Invited Paper

Confidentiality and Data Protection Through Disclosure Limitation: Evolving Principles and Technical Advances[1]

Stephen E. Fienberg[2]

ABSTRACT

Confidentiality and privacy are widely conceived of as ethical matters and they impinge directly upon the work of statisticians and statistical agencies. But providing access to publicly-collected data is also an ethical matter and the goal of agencies should be to release the maximal amount of information without undue risk of disclosure of individual information. Assessing this tradeoff is inherently a statistical one as are the development of methods to limit disclosure risk. Yet, until recently, they have received limited attention from statistical methodologists. That situation has changed considerably in the past decade. This paper addresses some ethical dimensions of confidentiality and privacy as they relate to statistical activities and we outline some of the evolving principles that are guiding the development of statistical methodology in this area. The paper also describes how these research methods relate to a data access query system being developed for use by statistical agencies in the United States.

KEY WORDS: confidentiality, privacy, access, disclosure limitation, contingency tables

1. Introduction and Themes

The explosion of computerized data bases containing financial and health care records and the vulnerability of data bases accessible via the Internet has heightened public attention and generated fears regarding the privacy of personal data. In particular, the public is wary of what government might do with such data (e.g., see a recent commentary on the issue which appeared in *The New York Times* by Berke (2000)). Little or no distinction is made in the public eye between privately held data bases, administrative data bases, and statistical data bases (especially those gathered and managed by statistical agencies). The general public disquiet regarding privacy has heightened attention to issues of confidentiality and privacy in government statistical agencies, even though agencies long ago recognized the ethical dimensions of confidentiality and privacy and legislatures wrote protection for confidentiality into the laws governing the operation of agencies.

But providing access to data collected either directly under government auspices or at public expense is also an ethical matter as Fienberg, Martin, and Straf (1985) and others argue, especially since such data can be viewed as a public good. These two sets of ethical concerns conflict, at least somewhat, and most statisticians have argued that the goal of statistical agencies should be to release the maximal amount of information without undue risk of disclosure of individual information. Assessing this tradeoff is inherently a statistical matter as are the development of methods to limit disclosure risk. As a consequence, I believe that confidentiality

[1] Presented at the 2000 International Conference on Statistics, Development and Human Rights, 4-8 September 2000, Montreux, Switzerland.

[2] Professor, Department of Statistics and Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213-3890, U.S.A.. Email: fienberg@stat.cmu.edu

and disclosure limitation are inherently statistical issues even though they have not always been treated as such by those concerned with privacy and related ethical matters, or by official statisticians charged with protecting confidentiality. This makes the topic an especially appropriate one for inclusion in the program of a conference dealing with statistics and human rights.

Government statisticians and others have often argued that confidentiality of data provided by establishments is a greater concern than the confidentiality of data provided by individuals. The question is whether this is an ethical, legal, or simply practical concern. I can see no obvious human rights issue associated with the release of information on establishments per se, and I believe that it would be easy to argue that there is no inherent right to privacy for establishments as there is for individuals. If one accepts this argument, then confidentiality for establishments raises no major ethical issue for statisticians. But in many ways data on enterprises pose greater technical problems for disclosure limitation because some classifications of enterprises are often dominated by one or two of their members, and reported data are typically weighted by establishment size. Further longitudinal data on establishments need to cope with changes such as mergers and acquisitions, etc. Further agencies often find themselves in the ironic position of attempting to conceal through isclosure limitation methodology information about specific establishments that the latter have already released publicly, e.g., through annual reports to shareholders or filings with the Securities and Exchange Commission or some other government regulatory agency. For the remainder of this paper I restrict attention to issues associated with confidentiality and disclosure limitation for individuals or groups of individuals such as families.

In section 2 I elaborate on the ethical themes associated with confidentiality and privacy and address, albeit briefly, the issue of whether we should release restricted data to achieve confidentiality objectives or whether we should simply restrict access. Both approaches have as their goal disclosure limitation. I am an advocate for unrestricted access to as much data as it is possible to release. Thus, I attempt to summarize the case for unlimited access to restricted data as an approach to limit disclosure risk, but not so much as to impair the vast majority of potential research uses of the data.

For far too long, confidentiality and disclosure limitation were relegated to the non-statistical part of large-scale data collection efforts and, as a consequence, the methods used to address them were ad hoc and conservative. Beginning with a 1977 paper by Dalenius (1977) and a detailed and forward-looking 1978 report of the Subcommittee on Disclosure-Avoidance Techniques of the Federal Committee on Statistical Methodology Subcommittee on Disclosure-Avoidance Techniques (1978)), statisticians slowly began to address the issues in a systematic fashion. Twenty years later, we can look back and take stock of the growth of statistical activity and ideas in this area--e.g., by examining the recent 1994 report of the Federal Statistical Methodology Subcommittee on Disclosure-Avoidance Techniques, the report of a panel of the Committee on National Statistics (1993), and two special issues of the *Journal of Official Statistics*, in 1993 and 1998, as well as the 1999 proceedings of the 1998 Lisbon, Portugal Conference on Statistical Data Protection.

In section 3, I provide an overview of some current methods in use for data disclosure limitation and statistical principles that underlie them. In section 4, I relate recent research ideas on bounds for categorical data (one of my special research interests) to a new database query system in development for use by US statistical agencies under the auspices of the National Institute of

Statistical Sciences. Sections 3 and 4 draw directly on material from Fienberg (2000). I end with an overview of disclosure limitation methodology principles and a discussion of ethical issues and confidentiality concerns raised by new forms of statistical data.

2. Ethical Issues in Confidentiality: Restricted Access Versus Restricted Data

Many authors have talked about the inherent tradeoff between data protection and data access. The discussion in this section draws heavily on related discussions in Fienberg ((1997), (1998a)).

The probabilistic notion of disclosure, due originally to Dalenius (1977), suggests that any release of actual data should produce disclosure at some level, since the released data should increase the information available about individuals in the database. This is in essence a statement about the changing conditional probability of identification of individuals as one condition on increasing amounts of information.

There are actually two types of disclosure, exact and inferential. For exact disclosure we talk about disclosing, with probability one, the identity of an individual respondent and thus various attributes of that individual, or simply disclosure resulting from attributes possessed by a group of individuals of whom the target is one. This can happen in various ways. But more often than not we infer such identity and/or attributes, but with probability less than 1. Implicit in almost all of the recent research on the topic is the role of the unidentified intruder who had data to match against the released data files (e.g., see Lambert (1993) and Fienberg, Makov and Sanil (1997)). Thus, the intruder's goal is to effect identification and thereby create linked files.

As Figure 1 from Fienberg (2000) depicts in a schematic fashion, not all data disclosures breach promises of confidentiality to respondents, since release of data for those in a sample increases the information available for those not in the sample, and an intruder can cause harm to those whose data are not released by falsely identifying someone in a data base. See Fienberg (1997) for further discussion along these lines.

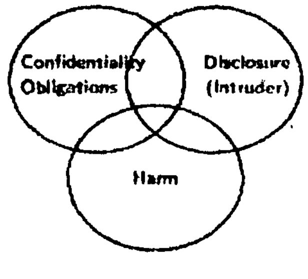

Figure 1. Relationships among confidentiality, disclosure, and harm

Further, promises of confidentiality represent a form of contractual arrangement and they can be breached either by the agency or by the respondent. For example, if I announce publicly that I was a member of the Current Population Survey (CPS) sample during a particular period of time, as I was, then I have broken the contractual arrangement. My survey records will not necessarily be identifiable from a CPS data release, but the probability of an intruder being able to identify my records will have increased by a factor roughly proportional to the inverse of the sampling fraction! As a consequence, the U.S. Bureau of the Census, which collects the CPS data, has a somewhat different obligation to me than it did before I announced my participation in the survey. What is even more problematic, is that by announcing my participation in the CPS I have affected the probability of identity of others in the sample because of the specific cluster design utilized in the CPS, and this raises ethical issues about my behavior and problems for the agency.

There are two different philosophies that people adopt with regard to the preservation of confidentiality associated with individual-level data: (1) *restricted* or *limited information*, wherein the amount or format of the data released is subject to restrictions, and (2) *restricted* or *limited access*, wherein the access to the information is itself restricted. I have argued elsewhere (e.g., see Fienberg (1997), (1998a)) that federal statistical data are a public good and that the federal statistical agencies have a *responsibility* to provide wide and unrestricted access to data that might be of value to secondary users outside the agencies themselves. Restricted access should only be justified in extreme situations where the confidentiality of data in the possession of an agency cannot be protected through some form of restriction on the information released. For a discussion of some of the benefits of restricted access, see David (1998), and for a response regarding why restricted access approaches leave far too much to be desired, see Fienberg (1998a).

Government statistical data such as those gathered as part of censuses and major sample surveys meet two key tests that are usually applied to quantities labeled as public goods: jointness of consumption (consumption by one person does not diminish their availability to others), and benefit to the nation as a whole (statistical data are used to inform public policy and as the basis for democratic representation). The only issue, then, is whether or not there is non-exclusivity, i.e., whether or not it makes sense to provide these statistical data to some citizens and not to others. If we have means for providing access to all or virtually all in society, e.g., via the Internet and the World Wide Web, then the costs of providing the data to all is often less than the costs of restricting access. There are other perhaps hidden costs, however, that result from expanded use to those who produce the data. Fienberg, Martin, and Straf (1985) provide a general discussion of the costs and benefits of data sharing. Duncan, Jabine, and de Wolf (1993) and Duncan (1995) give a more focussed discussion relevant to the present context. My view is not only that restricting access to a public good produces bad public policy but that it cannot work effectively. This is primarily because the gate keepers for restricted data systems have little or no incentive to widen access or to allow research analysts the same freedom to work with a data set and share results as they are accustomed to having with unrestricted access. Just imagine the difficulty the researchers would have if they are accustomed to reporting residual plots and other information that allows for a partial reconstruction of the original data, at least for some variables, since restricted data centers typically do not allow users to take such information away. Thus, for me, the question is not if we should continue to supply public-use microdata, but how. For that we need tools for disclosure limitation that have as their output usable statistical data bases.

3. Methodology for Disclosure Limitation

Duncan (2000) categorizes the methodology used for disclosure limitation in terms of *disclosure limiting masks*, i.e., transformations of the data where there is a specific functional relationship (possibly stochastic) between the masked values and the original data. For example, here is a general class of methods for disclosure limitation that is referred to as *matrix masking* by Duncan and Pearson (1991). The idea is to think in terms of transforming an $n \times p$ data matrix Z through pre- and post-multiplication and possible addition of noise, i.e.,

$$Z \rightarrow AZB+C \qquad (1)$$

where A is a matrix that operates on cases, B is a matrix that operates on variables, and C is a matrix that adds perturbations or noise. Matrix masking includes a wide variety of standard approaches to disclosure limitation:

-adding noise,
-releasing a subset of observations (delete rows from Z),
-*cell suppression* for cross-classifications,
-including simulated data (add rows to Z),
-releasing a subset of variables (delete columns from Z), and
-switching selected column values for pairs of rows (*data swapping*).

Even when one has applied a mask to a data set, the possibilities of both identity and attribute disclosure remain, although the risks may be substantially diminished.

Duncan suggests that we can categorize most disclosure limiting masks as suppressions (e.g., cell suppression), recodings (e.g., collapsing rows or columns, or swapping), or samplings (e.g., releasing subsets), although he also allows for simulations (as above). Further some masking methods alter the data in systematic ways, e.g., through aggregation or through cell suppression, and others do it through random perturbations, often subject to constraints for aggregates. Examples of perturbation methods are *controlled random rounding, data swapping*, and the recently proposed *post-randomization method* or PRAM of Gouweleeuw, et al. (1998) and generalized by Duncan and Fienberg (1999). One way to think about random perturbation methods is as a restricted simulation tool, and thus we can link them to other types of simulation approaches that have recently been proposed.

Fienberg, Makov, and Steele (1998) pursue this simulation strategy and present a general approach to "simulating" from a constrained version of the cumulative empirical distribution function of the data. In the case when all of the variables are categorical, the cumulative distribution function is essentially the same as the counts in the resulting cross-classification or contingency table. As a consequence, we think of this general simulation approach as equivalent to simulating from a constrained contingency table, e.g., given a specific set of marginal totals and replacing the original data by a randomly generated one drawn from the "exact" distribution of the contingency table under a log- linear model that includes "confidentiality-preserving" margins among its minimal sufficient statistics. Actually, Fienberg, et al. (1998) propose retaining the simulated table only if it is consistent with some more complex log-linear model. This approach offers the prospect of simultaneously smoothing of the original counts *and* providing disclosure limitation protection.

Rubin (1993) asserted that the risk of identity disclosure can be eliminated by the use of synthetic data (in his case using Bayesian methodology and multiple imputation techniques) since there is no direct function link between the original data and the released data. Or said another way, we have no confidentiality problem since we have replaced all of the real individuals with simulated ones. But some simulated individuals may be virtually identical to original sample individuals in terms of their data values, or at least close, and thus the possibility of both identity and attribute disclosure remain.

Another extremely important feature of the simulation methodology just described is that information on the variability is directly accessible to the user. For example in the Fienberg, Makov, and Steele (1998) approach for categorical data, anyone can begin with the reported table and information about the margins that are held fixed, and then run the Diaconis-Sturmfels Markov chain algorithm to regenerate the full distribution of all possible tables with those margins. This then allows the user to make inference about the added variability in a modeling context in a form that is similar to the approach to. inference in Gouweleeuw, et al. (1998). Similarly, Rubin (1993) proposes the use of multiple imputations to get direct measure of variability associated with the posterior distribution of the quantities of interest. As a consequence, simulation and perturbation methods represent a major improvement from the perspective of access to data over cell suppression and data swapping. And they conform to a statistical principle of allowing the user of released data to apply standard statistical operations without being misled.

Many practical questions remain regarding the use and efficacy of such simulation methods for generating disclosure-limited public-use samples. For example,

-How effective are such devices for limiting disclosure, i.e., protecting against intruder?
-What is information loss when we compare actual data with those released?
-In the case of categorical data, how can they be used when the full cross-classification of interest is very sparse, consisting largely of 0s and 1s?
-How can we use models to generate the simulated data when the users have a multiplicity of models and even classes of models which they would like to apply to the released data?

Among the tools for risk assessment are various approaches for estimating the number of uniques in a population or more precisely the conditional probability of an individual being unique in the population given that he/she is unique in a sample. For two quite different but nonetheless related approaches to this problem, see Fienberg and Makov (1998) and Skinner and Holmes (1998) who actually provide per-record assessments of risk for the categorical response case. Samuels (1998) offers a novel way to look at the problem of uniqueness via some urn models arising in genetics problems and his approach is extended by Fienberg and Makov (2000). To go with risk assessment we also need information on the trade-off of gains versus the risks. Few have attended to this issue. The most interesting example arises in the context of a paper by Pannekoek and de Waal (1998), who suggest reporting empirical-Bayes-like mixtures of the observed data and smoothed versions of them for small area categorical data. Following their paper, Zaslavsky and Horton (1998) discuss how to evaluate the trade-off between disclosure risk in their approach and the loss due to non-publication. Much more work needs to be done, however, on both risk assessment and its trade-off with the gains resulting from expanded access.

It is worth noting that the perspective and principles elucidated here explain a special dimension of the current debate in the United States over the use of sample-based adjusted counts in the 2000 decennial census. In 1990, approximately 1 in 10 persons was not properly counted as a result of both errors of omissions and erroneous enumerations and other counting errors (e.g., see Anderson and Fienberg (1999) and Anderson et al. (2000)). The problem with such errors is that they are systematic and not uniformly distributed, either geographically or demographically. The U.S. Bureau of the Census plans to adjust the enumeration counts using the results of a large sample survey of households in randomly selected blocks. The resulting adjusted counts should not only have reduced bias but also should provide direct disclosure limitation protection for individuals and families in planned census data releases. The Census Bureau still plans to apply additional disclosure methods to the data as well (e.g., see Steel and Zayatz (1999)).

4. A Pilot Query System for Public Data Access

The National Institute of Statistical Sciences (NISS) has recently assembled a team of statistical researchers from multiple universities who have begun to work with statisticians in U.S. statistical agencies. These researchers are developing a Web-based query system that allows the use of disclosure limitation methods applied sequentially in response to a series of statistical queries in which the public knowledge of releases is cumulative.

The query system idea draws in part on a pilot project described in Keller- McNulty and Unger (1998), and it will use as tools the various disclosure limitation methods being developed in the literature. The idea is to fully automate the methods through algorithms and explore intruder behavior (c.f., Fienberg, Makov, and Sanil (1997)) and to utilize alternative approaches to risk assessment.

To get a sense of how this system *might* use the ideas on simulated data bases, consider a database consisting of a k-dimensional contingency table, for which the queries are only allowed to come in the form of requests for marginal tables. What is intuitively clear from statistical theory is that, as margins are released and cumulated by users, there is increasing information available about the table entries.

In response to a new query, the system now examines it in combination with all those previously released margins and decides if the risk of disclosure of individuals in the full unreleased table is too great. Then it might offer one of three responses: (1) yes---release; (2) no---don't release; or perhaps (3) simulate a new table, which is consistent with the previously released margins, and then release the requested margin table from it. Because released margins need to be consistent and even simulated, releases become highly constrained.

How might such a system evaluate the risk of disclosure from the release of a new margin? A number of researchers have recently been working on the problem of determining upper and lower bounds on the cells of the cross-classification given a set of margins. This is in one sense an old problem (at least for two-way tables) but it is also deeply linked to recent mathematical statistical developments and thus has generated a flurry of new research (e.g., see Buzzigoli and Giusti (1999), Fienberg (1999), and Roehrig, et al. (1999)).

Consider a 2 x 2 table of counts, $\{n_{ij}\}$, with given the marginal totals, $\{n_{1+}, n_{2+}\}$ and $\{n_{+1}, n_{+2}\}$:

$$\begin{array}{cc|c} n_{11} & n_{12} & n_{1+} \\ n_{21} & n_{22} & n_{2+} \\ \hline n_{+1} & n_{+2} & n \end{array}$$

The marginal constraints, i.e., that the counts in any row add to the corresponding one-way total, plus the fact that the counts must be non-negative imply bounds for the cell entries. Specifically, for the (i,j) cell, we have

$$\min\{n_{i+},n_{+j}\} \geq n_{ij} \geq \max\{n_{i+}+ n_{+j}-n,0\}. \qquad (2)$$

Bounds such as those in equation (2) usually are referred to as Fréchet bounds after the French statistician Maurice Fréchet who described them in a 1940 paper (see Fréchet (1940)), but they were independently described by Bonferroni and Hoeffding at about the same time. They have been repeatedly rediscovered by a myriad of others. Such bounds and their generalizations lie at the heart of a number of different approaches to disclosure limitation including cell suppression, data swapping and other random perturbation methods, and controlled rounding (e.g., see the discussion by Cox (1999)).

Fienberg (1999) describes these bounds and several of their multi-dimensional generalizations, and explains some of the links between them and the modern statistical theory of log-linear models for the analysis of contingency tables (see Bishop, Fienberg, and Holland (1975), Haberman (1974), and Lauritzen (1996)). Dobra and Fienberg (2000a) have now provided full details for the explicit calculation of sharp upper and lower bounds when the released margins correspond to the sufficient statistics of the *decomposable* sub-family of log-linear models (called *direct* models by Bishop et al. (1975)) as well as a number of important extensions, especially within the general family of *graphical* log-linear models described in Lauritzen (1996).

For some *non-graphical* models we can also construct explicit and sharp bounds. For example, Dobra and Fienberg (2000b) construct a non-iterative algorithm for bounds for a k-dimensional table given all (k-1)-dimensional margins. In the special case of 2^i tables with (2^{i-1})-dimensional margins fixed, they explain how these bounds result from a natural extension of the Fréchet bounds which ties directly to log-linear model theory. In Figure 2, we depict the constraining nature of such bounds for a simple 3-dimensional cross-classification with fixed two-dimensional margins.

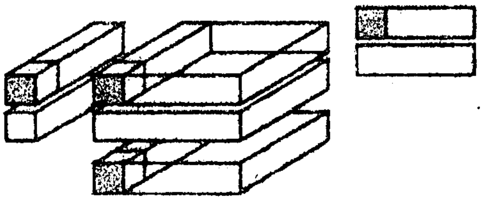

Figure 2. The marginal constraints for cells in a three-way table given two-way marginals

This brings us back to the notion of the development of a query system for cross-classifications of non-negative counts and the release of margins in response to successive queries. Such a system will rely on computationally efficient methods for the calculation of bounds, which we expect will rely on some of the theory outlined above, as well as other measures of disclosure risk. And it will need to examine methods for disclosure limitation going well beyond the simulation methods described briefly in Section 3.

It is important to note that a sequential query system need not be restricted to categorical variables nor to queries that come in the form of requests for tables. NISS plans to develop a basic query system and test it with one or more public-use microdata files, exploring intruder behavior in a variety of ways as well as different tools for disclosure limitation and the assessment of risk, e.g., the Argus approach developed at Statistics Netherlands by Hundepool, et al. (1998a), (1998b) and Willenborg and De Waal (1996). While there are many theoretical and empirical issues to explore and many exciting research questions to address, making such a system function, with actual agency data bases, offers the real future prospect of improved disclosure limitation *and* increased data access.

5. Conclusions and Further Issues

In this paper, I have focused on the interplay between the ethical issues of confidentiality and privacy, on the one hand, and access to publicly-collected data on the other. I explained why disclosure is an inherently statistical issue, i.e., one cannot eliminate the risk of disclosure, simply reduce it, unless one restricts access to the data. Then I outlined some of the complex relationships between promises of confidentiality to respondents in surveys or participants in studies and the nature of disclosure of information about those respondents. Because techniques for disclosure limitation are inherently statistical in nature I explained why they must be evaluated using statistical tools for assessing the risk of harm to respondents.

I then turned to the current array of statistical methods used to limit disclosure, especially those representable in terms of disclosure limitation masks, distinguishing among suppression, recoding, and sampling approaches, on the one hand, and systematic versus perturbational approaches on the other. Among the principles that have been the focus of much of the recent effort in disclosure limitation methodology are:

-*usability*, i.e., the extent to which the released data are free from systematic distortions that impair statistical methodology and inference.

-*transparency*, i.e., the extent to which the methodology and practice of it provide direct or even implicit information on the bias and variability resulting from the application of a disclosure limitation mask.

-*duality*, i.e., the extent to which the methods aim at both disclosure limitation and making the maximal amount of data available for analysis.

In particular, I described how these principles fit with recent proposals for the release of simulated data for release.

The role of marginal bounds for multi-way contingency tables raises new statistical issues, and I outlined a few of these in Section 4 in the context of a project organized by the National Institute of Statistical Sciences for evaluating competing approaches using a real-time sequential query-based system.

At the outset we pointed to the public concerns about data bases financial and health records. Such concerns are well-founded even with regard to survey data gathered by statistical agencies. For many years various health surveys have also included direct measurements of health status based on tests including those involving the drawing of blood samples. In the past, statisticians have taken the samples and then recorded simple summaries such as white or red blood cell counts. But technology and biological knowledge have advanced and we now must face the prospect of including in statistical data bases genetic sequencing information which, in principle, can uniquely identify individuals. These issues have already provoked considerable controversy in the context of data associated with stored tissue samples at the U.S. Centers for Disease

Control as Clayton, et al. (1995) have described, and privacy issues regarding genetic information have been the subject of a set of papers in a special 1999 issue of the legal journal *Jurimetrics*. New issues of access and confidentiality loom on the horizon as those engaged in health research consider data elements that consist of functional Magnetic Resonance Imaging or full body scan images, as well as blood, tissue, and other genetics-related samples (see the discussion of multiple media data in Fienberg (1998a)). It is essential that statisticians begin thinking about how to handle their "release," or limit their disclosure possibilities through restriction of what is released as part of a public-use data file. Such issues pose enormous methodological challenges for disclosure limitation research and for official statistics more broadly.

ACKNOWLEDGMENTS

This work was supported in part by Grant No. EIA-9876619 from the U.S. National Science Foundation to the National Institute of Statistical Sciences.

REFERENCES

ANDERSON, M., DAPONTE, B.O., FIENBERG, S.E., KADANE, J.B., SPENCER, B.D., & STEFFEY, D.L. (2000). Sample-based adjustment of the 2000 census-a balanced perspective. *Jurimetrics*, 341-356.

ANDERSON, M., & FIENBERG, S.E. (1999). *Who Counts: The Politics of Census-taking in Contemporary America.* New York: Russell Sage Foundation.

BERKE, R.L. (2000). What are you afraid of? A hidden image emerges. *The New York Times*, p. 1. (Week in Review, Section 4)

BISHOP, Y.M.M., FIENBERG, S.E., & HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

BUZZIGOLI, L., & GIUSTI, A. (1999). An algorithm to calculate the lower and upper bounds of the elements of an array given its margin. In *Statistical Data Protection, Proceedings of the Conference, Lisbon* (pp.131-147). Luxembourg: Eurostat.

CLAYTON, E.W., STEINBERG, K.K., KHOURY, M.J., THOMSON, E., ANDREWS, L., KAHN, M.J.E., KOPELMAN, L.M., & WEISS, J.O. (1995). Informed consent for genetics research on stored tissue samples. *Journal of the American Medical Association*, 274(22), 1786-1792.

COX, L. (1999). Some remarks on research directions in statistical data protection. In *Statistical Data Protection, Proceedings of the Conference, Lisbon* (pp.163-176). Luxembourg: Eurostat.

DALENIUS, T. (1977). Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 5, 429-444.

DAVID, M.H. (1998). Killing with kindness: The attack on public use data. *Proceedings of the Section on Government Statistics*, 3-7. (American Statistical Association).

DOBRA, A., & FIENBERG, S.E. (2000a). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences, 97*, 11885-11892.

DOBRA, A., & FIENBERG, S.E. (2000b). *Computing bounds for entries in k-dimensional cross-classifications given all (k-1)-dimensional marginals.* Unpublished Technical Report, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

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

DUNCAN, G.T. (2000). Confidentiality and statistical disclosure limitation. Forthcoming in N. Smelser & P. Baltes (Eds.), *International Encyclopedia of the Social and Behjavioral Sciences.* New York: Elsevier.

DUNCAN, G.T., & FIENBERG, S.E. (1999). Obtaining information while preserving privacy: A markov perturbation method for tabular data. In *Statistical Data Protection, Proceedings of the Conference, Lisbon* (pp. 351-362). Luxembourg: Eurostat.

DUNCAN, G.T., JABINE, T.B., & DE WOLF, V.A. (Eds.). (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Washington, DC: National Academy Press. (Panel on Confidentiality and Data Access, Committee on National Statistics)

DUNCAN, G.T., & PEARSON, R.B. (1991). Enhancing access to microdata while protecting confidentiality: prospects for the future (with discussion). *Statistical Science*, 6, 219-239.

FIENBERG, S.E. (1997). Confidentiality and disclosure limitation methodology: Challenges for national statistics and statistical research. (Background paper prepared for the Committee on National Statistics)

FIENBERG, S.E. (1998a). Should we continue to release public-use microdata files? Yes,yes, yes! *Proceedings of the Section on Government Statistics*, 8-12. (American Statistical Association).

FIENBERG, S.E. (1998b). Towards multiple-media survey and census data: Rethinking fundamental issues of design and analysis. In *Symposium 97: New Directions in Surveys and Censuses* (pp.7-18). Ottawa, Canada. (Keynote Address)

FIENBERG, S.E. (1999). Fréchet and bonferroni bounds for multi-way tables of counts with applications to disclosure limitation. In *Statistical Data Protection, Proceedings of the Conference,* Lisbon (pp. 115-129). Luxembourg: Eurostat.

FIENBERG, S.E. (2000). Statistical perspectives on confidentiality and data access in public health. Forthcoming in a 2001 issue of *Statistics in Medicine.*

FIENBERG, S.E., & MAKOV, U.E. (1998). Confidentiality, uniqueness, and disclosure avoidance for categorical data. *Journal of Official Statistics*, 14, 385-397.

FIENBERG, S.E., & MAKOV, U.E. (2000). Uniqueness and disclosure risk: Urn models and simulation. In *ISBA 2000 proceedings.* Luxembourg:Eurostat.

FIENBERG, S.E., MAKOV, U.E., & SANIL, A.P. (1997). A Bayesian approach to data disclosure. Optimal intruder behavior for continuous data. *Journal of Official Statistics*, 13, 75-90.

FRÉCHET, M. (1940). *Les Probabilitiés, Associées aun Système d'Évènments Compatibles et Dépendants.* Paris: Hermann & Cie.

GOUWELEEUW, J.M., KOOIMAN, P., WILLENBORG, L.C.R.J., & WOLF, P.P. DE. (1998). Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14, 463-478.

HABERMAN, S.J. (1974). *The Analysis of Frequency Data.* Chicago: University of Chicago Press.

HUNDEPOOL, A., WILLENBORG, L., VAN GEMERDEN, L., WESSELS, A., FISCHETTI, M.M SALAZAR, J.-J., & CAPRARA, A. (1998b). *τ-Argus User's Manual.* Department of Statistical Methods, Statistics Netherlands.

HUNDEPOOL, A., WILLENBORG, L., WESSELS, A., VAN GEMERDEN, L., TIOURINE, S., & HURKENS, C. (1998a). *μ-Argus User's Manual.* Department of Statistics, Statistics Netherlands.

KELLER-MCNULTY, S., & UNGER, E.A. (1998). A data system prototype for remote access to information based on confidential data. *Journal of Official Statistics,* 14, 347-360.

LAMBERT, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics,* 9, 313-331.

LAURITZEN, S. (1996). *Graphical Models.* New York: Oxford University Press.

PANNEKOEK, J., & WAAL, T. DE. (1998). Synthetic and combined estimators in statistical disclosure control. *Journal of Official Statistics,* 14, 399-410.

ROEHRIG, S.F., PADMAN, S., DUNCAN, G., & KRISHNAN, R. (1999). Disclosure detection in multiple linked categorical datafiles: A unified network approach. *Staitistical Data Protection, Proceedings of the Conference,* 149-162. (Eurostat)

RUBIN, D.B. (1993). Satisfying confidentiality constraints through the use of synthetic multiply imputed microdata. *Journal of Official Statistics,* 9, 461-468.

SAMUELS, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniqueness problem in microdata disclosure risk assessment. *Journal of Official Statistics,* 14, 373-383.

SKINNER, C.J., & HOLMES, D.J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics,* 14, 361-372.

STEEL, P., & ZAYATZ, L. (1999). Disclosure limitation for the 2000 census of housing and population. *Statistical Data Protection, Proceedings of the Conference,* 362-368. (Eurostat)

SUBCOMMITTEE ON DISCLOSURE-AVOIDANCE TECHNIQUES (1978). *Statistical Policy Working Paper No. 2: Report on Statistical Disclosure and Disclosure Avoidance Techniques.* Washington, DC: Federal Committee on Statistical Methodology, Office of Federal Policy and Standards, U.S. Dept. of Commerce.

SUBCOMMITTEE ON DISCLOSURE-AVOIDANCE TECHNIQUES (1994). *Statistical Policy Working Paper No. 22: Report on Statistical Disclosure Limitation Methodology.* Washington, DC: Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

WILLENBORG, L., & DE WAAL, T. (1996). *Statistical Disclosure Control in Practice* (Vol. 111). New York: Springer Verlag.

ZASLAVSKY, A.M., & HORTON, N.J. (1998). Balancing disclosure risk against the loss of non-publication. *Journal of Official Statistics,* 14, 411-419.