Statistical Data Stewardship in the 21st Century: An Academic Perspective

George T. Duncan

Carnegie Mellon University

Joint Statistical Meetings
New York City

2002 August 11

Statistical Data Stewardship in the 21st Century: An Academic Perspective[1]

By George T. Duncan

Abstract

This paper presents an academic perspective on a broad spectrum of ideas and best practices for statistical data collectors to ensure proper stewardship for personal information that they collect, process and disseminate. Academic researchers in confidentiality address statistical data stewardship both because of its inherent importance to society and because the mathematical and statistical problems that arise challenge their creativity and capability. To provide a factual basis for policy decisions, an information organization (IO) engages in a two-stage process: (1) It gathers sensitive personal and proprietary data of value for analysis from respondents who depend on the IO for confidentiality protection. (2) From these data, it develops and disseminates data products that are both useful and have low risk of confidentiality disclosure. The IO is a broker between the respondent who has a primary concern for confidentiality protection and the data user who has a primary concern for the utility of the data. This inherent tension is difficult to resolve because deidentification of the data is generally inadequate to protect their confidentiality against attack by a data snooper. Effective stewardship of statistical data requires restricted access or restricted data procedures. In developing restricted data, IOs apply disclosure limitation techniques to the original data. Desirably, the resulting restricted data have both high data utility U to users (analytically valid data) and low disclosure risk R (safe data). This paper explores the promise of the *R-U confidentiality map*, a chart that traces the impact on R and U of changes in the parameters of a disclosure limitation procedure. Theory for the R-U confidentiality map is developed for additive noise. By an implementation through simulation methods, an IO can develop an *empirical R-U confidentiality map*. Disclosure limitation for tabular data is discussed and a new method, called *cyclic perturbation*, is introduced. The challenges posed by on-line access are explored.

Keywords: Privacy, Confidentiality, Policy

---

[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 here 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."

1. Introduction

Academic researchers in confidentiality address statistical data stewardship both because of its inherent importance to society and because the mathematical and statistical problems that arise challenge their creativity and capability. To provide a factual basis for policy decisions, an information organization (IO) engages in a two-stage process:

1. It gathers sensitive personal and proprietary data of value for analysis from respondents who depend on the IO for confidentiality protection
2. From these data, it develops and disseminates data products that are both useful and have low risk of confidentiality disclosure.

The IO is a broker between the respondent, who has a primary concern for confidentiality protection, and the data user, who has a primary concern for the utility of the data. An IO cannot simply erect firewalls around its data, because the IO has a mandate to disseminate products based on these data. This mandate is based on the IOs awareness that its data products contribute legitimate and valuable information to its clients. In a democratic and free market society, the client base of many IOs is broad. Statistical agencies, for example, 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). 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 IO 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, an IO 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 of higher resolution.

In fulfilling their stewardship responsibilities, information organizations must manage the tension between ensuring confidentiality and providing access to useful data (Duncan, Jabine and de Wolf 1993, Kooiman, Nobel and Willenborg 1999, Marsh *et al* 1991). Resolving this tension requires policies under which an IO can disseminate data products that have both

1. high data utility U, so faithful in critical ways to the original data (analytically valid data)
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 seek 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, say R, is a numerical assessment of the risk of unintended disclosures following dissemination of the data. A measure of data utility, say U, is a numerical assessment of the usefulness of the released data for legitimate purposes. An R-U confidentiality map quantifies the 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, information organizations 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). 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.

Just to illustrate the ideas, shown below is a simple example of the construction of an R-U confidentiality map:
For the realized values, $x_1, \ldots, x_n$, the masked data has the additive noise form,
$Y_i = x_i + \varepsilon_i, \varepsilon_i \sim iid(0, \lambda^2),\ i = 1, \ldots, n.$

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

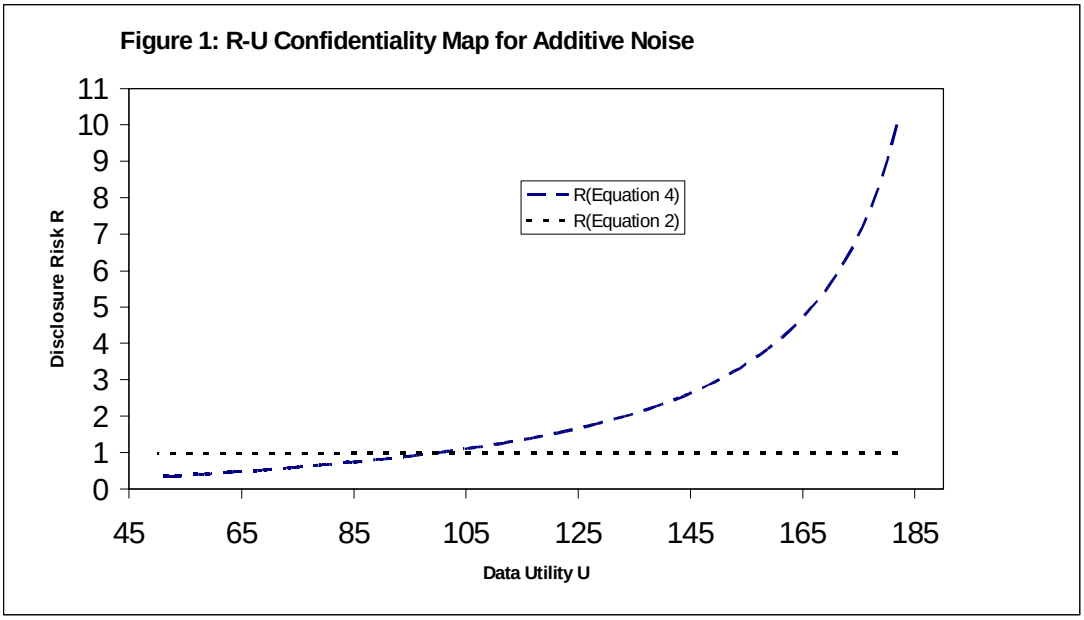
$\hat{\mu} = \bar{Y}$, the sample mean of the masked data. Therefore, $E(\hat{\mu}) = \mu,\ Var(\hat{\mu}) = \frac{1}{n}\left(\sigma^2 + \lambda^2\right)$, and the data utility is $U = \frac{n}{\sigma^2 + \lambda^2}$.

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 IO 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 2 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.



In general, an R-U confidentiality map can be used for, at least, the following purposes:

- inform the IO about whether or not proposed disclosure limitation methods are adequate in lowering disclosure risk and maintaining data utility
- facilitate comparisons between various disclosure limitation methods
- examine the risk of particular types of data snooper knowledge.

3. Empirical R-U Confidentiality Map

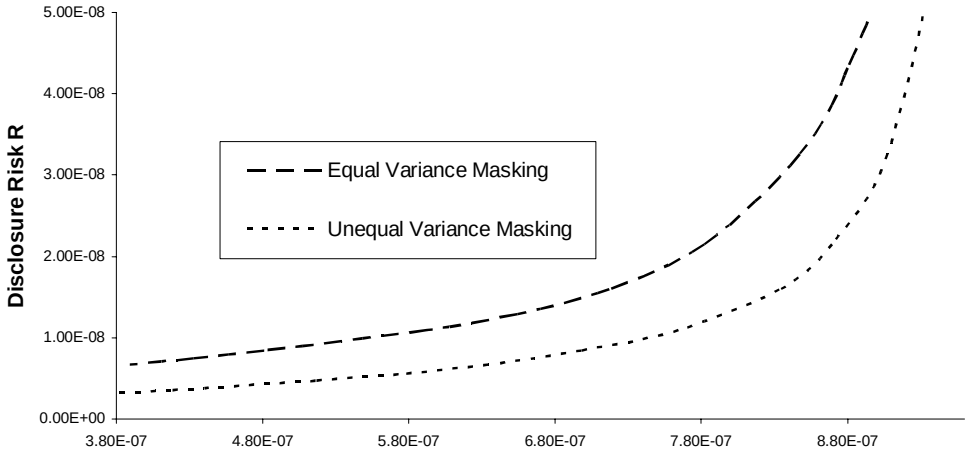
Analytical methods, such as the one in the previous section, can be used to investigate general properties of disclosure limitation methods and snooper strategies. In practice, nonetheless, any probabilistic structure, assumed for the analytic development, may not be fully adequate to depict consequential features of the actual data. Also, it may be difficult to derive the impact of disclosure limitation methods on disclosure risk and data utility. For such real-life examples of both practical size and realistic complexity, an *empirical R-U confidentiality map* can be used, providing an agency with a support tool in the developing safe and analytically valid data products from a specific database. To illustrate the empirical approach, we use confidential data from a survey by the National Center for Education Statistics (NCES)—the Teacher Followup Survey (TFS) of 1994-95.

To obtain the empirical R-U confidentiality map, we simulate the masking process of additive noise. For each of a range of values of λ^2, the agency can generate a number, say M, of masked datasets. From these simulated datasets, the agency can estimate the disclosure risk and the data utility.

Figure 2 was constructed from the actual confidential data for additive noise truncated at zero. It shows how an empirical R-U confidentiality map can help compare two alternative masking

strategies. In "unequal variance masking", the variance of the noise is doubled for sensitive data values. From the empirical R-U confidentiality map we clearly see that unequal variance masking dominates, because it results in higher data utility at any specified level of disclosure risk (and hence confidentiality protection).

Figure 2. Empirical R-U Confidentiality Map: Comparing Risk for Two Masking Methods



4. Tabular Data

Considerable effort has gone into developing disclosure limitation methods for tabular data (Duncan 2001, Duncan, Jabine and de Wolf 1993, Willenborg and de Waal 1996, 2000). These techniques include cell suppression, local suppression, global recoding, rounding, and various forms of perturbation (Federal Committee 1994). Under cell suppression, for example, the values of table cells that pose confidentiality problems are determined and suppressed (as primary suppressions) as well as values of additional cells that can be inferred from released table margins (as secondary or complementary suppressions) (Cox 1980). Cell suppression is the most common method of disclosure limitation for tabular data, but it has been seriously criticized by Duncan and Fienberg (1999) and others. A basic alternative to cell suppression are perturbation methods. A discussion of such methods is contained in Duncan *et al* (2001). Perturbation is used through controlled rounding (Cox 1982), versions of post-randomized response (Gouweleeuw, Kooiman, Willenborg, and de Wolf, 1998), and Markov perturbation approaches have been proposed in various forms by Duncan and Fienberg (1999), Fienberg, Makov, and Steele (1998), and Fienberg, Makov, Meyer, and Steele (2000). Many of these methods can be represented in the form of matrix masks (Duncan and Pearson 1991). In most of this methodological development, the focus has been on procedures that mask the data in ways that seem to effectively lower disclosure risk while hopefully leaving the data product useful for statistical purposes. The R-U confidentiality map is a tool that can provide a more rigorous analytical

framework for this intuitive notion. It can also provide a systematic way to examine new disclosure limitation methods.

Recently, I have begun work with Stephen Roehrig on a new method we call *cyclic perturbation*. Like other perturbation methods, such as controlled rounding, cyclic perturbation preserves marginal totals of the table. A distinguishing characteristic of the procedure is that it allows both the data user and the data snooper to employ Bayesian methods. In a straightforward way, a data user can determine the posterior distribution over possible cell values, given the user's prior probabilities over the set of possible true tables. With this approach, the IO can construct an R-U confidentiality map and so know the level of confidentiality protection afforded and the degree of data utility.

Under cyclic perturbation, the original table is transformed through a sequence of stochastic elementary moves or data squares. As in Duncan and Fienberg (1999), such an elementary move perturbs cell entries according to an alternating cycle of "+"s and "-"s arranged in a square. For an $m \times n$ table, there are $\binom{m}{2}\binom{n}{2}$ squares, choosing two rows and two columns from the table, and more elaborate data cycles can be constructed by forming sums and differences of them. Once a data cycle has been chosen, cyclic perturbation proceeds by flipping a three-sided coin, whose probabilities for sides A, B, and C are α, β, and $\gamma \equiv 1-(\alpha+\beta)$, respectively. If the coin shows side A, cells marked + in the data cycle are increased by one, and cells in positions marked – are decreased by one. If the coin shows side B, the reverse happens, while if the coin shows side C, all data cell values remain unchanged. Clearly, each such iteration leaves the row and column sums unchanged. A cell with entry zero is left unchanged. If α and β are set equal, the expected value of any cell is just its original value. For a given cell, the process represents a random walk with two absorbing states (given the marginal totals, each cell has a minimum value and a maximum value). We can prove a result that is important for simplifying data analysis: With a uniform prior distribution on possible tables, the posterior mode for a cell value equals the perturbed (and published) cell value, for any disclosure limitation parameter values $\gamma > \alpha, \beta$. The R-U confidentiality map allows comparison of this method with other methods such as cell suppression and random rounding.

5. On-Line Access

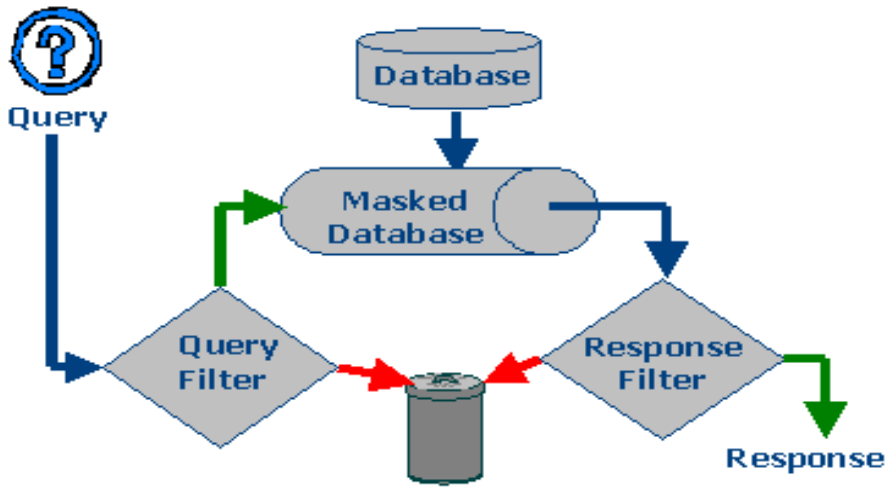
Increasingly, the way information is disseminated is through online databases. The following are examples of websites maintained by national statistical offices:

American FactFinder: http://factfinder.census.gov/servlet/BasicFactsServlet (see Zayatz and Rowland 1999)

Office of National Statistics (the UK Government Site): http://www.statistics.gov.uk/

Statistics Netherlands: http://www.cbs.nl/en/figures/keyfigures/index.htm

Disclosure limitation has a number of options in this case, as the following graphic illustrates:



Query filters include, for example, restrictions on the dimensionality of a requested table, say to no more than three-dimensional tables. The database may be masked through topcoding or data swapping. The response filter may not allow tables where the value (say, R&D expenditures) for a cell came from only one or two entities (say privately held optical networking firms).

A basic work that lays out key ideas for on-line disclosure limitation is Adam and Wortman (1989). Methodology for additive noise in repeated query systems in developed in Duncan and Mukherjee (2000). A variety of general concerns about remote access are laid out by Blakemore (2001). Some key issues for methodological research that are currently being addressed, for example in the Digital Government Initiative to the National Institute of Statistical Science, or should be addressed include the following:

1. Understanding the interplay between query filters, masking of a database, and response filters and its impact on disclosure risk and data utility
2. Appropriate models for disclosure risk in repeated query systems
3. The detection of anomalous behavior on the network
4. R-U confidentiality maps for dynamic databases (Increasingly databases will be based on real-time data captures, and so will change rapidly.)
5. Confidentiality issues for multimedia data, e.g., video
6. Consideration of disclosure limitation procedures appropriate for a hierarchy of data users, e.g., in the health care domain, some researchers may have obtained clearance to more detailed data.

6. Conclusions

I see the following areas as ones where some research 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. 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.

Besides those areas specifically considered in this paper, there are a number of promising areas for future research:

Data swapping. In data swapping (Dalenius and Reiss 1982, Reiss 1980, Spruill 1983) 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

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).

In each of these areas, the R-U confidentiality map may provide a unifying framework for consideration of tradeoffs in disclosure risk and data utility, while motivating the systematic comparison of alternative disclosure limitation methods.

References

Adam, N.R. and Wortmann, J.C. (1989) Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys* **21** 515-556.

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.

Chowdhury, S. D., Duncan, G. T., Krishnan, R., Roehrig, S. F., and Mukherjee, S. (1999) Disclosure detection in multivariate categorical databases: Auditing confidentiality protection through two new matrix operators. *Management Science* **45** 1710-1723.

Cox, L. H. (1980) Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association* **75** 377-385.

Cox, L. H. (1981) Linear sensitivity measures and statistical disclosure control. *Journal of Statistical Planning and Inference* **5** 153-164.

Cox, L. H. (1987) A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association* **82** 38-45.

Cox, L. H. (1995) Network models for complementary cell suppression. *Journal of the American Statistical Association* **90** 1453-1462.

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.

Duncan, G. T. (2001) Confidentiality and statistical disclosure limitation. *International Encyclopedia of the Social and Behavioral Sciences.*

Duncan, G. T. and Fienberg, S. E. (1999) Obtaining information while preserving privacy: a Markov perturbation method for tabular data. In *Statistical Data Protection (SDP'98) Proceedings*, Eurostat, Luxembourg, 351-362.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. and Roehrig, S. F. (2001) Disclosure limitation methods and information loss for tabular data. In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (Doyle, P. Lane, J. I., Theeuwes, J. M. and Zayatz, L. V. eds) Amsterdam: Elsevier, 135-166.

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., Krishnan, R., Padman, R., Reuther, P., Roehrig, S. (2001), Exact and heuristics methods for cell suppression in multi-dimensional linked tables, *Operations Research,* Forthcoming.

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 Mukherjee, S. (2000) Statistical database security against tracker and repeated query attack. *Journal of the American Statistical Association* **95** 720-728.

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.

Fellegi, I. P. (1975) Controlled random rounding. *Survey Methodology* **1** 123-133.

Fellegi, I.P., and Suntner, A.B. (1969) A theory for record linkage. *Journal of the American Statistical Association* **64** 1183-1210.

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).

Fienberg, S.E. and Makov, E.U. (1998) Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics* 14 385-398.

Fienberg, S. E., Makov, U. E. and Steele, R. J. (1998) Disclosure limitation using perturbation and related methods for categorical data (with discussion). *Journal of Official Statistics* **14** 485-512.

Fischetti, M. and Salazar-González, J. J. (1998) Experiments with controlled rounding for statistical disclosure control in tabular data with linear constraints. *Journal of Official Statistics* **14** 553-566.

Fischetti, M. and Salazar-González, J. J. (1999) Models and solving the cell suppression problem for linearly constrained tabular data. In *Statistical Data Protection (SDP'98) Proceedings*, Eurostat, Luxembourg, 401-409.

Fischetti, M. and Salazar-González, J.J (2000), Models and algorithms for optimizing cell Suppression in tabular data with linear constraints. *Journal of the American Statistical Association* **95**, 916-928.

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

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf, P.-P. (1998) Post randomisation for statistical disclosure control: Theory and implementation (with discussion). *Journal of Official Statistics* **14** 463-484.

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.

Kelly, J.P., Assad, A.A. and Golden, B.L. (1990) The Controlled Rounding Problem: Relaxations and Complexity Issues. *OR Spektrum* **12** pp. 129-38.

Kelly, J., Golden, B., and Assad, A. (1990) Controlled rounding of tabular data. *Operations R*esearch **38** 760-772.

Kelly, J., Golden, B., and Assad, A. (1992) Cell suppression: disclosure protection for sensitive tabular data. *NETWORKS* **22** 397-417.

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

Nargundkar, M. S. and Saveland, W. (1972) Random rounding to prevent statistical disclosure. *Proceedings of the American Statistical Association, Social Statistics Section* 382-385.

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.

Özsoyoğlu and Chung (1986) Information loss in the lattice model of summary tables due to cell suppression. *Proceedings of IEEE Symposium on Security and Privacy*, 160-173.

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

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.