

# Bounds for Cell Entries in Two-way Tables Given Conditional Relative Frequencies

Aleksandra B. Slavkovic<sup>1</sup> and Stephen E. Fienberg<sup>2\*\*</sup>

<sup>1</sup> Department of Statistics  
Carnegie Mellon University, Pittsburgh, PA 15213-3890, U.S.A.  
`sesa@stat.cmu.edu`

<sup>2</sup> Department of Statistics  
Center for Automated Learning and Discovery  
Center for Computer Communications and Security  
Carnegie Mellon University, Pittsburgh, PA 15213-3890, U.S.A.  
`fienberg@stat.cmu.edu`

**Abstract.** In recent work on statistical methods for confidentiality and disclosure limitation, Dobra and Fienberg (2000, 2003) and Dobra (2002) have generalized Bonferroni-Fréchet-Hoeffding bounds for cell entries in  $k$ -way contingency tables given marginal totals. In this paper, we consider extensions of their approach focused on upper and lower bounds for cell entries given arbitrary sets of marginals and conditionals. We give a complete characterization of the two-way table problem and discuss some implications to statistical disclosure limitation. In particular, we employ tools from computational algebra to describe the locus of all possible tables under the given constraints and discuss how this additional knowledge affects the disclosure.

*Keywords:* Confidentiality; Contingency tables; Integer programming; Linear programming; Markov bases; Statistical disclosure control; Tabular data.

## 1 Introduction

Current disclosure limitation methods either change the data (e.g. cell suppression, controlled rounding, and data swapping) or release partial information (e.g. releasing marginals). Methods falling under the first category either have limited statistical content (e.g. cell suppression) or modify existing statistical correlations (e.g. controlled rounding), thus modifying the proper statistical inference. The third method insures confidentiality by restricting the level of detail at which data are released, as opposed to changing the data as with other methods. Releases of partial information such as releasing only the margins often allows for proper statistical inference. In this setting, Dobra and Fienberg (2000, 2002, 2003) used undirected graphical representations for log-linear models as a framework to compute bounds on cell entries given a set of margins as

---

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

input to assessing disclosure risk. Natural extensions of this work involve other types of partial data releases such as conditionals, regressions, or any other data summaries. Government agencies in the U.S. are already releasing tables of rates, that is conditional relative observed frequencies (for example, see Figure 1). However, not much is known on their effect on confidentiality and data privacy. Furthermore, releasing of conditional distributions for higher-dimensional contingency tables could be useful for researchers interested in assessing causal inference with observed data while still maintaining confidentiality.

**Table 2. Volunteer rates by sex, race, Hispanic origin, and selected characteristics, September 2002**

Selected characteristics	White			Black			Hispanic		
	Total	Men	Women	Total	Men	Women	Total	Men	Women
<b>Age</b>									
Total, 16 years and over	29.4	25.1	33.4	19.2	16.7	21.1	15.7	12.9	18.4
16 to 19 years	28.6	24.3	33.0	18.8	16.3	21.1	18.1	15.3	20.9
20 to 24 years	19.3	15.7	22.9	13.1	9.9	15.8	9.4	7.6	11.3
25 to 34 years	26.8	20.8	32.7	20.2	15.6	24.0	16.9	12.9	21.0
35 to 44 years	37.1	30.8	43.4	22.4	19.1	25.2	20.6	15.7	25.4
45 to 54 years	33.5	29.4	37.6	20.4	19.3	21.2	18.1	15.1	17.1
55 to 64 years	28.8	26.1	31.4	20.6	19.1	21.7	13.2	12.0	14.2
65 years and over	23.9	22.2	25.2	13.9	14.9	13.3	6.9	6.2	7.4
<b>Employment status among persons aged 16 years and over</b>									
Employed	31.4	27.1	36.6	21.9	18.9	24.6	17.0	14.0	21.1
Unemployed	26.5	21.3	32.6	21.5	18.2	24.6	17.9	12.3	25.5
Not in the labor force	25.6	20.1	28.9	14.1	12.1	15.5	12.6	9.0	14.4
<b>School enrollment status among persons aged 16 to 24 years</b>									
Enrolled in high school	32.3	26.0	39.5	18.2	17.0	19.4	19.6	15.6	23.9
Enrolled in college	28.3	25.2	31.1	23.9	19.7	26.4	19.6	19.4	19.7
Not enrolled in school	16.0	13.0	19.1	10.5	8.4	12.7	8.6	7.2	10.3
<b>Educational attainment among persons aged 25 years and over</b>									
Less than a high school diploma	10.5	9.0	11.8	9.2	8.6	9.6	8.4	5.8	11.0
High school graduate, no college <sup>1</sup>	22.8	18.2	26.8	14.1	12.7	15.4	16.3	13.4	19.2
Less than a bachelor's degree <sup>2</sup>	34.5	28.9	39.4	26.1	23.2	28.0	25.2	22.9	27.2
College graduate	46.0	40.9	51.4	36.6	33.4	39.1	31.9	27.2	36.4

<sup>1</sup> Includes high school diploma or equivalent.  
<sup>2</sup> Includes the categories of some college, no degree; and associate's degree.

NOTE: Data on volunteers relate to persons who performed unpaid volunteer activities for an organization at any point from September 1, 2001, through the survey week in September 2002. Details for the above race and Hispanic-origin groups will not sum to totals because data for the "other races" group are not presented and Hispanics are included in both the white and black population groups.

**Fig. 1.** An example of a published 3-way table with rates from the U.S. Department of Labor Bureau of Labor Statistics website. Data are from 2002 CPS supplement. Source: Boraas, S. Volunteerism in the U.S. Monthly Labor Review, August 2003.

In this paper we develop nonparametric upper and lower bounds for cell entries in two-way contingency tables given an arbitrary set of marginals and conditionals for purpose of evaluating disclosure risk. In Section 2, we provide technical background and introduce some notation. In Section 3, we discuss the complete characterization of the joint distribution and present new results on uniqueness of the entries in the table given a marginal and a conditional. In Section 4, we estimate the bounds on cell entries via optimizations techniques such as linear and integer programming and discuss issues relevant to disclosure limitation. For two-way tables we give a complete characterization of the space

of possible tables given different margins and/or conditionals. We also employ a representation from algebraic geometry to describe the locus of all possible tables under the given constraints and discuss possible implications for disclosure.

## 2 Technical Background and Notation

Beyond the disclosure setting, bounds for cell entries in tables arise in a variety of other statistical contexts. They can be found in mass transportation problems (Rachev and Rüschendorf 1998), computer tomography (Gutmann et al. 1991), ecological inference (King 1997), and causal inference of imperfect experiments (Balke and Pearl 1997). Most of this work focuses on bounds induced by given marginals, with very little attention to bounds induced by sets of conditionals and marginals. For example, Balke and Pearl (1997), Tian and Pearl (2000), and Pearl (2000) use linear programming formulations to establish sharp bounds on causal effect in experiments with imperfect compliance. These are non-parametric bounds limited to a single cause analysis and the cases are limited to bivariate random variables. Pearl (2000) also describes a Gibbs sampling technique to calculate the posterior distribution of the causal quantity of interest, but he points out that the choice of prior distributions in this setting can have a significant influence on the posterior.

While statisticians have long been interested in combining marginal and conditional distributions, results have been developed mainly for complete specification of the joint as in the work of Gelman and Speed (1993, 1999), Arnold et al. (1996, 1999), and Besag (1974). The Hammersley-Clifford Theorem (1974) establishes a connection between the joint distribution and the full conditionals, and often describes conditional statements of an undirected graph. 1993 defined conditions under which a collection of marginal and conditional distributions uniquely identifies the joint distribution. Their key assumption is positivity in the Hammersley-Clifford Theorem (Besag 1974) which, for the discrete case, means that there are no cells with zero probability. Arnold et al. (1996, 1999) extended the sets defined by the Gelman and Speed theorem by relaxing the positivity condition such that in the discrete case the sets never involve conditioning on an event of zero probability, i.e., appropriate marginal distributions are strictly positive. Arnold et al. (1996, 1999) also address the question of whether the given set of densities are consistent (compatible) and whether they uniquely specify the joint density. When sets of queries in the form of marginals and conditionals satisfy the conditions of the Gelman and Speed (1993) and/or Arnold et al. (1999), the released information will then reveal all of the information in the table since the joint distribution will be uniquely identified. This uniqueness poses a problem in tables with small counts as it increases a risk of uniquely identifying individuals who provided the data. We are interested in identifying such situations as part of a broader goal of developing safe tabular releases in terms of arbitrary sets of marginals and conditionals that would still allow for proper statistical inferences about models for the original table.

Arnold et al. (1999) describe an algorithm for checking the compatibility of a set of conditional probability densities. In the disclosure-type setting, we know that the unique joint distribution exists, and released conditionals are compatible in the sense that they come from the same joint distribution over the space of all variables. The issue of compatibility may become relevant if we are to consider that an intruder has outside additional knowledge and would like to incorporate some prior information with the data provided by the agency. We are interested in assessing the uniqueness condition. Arnold et al. (1999) provide a uniqueness theorem (see Section 3) and describe an iterative algorithm due to Vardi and Lee (1993) for determining the missing marginal distribution in order to establish the joint distribution. In cases where there is more than one solution, the algorithm obtains one of the solutions but is unable to detect that there is more than one solution. This algorithm is like the iterative proportional fitting algorithm for maximum likelihood estimation from incomplete contingency tables (e.g., see Bishop, Fienberg, and Holland (1975)), and solves linear equations subject to non-negativity constraints and cell probabilities summing to one. While the algorithm always converges for compatible distributions, the convergence is quite slow even for two-way tables and it is unclear how successfully the algorithm deals with boundary cases.

Algebraic statistics is an emerging field that advocates use of tools of computational commutative algebra such as Gröbner bases in statistics (Pistone et al. 2001). Diaconis and Sturmfels' (1998) seminal work had a major impact on developments in statistical disclosure techniques. Their paper applies the algebraic theory of toric ideals to define a Markov Chain Monte Carlo method for sampling from conditional distributions, through the notion of Markov bases. Given a set of marginal constraints, Dobra and Fienberg (2002, 2003) and Dobra et al. (2003) have utilized Gröbner bases in connection with graphs to fully describe the space of possible tables. Most recently, Slavkovic (2003) described Markov bases for two-way tables given the sets of conditionals and marginals, a topic closely related to that of the present paper, and to which we return later.

## 2.1 Notation

Let  $X$  and  $Y$  be discrete random variables with possible values  $x_1, x_2, \dots, x_I$  and  $y_1, y_2, \dots, y_J$ , respectively. Let  $n_{ij}$  denote the observed cell counts in the  $I \times J$  table  $\mathbf{n}$ . We represent their *joint probability distribution* as the  $I \times J$  matrix  $P = (p_{ij})$  which is the normalized table of counts such that all cell entries  $p_{ij} = P(X = x_i, Y = y_j), i = 1, 2, \dots, I, j = 1, 2, \dots, J$ , are nonnegative and sum to 1. Let  $p_{i.} = \sum_{j=1}^J p_{ij} = P(X = x_i)$  and  $p_{.j} = \sum_{i=1}^I p_{ij} = P(Y = y_j)$  be the *marginal probability distributions* for  $X$  and  $Y$ , respectively. We can also represent the *conditional probability distributions* as  $I \times J$  matrices  $A = (a_{ij})$  and  $B = (b_{ij})$  where

$$a_{ij} = P(X = x_i | Y = y_j) = \frac{p_{ij}}{p_{.j}} = \frac{n_{ij}}{n_{.j}}, \quad i = 1, 2, \dots, I, j = 1, 2, \dots, J, \quad (1)$$

$$b_{ij} = P(Y = y_j | X = x_i) = \frac{p_{ij}}{p_{i.}} = \frac{n_{ij}}{n_{i.}}, \quad i = 1, 2, \dots, I, j = 1, 2, \dots, J. \quad (2)$$

### 3 Complete Specification of the Joint Distribution

In the disclosure limitation setting, we say that sets of conditionals and/or marginals are compatible if they come from the same joint distribution. But when there is compatibility, we want to check if the given set of conditionals and/or marginals is sufficient to uniquely identify the joint distribution. We allow cell entries to be zero as long as we do not condition on an event of zero probability. Then the uniqueness theorem of Gelman and Speed (1993) as amended by Arnold et al. (1999) tells us that the joint distribution for any two-way table is uniquely identified by any of the following sets of distributions:

- (1)  $f(x|y)$  and  $f(y|x)$ ,
- (2)  $f(x|y)$  and  $f(y)$ ,
- (3)  $f(y|x)$  and  $f(x)$ ,

(Note that these are equivalent under the independence of  $X$  and  $Y$ .)

In addition, Arnold et al. (1996, 1999) show that sometimes collections of the type  $\{f(x|y), f(x)\}$  or  $\{f(y|x), f(y)\}$  also uniquely identify the joint distribution as long as we do not condition on a set of probability zero. When the positivity condition on the Cartesian product assumed by Gelman and Speed does not hold, Arnold et al. (1999) suggest that uniqueness can be checked by determining the uniqueness of a missing marginal given the provided information. Besides using the Vardi and Lee (1993) algorithm, they also suggest running a Markov process that generates  $X$ 's by cycling through a list of conditional distributions. If the process is irreducible, then there exists a unique marginal that together with the given conditionals uniquely determines the joint distribution. If we already know that we are given proper conditional distributions for tabular data, however, the proposed algorithm can be replaced by simply checking the number of levels of two variables and the rank of the conditional matrix. For a subset of cases we can apply a simple formula for finding the cell probabilities as well, as the following results imply.

**Theorem 1.** *For two discrete random variables,  $X$  and  $Y$ , either the collection  $\mathcal{C}_x = \{f(x|y), f(x)\}$  or the collection  $\mathcal{C}_y = \{f(y|x), f(y)\}$  uniquely identifies the joint distribution if matrices  $A$  and  $B$  have full rank and if  $I \geq J$  for  $\mathcal{C}_x$  or  $J \geq I$  for  $\mathcal{C}_y$ .*

The unique joint distribution in Theorem 1 takes a simple form for the entries of the  $I \times 2$  table.

**Theorem 2.** *Suppose that for an  $I \times 2$  tables where  $I \geq 2$  we are given  $f(x|y)$  and  $f(x)$ . Then the unique probabilities for the cell entries are given by*

$$p_{ij} = a_{ij} \frac{p_{i.} - a_{i\{\mathcal{J}\setminus j\}}}{a_{ij} - a_{i\{\mathcal{J}\setminus j\}}}. \quad (3)$$

*Example 1.* The data in Table 1 are from a fictitious survey of 50 students about illegally downloading MP3s. Suppose that the only information available about the survey are  $P(\text{Download}) = \{0.4, 0.6\}$  and

$$P(\text{Download}|\text{Gender}) = \{b_{ij}\} = B = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}$$

Applying Theorem 2, we get the exact cell probabilities in the original table  $\{p_{11}, p_{12}, p_{21}, p_{22}\} = \{0.3, 0.2, 0.1, 0.4\}$ . For example,  $p_{12} = b_{12} \frac{p_{.2} - b_{22}}{b_{12} - b_{22}} = 0.2$ .

**Table 1.** Fictitious example: Number of students illegally downloading MP3s by gender

	Download Yes	Download No	Total
Male	15	10	25
Female	5	20	25
Total	20	30	50

Note that these results hold regardless of the value of sample size  $N$ . Knowledge of the sample size will give us the original (unique) table of counts.

## 4 Bounds for Cell Entries in $I \times J$ Tables

Any contingency table with non-negative integer or real entries and fixed marginal and/or conditionals is a point in the convex polytope defined by a linear system of equations induced by released conditionals and marginals. When the table does not satisfy the Gelman and Speed (1993) Theorem or Theorem 1, we have the possibility of more than one realization for the joint distribution  $(X, Y)$ , i.e., there is more than one possible table that satisfies the constraints implied by the margins and conditionals. One way to evaluate the safety of released tabular data is to determine bounds on cell entries given the margins and conditional. To fully describe the space of  $I \times J$  tables subject to marginal and conditional constraints, we need to consider combinations of marginals and conditionals of the following forms:

1.  $f(x)$  or  $f(y)$ ,
2.  $f(x|y)$  or  $f(y|x)$ ,
3.  $\{f(x|y), f(x), I < J\}$  or  $\{f(y|x), f(y), J < I\}$ .

### 4.1 Linear/Integer Programming Bounds

The simplest and most natural method for solving system of linear equations is the simplex method. Dobra and Fienberg (2003) discuss some inadequacies of this method for addressing bounds given sets of marginals, but this method works relatively well for tables with small dimensions, Hence we explore its feasibility and usefulness for the other two listed cases.

**Unknown sample size  $N$ .** We first consider a setting where sample size  $N$  is unknown. It is trivial to see that first two sets do not give us the joint distribution. When a single marginal is given, the cell probabilities are bounded below by zero and above by the corresponding marginal value, e.g.  $0 \leq p_{ij} \leq p_i$ . (Dobra and Fienberg 2000). When a single conditional distribution is given, the cell probabilities are bounded by zero and associated conditional probability; for example,  $0 \leq p_{ij} \leq a_{ij}$ , for  $f(x|y)$ . This can be verified by setting up a linear programming (LP) problem since the observed conditional frequencies are a linear-fractional map of either the original cell counts or cell probabilities:

$$\text{Max } p_{ij}, \quad (4)$$

$$\text{subject to } \sum_i \sum_j p_{ij} = 1, \quad (5)$$

$$(1 - a_{ij})p_{ij} - a_{ij} \sum_{k \neq i} p_{kj} = 0, \quad \sum_i p_{ij} > 0, \quad (6)$$

$$\text{and } p_{ij} \geq 0, \forall i, j. \quad (7)$$

For  $2 \times 2$  tables, for example, there are four equations of the form (6) but linear dependencies reduce the set of these equations to two, each involving two conditional values that add to one and their respective  $p_{ij}$  entries. In addition, the cross-product ratios of all  $2 \times 2$  positive submatrices of  $A$ ,  $B$  and  $\mathbf{n}$  are equal, i.e.,

$$\alpha = \frac{a_{11}a_{22}}{a_{12}a_{21}} = \frac{b_{11}b_{22}}{b_{12}b_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

(see Bishop, Fienberg and Holland (1975)). We can replace some of the constraints in Equation (6) by a linearized version of the odds ratios to obtain the equivalent bounds.

To complete the characterization for two-way tables, we need to consider the third set of combination of marginals and conditionals. For example, for  $\{f(x|y), f(x), I < J\}$ , the following linear programming problem produces optimal solutions:

$$\text{Max } p_{ij}, \quad (8)$$

$$\text{subject to } \sum_i \sum_j p_{ij} = 1, \quad (9)$$

$$\sum_j p_{ij} = p_i, \quad (10)$$

$$a_{ij} = \frac{p_{ij}}{p_{ij} + \sum_{k \neq i} p_{kj}}, \forall i = \{1, 2, \dots, I\}, j = \{1, 2, \dots, J\}, \quad (11)$$

$$\text{and } p_{ij} \geq 0, \forall i = \{1, 2, \dots, I\}, j = \{1, 2, \dots, J\}. \quad (12)$$

**Theorem 3.** For an  $I \times 2$  table given  $f(x|y)$  and  $f(x)$ , sharp upper bounds on the cell probabilities are given by

$$UB = \begin{cases} a_{ij} \frac{p_{i.} - \max_{k \neq j} \{a_{ik}\}}{a_{ij} - \max_{k \neq j} \{a_{ik}\}} & \text{if } p_{i.} \geq a_{ij}, \\ a_{ij} \frac{p_{i.} - \min_{k \neq j} \{a_{ik}\}}{a_{ij} - \min_{k \neq j} \{a_{ik}\}} & \text{if } p_{i.} < a_{ij}, \end{cases} \quad (13)$$

and sharp lower bounds are given by

$$LB = \begin{cases} \max\{0, a_{ij} \frac{p_{i.} - \min_{k \neq j} \{a_{ik}\}}{a_{ij} - \min_{k \neq j} \{a_{ik}\}}\} & \text{s.t. } a_{ij} \frac{p_{i.} - \min_{k \neq j} \{a_{ik}\}}{a_{ij} - \min_{k \neq j} \{a_{ik}\}} \leq UB \} & \text{if } p_{i.} \geq a_{ij}, \\ \max\{0, a_{ij} \frac{p_{i.} - \max_{k \neq j} \{a_{ik}\}}{a_{ij} - \max_{k \neq j} \{a_{ik}\}}\} & \text{s.t. } a_{ij} \frac{p_{i.} - \max_{k \neq j} \{a_{ik}\}}{a_{ij} - \max_{k \neq j} \{a_{ik}\}} \leq UB \} & \text{if } p_{i.} < a_{ij}. \end{cases} \quad (14)$$

These results also generalize to  $I \times J$  tables where  $I, J \geq 3$ , such that the first part of the bounds formula (e.g.  $a_{ij} \frac{p_{i.} - \max_{k \neq j} \{a_{ik}\}}{a_{ij} - \max_{k \neq j} \{a_{ik}\}}$  if  $p_{i.} \geq a_{ij}$ ) holds, but there is an additional factor that we need to adjust the bounds by. For example, in a  $3 \times 4$  table, let  $k$  be the index for the conditional value which satisfies  $\max_{k \neq j} \{a_{ik}\}$ ,  $l$  be the index for the conditional value that satisfies  $\min_{k \neq j} \{a_{ik}\}$ , and  $r \in \{\mathcal{I} \setminus i\}$ , and let

$$d_{ij} = \frac{e_{ij} - g_{ij}}{f_{ij} - g_{ij}},$$

where

$$e_{ij} = \frac{p_{r.} - a_{rk}}{p_{i.} - a_{ik}}, \quad f_{ij} = \frac{a_{rj} - a_{rk}}{a_{ij} - a_{ik}}, \quad g_{ij} = \frac{a_{rl} - a_{rk}}{a_{il} - p_{ik}}.$$

Then the solution for the upper bound is:

$$UB = \begin{cases} a_{ij} \frac{p_{i.} - \max_{k \neq j} \{a_{ik}\}}{a_{ij} - \max_{k \neq j} \{a_{ik}\}} \times d_{ij} & \text{if } p_{i.} \geq a_{ij}, \\ a_{ij} \frac{p_{i.} - \min_{k \neq j} \{a_{ik}\}}{a_{ij} - \min_{k \neq j} \{a_{ik}\}} \times d_{ij} & \text{if } p_{i.} < a_{ij}. \end{cases} \quad (15)$$

More research remains to be done to understand the structures of these bounds and how they extend to  $k$ -way tables.

Given these constraints and the LP approach, in the limit these are the sharpest bounds we can obtain. For inferential purposes this may be sufficient. But these bounds fail to preserve two conditions (when none of the conditional probability values are zero) then:

1. None of the individual cell probabilities (or counts) can be zero, and
2. Cross-product ratios of  $2 \times 2$  subtables cannot be zero or infinite.

Therefore,  $0 \leq p_{ij} \leq a_{ij}$ , for  $f(x|y)$  does not give sharp bounds for the table of counts. Without some other prior information relevant to the observed data, however, we cannot do better using linear programming. The example in Section 4.3 demonstrates how these bounds can lead to a false sense of data security.

**Known sample size  $N$ .** When we know the sample size  $N$ , we can obtain tighter bounds for the cell probabilities, and hence for cell counts, but we are also likely to get fractional bounds by applying linear programming methods. One way to obtain tighter bounds on cell probabilities is assure nonzero values of the entries, e.g., by modifying the constraints in equations (7) and (12) to become  $p_{ij} \geq \frac{1}{N}, \forall i, j$ . However, even this approach is not always sufficient to produce the sharpest bounds on the tables of counts and is likely to lead to fractional bounds on the tables with integer entries as well.

When we know  $N$ , we can also apply integer programming (IP):

$$\text{Max } n_{ij}, \tag{16}$$

$$\text{subject to } \sum_i \sum_j n_{ij} = N, \tag{17}$$

$$(1 - a_{ij})n_{ij} - a_{ij} \sum_{k \neq i} n_{kj} = 0, \quad a_{ij} = \frac{n_{ij}}{\sum_i n_{ij}} \tag{18}$$

$$\text{and } n_{ij} \geq 1, \forall i, j. \tag{19}$$

IP methods such branch-and-bound rely on implicit enumeration of possible solutions and will either give the sharpest bounds for the cell counts or will not find a feasible solution. For larger tables, however, IP methods can be computationally prohibitive.

*Example 2.* Suppose that the only information we have about the data from Table 1 is that  $P(\text{Download}|\text{Gender})$  is given by the matrix  $B$ . The following IP model will give the sharp bounds (see Table 2) for the integer entries of the table.

$$\text{Max } n_{ij}, \tag{20}$$

$$\text{subject to } n_{11} + n_{12} + n_{21} + n_{22} = 50, \tag{21}$$

$$0.4n_{11} - 0.6n_{12} = 0, \tag{22}$$

$$0.8n_{21} - 0.2n_{22} = 0, \tag{23}$$

$$\text{and } n_{ij} \geq 1, \forall i, j. \tag{24}$$

However, conditional frequencies are often reported as floating point approximations. This rounding typically leads to infeasible IP solutions, and we need to apply linear programming relaxation which in turn may again give us fractional bounds that are not tight, as the following example illustrates.

**Table 2.** Data from Table 1 and IP bounds given  $P(\text{Download}|\text{Gender})$

	Download Yes	Download No
Male	15 [3, 27]	10 [2, 18]
Female	5 [1, 9]	20 [4, 36]

*Example 3.* Suppose that the only knowledge we have about the original Table 1, consists of the conditional frequencies

$$P(\text{Gender}|\text{Download}) = a_{ij} = A = \begin{pmatrix} 0.75 & 0.33 \\ 0.25 & 0.67 \end{pmatrix}$$

The left panel in Table 3 gives the LP fractional bounds for this problem. There is a significant discrepancy between these upper and lower bounds for some of the cells when compared with the sharp bounds provided in the right side panel of the same table. Moreover errors due to rounding of the reported conditional frequencies would lead to a very different set of bounds. The issue of rounding in reporting statistical data has been received very limited attention in the disclosure literature and we are just starting to explore its effects on bounds.

**Table 3.** Data from Table 1 with LP relaxation bounds given  $P(\text{Gender}|\text{Download})$  in the left panel and the sharp bounds obtained via Markov basis in the right panel.

	Download Yes	Download No	Download Yes	Download No
Male	15 [3, 35.25]	10 [1, 15.33]	[6, 33]	[2, 14]
Female	5 [1, 11.74]	20 [2, 30.67]	[2, 11]	[4, 28]

## 4.2 Markov Bases and Bounds

We can find feasible solutions to the constrained maximization problem by using tools from computational algebra, such as Gröbner or Markov bases. A set of minimal Markov bases (moves) allows us to build a connected Markov chain and perform a random walk over the space of tables of counts that have the same fixed marginals and/or conditionals. A technical description of calculation and structure of Markov bases given fixed conditional distributions for two-way tables can be found in Slavkovic (2003).

Consider the information provided in *Example 3*. The minimal Markov basis is represented by the following binomial  $n_{11}^9 n_{21}^3 - n_{12}^4 n_{22}^8$ . This implies two possible moves on our  $2 \times 2$  table of counts. These moves together with the sample size  $N$  describe the space of tables containing only four possible tables with non-negative integer counts (see Table 4). This procedure not only gives the sharp bounds listed in right panel of Table 3, but also provides more insight into the structure of the table. A number of possible table realizations could also

be used as a measure of disclosure risk. In this case, the space of tables satisfying the constraints is too small and would easily lead to a full disclosure of any of the cells, and the whole table. These observations hold for a higher dimensional tables as well and could have implications on the assessment of disclosure risk. In particular, they suggest that reporting conditional distributions for a subset of variables may be essentially equivalent from the bounds perspective to the reporting of the corresponding marginal. Although we need to investigate this issue further it may be welcome news since the calculation of Markov bases is typically computationally very expensive for relatively large multi-way tables.

**Table 4.** The space of contingency tables with non-negative integer entries given fixed conditional distribution  $P(\text{Gender}|\text{Download})$

	Y <sub>1</sub>	Y <sub>2</sub>	Total
X <sub>1</sub>	6	14	20
X <sub>2</sub>	2	28	30
Total	8	42	50

	Y <sub>1</sub>	Y <sub>2</sub>	Total
X <sub>1</sub>	15	10	25
X <sub>2</sub>	5	20	25
Total	20	30	50

	Y <sub>1</sub>	Y <sub>2</sub>	Total
X <sub>1</sub>	26	6	30
X <sub>2</sub>	8	12	20
Total	32	18	50

	Y <sub>1</sub>	Y <sub>2</sub>	Total
X <sub>1</sub>	33	2	35
X <sub>2</sub>	11	4	15
Total	44	6	50

### 4.3 Example: Delinquent Children Data

In this section we consider a  $4 \times 4$  table of counts used in the Disclosure Limitation Methodology Report of the Federal Committee on Statistical Methodology (1994). Titles, row and column headings are fictitious. Table 5 shows number of delinquent children by county and education level of household head. We compare outcome of releasing margins versus conditionals on disclosure. First, suppose we consider releasing both margins (row sums and columns sums). we give Fréchet bounds for the counts in this case in the left panel of Table 5. Observe that the cells with small counts (i.e., sensitive cells) are well protected. For example, cell [Alpha, Medium] with count “1” is bounded below by 0 and above by 20. Utilizing tools from computational algebra and Markov bases, one can determine that there are 18,272,363,056 possible tables given the fixed margins.

Next, suppose that instead of margins we only release conditional frequencies:

$$P(\text{Education}|\text{County}) = \begin{pmatrix} 0.750 & 0.050 & 0.150 & 0.050 \\ 0.364 & 0.182 & 0.182 & 0.272 \\ 0.120 & 0.400 & 0.400 & 0.080 \\ 0.343 & 0.400 & 0.200 & 0.057 \end{pmatrix}$$

Integer programming has no feasible solution for this problem and we give the linear programming relaxation bounds in the right panel of Table 5. The bounds are significantly tighter than those associated with releasing the margins,

**Table 5.** Delinquent children data by county and education level. The left panel contains the cell counts and the Fréchet bounds given the margins. The right panel contains the LP relaxation bounds given  $P(\text{Education}|\text{County})$ .

County	Low	Medium	High	Very High	Low	Medium	High	Very High
Alpha	15[0,20]	1[0,20]	3[0,20]	1[0,20]	[15,74.6]	[1,4.97]	[3, 14.9]	[1,4.97]
Beta	20[0,50]	10[0,35]	10[0,30]	15[0,20]	[1.99,30.8]	[1,15.5]	[1,15.5]	[1.5,23.2]
Gamma	3[0,25]	10[0,25]	10[0,25]	2[0,20]	[1.5,11]	[5,36.8]	[5,36.8]	[1,7.36]
Delta	12[0,35]	14[0,35]	7[0,30]	2[0,20]	[6.02,33.27]	[7.02,38.8]	[3.51,19.4]	[1,5.53]

indicating higher disclosure risk. An even more surprising result comes from calculating Markov bases and using the tools from computational algebra to determine the space of tables. In this case, there is *only one* table with non-negative integer entries satisfying the given conditional and the sample size. Hence we have full disclosure of the counts which was masked by the bounds obtained from linear and integer programming. In this case, and in most two-way table with sensitive cells, it is not safe to release the conditionals as they carry almost full information about the table itself.

## 5 Conclusions

To date statistical disclosure limitation methodologies for tables of counts have been heavily intertwined with the release of unaltered marginal totals from such tables, and methods have focused in part on inferences that are possible by an intruder from such releases as well as the use of marginals for inferences about underlying relationships among variables that make up the table, e.g., based on log-linear models. Many statistical agencies also release other forms of summary data from tables, such as tables of rates or observed conditional relative frequencies. These are predominantly released as two-way and three-way tables, with conditioning on a single variable. Preserving conditionals and marginals from a table puts us into a constrained subset of the probability simplex for the the space of possible tables. It is only with the knowledge of sample size  $N$  (or a 1-way marginal in addition to a set of conditionals) that we can calculate the bounds on the actual counts. For two-way tables this often leads to full disclosure, i.e., the complete specification of the original table. Zero cells in a table become zeros again when we condition on one or more of the variables, and thus the release of such conditionals reveals extra information about the full cross-classification.

We are just beginning to understand the implications of rounding in the construction of conditionals on disclosure limitation. But the work we have done to date suggests that in two dimensions the impact of rounding on identification of feasible tables is likely to increase dramatically the disclosure of sensitive cells in the original table.

Finally, we note that the work reported here is illustrative of the kinds of statistical calculations that are possible in higher-way tables with the release of some combination of marginals and conditionals. In some cases the released data reduce to a set of marginals and the results of Dobra and Fienberg (2000, 2002) then can be used directly. In other cases, the release of a conditional instead of a marginal can yield larger bounds and looser inferences about the cells in the table by an intruder. We hope to report on extensions of the methodology introduced here to the multi-way case in the near future.

## Acknowledgments

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

## References

- Arnold, B., Castillo, E. and Sarabia, J.M. Specification of distributions by combinations of marginal and conditional distributions. *Statistics and Probability Letters*, 26:153–157, 1996.
- Arnold, B., Castillo, E. and Sarabia, J.M. *Conditional Specification of Statistical Models*. Springer-Verlag, New York, 1999.
- Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of American Statistical Association*, 92(439):1171–1176, 1997.
- Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36(2):192–236, 1974.
- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge MA.
- Dobra, A. *Statistical Tools for Disclosure Limitation in Multi-Way Contingency Tables*. Ph.D. Thesis, Department of Statistics, Carnegie Mellon University, 2002.
- Dobra, A. and Fienberg, S.E. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences*, 97(22):11885–11892, 2000.
- Dobra, A. and Fienberg, S.E. Bounding entries in multi-way contingency tables given a set of marginal totals. In Y. Haitovsky, H.R. Lerche, and Y. Ritov, editors, *Foundations of Statistical Inference: Proceedings of the Shores Conference 2000*, pages 3–16. Springer-Verlag, Berlin, 2003.
- Dobra, A., Fienberg, S.E. and Trottini, M. Assessing the risk of disclosure of confidential categorical data. In J. Bernardo et al., editors, *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting on Bayesian Statistics*, pages 125–14. Oxford University Press, Oxford, 2003.
- Edwards, D. *Introduction to Graphical Modeling. 2nd Edition*. Springer-Verlag, New York, 2000.

- Federal Committee on Statistical Methodology (1994).  
*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. <http://ntl.bts.gov/docs/wp22.html>.
- Fienberg, S.E., Makov, U.E., Meyer, M.M. and Steele, R.J. Computing the exact distribution for a multi-way contingency table conditional on its marginal totals. In P.K.M.E. Saleh, editor, *Data Analysis from Statistical Foundations: A Festschrift in Honor of the 75th Birthday of D.A.S. Fraser*, pages 145–165. Nova Science Publishers, Huntington, NY, 2001.
- Gelman, A. and Speed, T.P. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society. Series B*, 55(1):185–188, 1993.
- Gelman, A. and Speed, T.P. Corrigendum: Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society. Series B*, 61(2):483, 1999.
- Gutmann, S., Kemperman, H.J.B, Reeds, J.A. and Shepp, L.A. Existence of probability measures with given marginals. *The Annals of Probability*, 19(4):1781–1797, 1991.
- Karr, A.F., Dobra, A., Sanil, A. and Fienberg, S.E. Software systems of tabular data releases. *The International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10:529–544, 2002.
- King, G. *A Solution to the Ecological Inference Problem*. Princeton University Press, Princeton, 1997.
- Lauritzen, S.L. *Graphical Models*. Oxford University Press, Oxford, 1996.
- Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge 2000.
- Pistone, J., Riccomagno, E. and Wynn, H.P. *Algebraic Statistics - Computational Commutative Algebra in Statistics*. Chapman & Hall/CRC, Boca Raton, 2001.
- Rachev, S.T. and Rüschendorf, L. *Mass Transportation Problems*, Volumes 1 and 2. Springer-Verlag, New York, 1998.
- Slavkovic, A.B. Markov bases given fixed conditional distributions for two-way contingency tables. *In preparation*, 2003.
- Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Technical Report*, (R-271), April 2000.
- Whittaker, J. *Graphical Models in Applied Mathematical Multivariate Statistics*. Wiley, New York, 1990.