NISS

A Risk-Utility Framework for Categorical Data Swapping

Shanti Gomatam, Alan F. Karr and Ashish Sanil

Technical Report Number 132
February, 2003

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

A Risk-Utility Framework for Categorical Data Swapping

Shanti Gomatam, Alan F. Karr and Ashish Sanil*
National Institute of Statistical Sciences
Research Triangle Park, NC 27709–4006
{sgomatam,karr,ashish}@niss.org

February 28, 2003

Abstract

Data swapping is a statistical disclosure limitation method used to protect the confidentiality of data by interchanging variable values between records. We propose a risk-utility framework for selecting an optimal swapped data release when considering several swap variables and multiple swap rates. Risk and utility values associated with each such swapped data file are traded off along a frontier of undominated potential releases, which contains the optimal release(s). Current Population Survey data are used to illustrate the framework for categorical data swapping.

Key words: constrained swaps; data confidentiality; Hellinger distance; optimal release; risk measure; risk-utility frontier; statistical disclosure limitation; swap rate; swapping attribute; unconstrained swaps, utility measure.

1 Introduction

Data swapping is a statistical disclosure limitation technique that works at the microdata level. Confidentiality protection is achieved by modifying a fraction of the records in the database by switching a subset of variables across selected pairs of records (known as swap pairs). The goal is to make it impossible for any intruder to be certain of having identified an individual or entity in the database. The term "data swapping" was first used by [4]. Other papers that discuss data swapping in some detail include [1], [5], [6], [8], [14], [15], [18] and [19], and [23]. Special cases of data

*Support for this research was provided by National Science Foundation grant EIA–9876619 to NISS, and by the National Center for Education Statistics.

Variable Name	Abbreviation	Categories
Age (in years) (*Age*)	A	<25, 25–55, >55
Employer Type (*WrkTyp*)	W	Govt., Priv., Self-Emp., Other
Education (*Educ*)	E	<HS, HS, Bach, Bach+, Coll
Marital Status (*MarStat*)	M	Married, Other
Race (*Race*)	R	White, Non-White
Sex (*Sex*)	S	Male, Female
Average Weekly Hours Worked (*Hours*)	H	< 40, 40, > 40
Annual Salary (*Income*)	I	<\$50K, $\$50K+$

Table 1: Variable categories for CPS-8d data.

swapping have been referred to as "record swapping," "rank swapping" and the "confidentiality edit," and various terminologies have been used in these papers. We use terminology used in [8].

This paper is written from the point of view of a statistical agency that wishes to determine which of several candidate data releases, corresponding to different choices of swap variables, different values of the swap rate and even different realizations of the swapping (which involves randomization), best serves its purposes. We conceive these choices as taking place within a *risk-utility* framework: each release carries both a risk of compromising confidentiality and a data utility to users. Ideally, the agency would select a release with no risk and maximum possible utility. However, in practice one must balance risk and utility in picking an "optimal" release. We propose an approach that trades off risk and utility, and illustrate how it works for particular choices of the risk and utility measures.

Model-based frameworks for trading off risk and utility are described in [7, 22]. The terminology "R-U confidentiality map" is used [7] to describe this tradeoff in the contexts of top-coding and simulation experiments for perturbed multivariate data [11]. A Bayesian approach to contrasting risk and utility for cell suppression is proposed in [22].

In the sections that follow, we use 1993 Current Population Survey data (referred to as CPS-8d henceforth) [2] to illustrate our framework. The CPS-8d database contains 48,842 observations of 8 categorical variables, with categories as shown in Table 1.

Section 2 explains the risk-utility framework, and defines the risk and utility measures used. Section 3 illustrates an application of the risk-utility framework to select an "optimal" release from unconstrained swap collections for different swap rates, as well as two different constrained swaps. We assume throughout that all variables are categorical.

2 Risk-Utility Frontiers

To employ data swapping as a disclosure limitation procedure, an agency must choose appropriate parameter values, in particular the swap variables and swap rate, for the actual release. For

example, suppose that the swap rate is fixed and it has been decided that only one variable will be swapped. Then, ignoring differences caused by the random seed, there are as many possible releases as there are variables in the database.

In our risk-utility framework, each candidate release (for simplicity, we usually omit "candidate" below) is characterized by numerical values of risk and utility. When choosing among releases, the agency would like to pick the one that has both minimum risk and maximum utility. However, as Figure 1—a scatterplot of 10 hypothetical (Utility, Risk) pairs—illustrates, this is ordinarily not possible: higher utility entails higher risk. However, not all releases are sensible. Specifically, any release dominates all releases lying to its northwest, in the sense of having both higher utility and lower risk than they do. Therefore, the agency should consider only releases on the *frontier* (in economics, the *efficient frontier*)— those with no other releases to the southeast. In Figure 1, points on the risk-utility frontier are connected with a solid line.

If the frontier consisted of a single release, that release would be optimal. However, because risk and utility increase together, in practice the frontier contains several releases. Selection of a release on the frontier can be done by assessing the risk-utility balance subjectively or quantitatively, by means of an objective function that relates risk and utility. To illustrate, the dashed line in Figure 1 corresponds to a linear risk-utility relationship of the form

$$\text{Risk} = a \times \text{Utility} + c,$$

and the figure identifies the release on the frontier that is optimal for a particular value of a. Similar approaches have been used in economics to maximize consumer utility for the purchase of a combination of two commodities [10].

Operationally, instead of risk and utility, we use an equivalent risk-distortion (distortion being a form of "dis–utility") formulation of the problem. The risk-distortion plots in §3 are more similar to those in the economics literature than Figure 1 because risk and distortion are substitutes. However, in contrast to the economics literature where the frontiers are usually modeled as smooth functions, our problem is essentially discrete, because there is only a finite number of release candidates. (The swap rate is, of course, a continuous variable, but swapping can be done for only finitely many choices.) Our approach, which is necessarily empirical, consists of three steps: First, perform swapping with various swap variables and swap rates to obtain n candidate releases, $\text{RC}_1, \text{RC}_2, \ldots, \text{RC}_n$. Then evaluate $(\text{Risk}(\text{RC}_i), \text{Distortion}(\text{RC}_i))$ pairs for each of the RC_i. Finally, determine the frontier and select the best release. If desired, as in §3 and [9], plots or visualization tools can be used to increase understanding of the candidate releases and the frontier.

2.1 Risk Measure

We restrict attention to releases of the entire database. (Sampling is another utility-reducing strategy to reduce risk.) In this case, database uniques or near uniques are potentially riskier than other elements. For categorical data, these elements are contained in small count cells in the contingency table created by using all variables in the data. The n-rule, which is widely used in statistical disclosure limitation, considers records that fall in cells with a non-zero count less than n (typically

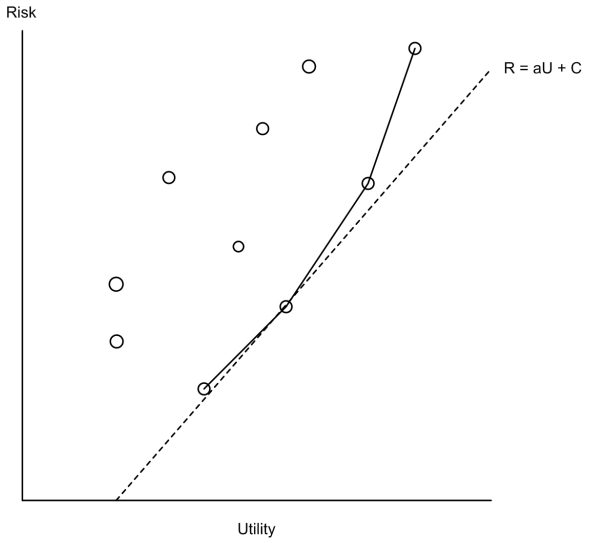

Figure 1: Conceptual example of risk-utility frontier and an optimal release for a linear tradeoff between risk and utility.

$n = 3$) to be at risk. Thus, one measure of risk is the proportion of *unswapped* records in small count cells in the table created from post-swap data:

$$\text{Risk} = \frac{\sum_{C_1,C_2} \text{Number of unswapped records}}{\text{Total number of unswapped records}},$$

where C_1 and C_2 are the cells in the full data table with counts of 1 and 2.

Other measures of risk are based primarily on re-identification through record linkage to an external database [3, 5, 12, 17, 20, 21, 19]. Such methods define risk as the fraction of total unswapped records that can be linked either uniquely, or nearly uniquely.

2.2 Utility Measure

Our basic approach is to measure utility by comparing the pre- and post-swap databases: the "closer" these are, the higher the utility. For clarity, for the remainder of the paper, we term lack of closeness *distortion*, which is therefore dis-utility.

Several utility measures used to quantify distortion due to data swapping have been compared in [8] using the CPS-8d data. Of these, Hellinger distance appeared to capture most successfully the general trend indicated by all measures. Hellinger distance [13] between distributions with probability mass functions f and g on a countable set is defined as

$$H(f,g) = \sqrt{\frac{1}{2}\sum_{C}\left(\sqrt{f(C)} - \sqrt{g(C)}\right)^2}, \tag{1}$$

where the sum is over cells C in the full data table.

In this paper we use the 8-way (empirical) Hellinger distance between the pre- and post-swap tables to measure distortion: in (1), f and g correspond to the pre-swap and post-swap frequencies in the CPS-8d data table.

Other utility measures applicable to data swapping have been proposed and studied [1, 5, 14, 15, 19]. With the exception of conditional entropy, variations or generalizations of these methods are considered in [8]. Measures of utility that are tied more closely to uses of the data for statistical inference are a topic for future research.

3 Performance Comparisons

In this section we present (Risk, Distortion) values and frontiers for unconstrained swaps ("random true swaps" [8]) and constrained swaps of the CPS-8d database. Constrained swaps involve restrictions on variables on other than those swapped: for instance, we may only allow swaps between records with the same value for *Sex*.

Collections of swaps in the figures given below include all single variable swaps and all simultaneous swaps of two variables.[1] The abbreviations in Table 1 are used to indicate variables swapped, with concatenated abbreviations indicating two-variable swaps.

3.1 Unconstrained Swaps

To study the dependence of risk distortion on swap variable(s) and swap rate, we consider three collections of unconstrained swaps, corresponding to swap rates 0.005, 0.01 and 0.05. The swap rate indicates the approximate half-fraction of observations swapped, so a 0.005 rate corresponds to a swap of approximately 1% of the data [16]. Each collection contains 8 single-variable swaps and 28 two-variable swaps. Hence we have $108 = 36 \times 3$ possible swapping conditions,[2] each resulting in a (Risk, Distortion) pair of values.

Figure 2 contains risk-distortion scatterplots of the single- and two-variable swaps for each of the three rates separately. Lines connect the releases on each frontier. A user who has decided on a rate need only look at the plot corresponding to that rate and decide which single- or two-variable swap release on the frontier best captures relevant risk and distortion tradeoffs.

Alternatively, a user who is undecided about the swap rate would select from the frontier generated by combining swaps for all rates of interest. Figure 3 shows the combined plot for the three swap rates. The frontier for the combined plot is a strict subset of the union of all three frontiers. In particular, notice that the 0.05 swap of *Educ*, which was on the frontier for the 0.05 swap rate, is dominated by many of the 0.01 swaps.

[1] In a simultaneous swap, all swap variables are exchanged between the two records at the same time [8].

[2] The 0.05 rate swap involving *Race* and *Income* was not feasible and has been left out of the plots.

Figure 3 shows clearly that distortion increases and risk decreases as the swap rate increases. Collectively, single-variable swaps tend to be riskier than two-variable swaps. As swap rate increases, variability (over choice of swap variables) of both risk and utility increases.

Since the data swapping algorithm involves randomization [16], different random seeds result in different post-swap databases, and hence different risk and distortion values, even for the same swap variables and swap rate. In (small numbers of) replications with multiple seeds, we observed overall consistency, with some variability, in the risk-distortion scatterplots. In principle, an agency could compare collections resulting from different randomizations in order to pick a release using the same risk-utility framework: combine the releases from all randomizations in a single plot and pick an optimal release from the resulting frontier.

3.2 Constrained Swaps

Two constrained swap collections for the 1% swap rate are presented. Note that constrained swaps cannot be performed if the number of record pairs to be swapped is larger than the number of record pairs (picked without replacement) satisfying the constraints. Multiple, stringent constraints may therefore be problematic.

For the first collection, *Sex* is constrained to be the same for any feasible swap pair. In the second, *MarStat* was constrained to differ. Figure 4 contains the resulting risk-distortion scatterplots. The range of both risk and distortion values is nearly the same as for the unconstrained swap at the 0.01 rate. Although the frontier has changed, we see no obvious systematic effects, especially in light of randomization-induced variability.

4 Conclusion

We have presented a framework that can be used to discriminate among post-swap data releases on the basis of risk and utility. This technique has been illustrated for candidate releases corresponding different choices of swap variable(s) and swap rate, but it applies equally well to other parameters, for instance, the choice of random seeds for the swaps. It can even be used to compare multiple disclosure methods for which the same risk and utility measures are reasonable.

References

[1] M. Boyd and P. Vickers. Record swapping—a possible disclosure control approach for the 2001 UK Census. In *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, 1999.

[2] Census Bureau. Current Population Survey, 2002. Information available on-line at www.bls.census.gov/cps/cpsmain.htm.

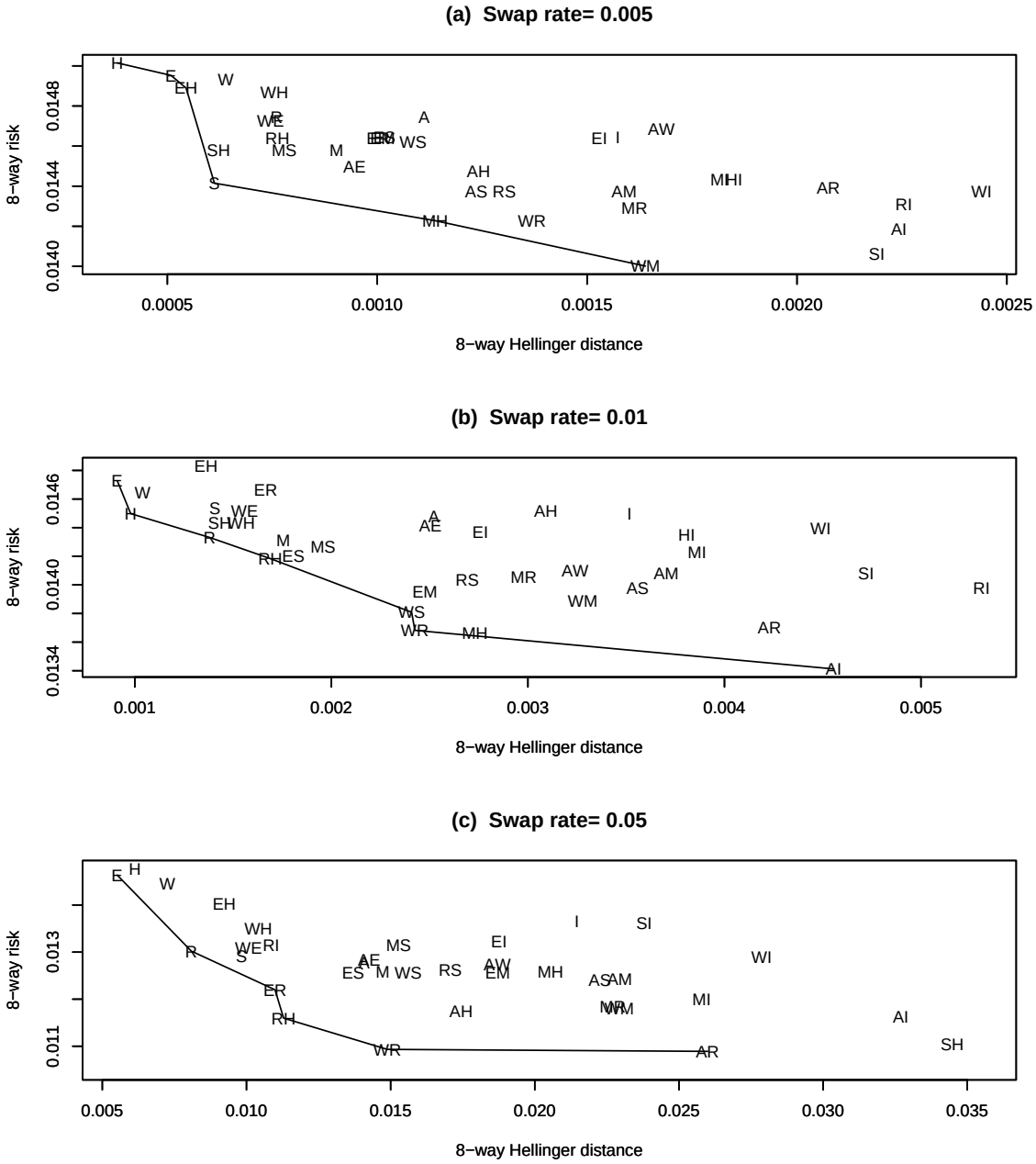

Figure 2: Risk-distortion scatterplots showing frontiers of 8-way Hellinger distance versus 8-way risk for unconstrained swaps.

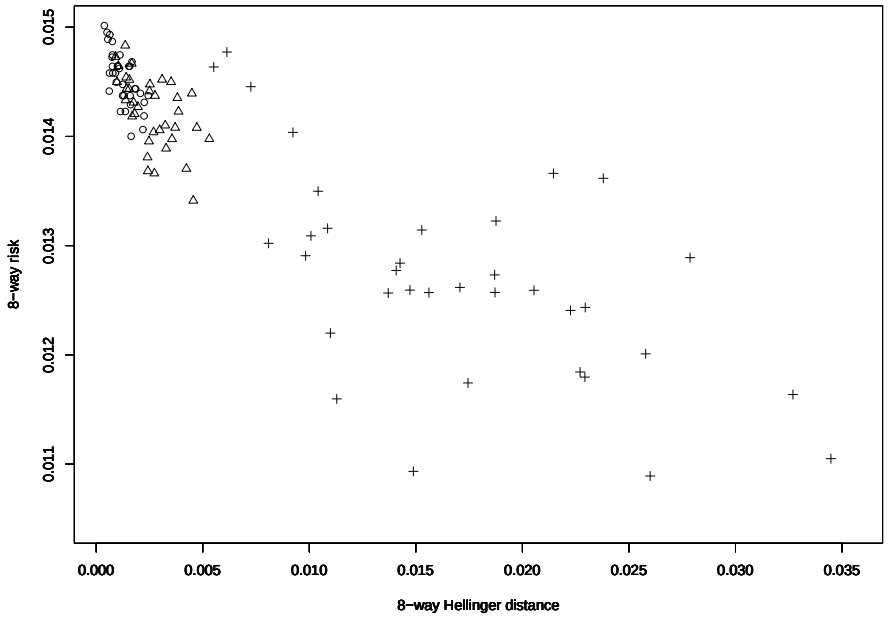

Figure 3: Graph of 8-way Hellinger distance versus 8-way risk for all 3 rates. Circles represent 0.005 rate swaps, triangles represent 0.01 rate swaps, and plusses represent 0.05 rate swaps.

[3] L. H. Cox. Confidentiality problems in microdata release. *Proceedings of the Third Annual Symposium on Computer Applications in Medical Care, IEEE Computer Society*, pages 397–402, 1979.

[4] T. Dalenius and S. P. Reiss. Data-swapping: A technique for disclosure limitation. *J. Statist. Planning Inf.*, 6:73–85, 1982.

[5] J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing SDC methods for microdata on the basis of information loss and disclosure risk, 2001. Presented at UNECE Workshop on Statistical Data Editing, Skopje, Macedonia.

[6] G. T. Duncan and S. A. Keller-McNulty. The impact of data swapping on confidentiality and data utility, 2000. Talk presented to Institute for Social Research, University of Michigan.

[7] G. T. Duncan, S. A. Keller-McNulty, and S. L. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map. *Manag. Sci.*, 2002. Submitted for publication.

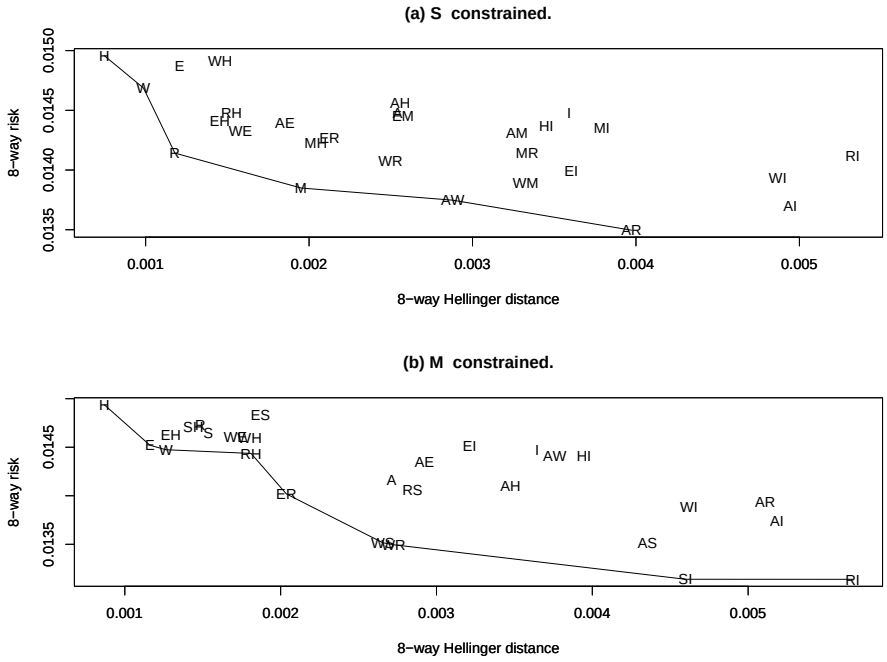

Figure 4: Risk-distortion scatterplots for constrained swaps at the 0.01 swap rate. Figure 4 (a) contains the collection when *Sex* was constrained to be the same, and Figure 4 (b) contains the one where *MarStat* was constrained to be different.

[8] S. Gomatam and A. F. Karr. Distortion measures for categorical data swapping. *J. Official Statist.*, 2003. Submitted for publication; PDF version available on-line as Technical Report 131 at www.niss.org/publications.html.

[9] S. Gomatam, A. F. Karr, C. Liu, and A. P. Sanil. Data swapping: A risk–utility framework and web service implementation, 2003. Submitted for presentation at dg.o2003.

[10] J. M. Henderson and R. E. Quandt. *Microeconomic Theory: A Mathematical Approach*. McGraw-Hill, New York, 1958.

[11] J. J. Kim and W. Winkler. Masking microdata files. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 114–119, 1995.

[12] D. Lambert. Measures of disclosure risk and harm. *J. Official Statist.*, 9(2):313–331, 1993.

[13] L. Le Cam and G. L. Yang. *Asymptotics in Statistics*. Springer–Verlag, New York, 1990.

[14] R. A. Moore. Controlled data-swapping techniques for masked public use microdata sets, 1996. US Census Bureau, Statistical Research Division, Washington.

[15] R. A. Moore. Preliminary recommendations for disclosure limitation for the 2000 census: Improving the 1990 confidentiality edit procedure, 1996. US Census Bureau, Statistical Research Division, Washington.

[16] A. Sanil, S. Gomatam, A. F. Karr, and C. Liu. NISS WebSwap: A web service for data swapping. *J. Statist. Software*, 2003. Submitted for publication; PDF version available online as Technical Report 126 at www.niss.org/publications.html.

[17] N. L. Spruill. Measure of confidentiality. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 260–265, 1982.

[18] L. C. R. J. Willenborg and T. de Waal. *Statistical Disclosure Control in Practice*. Springer–Verlag, New York, 1996.

[19] L. C. R. J. Willenborg and T. de Waal. *Elements of Statistical Disclosure Limitation*. Springer–Verlag, New York, 2000.

[20] W. E. Winkler. Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Res. Official Statist.*, 1:87–104, 1998.

[21] W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata. *Inf. Control in Statist. Databases*, 2002.

[22] A. M. Zaslavsky and N. J. Horton. Balancing disclosure risk against the loss of nonpublication. *J. Official Statist.*, 14(4):411–419, 1998.

[23] L. Zayatz, P. Steel, and S. Rowland. Disclosure limitation for Census 2000, 2000. US Census Bureau, Statistical Research Division, Washington.