DATA-SWAPPING — A TECHNIQUE FOR DISCLOSURE CONTROL
(Extended Abstract)

Tore Dalenius, Brown University[1]
and University of Stockholm

and

Steven P. Reiss, Brown University[2]

INTRODUCTION

In recent years there has been increasing concern about the confidentiality of computerized databases. This has led to a fast growing interest in the development of techniques for controlling the disclosure of information from such databases about individuals, both natural and legal persons.[3] Much of this interest has been focused on the release of statistical tabulations and microdata files. In this paper, we introduce a new technique, data-swapping, that addresses these interests. This technique can be used to both produce microdata and release statistical tabulations so that confidentiality is not violated.

The ideas presented in this report are tentative and have not been tested on real-life data. The illustrations provided in the text are based on small-scale numerical experiments, and serve to illustrate the use of data-swapping for disclosure control. Before the proposed techniques can be used on a large scale, more research will be needed. Directions for this research are pointed out throughout the report.

This report is organized in two parts. In Part I, we present the underlying idea of the technique. In Part II, we present some theory and methods.

I. THE UNDERLYING IDEA

1. DATA TO BE PROTECTED

Consider a set of N individuals. With each individual, we associate data observed with respect to V variables, $X, Y, \ldots, U$, some of which may be sensitive. We denote the data by $x, y, \ldots, u$, respectively, and assume that it is all categorical. Thus, each of the variables assumes values in the domain $\{0, 1, \ldots, (r-1)\}$ for some r. Here $0, 1, \ldots, (r-1)$ denote the categories into which the N individuals are classified. In the interest of clarity, much of our discussion will be carried out for the special case of dichotomous data, i.e. $r = 2$. We represent such a categorical database by an $N \times V$ matrix, the original data matrix m_o.

The basic idea of data-swapping is to use this matrix, m_o, as the "input" for producing another data matrix, m_e, which is then to be used as the basis for producing statistics by way of tabulations (including tabulations based on released microdata). We classify these statistics by means of their order. In particular, a t-order statistic is defined as follows:

i. if the tabulation involves data for one variable only, the statistic will be referred to as an 1-order statistic; the simplest case in kind is $\#(x = 0)$;

ii. if the tabulation involves data for two or more variables, the statistic will be referred to as a 2-order statistic, a 3-order statistic, etc., as the case may be; a simple example is $\#(x = 0, y = 0)$, which is a 2-order statistic.

In this paper we consider in detail only the cases $t = 1$ and $t = 2$. However, the techniques that are discussed can be used for any value of $t \leq V$.

In discussing disclosure control we are concerned with the possibility of a person inferring sensitive information from a database. More formally, we say that a value of some variable for some individual is compromisable if it can be uniquely determined from the information that is released. Here we assume that this individual can be identified from the values of its other variables, and that we are interested in determining the value of some sensitive variable. We extend this definition naturally to say that a database is compromisable if any of its values can be compromised.

2. THE RELATION BETWEEN m_o AND m_e

The statistics referred to in Sec. 1 are to be computed from m_e, a data matrix with data $x', y', \ldots, u'$ derived by subjecting m_o to data-swapping. This matrix m_e — while not identical with m_o — is, by specification, "t-order equivalent" with m_o, so that t-order statistics are preserved.

2.1 The Notion of "Equivalence"

We generally define "equivalence" in terms of the following invariance condition: m_o and m_e are "t-order equivalent" if they yield the same t-order statistics. For example, in the case of dichotomous data, the invariance condition may demand that $\#(x = 0) = \#(x' = 0)$ and the like for all other variables and their values; this would be 1-order equivalence.

2.2 The Signification of $m_o \neq m_e$

The matrix m_e differs from m_o with respect to the data for one or more variables. In most illustrations to be given here, we consider the case where m_e differs from m_o with respect to only data for one variable, which we choose throughout to be the X-variable. The fact that m_e differs from m_o is the very basis for the protection provided by our scheme for disclosure control. Thus, the fact that for some individual $x' = 0$ does not imply that $x = 0$.

3. THE DATA-SWAPPING TECHNIQUE

We describe the data-swapping technique by way of a simple example. Consider the following matrix m_o of data for $V = 4$ variables for $N = 10$ individuals:

Individual Number	Data x	y	z	u
1	0	1	1	1
2	0	1	0	1
3	0	0	0	0
4	0	1	1	1
5	0	0	1	1
6	1	0	1	1
7	1	1	1	1
8	1	1	1	0
9	1	0	0	1
10	1	0	0	1

Table 1

3.1 Swapping Data for One Variable Only

We swap data for the X-variable for $k = 4$ individuals, number 1 and three others, subject to the invariance condition that there be 2-order equivalence.

There are, in this specific case, five possible matrices m_e fulfilling the condition just stated, namely:

Matrix	x-Data Swapped for Individual Numbers			
m_{e1}	1	3	8	9
m_{e2}	1	3	8	10
m_{e3}	1	4	8	9
m_{e4}	1	4	8	10
m_{e5}	1	5	6	7

Table 2

More specifically, m_{e1} is given by:

Individual Number	Data x'	y'	z'	u'
1	1	1	1	1
2	0	1	0	1
3	1	0	0	0
4	0	1	1	1
5	0	0	1	1
6	1	0	1	1
7	1	1	1	1
8	0	1	1	0
9	0	0	0	1
10	1	0	0	1

Table 3

Clearly, this matrix yields 1- and 2-order statistics identical with the corresponding statistics computed from m_o. (To check that this is true, we have only to check statistics involving data for the X-variable.)

The matrix m_{e1} differs from m_o with respect to the x-data. We note, for example, that:

$$\#(x = 0,\ x' = 1) = 2$$

$$\#(x = 1,\ x' = 0) = 2$$

with $2 + 2 = k$.

It is worth noting that *no* uncertainty is introduced into the resulting 1-order and 2-order statistics as a consequence of using data-swapping. In this respect data-swapping compares favorably with several other techniques for disclosure control.

It remains to discuss the amount of control imposed. We discuss this in detail in Part II. At this stage we just point out that

$$\frac{\#(x = 0,\ x' = 0)}{\#(x = 0)} = .6$$

which, in a crude sense, may be looked upon as a measure of the degree of protection involved.

3.2 Swapping Data for Two or More Variables

So far we have considered only protecting d data for one variable. The technique illustrated can easily be extended to cases involving two or more variables. To protect data in both the X and Y variables, for example, we swap data for both. Starting with m_o, we swap data for the X-variable and get $m_e = m_{ex'}$. Next, starting with $m_{ex'}$, we swap data for the Y-variable and get $m_{ex'y'}$. It is immediately clear that $m_{ex'y'}$ satisfies the condition of 1- and 2-order equivalence and offers the necessary protection. By the same token the data-swapping may involve three or more variables.

It is interesting to note that we do not have to swap values for the same individuals, or even the same number of individuals, to generate $m_{ex'y'}$ as we used to generate $m_{ex'}$.

II. THEORY AND METHODS

4. MATHEMATICAL CONSIDERATIONS

So far we have illustrated a technique whereby data in a statistical database can be protected from compromise. The basic idea is that the value of a sensitive variable for a particular individual cannot be compromised if there are (at least) two distinct databases that are consistent with the underlying statistics and that assign different values to that variable. This notion was extended to a complete database by noting that a database is protected if and only if each of the sensitive values is protected. The purpose of this section is to provide a mathematical framework that supports the use of the technique of data-swapping to protect the data in a database. In particular, we show that in a database that has sufficiently many individuals relative to the number of variables, each sensitive value is involved in a large number of potential data-swaps. Moreover, we present conditions that will insure, with any probability $p < 1$, that every sensitive value in a database is protected.

4.1 Notation

We begin by casting our model of a database into slightly more mathematical terms. The database m_o is defined as an $N \times V$ matrix over the domain $\{0, 1, \ldots, (r-1)\}$, $r \geq 2$. Here N is the number of individuals and V is the number of variables. The case $r = 2$ is the simple 0-1 case from which our previous examples were drawn. We also define a set of parameters that describe the distribution of values within the database. These are labeled a_i, $i \geq 1$, and characterize the minimum fraction of the database that occurs as the count of an *ith* order-statistic. That is, every value of every variable must be associated with at least N/a_1 individuals. Similarly, every pair of values for every pair of variables must occur for at least N/a_2 individuals, and so on.

We define the notion of safety by means of alternative databases in terms of this model. A datum is called t-safe if there are two databases that have the same t-order statistics and that assign different values to that datum. A k-data-swap in this model is the process of changing the values for a single variable for a set of k individuals. A datum is called (k,t)-safe if there are two databases with the same t-order statistics that assign different values to the datum such that one database can be obtained from the other by a single k-data-swap. It is clear that if a datum is (k,t)-safe, then it must be t-safe as well. This is the basis of our study.

While the results in this section and this paper apply, with minor variations, to t-order statistics for any $t \geq 2$, only the results for $t = 2$ are derived and presented below.

4.2 The Mathematics of Data-Swapping

The technique of data-swapping can be used statically to demonstrate that a database released only as t-order statistics is well protected, and dynamically to allow microdata that is statistically accurate to be freely released. Both of these applications depend on the possibility of data-swaps in a particular database. In this section we present the basic results in this respect.

We first determine the probability of a random choice of k individuals being valid, i.e. 2-order statistic preserving, and the number of data-swaps involving k out of N individuals.

Lemma: The probability of a random choice of k individuals being a valid data-swap is

$$p \approx \frac{r^{(V-1)r}}{(\pi k)^{(V-1)(r-1)}}$$

Lemma: In a database of N individuals, the number of potential 2-data-swaps of k individuals that involve a fixed individual is at least

$$\binom{N/a_1}{k/2}^2 \binom{k/2-1}{N/a_1-1}$$

The number that do not involve a fixed individual is at least

$$\binom{N/a_1}{k/2}^2$$

Corollary: The number of ways of selecting a data-swap of k individuals that involve a fixed individual with $k << N/a_1$ is about

$$\frac{N^{k-1}\, e^k\, 2^{k-1}}{k^k\, a_1^{k-1}\, \pi}$$

5. SAFETY CONSIDERATION INVOLVING t-ORDER STATISTICS

In order to produce t-order statistics, we may, of course, use either m_o or m_e. Using m_e has the advantage of providing some protection against "accidental" disclosure, for example, in the course of the data processing operation.

The main role of data-swapping in the present context is, however, to show that a database presented only in terms of t-order statistics is unlikely to be compromised. To do this we need to show that every sensitive variable of every individual is almost certainly involved in at least one swap and hence cannot be determined.

We first note that the expected number of data-data-swaps can be large provided that N grows as a small polynomial in V. In particular,

THEOREM: If $V < N/a_2$, $V \geq 4$, and $N \geq \frac{1}{4} a_1 F^{1/(V-1)} V^{(Vr-r+1)/(V-1)}$, for some function F, then the expected number of data-swaps of $k = V$ individuals involving a fixed individual is $\geq F$.

Lemma: A database is 2-safe with probability p if

$$F \geq \log (5NV p^*)$$

where $p^* = \log(1-p)/\log(p)$ and F is the expected number of data-swaps of V individuals involving a fixed individual.

THEOREM: If $V < N/a_2$, and

$$\frac{N}{\log(5NVp^*)^{1/V-1}} \geq \frac{1}{4} a_1 V^{Vr-r+1/V-1}$$

where $p^* = \log(1-p)/\log(p)$. Then, with probability p, every value in the database is 2-safe.

This last statement shows that, under assumptions that are not restrictive and that are verifiable, one can be quite certain that a database presented solely as a set of t-order statistics is completely protected. Moreover, even if some value is not protected, the results in Reiss (1977) show that the problem of determining this value is extremely difficult. Finally, the above theorem can be contrasted with the other result in Reiss (1977) where it was shown that the problem of checking if such a database is protected is intractable.

6. SAFETY CONSIDERATIONS INVOLVING THE RELEASE OF MICRODATA

In the previous section we proved that a database presented only in terms of t-order statistics probably could not be compromised provided that the number of individuals is sufficiently large. In some applications it is desirable to provide microdata as a basis for statistical inquiry. If this microdata is derived directly from the actual database, then compromise is trivial. As illustrated in the earlier section of this paper, the method of data-swapping provides a viable alternative. The basic idea is to construct a new database that is equivalent to the original one in terms of t-order statistics, and where the new data is sufficiently different from the original so that compromise is not possible.

One way of constructing an alternative database is to start with the t-order statistics and find an arbitrary database that is consistent with them. In Reiss (1977) it was shown that this approach is not feasible. The alternative approach suggested here is to take the actual database and use data-swapping to change enough sensitive values so that compromise is not possible. In this section we show that such microdata can be safely released, and hence, that this is a viable alternative. We conclude by investigating the computational feasibility of such an approach.

6.1 The Information Content of the Microdata

We first show that we can control the amount of information provided by microdata. Our notation is the same as in the previous section.

Lemma: Consider a sensitive 0-1 variable containing $a\alpha$ zeros and $(N-\alpha)$ ones. Let

$$p = \begin{cases} \alpha/N & \text{if only } 0 \text{ is a sensitive value} \\ (N-\alpha)/N & \text{if only } 1 \text{ is a sensitive value} \\ \text{MAX}(\alpha/N,\ (N-\alpha)/N) & \text{otherwise.} \end{cases}$$

Then we can insure that no information about the original sensitive values can be determined from the values after a data-swap with probability greater than p.

From the 1-order statistics, we know that a value is zero with probability α/N and is 1 with probability $N-\alpha/N$. Hence we get

THEOREM: The resultant database after a $\alpha(N-\alpha)/N$ data-swap yields no information that is not implicit in the t-order statistics, $t \geq 1$.

6.2 Computational Considerations

The above theorem shows that microdata obtained from the data-swapping can be safely released. The question that remains is whether it is computationally feasible to find the necessary data-swap in the original database. While actually determining this data-swap seems to be an inherently difficult problem, we illustrate a technique that might, in limited circumstances, make it feasible.

We first note that the number of variables is more of a computational consideration than is the number of individuals in the database. This follows since a data-swap of k values can be constructed from a number of smaller data-swaps on disjoint portions of the database. Thus, we can swap $\alpha(N-\alpha)/N$ values as in the previous lemma by performing $\alpha(N-\alpha)/NV$ disjoint data-swaps of V values each. These disjoint sets can be chosen at random to enhance safety. Moreover, the choice of a set of V values insures, by the results of the previous sections, that such swaps very likely exist. Finally, we note that if $T(V)$ is the time required to find a data-swap of V values, we can find a swap of $\alpha(N-\alpha)/N$ values in $\alpha(N-\alpha)/NV$ $T(V) \leq N/V\ T(V)$ time.

This leaves the more difficult problem of finding a data-swap of V values. This can be done in a simple manner by randomly selecting V values from the original database that have not already been involved in a data-swap, and testing if the resultant data-swap is valid. The expected number of tests required here is the reciprocal of the probability of such a test succeeding, or

$$\approx \frac{(\pi V)^{(V-1)(r-1)}}{r^{(V-1)r}} .$$

The total amount of work necessary to compute a $\alpha(N-\alpha)/N$ data-swap is then on the order of

$$\frac{N}{V} \frac{(\pi V)^{(V-1)(r-1)}}{r^{(V-1)r}}$$

This result is significantly better than a naive approach, but does not represent a tractable method for realistic N and V. The problem of finding such a method, or of determining if one exists at all, is an interesting one for further research.

ACKNOWLEDGMENTS

The authors would like to acknowledge the helpful comments of the discussant, Dr. Zalkind, and the typing of K. Kalia and J. Follansbee.

REFERENCES

OFFICE OF FEDERAL STATISTICAL POLICY AND STANDARDS
Statistical Policy Working Paper 2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques, prepared by Subcommittee on Disclosure and Disclosure=Avoidance Techniques, Federal Committee on Statistical Methodology. Government Printing Office, Washington, DC, 1978.

REISS, S. P.
Statistical database confidentiality. Report No. 25 of the research project Confidentiality in Surveys.
Department of Statistics, University of Stockholm, Stockholm, 1977.

FOOTNOTES

[1]The research reported here has been supported by a grant from the Bank of Sweden Tercentenary Foundation.

[2]Program in Computer Science, Brown University, Providence, RI 02912.

[3]For a review of this development, see Office of Federal Statistical Policy and Standards (1978).