Journal of Official Statistics, Vol. 28, No. 1, 2012, pp. 121–140

Cell Bounds in *k*-way Tables Given Conditional Frequencies

Byran J. Smucker[1], *Aleksandra Slavković*[2], *and Xiaotian Zhu*[3]

For data summarized and released as a contingency table, considerable attention has been accorded to cell bounds given marginal totals. Here, we consider bounds on cell counts for *k*-way tables when observed conditional probabilities and total sample size are released. If this information implies narrow bounds, a disclosure risk may result. We compute sharp integer bounds using integer programming and demonstrate that, in some cases, they can be unacceptably narrow. We also derive closed-form solutions for linear relaxation bounds, and show that they can be improved via a method that can also account for rounding uncertainty. The gaps between the sharp bounds and those of their linear relaxations are often large, which implies the utility of the latter is limited, especially if the sharp bounds can be computed quickly. Our formulations can solve small tables with small sample sizes quite quickly, but large instances can take on the order of hours.

Key words: Confidentiality; contingency tables; integer programming; linear programming; statistical disclosure control; tabular data.

1. Introduction

Data privacy has become an important problem of the modern society with both government and nongovernment agencies collecting, archiving and releasing a growing amount of personal and sensitive data. The statistical disclosure limitation (SDL) literature is rooted in official statistics and is concerned with increasing public data access and utility while maintaining data confidentiality. Data stewards wish to release data so meaningful statistical inference can be performed, but haphazard data releases are inappropriate because of confidentiality guarantees.

The goal of SDL is to develop methods and tools for evaluating the trade-off between privacy and the release of useful data. We consider the problem of establishing bounds on contingency table cells given tabular sample size and observed conditional probabilities (i.e., rates) derived from underlying contingency table of counts. These bounds, called *feasibility intervals* (Willenborg and de Waal 1996), are one of many ways to measure disclosure risk, which may result when cells have small counts

[1] Miami University, Department of Statistics, Oxford, OH 45056, U.S.A. Email: absmucker@gmail.com
[2] The Pennsylvania State University, Department of Statistics, University Park, PA 16802, U.S.A. Email: sesa@psu.edu
[3] The Pennsylvania State University, Department of Statistics, University Park, PA 16802, U.S.A. Email: xxz131@psu.edu
Acknowledgments: The research reported here was supported in part by NSF Grant SES-0532407 to the Department of Statistics, Pennsylvania State University. The first author also gratefully acknowledges Steven E. Wright of Miami University for his helpful discussion and assistance with the interfaces between Matlab and CPLEX.

with narrow bounds. Indeed, cell counts can even be directly disclosed when the upper and lower bounds are the same. Many tabular data releases are in the form of marginal tables (sums of rows and/or columns), but data stewards may publish observed conditional probabilities as an alternative way of releasing summary statistics, representing proportions of individuals who fall into a certain category given particular characteristics.

Feasibility intervals also have application beyond a direct measurement of disclosure risk. For instance, Chen et al. (2006) have used sequential importance sampling to simulate draws from the distribution of associated contingency tables, given marginal totals. This procedure may be used as a tool for inference for tabular data and requires bounds to be calculated. Slavković and Lee (2010) have recently proposed an MCMC approach to simulate draws from a distribution of contingency tables that preserveconditional probabilities, which would require priors that might be informed by these bounds.

Much of the recent work in SDL combines elements from multiple disciplines, including statistics and operations research (see a review in Salazar-Gonzalez 2008). In this article, we add to this confluence by considering integer programs which can be solved to produce sharp cell bounds given unrounded conditional probabilities and sample size. In addition, we assume throughout that no external sources of information are available. We find that these exact bounds can uniquely identify cell counts in certain situations and, even when unique identification does not occur, often produce lower bounds that are identical to the underlying count. However, when the released conditionals are rounded, the exact bounds often cannot be computed, possibly giving additional disclosure protection. We also derive closed-form expressions for the bounds in the case that the integer requirement for cells is relaxed. These results are extensions of Smucker and Slavković (2008), relying on the fact that a k-way table can be represented as a two-way table. As in Smucker and Slavković (2008), we use a mathematical programming formulation that allows for sampling zeros and satisfies the definition of conditional probability. Beyond that, we present an easily computable result which improves the linear relaxation bounds in two ways. First, if rounding uncertainty is ignored, the bounds will be tighter. Second, these bounds can account for the uncertainty introduced by rounding. Despite these improvements, we show empirically that the difference between the sharp integer bounds and those based on linear relaxations is most often dramatic. We also consider the situation in which partial conditional probabilities (defined later) are released instead of full conditionals.

Though data releasers have traditionally been large national agencies, smaller data owners may wish to share their collected data as well. Thus, it is important to understand the effects of these types of releases for small-scale, as well as large-scale, data. Consequently, we explore both small and large tables in this article, and the associated bounds given conditional probabilities and total sample size.

In Section 2, we give some background on optimization and review the literature as it pertains to contingency table cell bound calculation. In Section 3, we present the mathematical programming formulations and closed-form results for the linear relaxations, for both full and partial conditional frequencies, as well as improved bounds

that can account for rounding. We demonstrate their application with two examples in Section 4, and give a general discussion in Section 5.

2. Optimization and Cell Bound Calculation

An integer program (IP) consists of a linear objective function, optimized subject to linear constraints, with the additional constraint that all decision variables are integral. (We point out that in the optimization literature, decision variables are variables within a particular optimization structure whose values are to be manipulated in the course of the optimization; contrastingly, a random variable is a variable whose value is determined by some random process.) IPs are solved using methods like Branch-and-Bound and Branch-and-Cut (see Nemhauser and Wolsey 1988). In this work, the commercial solver CPLEX (2009) was used to solve the integer programs.

Calculating cell bounds for two-way tables given marginal totals is an old problem dating back to Bonferroni (1936), Fréchet (1940), and Hoeffding (1940). The extension to k-way tables has proven challenging (Cox 2002). Fienberg (1999) provides generalizations of the two-way bounds, and others have studied special cases (Dobra and Fienberg 2001; Cox 2002; 2007). Buzzigoli and Gusti (1998) introduced the "shuttle algorithm" which computes bounds—not necessarily sharp—for k-way tables given an arbitrary set of marginals. Dobra and Fienberg (2003; 2010) generalize this procedure, the latter in particular producing exact bounds.

Less work has been done for bounds given observed conditional probabilities. Using both mathematical programming and tools from algebraic statistics, Slavković and Fienberg (2004) and Fienberg and Slavkovic (2005) first examined these bounds (see also Dobra et al. 2008). Later, Smucker and Slavković (2008) presented an alternative (see Section 3) to this original optimization formulation, but only considered two-way tables. Slavković and Lee (2010) have utilized these bounds in conjunction with algebraic tools for creating synthetic two-way contingency tables, and for assessing both disclosure risk and data utility associated with such synthetic data releases. In this article, we make the extension to k-way tables, considering calculations given both full and partial conditional rates.

To calculate sharp bounds we use IP, as mentioned above, but this is often computationally expensive for large tables (see the end of Section 4.2). Thus we attempt to use linear relaxation bounds as a cheap approximation. Given the marginals, the maximal gap between an IP and its linear relaxation has been studied and theoretically has been shown to be exponentially large (Sullivant 2005; Hosten and Sturmfels 2003), and Onn (2006) shows that there could be arbitrary gaps within these bounds. For the marginal case, then, it could be misleading to assess disclosure risk by using the linear relaxation as an approximation to the integer bounds. In this article, we demonstrate that the same is true in the case of given conditional probabilities, and that for k-way tables is perhaps even more pronounced than in the two-way case.

3. Bounds for Cells In k-way Tables Given Conditional Probabilities

In this section we formulate the integer program used to compute sharp cell bounds, and also give easily computed expressions for linear relaxations of the bounds. We note here

that our concern is simply with the information available given a table of conditional probabilities and sample size. How the data are collected—for instance, whether zeros in the table are structural or arise via the chance of sampling—is beyond the scope of this work. We do note, however, that the formulations which follow can easily account for observed or structural zeros in the data, by simply setting those cells to zero and ignoring them.

3.1. Setting and Notation

Let $X = \{X_1, \ldots, X_k\}$ be a vector of categorical random variables and let $\{i_1, i_2, \ldots, i_k\}$ be the index sets corresponding to each of the random variables, where $i_1 = 1, \ldots, I_1$ (I_1 is the number of categories in the first random variable), $i_2 = 1, \ldots, I_2$, all the way up to $i_k = 1, \ldots, I_k$. Define $f_X(x)$ as the joint density of these variables and define mutually disjoint sets of indices I, J, and K, such that $I, J, K \subset \{i_1, \ldots, i_k\}$ and $I \cup J \cup K = \{i_1, \ldots, i_k\}$. Also let X_I be the vector of random variables corresponding to those represented in I, X_J those represented in J, and X_K those represented in K. X_I corresponds to those variables upon which we are conditioning, X_J corresponds to the response variables, and X_K are those variables that are not being considered at all.

For instance, in a four-way table there are four random variables—X_1, X_2, X_3, and X_4—with indices i_1, i_2, i_3, and i_4, where the last variable, X_4 corresponds to the response variable. In that case, we might condition upon the first three variables (corresponding to X_1, X_2, and X_3 indices) so that $X_I = \{X_1, X_2, X_3\}$, $X_J = \{X_4\}$, and $X_K = \emptyset$. Alternatively, we might have that $X_I = \{X_1, X_2\}$, $X_J = \{X_4\}$, and $X_K = \{X_3\}$ so that the given table is aggregated over X_3.

Using this notation we define *full* and *partial* conditional probabilities.

Definition 3.1 *Let $X_K = \emptyset$ and $O = \{o_{IJ}\}$ be the observed count for cell IJ. The observed full conditional probabilities are defined as*

$$\hat{d}_{IJ\cdot} = \frac{o_{IJ}}{o_{I\cdot}} = \frac{o_{IJ}}{\sum_{J'} o_{IJ'}}$$

which is an estimate of the actual conditional probability $P(X_J = x_J | X_I = x_I)$ where x_I and x_J are realizations of the associated random variables.

If X_K is nonempty, we collapse over the variables in X_K to get to a two-way table (see Section 4 for examples of this).

Definition 3.2 *The observed partial conditional probabilities, estimating $P(X_J|X_I)$, are defined as*

$$\hat{d}_{IJ\cdot} = \frac{\sum_{K'} o_{IJK'}}{\sum_{J'} \sum_{K'} o_{IJ'K'}} \tag{1}$$

3.2. Cell Bounds Based on Mathematical Programs

To calculate integer bounds given full conditionals and table sample size, the following integer program is constructed:

$$Min\, n_{AB} \tag{2}$$

$$s.t. \sum_{I'}\sum_{J'} n_{I'J'} = N \tag{3}$$

$$o_{IJ}\sum_{J'\neq J} n_{IJ'} + \left(o_{IJ} - \sum_{J'} o_{IJ'}\right) n_{IJ} = 0, \quad \forall\ I,\ \ J = 1, \ldots, I_J - 1 \tag{4}$$

$$\sum_{J'} n_{IJ'} \geq 1\ \ \forall\ I \tag{5}$$

$$n_{IJ} \geq 0\ \ \forall I, J \tag{6}$$

$$n_{IJ}\ \ \text{integer}\, \forall I, J \tag{7}$$

where AB represents a particular cell to be minimized, and in (4) I_J is the total number of categories in the response variable(s). To calculate the lower bound for cell AB, the above IP is solved; the upper bound is found by maximizing it. This process is repeated for each cell to obtain its bounds, which are known as sharp, or exact, because they are integer-constrained.

The first constraint, (3), is to enforce the sample size. We ensure positive marginal sums by (5), nonnegative cell entries by (6), and integer entries by (7). For (4), we use

$$\hat{d}_{IJ} = \frac{o_{IJ}}{\sum_{J'} o_{IJ'}} = \frac{n_{IJ}}{\sum_{J'} n_{IJ'}}$$

where o_{IJ} is the observed count and n_{IJ} is the decision variable for cell IJ used in the optimization program. To derive (4),

$$\begin{aligned} 0 &= \frac{o_{IJ}}{\sum_{J'} o_{IJ'}} - \frac{n_{IJ}}{\sum_{J'} n_{IJ'}} = o_{IJ}\sum_{J'} n_{IJ'} - n_{IJ}\sum_{J'} o_{IJ'} \\ &= o_{IJ}\sum_{J'\neq J} n_{IJ'} + o_{IJ}n_{IJ} - n_{IJ}\sum_{J'} o_{IJ'} \quad = o_{IJ}\sum_{J'\neq J} n_{IJ'} + n_{IJ}\left(o_{IJ} - \sum_{J'} o_{IJ'}\right) \end{aligned}$$

Note that (4), as formulated here, requires that o_{IJ}, the actual counts, be given. We are, of course, trying to bound the cells even as we assume that we know them. This is necessary, though, to calculate the sharp integer bounds for most datasets, because the released conditional probabilities, $\hat{d}_{IJ}$, are rounded and thus preclude a feasible optimal integer solution if used directly. Therefore, in most realistic cases intruders would be unable to calculate the bounds based on the above IP. Integer programming formulations which account for rounding may or may not provide similar bounds, but are beyond the scope of this article. The present formulation, however, can be used by data agencies to assess true disclosure risk.

We follow Smucker and Slavković (2008) and do not require a count of at least one in each cell for which a positive conditional probability is released (as done in Slavković and Fienberg 2004), instead requiring only that each row has a count of at least one (if rows are fully composed of zeros, they are skipped in the optimization). This results in wider bounds than those in Slavković and Fienberg (2004), but in Section 3.3 we offer improved, closed-form linear relaxation bounds that are similar to the tighter bounds of Slavković and Fienberg (2004) while accounting for rounding uncertainty.

As in Smucker and Slavković (2008), a closed-form expression can be developed for the linear relaxation to the integer program in (2)–(7). The case of k-way tables is much the same as that of two-way tables because by considering full conditional probabilities we have essentially collapsed the problem from a k-way table to a two-way table with dimensions $I_1 \cdot I_2 \cdot \ldots \cdot I_{k-1} \times I_k$ (this assumes the *kth* random variable is the response; if there is more than one response variable, the dimensions of the table will be $I_I \times I_J$, where I_I is the total number of categories in the variables upon which we are conditioning, and I_J is the total number of categories in the response variables). Thus, the proof is very similar to that in Smucker and Slavković (2008) and is omitted here. We define R to be the number of nonzero marginals (i.e., the number of rows with nonzero sums) in the two-way table that are constructed from the k-way table. If there are no nonzero marginals, $R = I_I$.

Theorem 3.3 *Assume we have a k-way contingency table. Based on the full conditional probabilities, $\hat{d}_{IJ}$ and the sample size, N, we can construct a linear program of the form (2)–(6). This linear program is minimized at $n_{AB} = \hat{d}_{AB}$ and maximized at $(N - (R - 1))\hat{d}_{AB}$. That is,*

$$\hat{d}_{AB} \leq n_{AB} \leq (N - (R - 1))\hat{d}_{AB} \tag{8}$$

If partial conditionals are given instead of full conditionals, the contingency table of interest will be of lower dimension, but otherwise it will have the same character. The table of partial conditionals can be flattened to a two-way table, and the IP (2)–(7) or its linear relaxation can be applied (see Section 4.1.2 and Section 4.2.2 for examples). This will result in bounds on the cells of the table of partial conditionals, not the underlying cells of the full table. However, once the bounds on the partial table are known, the bounds on the cells in the underlying full table follow immediately, because of two observations. First, given a table of partial conditional probabilities (i.e., a table with variables I and J, summed over K), the lower bound on each cell of the underlying table (i.e., the original table with all variables I, J, and K) is 0. Second, the upper bound for a cell in the partial table is the same as the upper bound for all corresponding cells in the underlying table. Consequently, the cell bounds on the underlying cells of the original contingency table (i.e., the cells defined by the variables in I, J, and K) can be completely specified by finding the upper bounds on the table defined by the partial conditional probabilities (i.e., the variables in I and J). We formalize this as follows.

Theorem 3.4 *Assume that K is nonempty, and let n^*_{AB} be an upper bound on cell AB in the table of partial conditionals.*

(a) The lower bound for any cell ABC is 0.
(b) The upper bound for any cell ABC is equal to n^*_{AB}.

Proof. First, write the partial table—defined by summing across the variables K—as a 2-way table with the J variables defining the columns and the I variables defining the rows. Thus, using (2)–(7), nonzero lower and upper bounds can be calculated for each cell in the partial table.

To show (a), notice that each cell in the partial table represents a sum over the categories defined by K. Thus, the lower bound on a cell AB in the partial table is a bound on $\sum_{K'} n_{ABK'}$. Since there are only nonnegativity and integer constraints on individual cells n_{IJK}, the lower bound on this sum can be distributed to each element in an arbitrary way. Thus, set $n_{ABC} = 0$ and distribute the lower bound among the other elements of the sum.

For (b), a similar argument shows that the upper bound for cell AB can be distributed arbitrarily to the underlying cells. Thus, cell ABC is maximized when n^*_{AB} is dedicated solely to ABC, and the rest of the underlying cells are 0. □

These results apply both to the sharp bounds as well as the linear relaxation bounds. In both cases, bounds can be found for the cells in the table of partial conditionals, and bounds on the cells of the underlying full table follow from Theorem 3.4.

3.3. Tightened Cell Bounds

The linear program defined by (2)–(6) requires only that each marginal total is positive, instead of stipulating that each nonzero cell is positive. This produces wider bounds than necessary because it fails to account for each cell with positive conditional probability. Therefore, we now present a procedure which incorporates this information and can quickly find tightened linear relaxation bounds. Furthermore, we also account for the uncertainty introduced by the rounding which is inevitable when conditional probabilities are released in the form of a contingency table. We denote these as LP* bounds.

Theorem 3.5 *Assume we have a k-way contingency table for which is released the full conditional probabilities* $\hat{d}_{IJ} > 0$*, and the sample size, N. Let* n^-_{AB} *and* n^+_{AB} *be guaranteed lower and upper bounds for cell AB. Also, let* l_I *be the smallest positive number in row I, and r be a specified, positive rounding constant. Then,*

$$n^-_{AB} = \left\lceil \frac{\hat{d}_{AB} - r}{l_B + r} \right\rceil \tag{9}$$

and

$$n^+_{AB} = \left\lfloor \left(N - \sum_{I' \neq A} \sum_{J'} n^-_{I'J'} \right) (\hat{d}_{AB} + r) \right\rfloor \tag{10}$$

where $\lceil \cdot \rceil$ *and* $\lfloor \cdot \rfloor$ *are the ceiling and floor operator, respectively.*

Proof. Suppose $\hat{d}_{AJ'} = l_A$ so that $\hat{d}_{AJ'} < l_A + r$. Since $\hat{d}_{AJ'} > 0$, $n_{AJ'} \geq 1$. Now,

$$n_{A\cdot} = \frac{n_{AJ'}}{\hat{d}_{AJ'}} \geq \frac{1}{\hat{d}_{AJ'}} > \frac{1}{l_A + r} \tag{11}$$

where the first inequality comes from $n_{AJ'} \geq 1$ and the second from $\hat{d}_{AJ'} < l_A + r$. Thus,

$$n_{AB} = \hat{d}_{AB} \cdot n_{A\cdot} \geq (\hat{d}_{AB} - r) \cdot n_{A\cdot} \geq \frac{\hat{d}_{AB} - r}{l_A + r} \tag{12}$$

where the final inequality follows from (11). Since n_{AB} is integer,

$$n^-_{AB} = \left\lceil \frac{\hat{d}_{AB} - r}{l_A + r} \right\rceil$$

For the upper bound, we know that

$$\hat{d}_{AB} = \frac{n_{AB}}{n_{A\cdot}} = \frac{n_{AB}}{N - \sum_{I' \neq A} n_{I'}} \tag{13}$$

so

$$n_{AB} = \left(N - \sum_{I' \neq A} n_{I'}\right)\hat{d}_{AB} \leq \left(N - \sum_{I' \neq A} n^-_{I'\cdot}\right)\hat{d}_{AB} \tag{14}$$

$$\leq \left(N - \sum_{I' \neq A} n^-_{I'\cdot}\right)(\hat{d}_{AB} + r) \tag{15}$$

where $n^-_{I'\cdot} = \sum_{J'} n^-_{I'J'}$.

Since n_{AB} is integer, $n^+_{AB} = \left\lfloor \left(N - \sum_{I' \neq I} \sum_{j'} n^-_{I'J'}\right)(\hat{d}_{AB} + r) \right\rfloor$.

Remark. The constant r accounts for the uncertainty introduced by rounding. For instance, if $\hat{d}_{AB}$ is rounded to the third decimal place, $r = 0.0005$. This means that for a given $\hat{d}_{AB}$, the true value is actually in the interval $[\hat{d}_{AB} - r, \hat{d}_{AB} + r)$; e.g., if we are given in a two-way table that $\hat{d}_{11} = .403$, the true value of $\hat{d}_{11}$ is somewhere in the interval [.4025, .4035).

Remark. If a cell AB is such that $r > \hat{d}_{AB} > 0$, the table entry will be rounded to zero. In this case, we would assume $n_{AB} = 0$ and so the procedure breaks down. However, in the case of rounding to three decimal places so that $r = 0.0005$, any table with a sample size less than $1/.0005 = 2,000$ is guaranteed that any positive count will not be rounded to zero. Similarly, for two decimal places, the sample size must be less than 200. But even if these sample size restrictions are not met, as long as no individual row has a count greater than $1/r$, the result will hold. The data agency can easily check this.

Remark. Although this result gives integer bounds, they do not constitute "exact" bounds because we are not enforcing integer entries on the cells except at the end, artificially, by the floor or ceiling function.

For these LP* bounds when partial conditionals are given, the same arguments made in the previous section apply here. The LP* bounds can be calculated for the table of partial conditionals, and Theorem 3.4 can be invoked to give bounds on the cells of the underlying full table.

4. Illustrative Examples

In this section we calculate bounds for two examples, a four-way and an eight-way table, given full conditional probabilities and table sample size. We also find bounds given partial conditional probabilities.

4.1. Example 1: Four-way Table

This $2 \times 2 \times 2 \times 3$ dataset, with $N = 193$, is due to Koch et al. (1983) and shows the number of patients who recover in a clinical trial for an analgesic drug. These patients are given one of two treatments, have one of two possible statuses, and are treated in one of two centers. There are small counts in this dataset, and even some sampling zeros, as can be seen in Table 1. This example has been used previously to demonstrate an alternative mathematical programming formulation (see Section 3.2) in Slavković (2004), and we will make a comparison to those results.

Let $X_1 = Center$ ($i_1 = 1, 2$), $X_2 = Status$ ($i_2 = 1, 2$), $X_3 = Treatment$ ($i_3 = 1, 2$), and $X_4 = Recovery$ ($i_4 = 1, 2, 3$), and let $o_{i_1 i_2 i_3 i_4}$ be the observed value in the appropriate cell. Thus, $I = \{i_1, i_2, i_3\}$ and $J = \{i_4\}$. Then, the observed full conditionals are $\hat{a}_{IJ} = \hat{P}(Recovery|Center, Status, Treatment)$. For instance,

$$\hat{a}_{1111} = \hat{P}(R = Poor|C = 1, S = 1, T = 1) = \frac{3}{3 + 20 + 5} = 0.107$$

and

$$\hat{a}_{2213} = \hat{P}(R = Excellent|C = 2, S = 2, T = 1) = \frac{4}{3 + 9 + 4} = 0.25$$

4.1.1. Full Conditionals

To calculate the sharp integer bounds for cells in Table 1, we use the integer program defined in (2)–(7). Because of the rounding issue discussed in Section 3.2, the only way to calculate these integer bounds is to use the original tabular counts. We give sharp IP bounds, as well as LP and LP* relaxation bounds, in Table 2.

Table 1. Clinical Trial Data and Full Conditional Probabilities (in parentheses)

			Recovery		
Center	Status	Treatment	Poor	Modest	Excellent
1	1	1	3 (0.107)	20 (0.714)	5 (0.179)
		2	11 (0.333)	14 (0.424)	8 (0.243)
	2	1	3 (0.103)	14 (0.483)	12 (0.414)
		2	6 (0.25)	13 (0.542)	5 (0.208)
2	1	1	12 (0.5)	12 (0.5)	0 (0)
		2	11 (0.524)	10 (0.476)	0 (0)
	2	1	3 (0.188)	9 (0.562)	4 (0.25)
		2	6 (0.333)	9 (0.5)	3 (0.167)

Table 2. IP, LP, and LP bounds for Clinical Trial data given full conditionals and sample size*

Center	Status	Treatment	Poor	Modest	Excellent
1	1	1	[3, 6], [1, 16], [0.11, 19.93]	[20, 40], [7, 110], [0.71, 132.86]	[5, 10], [2, 27], [0.18, 33.21]
		2	[11, 11], [2, 50], [0.33, 62]	[14, 14], [2, 63], [0.42, 78.91]	[8, 8], [1, 36], [0.24, 45.09]
	2	1	[3, 3], [1, 16], [0.10, 19.24]	[14, 14], [5, 74], [0.48, 89.79]	[12, 12], [4, 64], [0.41, 76.97]
		2	[6, 12], [2, 37], [0.25, 46.5]	[13, 26], [3, 81], [0.54, 100.75]	[5, 10], [1, 31], [0.21, 38.75]
2	1	1	[1, 18], [1, 73], [0.5, 93]	[1, 18], [1, 73], [0.5, 93]	0
		2	[11, 11], [2, 77], [0.52, 97.43]	[10, 10], [1, 70], [0.48, 88.57]	0
	2	1	[3, 9], [1, 28], [0.19, 34.88]	[9, 27], [3, 84], [0.56, 104.63]	[4, 12], [2, 37], [0.25, 46.5]
		2	[2, 12], [2, 50], [0.33, 62]	[3, 18], [3, 75], [0.5, 93]	[1, 6], [1, 25], [0.17, 31]

In this table, there are three cells with a count of three. The sharp bounds uniquely identify one of those cells, and while the other cells are more protected the intervals are still fairly narrow. However, keep in mind that these sharp bounds would not be directly available to intruders because no uncertainty due to rounding is assumed. In contrast, even the tightened LP$*$ bounds are much less informative and it is doubtful that they pose a significant disclosure risk, assuming no external information is available.

To demonstrate the computation of the LP bounds, based on Theorem 3.3, we first recall that R is the number of nonzero marginals in the two-way table which is constructed from the k-way table. In this example (Tables 1 and 2), there are no zero marginals, so $R = 2^3 = 8$ since each of the three variables upon which we are conditioning have two categories. Then, each linear relaxation upper bound can be calculated by

$$(N - (R - 1))\hat{d}_{AB} = (193 - 7)\hat{d}_{AB} = 186\hat{d}_{AB}$$

The LP$*$ bounds are not much more difficult to calculate. For instance the smallest entry in the first row of Table 1 is $\ell_1 = 0.107$. Based on (9), the lower bound for cell 1112 is

$$n^{-}_{1112} = \left\lceil \frac{.714 - .0005}{.107 + .0005} \right\rceil = 7 \tag{16}$$

To calculate the upper bound on this cell, we must aggregate the lower bounds for all cells not in row 111. In this case that sum is

$$\sum_{I' \neq 111} \sum_{J'} n^{-}_{I'J'} = 38 \tag{17}$$

so that

$$n^{+}_{1112} = \lfloor (193 - 38)(.714 + .0005) \rfloor = 110 \tag{18}$$

Notice also that since $N = 193 > 1/.0005$, there is no chance that a positive conditional probability was rounded to 0.

Comparing these relaxation bounds to those calculated using the formulation of Slavković (2004), we find that the LP bounds are wider, but the LP* bounds are generally narrower. For instance, in the first cell the LP bounds are [0.11, 19.93] and the LP* bounds are [1, 16], while those in Slavković (2004 p. 145) are [1, 17.03]. In the second cell the LP bounds are [0.71, 132.86], the LP* bounds are [7, 110] while those in Slavković (2004) are [6.67, 113.55]. For this example, the only cell for which the bounds of Slavković (2004) are tighter than the LP* bounds is 2112 ([1, 72.26] versus [1, 73], respectively).

Looking more closely at the integer bounds, we notice that even if the original counts are not uniquely identified, many of the cells have lower bounds that are equal to the actual cell count. Further, in the cases in which the lower bound is less than the actual cell count, the actual cell count is a multiple of the lower bound. In fact, the lower bound is reduced by the greatest common divisor (gcd) of its row. This is most easily seen via an example.

Consider the last row in the four-way example in Tables 1 and 2, which corresponds to a patient with Status 2 who had received Treatment 2 at Center 2. From Table 1, we see that there were 6 patients with a poor recovery, 9 with a modest recovery, and 3 with an excellent recovery. Thus, the conditional probabilities can be calculated as $\hat{d}_{2221} = 6/18$, $\hat{d}_{2222} = 9/18$, and $\hat{d}_{2223} = 3/18$. Now, the greatest factor by which each of these fractions can be reduced is 3, to 2/6, 3/6, and 1/6, respectively, and as can be seen in Table 2 the lower integer bounds for these three cells are 2, 3, and 1. In the fifth row, the gcd is 12, so that the lower bound is 1 for cells 2111 and 2112. In all the other rows the gcd is 1 and so the lower bound is equal to the actual cell count.

Also, the sharp upper bounds calculated via the integer programs seem to be an integer multiple of the lower bound and this multiple seems to be constant among rows. So for instance, the multiple for the last row is 6, but for the first row is 2 (see Table 2).

4.1.2. Partial Conditionals

In addition to examining the cell bounds produced given the full conditionals, $P(R|CST)$, we also look at conditionals involving subsets of the data. Using these "small conditionals" we formulate a linear or integer program using (2)–(7).

Table 3. Treatment|center, status counts and conditional probabilities for clinical trial data

		Treatment	
Center	Status	1	2
1	1	28 (0.459)	33 (0.541)
	2	29 (0.547)	24 (0.453)
2	1	24 (0.533)	21 (0.467)
	2	16 (0.471)	18 (0.529)

Suppose the data owner releases the small conditional, $\hat{P}(Treatment|Center, Status)$. Estimates of these conditional probabilities are calculated from the original data by:

$$\hat{d}_{i_1 i_2 i_3 \cdot} = \frac{\sum_{i_4} o_{i_1 i_2 i_3 i_4}}{\sum_{i_3} \sum_{i_4} o_{i_1 i_2 i_3 i_4}}$$

Using the notation in Section 3.1, $I = \{i_1, i_2\}$, $J = \{i_3\}$, and $K = \{i_4\}$.

These partial conditionals are calculated and shown in Table 3, along with the actual counts for each of these cells. The bounds on this table of partial conditionals are in Table 4.

Interestingly, this table of partial conditionals is completely disclosed. Though there are no very small counts, it demonstrates that small tables are often at risk of disclosure. We can appeal to Theorem 3.4 to determine the bounds for the cells in the full table. Thus, the lower bounds for all underlying cells are 0 and the upper bounds are the same as the counts for the associated cell in the partial table. For instance, for the cell $(Center, Status, Treatment, Response) = (1, 1, 1, 1)$, the lower bound is 0 and the upper

Table 4. IP, LP and LP bounds for clinical trial data, given T|CS conditional probabilities and original data*

		Treatment	
Center	Status	1	2
1	1	[28, 28], [1, 84], [0.46, 87.21]	[33, 33], [2, 99], [0.54, 102.79]
	2	[29, 29], [2, 100], [0.55, 103.96]	[24, 24], [1, 83], [0.45, 86.04]
2	1	[24, 24], [2, 98], [0.53, 101.33]	[21, 21], [1, 86], [0.47, 88.67]
	2	[16, 16], [1, 86], [0. 47, 89.41]	[18, 18], [2, 97], [0.53, 100.59]

Table 5. CPS variables and number of levels

Variable	Name	Num. of levels	Levels	Index
Age	A	3	<25, 25–55, >55	i_1
Employment	B	4	Government, private, self-employed, other	i_2
Education	C	5	$<$HS, HS, college, bachelor, bachelor +	i_3
Marital status	D	2	Married, unmarried	i_4
Race	E	2	Non-White, White	i_5
Sex	F	2	Female, male	i_6
Hours worked	G	3	<40, 40, >40	i_7
Salary	H	2	<50, 50 +	i_8

bound is 28, which is the upper bound for the cell $(Center, Status, Treatment) = (1, 1, 1)$ in Table 4. Notice that the information concerning the two zero cells in the full table (Table 1) is lost.

4.2. Example 2: Eight-way Example

This dataset comes from the U.S. 1993 Current Population Survey (CPS), which is a survey conducted by the U.S. Census Bureau on behalf of the U.S. Bureau of Labor Statistics. It is a collection of data that, according to the U.S. Bureau of Labor Statistics website, is a "monthly survey of households. . . [providing] a comprehensive body of data on the labor force, employment, unemployment, persons not in the labor force, hours of work, earnings, and other demographic and labor force characteristics." This particular dataset includes eight variables and a sample size $N = 48,842$. Table 5 gives the variables, the names we use to represent them, the numbers of levels, the levels themselves, and the index for each variable.

We take Salary to be the response variable, and so the full conditionals will be

$$P(H|A, B, C, D, E, F, G) \tag{19}$$

Table 6. Select data for CPS example

A	B	C	D	E	F	G	H	<50	50 +
>55	Gov't	HS	Married	Non-White	Male	<40		1	1
						>40		0	1
						40		7	3
				White	Female	<40		5	2
						>40		2	0
						40		0	3
					Male	<40		22	3
						>40		10	4
						40		56	24

Table 7. IP/LP/LP results for CPS data given in Table 6*

	IP Bounds		LP* Bounds		LP Bounds	
H	<50	$50+$	<50	$50+$	<50	$50+$
	[1, 8,528]	[1, 8,528]	[1, 20,063]	[1, 20,063]	[0.50, 23,852.50]	[0.50, 23,852.50]
	0	[1, 17,056]	0	[1, 40,105]	0	[1, 47,705]
	[7, 11,942]	[3, 5,118]	[3, 28,081]	[1, 12,046]	[0.70, 33,393.50]	[0.30, 14,311.50]
	[5, 12,185]	[2, 4,874]	[3, 28,642]	[1, 11,485]	[0.71, 34,075.00]	[0.29, 13,630.00]
	[1, 17,056]	0	[1, 40,105]	0	[1, 47,705]	0
	0	[1, 17,056]	0	[1, 40,105]	0	[1, 47,705]
	[22, 15,026]	[3, 2,049]	[8, 35,301]	[1, 4,831]	[0.88, 41,980.40]	[0.12, 5,724.60]
	[5, 12,185]	[2, 4,874]	[3, 28,642]	[1, 11,485]	[0.71, 34,075.00]	[0.29, 13,630.00]
	[7, 11,942]	[3, 5,118]	[3, 28,081]	[1, 12,046]	[0.70, 33,393.50]	[0.30, 14,311.50]

4.2.1. Full Conditionals

In the full conditionals for this dataset, we have many margins which are zero. This presents the question of how to treat these rows, since most relevant inferential procedures proceed under the assumption that the margins are greater than zero. One possibility is to collapse certain variables to fewer categories, while another option is to treat those variables as identical to zero.

From the perspective of the agency releasing the data, conditionals with zero marginals are undefined, and would be of little inferential use. Thus, collapsing the table in such a way that no margins are zero is probably the best solution. However, for this dataset, significantly collapsing those variables with four and five categories still does not result in exclusively nonzero margins. Also, there seems to be no guidance in the Statistical Policy Working Paper 22 by the Federal Committee on Statistical Methodology that covers this. Thus, we assume that any "conditional probability" from a row with a margin of zero is zero itself. In fact, we do not even optimize it, setting $n_{IJ} = 0$ if $\sum_J o_{IJ} = 0$. Besides this, the structure of the integer program is as (2)–(7), with $I = \{i_1, i_2, i_3, i_4, i_5, i_6, i_7\}$ and $J = \{i_8\}$.

The results are too numerous to give in their entirety. Instead, we present an interesting subset of the results in Table 7. For convenience, we present the corresponding original data in Table 6. For instance, in the third row of the displayed results, the actual counts for the two cells are 7 and 3, and the sharp IP bounds are [7, 11,942] and [3, 5,118] the LP bounds are [0.7, 33,393.5] and [0.3, 14,311.5], and the LP* bounds are [3, 28,081] and [1, 12,046].

These results are consistent with those from the other example. Overall, the IP results are often significantly narrower than the bounds from the linear relaxation but, as can be seen, even for cells with counts of 1 there are no narrow bounds which in and of themselves would cause a disclosure risk. However, the IP lower bounds are often the same as the original counts. Because of this, care should be taken before releasing the full conditional probabilities. On the other hand, for this example as with the previous, exact bounds can only be calculated using the original counts (because of inexactness due to rounding the released rates; see Section 3.2) and so would not be computable to data snoopers. As in the four-way table, the lower bound for a cell is different than the actual counts when its row has a greatest common divisor other than 1.

Table 8. The first six and last six rows for table of partial conditional probabilities $P(H|A, C, D, E, F, G)$

A	C	D	E	F	G	H	<50	$50+$
<25	$<HS$	Married	Non-White	Female	<40		6	0
					40		1	0
					>40		5	0
				Male	<40		1	0
					40		1	0
					>40		2	0
...	...	...	...	...	...		...	...
25–55	HS	Unmarried	White	Female	<40		542	12
					40		311	19
					>40		1,022	12
				Male	<40		263	3
					40		647	55
					>40		1,169	46

Because thousands of integer programs have to be solved to compute all of the exact bounds, and because each of the IPs is fairly large (more than 1,000 constraints and decision variables), considerable time was required for Cplex to solve them. In all, using an API interface between Matlab and CPLEX V12.1, it took more than five hours to compute all bounds via a compute node on Miami University's Redhawk cluster. Such a node has dual Quad-core 2.26 GHz Intel E5520 processors with Intel64 technology, and 24 GB of memory and 160 GB of local disk space. (Note: We found that the large table revealed some memory management issues in this Matlab/CPLEX interface; it could not solve all of the cells consecutively; however, when broken into smaller subgroups of cells and run as separate jobs, each IP was solved.)

4.2.2. Partial Conditionals

It is quite possible with such a large dataset, that only a portion may be released. For instance, suppose a researcher was interested in only a subset of the eight variables.

Table 9. Table of partial conditional probabilities $P(H|D, F, G)$

D	F	G	H	<50	$50+$
Married	Female	<40		689	369
		40		233	257
		>40		748	513
	Male	<40		1,740	570
		40		3,767	4,579
		>40		5,811	3,768
Unmarried	Female	<40		5,041	90
		40		1,827	311
		>40		5,885	229
	Male	<40		3,122	66
		40		2,783	595
		>40		5,509	340

Table 10. IP, LP, and LP bounds for the cells in Table 9, given partial conditional probabilities $P(H|D,F,G)$ and total sample size*

	IP Bounds		LP* Bounds		LP Bounds	
H	<50	$50+$	<50	$50+$	<50	$50+$
	[689, 5,512]	[369, 2,952]	[2, 31,705]	[1, 17,008]	[0.65, 31,800.15]	[0.35, 17,030.85]
	[233, 2,563]	[257, 2,827]	[1, 23,188]	[2, 25,524]	[0.48, 23,219.64]	[0.52, 25,611.36]
	[748, 5,984]	[513, 4,104]	[2, 28,882]	[1, 19,830]	[0.59, 28,965.57]	[0.41, 19,865.43]
	[174, 4,060]	[57, 1,330]	[4, 36,670]	[1, 12,045]	[0.75, 36,781.79]	[0.25, 12,049.21]
	[3,767, 3,767]	[4,579, 4,579]	[1, 21,972]	[2, 26,741]	[0.45, 22,040.06]	[0.55, 26,790.94]
	[1,937, 5,811]	[1,256, 3,768]	[2, 29,563]	[1, 19,149]	[0.61, 29,622.81]	[0.39, 19,208.19]
	[5,041, 10,082]	[90, 180]	[54, 47,864]	[1, 901]	[0.98, 47,974.48]	[0.02, 856.52]
	[1,827, 5,481]	[311, 933]	[6, 41,636]	[1, 7,081]	[0.85, 41,727.89]	[0.15, 7,103.11]
	[5,885, 11,770]	[229, 458]	[26, 46,911]	[1, 1,825]	[0.96, 47,002.03]	[0.04, 1,828.97]
	[1,561, 7,805]	[33, 165]	[46, 47,710]	[1, 1,047]	[0.98, 47,820.07]	[0.02, 1,010.93]
	[2,783, 5,566]	[595, 1,190]	[5, 40,126]	[1, 8,589]	[0.82, 40,229.92]	[0.18, 8,601.08]
	[5,509, 5,509]	[340, 340]	[17, 45,880]	[1, 2,847]	[0.94, 45,992.47]	[0.06, 2,838.53]

Alternatively, perhaps the data agency decides to release a smaller table that consists of a subset of the original variables, to ensure cell counts are sufficiently large for privacy purposes. In these cases, as with the first example, there is interest in determining bounds on the released partial conditional probabilities. As before, there are two sorts of bounds: Bounds on the cells in the underlying full contingency table cells, and bounds on the cells of the table of partial conditionals.

We will consider two tables of partial conditional probabilities based on the CPS data. The first sums over the second variable, employment, so that the probabilities of interest are $P(H|A,C,D,E,F,G)$. A portion of this data is given in Table 8. The second, $P(H|D,F,G)$, collapses significantly further and results in a 12×2 table in which all cells are not only nonzero but in the hundreds and thousands. These data are shown in Table 9.

The IP, LP$*$, and LP bounds for the cells in Table 9 are given in Table 10. Since Table 9 has been collapsed from the full table to the extent that all cells have large counts, there is no significant risk of disclosure from a small count. Note, however, that the cell counts in the fifth and twelfth rows, while not small, are fully disclosed.

In terms of the underlying cells in the full 8-way table, we can appeal to Theorem 3.4 to conclude that the lower bounds for all the cells in the full table are 0, and the upper bounds are the same as the associated upper bounds in Table 10. For instance, if the upper bound for cell $(D,F,G,H) =$ (Unmarried, Female, 40, 50+) is 933, then the upper bounds for all cells in the full table with $(D,F,G,H) =$ (Unmarried, Female, 40, 50+) are the same.

We also include the LP$*$ (assuming the partial conditional probabilities given with three decimal places) and LP bounds in Table 10, and they are quite wide. Note that in several cases, the LP$*$ bounds are actually wider than the LP bounds (for instance, in Row 7, in the $50+$ category). This is because the LP$*$ bounds account for rounding uncertainty. If rounding considerations were eliminated, the LP$*$ bounds would be tighter. Also, there are two rows in this table that sum to more than 2,000. Thus, if either of them included a count of 1, its conditional probability might be rounded to 0. In this case, all of the cell counts in these rows were much higher than 1.

For the other set of conditional probabilities, $P(H|A,C,D,E,F,G)$ (Table 8), the original table is collapsed only over variable B, employment, which has four categories.

This results in a flattened table that is 360×2, and still contains many cells with zero counts, as well as rows with zero counts. Compared to the smaller table of partial conditional probabilities in Table 9, the exact bounds are much wider, because the rows with small sums give much more flexibility for the bounds. In fact, excluding the zero cells, the narrowest set of bounds is in the row which corresponds to (A, C, D, E, F, G) = (<25, College, Unmarried, White, Female, <40), for the 50 + salary, with bounds [1, 18]. The next narrowest bounds are 44. The associated LP and LP$*$ bounds are much wider even than the exact bounds. Again, for the underlying cells in the full table, the lower bounds are 0 and the upper bounds are the same as the associated upper bounds in the partial table. For instance, in the underlying table there would be four cells corresponding to $(A, C, D, E, F, G) = (< 25, \text{College}, \text{Unmarried}, \text{White}, \text{Female}, < 40)$, and those four cells would have the same values of (A, C, D, E, F, G) and each of the values of variable B (employment).

5. Discussion

To date statistical disclosure limitation methodologies for tables of counts have been heavily focused on the release of unaltered marginal totals from such tables, and in part on inferences that are possible by an intruder from such releases. Many statistical agencies also release other forms of summary data from tables, such as tables of rates or observed conditional frequencies. These are predominantly released as two-way and three-way tables, with conditioning on a single variable.

In this article, we have extended Smucker and Slavković (2008) to k-way contingency tables, using the same basic IP formulation. We also give closed-form procedures for linear relaxation bounds which extend those given in Smucker and Slavković (2008). We then develop improved linear relaxation bounds which also account for the uncertainty due to the rounding of the conditional probabilities in the released data.

The bounds calculated in this article (see Tables 2, 4, 7, 10, and 11) show that generally the linear relaxation bounds are significantly wider than the sharp integer bounds, and that

Table 11. IP, LP$*$, and LP bounds for the counts in Table 8, given partial conditional probabilities $P(H|A, C, D, E, F, G)$ and total sample size

	IP Bounds		LP* Bounds		LP Bounds	
H	<50	50 +	<50	50 +	<50	50 +
	[1, 16,334]	0	[1, 42,829]	0	[1.00, 48,496.00]	0
	[1, 16,334]	0	[1, 42,829]	0	[1.00, 48,496.00]	0
	[1, 16,334]	0	[1, 42,829]	0	[1.00, 48,496.00]	0
	[1, 16,334]	0	[1, 42,829]	0	[1.00, 48,496.00]	0
	[1, 16,334]	0	[1, 42,829]	0	[1.00, 48,496.00]	0
	[1, 16,334]	0	[1, 42,829]	0	[1.00, 48,496.00]	0
	. . .	. . .	. . .	. . .	. . .	. . .
	[271, 15,989]	[6, 354]	[44, 41,930]	[1, 964]	[0.98, 47,445.55]	[0.02, 1,050.45]
	[311, 15,550]	[19, 950]	[17, 40,362]	[1, 2,505]	[0.94, 45,703.81]	[0.06, 2,792.19]
	[511, 16,352]	[6, 192]	[79, 42,393]	[1, 536]	[0.99, 47,933.18]	[0.01, 562.82]
	[263, 16,306]	[3, 186]	[86, 42,443]	[1, 493]	[0.99, 47,949.05]	[0.01, 546.95]
	[647, 15,528]	[55, 1,320]	[12, 39,501]	[1, 3,361]	[0.92, 44,696.46]	[0.08, 3,799.54]
	[1,169, 16,366]	[46, 644]	[25, 41,226]	[1, 1,649]	[0.96, 46,659.94]	[0.04, 1,836.06]

the relaxation bounds in particular are not narrow enough to pose significant disclosure risk. The IP bounds, for the smaller example as well as the small partial conditional table for the larger example, resulted in some narrow bounds and some cell counts that were uniquely identified. Additionally, as noted in Section 4, the sharp lower bounds for cells are often the same as the actual count, and when they are not it is because the greatest common divisor in the cell's row is larger than 1. Further, the upper bounds seem to be an integer multiple of the lower bound and this multiple seems to be constant among rows.

As given here, the computational requirements of calculating integer bounds for a table do not necessarily increase linearly in the number of cells. The sparsity of large tables appears to make the individual optimizations easier. For instance, the full eight-way table of Section 4.2.1 has over a thousands cells which took CPLEX a total computation time of about 5 hours; the much smaller partial table of Section 4.2.2 has only 24 cells, but took 0.834 hours to solve. There may be other factors influencing this sublinear (in the number of cells) relationship, but we note that a different, noncommercial Matlab/CPLEX interface solved the full table in less than 2 hours, but took over 12 hours for the small partial table. Overall, it appears that sparsity, as much as or more than the number of cells, plays an important role in the computational difficulty.

In summary, the IP bounds may provide a substantial amount of information and should be examined by data agencies before rates are released. On the other hand, linear relaxation bounds are quite wide and seem to offer little insight into disclosure risk for particular cells. From a data snooper's perspective, the IP's are generally infeasible unless the original data is given. However, it remains to be seen whether introducing rounding uncertainty to the IP formulation will be a reasonable obfuscatory mechanism. This is also the source of ongoing work.

We address only bounds based on full and partial conditional probabilities in this article. Alternatively, an agency might release a combination of partial conditionals and marginal totals (see Slavković et al. 2012). The mathematical programs presented here might be adapted to this situation, and it is possible that closed-form linear relaxation bounds, similar to those in this article, could also be derived.

We note also that in addition to the gaps between exact bounds and their linear relaxation, there also are gaps *within* established integer bounds. In other words, given a set of released conditional probabilities, feasible tables may not exist allowing a cell to take on certain counts within the computed sharp bounds. An algebraic approach, rather than the mathematical programming approach of this article, may be better equipped to explore such gaps (see Dobra et al. 2008; Slavković 2010).

6. References

(2005). Report on statistical disclosure limitation methodology. Federal Committee on Statistical Methodology, Statistical Policy Working Paper 22 (Version Two).

(2009). IBM ILOG CPLEX V12.1 User's Manual for CPLEX.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilitá. Publicazioni del R. Instituto Superiore di Scienze Economiche e Commerciali di Firenze, 8. [In Italian]

Buzzigoli, L. and Gusti, A. (1998). An Algorithm to Calculate the Upper and Lower Bounds of the Elements of an Array Given its Marginals. Statistical Data Protection (SDP '98) Proceedings, 131–147, Luxemburg. Eurostat.

Chen, Y., Dinwoodie, I., and Sullivant, S. (2006). Sequential Importance Sampling for Multiway Tables. Annals of Statistics, 34, 523–545.

Cox, L. (2002). Bounds on Entries in 3-dimensional Contingency Tables. Inference Control in Statistical Databases – From Theory to Practice. LNCS, J. Domingo-Ferrer (ed.). Volume 2316, Springer-Verlag, 21–33.

Cox, L. (2007). Contingency Tables of Network Type: Models, Markov Basis and Applications. Statistica Sinica, 17, 1371–1393.

Dobra, A. and Fienberg, S.E. (2001). Bounds for Cell Entries in Contingency Tables given Marginal Totals and Decomposable Graphs. Statistical Journal of the United Nations Economic Commission for Europe, 18, 363–371.

Dobra, A. and Fienberg, S.E. (2003). Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals. Statistical Journal of the United Nations ECE, 18, 363–371.

Dobra, A. and Fienberg, S. (2010). The Generalized Shuttle Algorithm. In Algebraic and Geometric Methods in Statistics, P. Gibilisco, E. Riccomagno, and M.-P. Rogantin (eds). Cambridge University Press, 135–173.

Dobra, A., Fienberg, S., Rinaldo, A., Slavković, A., and Zhou, Y. (2008). Algebraic Statistics and Contingency Table Problems: Log-linear Models, Likelihood Estimation and Disclosure Limitation. IMA Volumes in Mathematics and its Applications: Emerging Applications of Algebraic Geometry, M. Putinar and S. Sullivant (eds). Volume 149, Springer Science + Business Media, Inc, 63–88.

Fienberg, S.E. (1999). Fréchet and Bonferroni Bounds for Multi-way Tables of Counts with Applications to Disclosure Limitation. Statistical Data Protection: Proceedings of the Conference. Luxembourg: Eurostat, 115–129.

Fienberg, S.E. and Slavkovic, A.B. (2005). Preserving the Confidentiality of Categorical Statistical Data Bases when Releasing Information for Association Rules. Data Mining and Knowledge Discovery, 11, 155–180.

Fréchet, M. (1940). Les Probabilitiés Associées a un Système d'Événments Compatibles et Dépendants, Vol. Premiere Partie. Paris: Hermann & Cie. [In French]

Hoeffding, W. (1940). Scale-invariant Correlation Theory. Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universität Berlin, 5, 181–233.

Hosten, S. and Sturmfels, B. (2003). Computing the Integer Programming Gap. Combinatorica, 27, 367–382.

Koch, G., Amara, J., Atkinson, S., and Stanish, W. (1983). Overview of Categorical Analysis Methods. SAS-SUGI, 8, 785–795.

Nemhauser, G.L. and Wolsey, L.A. (1988). Integer and Combinatorial Optimization: Wiley-Interscience.

Onn, S. (2006). Entry Uniqueness in Margined Tables. Privacy in Statistical Databases – PSD 2006. LNCS, J. Domingo-Ferrer and L. Franconi (eds). Volume 4302, Springer-Verlag, 94–101.

Salazar-Gonzalez, J.-J. (2008). Statistical Confidentiality: Optimization Techniques to Protect Tables. Computers and Operations Research, 35, 1638–1651.

Slavković, A.B. (2004). Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables. PhD thesis, Carnegie Mellon University.

Slavković, A.B. (2010). Partial Information Releases for Confidentiality Contingency Table Entries: Present and Future Research Efforts. Journal of Privacy and Confidentiality, 1.

Slavković, A.B. and Fienberg, S.E. (2004). Bounds for Cell Entries in Two-way Tables Given Conditional Relative Frequencies. In Privacy in Statistical Databases – PSD 2004, Lecture Notes in Computer Science No. 3050, J.J. Domingo-Ferrer and V. Torra (eds). Springer-Verlag, 30–43.

Slavković, A.B. and Lee, J. (2010). Synthetic Two-way Contingency Tables that Preserve Conditionals Frequencies. Statistical Methodology, 7, 225–239.

Slavković, A.B., Zhu, X., and Petrović, S. (2012). A Sample Space of k-way Tables given Conditionals and Their Relations to Marginals: Implications for Cell Bounds and Markov Bases. Forthcoming.

Smucker, B. and Slavković, A.B. (2008). Cell Bounds in Two-way Contingency Tables Based on Conditional Frequencies. PSD 2008. LNCS, J. Domingo-Ferrer and Y. Saygin (eds). volume 5262, Berlin Heidelberg: Springer-Verlag, 64–76.

Sullivant, S. (2005). Small Contingency Tables with Small Gaps. Siam J. Discrete Math, 18, 787–793.

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

Received July 2010
Revised August 2011