

19th CEIES SEMINAR

Innovative solutions in providing access to micro-data

Lisbon, 26 and 27 September 2002

Session 1

**State of the art: An overview of policy and practice on release of
microdata**

Presented by George T. Duncan

Carnegie Mellon University, United States

POLICY AND PRACTICE ON RELEASE OF MICRODATA

By George T. Duncan

Executive Summary

Statistical offices must provide data products that are both useful and have low risk of confidentiality disclosure. Recognizing that deidentification of data is generally inadequate to protect their confidentiality against attack by a data snooper, agencies can release microdata under policies of *restricted access* or can release products of *restricted data*.

Under a policy and practice of *restricted access*, administrative procedures impose conditions on user access to data. These conditions may depend on the type of data user; conditions may be different for interagency data sharing than for external data users. Various restricted access policies (Jabine 1993a,b) have been implemented in the last twenty years. A prospective data user may apply, for example, for sworn employee status, agreeing to be bound by similar conditions to regular employees of the agency. If approved, data users may be required to relocate to the agency to gain access to unrestricted data. In some cases of restricted access, for example to the Panel Study of Income Dynamics and the National Longitudinal Survey of Youth (Jabine 1993b), the researcher must post bond. The money will be forfeited if the researcher fails to honor the release agreement. A failure may be occasioned by, say, unauthorized sharing of the data or performing analyses not specified in the proposal. An example of an institutional arrangement for restricted access by external data users is the Census Research Data Center at granted Carnegie Mellon University's H. John Heinz III School of Public Policy and Management (see <http://www.heinz.cmu.edu/census/>). Through access to such valuable data, the Center has attracted nationally renowned scholars to engage in interdisciplinary, collaborative research on important policy issues.

Under a policy and practice of release of *restricted data*, the agency transforms the original data to lower disclosure risk. For microdata, this is accomplished through disclosure limitation techniques such as (1) release of only a sample of the data, (2) including simulated data, (3) "blurring" of the data by grouping or adding random error, (4) excluding certain attributes, and (5) swapping data by exchanging the values of just certain variables between data subjects. Desirably, the resulting restricted data have both high data utility U to users (analytically valid data) and low disclosure risk R (safe data).

Agencies need to tools to evaluate current and prospective policies regarding the release of restricted data. Such a tool is the *R-U confidentiality map*, a chart that traces the impact on disclosure risk R and data utility U of changes in the parameters of a disclosure limitation procedure. As an illustration of its promise, theory for the R-U confidentiality map is developed for additive noise applied to univariate data under different scenarios of data snooper attack.

A detailed discussion of data swapping is provided.

POLICY AND PRACTICE ON RELEASE OF MICRODATA

By George T. Duncan¹

1. Restricted Access vs. Restricted Data

Wide-ranging mechanisms exist to deal with conflicts about the capture and dissemination of data. They span legislation, interorganizational contractual arrangements, intraorganizational administrative policies, and ethical codes. They also include technological remedies such as the release of masked data that may satisfy data users needs for statistical information while posing little risk of disclosure of personal information (Duncan and Pearson 1991). In managing confidentiality and data access functions, statistical offices can employ two fundamentally different approaches for responsible provision of information: *restricted data* and *restricted access*. As developed in Duncan, Jabine and de Wolf (1993), these approaches have the following interpretations:

- **Restricted data.** Data are transformed to lower disclosure risk. This is accomplished through disclosure limitation techniques such as (1) release of only a sample of the data, (2) including simulated data, (3) "blurring" of the data by grouping or adding random error, (4) excluding certain attributes, and (5) swapping data by exchanging the values of just certain variables between data subjects.

Restricted access. Administrative procedures imposing conditions on access to data. These conditions may depend on the type of data user; conditions may be different for interagency data sharing than for external data users. Various restricted access policies (Jabine 1993a,b) have been implemented in the last twenty years. Notable has been the fellowship programs run jointly by the American Statistical Association, the National Science Foundation together with four agencies, the Bureau of Labor Statistics, the Bureau of the Census, the US Department of Agriculture, and the National Institute of Standards and Technology. If approved, data users relocate to the agency to gain access

¹ This work was partially supported by grants from the National Science Foundation under Grant EIA-9876619 to the National Institute of Statistical Sciences, the National Center for Education Statistics under Agreement EDOERI-00-000236 to Los Alamos National Laboratory, and the National Institute on Aging under Grant 1R03AG19020-01 to Los Alamos National Laboratory. Many of the ideas described originated out of joint research with Sallie Keller-McNulty, Lynne Stokes and Stephen Roehrig. The author thanks the National Center for Education Statistics for providing—under nondisclosure license 010207550—access to the individually identifiable survey database entitled, “National Education Longitudinal Study of 1988 (NELS), Schools and Staffing Survey (SASS), and all follow-ups.”

to unrestricted data. In some cases of restricted access, for example to the Panel Study of Income Dynamics and the National Longitudinal Survey of Youth (Jabine 1993b), the researcher must post bond. The money will be forfeited if the researcher fails to honor the release agreement, say by unauthorized sharing of the data or performing analyses not specified in the proposal. An example of an institutional arrangement for restricted access by external data users is the Census Research Data Center at Carnegie Mellon University (see <http://www.heinz.cmu.edu/census/>). The Bureau of the Census had long sought a mechanism by which it could make detailed Census information more readily available to researchers, as well as connect Census data to other important national datasets, such as those housed at the Environmental Protection Agency and the Department of Justice, while maintaining the integrity and confidentiality of that data. With this in mind, the Bureau granted Carnegie Mellon University's H. John Heinz III School of Public Policy and Management the first Census Center to be housed at a university. Through access to such valuable data, the Center has attracted nationally renowned scholars to engage in inter-disciplinary, collaborative research on important policy issues. Additional restricted access sites of the Census are now maintained in collaboration with the University of California and with Duke University. Further sites have been opened by the Netherlands, the U.S. National Center for Health Statistics, and Eurostat.

To some extent, the confidentiality promised by a statistical office is necessarily at risk. A statistical office cannot simply erect firewalls around its data, because the has a mandate to disseminate products based on these data. This mandate is based on the IOs awareness that its data products contribute legitimate information to its clients. In a democratic and free market society, the client base of many statistical offices is broad. They not only provide data to guide government policy making, but they also provide data products to individuals, firms, non-governmental organizations, the media, and interest groups. As a most desirable result, public policy debate and decentralized economic decision making are informed. On the other hand, unintended consequences of dissemination can occur if the released information allows the confidentiality pledge to be compromised by a *data snooper*. The term *data snooper* refers to anyone with legitimate access to the data product and whose goals and methods in the use of the data are not consonant with the mission of the agency. Thus, a hacker who tries to break into a protected computer system is not a data snooper. Nor is a researcher who uses exploratory data analysis to discover statistical relationships. Other terms in the literature for “data snooper” include “data spy,” “intruder,” or “attacker.” Compromise of confidentiality by a data snooper constitutes a statistical confidentiality disclosure (Elliot and Dale 1999). Such a compromise occurs when the data dissemination permits a data snooper to gain illegitimate information about a respondent.

Ensuring confidentiality is not a simple task. For most of the census or survey data collected by statistical agencies, deidentification—removal of apparent identifiers like name, social security number, email address, etc. (although an obvious first step)—is not adequate to lower disclosure risk to an acceptable level (Paass 1988, Winkler 1998). Also, most health care information, such as hospital discharge data, cannot be

anonymized through deidentification. The key reason that removing identifiers does not assure sufficient anonymity of respondents is that, today, a data snooper can get inexpensive access to databases with names attached to records. Marketing and credit information databases and voter registration lists are exemplars. Having this external information, the data snooper can employ sophisticated, but readily available, record linkage techniques. The resultant attempts to link an identified record from the public database to a deidentified record provided by the are often successful (Winkler 1998). With such a linkage, the record would be reidentified.

To publicly disseminate a data product safe from attack by a data snooper, a statistical office must go beyond deidentification; it must restrict the data by employing a disclosure limitation method. An easily interpreted and implemented method is to coarsen the data, essentially creating bins and counting the number of occurrences in the data. For example, recode income in increments of \$5,000 and release a table giving, say, how many earned between \$60,000 and \$65,000. Coarsening provides a good example of an approach that while effective in lowering disclosure risk also lowers the data's utility. By fuzzing the target of a record linkage, such coarsening clearly makes reidentification through record linkage less likely. On the other hand, data utility becomes a problem with this coarsening to tabular form because releasing such tables no longer satisfies many users of statistical data. Coarsen gender, for instance, and you've lost the attribute entirely. Those data users who command the latest computer technology and who can make the most important research and policy contributions typically need data that are more distinguished. To be able to assess alternative disclosure limitation methods, we first need a framework for assessing how good a disclosure limitation procedure is.

In fulfilling their stewardship responsibilities, statistical offices must manage the not easily resolved tension between ensuring confidentiality and providing access (Duncan, Jabine and de Wolf 1993, Kooiman, Nobel and Willenborg 1999, Marsh *et al* 1991). Resolving this tension requires policies under which a statistical office can disseminate data products that have both

1. high data utility U , so faithful in critical ways to the original data (analytically valid data), and
2. low disclosure risk R , so confidentiality is protected (safe data).

Statistical disclosure limitation techniques (Chowdhury et al 1999; Duncan and Lambert 1986; Zayatz et al 1999) provide classes of transformations that lower disclosure risk. Complicating the IO's task is the cornucopia of available statistical disclosure limitation methods, each with different impacts on data utility and disclosure risk. Major methods include suppressing attributes, swapping attributes, releasing only a sample of the population, topcoding, adding noise, various forms of aggregation, and cell suppression. General references to the literature in disclosure limitation include Duncan (2001), Duncan, Jabine and de Wolf (1993), Eurostat (1996), Fienberg (1994, 1997), Jabine (1993b), Mackie and Bradburn (2000), and Willenborg and de Waal (1996).

We can look systematically at the simultaneous impact on disclosure risk and data utility of implementing disclosure limitation techniques and choosing their parameter values. A measure of statistical disclosure risk, R , is a numerical assessment of the risk of

unintended disclosures following dissemination of the data. A measure of data utility, U , is a numerical assessment of the usefulness of the released data for legitimate purposes. By fully developing the concept of an R-U confidentiality map and applying it in some important contexts, this article provides a quantified link between R and U directly through the parameters of the disclosure limitation procedure. As noted in the Report on the Work Session on Statistical Data Confidentiality (2001), “Assessing and limiting the effects of disclosure limitation on data analysis and usefulness is an extremely important but difficult and under-explored area.” With an explicit representation of how the parameters of the disclosure limitation procedure affect R and U, the tradeoff between disclosure risk and data utility is apparent. With the R-U confidentiality map, statistical offices have a workable new tool to frame decision making about data dissemination under disclosure limitation.

2. R-U Confidentiality Map

The rudiments of the R-U confidentiality map were presented by Duncan and Fienberg (1999), and further explored for categorical data by Duncan *et al.* (2001). An R-U confidentiality map provides a quantified link between R and U directly through the parameters of the disclosure limitation procedure. With an explicit representation of how the parameters of the disclosure limitation procedure affect R and U, the tradeoff between disclosure risk and data utility is apparent. With the R-U confidentiality map, we have a workable new tool to frame decision making about data dissemination under disclosure limitation.

In its most basic form, an R-U confidentiality map is the set of paired values, (R, U) , of disclosure risk and data utility that correspond to various strategies for data release. Typically, these strategies implement a disclosure limitation procedure, like masking through the addition of random error. Such procedures are determined by parameters, for instance, the magnitude of the error variance λ^2 for noise addition. As λ^2 is changed, a curve is mapped in the R-U plane. Visually, the R-U confidentiality map portrays the tradeoff between disclosure risk and data utility as λ^2 increases, and so more extensive masking is imposed. Shown below is a simple example of the construction of an R-U confidentiality map:

For the realized values, x_1, \dots, x_n , the masked data has the additive noise form,

$$Y_i = x_i + \varepsilon_i, \varepsilon_i \sim iid(0, \lambda^2), i = 1, \dots, n.$$

Data Utility: The data user estimates the population mean μ using

$$\hat{\mu} = \bar{Y}, \text{ the sample mean of the masked data. Therefore, } E(\hat{\mu}) = \mu, \text{ Var}(\hat{\mu}) = \frac{1}{n}(\sigma^2 + \lambda^2),$$

and the data utility is $U = \frac{n}{\sigma^2 + \lambda^2}$.

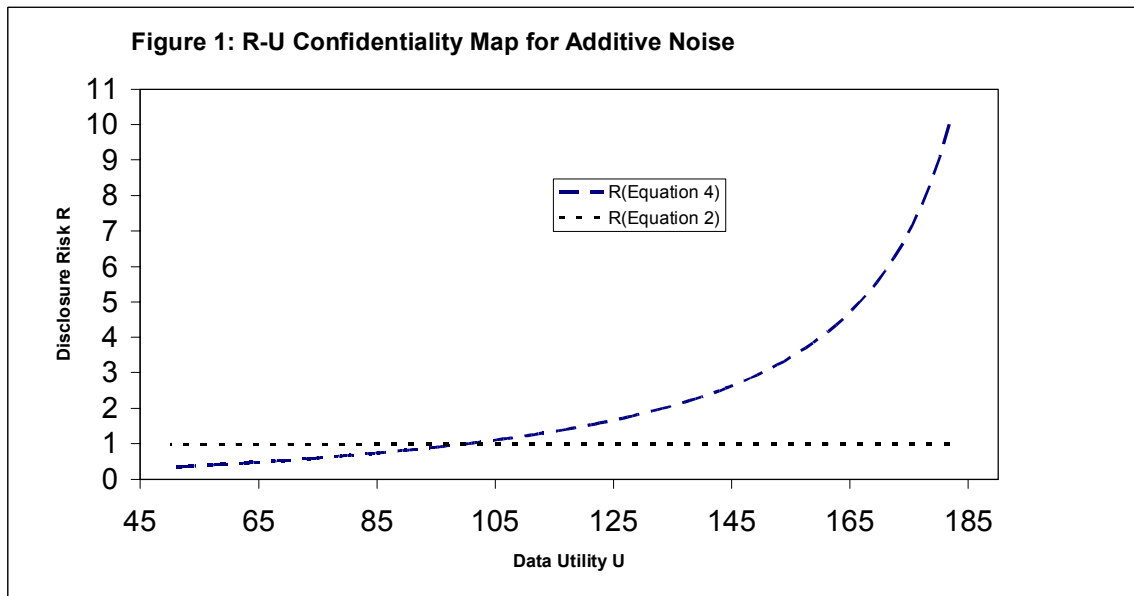
Disclosure Risk: The first two states of knowledge of the data snooper, assuming the snooper’s goal is to compromise a specific entity, have the same disclosure risk. In both

states the data snooper is simply after a specific target value τ and will use $\hat{\tau} = \bar{Y}$ as the estimator for τ . This gives risk of

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{n}{\sigma^2 + \lambda^2 + n(\mu - \tau)^2}.$$

Given this risk specification, the statistical office can determine what entities lead to the maximum risk across either the sample or the entire population.

Displayed in Figure 1 is an R-U confidentiality map for two risk measures in this example. One risk measure is a special case of the one above. The resulting map is labeled in Figure 1 as Equation (2). The other risk measure is based on the assumption that the data snooper knows the index of the target. The resulting map is labeled as Equation (4). The figure displays the impact on data utility and disclosure risk for changes in the disclosure limitation parameter λ^2 , under each of these two scenarios.



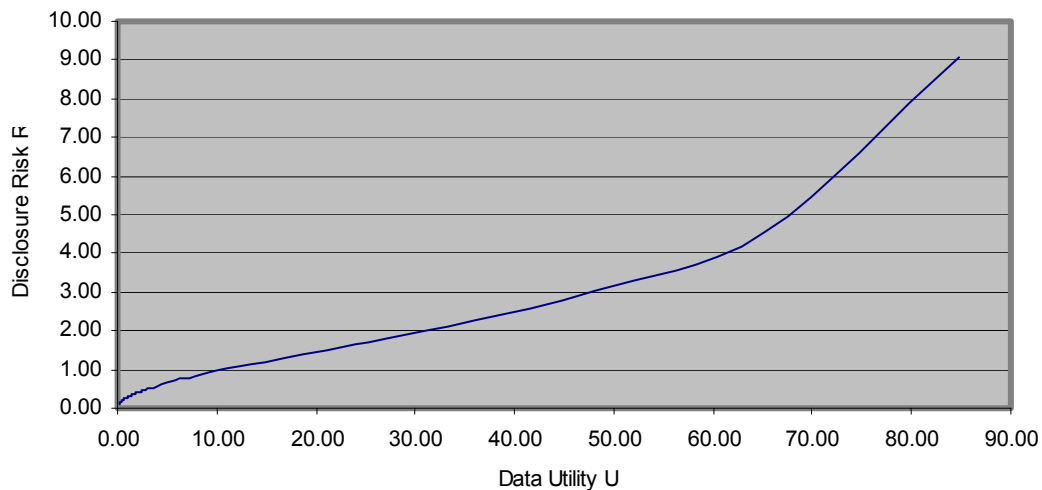
3. Data Swapping

As a method for disclosure limitation, data swapping switches certain fields in a record with the corresponding fields in another record. The statistics literature suggests that swapping observations “at random” was first proposed by Dalenius and Reiss (1978) and then developed more fully by Dalenius and Reiss (1982), both for the special case of categorical data. Also, see Spruill (1983). Earlier, in the computer science literature, Conway and Strip (1976) suggested a method they called *value disassociation*—a database query for the value of a record field yields the value of the same field in some other record. They propose that if a field for record i is included in a database query that it be replaced by the same field of another record j . The record j would be selected at random and with replacement. Data swapping is also referred to as *multidimensional transformation* (Schlörer 1981) and *data switching* (Navarro, Flores-Baez and Thompson

1988). A deficiency of the current literature is that it does not address the performance characteristics of data swapping. We remedy this deficiency by providing measures of disclosure risk or data utility and examining tradeoffs between them through the R-U confidentiality map.

Data swapping transforms the original data matrix X of n records (rows), each with p attributes (columns), into another $n \times p$ matrix M . The data product to be released is M or some statistics based on M . In its early expression in the literature, data swapping was argued to intuitively lower disclosure risk while explicitly maintaining, or approximately maintaining, certain statistics calculated from M . In this sense, M is equivalent or near equivalent to X , and for some applications would have high data utility. An example of an R-U confidentiality map in the case of data swapping is displayed as Figure 4.

Figure 3. R-U Confidentiality Map: Estimating a Mean Under Data Swapping



4. Conclusions

I see the following areas as ones where some contributions have been made, but additional work is needed to meet pressing needs of statistical agencies:

- Obtaining a better understanding of the empirical disclosure risks that arise when data are being disseminated.** Most research to date in this area has focused on the possibilities a data intruder (also called a data snooper) has to compromise confidential data. Thus, the emphasis has been on the potential for disclosure. Little empirical work has addressed what might motivate someone to act as a data snooper. Nor has there been much beyond speculation about what the actual harm might be to a statistical office of a compromise of confidentiality. As was recognized in the project planning, solid empirical work in these areas will

lead to better understanding of the real disclosure risk involved in the release of data products.

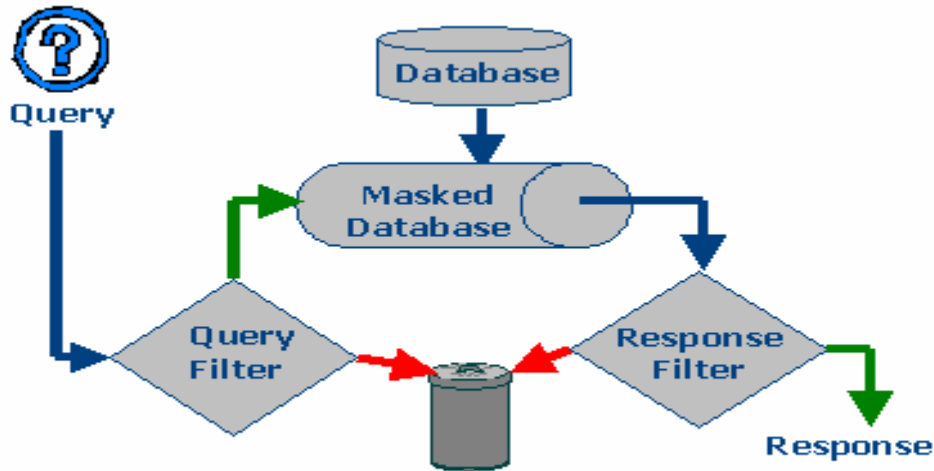
- **Quantifying information loss due to the implementation of statistical disclosure limitation procedures.** I believe that progress on this can be made using two parallel strategies. One strategy is to pursue the information-theoretic framework first put forth in Duncan and Lambert (1986). The other strategy is to empirically examine how different classes of researchers actually use data. Certain researchers, for example, those in academia with abundant computing resources and strong methodological skills, want data that is longitudinal and that has fine geographical detail. Other researchers may be content with more aggregate data. It is likely difficult to serve the needs of the first class of researchers with publicly available data products. Instead, they may be required to obtain the data they need under restricted access conditions.
- **Disseminating microdata with detailed geographical information.** An important issue is then how to release such information, on the one hand with the aim to allow for detailed regional studies of various kinds, while guarding the privacy of the individuals represented in the data file. The main problem against which to guard is the differencing of tables with different geography, and thereby disclosing information

There are a number of promising areas for future research:

Data swapping. In data swapping () some fields of a record are swapped with the corresponding fields in another record.

Virtual data. Synthetic data sets consist of records of individual synthetic units rather than records the agency holds for actual units. Rubin (1993) suggested synthetic data construction through a multiple imputation method

Online access to statistical databases. Increasingly the normal way of disseminating information is through online databases. Statistical disclosure limitation has a number of options in this case, as the following graphic illustrates:



Issues of online access are explored by Adam and Wortmann (1989) and Blakemore (2001). Work is being done in this area by the National Institute of Statistical Sciences under the NSF's Digital Government Initiative.

Methods for longitudinal data. As many have noted (e.g., Benedetti *et al* (1997) and the Federal Committee on Statistical Methodology (1994)), adequate statistical disclosure limitation methods for longitudinal data do not exist. This is a serious lack because of the immense value of longitudinal data in empirical policy-related research (Mackie and Bradburn 2000).

Use of Bayesian methods. Bayesian methods can be powerful tools for statistical disclosure limitation (Duncan and Lambert (1986), Fienberg, Makov and Sanil (1995)). A key element in this is the exploration of the relationship between disclosure risk and data utility.

References

- Adam, N.R. and Wortmann, J.C. (1989). Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys* **21** 515-556.
- Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990) Disclosure control of microdata. *Journal of the American Statistical Association* **85** 38-45.
- Blakemore, M. (2001) The potential and perils of remote access. *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*. Doyle, P, Lane, J., Theeuwes, J., and Zayatz, L. (Eds.) Amsterdam: Elsevier.
- Chen, G. and Keller-McNulty, S. (1998) Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* **14** 79-95.

- Dalenius, T. and Reiss, S.P. (1978). Data-swapping: A technique for disclosure control (extended abstract). *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 191-194.
- Dalenius, T. and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6** 73-85.
- De Waal, A. G. and Willenborg, L. C. R. J. (1994) Minimizing the number of local suppressions in a microdata set. Report. Statistics Netherlands, Voorburg.
- De Waal, A.G. and Willenborg, L.C.R.J. (1996). A View on Statistical Disclosure for Microdata. *Survey Methodology* **22** 95-103.
- De Waal, A.G. and Willenborg, L.C.R.J. (1998). Optimal local suppression in microdata. *Journal of Official Statistics* **14** 421-435.
- Domingo-Ferrer, Josep (1999) Microdata masking methods. Workshop on Confidentiality Research. May 3-4. U.S. Census Bureau. Alexandria, VA.
- Duncan, G. T. (2001) Confidentiality and statistical disclosure limitation. *International Encyclopedia of the Social and Behavioral Sciences*. To appear.
- Duncan, G. T., Jabine, T. B., and de Wolf, V. A. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* Panel on Confidentiality and Data Access, Committee on National Statistics, National Academy Press, Washington, DC.
- Duncan, G. T. and Keller-McNulty, S. (2001) Disclosure risk vs. data utility: The R-U confidentiality map. Technical Report. Statistical Sciences Group. Los Alamos National Laboratory. Los Alamos, New Mexico.
- Duncan, G. T. and Lambert, D. (1986) Disclosure-limited data dissemination (with discussion) *Journal of the American Statistical Association*. **81** 10-28.
- Duncan, G. T. and Lambert, D. (1989) The risk of disclosure of microdata. *Journal of Business and Economic Statistics* **7** 207-217.
- Duncan, G. T. and Pearson, R. (1991) Enhancing access to microdata while protecting confidentiality: Prospects for the future (with discussion). *Statistical Science* **6** 219-239.
- Elliot, M. and Dale, A. (1999) Scenarios of attack, the data intruders' perspective on statistical disclosure risk. *Netherlands Official Statistics* **14** 6-10.

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.

Fellegi, I. P. (1972) On the question of statistical confidentiality. *Journal of the American Statistical Association* **67** 7-18.

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

Fienberg, S.E., Steele, R.J., and Makov, U.E. (1996) Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and log-linear models. *Proceedings of Bureau of the Census 1996 Annual Research Conference*. US Bureau of the Census, Washington, DC, 87-105.

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

Fienberg, S. E. (2001) Statistical perspectives on confidentiality and data access in public health. *Statistics in Medicine* **20** (in press).

Fuller, W. (1993) Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9** 383-406.

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

Jabine, T. B. (1993a). Statistical Disclosure Limitation Practices of United States Statistical Agencies. *Journal of Official Statistics*, **9**, 427-454.

Jabine, T. B. (1993b). Procedures for Restricted Data Access. *Journal of Official Statistics*, **9**, 537-590.

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

Moore, R.A. (1996). Controlled data swapping techniques for masking public use microdata sets. RR 96-05. U.S. Bureau of the Census, Washington, DC.

Navarro, A., Flores-Baez, L., and Thompson, J. (1988) Results of Data Switching Simulation. Presented at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees.

Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics* 6 487-500.

Schrijver, A. (1986) *Theory of Linear and Integer Programming*. Wiley, New York.

Spruill, N. L. (1983) The confidentiality and analytic usefulness of masked business microdata. *Proceedings of the Section on Survey Research Methods*, American Statistical Association 602-607.

Strudler, M., Oh, H. L., and Scheuren, F. (1986) Protection of taxpayer confidentiality with respect to the tax model. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 375-381.

Trottini, M. (2001) A decision-theoretic approach to data disclosure problems. Paper prepared for 2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality 14-16 March 2001, Skopje, Macedonia.

Willenborg, L. and de Waal, T. (1996) *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics **111** Springer, New York.

Willenborg, L. and de Waal, T. (2000). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics **155** Springer-Verlag, New York.

Winkler, W. E. (1998) Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Research in Official Statistics* **1** 87-104.

Zaslavsky, A.M. and Horton, N.J. (1998) Balancing disclosure risk against the loss of nonpublication. *Journal of Official Statistics*, **14**, 411-419.

Zayatz, L. V. and Rowland, S. (1999) Disclosure limitation for American FactFinder. Paper presented at the American Statistical Association Joint Statistical Meetings, Baltimore, MD, August 8.