Draft 2.7 Friday, June 15, 2001,SFR

Confidentiality, Disclosure and Data Access:
Theory and Practical Applications for Statistical Agencies

Chapter 2.3

Disclosure Limitation Methods and Information Loss for Tabular Data

George T. Duncan, Stephen E. Fienberg, Ramayya Krishnan,
Rema Padman and Stephen F. Roehrig

1. Introduction

Even in the age of electronic dissemination of statistical data, tables are central data products of statistical agencies. For prominent examples, see the *American FactFinder* (http://factfinder.census.gov/servlet/BasicFactsServlet) from the U.S. Bureau of Census, the Office of National Statistics (http://www.statistics.gov.uk/) in the U.K., and Statistics Netherlands (http://www.cbs.nl/en/figures/keyfigures/index.htm). Much survey and census data is categorical in nature and thus the representation of survey results in the form of cross-classifications or tables is a natural device for statistical reporting. But even when they collect measurement data, statistical agencies often represent the information from them in the form of discretized quantities. As a result, tables of counts represent a primary unit of reporting and analysis. Sometimes these tables represent simple cross-classifications of the counts of survey and census elements. Other times the sample units are weighted according to probabilities of selection and /or are interpretable as the numbers of people in the population (based on the sample). In such tables of counts, the occurrence of small values is usually taken to present the possibility of a disclosure risk, since data for individuals who are unique in the population may be used in matching against other databases by an intruder or data snooper.

Considerable effort has gone into developing disclosure limitation methods for tabular data that effectively lower disclosure risk and provide products with high utility to legitimate data users (Duncan 2001, Duncan, Jabine and de Wolf 1993, Willenborg and de Waal 1996, 2000). These techniques include cell suppression, local suppression, global recoding, rounding, and various forms of perturbation (Federal Committee 1994). Under cell suppression, for example, the values of table cells that pose confidentiality problems are determined and suppressed (as primary suppressions) as well as values of additional cells that can be inferred from released table margins (as secondary suppressions) (Cox 1980). Perturbation is used through controlled rounding (Cox 1982),

versions of post-randomized response (Gouweleeuw, Kooiman, Willenborg, and de Wolf, 1998), and Markov perturbation approaches have been proposed in various forms by Duncan and Fienberg (1999), Fienberg, Makov, and Steele (1998), and Fienberg, Makov, Meyer, and Steele (2000). Many of these methods can be represented in the form of matrix masks (Duncan and Pearson 1991). The computational problems associated with these approaches have been widely explored in recent years through such techniques as (1) network methods by Cox (1995), (2) mathematical programming (IP, LP) and graph theory as addressed by Fischetti and Salazar (1996, 1998, 1999), Chowdhury, Duncan, Krishnan and Roehrig (1999), and Duncan, Krishnan, Padman, Reuther, and Roehrig (2001), and (3) branch and bound methods by Fienberg (1998), Dobra (2001), and Dobra and Fienberg (2001).

In Section 2 of this chapter, we describe a framework for simultaneously examining the impact of disclosure limitation techniques on the two attributes of confidentiality protection and information loss. The first attribute is characterized as inverse to disclosure risk, and measures the extent to which confidentiality is protected from the attacks of a data snooper. The second attribute is characterized as data utility, and measures the extent to which data users will still find the tabular data product useful even though there may be some information loss. In Section 3, we describe a variety of techniques, some quite new and under development, for limiting disclosure for tabular data. In Section 4, we consider the topic of disclosure auditing for tabular data. Disclosure auditing involves procedures for examining a proposed data product and assessing its vulnerability to attack by an intruder or data snooper. In Section 5, we show how the framework in Section 2 can provide what we call an R-U confidentiality map for evaluating and analyzing disclosure risk and data utility of tabular data. We devote our attention to tables of counts, and look both at two-way and multi-way tables. Most of the methods we describe, however, are applicable in related form to weighted tables of various kinds.

The methods surveyed in this chapter comprise a substantial part of the working arsenal of disclosure limitation practitioners at statistical agencies. Many of the references given at the end of the chapter point to seminal works in the field, and thus will be useful to both practitioners and researchers. The newer techniques described here are indicative of directions currently being pursued, and so will be of interest to researchers wishing to extend the state of the art.

To lend concreteness to our exposition throughout the chapter, we make use of the three-dimensional table presented in Table 1. It will illustrate the various issues and disclosure limiting methods. Rows are indexed by $i = 1,2,3, 4$, columns by $j = 1,2,3,4$ and levels by $k = 1,2,3$. We refer to the three two-way marginal totals derivable from this table by summing over variables as $IJ+$ (for the row by column totals), $I+K$ (for the row by layer totals), and $+JK$ (for the column by layer totals). Similarly, $I++$, $+J+$, and $++K$ represent the corresponding one-way marginal totals. The tables above the horizontal line are the three (i, j) levels, while the table below it is the $IJ+$ marginal.

If we consider this example a three-way population table, then the six cells with entries of "1" represent individuals who are *unique* in the population and thus pose a confidentiality problem. The six cells with entries of "2" would also be considered by most to pose serious disclosure risk, since one individual recorded in such a cell sees the other as unique. There are no such entries, however, in any of the two-way marginals, and this fact may, as we show below, generate a false sense of security on the part of a data administrator if the marginals alone were published.

k = 1

1	4	66	3	74
1	2	0	0	3
0	4	3	1	8
3	0	0	3	6
5	10	69	7	91

k = 2

2	3	2	68	75
0	4	78	3	85
0	0	0	61	61
0	3	1	0	4
2	10	81	132	225

k = 3

0	80	0	1	81
4	2	2	1	9
3	0	4	45	52
61	3	55	4	123
68	85	61	51	265

3	87	68	72	230
5	8	80	4	97
3	4	7	107	121
64	6	56	7	133
75	105	211	190	581

Table 1. Our Illustration.

2. A Framework for Disclosure Risk and Information Loss

Some argue that the legitimate objects of inquiry for statistical research are inferences drawn from aggregates over individual records; for example, the proportion of pilots for commercial airlines in the United States who have an alcohol abuse problem, perhaps given a set of additional demographic and occupational characteristics. The statistical agency often seeks to provide users with data that will allow accurate inferences about such population characteristics. Unfortunately, the additional characteristics of interest may well make the resulting cross-classification quite sparse, possibly replete with entries of "1" and "2." Because of confidentiality promises—whether explicit or implicit—the statistical agency seeks to thwart the data snooper who might seek to use the disseminated data to draw accurate inferences about, say, the alcohol abuse status of a particular pilot for American Airlines. This capability by a data snooper represents a statistical disclosure, but that level of statistical detail may still be of legitimate statistical interest on the part of a careful analyst, who has no interest in

utilizing the information about this particular pilot beyond the context of this statistical analysis.

There are two major types of disclosures—identity disclosure and attribute disclosure. *Identity disclosure* occurs with the association of a respondent's identity with a disseminated data record (Spruill 1983, Paass 1988, Strudler et al. 1986). *Attribute disclosure* occurs when the respondent can be associated with either an attribute value in the disseminated data or an estimated attribute value based on the disseminated (Duncan and Lambert 1989, Lambert 1993). In the case of identity disclosure, the association is assumed exact. In the case of attribute disclosure, the association can be approximate. Most statistical agencies place emphasis on limiting the risk of identity disclosure, perhaps because of its substantial equivalence to the inadvertent release of an identified record, a clear administrative slipup. On the other hand, an attribute disclosure, even though it invades the privacy of a respondent, may not be so easily traceable to actions of the agency.

We introduce a conceptual framework to provide context to our discussion of disclosure limitation methods. We take the data user to be primarily interested in the estimation of a conditional or a joint probability or population proportion based on the tabular data. We assume that an intruder or data snooper has access to external information that will make it likely that the snooper can compromise confidentiality when a cell count is small. Our framework establishes quantitative measures for two basic attributes:

1. **Disclosure Risk**—the measure of risk to confidentiality that the data trustee, such as a statistical agency, would experience by a data release.

2. **Data Utility**—a measure of the value of information to a legitimate data user.

Generally, the application of a disclosure limitation method would have the desirable effect of lowering disclosure risk while, concomitantly, have the undesirable effect of lowering data utility.

Disclosure risk is determined, most generally, based on how the agency envisions the data snooper making a disclosure. More simply it may be based on some measure of the percent of the population that could be easily compromised due to the uniqueness of their attribute values. Duncan and Lambert (1986, 1989), Chen and Keller-McNulty (1998), and Fienberg and Makov (1998), among others, develop specific disclosure risk models. Finally, the released data will have more or less utility for the user depending on the degree of perturbation from the original data and the intended use of the data. The statistical disclosure limitation problem is to choose a methodology for data release so that disclosure risk is adequately low while statistical information (data utility) in the disseminated data are as high as possible (Duncan and Fienberg 1999). This characterization of the problem is displayed in Figure 1. Data utility is the value of the statistical information that the agency provides to a legitimate user.

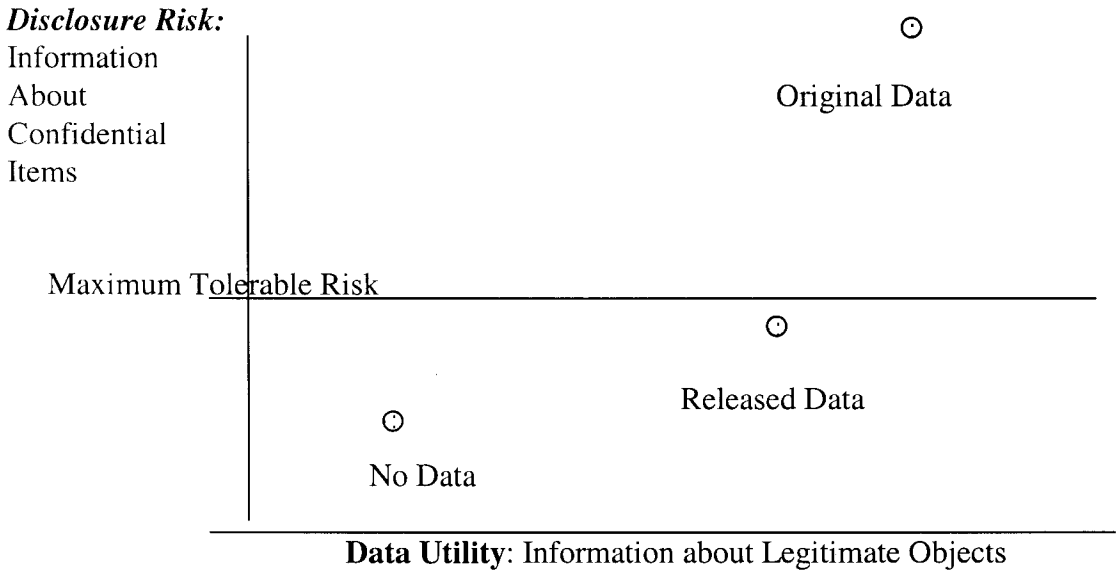

Figure 1: The Statistical Disclosure Limitation Problem

In addressing the statistical disclosure limitation problem, the agency is shoring the two pillars of its foundation: satisfying the data users who depend on its products and reassuring the respondents who provide it with data. Data snoopers pose a threat to the agency's ability to deliver on its promise of confidentiality to respondents. Thus, Figure 1 is a graphical assessment of how an agency provides data utility to users and lowers disclosure risk in the face of attack by data snoopers. Domingo-Ferrer (1999) terms a data product of high data utility to be both *analytically valid* (key statistical characteristics, like means and covariances preserved) and *analytically interesting* (several variables on important subdomains provided) and notes that it involves low information loss; he terms data products with low disclosure risk to be *safe*. Zaslavsky and Horton (1998) use a decision-theoretic approach based on the structure of Figure 1 to derive an optimal disclosure limitation scheme for minimum cell size in tabular data.

Generally, a data snooper has *a priori* knowledge about a target (Duncan and Lambert, 1986). Typically, this knowledge would take the form of a database with identified records (Adam and Wortmann, 1989). Certain variables may be in common with the subject database. These variables are called *key* or *identifying* (de Waal and Willenborg, 1996, 1998). When a single record matches on the key variables, the data snooper has a candidate record for identification. This candidacy is promoted to an actual identification if the data snooper is convinced that the individual is in the target database. This would be the case either if the data snooper has auxiliary information to that effect or if the data snooper is convinced that the individual is unique in the population. The data snooper may find that according to certain key variables, a sample record is unique. The question then arises as to whether the individual is also unique on these key variables in the population. Bethlehem, Keller and Pannekoek (1990) have examined detection of records agreeing on simple combinations of keys based on discrete variables in the files.

Record linkage methodologies have been examined by Fuller (1993) and extensively by Winkler (1998), who uses a version of the matching algorithm originally proposed by Fellegi and Suntner (1969). Trottini (2001) presents an even more comprehensive framework for the decision-theoretic trade-offs between the perspectives of the agency and the perspectives of users, in light of the extent to which an intruder is able to infer target values from a released dataset.

Elliot and Dale (1999) and Paass (1988) explore the psyche and motivations of the data snooper. The snooper may or may not be someone with limited access to the data and may or may not be motivated for malicious reasons. Prudently, however, the database administrator must assume a worst-case scenario, i.e., that the data snooper has access to sophisticated analytical tools; is knowledgeable about the data and has ready access to relevant external data sources; and has the necessary computational power to attempt an attack on the data.

Data utility is a positive expression of information loss. A variety of measures of data utility have been proposed. For example, Özsoyoğlu and Chung (1986) suggested a measure for tabular data under disclosure limitation through cell suppression as simply the percentage of suppressed cells. This particular measure is crude at best. Similarly, de Waal and Willenborg (1998) consider a variety of options for choosing local suppressions (i.e., values for specific variables in specific records) by focussing on the total number of such suppressions, or the number of categories affected by the local suppressions.

More generally, we presume that once the database administrators are able to hold disclosure risk to an adequately low level, they should then seek to maximize data utility. This follows from the perspective that all disclosure limitation methods are attempting to maximize data utility for a given user task subject to a constraint on disclosure risk. We examine the fundamental tradeoff between data utility and disclosure risk. As information loss increases because of disclosure limitation, an estimate of a conditional probability becomes less precise, and data utility consequently goes down. Simultaneously, disclosure risk also decreases. This conceptual framework can be used to compare alternative disclosure limitation methods. Typically one can control the use of any particular disclosure limitation method by choosing certain parameter values. For example, with noise addition the parameter is the variance, say τ^2, of the added noise. As τ^2 is changed, the disclosure risk R and the data utility U change. These changes can be presented graphically in what is called an R-U confidentiality map (Duncan and Keller-McNulty 2001, also Section 5 below).

Here we model disclosure risk for any individual table cell as the sum

$$\sum_k r(k)p(k),$$

where $r(k)$ is the risk associated with the data snooper obtaining knowledge that a cell entry has true value k, and $p(k)$ is the probability that the cell value is k, given the table and knowledge held by the data snooper about the disclosure-limiting method employed.

Thus the sum ranges over the possible true values a cell could have, given the published value.

To illustrate how this model could be applied, consider a table protected by rounding to base 3. For example, if from Table 1 the *IJ+* marginal table (the 4×4 table below the horizontal line) is rounded to base three, Table 2 results. (We discuss table rounding in more detail in Section 3.) A cell with a published value 3 could have true value 1, 2, 3, 4, or 5. The data disseminator would assign r*(k)*, $k \in \{1, 2, 3, 4, 5\}$ according to the perceived risk associated with an intruder determining that k was the true value. Generally, *r(k)* would be a decreasing function of k, since the disclosure risk would be higher with smaller cell counts.

3	87	69	72	231
6	9	81	3	99
3	3	6	108	120
63	6	57	6	132
75	105	213	189	582

Table 2. A Table Rounded to Base Three

In determining *p(k)* for each k, the data disseminator would attempt to assess the probability distribution that the data snooper would use. There are several possible approaches to this. A conservative approach is to act as though the data snooper places high probability on the lower possible values of k. Another approach is to base assessments of *p(k)* on the actual frequency distribution of cell counts in some reference population. Then the data disseminator can model information loss (the complement of data utility) as, for example, the mean square error in estimating conditional probabilities.

As we noted above, this framework allows us, in principle, to examine the various disclosure limitation schemes with regard to both data utility and disclosure risk. In the following sections, we illustrate this framework with a numerical example—an example that we will use in the evaluation and analysis section to compare some of the disclosure limitation methods, and show how it can quantify the utility/risk tradeoff.

3. Disclosure Limitation

In this section, we discuss a number of protection schemes that have been proposed for tables of counts. All of the methods discussed have appeared in the statistical literature with applications to two-dimensional tables. We will discuss the applicability of all of them to three- and higher-dimensional tables, since many desirable properties of statistical disclosure limitation techniques for two-dimensional tables are absent from the corresponding techniques for tables of higher dimension (Cox 1999).

Sampling

One of the surest ways to limit disclosure is to release only part of the data. Thus releasing a table whose counts are based on a sample of the units in the original table is a way to provide a serious measure of protection for the original reporting units. The statistical agency practice of releasing microdata samples is essentially based on this approach. The virtue of sampling in this context is that if the details of the sampling procedure are available, a user can make valid inferences about the population underlying the sample table, albeit with less precision than would have been possible with the original table.

Cell Suppression

Cell suppression has been used for many years by a large number of statistical agencies (Cox 1980, de Carvalho, Dellaert and Osório 1994, Cox 1985, Kelly, Golden and Assad 1992, Fischetti and Salazar González 2000, Cox 2001, Giessing 2001). On the face of it, the idea is simplicity itself. If a table contains an entry that is deemed sensitive, the disseminator simply does not provide a value for it. This has no effect at all on other table entries, so it effectively localizes the distortion of the data to individual cells. A suppressed cell, in isolation, would appear to be able to take on any value whatsoever (and so provide complete protection), but in the context of the table as a whole, there are evident constraints that arise from marginal values and algebraic relationships between cell entries. Furthermore, data disseminators sometimes publish the rules used to determine suppressions, thus providing further clues to their likely values.

Under cell suppression, each sensitive cell in the table is suppressed; these are called *primary* suppressions. If marginal totals or other linked tables are also to be published, it may be necessary to remove additional cell values (*secondary* or *complementary* suppressions) that would allow an intruder to use algebraic or other means to identify the sensitive cell values. In terms of our R-U framework, the goal of cell suppression is to find secondary suppressions that maximize the utility of the resultant table while affording sufficient protection. Often the total number of suppressed cells is taken as the measure of utility, but other measures (e.g., entropy) have been used as well.

Except in special circumstances, the secondary cell suppression problem (CSP) is computationally NP-hard (Kelly, et al., 1992), suggesting that any solution procedure will grow exponentially in complexity with increasing problem size. Recent work by Fischetti and Salazar González (2000) has increased considerably the size of tables that can be protected optimally by suppression, but it is still quite common for heuristics to be used instead of procedures that are provably optimal. In addition, many of the heuristics in current use do not guarantee a specific level of protection. For such heuristics, one of the disclosure auditing techniques described in Section 4 should always be applied before a table is made public.

Many of the published heuristic methods for cell suppression (Cox, 1980, 1995, Kelly, et al., 1992) rely on the simple structure of two-dimensional tables. Algebraic

relationships between cells in two-dimensional tables are effectively captured in network flow models, which typically have fast solution routines. Unfortunately, once the transition is made to three- and higher-dimensional tables, the network representation breaks down, voiding many useful theoretical results and algorithms. The exact methods of Fischetti and Salazar González, (2000), however, do not depend on network structure, and so can find optimal suppression patterns for arbitrary n-dimensional tables, at least those of moderate size. When the size of the tables is such that optimal suppressions cannot be computed, meta-heuristic approaches such as tabu search (Glover and Laguna, 1997) provide a way of looking for "good" solutions within a reasonable time. Essentially, tabu search is a neighborhood search method with a built-in mechanism to prevent the algorithm from becoming stuck at local optima. Duncan et al. (2001) apply the tabu search approach in conjunction with the fast disclosure auditing approach of Chowdhury, et al. (1999) to solve cell suppression problems in three-dimensional tables. The quality of the tabu search solutions was comparable to the quality of the IP solutions.

Unfortunately, there is only a limited set of circumstances where the special structure of the tables to be protected permits an efficient solution algorithm. For example, see Duncan, et al. (2001), and the decomposition and reducibility results of Dobra and Fienberg (2000a), which characterize an important class of linked tables that give rise to these simplifications. In the absence of these special structures, however, large cell suppression problems (especially ones involving large tables and substantial numbers of primary suppressions) are computationally difficult.

Duncan and Fienberg (1999) and Fienberg (1997, 2000) have criticized cell suppression because it causes unnecessary loss of statistical information. In particular, they note that complementary suppressions destroy data that are not themselves sensitive, and the resulting tables greatly reduce the ability of the user to make correct inferences about relationships in the original unsuppressed table. Thus cell suppression achieves disclosure limitation at the expense of elimination of some data.

Rounding

Rather than simply suppress a sensitive cell, one might disguise its true value by modifying it in a principled way. One way to do this is to choose a positive integer b and round table entries to an integer multiple of it. This is usually done for *all* cells in the table. Rounding has the general advantage of providing at least a roughly correct value for every cell (assuming b is small), and thereby helps the data user to avoid badly incorrect inferences about cell values. Cell suppression, on the other hand, does open the possibility that the user may draw false inferences about the suppressed cell values (Duncan and Fienberg 1999). Less positively, with all cell values rounded, many more would typically be changed from their true values than would be the case with cell suppression. Also, when multiple, overlapping tables are rounded individually, a common cell may end up being rounded to two different values.

Rounding can be done with more or less sophistication. Fellegi (1972, 1975) introduced the notion *of controlled rounding* which insists that the rounded table is additive, meaning that rows, columns, etc. sum to the their respective (rounded) marginals (see also Cox and Ernst 1982). *Zero-restricted* controlled rounding (Kelly, Golden and Assad 1990) further requires that cell values in the unmodified table that are already multiples of b (in particular, zeros) remain so. Finally, *unbiased* controlled rounding (Causey, Cox and Ernst 1985, Cox 1987) specifies that the expected value of a rounded cell value equals its unrounded value. The requirements imposed by these different rounding methods are all related to attempts to improve the data utility for users while still minimizing disclosure risk in some formal sense.

Simple and efficient polynomial-time algorithms have been devised for all of these flavors of rounding, at least for two-dimensional tables. Polynomial-time algorithms have the property that the worst-case solutions time grows only as a polynomial function of the size of the problem (here, the number of cells to be rounded). More difficult computational problems are classified as NP-hard, and for such problems solution times may increase exponentially with size. The controlled rounding problem for three dimensions has been shown to be NP-hard (Kelly, Assad and Golden, 1990), though the heuristic given in Kelly, Golden and Assad (1990) has proven to be effective when unbiased solutions are not required. Fischetti and Salazar-González (1998) provide advice for implementing controlled rounding approaches empirically in three and more dimensions. The method of Dobra (2001) for bounding cell values implicitly or explicitly generates feasible tables and thus has the potential of identifying controlled rounding solutions or near solutions in higher dimensions.

If natural assumptions are made about the distribution of cell values in a table (i.e., that it is nearly uniform across local intervals of length b), it is often easy to specify probabilities for each possible cell value in a rounded table. In many cases, even if the value of the rounding base b is not explicitly announced, it can be easily deduced from the published table itself.

Data Swapping, Confidentiality Edit, and Simulated Tables

Dalenius and Reiss (1978) first proposed a method for swapping observations "at random" while preserving marginal totals. In Dalenius and Reiss (1982) they illustrate the implementation of a kth-order swap in which all k-dimensional margins of a p-dimensional table are preserved. This is similar to the notion of randomly selecting a replacement table among a restricted set of alternative tables with the same k-dimensional marginal totals. An obvious issue is how to choose k. Dalenius and Reiss illustrate their proposal with k=2. There is also the issue about what fraction of records to swap. Without going into details, Dalenius and Reiss were unable to come up with a general method for accomplishing data swapping. Nor were they able to assess the increase in variability associated with the added randomness.

The U.S. Census Bureau used a variant of data swapping in the context of the *Confidentiality Edit* as part of the 1990 decennial census. They wanted to interchange a subset of households in different census blocks who shared a number of characteristics, say k, in common. This has the result of holding the corresponding k-dimensional totals for those blocks fixed, as well as the $(p\text{-}k)$-dimensional margin adding across blocks and across the variables held constant under swapping. The primary method they focused on matched records on k=6 variables. For further details, see Navarro *et al* (1988), Griffin *et al* (1989), and Fienberg, Steele, and Makov (1996). Again we have a statistical issue about the choice of k, as well as the issue about the fraction of records to be swapped.

To illustrate this alternative notion of data swapping pictorially, we consider a 3×2×2 contingency table with entries $\{n_{ijk}\}$ as follows

n_{111}	n_{121}	n_{1+1}	n_{112}	n_{122}	n_{1+2}
n_{211}	n_{221}	n_{2+1}	n_{212}	n_{222}	n_{2+2}
n_{311}	n_{321}	n_{3+1}	n_{312}	n_{322}	n_{3+2}
n_{+11}	n_{+21}	n_{++1}	n_{+12}	n_{+22}	n_{++2}

We want to track what happens when we swap the values for a randomly selected pair of individuals, one in layer 1 and the other in layer 2. Suppose that the individual selected from layer 1 is in the (1,2,1) cell and that we are swapping his/her characteristics with a randomly selected individual in the (3,1,2) cell. The result is as follows:

n_{111}	$n_{121}-1$	$n_{1+1}-1$	n_{112}	$n_{122}+1$	$n_{1+2}+1$
n_{211}	n_{221}	n_{2+1}	n_{212}	n_{222}	n_{2+2}
$n_{311}+1$	n_{321}	$n_{3+1}+1$	$n_{312}-1$	n_{322}	$n_{3+2}-1$
$n_{+11}+1$	$n_{+21}-1$	n_{++1}	$n_{+12}-1$	$n_{+22}+1$	n_{++2}

Note that the two-dimensional total for the first two variables (adding over layers) is unchanged, as is the one-dimensional total for the third variable. This process is now repeated for pairs of randomly selected units in the two layers, thus producing a confidentiality edit that continues to preserve the same marginal totals. One variant of this, used by the Census Bureau for the 2000 census, is to select swapping "partners" by targeting certain unique records. Moore (1996) describes additional applications of variants on data swapping and contrasts them with the simple matrix masking technique of adding noise.

We note that data swapping is, like cell suppression, a method for altering cell counts in a multi-dimensional cross-classification while maintaining fixed marginal totals. This observation led Fienberg, Makov, and Steele (1996, 1998) to propose a more elaborate version of repeated data swapping which in essence allows for a series of moves from one

table to another subject to marginal constraints. Their method utilizes the tool of Gröbner bases described in Section 4 below, and in essence replaces the original table by a random draw from the exact distribution under the log-linear model whose minimal sufficient statistics correspond to the released marginals, subject to those marginals being fixed. Thus probabilistic simulation yields a replacement table with the same marginal totals as the original table. Actually, Fienberg, Makov, and Steele (1998) go further in proposing the retention of the simulated table only if it is consistent with some more complex log-linear model.

The data-swap transformation described above represents one of a subclass of possible "moves" in a Markov chain algorithm proposed by Diaconis and Sturmfels (1998). Such moves alone, however, do not always suffice to generate the exact distribution. Even in the cases where they do suffice, however, one needs to run the Markov chain a very long time in order to simulate the exact distribution as explored by Fienberg, Makov, and Steele (1998). Making a small proportion of swaps, as is done in practice, is not sufficient to rest the methodology on a firm statistical foundation that a user can invoke in order to assess the added uncertainty that results from the alteration of the data.

An extremely important feature of this simulation methodology is that information on the variability which it introduces into the data is directly accessible to the user, since anyone can begin with the reported table and information about the margins that are held fixed, and then run the Diaconis-Sturmfels Markov chain algorithm to regenerate the full distribution of all possible tables with those margins. This then allows the user to make inference about the added variability in a formal modeling context in a form that is similar to the approach to inference in Gouweleeuw, et al. (1998). As a consequence, simulation and perturbation methods represent a major improvement from the perspective of access to data over cell suppression and data swapping.

This approach offers the prospect of simultaneously smoothing the original counts *and* providing disclosure limitation protection. But there remain many practical issues regarding the use and efficacy of such methods for generating disclosure-limited public-use samples. For example,

- How effective are such devices for limiting disclosure, i.e., protecting against attack by a data snooper?
- What is the data utility (correspondingly, information loss) when we compare actual data with those released?
- How can they be used when the full cross-classification of interest is very sparse, consisting largely of 0s and 1s?
- How can we use models to generate the simulated data when the users have a multiplicity of models and even classes of models that they would like to apply to the released data?
- What if a release involves thousands of tables with overlapping cells?

We discuss some of the implications of using Markov perturbations in more detail below.

Markov Perturbation

The method of simulating from the exact distribution of a table given a set of marginals is intimately related to the notion of Markov perturbation described by Duncan and Fienberg (1999). Thinking of cell values as counts of entities classified in a particular category, Markov perturbation deliberately misclassifies by selectively moving entities from one classification to another. This is done in such a way that marginal totals are preserved, and the expected values of all cells are unchanged. In employing Markov perturbation, the statistical agency (1) lowers disclosure risk by increasing the uncertainty of a data snooper about the true cell value, and (2) gives the legitimate data user a value for analysis, albeit one that is subject to misclassification error—an error process that any good data analyst of categorical data must contend with anyway.

The procedure of Markov perturbation works as follows, described for a two-dimensional table. An *elementary data square* is chosen as a 2x2 submatrix of the table. Then each entity (i.e., each individual contributor to the counts in the four cells in the submatrix) moves to an adjacent cell (up or down, left or right) according to a Markov transition matrix chosen to be stationary. The transition matrix is chosen so that row sums and column sums are unchanged. To protect the entire table, the process is repeated with a random sequence of the possible elementary data squares.

In this section we have surveyed the principle techniques used for disclosure limitation. Since some of these techniques may in practice rely on heuristic algorithms to provide the desired level of protection, the next section complements this one by discussing methods that test whether this protection has been realized.

4. Disclosure Auditing

Disclosure auditing is a process of examining a proposed data product to assess its vulnerability to attack by a data snooper. As part of a sensible procedure for evaluating security implementations, protectors should play the role of those who might attempt to compromise the security. They should search for weak points. Prudently, they should assume that the attacker has adequate resources to similarly identify and exploit such weak points. In this section we present methods for disclosure audit that have been available for some time, as well as new methods. Special attention is given to higher-dimensional tables.

A data disseminator might wish to publish an entire, original, table. This table may well have cells, that we call sensitive cells, which are deemed to pose unacceptable disclosure risks. It is common to declare a cell in a population table whose value is small, say 1 or 2, as posing an unacceptable disclosure risk and hence sensitive. In that case, the table should not be disseminated in its original form. Instead, it should first be

transformed through some disclosure limitation procedure, such as one of the techniques described in Section 3. If the technique used cannot inherently guarantee the requisite level of protection, it is necessary to audit the proposed release, that is, apply a procedure to test the level of protection actually afforded. For example, cell suppression patterns are often determined through heuristic methods that don't provide such guarantees.

Alternatively, for reasons of brevity or protection, the disseminator might wish to publish tables—say one or more of the two-dimensional marginal tables *IJ+, I+K or +JK* in our running example—that are *derived* from the original higher-dimensional table. If the goal is to protect values in the original table, or values in an unpublished margin, auditing is again necessary to ensure confidentiality protection.

Before an audit can proceed, it is necessary for the disseminator to decide what constitutes a sensitive cell. Sufficient protection exists for a sensitive cell provided that in the released data product the true value of this cell entry is sufficiently ambiguous to a data snooper. A common and useful scheme is to define a protection range and demand that protection be such that any value in the range is potentially the correct cell value.

If such a protection range is given for a cell, then an audit verifies that there indeed exist realizations of the table that agree with the published data but have the sensitive cell with a value anywhere in the protection range. For example, Table 1 might be protected by suppressing all cells with values 1 or 2, (and also additional cells if necessary—see Section 4) so that for each sensitive cell, any value in the range [0, 4] is feasible.

Linear and Integer Programming

To verify a protection range for a sensitive cell of a table with published marginal totals, the obvious technique is linear programming (LP) or integer linear programming (IP) (Zayatz 1993). The published cells are used to form linear constraints on the possible values of the cell. Then the sensitive cell value is both minimized, to obtain the lower bound on the cell value, and maximized, to obtain the upper bound on the cell value. The lower and upper bounds then provide the protection range for that cell. In the case of multiple sensitive cells, the procedure is repeated for each. Since the constraints implied by the published cell values apply to every sensitive cell, the max/min pair for one cell can be calculated independently of that for any other cell.

Several difficulties arise in the use of standard linear programming (LP), and consequently there is considerable interest in finding alternative techniques. This is especially true for implementing procedures on large tables, which can require considerable computational effort, depending on the number of sensitive cells needing protection. For the example given above, 11 cells are sensitive, so an equal number of maxima and minima must be calculated, and this is computationally feasible. For much larger tables, and many more suppressions, the task can become daunting. As an example, the Bureau of Labor Statistics ES-202 Employment and Wages quarterly

publication has, for some county/state aggregations, nearly 40,000 cells, more than half of them suppressed. A "brute force" LP approach makes little sense in such a case.

Fortunately, it is often possible to decompose the overall audit problem into smaller pieces. For example, large tables often have one or more attributes arranged hierarchically (e.g., counties within a state, or 4-digit SIC codes "rolling up" to 3-digit codes). In such cases, auditing can be done at each level separately—if due care is given to the fact that margins at one level correspond to internal table entries at the next higher level. With such a decomposition, the number of LP problems to be solved increases, but the size of each problem (as measured by the number of variables plus constraints) decreases. Because the average time requirements of most linear programming algorithms increase roughly as the cube of the problem size, a large net savings in total computation can accrue. As an illustration, a two-level hierarchical table with a total of 8 rows and 8 columns contains 64 internal cells and 16 marginal constraints. If this table can be decomposed hierarchically into four 4×4 tables, then each of the four has 16 internal entries and 8 marginal constraints. Auditing the original table would consume computing resources proportional to $(64 + 16)^3$, while the four smaller ones would need resources proportional to $4\times(16 + 8)^3$, about a 90 percent savings.

Another way in which LP-based auditing procedures can be improved takes advantage of the linked structure of the table. In the process of maximizing, say, one cell value, a simplex-based LP algorithm will incrementally increase the current value for that cell, subject to the imposed constraints. At each step, the value of every other cell is recorded in a data structure (the "simplex tableau") that can be easily examined to see if any cell (other than the one that is currently being maximized) is at its guaranteed maximum or minimum. Detecting these occurrences is a simple matter, and it obviates the need to perform a separate optimization on that cell. Empirically, this can speed up the auditing process considerably.

A second potential difficulty with linear programming lies in the nature of some auditing problems. For tables of dimension greater than two, there is no general guarantee that the optima produced by linear programming will be integers. Two-dimensional tables can be represented as a network, with rows and columns as nodes, and internal cells as arcs. Because of a special property of networks, optimal solutions to max/min problems are integer, provided the known values (unsuppressed cells, row and column sums) are integer. Unfortunately, three- and higher-dimensional tables can no longer be represented as a network, so the integrality property of optima is lost (Roehrig 1999). For example, using Table 1 as the base table, suppose that we wish to publish the three two-dimensional margins *IJ+*, *I+K*, and *+JK*. If cells in the underlying three-dimensional table are considered sensitive, the LP approach to finding inferable bounds will result in an upper bound of 13.5 for cell (1, 3, 3). This is clearly unobtainable in any table of counts, so it cannot be right. Thus auditing techniques that give sharp *integer* bounds are needed. Integer programming (IP) is the obvious solution, but in general this is extraordinarily more difficult computationally than LP, which permits continuous solutions. In many circumstances, LPs giving non-integer optima can be augmented with

additional, legitimate constraints that force integrality ("Gomory cuts," for example (Schrijver 1986)). A general theory has yet to be developed, but Roehrig (2001b) has obtained results for special cases. Alternative approaches such as meta-heuristic search (e.g., Tabu Search described by Glover and Laguna, 1997) that have performed well on difficult combinatorial optimization problems can also be applied to these difficult auditing problems. Duncan, et al. (2001) report on the application of Tabu search to a related disclosure limitation problem.

Alternative Approaches

Because of the difficulties outlined above, other approaches to auditing disclosure in tables of counts have been pursued. We describe three of them here, but this is an active area of research, so the list is growing. The first alternative derives from generalizations of the well known Fréchet and Bonferroni bounds on joint probability distributions given lower-dimensional marginal distributions. It applies to certain situations in which lower-dimensional marginal totals (themselves tables) are to be published, and bounds are needed on entries in the original higher-dimensional table. Dobra and Fienberg (2000) show that when the released marginals form a *decomposable graph*, one can combine information from the subgraphs to realize sharp bounds for entries in the original table. Further, the same structure can be used to break the problem of computing bounds for a large table into sets of much smaller ones corresponding to irreducible components.

In our three-dimensional example, suppose that the margins *IJ*+ and +*JK* are published, but we wish to protect the underlying table *IJK*. The published marginals directly associate *I* and *J*, and *J* and *K*, but the interaction between the dimensions *I* and *K* are indirect. This can be visualized by drawing a graph with nodes representing the dimensions and arcs that indicate which nodes are joined by the published marginal tables. In this example, the graph is especially simple and is shown in Figure 2.

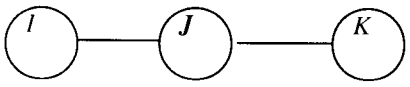

Figure 2. Graph Representing Released Marginal Tables.

Node *J* is a "separator" of nodes *I* and *K*. Dobra and Fienberg show that if a separator (which may be larger than a single node, depending on the released margins) is a clique, then a variant of the normal Fréchet bounds can be used to calculate bounds for entries in the base table (*IJK* in this case). These results are powerful because the bounds can essentially be "read off" from the published tables. Dobra and Fienberg give closed-form expressions for these bounds, so that no iterative mechanism is required.

Chowdhury, et al. (1999) developed an equivalent network-based bounding scheme for another three-dimensional case. Suppose once again that the marginals *IJ*+ and +*JK* from Table 1 are to be released, but now the third two-dimensional marginal *I*+*K* is considered sensitive. Here are the three two-dimensional marginal tables:

IJ+

3	87	68	72
5	8	80	4
3	4	7	107
64	6	56	7

+JK

5	2	68
10	10	85
69	81	61
7	132	51

I+K

74	75	81
3	85	9
8	61	52
6	4	123

Bounds on *I+K* can be quickly determined by either the Dobra and Fienberg or Chowdhury et al. procedures, and compared with the desired bounds from the confidentiality intervals for each sensitive cell. For the marginal table *I+K*, we show below upper bounds of the desired protection range determined as the actual cell value plus 20 percent and the upper bounds of the computed protection range. Note that disclosures occur when the computed upper bound is below the desired upper bound; there are three disclosures—at cells (1,1), (2,2) and (3,3).

Desired Upper Bounds

89	90	98
4	102	11
10	74	63
8	5	148

Computed Upper Bounds

88	152	200
86	94	78
21	120	65
74	71	133

The Dobra-Fienberg approach for the decomposable case has some natural extensions to tables that correspond to reducible graphs and this class of extensions can reduce substantially the computational demands of the calculation of bounds in large numbers of dimensions, e.g., see Dobra and Fienberg (2001). A third and closely related bounding technique, suggested in Fienberg (1999), elaborated upon by Dobra (2001), and illustrated in Dobra and Fienberg (2001), can also be thought of as a generalization of Buzzigoli and Giusti's (1999) "shuttle" algorithm. In its basic form, the Dobra procedure starts with loose upper and lower bounds for each cell, then iteratively narrows the bounds by taking advantage of cell relationships inherent in the tabular structure. The resulting bounds are sharp for a well-characterized group of problems, but for the general case the procedure uses a variant of the integer programming technique of implicit enumeration to find a table realization that provably achieves the sharpest bounds. The procedure's especially attractive feature is that it is relatively efficient in computing sharp bounds for the special cases such as when the marginals can be used to describe a decomposable graph, or when the corresponding graph has a reducible structure. The general method also works in the presence of "structural zeros" and so may be of use in connection with other disclosure limitation approaches such as identifying secondary suppressions.

Using either LP or the methods described above to find bounds on cell entries specifies *extremal* values; the process in and of itself does not give the likelihood for any individual value in the range. To fit into our disclosure risk framework, we need a way to specify or estimate the probability associated with each feasible value. This is possible, at least in principle. Diaconis and Sturmfels (1998) show how one can systematically

sample from the set of all tables that agree with the published marginal values. They provide a list of "moves"—changes to the internal cell entries—that leave the published marginal values unchanged. Such moves (often called "Gröbner basis" moves because of the method used to generate them) can be described as a set of cell increments and decrements; one example for a 3×3×3 table is the following.

0	0	0
0	-	+
0	+	-

+	0	-
-	+	0
0	-	+

-	0	+
+	0	-
0	0	0

Applying a move from this list to a feasible table (that is, one that agrees with the published values) results in a new feasible table. The list is generated in such a way that the set of all feasible tables can be traversed uniformly if one chooses moves randomly with equal probability from the list. Thus an estimate of the probability of a particular cell value can be obtained by moving randomly through the set of feasible tables and tallying the proportion of time that one lands on a table having that cell value. Fienberg, Makov, and Steele (1998) apply this work in the context of disclosure limitation problems and link it to Markov perturbations. Fienberg, Makov, Meyer, and Steele (2000) present an expository treatment of the theory in the context of contingency tables, making explicit links to the theory of log-linear models and they provide heuristic descriptions of the role of Gröbner bases (the moves described above) when the MCMC procedure approaches the extremal values.

Diaconis and Sturmfels applied this idea to sampling from the space of k-way tables when all (k-1)-way margins are known, and Fienberg, Makov, Meyer, and Steele (2000) make clear how it generalizes to complete k-way tables with any set of marginals fixed, but the idea easily generalizes to other situations, in particular to cell suppression. Cell suppression merely adds some constraints and removes some possible moves, so the basic plan of constructing the Gröbner basis to find legitimate moves still applies.

While elegant in principle, the Gröbner basis idea is limited in practice at present. The difficulty is computational. Currently, the best general-purpose computer programs take many hours to find the Gröbner basis moves for a 3×3×3 table when the three 2-way margins are given. Specialized programs can solve the same problem in a fraction of a second, but still take *months* to solve the analogous 5×4×3 problem (Roehrig 2001a). Larger problems are, at least in general, simply out of the question. Recent work (Dobra, 2000), however, shows how to calculate a basis quickly for the class of auditing problems whose released marginals form a decomposable graph. His construction extends to related problems and allows for the combination of Gröbner bases for component subtables in regular graphs.

To apply our framework for disclosure risk, the statistical agency might assume that the data snooper holds a particular probability distribution for the values within the protection range. A reasonable procedure might be to assume a unimodal distribution

spanning the interval, with its peak at the center and tails of decreasing probability extending to the endpoints. This is reasonable because although it is possible to have multiple feasible tables achieving the interval endpoint values, interior values allow more freedom for other cells to change, increasing the number of feasible tables possible. Thus one would not expect a distribution uniform over the protection range.

Other Disclosure Limitation Methods

Thus far we have discussed auditing tables that have either not been modified, or have been modified using cell suppression. In both cases, the published cells (i.e., the unsuppressed cells) imply constraints that serve to bound the sensitive cells. Some other forms of disclosure limitation may also be audited using LPs. The various forms of rounding (Section 3) all result in cells whose true values are unknown (thus affording protection), yet are known to lie in a well-defined interval. Thus, each published cell value gives rise to a pair of inequality constraints. Just as before, LP can find extreme values for a cell. Of course, the agency only needs to invoke LP to find bounds on quantities not in the rounded table; each rounded cell has obvious bounds. Nonetheless, a data disseminator may still want to know how bounds on sensitive cells in other, linked tables are influenced by the release of the rounded tables. In our example, the tables $IJ+$ and $+JK$ may be released in rounded form, but the disseminator might wish to know the resulting bounds for the table $I+K$.

Tables that have been protected by perturbation, either by the addition of random noise, data swapping, or Markov perturbation, rely on the theoretical properties of the method. As an example of this we now give an analysis of the protection provided by Markov perturbation.

To use the disclosure risk formula given in Section 2, we need a technique for determining the probabilities associated with possible cell values. We begin that process here, by showing the steps necessary to model a data snooper and incorporate that model into the analysis. A procedure along the same lines can be used to find a snooper's beliefs under other disclosure limitation schemes like cell suppression and rounding.

Consider the following elementary data square, taken as part of a larger two-dimensional table.

1	14	15
17	83	100
18	97	115

This square is altered using Markov perturbation, and the resulting (published) square is the snooper's starting point. (The full Markov perturbation process will, as described above, alter cells within this square as other, intersecting, squares are perturbed. We treat the simple single-square case here, and defer the full analysis for later research.) The top-left cell has the true value $\omega=1$. Modifications to the square that leave the

margins fixed and the internal entries non-negative are restricted; the masked value M for the top-left cell can be no larger than 15 and no smaller than 0. We can think of this square as being composed of a number of entities (0) that must remain there because of the marginal and non-negativity restrictions, and entities currently classified there but free to move out (1). Similarly, the lower-left cell can be thought of as consisting of unmoving entities (3) and movable ones (14). During the perturbation, some proportion of the top-left cell's movable entities move out, and some of the lower-left cell's movable entities move in. Under moves that are independent and identically distributed for a particular cell, the resulting number entities in the top-left cell is a random variable that can be expressed as a constant plus the sum of two independent binomial random variables. It has the form:

$M = c + Binom(1,1-\theta) + Binom(14, \frac{r}{1-r}\theta)$

where $r = 1/18$ and is included to preserve stationarity (see Duncan and Fienberg, 1999, for the details). The parameter $\theta \in [0, \min((1-r)/r, 1)]$ determines the extent of disclosure limitation: it controls the probability of movement by the entities. If θ is zero, the table remains unchanged, while larger values provide increasing protection.

Suppose that after perturbation, the published data square is the following:

3	12	15
15	85	100
18	97	115

A data snooper's view of the protected top-left cell is the following:

> I can think of the published cell value M as what I've "observed", and the true "state of nature" ω as being the unperturbed value. I want to construct the probabilities of the various true states of nature, given the evidence provided me by the published table. I know enough to construct probabilities for the observed value given the various possible true states of nature. But this is the wrong way around. Yet, with my prior distribution on the possible cell values, I can use Bayes' theorem to "invert" the conditional probabilities.

Let's see how this can be done.

Because the Markov perturbation process leaves margins fixed, it is easy to determine the range of possible true values ω for the top-left cell (the possible states of nature). For each of these, we first construct *P(Observed | State of Nature)*. For example, one possible state of nature is the value ω=1 (which happens to be the true state, although unknown to the snooper). There are a number of ways to move from 0 (the count of unmovable entities) to 3 (the published value), each a sum adding to 3 of the two kinds of movement described above. Not all pairs are permissible, however, because the (1,1) cell can "give up" only one entity. So the feasible set of "stayer-mover" pairs is $A = \{(0,3),$

(1,2)}. We write 3 = $X + Y$, where X ~ Binom(1, 1 - θ) and Y ~ Binom(14, $r/(1-r)\theta$). Then the likelihood function value at ω=1 is given by

$$P(Observed\ M = 3 \mid State\ of\ Nature\ \omega = 1) = \sum_{A} P(X = x)P(Y = y)$$

$$= \sum_{A}\binom{1}{x}(1-\theta)^{x}\theta^{1-x} \bullet \binom{14}{y}\left(\frac{r}{1-r}\theta\right)^{y}\left(1-\frac{r}{1-r}\theta\right)^{14-y}.$$

The general form of the likelihood function is given by

L(ω)=P(Observed M | State of Nature ω)=

$$\sum_{A}\binom{\omega}{x}(1-\theta)^{x}\theta^{\omega-x}\binom{15-\omega}{y}\left(\frac{r}{1-r}\theta\right)^{y}\left(1-\frac{r}{1-r}\theta\right)^{15-\omega-y},$$

where $A = \{(x, y) : x + y = M; 0 \le x \le \omega; 0 \le y \le 15-\omega\}$ and $r = \frac{\omega}{18}$.

To calculate this likelihood value the data snooper must know, or assume, the value of θ. This would be the case if the agency publicly released the value of this disclosure limitation parameter. If the agency chose not to release the value of θ, then the data snooper would be uncertain about the appropriate likelihood value. The effect of this would be to raise the data snooper's perceived chances of error and hence lower the disclosure risk. Hence, calculations based on assuming the data snooper knows the value of θ can be taken to provide upper bounds on the actual disclosure risk. Similarly, without knowing for sure the value of θ the data user also has increased uncertainty. Based on the value of θ, the value of r can be computed by
the snooper because of the conditioning assumption that the state of nature has value ω, so r = ω/18. Calculating these conditional probabilities for the possible states of nature ω and based on an observed value of M=3, we get Table 3. These entries can be interpreted as the likelihood values for each of the given θ values.

θ ω	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.05	0	.0007	.0683	.7746	.1482	.0178	.0017	.001	.0000	0	0	0	0	0	0	0
0.1	0	.0026	.1151	.6193	.2235	.515	.0096	.0016	.0002	0	0	0	0	0	0	0
0.2	0	.0088	.1655	.4382	.2708	.1121	.380	.0114	.0032	008	.002	0	0	0	0	0

Table 3. Likelihood Values

In keeping with our overall plan, we now need to specify the data snooper's prior beliefs for the various states of nature.

In the simplest case, we might assume that the data snooper holds a uniform distribution over ω. The posterior probabilities P(State of Nature = ω | Observed M = 3) are just those in Table 3; the prior distribution is uninformative. Suppose, instead, that the data snooper's prior distribution on ω were the following, anticipating a true state of nature close to 1 (note that the data snooper has prior probability 0 that ω = 0, because true zeroes would be unperturbed).

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	.3	.3	.15	.1	.05	.04	.03	.02	.01	0	0	0	0	0	0

Table 4. Data Snooper's Prior Distribution on the State of Nature ω

Then the posterior probabilities for ω are as in Table 5.

θ	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.05	0	.001	.146	.751	.096	.006	0	0	0	0	0	0	0	0	0	0
0.1	0	.005	.259	.579	.139	.016	.002	0	0	0	0	0	0	0	0	0
0.2	0	.015	.403	.381	.157	.032	.009	.002	0	0	0	0	0	0	0	0

Table 5. Posterior Probabilities for Prior Distribution Given in Table 4.

A tractable family of distributions for basing the data snooper's prior distribution for ω is the beta-binomial family, conditioned on the known upper and lower bounds for n. A special case of the beta-binomial distribution is the discrete uniform distribution discussed above.

The general picture is this. Different choices of the disclosure limitation parameter θ produce differing amounts of "blurring" of the probabilities of the true state of nature, and so provide varying degrees of disclosure risk. At the same time, these different values of θ cause different amounts of data distortion, and consequently affect data utility. In Section 5 we will illustrate some of these tradeoffs.

5. Evaluation and Analysis

We are now in a position to show how two disclosure limitation methods can be compared using the R-U confidentiality map described initially in Section 2. For ease of exposition, we consider the simple 2×2 table with marginal totals of Section 3, and compare cell suppression with Markov perturbation. The technique is extensible to larger tables and different limitation methods.

Since our 2×2 table contains a 1 in cell (1,1), we assume that this cell is a primary suppression. This decision necessitates complementary suppressions, which for this

simple case must obviously include the remaining three interior cell entries. Thus under cell suppression, all that can be published are the marginal totals. For Markov perturbation, we assume that the published table is the one discussed in Section 4 above.

To trade off disclosure risk and data utility, we require specific measures. To assess disclosure risk, for this illustration we use the reciprocal of entropy. Specifically, we take $R = 1 / (-\Sigma(p_\omega \log p_\omega))$, where p_ω is the snooper's probability that the (1,1) cell value is ω. This measure assumes that disclosure risk is reduced as the snooper's probability function over the possible true cell values ω spreads out. To measure data utility, we use mean squared precision, specifically the reciprocal of the mean squared error based on the probability distribution of ω available to the data user and the fact that the true value of ω is 1.

In the case of Markov perturbation, we assume both the data user and the snooper are aware of the form of disclosure limitation that has been applied to the table, and for convenience we assume that both parties have the same prior distributions on the disguised value ω in cell (1,1). For cell suppression, we measure both disclosure risk and data utility according to the bounds that can be computed for the missing (1,1) cell. As noted in the previous section, the value ω is easily seen to be in the range [0, 15]. For simplicity of illustration, let us suppose that both the data snooper and the data user are interested in the value of ω. In further developments we will take the data snooper to be primarily interested in whether ω can be taken to be 1, and the data user to be interested in inference about the probability of falling in the (1, 1)-cell according to a probability model.

With these measures, we find the following data utilities and disclosure risks, also depicted in Figure 3.

	Data Utility	**Disclosure Risk**
Markov Perturbation		
$\theta = 0.05$	.240	1.263
$\theta = 0.1$	.232	0.969
$\theta = 0.2$	.216	0.713
Cell suppression	0.159	0.255
Rounding	0.346	0.919
Original Data	∞	2

Included in the table and graph are the results of an identical risk/utility analysis of our simple 2×2 table after rounding to base three. If we assume that the "No Data" case (as shown in Figure 1) amounts to publishing just the marginal totals, the risk and utility values coincide with those for cell suppression. On the other hand, if "No Data" means that not even the margins are published, risk and utility are both zero. The relative performance of the various disclosure limitation methods examined in this simple

example should not be taken as suggesting that one method is universally superior to another. These and other methods need to be examined in the context of their actual use, with actual data products.

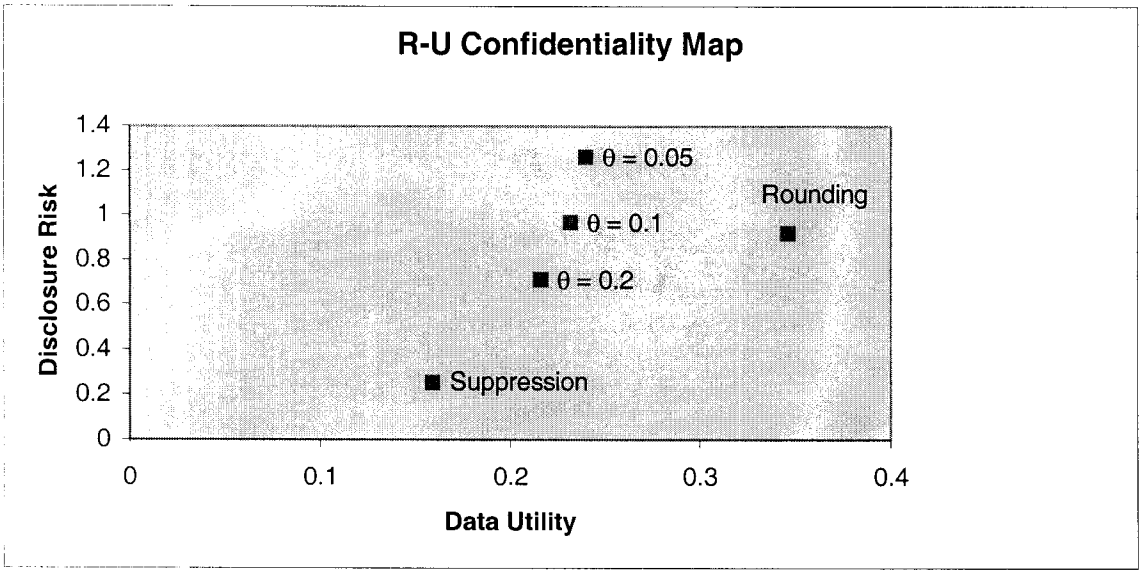

Figure 3. R-U Confidentiality Map for Suppression and Markov Perturbation

With each data product there is a disclosure threshold, above which the disclosure risk is too great. As Figure 2 illustrates, for different risk thresholds, different disclosure limitation methods, or their parameters, may be preferred. In this example, by varying the Markov exchange parameter θ it is possible to move from no protection ($\theta = 0$) to protection essentially equivalent to that provided by cell suppression, yet with higher data utility. This example is extreme, as it only analyzes an elementary 2×2 table. Nonetheless, it is clear that under reasonable assumptions, very different forms of disclosure limitation can be successfully compared.

There are several ways the R-U confidentiality map might be used within an organization. First, it may be the case that the organization is unclear on the actual level of disclosure risk that has been borne in the past, and therefore unsure how to proceed in the future. Generating R-U maps for previous data releases would enable it to quantify the risks taken in the past, and compare such risks among different data products. Such a program could enable the organization to develop a coherent strategy of dissemination, one in which comparable (or perhaps justifiably different) risks exist over the various releases. Knowledge of risk in past releases could be further used as a benchmark for the risk associated with a new release, especially one using a new disclosure limitation technique.

Along this same line, a comparison of two competing limitation techniques might result in an R-U confidentiality map like the one in Figure 4. In this figure, the choice of

disclosure limitation method depends crucially on the disclosure threshold. As the threshold is raised above the point of intersection of the two curves, method 1 provides considerably more utility for a given threshold increase, suggesting new possibilities for the risk-utility tradeoff. Our central theme, that the R-U confidentiality map allows much more informed decision making, is especially apparent in this example.

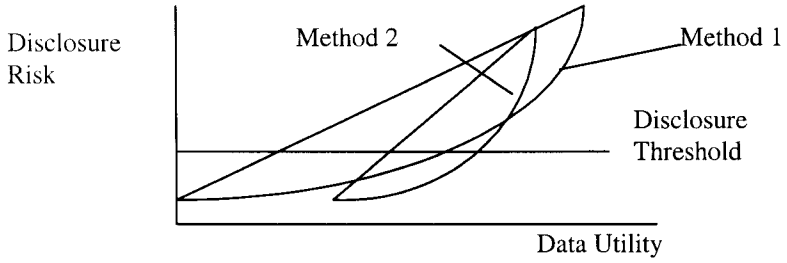

Figure 4: A Hypothetical R-U Confidentiality Map

6. Conclusions

Here we summarize our results, provide some perspectives, and indicate areas where additional research might provide improved disclosure limiting techniques.

Statistical agencies have a variety of disclosure limitation methods available for use in their efforts to protect the confidentiality of tabular data. A systematic way of comparing the merits or these methods is through the R-U confidentiality map. An important further consideration is the computational burden of the procedure.

Perturbation methods are attractive because they have the prospect of providing users with more data and in a form that allows for proper statistical inferences. We discussed several related versions in this chapter.

Recent advances in computational algorithms and the new statistical perspectives that have been brought to bear on disclosure limitation problems suggest that we may soon be in a position to do a much more thorough job of examining tabular data for possible disclosures and then applying disclosure limitation methods in such a form as to give users greater access to data for analysis.

Towards this end, we see the need for further research to identify procedures that have increased data utility while maintaining low disclosure risk, and more attention to the development of efficient computational algorithms that scale to the high-dimensional tabular problems typical of much statistical agency data.

Acknowledgements

Original research reported in this chapter was supported in part by the National Science Foundation under grant EIA-9876619 to the National Institute of Statistical Sciences.

URLs Referred to in the Paper

American FactFinder: http://factfinder.census.gov/servlet/BasicFactsServlet

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

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

References

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

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990) Disclosure control of microdata. *Journal of the American Statistical Association* **85** 38-45.

Buzzigoli, L. and Giusti, A. (1999) An algorithm to calculate the lower and upper bounds of the elements of an array given it marginals. In *Statistical Data Protection (SDP'98) Proceedings*, Eurostat, Luxembourg, 131–147.

Carvalho, F. de, Dellaert, N. and Osorio, M. de Sanches (1994) Statistical disclosure in two-dimensional tables: General tables. *Journal of the American Statistical Association* **89** 1547-1557.

Causey, B., Cox, L. and Ernst, L. (1985) Applications of transportation theory to statistical problems. *Journal of the American Statistical Association* **80** 903-909.

Chen, G. and Keller-McNulty, S. (1998) Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* **14** 79-95.

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

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

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

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

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

Cox, L.H. (1999) On properties of multi-dimensional statistical tables. Unpublished manuscript.

Cox, L.H. (2002) Disclosure Risk for Tabular Economic Data. Chapter 3.3 of this volume.

Dalenius, T. and Reiss, S.P. (1978). Data-swapping: A technique for disclosure control (extended abstract). *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 191-194.

Dalenius, T. and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6** 73-85.

De Vries, R. E. (1993) Disclosure control of tabular data using subtables. Report. Statistics Netherlands, Voorburg.

De Waal, A. G. and Pieters, A. J. (1995) ARGUS User's Guide Report, Department of Statistical Methods, Statistics Netherlands, Voorburg.

De Waal, A. G. and Willenborg, L. C. R. J. (1994) Minimizing the number of local suppressions in a microdata set. Report. Statistics Netherlands, Voorburg.

De Waal, A.G. and Willenborg, L.C.R.J. (1996). A View on Statistical Disclosure for Microdata. *Survey Methodology* **22** 95-103.

De Waal, A.G. and Willenborg, L.C.R.J. (1998). Optimal local suppression in microdata. *Journal of Official Statistics* **14** 421-435.

Diaconis, P. and Sturmfels, B. (1998) Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics* **26** 1 pp. 363-97.

Dobra, A. (2000). Measuring the disclosure risk for multi-way tables with fixed marginals corresponding to decomposable log-linear models. Technical Report, Department of Statistics, Carnegie Mellon University.

Dobra, A. (2001). Computing sharp integer bounds for entries in contingency tables given a set of fixed marginals. Technical Report, Department of Statistics, Carnegie Mellon University.

Dobra, A. and Fienberg, S. E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences* **97**, 11185-11192.

Dobra, A. and Fienberg, S. E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals. Paper prepared for 2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality 14 - 16 March 2001, Skopje, Macedonia.

Domingo-Ferrer, Josep (1999) Microdata masking methods. Workshop on Confidentiality Research. May 3-4. U.S. Census Bureau. Alexandria, VA.

Duarte de Carvalho, F., Dellaert, N. P., de Sanches Osório, M. (1994) Statistical disclosure in two-dimensional tables: General tables. *Journal of the American Statistical Association* **89** 1547-1557.

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

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

Duncan, G. T., Jabine, T. B., and de Wolf, V. A. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* Panel on Confidentiality and Data Access, Committee on National Statistics, National Academy Press, Washington, DC.

Duncan, G. T. and Keller-McNulty, S. (2001) Disclosure risk vs. data utility: The R-U confidentiality map. Technical Report. Statistical Sciences Group. Los Alamos National Laboratory. Los Alamos, New Mexico.

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

Duncan, G. T. and Lambert, D. (1986) Disclosure-limited data dissemination (with discussion) *Journal of the American Statistical Association.* **81** 10-28.

Duncan, G. T. and Lambert, D. (1989) The risk of disclosure of microdata. *Journal of Business and Economic Statistics* **7** 207-217.

Duncan, G. T. and Pearson, R. (1991) Enhancing access to microdata while protecting confidentiality: Prospects for the future (with discussion). *Statistical Science* **6** 219-239.

Elliot, M. and Dale, A. (1999) Scenarios of attack, the data intruders' perspective on statistical disclosure risk. *Netherlands Official Statistics* **14** 6-10.

Ernst, L. R. (1989) Further applications of linear programming to sampling problems. Technical Report Census/SRD/RR-89-05. Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Federal Committee on Statistical Methodology (1994) Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology. Washington, DC: U.S. Office of Management and Budget.

Fellegi, I. P. (1972) On the question of statistical confidentiality. *Journal of the American Statistical Association* **67** 7-18.

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

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

Fienberg, S. E. (1994) Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics* **10** 115-132.

Fienberg, S.E., Steele, R.J., and Makov, U.E. (1996) Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and log-linear models. *Proceedings of Bureau of the Census 1996 Annual Research Conference*. US Bureau of the Census, Washington, DC, 87-105.

Fienberg, S. E. (1997) Confidentiality and disclosure limitation methodology: challenges for national statistics and statistical research. Paper commissioned by the Committee on National Statistics for presentation at its 25th anniversary meeting.

Fienberg, S. E. (1999) Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation. In *Statistical Data Protection (SDP'98) Proceedings*, Eurostat, Luxembourg, 115?129.

Fienberg, S. E. (2001) Statistical perspectives on confidentiality and data access in public health. *Statistics in Medicine* **20** (in press).

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

Fienberg, S. E., Makov, E. U., Meyer, M. M., and Steele, R. J. (2001) Computing exact distribution for a multi-way contingency table conditional on its marginal totals. In *Data Analysis from Statistical Foundations: Papers in Honor of D.A.S. Fraser*, ed. A, Saleh. Nova Science Publishing.

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

Fischetti, M. and Salazar-González, J. J. (1996) Models and algorithms for the cell suppression problem. *Proceedings of the Third International Seminar on Statistical Confidentiality*. EUROSTAT, Luxembourg, 114-122.

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

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

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

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

Giessing, S. (2002) A practitioner's guide to non-perturbative disclosure control methods for tabular data. Chapter 3.1 of this volume.

Glover, F. and Laguna, M. (1997), *Tabu Search*, Kluwer Academic Publishers, Boston, MA.

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

Griffin, R., Navarro, A., and Flores-Baez, L. (1989) Disclosure avoidance for the 1990 census. *Proceedings of the Section on Survey Research*, American Statistical Association, 516-521.

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

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

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

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

Moore, R.A. (1996). Controlled data swapping techniques for masking public use microdata sets. RR 96-05. U.S. Bureau of the Census, Washington, DC.

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

Navarro, A., Flores-Baez, L., and Thompson, J. (1988) Results of Data Switching Simulation. Presented at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees.

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

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

Roehrig, S.F. (1999) Auditing disclosure in multiway tables with cell suppression: simplex and shuttle solutions. Paper presented at the American Statistical Association Joint Statistical Meetings, Baltimore, MD, August 8.

Roehrig, S.F. (2001a) Computing Gröbner bases for statistical disclosure limitation. To be presented at Grostat 2001, New Orleans, September 2001.

Roehrig, S.F. (2001b) Finding integer solutions to disclosure limitation problems using strong inequalities, Working Paper, The Heinz School of Public Policy and Management, Carnegie Mellon University.

Schrijver, A. (1986) *Theory of Linear and Integer Programming*. Wiley, New York.

Spruill, N. L. (1983) The confidentiality and analytic usefulness of masked business microdata. *Proceedings of the Section on Survey Research Methods,* American Statistical Association 602-607.

Strudler, M., Oh, H. L., and Scheuren, F. (1986) Protection of taxpayer confidentiality with respect to the tax model. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 375-381.

Trottini, M. (2001) A decision-theoretic approach to data disclosure problems. Paper prepared for 2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality 14-16 March 2001, Skopje, Macedonia.

Willenborg, L. and de Waal, T. (1996) *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics **111** Springer, New York.

Willenborg, L. and de Waal, T. (2000). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics **155** Springer-Verlag, New York.

Winkler, W. E. (1998) Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Research in Official Statistics* **1** 87-104.

Zaslavsky, A.M. and Horton, N.J. (1998) Balancing disclosure risk against the loss of nonpublication. *Journal of Official Statistics*, **14**, 411-419.

Zayatz, L. (1993) Using linear programming methodology for disclosure avoidance purposes. Proceedings of the International Seminar on Statistical Confidentiality. EUROSTAT, Luxembourg, 341-351.

Zayatz. L. V. and Rowland, S. (1999) Disclosure limitation for American FactFinder. Paper presented at the American Statistical Association Joint Statistical Meetings, Baltimore, MD, August 8.