

Datamining and Disclosure Limitation for Categorical Statistical Databases

Stephen E. Fienberg

Department of Statistics, Center for Automated Learning and Discovery, and CyLab
Carnegie Mellon University
Pittsburgh, PA 15213-3890, U.S.A.
fienberg@stat.cmu.edu

Abstract

There are many distinctions between statistical research databases and those arising in commercial or administrative settings, and thus different issues regarding confidentiality and privacy protection on the one hand and data access and the use of databases on the other. Data integration across multiple databases raises issues in both domains, especially with regard to protection against intruders. This paper highlights some methods developed to limit possible disclosure of confidential information from statistical databases while at the same time publicly releasing sufficient information to allow users, whether dataminers or other more traditional statistical analysts, sufficient data to reach proper statistical conclusions from their analyses. The disclosure limitation tools discussed include: data perturbation and simulation, partial releases, and sampling, with a special focus on partial release of data from multi-dimensional cross-classifications or contingency tables.

1. Introduction

Disclosure limitation is a term that is usually used to denote a set of statistical techniques aimed to protect confidentiality of individual respondents in the release of statistical data. The goal of those applying disclosure limitation techniques, often in government statistical agencies, is to release publicly data that are both useful for secondary analysts and do not violate the privacy and confidentiality of respondents who have provided the data. This paper focuses on tools for statistical disclosure limitation and their role in assessing the tradeoff between confidentiality protection and utility. For a review of the earlier literature on the topic of statistical disclosure limitation, see [27, 33] and, for an overview primarily from the perspective of statistical agencies, see [24]. Our focus here and the statistical methods related to it are somewhat different from those which one finds in the computer science and data mining literatures,

c.f., Adam and Wortman [1] and Agrawal and Srikant [2]. In particular we are thinking of datamining largely in terms of the application of statistically valid search and inference tools for large-scale data bases.

At the outset we emphasize our choice of the label *disclosure limitation* as opposed to *disclosure avoidance* to describe the relevant methodology. This is because the fundamental notion of disclosure is considered to be probabilistic:

If the release of the statistics S makes it possible to determine the value [of confidential statistical data] more accurately than is possible without access to S , a disclosure has taken place. [30, pp. 7 and 9]

Because any release is likely to carry some information about confidential elements in a statistical database, the only way to avoid all risk is not to release any data. But this runs counter to the mission of a statistical agency which is to release data for public and government statistical use. As a consequence, the goal of disclosure limitation methodology then becomes managing the tradeoff between disclosure risk and the utility of the released data for inferential purposes (c.f., Duncan et al. [26, 28] and Trottini [51, 52, 53]). This paper discusses the tradeoff between confidentiality protection and utility in a more informal fashion than that which is found in these other sources and attempts to take stock of some of the progress achieved in the work on disclosure limitation over the past decade.

We begin, in Section 2, with a brief review of the statistical literature on disclosure limitation including methods for assessing the tradeoff between disclosure risk and utility. Then we focus on methods for categorical data in the form of multi-dimensional cross-classifications or contingency tables. This serves several purposes but most importantly allows us to see the key role of statistical models from the disclosure limitation perspective. We introduce log-linear models in Section 3 and discuss the key properties for both disclosure limitation and making statistical inferences. In Section 4, we describe the method of partial re-

lease of marginal tables and a somewhat simplistic method of disclosure risk assessment through bounds and distributions over possible tables satisfying the margins. We use as an example a 2^6 table. In Section 5, we focus on the trade-off issue for the example as a way of illustrating the more formal approaches suggested in the literature. In Section 6, we describe very briefly the related methods associated with the calculation of distributions over possible tables given the partial release of a set of margins, as well as the counting of tables. In the final section of the paper, we suggest some basic principles that should underpin future research developments in the area of disclosure limitation and we describe some of the research challenges that remain.

2 Methodology for Disclosure Limitation

Duncan and Pearson [27] categorize the methodology used for disclosure limitation in terms of disclosure limiting masks, i.e., transformations of the data where there is a specific functional relationship (possibly stochastic) between the masked values and the original data. The basic idea of data masking involves thinking in terms of transformations. In essence it involves transforming an $n \times p$ (cases by variables) data matrix Z through pre- and post-multiplication and the possible addition of noise, i.e.,

$$Z \longrightarrow AZB + C, \quad (1)$$

where A is a matrix that operates on the n cases, B is a matrix that operates on the p variables, and C is a matrix that adds perturbations or noise. Matrix masking includes a wide variety of standard approaches to disclosure limitation:

- adding noise,
- releasing a subset of observations (delete rows from Z),
- cell suppression for cross-classifications,
- including simulated data (add rows to Z),
- releasing a subset of variables (delete columns from Z), and
- switching selected column values for pairs of rows (data swapping).

Even when one has applied a mask to a data set, the possibilities of both identity and attribute disclosure remain, although the risks may be substantially diminished.

There are different ways to categorize disclosure limiting masks, e.g., as suppressions (e.g., local suppression [54] or cell suppression, perhaps subject to marginal constraints, [24]); recodings (e.g., collapsing rows or columns,

sometimes referred to as global recoding [54, 28], or data swapping [11, 38]); or sampling (e.g., releasing subsets of observations).

Sampling clearly provides a measure of direct protection from disclosure provided that there is no information of which individuals or units are included in the sample. An intruder wishing to identify an individual in the sample and link that person's information to data in external files, using "key" variables such as age and geography available in both databases, needs first to determine whether a record is unique in the sample, and then if so, the extent to which a record that is unique in the sample is also unique in the population. For continuous variables, virtually all individuals are unique in the sample, and we need to understand the probability that an intruder would correctly match records, e.g., in the presence of error in the key variables (e.g., see Fienberg et al. [36]). For categorical data, uniqueness corresponds to counts of "1" and various authors [35, 45, 47] have examined how to reason about the probability that a record which is unique in the sample is also unique in the population as well as about other measures of disclosure risk [46].

Some masking methods alter the data in systematic ways, e.g., through aggregation or through cell suppression, and others do it through random perturbations, often subject to constraints for aggregates. Controlled random rounding [8] is an example of a "systematic" perturbation method, whereas data swapping [11] and the post-randomization method (PRAM) of Gouweleeuw, et al. [40] involve "random" perturbations. One way to think about random perturbation methods is as a restricted simulation tool, and thus we can link them to other types of simulation approaches that have recently been proposed.

Some masking methods alter the data in systematic ways, e.g., through aggregation or through cell suppression, and others do it through random perturbations, often subject to constraints for aggregates. Controlled random rounding [8] is an example of a "systematic" perturbation method, whereas data swapping [11] and the post-randomization method (PRAM) of Gouweleeuw, et al. [40] involve "random" perturbations. One way to think about random perturbation methods is as a restricted simulation tool, and thus we can link them to other types of simulation approaches that have recently been proposed.

Fienberg, Makov, and Steele [37] pursue a related simulation strategy and present a general approach to "simulating" from a constrained version of the cumulative empirical distribution function of the data. In the case when all of the variables are categorical, the cumulative distribution function is essentially the same as the counts in the resulting cross-classification or contingency table. As a consequence, their simulation approach is equivalent to simulating from a constrained contingency table, given a specified

set of marginal totals, and replacing the original data by a randomly generated one drawn from the “exact” distribution of the contingency table under a log-linear model given its minimal sufficient statistics (see Section 6).

In 1993, Rubin [44] asserted that the risk of identity disclosure can be eliminated by the use of synthetic data (in his case using Bayesian methodology and multiple imputation techniques) since there is no direct functional link between the original data and the released data. Or said another way, we have no confidentiality problem since we have replaced all of the real individuals with simulated ones. Raghunathan, Reiter, and Rubin [43] provide details on the implementation of this approach. But with both simulation or multiple-imputation methodology, it is still possible that some simulated individuals may be virtually identical to original sample individuals in terms of their data values, or at least close enough that the possibility of probabilistic identity disclosure remains. Thus, one still needs to carry our checks for “near identification” of original individuals.

Another important feature of the statistical simulation approach is that information on the added variability associated with the transformation of the data is directly accessible to the user. For example in the Fienberg, Makov, and Steele approach for categorical data, anyone can begin with the reported table and information about the margins that are held fixed, and then run the Diaconis-Sturmfels [14] Metropolis algorithm to regenerate the full distribution of all possible tables with those margins. This then allows the user to make inferences taking into account the added variability from sampling in a form that is similar to the approach to inference in PRAM. Similarly, multiple imputations produce a direct measure of variability associated with the posterior distribution of the quantities of interest [43]. As a consequence, simulation and random perturbation methods represent a major improvement over cell suppression (see the description below) and data swapping. And they conform to a statistical principle of allowing the user of released data to apply standard statistical operations without being misled.

There has been considerable research on disclosure limitation methods for tabular data, especially in the form of multi-dimensional tables of counts (contingency tables). The most popular methods include a process known as cell suppression which systematically deletes the values in selected cells in the table, and collapsing categories (a form of aggregation). While cell suppression methods have been very popular with the U.S. government statistical agencies and they are useful for tables with non-negative magnitude entries rather than simply counts, they also have major drawbacks. First, there are not yet good algorithms for the methodology associated with high dimensional tables. But more importantly, the methodology systematically distorts the information about the cells in the table for users and as a

consequence it makes it difficult for secondary users to draw correct statistical inferences about the relationships among the variables in the table. For further discussion on cell suppression and extensive references, see the various chapters in Doyle et al. [24], especially the one by Duncan et al. [25].

A special example of collapsing involves summing over variables to produce marginal tables. Thus instead of reporting the full multi-way contingency table we might report one or more collapsed versions of it. The release of multiple sets of marginal totals has the virtue of allowing statistical inferences about the relationships among the variables in the original table using log-linear model methods (e.g., see Bishop, Fienberg, and Holland [5]). What is also intuitively clear from statistical theory is that, with multiple marginals, one may have highly accurate information about the actual cell entries in the original table, and thus we still need to investigate the possibility of disclosure as is the case with simulation methods.

While there are many papers that claim to choose parameters or settings in disclosure limitation procedures in a way to minimize “information loss”, e.g., see Willenborg and de Waal [54], or maximize data utility, e.g., see Dandekar [12] and Cox, Kelly, and Patil [10], these methods rarely address the analytical utility of the released data and at best focus on “preserving a small set of data summaries from the original database.”

There are in fact formal frameworks for assessing the trade-offs between risk and utility where we measure utility in terms of formal quantities associated with inference. Duncan with a variety of coauthors has stressed a graphical representation for this trade-off which they call the R-U map, e.g., see [25, 26, 28]. They illustrate trade-off choices for disclosure limitation techniques such as adding noise to the data and topcoding (truncation) for variables like income. Trottni [51, 52, 53] takes the trade-off formalism several steps further and embeds it in a fully Bayesian decision theoretic framework. A key feature of both of these approaches is that one makes different choices for different uses and these inevitably depend on a formal statistical modeling and inference framework. In this paper we adopt a somewhat more informal assessment process but within a similar modeling framework. For a somewhat different perspective on this problem using aggregating as a disclosure limitation method, see Chawla et al. [6].

3 Some Basic Theory for Log-Linear Models

Here we summarize the basic statistical theory that we use for both determining the usefulness of partial releases of multi-dimensional contingency table data as well as assessing the disclosure risk of such releases. Indeed, part of the message of the paper is that these two uses of marginals are inextricably intertwined.

Consider a $2 \times 2 \times 2$ table of observer counts $\{n_{ijk}\}$, with corresponding estimated expected values, $\{m_{ijk}\}$, under a standard sampling model such as multinomial, product-multinomial, or Poisson—see [3, 5]):

m_{111}	m_{121}	m_{112}	m_{122}
m_{211}	m_{221}	m_{212}	m_{222}

The general log-linear model for $\{m_{ijk}\}$ takes the form

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}, \quad (2)$$

where each subscripted u -term sums to zero over any subscript, e.g.,

$$\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0. \quad (3)$$

We get *unsaturated* models from (2) by setting sets of u -terms equal to zero, e.g., if we set

$$u_{123} = 0 \text{ for all } i, j, k, \quad (4)$$

we have the model of no second-order interaction which is equivalent to equating the values of the odds ratio in each layer of the table, i.e.,

$$\frac{m_{111}m_{221}}{m_{121}m_{211}} = \frac{m_{112}m_{222}}{m_{122}m_{212}}. \quad (5)$$

The *minimal sufficient statistics* (MSSs) are the two-dimensional marginal totals, $\{n_{ij+}\}$, $\{n_{i+k}\}$, and $\{n_{+jk}\}$ (except for linearly redundant statistics included for purposes of symmetry), where a “+” indicates summation over the corresponding subscript. The MLEs of the $\{m_{ijk}\}$ under model (4) must satisfy the likelihood equations,

$$\begin{aligned} \hat{m}_{ij+} &= n_{ij+}, & i, j &= 1, 2, \\ \hat{m}_{i+k} &= n_{i+k}, & i, k &= 1, 2, \\ \hat{m}_{+jk} &= n_{+jk}, & j, k &= 1, 2, \end{aligned} \quad (6)$$

usually solved by some form of iterative procedure. When the data are generated by a product-multinomial sampling model one set of set of equations may be fixed by design, e.g., with the third set of equations in (6) corresponds to binomial sampling constraints when we are given the totals in the two-way margin for variables 2 and 3.

For higher-dimensional tables the structure of the log-linear model in equation 2 generalizes directly, with subscripted u -terms for all possible subsets of variables in the table. We typically work only with hierarchical log-linear models such that when we set a u -term equal to zero, all of its higher or relatives are also equal to zero, e.g., in a 3-dimensional table,

$$u_{12} = 0 \text{ for all } i, j \Rightarrow u_{123} = 0 \text{ for all } i, j, k. \quad (7)$$

The MSSs correspond to the highest order u -terms in the model and the likelihood equations are found by setting them equal to their expectations. As a short hand notation we describe a log-linear model in terms of its minimal sufficient statistics, e.g., [12][13][23] corresponds to the no second-order interaction model in three dimensions.

Graphical log-linear models are those satisfying a set of conditional independence relationships. Consider a k -dimensional table corresponding to k discrete random variables $\mathbf{X} = (X_1, X_2, \dots, X_k)$ and let $K = \{1, 2, \dots, k\}$ be the corresponding set of vertices. Further, let $\mathbf{X}_{\mathbf{K}/\{i,j\}}$ denote the set of $k - 2$ variables excluding X_i and X_j . A standard notation for representing the conditional independence of X_i and X_j given the remaining variables is

$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\mathbf{K}/\{i,j\}}. \quad (8)$$

This conditional independence model is also itself a log-linear model for the k -dimensional table. Graphical log-linear models correspond to the simultaneous occurrence of several such conditional independence models. They have special attractive properties in addition to the interpretation associated with these conditional independence relationships (for details, see Lauritzen [41]) and they can be depicted, as the name suggests, using graphs.

The conditional independence graph of \mathbf{X} is the undirected graph $\mathcal{G} = (K, E)$ such that the edge between i and j is not in the edge set E if and only if $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\mathbf{K}/\{i,j\}}$. Given an independence graph, \mathcal{G} , the corresponding graphical log-linear model is the one in which all u -terms containing the pairs of coordinates corresponding to edges not in E are taken to be identically zero.

When the graph corresponding to a graphical log-linear model is triangulated, the model is said to be decomposable and the estimated expected values can be written directly as a function of the expectations of the MSSs, and thus they have a simple closed-form expression that is easy to compute.

If $\hat{\mathbf{m}}$ is the MLE of \mathbf{m} under a log-linear model, and if the model is correct, then the likelihood ratio statistic,

$$G^2 = 2 \sum_{i=1}^t n_i \log \left(\frac{n_i}{\hat{m}_i} \right), \quad (9)$$

has an asymptotic χ^2 distribution with $t - s$ degrees of freedom, where s is the total number of independent constraints implied by the log-linear model and the multinomial sampling constraints (if any). If the model is not correct, then G^2 is stochastically larger than χ_{t-s}^2 . Thus we can use G^2 to assess the goodness-of-fit of the model.

To compare two log-linear models, $M^{(1)}$ and $M^{(2)}$, where $M^{(1)}$ is a special case of $M^{(2)}$ found by setting additional u -terms equal to zero, we difference the likelihood

ratio statistics, and this difference takes a special form that only involves the estimated expected values:

$$\begin{aligned}\Delta G^2 &= 2 \sum_{i=1}^t n_i \log \left(\frac{n_i}{\hat{m}_i^{(1)}} \right) - 2 \sum_{i=1}^t n_i \log \left(\frac{n_i}{\hat{m}_i^{(2)}} \right) \\ &= 2 \sum_{i=1}^t \hat{m}_i^{(2)} \log \left(\frac{\hat{m}_i^{(2)}}{\hat{m}_i^{(1)}} \right).\end{aligned}\quad (10)$$

The last line of equation (10) follows from the multiplicative representation of the models and the the form of the likelihood equations. Since we can estimate the estimated expected values from the MSSs (marginals) associated with a model and assess the fit of that model using only the margins of another model, the release of partial information in the form of marginals from a multi-way contingency table is a viable method for providing useful information to statistical analysts.

4 Partial Releases of Margins from Tables

A number of researchers have recently been working on the problem of determining upper and lower bounds on the cells of a multi-way table given a set of margins, in part to address this problem, although other measures of risk may clearly be of interest. The problem of computing bounds is in one sense an old one (at least for two-way tables). For simplicity, we refer to these as *Fréchet bounds* after the French statistician M. Fréchet who derived them in the context of cumulative probability distributions, although they were independently described by both Bonferroni and Hoeffding at about the same time in 1940. Fréchet bounds and their generalizations lie at the heart of a number of different approaches to disclosure limitation including cell suppression, data swapping and other random perturbation methods, and controlled rounding. Here we use them directly for assessing disclosure risk in contingency tables.

Fréchet bounds for the cell entries in an $I \times J$ table with entries $\{n_{ij}\}$ and row margins $\{n_{i+}\}$ and column margins $\{n_{+j}\}$ are given by

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}.$$

These bounds are sharp.

Consider a situation where instead of releasing a full k -dimensional contingency table, we release a set of lower-dimensional marginal totals from it. Any contingency table with non-negative integer entries and fixed marginal totals is a lattice point in the convex polytope \mathbf{Q} defined by the linear system of equations induced by the released marginals. The constraints given by the values in the released marginals induce upper and lower bounds on the interior cells of the initial table. These bounds or *feasibility intervals* can be obtained by solving the corresponding linear programming

problems. The importance of systematically investigating these linear systems of equations should be readily apparent. If the number of lattice points in \mathbf{Q} is below a certain threshold, we have significant evidence that a potential disclosure of the entire dataset might have occurred. Moreover, if the induced upper and lower bounds are too tight or too close to the actual sensitive value in a cell entry, the information associated with the individuals classified in that cell may become public knowledge. For simplicity, we consider a set margins \mathcal{R} to be releasable if and only if the minimum difference between the upper and lower bounds for the small count cells of “1” or “2” in table \mathbf{n} is greater or equal to some threshold.

That marginal tables are in fact minimal sufficient statistics for log-linear models turns out to provide a crucial to the development of efficient ways to calculate bounds.

4.1 Decomposable and Reducible Cases

As we noted above, decomposable graphical models have closed form structure and special properties. The expected cell values can be expressed as a function of the fixed marginals. To be more explicit, the maximum likelihood estimates are the product of the marginals divided by the product of the separators. By induction on the number of MSSs, Dobra and Fienberg [18] developed generalized Fréchet bounds for decomposable log-linear models with any number of MSSs. These bounds are sharp in the sense that they are the tightest possible bounds given the marginals. In addition, we can determine feasible tables for which these bounds are attained using from tools from algebraic geometry to which we return in a later section.

Theorem 1 (Fréchet Bounds for Decomposable Models)
Assume that the released set of marginals for a k -way contingency table is the set of MSSs of a decomposable log-linear model. Then the upper bounds for the cell entries in the initial table are the minimum of relevant margins, while the lower bounds are the maximum of zero, or sum of the relevant margins minus the separators.

When the log-linear model associated with the released set of marginals is not decomposable, it is natural to ask ourselves whether we could reduce the computational effort needed to determine the tightest bounds by employing the same strategy used for decomposable graphs, i.e. decompositions of graphs by means of complete separators. An independence graph that admits a proper decomposition but is not necessarily decomposable, is said to be *reducible* and a *reducible log-linear model* in [18] is one for which the corresponding MSSs are marginals that characterize the components of a reducible independence graph. If we can calculate the maximum likelihood estimates for the log-linear models corresponding to every component of a reducible

graph \mathcal{G} , then we can easily derive explicit formulae for the maximum likelihood estimates in the reducible log-linear model with independence graph \mathcal{G} [18].

Theorem 2 (Bounds for Reducible Models) *Assume that the released set of marginals is the set of MSSs of a reducible log-linear model. Then the upper bounds for the cell entries in the initial table are the minimum of upper bounds of relevant components, while the lower bounds are the maximum of zero, or sum of the lower bounds of relevant components minus the separators.*

There is one more special case worthy of attention in the context of the example in this paper. In general, when we are given $(k - 1)$ -way marginal tables are given, the corresponding independence graph is complete, hence there are no conditional independence relationships to exploit. If the table is dichotomous, the log-linear model of no k th-order interaction has only one degree of freedom and consequently the counts in any cell can be uniquely expressed as a function of one single fixed cell alone [34]. By imposing the non-negativity constraints for every cell in our contingency table, we are then able to derive sharp upper and lower bounds.

4.2 A General Bounds Algorithm

For decomposable and reducible graphs, we took advantage of the special structure of the conditional independencies “induced” among the variables cross-classified in a table of counts by the set of fixed marginals. When the released margins correspond to a log-linear model that is neither decomposable nor reducible, we need a more elaborate form of bounds calculation. Dobra [15, 20] has developed an iterative algorithm for this situation, a variation on a branch-and-bound algorithm, and the algorithm sequentially improves the bounds for all the cells until no further adjustment can be made. Because of its structure it produces the bounds in the decomposable case without iteration and also takes advantage of the decomposition associated with regular graphs described above.

Others have used versions of network algorithms, the simplex method, and other LP algorithms to compute such bounds, but, as Dobra [15] and Cox [9] have observed, LP algorithms often produce fraction bounds for problems where we know that the bounds must be integer in value. Simply rounding may not suffice, and Sullivant [50] using results from algebraic geometry, has recently shown that, as the dimensionality of the table grows, the gap between the true sharp bounds and the LP solution cannot be bounded.

Unfortunately, as the dimensionality of the table grows, Dobra’s iterative algorithm is computationally elaborate and is not especially useful as the main component of a search for an optimal form of marginal release. Dobra et al. [17]

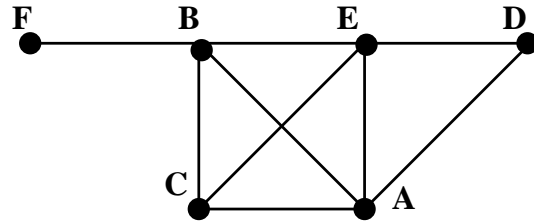


Figure 1. Independence graph for six-dimensional table induced by the marginals [ABCE], [ADE], and [BF].

suggest some simplified search strategies based on decomposable bounds and then use the algorithm only after focusing in on a small subset of sets of marginal releases. they apply these ideas to a large sparse 2^{16} table drawn from the National Long Term Care Survey.

4.3 Example: Prognostic Risk Factors for Czech Auto Workers

The data in Table 1 come from a prospective epidemiological study of 1841 workers in a Czechoslovakian car factory, as part of an investigation of potential risk factors for coronary thrombosis (see Edwards and Havranek [29]). Prior analyses of these data can be found in Dobra and Fienberg [19, 20] and Dobra et al. [21]. Here we integrate those results and, in the section that follows, expand upon them focusing in particular on issues of data utility and the risk-utility tradeoff.

In left-hand panel of Table 1, A indicates whether or not the worker “smokes,” B corresponds to “strenuous mental work,” C corresponds to “strenuous physical work,” D corresponds to “systolic blood pressure,” E corresponds to “ratio of β and α lipoproteins,” and F represents “family anamnesis of coronary heart disease.” Our focus for disclosure limitation is on the three cells in the table with counts of “1” and “2”. Because the data include all of the workers in the factory we in essence have a population and we *act* here as if the variables describing the dimensions of the table are key variables that could be used by intruders for record linkage, giving them possible access to other information on the workers not included in this particular table.

Decomposable Bounds Example. Suppose we are given three marginal tables, [ABCE], [ADE], and [BF]. from this 6-way table. These are the marginals corresponding to a graphical model whose independence graph is given in Fig. 1. Of the 15 possible edges in the graph 6 are absent.

Table 1. Prognostic factors for coronary heart disease as measured on Czech autoworkers. Source: [29]. The left-hand panel contains the original counts and the right-hand panel the bounds for the margins [ABCE], [ADE], and [BF].

F	E	D	C	B				B			
				A	no	yes	no	yes	A	no	yes
neg	< 3	< 140	no	44	40	112	67	[0,88]	[0,62]	[0,224]	[0,117]
			yes	129	145	12	23	[0,261]	[0,246]	[0,25]	[0,38]
		≥ 140	no	35	12	80	33	[0,88]	[0,62]	[0,224]	[0,117]
			yes	109	67	7	9	[0,261]	[0,151]	[0,25]	[0,38]
	≥ 3	< 140	no	23	32	70	66	[0,58]	[0,60]	[0,170]	[0,148]
			yes	50	80	7	13	[0,115]	[0,173]	[0,20]	[0,36]
		≥ 140	no	24	25	73	57	[0,58]	[0,60]	[0,170]	[0,148]
			yes	51	63	7	16	[0,115]	[0,173]	[0,20]	[0,36]
pos	< 3	< 140	no	5	7	21	9	[0,88]	[0,62]	[0,126]	[0,117]
			yes	9	17	1	4	[0,134]	[0,134]	[0,25]	[0,38]
		≥ 140	no	4	3	11	8	[0,88]	[0,62]	[0,126]	[0,117]
			yes	14	17	5	2	[0,134]	[0,134]	[0,25]	[0,38]
	≥ 3	< 140	no	7	3	14	14	[0,58]	[0,60]	[0,126]	[0,126]
			yes	9	16	2	3	[0,115]	[0,134]	[0,20]	[0,36]
		≥ 140	no	4	0	13	11	[0,58]	[0,60]	[0,126]	[0,126]
			yes	5	14	4	4	[0,115]	[0,134]	[0,20]	[0,36]

Table 2. Bounds for Czech auto-workers data from Table 1 given the marginals [BF], [BC], [BE],[AB], [AC], [AE], [CE], [DE], [AD]

F	E	D	C	B			
				A	no	yes	no
neg	< 3	< 140	no	[0,206]	[0,167]	[0,404]	[0,312]
			yes	[0,421]	[0,463]	[0,119]	[0,119]
		≥ 140	no	[0,206]	[0,167]	[0,404]	[0,312]
			yes	[0,416]	[0,333]	[0,119]	[0,119]
	≥ 3	< 140	no	[0,181]	[0,167]	[0,333]	[0,339]
			yes	[0,314]	[0,344]	[0,119]	[0,119]
		≥ 140	no	[0,181]	[0,167]	[0,363]	[0,339]
			yes	[0,314]	[0,341]	[0,119]	[0,119]
pos	< 3	< 140	no	[0,134]	[0,134]	[0,126]	[0,126]
			yes	[0,134]	[0,134]	[0,119]	[0,119]
		≥ 140	no	[0,134]	[0,134]	[0,126]	[0,126]
			yes	[0,134]	[0,134]	[0,119]	[0,119]
	≥ 3	< 140	no	[0,134]	[0,134]	[0,126]	[0,126]
			yes	[0,134]	[0,134]	[0,119]	[0,119]
		≥ 140	no	[0,134]	[0,134]	[0,126]	[0,126]
			yes	[0,134]	[0,134]	[0,119]	[0,119]

Since the graph in Fig. 1 is triangulated or decomposable, we can apply Theorem 1, to compute the upper and lower bounds for the cell entries induced by the marginals [BF], [ABCE], and [ADE], i.e., we get the upper bounds by computing the minimum of the corresponding entries in the fixed marginals, while the lower bounds are the sum of the same entries minus the sum of the corresponding entries in the marginals associated with the separators of the independence graph, [B] and [AE] (see the right-hand panel of Table 1). The bounds corresponding to the three small counts of “1” and “2” are [0,25], [0,38] and [0,20]. All three of these pairs of bounds differ quite substantially and thus we might conclude that there is little chance of identifying the individuals in the small cells.

Reducible Bounds Example. Now we step back and look at an even less problematic release involving the margins: [BF], [BC], [BE], [AB], [AC], [AE], [CE], [DE], [AD]. The independence graph associated with this set of marginals is the same as the graph in Fig. 1, but the log-linear model whose MSSs correspond to those marginals is not graphical. Since the independence graph decomposes into three components, [BF], [ABCE], and [ADE], and there are two separators, [B] and [AE], we can apply the result from Theorem 2.

The first component, [BF], is maximal and hence there is nothing to be done. The other two components are not maximal, however, and we need to compute upper and lower bounds for each of them using the general algorithm. For example, We calculated the marginal [ADE] given the marginals [AE], [DE], and [AD] using the simple approach for 2^k tables given their $(k - 1)$ -way margins with $k = 3$ to get the following bounds:

E	D	A	no	yes	A	no	yes
< 3	no	333	312		[182,515]	[130,463]	
	yes	265	151		[83,416]	[0,333]	
≥ 3	no	182	227		[0,333]	[76,409]	
	yes	181	190		[30,363]	[8,341]	

We next use the general algorithm to calculate bounds for the cell entries in the marginal [ABCE] given the marginals [BC], [BE],[AB], [AC], [AE], and [CE] as follows:

E	C	B		no		yes	
		A	no	yes	no	yes	
< 3	no		[0,206]	[0,167]	[0,404]	[0,312]	
	yes		[0,421]	[30,463]	[0,119]	[0,119]	
≥ 3	no		[0,181]	[0,167]	[0,363]	[0,339]	
	yes		[0,314]	[0,344]	[0,119]	[0,119]	

Since we have upper and lower bounds for each of the components of a reducible graph, Theorem 2 allows us to

piece together the bounds for the components [BF], [ABCE] and [ADE] to obtain the sharp integer bounds in Table 2 for the original 6-way table. Note that while some of the lower bounds for the 3-way marginal component [ADE] are non-zero, when we combine them with the other components the resulting lower bounds are all zero.

Theorem 2 in essence provides us with a method for replacing the original problem, namely, computing bounds for a 6-way table, by two smaller ones, i.e., computing bounds for a 4-way and a 3-way table. The computational effort required to use Theorem 2 is minimal once bounds for the components are available, and thus exploiting it in this fashion could lead to appreciable computational savings, especially when we consider large sparse tables.

Another Bounds Example. Finally, suppose we consider the release of all 5-way margins of Table 1, the space of tables \mathbf{Q} contains only two integer tables: the original table \mathbf{n} itself and a second table whose entries are found by adding or subtracting one unit from the corresponding entries in \mathbf{n} . Consequently, the feasibility intervals $[L(t), U(t)]$ for all the cells are of length one. This means that releasing all 5-way margins could well compromise the confidentiality of the individuals corresponding to the entries containing counts of “1” and “2”.

5 The Risk-Utility Tradeoff: Deciding What to Release

Rather than applying a full-scale version of the RU-map or the criteria proposed by Trottni, we approach the risk-utility tradeoff in the context of the example using the simple tools involving bounds for cell entries.

It turns out that there are more than 32,000 decomposable models for a 6-dimensional table. We computed the bounds for all of these using the result in Theorem 1, and ran ordered the models based on the feasibility intervals for the three target cells. The tightest bounds are (0,3), (0,6), and (0,3), and these are attained for 31 models, all of which involve the margin [ACDEF]. Another 30 models have bounds for these cells that differ by 5 or less and these involve [ABCDE] as well as possibly [ACDEF]. Thus the two 5-dimensional margins [ACDEF] and [ABCDE] appear to be the “source” of the disclosure risk regarding the small cell entries.

Suppose we release the remaining four 5-way margins and the one 4-way margin not implied by them, i.e., $\mathcal{R}^*=[ACDE][ABCDF][ABCEF][BCDEF][ABDEF]$. For these margins the feasible intervals for all cells are 10 or more and it is not unreasonable to conclude that an intruder would not be able to use the information in them to achieve accurate linkage to other databases.

What can the user do with these data? In fact, the user can fit all “reasonable” log-linear models to the data and assess their fit using the methods outlined in Section 3, in particular the formula for ΔG^2 in (10), and one or more model search criteria. In particular, the user will be able to determine that the model with MSSs [ABCE], [ADE], and [BF] fits the data extremely well, and corresponds to the model singled out by several widely-used search criteria such as BIC. Thus we have achieved a sensible combination of low disclosure risk AND high utility.

Finally, suppose that we accompany the release of \mathcal{R}^* with an announcement that the released margins are adequate for any user to make correct inferences from the data, i.e., the inferences should be same as those that users would make had the the entire table been released. This is extra information that the intruder can use to eliminate some of the possible tables in the convex polytope corresponding to the marginal constraints, and thus in principle tighten the bounds for the cell entries. Fortunately such conditioning will not materially reduce the space of possible tables for intruder’s inferences in this example.

Others have suggested different approaches to the release of tabular data. For example Cox et al. [10] and Dandekar [12] propose adaptations of a method that extends the notion of cell suppression which they refer to as *Controlled Tabular Adjustment* or CTA. While these methods may be useful for tables of magnitudes, it remains unclear how they could be used in a contingency table context to generate data releases that yield proper statistical inferences even in restricted circumstances, despite the claims in [12].

6 Perturbation Maintaining Marginal Totals

While we have focused here on the generation of bounds for tables with given marginals, there is a parallel statistics literature on the generation of distributions over the corresponding space of tables using Markov bases from the algebraic geometry representation for the toric ideal of polynomial rings. Diaconis and Sturmfels [14] give the basic structure for this representation and discuss the role of generators for the algebraic geometry structure. They explain how to generate such bases and propose a version of the Metropolis algorithm for generating the exact distribution of a log-linear model given its minimal sufficient statistic margins. For applications of their methods to disclosure limitation see [37, 31]. Fienberg and McIntyre [38] explain why this can be thought of as the natural representation for the limiting form of data swapping or constrained perturbation for contingency tables. Once we generate the Markov basis for such a contingency table problem, we can use the elements of it as moves that allow us to traverse the complete space of tables satisfying the constraints. Alternatively we can use Markov the basis in a computational algebra computing

package to count the tables. The difficult task for large contingency tables is the generation of the Markov basis.

For each of the special cases described in Section 3 where the calculation of sharp bounds for the cell entries does not require elaborate computation, there is a parallel simple form for the Markov basis where the generators, or tables of moves, only contain elements that equal 1, 0 or -1. e.g., see Dobra [16] and Dobra and Sullivant [22]. While one might expect that there would be simple bases for any log-linear model for a 2^k table, Aoki and Takemura [4] have demonstrated that this is not the case.

Other surprises has arisen from this nascent algebraic statistics literature. Aoki and Takemura [4] have further demonstrated the separation of components of the space of tables when we are given sets of marginals and De Loera and Onn [13] have shown that for some tables all values between the upper and lower bounds discussed above may not be possible. These “gaps” mean that there exist cell values between the upper and lower bounds for which no tables satisfy the marginal constraints. These results argue for going beyond bounds in assessing risk.

A somewhat different way to think about assessing risk, related to the calculation of bounds and implementing perturbations is simply counting the numbers of possible tables instead of putting arbitrary distributions over them. As we saw in the example, for the release of all 5-way marginals of Table 1, the space of tables \mathcal{Q} contains only two integer tables, one of which is the original table. As we releasing fewer higher order marginals the number of possible tables grows, often dramatically, thus making difficult for an intruder to identify the original table. In principle some of the specialized computer algebra software packages can be used to do these calculations, but for our example the computational burden is very high and we have been unable to implement the calculations to date.

One way to calculate the number of tables is to run the Metropolis algorithm as just mentioned. Another is to adapt Dobra’s general bounds algorithm to actually count the tables. Dobra et al. [23] use this approach report that for the released margins,

$$[ACDEF][ABDEF][ABCDE][BCDF][ABCF][BCEF],$$

there are 810 possible tables, whereas for the release of all of the 15 two-way margins there are 705,884 possible tables. Clearly for an intruder, picking out the original table from such a large number of tables would be like finding a needle in a haystack. Dobra et al. [23] and Chen et al.[7] describe different approaches to approximating these counts that are especially useful for approximation purposes.

Of course, an intruder might by accident happen across a table that accurately included a count of 1 in the (1,2,2,1,1,2) cell, but which didn’t get any of the other original counts correct. Dobra et al. also explore this issue using

simulations instead of exact counting.

It seems reasonably clear that, at least in this example, none of the alternative approaches discussed here suggest radically different assessments of disclosure risk and that however we choose to apply them, we will allow analysts to focus on the same models they would have identified using the full table.

7 Global Recoding: Combining Categories for Contingency Tables

Up to this point we have focused on disclosure limitation through the form of collapsing that involves summing over variables to produce marginal tables. Thus instead of reporting the full multi-way contingency table we might report one or more collapsed versions of it. In Section 2 we noted that a popular disclosure limitation technique for categorical variables was “global recoding” or combining categories. This too is a special form of collapsing that involves summing over categories within variables to produce a new k -dimensional table (unless of course we drop a variable!) with a reduced number of cells. This yields a *single releasable* marginal table.

The problem with global recoding is that it restricts the form of the analysis one can do—i.e., the data analyst had no ability to “inferentially” recover information about relationships in the full table using log-linear model methods. When the global recoding involves arbitrary combinations of categories for variables, this clearly can lead the analyst to inferential errors.

When the categorical variables are ordinal in nature, i.e., the categories have an explicit or implicit ordering, there are related models that may allow for proper inferences, e.g., those described by McCullagh [42], and the association models of Leo Goodman.

An alternative to the strategy of releasing a single collapsed table following global recoding, is to release multiple such tables using different choices of global recodings. If done with care, this would allow for methodologies for disclosure risk assessment and model fit parallel to those described in this paper. To our knowledge, no one has pursued this strategy.

8 Conclusions

In this paper, we have focused on the interplay between the issues of confidentiality, on the one hand, and access to statistical data bases on the other. Disclosure limitation is an inherently statistical issue because we cannot eliminate the risk of disclosure unless we allow no access to the data. One popular alternative strategy is to restrict access to the data to “approved individuals,” which is strictly speaking also a

form of disclosure albeit controlled, but it requires screening of abstracted information for broader release. Several chapters in the volume edited by Doyle et al. [24] describe several variants on the restricted access model in current use. Here we have focused on “restricted data” model of disclosure limitation instead of “restricted access.”

Because techniques for disclosure limitation are inherently statistical in nature we explained why they must be evaluated using statistical tools for assessing the risk of harm to respondents. We then outlined some of the current statistical methods used to limit disclosure, especially those representable in form of disclosure limitation masks. We illustrated the trade-off issues by reviewing methods for the release of partial information in the form of marginal tables from a k -dimensional contingency table.

Similar issues arise when we consider other statistical disclosure limitation methods for tabular magnitude data and microdata. Although there is not yet a consensus on how to evaluate disclosure limitation methods, there are at least three desirable features or principles for which we need to strive:

- *Usability*—the extent to which the released data are free from systematic distortions that impair statistical methodology and inference.
- *Transparency*—the extent to which the methodology and practice of it provide direct or even implicit information on the bias and variability resulting from the application of a disclosure limitation mask or other methodology.
- *Duality*—the extent to which the methods aim at both disclosure limitation and making the maximal amount of data available for analysis.

In this paper, we described how these principles fit with recent proposals for the release of marginals from multi-way contingency tables and the role of marginal bounds in evaluating the disclosure limitation possibilities. Even for contingency tables, the notion of partial releases need not be restricted to margins. For extensions to the release of margins and conditional tables, see the work of Slavkovic [48, 49] and Fienberg and Slavkovic [39, 32].

These principles and the special nature of the confidentiality-utility tradeoff associated with statistical databases pose at least two different sets of challenges for datamining. Can datamining tools be used to extract innovative forms of information from databases to which disclosure limitation methods have been applied? Can datamining methods lead to better forms for data protection?

Acknowledgments

The research reported here was supported in part by NSF grants EIA-9876619 and IIS-0131884 to the National Institute of Statistical Sciences, as well as by Grant R01-AG023141 from the NIH to the Department of Statistics and by Army contract DAAD19-02-1-3-0389 to CyLab, both at Carnegie Mellon University. I have benefited from extensive conversations with Adrian Dobra, George Duncan, Alan Karr, Steve Roehrig, Ashish Sanil, Aleksandra Slavkovic, Seth Sullivant, and Mario Trottni about the methods described here and the computations associated with the example. They bear no responsibility, however, for how I have represented their input.

References

- [1] N. R. Adam and J. C. Wortman. Security control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21:516–556, 1989.
- [2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, Dallas, Texas, 2000.
- [3] A. Agresti. *Categorical Data Analysis*. Wiley, 2nd edition, 2002.
- [4] A. Aoki and A. Takemura. Invariant minimal Markov basis for sampling contingency tables with fixed marginals. METR Technical Report, 03-25, 2003.
- [5] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. M.I.T. Press, Cambridge, MA, 1975.
- [6] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Towards privacy in public databases. unpublished manuscript, 2004.
- [7] Y. Chen, I. H. Dinwoodie, and S. Sullivant. Sequential importance sampling for multiway tables. submitted for publication, 2004.
- [8] L. H. Cox. A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82:520–524, 1987.
- [9] L. H. Cox. On properties of multi-dimensional statistical tables. *Journal of Statistical Planning and Inference*, 117:251–273, 2003.
- [10] L. H. Cox, J. P. Kelley, and R. Patil. Balancing quality and confidentiality for multivariate tabular data. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Data Bases*, volume 3050 of *Lecture Notes in Computer Science*, pages 87–98, 2004.
- [11] T. Dalenius and S. P. Reiss. Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6:73–85, 1982.
- [12] R. A. Dandekar. Maximum utility-minimum information loss table server design for statistical disclosure control of tabular data. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Data Bases*, volume 3050 of *Lecture Notes in Computer Science*, pages 121–135, 2004.
- [13] J. De Loera and S. Onn. All linear and integer programs are slim 3-way transportation programs. unpublished manuscript, 2004.
- [14] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26:363–397, 1998.
- [15] A. Dobra. *Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables*. PhD thesis, Department of Statistics, Carnegie Mellon University, 2002.
- [16] A. Dobra. Markov bases for decomposable graphical models. *Bernoulli*, 9:1–16, 2003.
- [17] A. Dobra, E. Erosheva, and S. E. Fienberg. Disclosure limitation methods based on bounds for large contingency tables with application to disability data. In H. Bozdogan, editor, *Proceedings of Conference on the New Frontiers of Statistical Data Mining*, pages 93–116. CRC Press, 2003.
- [18] A. Dobra and S. E. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences*, 97:11885–11892, 2000.
- [19] A. Dobra and S. E. Fienberg. Bounds for cell entries in contingency tables induced by fixed marginal totals. *Statistical Journal of the United Nations ECE*, 18:363–371, 2001.
- [20] A. Dobra and S. E. Fienberg. Bounding entries in multi-way contingency tables given a set of marginal totals. In Y. Haitovsky, H. Lerche, and Y. Ritov, editors, *Foundations of Statistical Inference: Proceedings of the Shores Conference 2000*, pages 3–16. Berlin, 2003. Springer-Verlag.
- [21] A. Dobra, A. Karr, A. Sanil, and S. E. Fienberg. Software systems for tabular data releases. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10:529–544, 2002.
- [22] A. Dobra and S. Sullivant. A divide-and-conquer algorithm for generating Markov bases of multi-way tables. *Computational Statistics*, 18:to appear, 2003.
- [23] A. Dobra, C. Tebaldi, and M. West. Data augmentation in multi-way contingency tables with fixed marginal totals. submitted for publication, 2004.
- [24] P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, editors. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, Amsterdam, 2001.
- [25] G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig. Disclosure limitation methods and information loss for tabular data. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 135–166. Elsevier, Amsterdam, 2001.
- [26] G. T. Duncan, S. Keller-McNulty, and S. L. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map. *Management Science*, page to appear, 2004.
- [27] G. T. Duncan and R. B. Pearson. Enhancing access to microdata while protecting confidentiality: Prospects for the future (with discussion). *Statistical Science*, 6:219–239, 1991.
- [28] G. T. Duncan and S. L. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map as applied to topcoding. *Chance*, 17(3):16–20, 2004.
- [29] D. E. Edwards and T. Havranek. A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72:339–351, 1985.

- [30] Federal Committee on Statistical Methodology. *Report on Statistical Disclosure Limitation Methodology*. Statistical Policy Working Paper 22. Subcommittee on Disclosure Limitation Methodology. Office of Management and Budget, Executive Office of the President, Washington, DC, 1994.
- [31] S. Fienberg, U. Makov, M. Meyer, and R. Steele. Computing the exact distribution for a multi-way contingency table conditional on its marginals totals. In A. K. M. E. Saleh, editor, *Data Analysis from Statistical Foundations: Papers in Honor of D. A. S. Fraser*, pages 145–165. Nova Science Publishing, Huntington, NY, 2001.
- [32] S. Fienberg and A. B. Slavkovic. Making the release of confidential categorical data count. *Chance*, 17(3):5–10, 2004.
- [33] S. E. Fienberg. Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, 10:115–132, 1994.
- [34] S. E. Fienberg. Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation. In *Statistical Data Protection (SDP'98) Proceedings*, pages 115–129, Luxembourg, 1999. Eurostat.
- [35] S. E. Fienberg and U. E. Makov. Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, 14:485–502, 1998.
- [36] S. E. Fienberg, U. E. Makov, and A. P. Sanil. A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics*, 13:75–89, 1997.
- [37] S. E. Fienberg, U. E. Makov, and R. J. Steele. Disclosure limitation using perturbation and related methods for categorical data (with discussion). *Journal of Official Statistics*, 14:485–502, 1998.
- [38] S. E. Fienberg and J. McIntyre. Data swapping: Variations on a theme by Dalenius and Reiss. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Data Bases*, volume 3050 of *Lecture Notes in Computer Science*, pages 14–29, 2004.
- [39] S. E. Fienberg and A. B. Slavkovic. Bounds for cell entries in two-way tables given conditional relative frequencies. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Data Bases*, volume 3050 of *Lecture Notes in Computer Science*, pages 30–43, 2004.
- [40] J. M. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P. P. d. Wolf. Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14:463–478, 1998.
- [41] S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1996.
- [42] P. McCullagh. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42:109–142, 1980.
- [43] T. E. Raghunathan, J. Reiter, and D. B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16, 2003.
- [44] D. B. Rubin. Discussion statistical disclosure limitation. *Journal of Official Statistics*, 9:461–468, 2003.
- [45] S. M. Samuels. A bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics*, 14:373–383, 1998.
- [46] C. J. Skinner and M. J. Elliot. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B*, 64:855–867, 2001.
- [47] C. J. Skinner and D. J. Holmes. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14:373–383, 1998.
- [48] A. B. Slavkovic. *Disclosure Limitation Beyond the Margins: Characterization of Joint Discrete distributions for Contingency Tables*. PhD thesis, Department of Statistics, Carnegie Mellon University, 2004.
- [49] A. B. Slavkovic. Statistical disclosure limitation with released marginals and conditionals for contingency tables. In these proceedings, 2004.
- [50] S. Sullivant. Small contingency tables with large gaps. *SIAM Journal on Discrete Mathematics*, page to appear, 2004.
- [51] M. Trottni. A decision-theoretic approach to data disclosure problems. *Research in Official Statistics*, 4:7–22, 2001.
- [52] M. Trottni. *Decision Models for Data Disclosure Limitation*. PhD thesis, Department of Statistics, Carnegie Mellon University, 2003.
- [53] M. Trottni and S. Fienberg. Modelling user uncertainty for disclosure risk and data utility. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10:511–528, 2002.
- [54] L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*, volume 155 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2000.