

Data Swapping: Variations on a Theme by Dalenius and Reiss

Stephen E. Fienberg^{1**} and Julie McIntyre²

¹ Department of Statistics

Center for Automated Learning and Discovery

Center for Computer Communications and Security

Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

fienberg@stat.cmu.edu

² Department of Statistics

Carnegie Mellon University, Pittsburgh PA 15213-3890, USA,

julie@stat.cmu.edu

Abstract. Data swapping, a term introduced in 1978 by Dalenius and Reiss for a new method of statistical disclosure protection in confidential data bases, has taken on new meanings and been linked to new statistical methodologies over the intervening twenty-five years. This paper revisits the original (1982) published version of the the Dalenius-Reiss data swapping paper and then traces the developments of statistical disclosure limitation methods that can be thought of as rooted in the original concept. The emphasis here, as in the original contribution, is on both disclosure protection and the release of statistically usable data bases.

Keywords: Bounds table cell entries; Constrained perturbation; Contingency tables; Marginal releases; Minimal sufficient statistics; Rank swapping.

1 Introduction

Data swapping was first proposed by Tore Dalenius and Steven Reiss (1978) as a method for preserving confidentiality in data sets that contain categorical variables. The basic idea behind the method is to transform a database by exchanging values of sensitive variables among individual records. Records are exchanged in such a way to maintain lower-order frequency counts or marginals. Such a transformation both protects confidentiality by introducing uncertainty about sensitive data values and maintains statistical inferences by preserving certain summary statistics of the data. In this paper, we examine the influence of data swapping on the growing field of statistical disclosure limitation.

Concerns over maintaining confidentiality in public-use data sets have increased since the introduction of data swapping, as has access to large, computerized databases. When Dalenius and Reiss first proposed data swapping, it was in many ways a unique approach the problem of providing quality data to users

** Currently Visiting Researcher at CREST, INSEE, Paris, France

while protecting the identities of subjects. At the time most of the approaches to disclosure protection had essentially no formal statistical content, e.g., see the 1978 report of the Federal Committee on Statistical Methodology, FCSM (1978), for which Dalenius served as a consultant.

Although the original procedure was little-used in practice, the basic idea and the formulation of the problem have had an undeniable influence on subsequent methods. Dalenius and Reiss were the first to cast disclosure limitation firmly as a statistical problem. Following Dalenius (1977), Dalenius and Reiss define disclosure limitation probabilistically. They argue that the release of data is justified if one can show that the probability of any individual's data being compromised is appropriately small. They also express a concern regarding the usefulness of data altered by disclosure limitation methods by focusing on the type and amount of distortion introduced in the data. By construction, data swapping preserves lower order marginal totals and thus has no impact on inferences that derive from these statistics.

The current literature on disclosure limitation is highly varied and combines the efforts of computer scientists, official statisticians, social scientists, and statisticians. The methodologies employed in practice are often ad hoc, and there are only a limited number of efforts to develop systematic and defensible approaches for disclosure limitation (e.g., see FCSM, 1994; and Doyle et al., 2001). Among our objectives here are the identification of connections and common elements among some of the prevailing methods and the provision of a critical discussion of their comparative effectiveness.³ What we discovered in the process of preparing this review was that many of those who describe data swapping as a disclosure limitation method either misunderstood the Dalenius-Reiss arguments or attempt to generalize them in directions inconsistent with their original presentation.

The paper is organized as follows. First, we examine the original proposal by Dalenius and Reiss for data swapping as a method for disclosure limitation, focusing on the formulation of the problem as a statistical one. Second, we examine the numerous variations and refinements of data swapping that have been suggested since its initial appearance. Third, we discuss a variety of model-based methods for statistical disclosure limitation and illustrate that these have basic connections to data swapping.

2 Overview of Data Swapping

Dalenius and Reiss originally presented data swapping as a method for disclosure limitation for databases containing categorical variables, i.e., for contingency tables. The method calls for swapping the values of sensitive variables among

³ The impetus for this review was a presentation delivered at a memorial session for Tore Dalenius at the 2003 Joint Statistical Meetings in San Francisco, California. Tore Dalenius made notable contributions to statistics in the areas of survey sampling and confidentiality. In addition to the papers we discuss here, we especially recommend Dalenius (1977, 1988) to the interested reader.

records in such a way that the t -order frequency counts, i.e., entries in the t -way marginal table, are preserved. Such a transformed database is said to be t -order *equivalent* to the original database.

The justification for data swapping rests on the existence of sufficient numbers of t -order equivalent databases to introduce uncertainty about the true values of sensitive variables. Dalenius and Reiss assert that any value of a sensitive variable is protected from compromise if there is at least one other database or table, t -order equivalent to the original one, that assigns it a different value. It follows that an entire database or contingency table is protected if the values of sensitive variables are protected for each individual. The following simple example demonstrates how data swaps can preserve second-order frequency counts.

Example: Table 1 contains data for three variables for seven individuals. Suppose variable X is sensitive and we cannot release the original data. In particular, notice that record number 5 is unique and is certainly at risk for disclosure from release of the three-way tabulated data. However, is it safe to release the two-way marginal tables?

Table 1b shows the table after a data-swapping transformation. Values of X were swapped between records 1 and 5 and between records 4 and 7. When we display the data in tabular form as in Table 2, we see that the two-way marginal tables have not changed from the original data. Summing over any dimension results in the same 2-way totals for the swapped data as for the original data. Thus, there are at least two data bases that could have generated the same set of two-way tables. The data for any single individual cannot be determined with certainty from the release of this information alone.

(a) Original Data				(b) Swapped Data			
Record	X	Y	Z	Record	X	Y	Z
1	0	1	0	1	1	1	0
2	0	1	0	2	0	1	0
3	0	0	1	3	0	0	1
4	0	0	1	4	1	0	1
5	1	1	1	5	0	1	1
6	1	0	0	6	1	0	0
7	1	0	0	7	0	0	0

Table 1. Swapping X values for two pairs of records in a 3-variable hypothetical example

An important distinction arises concerning the form in which data are released. Releasing the transformed data set as microdata clearly requires that enough data are swapped to introduce sufficient uncertainty about the true values of individuals' data. In simple cases such as the example in Table 1 above, appropriate data swaps, if they exist, can be identified by trial and error. However identifying such swaps in larger data sets is difficult. An alternative is to release the data in tabulated form. All marginal tables up to order t are unchanged by the transformation. Thus, tabulated data can be released by showing the existence of appropriate swaps without actually identifying them. Schlörer (1981)

(a) Original Data						(a) Swapped Data					
Z						Z					
0			1			0			1		
Y			Y			Y			Y		
X	0	1	X	0	1	X	0	1	X	0	1
0	0	2	0	2	0	0	1	1	0	1	1
1	2	0	1	0	1	1	1	1	1	1	0

Table 2. Tabular versions of original and swapped data from Table 1

discusses some the trade-offs between the two approaches and we return to this issue later in the context of extensions to data swapping.

Dalenius and Reiss developed a formal theoretical framework for data swapping upon which to evaluate its use as a method for protecting confidentiality. They focus primarily on the release of data in the form of 2-way marginal totals. They present theorems and proofs that seek to determine conditions on the number of individuals, variables, and the minimum cell counts under which data swapping can be used to justify the release of data in this form. They argue that release is justified by the existence of enough 2-order equivalent databases or tables to ensure that every value of every sensitive variable is protected with high probability.

In the next section we discuss some of the main theoretical results presented in the paper. Many of the details and proofs in the original text are unclear, and we do not attempt to verify or replace them. Most important for our discussion is the statistical formulation of the problem. It is the probabilistic concept of disclosure and the maintenance of certain statistical summaries that has proved influential in the field.

2.1 Theoretical Justification for Data Swapping

Consider a database in the form of an $N \times V$ matrix, where N is the number of individuals and V is the number of variables. Suppose that each of the V variables is categorical with $r \geq 2$ categories. Further define parameters a_i , $i \geq 1$, that describe lower bounds on the marginal counts. Specifically, $a_i = N/m_i$ where m_i is the minimum count in the i -way marginal table.

Dalenius and Reiss consider the release of tabulated data in the form of 2-way marginal tables. In their first result, they consider swapping values of a single variable among a random selection of k individuals. They then claim that the probability that the swap will result in a 2-equivalent database is

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

Observations:

1. The proof of this result assumes that only 1 variable is sensitive.

2. The proof also assumes that variables are independent. Their justification is: “each pair of categories will have a large overlap with respect to k .” But the specific form of independence is left vague. The 2-way margins for X are in fact the minimal sufficient statistics for the model of conditional independence of the other variables given X (for further details, see Bishop, Fienberg, and Holland, 1975).

Dalenius and Reiss go on to present results that quantify the number of potential swaps that involve k individuals. Conditions on V , N , and a_2 follow that ensure the safety of data released as 2-order statistics. However the role of k in the discussion of safety for tabulated data is unclear. First they let $k = V$ to get a bound on the expected number of data swaps. The first main result is:

Theorem 1. *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 possible data-swaps of $k = V$ individuals involving a fixed variable is $\geq F$.*

Unfortunately, no detail or explanation is given about the function F . Conditions on V , N , and a_2 that ensure the safety of data in 2-way marginal tables are stated in the following theorem:

Theorem 2. *If $V < N/a_2$, and*

$$\frac{N}{\{\log(5NVp^*)\}^{2/(V-1)}} \geq a_1 V^{(Vr-r+1)/(V-1)}$$

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

Observations:

1. The proof depends on the previous result that puts a lower bound on the expected number of data swaps involving $k = V$ individuals. Thus the result is not about releasing all 2-way marginal tables but only those involving a specific variable, e.g., X .
2. The lower bound is a function F , but no discussion of F is provided.

In reading this part of the paper and examining the key results, we noted that Dalenius and Reiss do not actually swap data. They only *ask about* possible data swaps. Their sole purpose appears to have been to provide a framework for evaluating the likelihood of disclosure.

In part, the reason for focusing on the release of tabulated data is that identifying suitable data swaps in large databases is difficult. Dalenius and Reiss do address the use of data swapping for release of microdata involving non-categorical data. Here, it is clear that a database must be transformed by swapping before it can safely be released; however, the problem of identifying enough swaps to protect every value in the data base turns out to be computationally impractical. A compromise, wherein data swapping is performed so that t -order frequency

counts are *approximately* preserved, is suggested as a more feasible approach. Reiss (1984) gives this problem extensive treatment and we discuss it in more detail in the next section.

We need to emphasize that we have been unable to verify the theoretical results presented in the paper, although they appear to be more specialized than the exposition suggests, e.g., being based on a subset of 2-way marginals and not on all 2-way marginals. This should not be surprising to those familiar with the theory of log-linear models for contingency tables, since the cell probabilities for the no 2nd-order interaction model involving the 2-way margins does not have an explicit functional representation (e.g., see Bishop, Fienberg, and Holland, 1975). For similar reasons the extension of these results to orders greater than 2 is far from straightforward, and may involve only marginals that specify decomposable log-linear models (c.f., Dobra and Fienberg, 2000).

Nevertheless, we find much in the authors' formulation of the disclosure limitation problem that is important and interesting, and that has proved influential in later theoretical developments. We summarize these below.

1. The concept of disclosure is probabilistic and not absolute:
 - (a) Data release should be based on an assessment of the probability of the occurrence of a disclosure, c.f., Dalenius (1977).
 - (b) Implicit in this conception is the trade-off between protection and utility. Dalenius also discusses this in his 1988 Statistics Sweden monograph. He notes that essentially there can be no release of information without some possibility of disclosure. It is in fact the responsibility of data managers to weigh the risks. Subjects/respondents providing data must also understand this concept of confidentiality.
 - (c) Recent approaches rely on this trade-off notion, e.g., see Duncan, et al. (2001) and the Risk-Utility frontiers in NISS web-data-swapping work (Gomatam, Karr, and Sanil, 2004).
2. Data utility is defined statistically:
 - (a) The requirement to maintain a set of marginal totals places the emphasis on statistical utility by preserving certain types of inferences. Although Dalenius and Reiss do not mention log-linear models, they are clearly focused on inferences that rely on t -way and lower order marginal totals. They appear to have been the first to make this a clear priority.
 - (b) The preservation of certain summary statistics (at least approximately) is a common feature among disclosure limitation techniques, although until recently there was little reference to the role these statistics have for inferences with regard to classes of statistical models.

We next discuss some of the immediate extensions by Dalenius and Reiss to their original data swapping formulation and its principal initial application. Then we turn to what others have done with their ideas.

2.2 Data Swapping for Microdata Releases

Two papers followed the original data swapping proposal and extended those methods. Reiss (1984) presented an approximate data swapping approach for

the release of microdata from categorical databases that approximately preserves t -order marginal totals. He computed relevant frequency tables from the original database, and then constructed a new database elementwise to be consistent with these tables. To do this he randomly selected the value of each element according to probability distribution derived from the original frequency tables and then updated the table each time he generated a new element.

Reiss, Post, and Dalenius (1982) extended the original data swapping idea to the release of microdata files containing continuous variables. For continuous data, they chose data swaps to maintain generalized moments of the data, e.g., means, variances and covariances of the set of variables. As in the case of categorical data, finding data swaps that provide adequate protection while preserving the exact statistics of the original database is impractical. They present an algorithm for approximately preserving generalized k th order moments for the case of $k = 2$.

2.3 Applying Data Swapping to Census Data Releases

The U.S. Census Bureau began using a variant of data swapping for data releases from the 1990 decennial census. Before implementation, the method was tested with extensive simulations, and the release of both tabulations and microdata was considered (for details, see Navarro, et al. (1988) and Griffin et al. (1989)). The results were considered to be a success and essentially the same methodology was used for actual data releases.

Fienberg, et al. (1996) describe the specifics of this data swapping methodology and compare it against Dalenius and Reiss' proposal. In the Census Bureau's version, records are swapped between census blocks for individuals or households that have been matched on a predetermined set of k variables. The $(k + 1)$ -way marginals involving the matching variables and census block totals are guaranteed to remain the same; however, marginals for tables involving other variables are subject to change at any level of tabulation. But, as Willenborg and de Waal (2001) note, swapping affects the joint distribution of swapped variables, i.e., geography, and the variables not used for matching, possibly attenuating the association. One might aim to choose the matching variables to approximate conditional independence between the swapping variables and the others.

Because the swapping is done between blocks, this appears to be consistent with the goals of Dalenius and Reiss, at least as long as the released marginals are those tied to the swapping. Further, the method actually swaps a specified (but unstated) number of records between census blocks, and this becomes a data base from which marginals are released. However the release of marginals that have been altered by swapping suggests that the approach goes beyond the justification in Dalenius and Reiss.

Interestingly, the Census Bureau description of their data swapping methods makes little or no reference to Dalenius and Reiss's results, especially with regard to protection. As for utility, the Bureau focuses on achieving the calculation of summary statistics in released margins other than those left unchanged by

swapping (e.g., correlation coefficients) rather than on inferences with regard to the full cross-classification.

Procedures for the U.S. 2000 decennial census were similar, although with modifications (Zayatz 2002). In particular, unique records that were at more risk of disclosure were targeted to be involved in swaps. While the details of the approach remain unclear, the Office of National Statistics in the United Kingdom has also applied data swapping as part of its disclosure control procedures for the U.K. 2001 census releases (see ONS, 2001).

3 Variations on a Theme—Extensions and Alternatives

3.1 Rank Swapping

Moore (1996) described and extended the *rank-based proximity swapping algorithm* suggested for ordinal data by Brian Greenberg in an 1987 unpublished manuscript. The algorithm finds swaps for a continuous variable in such a way that swapped records are guaranteed to be within a specified rank-distance of one another. It is reasonable to expect that multivariate statistics computed from data swapped with this algorithm will be less distorted than those computed after an unconstrained swap. Moore attempts to provide rigorous justification for this, as well as conditions on the rank-proximity between swapped records that will ensure that certain summary statistics are preserved within a specified interval. The summary statistics considered are the means of subsets of a swapped variable and the correlation between two swapped variables. Moore makes a crucial assumption that values of a swapped variable are uniformly distributed on the interval between its bottom-coded and top-coded values, although few of those who have explored rank swapping have done so on data satisfying such an assumption. He also includes both simulations (e.g., for skewed variables) and some theoretical results on the bias introduced by two independent swaps on the correlation coefficient.

Domingo-Ferrer and Torra (2001a, 2001b) use a simplified version of rank swapping and in a series of simulations of microdata releases and claim that it provides superior performance among methods for masking continuous data. Trotinni (2003) critiques their performance measures and suggests great caution in interpreting their results.

Carlson and Salabasis (2002) also present a data-swapping technique based on ranks that is appropriate for continuous or ordinally scaled variables. Let X be such a variable and consider two databases containing independent samples of X and a second variable, Y . Suppose that these databases, $S_1 = [X_1, Y_1]$ and $S_2 = [X_2, Y_2]$ are ranked with respect to X . Then for large sample sizes, the corresponding ordered values of X_1 and X_2 should be approximately equal. The authors suggest swapping X_1 and X_2 to form the new databases, $S_1^* = [X_1, Y_2]$ and $S_2^* = [X_2, Y_1]$. The same method can be used given only a single sample by randomly dividing the database into two equal parts, ranking and performing the swap, and then recombining.

Clearly this method, in either variation, maintains univariate moments of the data. Carlson and Salabasis' primary concern, however, is the effect of the data swap on the correlation between X and Y . They examine analytically the case where X and Y are bivariate normal with correlation coefficient ρ , using theory of order statistics and find bounds on ρ . The expected deterioration in the association between the swapped variables increases with the absolute magnitude of ρ and decreases with sample size. They support these conclusions by simulations.

While this paper provides the first clear statistical description of data swapping in the general non-categorical situation, it has a number of shortcomings. In particular, Fienberg (2002) notes that: (1) the method is extremely wasteful of the data, using 1/2 or 1/3 according to the variation chosen and thus is highly inefficient. Standard errors for swapped data are approximately 40% to 90% higher than for the original unswapped data; (2) the simulations and theory apply only to bivariate correlation coefficients and the impact of the swapping on regression coefficients or partial correlation coefficients is unclear.

3.2 NISS Web-based Data Swapping

Researchers at the National Institute of Statistical Science (NISS), working with a number of U.S. federal agencies, have developed a web-based tool to perform data swapping in databases of categorical variables. Given user-specified parameters such as the swap variables and the swap rate, i.e., the proportion of records to be involved in swaps, this software produces a data set for release as micro-data. For each swapping variable, pairs of records are randomly selected and values for that variable exchanged if the records differ on at least one of the unswapped attributes. This is performed iteratively until the designated number of records have been swapped. The system is described in Gomatam, Karr, Chunhua, and Sanil (2003). Documentation and free downloadable versions of the software are available from the NISS web-page, www.niss.org.

Rather than aiming to preserve any specific set of statistics, the NISS procedure focuses on the trade-off between disclosure risk and data utility. Both risk and utility diminish as the number of swap variables and the swap rate increase. For example, a high swapping rate implies that data are well-protected from compromise, but also that their inferential properties are more likely to be distorted. Gomatam, Karr and Sanil (2004) formulate the problem of choosing optimal values for these parameters as a decision problem that can be viewed in terms of a risk-utility frontier. The risk-utility frontier identifies the greatest amount of protection achievable for any set of swap variables and swap rate.

One can measure risk and utility in a variety of ways, e.g., the proportion of unswapped records that fall into small-count cells (e.g., with counts less than 3) in the tabulated, post-swapped data base. Gomatam and Karr (2003, 2004) examine and compare several "distance measures" of the distortion in the joint distributions of categorical variables that occurs as a result of data swapping, including Hellinger distance, total variation distance, Cramer's V , the contingency coefficient C , and entropy. Gomatam, Karr, and Sanil (2004) consider a

less general measures of utility — the distortion in inferences from a specific statistical analysis, such as a log-linear model analysis.

Given methods for measuring risk and utility, one can identify optimal releases empirically by first generating a set of candidate releases by performing data swapping with a variety of swapping variables and rates and then measuring risk and utility on each of the candidate releases and provide a means of making comparisons. Those pairs that dominate in terms of having low risk and high utility comprise a *risk-utility frontier* that leads optimal swaps for allowable levels of risk. Gomatam, Karr, and Sanil (2003, 2004) provide a detailed discussion of choosing swap variables and swap rates for microdata releases of categorical variables.

3.3 Data Swapping and Local Recoding

Takemura (2002) suggests a disclosure limitation procedure for microdata that combines data swapping and local recoding (similar to micro-aggregation). First, he identifies groups of individuals in the database with similar records. Next, he proposes "obscuring" the values of sensitive variables either by swapping records among individuals within groups, or recoding the sensitive variables for the entire group. The method works for both continuous and categorical variables.

Takemura suggests using matching algorithms to identify and pair similar individuals for swapping, although other methods (clustering) could be used. The bulk of the paper discusses optimal methods for matching records, and in particular he focuses on the use of Edmond's algorithm which represents individuals as nodes in a graph, linking the nodes with edges to which we attach weights, and then matches individuals by a weighting maximization algorithm. The swapping version of the method bears considerable resemblance to rank swapping, but the criterion for swapping varies across individuals.

3.4 Data Shuffling

Mulalidhar and Sarathy (2003a, 2003b) report on their variation of data swapping which they label as data shuffling, in which they propose to replace sensitive data by simulated data with similar distributional properties. In particular, suppose that \mathbf{X} represents sensitive variables and \mathbf{S} non-sensitive variables. Then they propose a two step approach:

- Generate new data \mathbf{Y} to replace \mathbf{X} by using the conditional distribution of \mathbf{X} given \mathbf{S} , $f(\mathbf{X}|\mathbf{S})$, so that $f(\mathbf{X}|\mathbf{S}, \mathbf{Y}) = f(\mathbf{X}|\mathbf{S})$. Thus they claim that the released versions of the sensitive data, i.e., \mathbf{Y} , provide an intruder with no additional information about $f(\mathbf{X}|\mathbf{S})$. One of the problems is, of course, that f is unknown and thus there is information in Y .
- Replace the rank order values of \mathbf{Y} with those of \mathbf{X} , as in rank swapping.

They provide some simulation results that they argue show the superiority of their method over rank swapping in terms of data protection with little or no loss in the ability to do proper inferences in some simple bivariate and trivariate settings.

4 Data Swapping and Model-Based Statistical Methods

We can define model-based methods in two ways: (a) methods that use a specific model to perturb or transform data to protect confidentiality; or (b) methods that involve some perturbation or transformation to protect confidentiality, but preserve minimal sufficient statistics for a specific model, thereby maintaining the data users' inferences under that model. The former is exemplified by post-randomization methodologies and the latter by work on the release of margins from contingency tables or perturbed tables from conditional distributions. We describe these briefly in turn.

4.1 Post Randomization Method—PRAM

The Post Randomization Method (PRAM) is a perturbation method for categorical databases (Gouweleeuw, et al., 1998). Suppose that a sensitive variable has categories $1, \dots, m$. In PRAM, each value of the variable in the database is altered according to a predefined transition probability (Markov) matrix. That is, conditional on its observed value, each value of the variable is assigned one of $1, \dots, m$. Thus, observations either remain the same or are changed to another possible value, all with known probability. This is essentially Warner's (1965) method of randomized response but applied after the data are collected rather than before. Willenborg and de Waal (2001) note some earlier proposals of a similar nature and describe PRAM in a way that subsumes data swapping.

The degree of protection provided by PRAM depends on the probabilities in the transition matrix, as well as the frequencies of observations in the original database. PRAM has little effect on frequency tables. Given the transition matrix, it is straightforward to estimate the univariate frequencies of the original data, as well as the additional variance introduced by the method. The precise effect on more complicated analyses, such as regression models, can be difficult to assess. See the related work in the computer science literature by Agrawal and Srikant (2000) and Evfimievski, Gehrke, and Srikant (2003).

4.2 Model-based Approaches for the Release of Marginals and Other statistics

Fienberg, Steele, and Makov (1996, 1998) suggest “bootstrap-like” sampling from the empirical distribution of the data, and then releasing the sampled data for analysis. Multiple replicates are required to assess the added variability of estimates when compared with the those that could be generated from the original data. In the case of categorical data, this procedure is closely related to the problem of generating entries in a contingency table given a fixed set of marginals. Preserving marginal totals is equivalent to preserving sufficient statistics of certain log-linear models. Diaconis and Sturmfels (1978) developed an algorithm for generating such tables using Gröbner bases. Dobra (2003) shows that such bases correspond to simple data swaps of the sort used by Delanius and Reiss when the corresponding log-linear model is decomposable, e.g., conditional independence

of a set of $k - 1$ variables given the remaining one in a k -way contingency table. See Karr, Dobra, and Sanil (2003) for a web-based implementation.

The Dalenius and Reiss data swap preserves marginal totals of tables up to order t , and so can be viewed as a model-based method with respect to a log-linear model. In general, the set of tables that could be generated by data swapping is a subset of those that could be generated by the Diaconis and Sturmfels algorithm because of non-simple basis elements required to generate the full conditional distribution, e.g., see Diaconis and Sturmfels (1998) and Fienberg, Makov, Meyer, and Steele (2001). By comparison, the resampling method of Domingo-Ferrer and Mateo-Sanz (1999) is not model-based and can only be used to preserve a single margin. It has the further drawback of fixing sampling zeros, thereby limiting its usefulness in large sparse contingency tables.

Burridge (2003) extended the approach in Fienberg, Makov and Steele, for databases with continuous variables. Denote the database by (X, S) , where X represents sensitive variables that cannot be disclosed and S denote the remaining variables. Let T be a minimal sufficient statistic for the distribution of X given S . Values of the sensitive variables are replaced with a random sample, Y , from the distribution of X given (T, S) and the database (Y, S) is released. The idea is that the minimal sufficient statistic T will be preserved in the released database.

Clearly this method makes strong assumptions about the distributional properties of the data. In the case of discrete variables where the distribution comes from the exponential family, the results of Diaconis and Sturmfels apply again. Burridge proposes the method specifically for the case where $X|S$ is multivariate normal with mean $x\beta$ and covariance matrix Σ . He estimates sufficient statistics by fitting a separate linear regression model to each column of X and constructing the matrices $\hat{\beta}$ and $\hat{\Sigma}$, and he describes methods for generating perturbed data Y that preserve the conditional mean and variance of $X|S$. He also discusses the level of protection realized by this procedure. how general the approach is and how it relates to the other kinds of data swapping objectives in the non-categorical case remains to be seen. Note the similarity here to ideas in Mulalidhar and Sarathy (2003a, 2003b) but with a more formal statistical justification.

5 Discussion

In this paper we have revisited the original work of Dalenius and Reiss on data swapping and surveyed the some of the literature and applications it has spawned. In particular, we have noted the importance of linking the idea of data swapping to the release of marginals in a contingency table that are useful for statistical analysis. This leads rather naturally to a consideration of log-linear models for which marginal totals are minimal sufficient statistics. Although Dalenius and Reiss made no references to log-linear models, they appear in retrospect to provide the justification for much of the original paper. A key role in the rele-

vant theory is played by the conditional distribution of a log-linear model given its marginal minimal sufficient statistics.

There is an intimate relationship between the calculation of bounds for cell entries in contingency tables given a set of released marginals (Dobra and Fienberg, 2000,2001) and the generation of tables from the exact distribution of a log-linear model given its minimal sufficient statistics marginals. Work by Aoki and Takemura (2003) and unpublished results of de Loera and Ohn effectively demonstrate the possibility that the existence of non-simple basis elements can yield multi-modal exact distributions or bounds for cells where there are gaps in realizable values. These results suggest that data swapping as originally proposed by Dalenius and Reiss does not generalize in ways that they thought. But the new mathematical and statistical tools should allow us to reconsider their work and evolve a statistically-based methodology consistent with their goals.

6 Acknowledgments

The preparation of this paper was supported in part by National Science Foundation Grant No. EIA-0131884 to the National Institute of Statistical Sciences and by the Centre de Recherche en Economie et Statistique of the Institut National de la Statistique et des Études Économiques, Paris, France.

Bibliography

- Agrawal, R. and Srikant, R. (200). Privacy-preserving data mining. *Proceedings of the 2000 IEEE Symposium on Security and Privacy*, 439–450.
- Aoki, Satoshi and Takemura, Akimichi (2003). Minimal basis for connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. *Australian and New Zealand Journal of Statistics*, 45, 229–249.
- Bishop, Yvonne M. M., Fienberg, Stephen E., and Holland, Paul W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.
- Burridge, Jim (2003). Information preserving statistical obfuscation. *Journal of Official Statistics*, 13, 321–327.
- Carlson, Michael and Salabasis, Mickael (2002). A data-swapping technique for generating synthetic samples; A method for disclosure control. *Research in Official Statistics*, 5, 35–64.
- Dalenius, Tore (1977). Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 5, 429–444.
- Dalenius, Tore (1988). *Controlling Invasion of Privacy in Surveys*. Statistics Sweden, Stockholm.
- Dalenius, Tore and Reiss, Steven P. (1978). Data-swapping: A technique for disclosure control (extended abstract). *American Statistical Association, Proceedings of the Section on Survey Research Methods*, Washington, DC, 191–194.

- Dalenius, Tore and Reiss, Steven P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73–85.
- Diaconis, Persi and Sturmfels, Bernd (1998). Algebraic algorithms for sampling From conditional distributions. *Annals of Statistics*, 26, 363–397.
- Dobra, Adrian (2003). Markov bases for decomposable graphical models. *Bernoulli*, 9, 1–16.
- Dobra, Adrian and Fienberg, Stephen E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences*, 97, 11885–11892.
- Dobra, Adrian and Fienberg, Stephen E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals. *Statistical Journal of the United Nations ECE*, 18, 363–371.
- Domingo-Ferrer, Josep and Mateo-Sanz, Josep M. (1999). On resampling for statistical confidentiality in contingency tables. *Computers & Mathematics with Applications*, 38, 13–32.
- Domingo-Ferrer, Josep and Torra, Vicenc (2001). Disclosure control methods and information loss for microdata. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.): *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*. North-Holland, Amsterdam, pp. 91–110.
- Domingo-Ferrer, Josep and Torra, Vicenc (2001). A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.): *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*. North-Holland, Amsterdam, pp. 111–133.
- Doyle, Pat, Lane, Julia I., Theeuwes, Jules J.M., and Zayatz, Laura V., eds. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. North-Holland, Amsterdam.
- Duncan, George T., Fienberg, Stephen E., Krishnan, Rammaya, Padman, Rema, and Roehrig, Stephen F. (2001). Disclosure Limitation Methods and Information Loss for Tabular Data. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.): *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* North-Holland, Amsterdam, 135–166.
- Evfimievski, A., Gehrke, J., and Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. *Proceedings 2003 ACM PODS Symposium on Principles of Database Systems*.
- Federal Committee on Statistical Methodology (1978). *Report on Statistical Disclosure and Disclosure-Avoidance Techniques*. Statistical Policy Working Paper 2. Subcommittee on Disclosure-Avoidance Techniques. U.S. Department of Commerce, Washington, DC.
- Federal Committee on Statistical Methodology (1994). *Report on Statistical Disclosure Limitation Methodology*. Statistical Policy Working Paper 22. Sub-

- committee on Disclosure Limitation Methodology. Office of Management and Budget, Executive Office of the President, Washington, DC.
- Fienberg, Stephen E. (2002). Comment on a paper by M. Carlson and M. Salabasis: 'A data-swapping technique using ranks - A method for disclosure control.' *Research in Official Statistics*, 5, 65–70.
- Fienberg, Stephen E., Makov, Udi E., Meyer, M. M., and Steele, Russell J. (2001). Computing the exact distribution for a multi-way contingency table conditional on its marginal totals. In A.K.E. Saleh (ed.); *Data Analysis from Statistical Foundations: Papers in Honor of D.A.S. Fraser*, Nova Science Publishing, 145–165.
- Fienberg, Stephen E., Steele, Russell J., and Makov, Udi E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and loglinear models. . *Proceedings of Bureau of the Census 1996 Annual Research Conference*. US Bureau of the Census, Washington, DC, 87–105.
- Fienberg, Stephen E., Steele, Russell J., and Makov, Udi E. (1998). Disclosure limitation using perturbation and related methods for categorical data (with discussion). *Journal of Official Statistics*, 14, 485–511.
- Gomatam, Shanti and Karr, Alan F. (2003). Distortion measures for categorical data swapping. *Technical Report 132*, National Institute of Statistical Sciences, Research Triangle Park, NC.
- Gomatam, Shanti, Karr, Alan F., and Sanil, Ashish. (2003). A risk-utility framework for categorical data swapping. *Technical Report 132*, National Institute of Statistical Sciences, Research Triangle Park, NC.
- Gomatam, Shanti, Karr, Alan F., Chunhua "Charlie Liu, and Sanil, Ashish. (2003). Data swapping: A risk-utility framework and web service implementation. *Technical Report 134*, National Institute of Statistical Sciences, Research Triangle Park, NC.
- Gomatam, Shanti, Karr, Alan F., and Sanil, Ashish. (2004). Data swapping as a decision problem. *Technical Report 140*, National Institute of Statistical Sciences, Research Triangle Park, NC.
- Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and Wolf, P. P. de. (1998). Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14, 463–478.
- Griffin, R., Navarro, A., and Flores-Baez, L. (1989). Disclosure avoidance for the 1990 census. *Proceedings of the Section on Survey Research*, American Statistical Association, 516–521.
- Karr, Alan F., Dobra, Adrian and Sanil, Ashish P. (2003). Table servers protect confidentiality in tabular data releases. *Communications of the ACM*, 46, 57–58.
- Muralidhar, Krishnamurty and Sarathy, Rathindra (2003a). Masking numerical data: Past, present, and future. Presentation to Confidentiality and Data Access Committee of the Federal Committee on Statistical Methodology, Washington DC, April 2003.

- Muralidhar, Krishnamurthy and Sarathy, Rathindra (2003b). Access, data utility and privacy. Summary from *NSF Workshop on Confidentiality*, Washington DC, May 2003.
- Moore, Richard A. (1996). Controlled data-swapping techniques for masking public use microdata sets. *Statistical Research Division Report Series*, RR96-04, U.S. Bureau of the Census.
- Navarro, A., Flores-Baez, L., and Thompson, J. (1988). Results of Data Switching Simulation. Presented at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees.
- Office of National Statistics (2001). 2001 census disclosure control. Memorandum AG(01)06 dated November 27, 2001.
- Reiss, Steven P. (1984). Practical data-swapping: The first steps. *ACM Transactions on Database Systems*, 9, 20–37.
- Reiss, Steven P., Post, Mark J. and Dalenius, Tore (1982). Non-reversible privacy transformations. In *Proceedings of the ACM Symposium on Principles of Database Systems, March 29-31, 1982, Los Angeles, California*, pages 139–146.
- Schlörer, Jan (1981). Security of statistical databases: multidimensional transformation *ACM Transactions on Database Systems*, 6, 95–112.
- Takemura, Akamichi (2002). Local recoding and record swapping by maximum weight matching for disclosure control of microdata sets. *Journal of Official Statistics*, 18, 275–289.
- Trottini, Mario (2003). *Decision Models for Disclosure Limitation*. Unpublished Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.
- Warner, Stanley L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63–69.
- Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, Vol. 155, Springer-Verlag, New York.
- Zayatz, Laura (2002). SDC in the 2000 U.S. Decennial census. In *Inference Control in Statistical Databases*, (ed. J. Domingo-Ferrer), 183-202, Springer-Verlag, Berlin.