlandica, 46, 21–32.

Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. (1990). Disclosure Avoidance for Census Microdata in Great Britain. Proceedings of the Sixth Annual Research Conference, U.S. Bureau of the Census, Washington, DC: U.S. Department of Commerce, 131–143.

Skinner, C. and Holmes, D. (1993). Modelling Population Uniqueness. Proceedings of the International Seminar on Statistical Confidentiality, EUROSTAT, 175–199.

Spruill, N. (1982). Measure of Confidentiality. Proceedings of the Section on Survey Research Methods, American Statistical Association, 260–265.

Strudler, M., Oh, H., 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, 375–381.

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

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

Zayatz, L. (1993). Using Linear Programming Methodology for Disclosure Avoidance Purposes. Proceedings of the International Seminar on Statistical Confidentiality, EUROSTAT, Luxemburg, 341–351.

Received November 1993
Revised December 1994