A Data–Swapping Technique Using Ranks - A Method for Disclosure Control

Michael Carlson* Mickael Salabasis†

December 13, 2002

Abstract

A data-swapping technique based on ranks is described and suggested as a possible approach to statistical disclosure control. The proposed method is intended to be applied to quantitative data and utilizes the rank structure of disjoint subsets of an original data set; values of one subset are exchanged for values of other subsets. The procedure retains the validity of a sample on an intravariate level but the association between pairs of variables is typically weakened. Theoretical and simulation results indicate that the proposed method performs reasonably well in the bivariate normal case.

Keywords: Concomitants; Data-swapping; Data dissemination; Disclosure control; Order statistics; Ranks.

1 Introduction

With the gradually improved means for researchers and others to conduct their own explorations there has followed an increase in the demand for the collection and dissemination of public-use microdata files. This creates a natural conflict. On one hand there is the demand for making data widely accessible as the principal goal of data collection is research. On the other the rights of the respondents, both natural and legal, need to be protected against unnecessary exposure, as much statistics are based on information that is by some definition sensitive. In such instances producers of statistical

*Corresponding author. Department of Statistics, Stockholm University SE-106 91 Stockholm, Sweden. E-mail: Michael.Carlson@stat.su.se.

†Department of Economic Statistics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, Sweden.

data have a justifiable reluctancy to make the data available and may instead offer to conduct the analysis themselves. This process may prove both time consuming and expensive for the end-user. Furthermore, the end-user may be reluctant to specify the exact nature of the questions being asked since this could in itself reveal sensitive information, for instance political intentions or market strategies. Under these circumstances it is easy to identify a demand for methods that will help to increase the availability of microdata while at the same time offering a reasonable level of protection against improper disclosure. Comprehensive overviews and discussions of the problems associated with the dissemination of data and statistical disclosure control are given by e.g. Dalenius (1974, 1977), Frank (1976, 1983), Duncan and Lambert (1989), Bethlehem et al. (1990), Duncan and Pearson (1991), Reynolds (1993), Lambert (1993), Fienberg (1994) and also in references cited in these papers. Willenborg and de Waal (1996, 2000) provide excellent introductions to the issues concerned. Recent publications include also the special issue of the Journal of Official Statistics (1998, vol. 14), Doyle et al. (2001) and Domingo-Ferrer (2002).

The method suggested in this paper is a variant of data-swapping. Data-swapping as a means for disclosure control was suggested in Dalenius (1979) and Dalenius and Reiss (1982) for the case where the attributes assume categorical values. Fienberg et al. (1998) relates data-swapping methods for categorical data to conventional statistical methods associated with loglinear models. A method for dealing with quantitative data was illustrated in Dalenius (1988). Greenberg (1987) and Moore (1996) describe a method for ordinal data based on rank-proximity. Variants of data-swapping have also been used in practice by the US Census Bureau e.g. for the release of tabular data from the 1990 census. We refer to the paper by Moore (1996) and to references cited in Fienberg et al. (1998). See also Paass (1988) and Spruill (1983). Recent work may be found in Domingo-Ferrer and Torra (2001), Dandekar et al. (2002) and Sebé et al. (2002).

Whereas most data-swapping procedures described in the literature involve swapping values within the same sample or set, the method proposed in this paper exchanges the values of a pre-specified subset of a larger data set with the values of other subsets from the same larger set. The exchange is based on the rank structure of each of the involved subsets and is intended for quantitative data. The resulting set which is made public instead of the original set is a combination of values pertaining to respondents from several disjoint sets, the purpose being to enhance the level of confidentiality while maintaining a reasonable level of congruity with the original data. It should however be noted that the ideas in this paper are tentative and have not been tested on real-life data. The focus here is limited to describing the

basic swapping procedure and on the asymptotic bivariate and univariate properties of the resulting data sets. The numerical examples throughout this paper are based on experiments and before the methods can be used on a larger scale, more research is needed.

This paper is organized as follows. Properties of data-swapping methods in general are discussed in section 2. In section 3 some basic notation is introduced and the swapping procedure is described by means of simple examples in section 4. Section 5 provides some theory regarding the effect the swapping operations have on the association between pairs of variables and illustrates the effect by means of simulation studies. The univariate effects are described in section 6. In section 7 we discuss the effect of the data-swapping procedure on disclosure limitation and on statistical analysis. Finally some concluding remarks and prospects for future research are given in section 8.

2 Data-Swapping Methods

Data-swapping transforms a data set into another by exchanging attribute scores between respondents so that the value of a variable for a respondent is not her own value but the value of some other respondent. Normally data-swapping affects the structure of the data, especially if the swapping procedure is performed without restrictions. In general any data-swapping procedure will have its benefits but also its shortcomings. The following list adapted from Moore (1996) briefly summarizes some of the properties:

- Data-swapping masks accurate information about each respondent.
- The relationship between the record and the respondent is weakened if performed on potential key variables.
- Swapping can be used on a select set of variables, without disturbing the responses for non-sensitive and non-identifying attributes.
- Swapping of continuous variables can provide protection where it is necessary since rare and unique attributes and combinations of attributes are generally used to identify respondents.
- The variables selected for swapping are subject to additional error, diminishing the analytical value of the data. Typically multivariate relationships are distorted.

- Arbitrary swaps can produce a large number of unusual observations in sub-domains, e.g. newly born with high incomes. Even unrealistic or inconsistent combinations may occur.

- The procedure is simple, the programming is straightforward and in the case of rank-based methods the execution is as fast as the sorting algorithm used.

- A large number of records in the original file and a large number of variables selected for swapping can take a significant amount of computer resources; rank-based methods are only as fast as the sorting algorithm used.

Swapping also has the psychological advantage that it is often possible to say that the sensitive properties assigned to a person are never his own. The fact that Mr. Brown is coupled with the property "tax evader" is due to the fact that some one else in the material is in fact a tax evader. Furthermore, one can not be sure that it is Mr. Brown at all since the attributes used to link records in the released data to individuals in the population may have been swapped as well.

3 Notation

Consider a vector of random variables $(X, Y, \dots, W)$ generated from a continuous multivariate distribution with finite first and second order moments. Denote the expectation of a general variable Z by μ_Z and the corresponding variance by σ_Z^2. The covariance between two variables Z, U is denoted by σ_{ZU}. The cumulative distribution function and probability density function of a random variable Z will be denoted by $F_Z(z)$ and $f_Z(z)$, respectively. The multivariate distribution and the respective marginal distributions of $(X, Y, \dots, W)$ will be referred to as the parent distributions.

Assume that several independent random samples $\mathbf{M}_a$, $a = 1, 2, \dots$ of equal size n are available. A set $\mathbf{M}_a$ may be arranged as

$$\mathbf{M}_a = \left[\begin{array}{cccc} \mathbf{X}_a & \mathbf{Y}_a & \cdots & \mathbf{W}_a \end{array} \right] = \left[\begin{array}{cccc} X_{1,a} & Y_{1,a} & \cdots & W_{1,a} \\ \vdots & \vdots & & \vdots \\ X_{n,a} & Y_{n,a} & \cdots & W_{n,a} \end{array} \right]$$

where each row $(X_{i,a}, Y_{i,a}, \dots, W_{i,a})$, $i = 1, \dots, n$ denotes one observation of $(X, Y, \dots, W)$ and each column is the $n \times 1$ vector corresponding to respective random variable. The latter is denoted in bold typeface, e.g. $\mathbf{Z}_a$.

Let $X_{r:a}$ denote the r'th order statistic of the X's in the set $\mathbf{M}_a$. The variable $Y_{i,a}$ associated with $X_{r:a}$ will in the following be denoted by $Y_{[r:a]}$ and termed the concomitant to the r'th order statistic of X, adopting the terminology of, among others, David (1981). The concomitant to the s'th order statistic of Y is correspondingly denoted by $X_{[s:a]}$. A more appropriate notation should perhaps reflect the specific variate to which the concomitant is associated, e.g. $Y_{[r:a]_X}$ is the concomitant to $X_{r:a}$. We will however in the following focus on relations between pairs of variables and the present notation will suffice. For example, an observation on X and Y in $\mathbf{M}_a$ may be denoted by either of the following

$$(X_{i,a}, Y_{i,a}) = (X_{r:a}, Y_{s:a}) = \left(X_{r:a}, Y_{[r:a]}\right) = \left(X_{[s:a]}, Y_{s:a}\right)$$

assuming Rank $[X_{i,a}] = r$, and Rank $[Y_{i,a}] = s$. A general theory for concomitants of order statistics is given in e.g. Yang (1977) and David (1981).

4 Description of the Data-Swapping Procedure

The basic idea that the data-swapping method builds on is simple and is perhaps best illustrated by simple examples involving only two variables, X and Y say. Since the available subsets (samples) $\mathbf{M}_1, \mathbf{M}_2 \ldots$ are independent realizations from the same distribution, the conditional distribution of the r'th concomitants in any two samples are identical. Thus, conditional on X we have

$$f_{Y_{[r:1]}|X_{r:1}}\left(y \mid X_{r:1} = x\right) \;=\; f_{Y_{[r:2]}|X_{r:2}}\left(y \mid X_{r:2} = x\right) \;=\; f_{Y|X}\left(y \mid X = x\right).$$

Furthermore, since $E\left[X_{r:1}\right] = E\left[X_{r:2}\right]$, one would expect $X_{r:1}$ to be approximately equal to $X_{r:2}$ for large n, therefore justifying the approximation

$$E\left[Y_{[r:1]} \mid X_{r:1} = x_1\right] \quad \approx \quad E\left[Y_{[r:2]} \mid X_{r:2} = x_2\right].$$

Hence a simple way to create artificial observations is to exchange the observed order statistics between independent samples i.e. $X_{r:1}$ is swapped for $X_{r:2}$, for all $r = 1, \ldots, n$.

Assume that three disjoint and equally sized subsets $\mathbf{M}_1 = [\mathbf{X}_1, \mathbf{Y}_1]$, $\mathbf{M}_2 = [\mathbf{X}_2, \mathbf{Y}_2]$ and $\mathbf{M}_3 = [\mathbf{X}_3, \mathbf{Y}_3]$ are available of which we select $\mathbf{M}_1$ to build on and refer to as the reference set. The other sets, $\mathbf{M}_2$ and $\mathbf{M}_3$ will provide the values that replace the original values of $\mathbf{M}_1$ and we refer to these as the auxiliary sets.

Define $\tilde{\mathbf{X}}_a$ as the vector consisting of the n entries in $\mathbf{X}_a$ rearranged in ascending order. The permutation of $\mathbf{X}_a$ to $\tilde{\mathbf{X}}_a$ is easily defined as a linear transformation

$$\mathbf{R}_{Xa}\mathbf{X}_a = \tilde{\mathbf{X}}_a$$

where $\mathbf{R}_{Xa}$ is the $n \times n$ orthogonal matrix with ones in the entries corresponding to the shifts in $\mathbf{X}_a$ and zeroes elsewhere. For example, $X_{i,a} = X_{r:a}$ entails a one in the i'th column of the r'th row of $\mathbf{R}_{Xa}$ and zeroes in the remaining positions of that same row. The reverse operation is easily obtained by $\mathbf{R}_{Xa}^T\tilde{\mathbf{X}}_a = \mathbf{X}_a$ where $\mathbf{R}_{Xa}^T$ is the transpose of $\mathbf{R}_{Xa}$. The matrices $\mathbf{R}$ will in the following be referred to as ordering permutations.

To exchange the values of $\mathbf{X}_1$ for those in $\mathbf{X}_2$, the observations in $\mathbf{X}_2$ are first rearranged so that the rank ordering is the same as the ordering given by the observations in $\mathbf{X}_1$. This is easily accomplished by the operation

$$\mathbf{R}_{X1}^T\mathbf{R}_{X2}\mathbf{X}_2 = \mathbf{X}_1^*. \tag{1}$$

The column vector $\mathbf{X}_1$ is thereafter exchanged for $\mathbf{X}_1^*$, i.e. $X_{r:1}$ is swapped for $X_{r:2}$, for all $r = 1, \ldots, n$. The star notation $(*)$ will throughout denote that the original values have been swapped for new ones.

Example 1 *To illustrate the procedure consider two sets of data given by*

$$\mathbf{M}_1 = \begin{bmatrix} 46 & 45 \\ 26 & 39 \\ 63 & 44 \end{bmatrix}, \quad \mathbf{M}_2 = \begin{bmatrix} 72 & 40 \\ 32 & 59 \\ 61 & 60 \end{bmatrix}.$$

Swapping the scores in $\mathbf{X}_1$ *for those in* $\mathbf{X}_2$ *results to*

$$\begin{aligned} \mathbf{M}_1^* &= \left[\ \mathbf{R}_{X1}^T\mathbf{R}_{X2}\mathbf{X}_2, \quad \mathbf{Y}_1\ \right] \\ &= \left[\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 72 \\ 32 \\ 61 \end{bmatrix}, \quad \begin{bmatrix} 45 \\ 39 \\ 44 \end{bmatrix} \right] = \begin{bmatrix} 61 & 45 \\ 32 & 39 \\ 72 & 44 \end{bmatrix}. \end{aligned}$$

Carrying the operation one step further we may consider swapping the Y scores as well. By rearranging the observations in $\mathbf{Y}_2$ with respect to $\mathbf{Y}_1$ and then swapping the observed values of the same, yet another set is created. Alternatively, we can consider swapping the scores in $\mathbf{Y}_1$ for those given by $\mathbf{Y}_3$,i.e. we use two separate sources for the two variables. The re-ordering transformations in either case are defined analogously to (1).

Extending the procedure to situations with more than two variables is straightfoward. A simple example involving six variables is depicted in figure

1. Here three of the variables, X, Y and W, have been selected to remain intact, i.e. the original observations of $\mathbf{M}_1$ are retained. The variables Z and U are swapped for values from the auxiliary source $\mathbf{M}_2$ and the variable V is swapped for values from $\mathbf{M}_3$. Formulating the operation in terms of re-ordering permutations would yield the following expression:

$$\mathbf{M}_1^* = \left[\begin{array}{ccccc} \mathbf{X}_1, & \mathbf{Y}_1, & \mathbf{R}_{Z1}^T\mathbf{R}_{Z2}\mathbf{Z}_2, & \mathbf{R}_{U1}^T\mathbf{R}_{U2}\mathbf{U}_2, & \mathbf{R}_{V1}^T\mathbf{R}_{V3}\mathbf{V}_3 \quad \mathbf{W}_1 \end{array} \right].$$

When assessing what happens with the association between pairs of variables we note that any given pair in this example (and in general) is described by one of the following: (a) swapping only one variable, e.g. the relationship between X and Z, (b) swapping both variables with values from the same source, e.g. Z and U, (c) swapping both variables with values from separate sources, e.g. Z and V and finally (d) no swap, e.g. the relationship between the unswapped variables X and Y. The first three cases are investigated further in section 5, the fourth is trivial. For higher order relationships the number of possible combinations one needs to consider will of course be larger.

On the univariate level it can be argued that the swapping procedure is equivalent to adding some kind of error or noise to the individual scores. On the other hand it should be clear that any univariate inference is equally valid after the swap, as before; one sample of scores has simply been exchanged for another of equal size. However, from the viewpoint of disclosure limitation it is important to investigate the properties of the added error and we return to these issues in sections 6 and 7.

On the univariate level it can be argued that the swapping procedure is equivalent to adding some kind of error or noise to the individual scores. On the other hand it should be clear that any univariate inference is equally valid after the swap as before; one sample of scores has simply been exchanged for another of equal size. However, from the viewpoint of disclosure limitation it is important to investigate the properties of the added error and we return to these issues in sections 6 and 7.

4.1 Related methods

At least two rank-based swapping techniques for continuous data have been proposed in the literature; Dalenius (1988, ch. 27, pp. 6-9) and Greenberg (1987), the latter being explored by Moore (1996). The procedure suggested by Dalenius was illustrated in the case with two variables, X and Y. The original data matrix $\mathbf{M}$ is arranged in ascending order of X, and in the first step, $\mathbf{M}$ is partitioned into k consecutive and continuous subsets of

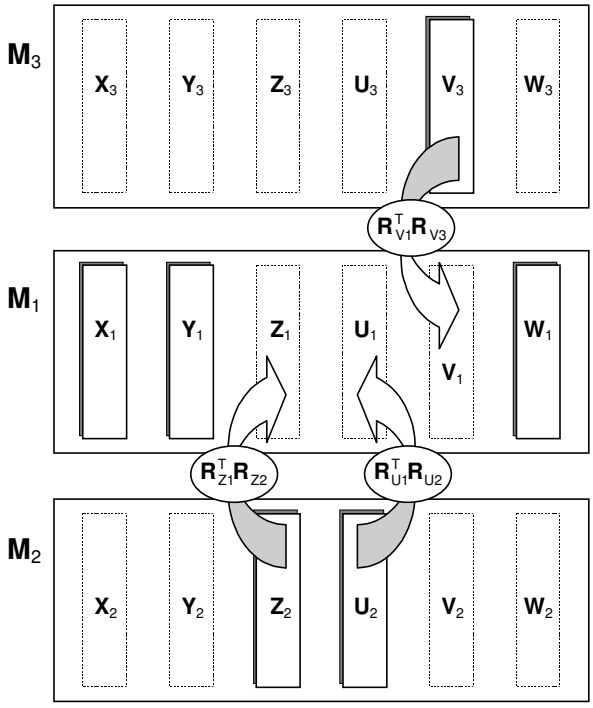

Figure 1: Example of the swapping procedure with six variables and three subsets.

equal size, with $1 < k < n$. Typically k should be much smaller than n. Note that the data in each subset is still arranged in increasing order of X. Next, the Y-values within each of the k subsets are permuted at random, generating the masked matrix $\mathbf{M}^*$. Permuting the X-values instead of the Y's, would have an equivalent effect. The procedure could then be repeated for each variable selected for swapping, given the initial partitioning. Dalenius gave the following two remarks: (1) the intravariate properties are exactly retained, since all the values originally present in $\mathbf{M}$ are retained in the released matrix. As for the correlation between X and Y, the resulting bivariate relationship will be subject to error which in general will decrease with increasing k. (2) The protection provided by disseminating $\mathbf{M}^*$ in place of $\mathbf{M}$ will clearly depend on the choice of k; increasing k will reduce protection since the number of available candidates to swap with is reduced.

The method proposed in this paper is apparently a reversed version of the Dalenius proposal. That is, Dalenius will first order the records in accordance with the ranks on the entire set, then partition the set into k subsets and finally randomly swap within each subset. Our proposal on the other hand

starts off by randomly partitioning the records of the entire set into k subsets, orders the records within each subset and then swaps values between subsets. In short, Dalenius stratifies the data and randomly swaps within strata, we randomly cluster the data and then swap between clusters. The implications noted by Dalenius will accordingly be the opposite.

The method proposed by Greenberg and Moore applies a different approach. Start with a data set $\mathbf{M}$ of size n and order the responses in ascending order of a single variable X. Determine a value P_X with $0 \leqslant P_X \leqslant 100$, the intention being to swap the value $X_{r:a}$ with that of $X_{s:a}$ so that the percentage difference of the indices (ranks) r and s is less than P_X of n. Initialize all records set to top- or bottom-code and records of all imputed and blank values as "swapped". All other records are initialized as "unswapped". Starting with the lowest unswapped rank, say r, randomly select a record with an unswapped rank from the interval $[r+1, m]$ where $m = \min\{N,\ r + nP_X/100\}$, hence motivating the term "rank-based proximity swap". Assuming the randomly selected rank is s, the values of $X_{r:a}$ and $X_{s:a}$ are swapped and labelled as "swapped". The procedure is then repeated until all ranks are labelled "swapped". The entire procedure can of course be performed on several additional variables. Furthermore the values P for the different variables need not equal each other. Moore defined criteria for choosing the P values with the purpose of preserving univariate and covariate properties of the original data.

The procedure of Greenberg and Moore resembles the proposal of Dalenius. The main difference is that the data is not initially partitioned into subsets, rather the data values are swapped on a record-by-record basis. The interval $[r+1, m]$ can be viewed upon as a window that moves across the records of the variable selected for swapping, determining the range of possible swapping candidates. The value P_X which determines the range of the interval is the counterpart of k in the method of Dalenius, the number of subsets which directly determines the size of the subsets.

4.2 Variations of the swapping procedure

Given the basic outline, variations of the described method are easily conceived. Although an in-depth investigation of such extensions is beyond the scope of this paper, we might hint at some possibilities. First, it is possible to make use of the entire data file and not just disjoint subsets of the records over separate variables. As seen in figure 1 only one third of all available values are used in the final set of this example. Extending the swap scheme to include two-way swaps between blocks of attributes or even chains of swaps, the entire set of values can be used. In this way, several subsets act as refer-

ence sets. By merging the resulting subsets, a final set $\mathbf{M}^*$ is created, which is equal in size to the original set. Furthermore, the univariate characteristics would remain intact since all of the original values are retained in the resulting set. Extending the example in figure 1 could for example result in the following swap-scheme:

$$\mathbf{M}^* = \begin{bmatrix} \mathbf{X}_1 & \mathbf{Y}_1 & \mathbf{R}_{Z1}^T\mathbf{R}_{Z2}\mathbf{Z}_2 & \mathbf{R}_{U1}^T\mathbf{R}_{U2}\mathbf{U}_2 & \mathbf{R}_{V1}^T\mathbf{R}_{V3}\mathbf{V}_3 & \mathbf{W}_1 \\ \mathbf{X}_2 & \mathbf{Y}_2 & \mathbf{R}_{Z2}^T\mathbf{R}_{Z1}\mathbf{Z}_1 & \mathbf{R}_{U2}^T\mathbf{R}_{U1}\mathbf{U}_1 & \mathbf{R}_{V2}^T\mathbf{R}_{V1}\mathbf{V}_1 & \mathbf{W}_2 \\ \mathbf{X}_3 & \mathbf{Y}_3 & \mathbf{Z}_3 & \mathbf{U}_3 & \mathbf{R}_{V3}^T\mathbf{R}_{V2}\mathbf{V}_2 & \mathbf{W}_3 \end{bmatrix}.$$

Here the scheme has been extended to include two-way swaps on the Z and U variables. Note also that the blocks $\mathbf{Z}_3$ and $\mathbf{U}_3$ have been left unswapped, the obvious effect being that the association between e.g. X and Z is lesser strained then if all the values of Z had been subjected to swapping. The drawback would of course be a lesser degree of masking of the original data. An example of a chain of swaps is given for the variable V where $\mathbf{V}_1$ is swapped for $\mathbf{V}_3$, $\mathbf{V}_2$ for $\mathbf{V}_1$ and $\mathbf{V}_3$ for $\mathbf{V}_2$. Once a masked set has been generated in this way it would of course be possible to release only a sample from it.

Allowing the subset size to depend on the degree of sensitivity is a variation suggested by the relationship between partition size, preservation of characteristics and disclosure control. Using smaller subsets for certain sensitive attributes the disclosure control can be enhanced. The increasing costs in terms of overall degradation induced could be checked by increasing the subset size of all other variables. This would in effect correspond to varying the value P_X of the method proposed by Greenberg (1987) and Moore (1996). It is also possible to consider partitioning anew for each variable selected for swapping.

A third variation is achieved by first taking a sample $\mathbf{M}_1$ of the original data set $\mathbf{M}$ and then recording the ranks with respect to $\mathbf{M}_1$. The size n of the sample would equal the number of records considered for release. The values of the sample are then exchanged for the corresponding quantiles of the remaining records, i.e. $\mathbf{M} \setminus \mathbf{M}_1$. The procedure might prove useful in cases were the size of the original data set is judged too large to be manageable as it is easier to implement and more rapidly executed compared to subjecting the entire set to a swapping scheme. Also the actual values in the resulting set will be subject to a lesser degree of variation; the values will depend less on the sample $\mathbf{M}_1$ and more on the sampling fraction. The degree of randomness is mainly attributed to the rank structure of the selected sample $\mathbf{M}_1$ which will vary between different samples.

5 Cross Product Moments

In this section we investigate the expected effects of the swapping procedure on the covariate association between pairs of variables within a subset as gauged by the cross product moment under some basic model assumptions. We consider the simple case with only two variables, X and Y, and explore the three different cases noted in section 4. Although exact expressions for the cross product moments are derived, the results are quite complicated. To illustrate the behavior we have used simulation studies under a bivariate normal distribution.

Assume that X and Y are generated from a continuous bivariate distribution with finite first and second order moments and that they are linked by a linear regression relation given by

$$Y_{i,a} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\left(X_{i,a} - \mu_X\right) + \varepsilon_{i,a}, \qquad |\rho| \leqslant 1 \tag{2}$$

where the $X_{i,a}$ and $\varepsilon_{i,a}$ are mutually independent. From (2) it follows that $E\left(\varepsilon_{i,a}\right) = 0$, $Var\left(\varepsilon_{i,a}\right) = \sigma_Y^2\left(1-\rho^2\right)$ and $Corr\left(X_{i,a}, Y_{i,a}\right) = \rho$. Without loss of generality we may assume in the following that the X's and Y's are standardized with zero expectations and unit variances. Thus, expressing the relationship in (2) in terms of order statistics and corresponding concomitants gives

$$Y_{[r:a]} = \rho X_{r:a} + \varepsilon_{[r:a]}. \tag{3}$$

In the following the index of the error term is dropped for notational ease.

5.1 Swapping one variable

With two available variables X and Y we first consider swapping only X. Assuming that $\text{Rank}\left[X_{i,1}\right] = R$, the swapping procedure exchanges $X_{i,1}$, $i = 1, \ldots, n$, for $X_{j,2} = X_{R:2} = X_{i,1}^*$ for some j.

Theorem 1 *For a randomly chosen pair $(X_{i,1}^*, Y_{i,1}) \in \mathbf{M}_1^*$, the cross product moment is given by*

$$E\left[X_{i,1}^* Y_{i,1}\right] = \frac{1}{n}\sum_{r=1}^{n} \mu_{X:r} E\left\{E\left[Y_{i,1} \mid X_{i,1} = X_{r:1}\right]\right\}$$

which under the assumption of (2) equals

$$E\left[X_{i,1}^* Y_{i,1}\right] = \mu_X\mu_Y + \rho\sigma_X\sigma_Y\left(\frac{1}{n\sigma_X^2}\sum_{r=1}^{n}\mu_{X:r}^2 - \frac{\mu_X^2}{\sigma_X^2}\right) \tag{4}$$

where $\mu_{X:r}$ denotes the expectation of the r'th order statistic of X (in a sample of size n). Under the assumption of (3), (4) simplifies to

$$E\left[X^*_{i,1}Y_{i,1}\right] = \frac{\rho}{n}\sum_{r=1}^{n}\mu^2_{X:r} = \rho_S. \tag{5}$$

Proof. We are satisfied in proving (5). Since the samples $\mathbf{M}_1$ and $\mathbf{M}_2$ are independent, $X_{R:2}$ and $Y_{[R:1]}$ are conditionally independent given the rank R. By conditioning on R, which is a random variable, we have

$$E\left[X^*_{i,1}Y_{i,1}\right] = E_R\left\{E\left[X_{R:2}Y_{[R:1]} \mid R\right]\right\} = E_R\left\{E\left[X_{R:2} \mid R\right]E\left[Y_{[R:1]} \mid R\right]\right\}$$

Since $\Pr(R = r) = n^{-1}$ the result follows. From (3) we have

$$\begin{aligned} E\left[Y_{[r:1]}\right] &= E\left\{E\left[Y_{i,1} \mid X_{i,1} = X_{r:1}\right]\right\} = E\left\{E\left[\rho X_{i,1} + \varepsilon \mid X_{i,1} = X_{r:1}\right]\right\} \\ &= E\left\{\rho X_{r:1}\right\} = \rho\mu_{X:r} \end{aligned}$$

and the result in (5) follows. Note that it is implied that the $\varepsilon \in \mathbf{M}_1$ are independent of the $X_{j,2} \in \mathbf{M}_2$. ■

The result above is of course valid only if the expectations $\mu_{X:r}$ exist for all r. However, this follows from the assumption of a finite first moment in the parent distribution of X (cf. David, 1981, pp. 32-33). Given that the variance of the parent distribution of X equals unity it is also concluded that

$$0 < \frac{1}{n}\sum_{r=1}^{n}\mu^2_{X:r} = \frac{1}{n}\sum_{r=1}^{n}\left(E\left[X^2_{r:a}\right] - \sigma^2_{X:r}\right) = 1 - \frac{1}{n}\sum_{r=1}^{n}\sigma^2_{X:r} < 1$$

where $\sigma^2_{X:r}$ denotes the variance of the r'th order statistic of X in a sample of size n. This provides a naive bound for the expected influence of the operation, i.e. $0 < \rho_S < \rho$. As the sum does not depend on ρ, the expected relative deterioration of the association induced by the scheme is constant for all ρ given the sample size n; that is, the absolute deterioration increases with the absolute magnitude of ρ and decreases with sample size. It should also be clear that the limit of (5) as $n \to \infty$ equals ρ; a proof is outlined in the appendix.

Although the calculation of (5) is straightforward it might be somewhat cumbersome in practice since it requires the computation of all n expectations of the order statistics. Saw and Chow (1966) showed that

$$\frac{1}{n}\sum_{r=1}^{n}\mu^2_{X:r} = \sum_{r=0}^{n-1}\frac{n!\,(n-1)!}{(n+r)!\,(n-1-r)!}\,(2r+1)\,J_r^2 \tag{6}$$

Table 1: The expected degree of deterioration (eq. 4) for various distributions and sample sizes. Values were calculated using the approximation in (6) with * $m = 23$ and ** $m = 14$; all other values were caculated using the sum of squared expectations in (4).

n	5	10	30	60	100	300	1000
$N(0,1)$	0.6390	0.7914	0.9186	0.9563	0.9726	0.9901^{*}	0.9968^{*}
$Unif(-\frac{1}{2},\frac{1}{2})$	0.6667	0.8182	0.9355	0.9672	0.9802	0.9934	0.9980
$Exp(1)$	0.5433	0.7071	0.8668	0.9220	0.9481	0.9791	0.9925
$LogN(0,1)$	0.3707	0.5157	0.6982	0.7825	0.8302^{**}	-	-

where

$$J_r = \int_{-\infty}^{+\infty} L_r\left[2F_X(x) - 1\right]\ x f_X(x)\ dx$$

where $L_r(z)$ is the r'th order Legendre polynomial in z. This representation can for some parent distributions efficiently reduce the computational burden. Saw and Chow also suggested a truncated version of the left hand side of (6) as an approximation, i.e. the summation is carried out only up to a number $m \ll n-1$.

Note that the expression inside the brackets of equation (4) can be seen as a measure of the expected deterioration of association between X and Y; when it is close to unity $E[X^{*}Y]$ will be close to $E[XY]$. Table 1 gives the values of this expression for a selection of sample sizes under four different distributions of X; a standardized normal, a uniform, an exponential distribution, and a lognormal distribution. The truncated version of (6) was used in three cases as indicated, using $m = 23$ for the $N(0,1)$ and $m = 14$ for $LogN(0,1)$. The absolute error due to the approximation (cf. Saw and Chow, 1966) is less than 10^{-4} for the normal distribution but somewhat larger for the lognormal, less than 5×10^{-3}. Furthermore, for the lognormal distribution the error of approximation was considered to large for sample sizes larger than 100 and these values have been omitted. We conclude from the table that under a standardized normal distribution, the expected degree of deterioration is less than 1% if the sample size is at least 300 and for $n > 60$ it is less than 5%. The uniform distribution yields a slightly better preservation whereas the exponential is only slightly inferior. The lognormal distribution exhibits a larger degree of degradation which most likely is due to the positive skewness and heavy right tail.

5.2 Swapping both variables for values from one source

When swapping both $X_{i,1}$ and $Y_{i,1}$ we must consider both their respective ranks. Assuming first that $\text{Rank}\,[X_{i,1}] = R$, the value of $X_{i,1}$ is exchanged for the value of $X_{j,2} = X_{R:2} = X^*_{i,1}$ for some j. Secondly, assuming that $\text{Rank}\,[Y_{i,1}] = S$, the value of $Y_{i,1}$ is exchanged for the value of $Y_{k,2} = Y_{S:2} = Y^*_{i,1}$ for some k. Note that the ranks R and S are dependent random variables due to the linear link between $X_{i,1}$ and $Y_{i,1}$ defined in (2)-(3). We introduce the following notation for the involved probabilities:

$$\Pr(R = r) = \pi_r = n^{-1}, \qquad \Pr(R = r, S = s) = \pi_{rs}$$

and

$$\Pr(S = s \mid R = r) = \pi_{s|r}.$$

The joint probability π_{rs} can be obtained from the relation $\pi_{rs} = \pi_r \cdot \pi_{s|r} = n^{-1}\pi_{s|r}$. Exact expressions for the conditional probabilities were given by David et al. (1977).

Theorem 2 *For a randomly chosen pair $(X^*_{i,1}, Y^*_{i,1}) \in \mathbf{M}^*_1$, the cross product moment is given by*

$$E\left[X^*_{i,1}Y^*_{i,1}\right] = \frac{1}{n}\sum_{r=1}^{n}\sum_{s=1}^{n}\pi_{s|r}\left(\mu_{X:r}\mu_{Y:s} + \sigma_{X:r,Y:s}\right) = \rho_{D1} \tag{7}$$

where $\sigma_{X:r,Y:s}$ denotes the covariance of the r'th order statistic of X and the s'th order statistic of Y, in a sample of size n.

Proof. By conditioning on the ranks R and S we have

$$E\left[X^*_{i,1}Y^*_{i,1}\right] = E_{R,S}\left\{E\left[X_{R:2}Y_{S:2} \mid R, S\right]\right\} = E_{R,S}\left\{\mu_{X:R}\mu_{Y:S} + \sigma_{X:R,Y:S}\right\}.$$

Using that $\pi_{rs} = n^{-1}\pi_{s|r}$ the result follows. ■

The product moment ρ_{D1} depends on ρ through both the covariance terms and the probabilities $\pi_{s|r}$. A reasonable conjecture under assumption (3) which however remains to be proven is that the limit of (7) as $n \to \infty$ equals ρ.

Although exact expressions for the conditional probabilities $\pi_{s|r}$ are available, the calculation of (7) is even more cumbersome compared to (5), even for moderate sample sizes. Some simplifications can possibly be attained by

using various recurrence relations (cf. David et al., 1977, David, 1981, pp. 46-49 and Lin, 1989). Also, the covariance term is quite involved in the general case (cf. David, 1981, pp. 25-26). It is however instructive at this stage to study the behavior of (7) under the assumption of (3) at the special cases when $\rho = 0$ or ± 1. The conditional probabilities in these cases are

$$\begin{array}{lll} \rho = 0 & \Longrightarrow & \pi_{s|r} = n^{-1}, \ \forall r, s \\ \rho = 1 & \Longrightarrow & \pi_{s|r} = \begin{cases} 1, & \text{for } r = s \\ 0, & \text{otherwise} \end{cases} \\ \rho = -1 & \Longrightarrow & \pi_{s|r} = \begin{cases} 1, & \text{for } r = n - s + 1 \\ 0, & \text{otherwise} \end{cases} \end{array} . \tag{8}$$

Thus, for $\rho = 0$, and keeping the assumption of zero expectations and unit variances in the parent distributions in mind, (7) reduces to

$$\rho_{D1} = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} \left(\mu_{X:r} \mu_{Y:s} + \sigma_{X:r,Y:s} \right) = \mu_X \mu_Y = 0$$

since $\sigma_{X:r,Y:s} = 0$ when X and Y are independent variates (cf. David, 1981, pp. 25-26). For $\rho = 1$

$$\rho_{D1} = \frac{1}{n} \sum_{r=1}^{n} \left(\mu_{X:r}^2 + \sigma_{X:r}^2 \right) = \frac{1}{n} \sum_{r=1}^{n} E\left[X_{r:a}^2 \right] = E\left[X^2 \right] = \mu_X^2 + \sigma_X^2 = 1$$

and for $\rho = -1$ the result is analogous. The latter result is intuitively reassuring; the swapping procedure can be seen as the imposing of the rank structure of $\mathbf{M}_1$ onto the observed values of $\mathbf{M}_2$ and in the case of $\rho = \pm 1$ the rank structures of the two sets are identical with probability equal to one as indicated by the conditional probabilities in (8).

5.3 Swapping both variables with values from separate sources

Here the reasoning is the same as in the preceding section with regard to the ranks R and S and their respective probabilities, unconditional and conditional. However, the $Y_{i,1}$ are now exchanged for $Y_{k,3} = Y_{S:3} = Y_{i,1}^{**}$ for some k instead of $Y_{S:2}$.

Theorem 3 *For a randomly chosen pair* $(X_{i,1}^{*}, Y_{i,1}^{**}) \in \mathbf{M}_1^{*}$*, the cross product moment is given by*

$$E\left[X_{i,1}^{*} Y_{i,1}^{**} \right] = \frac{1}{n} \sum_{r=1}^{n} \sum_{s=1}^{n} \pi_{s|r} \mu_{X:r} \mu_{Y:s} = \rho_{D2}. \tag{9}$$

Proof. Since the samples $\mathbf{M}_2$ and $\mathbf{M}_3$ are independent, $X_{R:2}$ and $Y_{S:3}$ are conditionally independent, given the ranks R and S. Thus, by conditioning on R and S we have

$$\begin{aligned} E\left[X^*_{i,1}Y^{**}_{i,1}\right] &= E_{R,S}\left\{E\left[X_{R:2}Y_{S:3} \mid R,S\right]\right\} \\ &= E_{R,S}\left\{E\left[X_{R:2} \mid R,S\right]E\left[Y_{S:3} \mid R,S\right]\right\} = E_{R,S}\left\{\mu_{X:R}\mu_{Y:S}\right\}. \end{aligned}$$

Using that $\pi_{rs} = n^{-1}\pi_{s|r}$ the result follows. ■

Comparing this result with (7) we first note that $|\rho_{D2}| \leqslant |\rho_{D1}|$ with equality if and only if $\rho = 0$. Reasonable conjectures that remain to be proven seem to be that $|\rho_{D2}| \leqslant |\rho_S|$ for all ρ and that the limit of (9) equals ρ as $n \to \infty$. Using the probabilities in (8) we find that for $\rho = 0$, (9) reduces to

$$\rho_{D2} = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}\mu_{X:r}\mu_{Y:s} = \mu_X\mu_Y = 0$$

and for $\rho = 1$

$$\rho_{D2} = \frac{1}{n}\sum_{r=1}^{n}\mu_{X:r}\mu_{Y:r} = \frac{1}{n}\sum_{r=1}^{n}\mu^2_{X:r}$$

under the assumption of zero expectations and unit variances in the parent distribution. For $\rho = -1$ the result is analogous. We note that when the correlation is at its extremes ± 1, the operation is evidently equivalent to swapping only one variable. This is reasonable since the rank structure of the two variables in both auxiliary sets and the reference set are identical with probability equal to one. Thus swapping $\mathbf{X}_1$ for $\mathbf{X}_2$ and $\mathbf{Y}_1$ for $\mathbf{Y}_3$ would be equivalent to using $\mathbf{M}_2$ (or $\mathbf{M}_3$) as reference set and swapping only the Y's (or X's).

5.4 Simulation study under a bivariate normal distribution

In order to provide an idea of the degree of expected deterioration on the association between pairs of variables, a simulation study was conducted under a bivariate normal distribution using the usual sample correlation coefficient as the measure of association. The study comprised sample sizes of $n = 30$, 100, 300 and 1000 and correlation coefficients $\rho = 0$, 0.1, $0.2, \ldots, 0.9$ and 0.95, 0.99 and 1.0, giving a total of 52 configurations with respect to sample size and correlation. The simulation study was carried out as follows:

1. Three random samples $\mathbf{M}_a = [\mathbf{X}_a, \mathbf{Y}_a]$, $a = 1, 2, 3$, were generated from a bivariate normal distribution with zero means and unit variances.

2. The sample correlation coefficient for the set $\mathbf{M}_1$, denoted r_U, was calculated.

3. $\mathbf{X}_1$ was swapped for the values in $\mathbf{X}_2$. The sample correlation coefficient for this set, denoted r_S, was calculated.

4. Both $\mathbf{X}_1$ and $\mathbf{Y}_1$ were exchanged for the values in $\mathbf{X}_2$ and $\mathbf{Y}_2$ respectively. The correlation coefficient for this set, denoted r_{D1}, was calculated.

5. Both $\mathbf{X}_1$ and $\mathbf{Y}_1$ were exchanged for the values in $\mathbf{X}_2$ and $\mathbf{Y}_3$ respectively. The correlation coefficient for this set, denoted r_{D2}, was calculated.

6. The values of r_U, r_S, r_{D1} and r_{D2} were subsequently stored in a file and the procedure was repeated $B = 50,000$ times for each combination of n and ρ.

The bias and standard error of each of the four resulting sample correlations were estimated by

$$Bias\left[r_{\mathcal{X}}\right] = \tfrac{1}{B}\sum_{i=1}^{B} r_{\mathcal{X},i} - \rho = \bar{r}_{\mathcal{X}} - \rho$$

and

$$SE\left[r_{\mathcal{X}}\right] = \sqrt{\tfrac{1}{B-1}\sum_{i=1}^{B}\left(r_{\mathcal{X},i} - \bar{r}_{\mathcal{X}}\right)^2}$$

respectively, for $\mathcal{X} = U$, S, $D1$ and $D2$. However, the ordinary sample correlation coefficient is a biased estimator of ρ. Since we are only interested in estimating the level of additional bias introduced by the swapping procedures, the estimates for the three swapped cases were adjusted by subtracting the estimated bias of r_U, i.e.

$$AddBias\left[r_{\mathcal{X}}\right] = Bias\left[r_{\mathcal{X}}\right] - Bias\left[r_U\right]$$

for $\mathcal{X} = S$, $D1$ and $D2$. The standard error of the additional bias is accordingly estimated by

$$SE_{AddBias[r_{\mathcal{X}}]} = \sqrt{SE\left[r_{\mathcal{X}}\right]^2 + SE\left[r_U\right]^2 - 2Cov\left[r_{\mathcal{X}}, r_U\right]}.$$

The estimated (additional) bias and standard error of the sample correlation coefficients are given in tables 3 - 4 together with the correlation between sample correlation coefficients, i.e. $Corr\left[r_{\mathcal{X}}, r_U\right]$. The additional bias introduced by the respective schemes is illustrated in figure 2 in order to provide a graphical presentation of the results in the tables. The figure also shows the estimated bias of r_U which was used to adjust the estimates. The simulation standard error is approximately given by $SE/\sqrt{B} \approx SE \cdot 0.0045$ and we conclude that the deviations of the bias estimates from zero at $\rho = 0$ are accounted for by simulation error; the true bias is of course exactly 0 at $\rho = 0$.

The general conclusion is that the performance with respect to retaining the original correlation ranges from reasonable to good, depending primarily on sample size. The expected behavior of the swapping schemes as discussed earlier seems to be confirmed. Swapping one variable results in an expected relative deterioration that is constant for all ρ given the sample size n. Swapping both variables for values from the same source adds to the level of expected deterioration. However, as ρ approaches unity, the effect of the swapping decreases. Finally, swapping both variables for values from separate sources resembles swapping from the same source but as ρ approaches unity the effect is equivalent to swapping only one variable. It should be noted that there is a slight discrepancy in the results of the simulation study when compared to the theoretically derived values in table 1. This is explained by the sample correlation coefficient which standardizes the observed covariance with the standard deviations of the variables involved. Hence, when a swap has taken place the corresponding observations in the denominator are changed as well which will adjust the level of deterioration to the better.

The effects of the data-swapping operations are also illustrated in figure 3 where the sample correlations of the first 500 unswapped samples are plotted against the resulting sample correlations after the respective swapping schemes were applied. The figure shows the results for sample size $n = 100$ and for a selection of the values of ρ. A diagonal line was imposed in each plot for reference. It is clear in particular for larger correlations, that a negative bias is introduced. Still, the association between the original and masked sets is obvious and indicates a reasonable degree of preservation of the association between X and Y. This is indicated further by the strong correlation between pre-swap and post-swap sample correlation coefficients as seen in tables 3 - 4. We note further that all three swap-schemes seem to produce similar results when the correlation is small or moderately large as illustrated by the cases $\rho = 0$ and 0.5. When viewing the spread conditional on the original sample correlation, it is seen that it is slightly smaller when

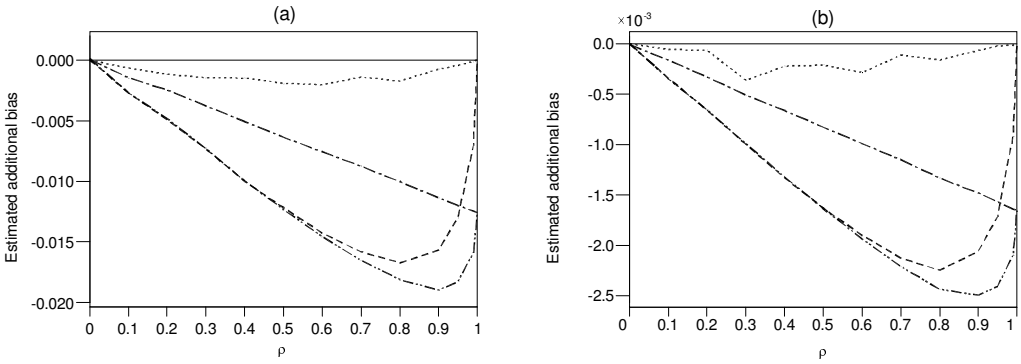

Figure 2: Estimated additional bias of the sample correlation coefficient introduced by the swapping procedures as a function of ρ under a standardized bivariate normal distribution; one variable swapped $r_S - r_U$, (dash - 1 dot), both swapped, same source, $r_{D1} - r_U$, (dashed) and both swapped, different sources, $r_{D2} - r_U$, (dash - 2 dots); estimated bias of the sample correlation coefficiet, unswapped, r_U, (dotted line). For sample sizes (a) $n = 100$ and (b) $n = 1000$; based on 50,000 simulations.

only X is swapped. By swapping both X and Y the spread tends to increase. Note however that the spread will decrease when both variables are swapped using the same auxiliary source.

6 Univariate Properties

A desirable property of any disclosure limiting technique is the ability to preserve univariate statistics such as means and variances. Although the original univariate sample moments of the reference set will not be retained, the swapping procedure generates at least equivalent samples of the same size. This means that any inference on single variables will be just as valid as if the analyst had access to the original data; we are simply exchanging one sample for an other. It is however evident that the procedure is equivalent to adding some kind of error or noise δ to the individual scores, i.e.

$$X_{j,2} = X^*_{i,1} = X_{i,1} + \delta_{i,1}$$

and it is essential to investigate the properties of this error. More specifically we are interested in finding the distribution of the error added to the original value after substitution, conditional on the original value and also conditional

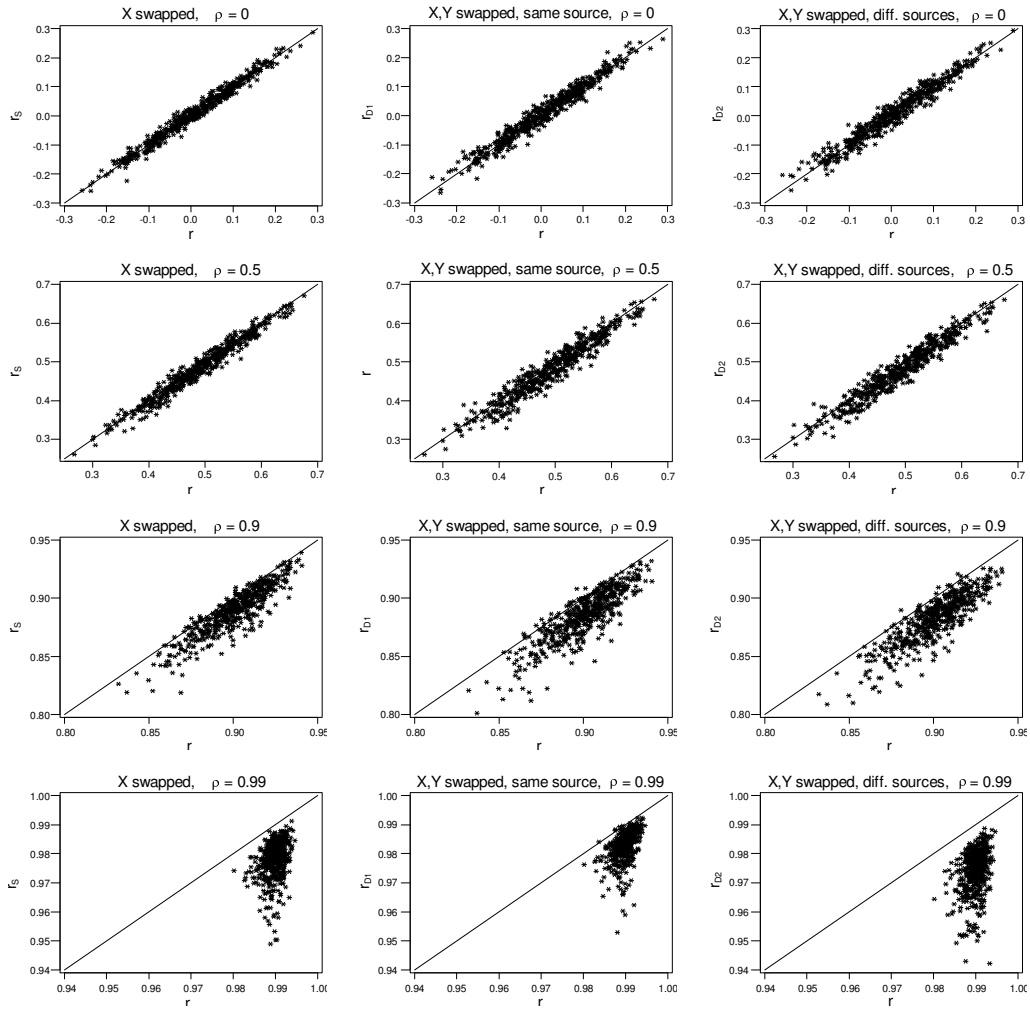

Figure 3: The effect of the swapping procedures on the sample correlation coefficient; original correlation r versus the correlation after swapping, r_S, r_{D1} and r_{D2} respectively; for sample size $n = 100$ and $\rho = 0$, 0.5, 0.9 and 0.99; first 500 simulated cases in each plot.

on the new swapped value the, i.e.

$$f_{\delta_{i,1}|X_{i,1}}(d \mid x) \qquad \text{and} \qquad f_{\delta_{i,1}|X^*_{i,1}}(d \mid x).$$

This is of interest when the variable is studied under different categories or sub-domains but also for assessing the effect on disclosure limitation, i.e the degree of added perturbation.

Note that if the swap scheme is extended to a two-way swap between $\mathbf{X}_1$ and $\mathbf{X}_2$ and the resulting sets $\mathbf{X}^*_1$ and $\mathbf{X}^*_2$ merged into one set, the univariate characteristics would remain intact since all of the original values are retained. Furthermore, the set of errors associated with $\mathbf{X}^*_2$ would be the same as those of X^*_1 with only a change in sign, i.e. $\delta_{j,2} = -\delta_{i,1}$ for some j. Thus, the errors in the merged set will sum to zero.

In the following the index of the variables concerned is dropped for notational ease; we simply denote the original variable by X, its substitute by X^* and the added error by δ.

6.1 The distribution of the added error

Since the two variables X and X^* originate from the same distribution we have that $f_X(x) = f_{X^*}(x)$. Furthermore, the exchange of X for X^* implies that they have the same rank R with respect to their respective sets $\mathbf{M}_1$ and $\mathbf{M}_2$. The unconditional distribution of R is as earlier noted given by $f_R(r) = n^{-1}$ and standard theory of order statistics yields

$$f_{R|X}(r \mid x) = f_{R|X^*}(r \mid x) = \tbinom{n-1}{r-1} F_x^{r-1} (1 - F_x)^{n-r}$$

and

$$f_{X|R}(x \mid r) = f_{X^*|R}(x \mid r) = n\tbinom{n-1}{r-1} F_x^{r-1} (1 - F_x)^{n-r} f_X(x)$$

where F_x is a shorter notation for $F_X(x)$. Thus, we have the following result:

Theorem 4 *The conditional distribution of the added error δ, given the value of the original value X, is*

$$\begin{aligned} & f_{\delta|X}(d \mid x) \\ = \; & n f_X(x+d) \sum_{r=1}^{n} \tbinom{n-1}{r-1}^2 F_{x+d}^{r-1} (1 - F_{x+d})^{n-r} F_x^{r-1} (1 - F_x)^{n-r}. \end{aligned} \quad (10)$$

Proof. Using that X and X^* are independent conditional on the rank R, since they originate from two independent samples, we have that

$$f_{X,X^*,R}(x, x^*, r) = f_R(r) f_{X|R}(x^* \mid r) f_{X|R}(x \mid r)$$

from which the joint distribution of X and X^* follows as

$$\begin{aligned} & f_{X,X^*}(x,x^*) \\ = \; & nf_X(x^*) f_X(x) \sum_{r=1}^{n} \binom{n-1}{r-1}^2 F_{x^*}^{r-1} (1-F_{x^*})^{n-r} F_x^{r-1} (1-F_x)^{n-r}. \end{aligned} \tag{11}$$

After transformation and conditioning on X, (10) follows. ■

The result is straightforward however impractical, especially for larger sample sizes. Conditional means and variances of (10) under a standardized normal distribution were however evaluated numerically together with the 10th and 90th percentiles for different n and for a selection of values on X. The results are presented in table 5; due to the symmetry in the joint distribution we are satisfied with reporting the values for $X \geqslant 0$; for negative X the results are analogous.

The basic findings are (a) the conditional mean is quite stable near the center but tends to curve off as X approaches the tails, (b) the conditional spread of the error increases as X increases in absolute magnitude but tends to decrease slightly when X is at its extreme; this is noticeable for $n = 30$ and 100, and (c) the conditional spread and curving of the conditional mean decrease with sample size. The properties are illustrated in figure 4 where the conditional means and percentiles are imposed over scatterplots of simulated cases (see the following subsection).

From (11) it is possible to derive the unconditional distribution of the error term which can be used to characterize the global properties of the perturbation. We are satisfied here by simply formulating the expectation and variance:

$$E[\delta] = E[X^* - X] = E[X^*] - E[X] = \mu_X - \mu_X = 0 \tag{12}$$

and

$$Var[\delta] = Var[X^*] + Var[X] - 2Cov[X^*, X] = 2\sigma_X^2 - 2E[X^*X] + \mu_X^2.$$

The cross product moment $E[X^*X]$ above is already given in (5) by setting $\rho = 1$. Thus, under the assumption of zero expectations and unit variances for the X's, the variance is given by

$$Var[\delta] = 2\left(1 - \frac{1}{n}\sum_{r=1}^{n} \mu_{X:r}^2\right). \tag{13}$$

6.2 Simulation study under a univariate normal distribution

In order to examine the expected behavior of the added error we conducted a simple simulation study. The simulation study was carried out for sample sizes $n = 30, 100, 300$ and 1000 and was performed as follows:

1. Two independent and equally sized univariate samples $\mathbf{M}_1$ and $\mathbf{M}_2$ were generated from a normal distribution with zero mean and unit variance.

2. A rank r was randomly selected from the interval $[1, n]$ and the difference $d = x_{r:2} - x_{r:1}$ was calculated.

3. The values of r, $x_{r:1}$, $x_{r:2}$ and d were subsequently stored in a file and the procedure was repeated $B = 50,000$ times for each sample size.

The resulting joint distribution of the error term and the original value of X is depicted in figure 4 and shows scatterplots of the simulated δ's against the simulated X's, showing the first 2000 values for each sample size. The numerically evaluated means of (10) are imposed on the plots together with the 10th and 90th percentiles.

The findings reported in the preceding subsection are largely confirmed. It is interesting to note the "tilting" hourglass shape in the distribution. This behavior is explained by the possibility of observing extreme outliers in either of the two variables, X and X^*. If an outlier is observed in the X's, the swapping procedure will tend to pull it back by exchanging it for a more probable observation found in the X^*'s. If the opposite occurs, i.e. an outlying value in the substitute X^*, this would tend to push the original observation, which with high probability is located closer to the center, further away. This also explains the curving of the conditional mean near the tails and the "bump" observed in the corresponding percentiles.

To conclude this section we report the unconditional behavior of the error terms. In table 2 the values of the means and standard deviations of the error terms, both theoretically derived and simulated, are given for all four sample sizes. The theoretical values were derived using (12) and (13) and for the latter we used the values of the sum of squared expectations of the order statistics listed in table 1.

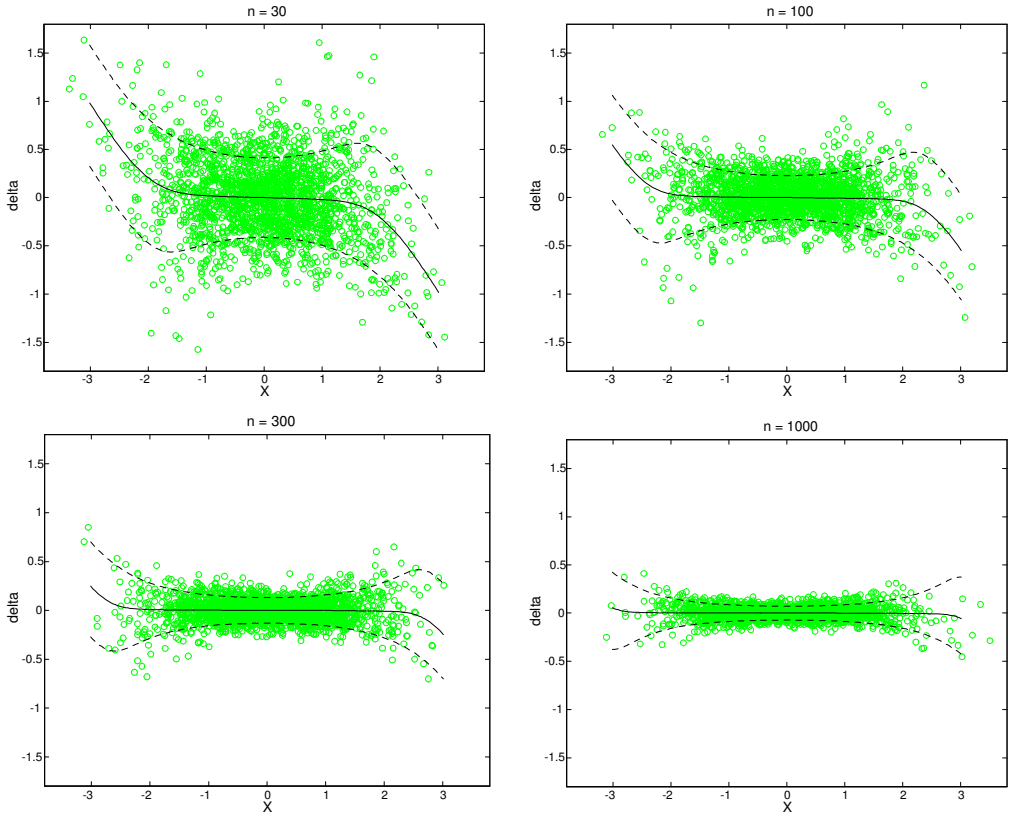

Figure 4: Relationship between the original values X and the added error δ under a standardized normal distribution for different sample sizes n. Scatterplot based on 2,000 simulated cases; conditional mean (solid line) and 10th and 90th percentiles (dashed lines) theoretically derived.

Table 2: Theoretical and simulated unconditional means and standard deviations of the added error δ for different sample sizes n; based on 50,000 simulations. (*) The values of the theoretical standard deviations for $n = 300$ and 1000 were calculated using the truncated version of (5) with $m = 23$.

	Theoretical		Simulated	
Sample size n	Mean	StDev	Mean	StDev
30	0.0	0.404	0.003	0.407
100	0.0	0.234	-0.000	0.234
300	0.0	0.141*	0.000	0.140
1000	0.0	0.080*	-0.000	0.079

7 The Effects of Data-Swapping

In the statistical analysis of a data set subjected to a masking technique such as data-swapping it is necessary to account for the amount of added perturbation. Thus, for an analyst to make inference about the parameters underlying the original data the data provider must also supply additional information so that the analyst can "statistically undo" the transformation and in doing so also include the additional uncertainty introduced. That is, given information about which variables that have been swapped, the sources of the swapped data (auxiliary sets), and the set size n used for the swap, an analyst should be able to adjust for the added bias and take into account the added uncertainty. Examples of this inferential approach are provided by e.g. Gouweleeuw et al. (1998) and Fienberg et al. (1998). See also the discussion by Rubin (1993).

As discussed in the preceding section the proposed swapping technique can be seen as a procedure where individual scores are perturbed by adding an error term. However, as already argued, any strictly univariate inference should not have to take the proposed swapping technique and the added error into account since the swapped data constitutes an equally valid sample compared to the original data; we have simply swapped one sample for another of the same size. This is even more evident after merging swapped subsets to the original set size since all original scores are still in the data. The procedure will however have an impact on any multivariate analysis. Typically associations between variables are weakened and the level of expected deterioration is seen to depend on the specific form of the assumed parent distribution and on the size of the subsets, as seen in e.g. table 1. For example, when the correlation coefficient is of interest, an approximate formula for the variance of the same is usually given by $Var\left[r_U\right] \approx n^{-1}(1-\rho^2)^2$. The added variability after swapping can be accounted for by adjusting the formula for the expected additional bias, i.e. $Var\left[r_S\right] \approx n^{-1}(1-(\rho+$ added bias$)^2)^2$, an approximation which holds for r up to about 0.95 based on the empirical results in tables 3-4. Other types of multivariate analysis such as means over domains will of course also require adjusting for bias and added variability. It should be noted that these results refer to expected degradation under specific models but that they still provide some insight in what to expect from applying the method. As pointed out by an anonymous referee, such asymptotic results are not always useful for an analyst in practical situations and further research is of course necessary. This is even more obvious when it comes to such statistical uses as model selection and estimation of casual effects.

The second property of the proposed method that must be evaluated be-

fore it can be used in practice, is its ability to limit the risk of disclosure. Many different conceptualizations of disclosure have been considered in the literature and a very useful typology for distinguishing the various conceptions is given in Duncan and Lambert (1989). Although large subsets were shown to retain associations between pairs of variables, the swap may not provide an acceptable level of perturbation and effectively mask the data since the changes on individual scores may not be large enough. This points to the problem of reaching a balance between the two conflicting aims: to minimize information loss and maximizing the reduction in disclosure risk.

With the method of data-swapping there are some special considerations. Assume for example that a record has been masked by exchanging the scores of all variables for new scores originating from one and the same other record. We then face two possibilities for a correct re-identification; linking to the original record or to the record from which the new scores originate. This lack of one-to-one mapping requires a redefinition of disclosure risk measures as discussed in e.g. in Dandekar et al. (2002). Another issue pertaining to disclosure limitation concerns extreme values since records containing such values on one or several attributes are more easily spotted in an identification process. In table 6 it is illustrated how the swapping procedure will result in large expected changes were it is most likely needed, in the tails of the distribution. E.g. an original value located at 2.5σ from the origin will change to something inside an approximate interval of $2\sigma \pm \sigma$ for $n = 30$ and to $2.3\sigma \pm 0.9\sigma$ for $n = 100$. It is also worth noting that the result in (10) can be used to infer the distribution of the original value given the swapped value simply by conditioning on the swapped value, i.e. $f_{X|X^*}(x \mid x^*)$. This shows that supplying the swapping scheme of the masked data provides additional information that can be used by a perceived intruder. A related issue concerns unique record, i.e. records with a unique combination of scores on the attributes (see Bethlehem et al. 1990). Although changes are most likely to be large on extreme individual scores there is no investigation of what will happen to combinations of extreme scores. A unique record may very well still be unique after the swap and if it in addition is closest to the original by some metric there will still be a risk of correctly linking the record.

Potential data uses are very diverse and it is difficult, if at all possible, to identify them all at the time of release. However, the need to measure information loss in some generic way reflecting the harm done to the original data is still desirable. For this purpose Domingo-Ferrer and Torra (2001) developed a metric for comparing masking techniques where the information loss and reduction of disclosure risk are combined into an overall score. Information loss is by their definition defined as the discrepancy between the original

and the masked data as measured by the individual record scores and a set of statistics such as means and covariances. In their paper, and in Dandekar et al. (2002), the rank-proximity-swap of Greenberg (1987) and Moore (1996) is compared to other masking methods such as multivariate microaggregation and additive noise, using the metric of Domingo-Ferrer and Torra. Although the metric in itself is perhaps not very useful for an analyst, the results are still encouraging since they appear to favor data-swapping over other.

8 Comments

We conclude from the results that the proposed method is able to retain basic properties of an original data set at least fairly well. It should however once again be stressed that the ideas in this paper are tentative and have not been tested on real-life data. Before the methods can be used on a larger scale more research is needed and in this section we might hint at some areas.

It remains to be investigated how the methods fare when faced with data comprising a large number of variables and coming from populations with more complex underlying properties. As discussed and exemplified by Lambert (1993), many interesting analyses do not involve just the mean vector and variance-covariance matrix of the joint distribution and a future area of investigation would be to ascertain the degree of congruity with respect to higher order associations. It is of course possible to approximately recover the sample moments after the swapping procedure has taken place in order to preserve the variance-covariance structure of the original data, e.g. by the post-masking optimization procedure of Sebé et al. (2002). But one could ask oneself, as Citteur and Willenborg (1993) argue, if yielding the sample moments directly would not be a more efficient alternative since the other information in the file would not necessarily pertain to reality.

The simulation studies reported on in this paper have only considered a standardized bivariate normal distribution. In practice data is often highly skewed as in e.g. business and economic or social data and such cases should be investigated as well. The theoretical results of theorems 1-4 however rely only on the very simple assumptions of finite first and second order moments in the parent distributions and that the variables are linked by a linear regression relation. The specific form of the parent distribution is not specified and the results hold in the general case as well. In table 1 the expected effect on four different parent distributions are reported. The lognormal performs worst of the four and a probable reason is a heavier upper tail compared to e.g. the exponential, a property which will yield more "extreme" values which in turn will be subjected to a larger amount

of added variability. However, more research is needed before any general conclusions can be stated.

Practical issues have to be investigated and resolved. In applications it will be necessary to consider the number of available observations and the size n of the sample considered for release which controls the number k of possible subsets. A partitioning of the original set into subsets has to be done and the reference set and auxiliary sets have to be chosen. Furthermore, it has to be determined which variables to swap and for each variable which auxiliary set to swap values from. Obviously, careful consideration would have to be given to these latter steps since there will be a trade-off between the degree of expected deterioration of information and the risk of disclosing sensitive information. For instance it might be considered that certain combinations of variables constitute a potential risk if left intact. In such cases a subset of the variables might have to be swapped, leaving the rest unswapped or alternatively swapped for values from another source, in order to achieve an acceptable level of perturbation.

Other practical issues pertain to data comprising ties or to data measured on an ordinal scale. The reordering permutations in (1) are not uniquely determined in the presence of ties and it remains to be investigated how different but valid reordering permutations affect the data. This also touches upon the issue of top-bottom coded data values and also missing data. One approach is to solve the problem in the same manner as Moore (1996). That is, leave the records containing top- or bottom codes or missing values unswapped and proceed only with those for which actual values are provided.

9 Acknowledgment

The authors wish to thank Professor Daniel Thorburn, Dep. of Statistics, Stockholm University, for bringing this area of research to their attention and for many valuable suggestions. The authors also acknowledge the valuable comments and suggestions of an anonymous referee and the editors.

A Limit of ρ_S

An outline for a proof, proposed by Thorburn (2000), that the limit of (5) as $n \to \infty$ equals ρ, is given by the following argument. Define the random variable ξ by choosing any number $\epsilon > 0$ and by truncating the parent distribution of X at the single points $a < 0 < b$ so that

$$\xi = \max\{a, \min\{X, b\}\} \qquad \text{and} \qquad E\left[\xi^2\right] > 1 - \epsilon/2.$$

Through the truncation of the parent distribution and using the asymptotic normality of the order statistics (e.g. David, 1981, pp. 254-257) it can be shown that

$$\frac{1}{n}\sum\nolimits_{r=1}^{n}\mu_{\xi:r}^{2}\rightarrow E\left[\xi^{2}\right]>1-\epsilon/2$$

as $n\rightarrow\infty$. Choose n_0 such that

$$\frac{1}{n}\sum\nolimits_{r=1}^{n}\mu_{\xi:r}^{2}>1-\epsilon,\quad\forall n\geqslant n_{0}.$$

This gives

$$1\geqslant\frac{1}{n}\sum\nolimits_{r=1}^{n}\mu_{X:r}^{2}\geqslant\frac{1}{n}\sum\nolimits_{r=1}^{n}\mu_{\xi:r}^{2}>1-\epsilon.$$

By letting $\epsilon\rightarrow 0$ the proof is complete.

B Tables

Table 3: Estimated bias and standard error (SE) of r_U and $r_{\mathcal{X}} - r_U$ and correlation (Corr) between $r_{\mathcal{X}}$ and r_U for $\mathcal{X} = S$, $D1$ and $D2$ (see sec. 5.4); based on 50,000 simulations, all values except Corr are $\times 10^{-3}$. Sample size $n = 30, 100$.

$n = 30$											
	r_U		$r_S - r_U$			$r_{D1} - r_U$			$r_{D2} - r_U$		
ρ	Bias	SE	Bias	SE	Corr	Bias	SE	Corr	Bias	SE	Corr
0.0	-0.59	185.9	-0.01	48.0	0.967	-0.02	66.9	0.935	-0.21	67.0	0.935
0.1	-0.49	183.4	-3.14	47.8	0.966	-6.28	67.0	0.933	-6.33	67.0	0.933
0.2	-3.90	179.2	-6.85	47.0	0.966	-12.92	66.0	0.933	-13.06	65.9	0.933
0.3	-4.74	170.1	-9.57	46.0	0.964	-18.89	64.4	0.929	-18.87	64.7	0.928
0.4	-6.43	157.4	-13.01	45.1	0.960	-25.15	62.7	0.922	-25.26	62.8	0.922
0.5	-6.20	142.5	-16.32	42.5	0.956	-31.13	60.2	0.914	-31.44	59.8	0.915
0.6	-7.23	122.2	-19.87	40.4	0.947	-37.02	57.3	0.897	-37.59	56.2	0.900
0.7	-6.69	99.1	-23.11	37.1	0.933	-41.47	52.7	0.873	-43.02	51.6	0.877
0.8	-5.52	71.2	-26.42	32.9	0.902	-44.16	47.1	0.827	-47.52	45.1	0.836
0.9	-2.81	38.1	-29.78	27.3	0.804	-41.02	37.5	0.725	-49.50	35.6	0.732
0.95	-1.76	19.8	-31.54	23.7	0.629	-34.66	30.4	0.586	-48.15	28.9	0.584
0.99	-0.39	4.1	-32.91	20.0	0.187	-18.99	18.4	0.327	-42.13	21.7	0.240
1.0	0.00	0.0	-33.02	19.0	-	0.0	0.0	-	-33.19	19.1	-
$n = 100$											
0.0	-0.54	100.2	0.04	15.9	0.987	0.02	22.5	0.975	0.13	22.5	0.975
0.1	-0.65	99.6	-1.45	15.8	0.987	-2.69	22.3	0.975	-2.70	22.3	0.975
0.2	-1.17	96.8	-2.47	15.7	0.987	-4.91	22.3	0.974	-4.83	22.2	0.974
0.3	-1.44	91.3	-3.78	15.4	0.986	-7.35	21.6	0.972	-7.31	21.6	0.972
0.4	-1.49	85.1	-5.08	14.9	0.985	-9.98	21.0	0.970	-9.97	20.9	0.970
0.5	-1.90	75.5	-6.36	14.3	0.982	-12.18	20.2	0.965	-12.32	20.1	0.965
0.6	-2.03	64.7	-7.56	13.3	0.979	-14.26	18.8	0.959	-14.55	18.7	0.959
0.7	-1.39	51.9	-8.77	12.4	0.972	-15.81	17.5	0.946	-16.52	17.2	0.948
0.8	-1.72	36.9	-10.02	10.9	0.958	-16.71	15.2	0.923	-18.10	14.7	0.927
0.9	-0.75	19.4	-11.35	9.1	0.903	-15.68	11.8	0.858	-18.97	11.5	0.861
0.95	-0.43	10.0	-11.98	8.0	0.779	-12.99	8.9	0.765	-18.28	9.3	0.741
0.99	-0.10	2.0	-12.46	6.8	0.280	-7.00	4.7	0.470	-15.89	7.1	0.313
1.0	0.00	0.0	-12.58	6.5	-	0.00	0.0	-	-12.58	6.5	-

Table 4: Estimated bias and standard error (SE) of r_U and $r_{\mathcal{X}} - r_U$ and correlation (Corr) between $r_{\mathcal{X}}$ and r_U for $\mathcal{X} = S$, $D1$ and $D2$ (see section 5.4); based on 50,000 simulations, all values except Corr are $\times 10^{-3}$. Sample size $n = 300, 1000$.

$n = 300$											
	r_U		$r_S - r_U$			$r_{D1} - r_U$			$r_{D2} - r_U$		
ρ	Bias	SE	Bias	SE	Corr	Bias	SE	Corr	Bias	SE	Corr
0.0	-0.22	57.9	0.00	5.7	0.995	-0.11	8.1	0.990	-0.03	8.1	0.990
0.1	-0.18	57.0	-0.51	5.7	0.995	-1.01	8.0	0.990	-1.02	8.0	0.990
0.2	-0.25	55.7	-0.98	5.6	0.995	-1.92	8.0	0.990	-1.94	8.0	0.990
0.3	-0.57	52.6	-1.44	5.5	0.995	-2.84	7.7	0.989	-2.86	8.0	0.989
0.4	-0.44	48.5	-1.93	5.3	0.994	-3.84	7.5	0.988	-3.85	7.6	0.988
0.5	-0.47	43.6	-2.44	5.1	0.993	-4.80	7.2	0.986	-4.80	7.0	0.987
0.6	-0.56	37.1	-2.91	4.8	0.992	-5.54	6.8	0.983	-5.63	6.6	0.984
0.7	-0.77	29.5	-3.44	4.4	0.989	-6.23	6.2	0.978	-6.49	6.0	0.979
0.8	-0.55	21.0	-3.93	3.9	0.983	-6.58	5.4	0.968	-7.12	5.1	0.970
0.9	-0.32	11.0	-4.38	3.3	0.958	-6.03	4.1	0.938	-7.29	3.9	0.938
0.95	-0.16	5.7	-4.61	2.8	0.895	-5.02	3.0	0.890	-7.09	3.2	0.873
0.99	-0.02	1.2	-4.84	2.4	0.424	-2.70	1.5	0.646	-6.15	2.5	0.442
1.0	0.00	0.0	-4.86	2.3	-	0.00	0.0	-	-4.88	2.3	-
$n = 1000$											
0.0	-0.17	31.6	-0.01	1.8	0.998	0.00	2.6	0.997	0.00	2.6	0.997
0.1	-0.06	31.1	-0.16	1.8	0.998	-0.34	2.6	0.997	-0.35	2.6	0.997
0.2	-0.06	30.4	-0.33	1.8	0.998	-0.66	2.5	0.997	-0.66	2.5	0.997
0.3	-0.36	28.9	-0.51	1.7	0.998	-1.00	2.5	0.996	-0.99	2.5	0.996
0.4	-0.22	26.5	-0.66	1.7	0.998	-1.33	2.4	0.996	-1.32	2.4	0.996
0.5	-0.21	23.8	-0.83	1.6	0.998	-1.63	2.3	0.995	-1.63	2.3	0.995
0.6	-0.29	20.3	-0.99	1.5	0.997	-1.90	2.2	0.994	-1.93	2.1	0.995
0.7	-0.11	16.1	-1.15	1.4	0.996	-2.12	2.0	0.992	-2.21	1.9	0.993
0.8	-0.16	11.4	-1.33	1.2	0.994	-2.24	1.7	0.989	-2.43	1.7	0.989
0.9	-0.07	6.0	-1.48	1.0	0.986	-2.06	1.2	0.979	-2.49	1.3	0.979
0.95	-0.02	3.1	-1.57	0.9	0.960	-1.71	0.9	0.960	-2.40	1.0	0.951
0.99	-0.01	0.6	-1.64	0.8	0.636	-0.92	0.4	0.842	-2.10	0.8	0.637
1.0	0.00	0.0	-1.65	0.7	-	0.00	0.0	-	-1.65	0.7	-

Table 5: Theoretical conditional means, variances and 10th and 90th percentiles for the added error δ given X under a standardized normal distribution for different values of X and sample sizes n.

	$n = 30$				$n = 100$			
X	Mean	Var	10th	90th	Mean	Var	10th	90th
0.0	-0.000	0.105	-0.413	0.413	-0.000	0.032	-0.227	0.227
0.5	-0.008	0.116	-0.433	0.429	-0.002	0.034	-0.238	0.237
1.0	-0.020	0.155	-0.494	0.483	-0.006	0.046	-0.272	0.272
1.5	-0.057	0.225	-0.612	0.559	-0.012	0.076	-0.343	0.343
2.0	-0.206	0.263	-0.817	0.472	-0.037	0.145	-0.472	0.453
2.5	-0.532	0.256	-1.140	0.136	-0.187	0.190	-0.697	0.391
3.0	-0.974	0.249	-1.573	-0.317	-0.539	0.189	-1.053	0.035
	$n = 300$				$n = 1000$			
X	Mean	Var	10th	90th	Mean	Var	10th	90th
0.0	-0.000	0.010	-0.131	0.131	-0.000	0.003	-0.072	0.072
0.5	-0.001	0.012	-0.137	0.137	-0.000	0.003	-0.075	0.075
1.0	-0.002	0.015	-0.158	0.158	-0.001	0.005	-0.086	0.086
1.5	-0.004	0.025	-0.200	0.201	-0.001	0.007	-0.110	0.111
2.0	-0.008	0.052	-0.280	0.286	-0.002	0.015	-0.156	0.159
2.5	-0.039	0.120	-0.430	0.410	-0.006	0.042	-0.246	0.256
3.0	-0.240	0.153	-0.695	0.278	-0.054	0.108	-0.420	0.376

References

[1] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990) Disclosure Control of Microdata. *Journal of the American Statistical Association*, **85**, pp. 38-45.

[2] Citteur, C.A.W. and Willenborg, L.C.R.J. (1993) Public Use Microdata Files: Current Practices at National Statistical Bureaus. *Journal of Official Statistics*, **9**, pp. 783-794.

[3] Dalenius, T. (1974) The Invasion of Privacy Problem and Statistics Production - an Overview. *Statistisk Tidskrift,* **3**, pp. 213-225.

[4] Dalenius, T. (1977) Towards a Methodology For Statistical Disclosure Control. *Statistisk Tidskrift,* **5**, pp. 429-444.

[5] Dalenius, T. (1979) Data-swapping; A Technique for Disclosure Control when Releasing Micro-Statistics. *Statistisk Tidskrift,* **4**, pp. 253-258.

[6] Dalenius, T. (1988) *Controlling invasion of privacy in surveys.* Monograph. Statistics Sweden, Stockholm.

[7] Dalenius, T. and Reiss, S.P. (1982) Data-swapping; A Technique for Disclosure Control. *Journal of Statistical Planning and Inference*, **6**, pp. 73-85.

[8] Dandekar , R.A., Domingo-Ferrer, J., and Sebé, F. (2002) LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection. In J. Domingo-Ferrer (Ed.) *Inference Control in Statistical Databases*, pp.153-162. LNCS 2316, Heidelberg: Springer-Verlag.

[9] David, H.A. (1981) *Order Statistics*, 2nd ed.. Monograph. John Wiley & Sons, New York.

[10] David, H.A., O'Connell, M.J. and Yang, S.S. (1977) Distribution and Expected Value of the Rank of a Concomitant of an Order Statistic. *The Annals of Statistics*, **5**, pp. 216-223.

[11] Domingo-Ferrer, J. (ed.) (2002) *Inference Control in Statistical Databases.* Monograph. Berlin: Springer.

[12] Domingo-Ferrer, J. and Torra, V. (2001) A Quantative Comparison of Disclosure Control Methods for Microdata. In P. Doyle, J. Lane, J. Theeuwes and L. Zayatz (Eds.) *Confidentiality, Disclosure and Data Access*, pp. 111-133. Amsterdam: North-Holland.

[13] Doyle, P., Lane, J., Theeuwes, J.J.M., and Zayatz, L. (eds.), (2001) *Confidentiality, Disclosure, and Data Access.* Monograph. Amsterdam: Elsevier.

[14] Duncan, G. and Lambert, D. (1989) The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics*, **7**, pp. 207-217.

[15] Duncan, G.T. and Pearson, R.W. (1991) Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future. With discussion. *Statistical Science*, **6**, pp. 219-239.

[16] Fienberg, S.E. (1994) Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality. *Journal of Official Statistics*, **10**, pp. 115-132.

[17] Fienberg, S.E., Makov, U.E. and Steel, R.J. (1998) Disclosure Limitation Using Perturbation and Related Methods for Categorical Data. With discussion. *Journal of Official Statistics*, **14**, pp. 485-502.

[18] Frank, O. (1976) Individual Disclosures from Frequency Tables. In T. Dalenius and A. Klevmarken (Eds.) *Personal Integrity and the Need for Data in the Social Sciences*. Swedish Council for Social Science Research, pp. 175-187.

[19] Frank, O. (1983) Statistical Disclosure Control. *Statistical Review*, **5**, *Essays in Honour of Tore E. Dalenius*, pp. 173-178.

[20] 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**, pp. 463-478.

[21] Greenberg, B. (1987) Rank Swapping for Masking Ordinal Microdata, US Bureau of the Census (unpublished manuscript).

[22] Lambert, D. (1993) Measures of Disclosure Risk and Harm. *Journal of Official Statistics*, **9**, pp. 313-331.

[23] Lin, G.D. (1989) The Product Moments of Order Statistics with Applications to Characterizations of Distributions. *Journal of Statistical Planning and Inference*, **21**, pp. 395-406.

[24] Moore, R.A. (1996) Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets. *Bureau of the Census, Statistical Research Division, Statistical Research Report Series,* No. RR96/04, US Bureau of the Census, Washington D.C..

[25] Paass, G. (1988) Disclosure Risk and Disclosure Avoidance for Microdata. *Journal of Business & Economic Statistics*, **6**, pp. 487-500.

[26] Reynolds, P.D. (1993) Privacy and Advances in Social and Policy Sciences; Balancing Present Costs and Future Gains. With discussion. *Journal of Official Statistics*, **9**, pp. 275-312.

[27] Rubin, D.B. (1993) Statistical Disclosure Limitation. Discussion on Jabine, T.B.. *Journal of Official Statistics*, **9**, pp. 461-468.

[28] Saw, J.G. and Chow, B. (1966) The Curve Through the Expected Values of Ordered Variates and the Sum of Squares of Normal Scores. *Biometrika*, **53**, pp. 252-255

[29] Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J.M., and Torra, R.A. (2002) Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in Masked Microdata Sets. In J. Domingo-Ferrer (Ed.) *Inference Control in Statistical Databases*, pp. 163-171. LNCS 2316, Heidelberg: Springer-Verlag.

[30] Spruill, N.L. (1983) The Confidentiality and Analytical Usefulness of Masked Business Microdata. In *Proceedings of the Section on Survey Research Methods*, pp. 602-607, American Statistical Association, Alexandria, VA.

[31] Statistics Sweden (1998) *Journal of Official Statistics*, Special issue on Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data, Vol. **14**, No. 4.

[32] Thorburn, D. (2000) Personal communication. Dep. of Statistics, Stockholm University.

[33] Willenborg, L. and de Waal, T. (1996) *Statistical Disclosure Control in Practice; Series: Lecture Notes in Statistics*, Vol. **111**. Monograph. Springer-Verlag, New York.

[34] Willenborg, L.C. and de Waal, T. (2000) *Elements of Statistical Disclosure Control; Series: Lecture Notes in Statistics*, Vol. **155**. Monograph. Springer-Verlag, New York.

[35] Yang, S.S. (1977) General Distribution Theory of the Concomitants of Order Statistics. *The Annals of Statistics*, **5**, pp. 996-1002.