Bounding Entries in Multi-way Contingency Tables Given a Set of Marginal Totals

Adrian Dobra and Stephen E. Fienberg
Department of Statistics and Center for Automated Learning and Discovery
Carnegie Mellon University, Pittsburgh, PA 15213-3890

October 8, 2001

Abstract

We describe new results for sharp upper and lower bounds on the entries in multi-way tables of counts based on a set of released and possibly overlapping marginal tables. In particular, we present a generalized version of the shuttle algorithm proposed by Buzzigoli and Giusti that computes sharp integer bounds for an arbitrary set of fixed marginals. We also present two examples which illustrate the practical import of the bounds for assessing disclosure risk.

Keywords: Statistical disclosure control; Log-linear models; Decomposable models; Reducible models; Integer programming.

1 Introduction

In this paper, we provide an overview of our recent work to develop bound for entries in contingency and other non-negative tables (see also Dobra and Fienberg (2001)). Our interest in this problem grows out of work to develop a Web-based table query system, coordinated by the National Institute of Statistical Sciences in the spirit of a pilot system described by Keller-McNulty and Unger (1998). The system is being designed to work with a database consisting of a k-way contingency table and it allows only those queries that come in the form of requests for marginal tables. What is intuitively clear from statistical theory is that, as margins are released and cumulated by users, there is increasing information available about the table entries. The system must examine each new query in combination with those previously released margins and decide if the risk of disclosure of individuals in the full unreleased k-way table is too great. Then it might offer one of three responses: (1) yes—release; (2) no—don't release; or perhaps (3) simulate a new table, which is consistent with the previously released margins, and then release the requested margin table from it (c.f., Duncan and Fienberg (1999); Fienberg, Makov, and Steele (1998); Fienberg, Makov, Meyer and Steele (2001)).

There are various approaches to assessing risk of disclosure and most of them relate to the inadvertent "release" of small counts in the full k-way table (e.g. see Skinner and Holmes (1998); Fienberg and Makov (1998); Samuels (1998)). Here we follow the approach of examining upper and lower bounds on the cell entries (see Buzzigoli and Giusti (1999); Cox (1999); Fienberg (1999) and Roehrig, et al. (1999)).

For more general background on related methods of disclosure limitation, we refer the interested reader to Willenborg and de Waal (1996; 2000).

The approach we outline in this paper draws heavily on the ideas associated with the theory of log-linear models for contingency tables (Bishop et al., 1975; Lauritzen, 1996), where the minimal sufficient statistics are in fact marginal totals corresponding to the highest-order terms in the model. In Section 2, we give some technical background and then, in Section 3, we present results from Dobra and Fienberg (2000) corresponding to decomposable and reducible graphical models. Then, in Section 4, we outline a general algorithm that computes sharp bounds for margins corresponding to any standard log-linear model. This algorithm generalizes the "shuttle" algorithm approach suggested by Buzzigoli and Giusti (1999). We apply our results to two examples, a 2^6 table and a 2^{16} table, and we discuss some of the possible implications for disclosure.

2 Technical Background

Bounds for entries in two-way contingency tables go back to seminal papers by Bonferonni (1936), Fréchet (1940), and Hoeffding (1940). For an $I \times J$ table with entries $\{n_{ij}\}$ and row margins $\{n_{i+}\}$ and column margins $\{n_{+j}\}$, these bounds take the form

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}, \tag{1}$$

For simplicity, we refer to these as *Fréchet bounds.* Until recently, the only multi-dimensional generalizations of this result that have been utilized involved non-overlapping fixed marginals. Our interest has been in deriving computationally efficient approaches to computing bounds when the marginals overlap (c.f. the related work described in Joe (1997)).

Any contingency table with non-negative integer entries and fixed marginal totals is a lattice point in the convex polytope $\mathbf{Q}$ defined by the linear system of equations induced by the released marginals. The constraints given by the values in the released marginals induce upper and lower bounds on the interior cells of the initial table. These bounds or *feasibility intervals* can be obtained by solving the corresponding linear programming problems. The importance of systematically investigating these linear systems of equations should be readily apparent. If the number of lattice points in $\mathbf{Q}$ is below a certain threshold, we have significant evidence that a potential disclosure of the entire dataset might have occurred. Moreover, if the induced upper and lower bounds are too tight or too close to the actual sensitive value in a cell entry, the information associated with the individuals classified in that cell may become public knowledge.

The problem of determining sharp upper and lower bounds for the cell entries subject to some linear constraints expressed in this form is known to be NP-hard (Roehrig et al., 1999). Several approaches have been proposed for computing bounds: however, almost all of them have drawbacks that show the need for alternate solutions. Network models (c.f. Cox (1999)) need formal structure to work even for 3-way tables and besides there is no general formulation for higher-way tables. In some ways, the most natural method for solving linear programming problems is the simplex method. For the bounds problem, we would have to run the procedure twice for every element in the table and consequently we overlook the underlying dependencies among the marginals by regarding the maximization/minimization problem associated with some cell as unrelated to the parallel problems associated with the remainder of the cells in the table. Although the simplex method works well for small problems and dimensions, by employing it we would ignore the special structure of the problem because we would consider every table as a linear list of cells. The computational inadequacy of the simplex approach is further augmented by the fact that we may get fractional bounds (Cox, 1999), which are very difficult to interpret. To avoid fractional bounds, one would have to make use of integer programming algorithms, but their computational complexity prevent their usage even for problems of modest size. These considerations suggest the need for more specialized, computationally inexpensive algorithms that could fully exploit the special structure of the problem we are dealing with.

3 Bounds When Marginals Characterize Decomposable and Reducible Graphical Models

We visualize the dependency patterns induced by the released marginals by constructing an independence graph for the variables in the underlying cross-classification. Each variable cross-classified in the table is associated with a vertex in this graph. If two variables are not connected, they are conditionally independent given the remainder. Models described solely in terms of such conditional independencies are said to be *graphical* (e.g., see Lauritzen (1996)).

3.1 Bound Results

Decomposable graphical models have closed form structure and special properties. The expected cell values can be expressed as a function of the fixed marginals. To be more explicit, the maximum likelihood estimates are the product of the marginals divided by the product of the separators. By induction on the number of MSSs, in Dobra and Fienberg (2000), we developed generalized Fréchet bounds for decomposable log-linear models with any number of MSSs. These generalized Fréchet bounds are sharp in the sense that they are the tightest possible bounds given the marginals. In addition, we can determine feasible tables for which these bounds are attained.

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

When the log-linear model associated with the released set of marginals is not decomposable, it is natural to ask ourselves whether we could reduce the computational effort needed to determine the tightest bounds by employing the same strategy used for decomposable graphs, i.e. decompositions of graphs by means of complete separators. An independence graph that is not necessarily decomposable, but still admits a proper decomposition, is called *reducible* (Leimer, 1993). Once again, we point out the link with maximum likelihood estimation in log-linear models. We define a *reducible log-linear model* (Dobra and Fienberg, 2000) as one for which the corresponding MSSs are marginals that characterize the components of a reducible independence graph. If we can calculate the maximum likelihood estimates for the log-linear models corresponding to every component of a reducible graph $\mathcal{G}$, then we can easily derive explicit formulae for the maximum likelihood estimates in the reducible log-linear model with independence graph $\mathcal{G}$ (Dobra and Fienberg, 2000).

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

3.2 Example 1: Risk Factors for Czech Auto Workers

The data in Table 1 come from a prospective epidemiological study of 1841 workers in a Czechoslovakian car factory, as part of an investigation of potential risk factors for coronary thrombosis (see Edwards and Havranek (1985)). In left-hand panel of Table 1, A indicates whether or not the worker "smokes", B corresponds to "strenuous mental work", C corresponds to "strenuous physical work", D corresponds to "systolic blood pressure", E corresponds to "ratio of β and α lipoproteins" and F represents "family anamnesis of coronary heart disease". Assume we are provided with three marginal tables [BF], [ABCE], and [ADE] of this 6-way table. These are the marginals corresponding to a graphical model whose independence graph is given in Fig. 1, and this model fits the data well.

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

Table 1: Czech autoworkers data from Edwards and Havranek (1985). The left-hand panel contains the cell counts and the right-hand panel contains the bounds given the margins [BF], [ABCE], and [ADE].

Using the result from Theorem 1, we see that the upper bounds for the cell entries induced by the marginals [BF], [ABCE], and [ADE] are the minimum of the corresponding entries in the fixed marginals, while the lower bounds are the sum of the same entries minus the sum of the corresponding entries in the marginals associated with the separators of the independence graph, i.e., [B] and [AE]. We give these bounds in the right-hand panel of Table 1. There are three cell entries containing non-zero "small" counts, i.e. counts of "1" and "2" in Table 1. The corresponding bounds are $[0, 25]$, $[0, 38]$ and $[0, 20]$. Since the latter two of these differ, we see that the upper and lower bounds are therefore dependent not only on the fixed marginals, but also on the position they occupy in the cross-classification. Moreover, the bounds for the entry of "1" are wider than the bound for one of the entries of "2". At any rate, all three of these pairs of bounds differ quite substantially and thus we might conclude that there is little chance of identifying the individuals in the small cells.

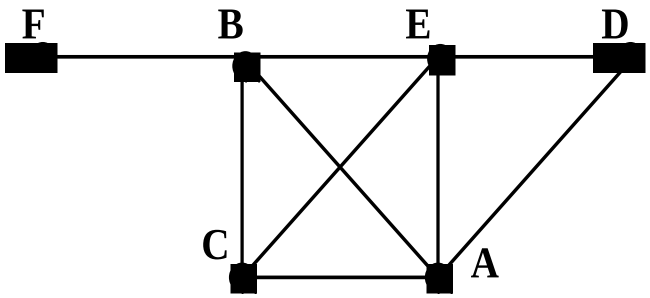

Figure 1: Independence graph induced by the marginals [BF], [ABCE] and [ADE].

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

The first component, [BF], is assumed fixed; hence there is nothing to be done. The other two components are not fixed, however, and we need to compute upper and lower bounds for each of them. Using the algorithm presented in the next section, we calculated bounds for the cell entries in the marginal [ABCE] given the marginals [BC], [BE],[AB], [AC], [AE], [CE] (see Table 2). We did the same for the marginal [ADE] given the marginals [AE], [DE], [AD] (see Table 3).

		B	no		yes		B	no		yes	
E	C	A	no	yes	no	yes	A	no	yes	no	yes
< 3	no		88	62	224	117		[0,206]	[0,167]	[0,404]	[0,312]
	yes		261	246	25	38		[0,421]	[0,463]	[0,119]	[0,119]
≥ 3	no		58	60	170	148		[0,181]	[0,167]	[0,363]	[0,339]
	yes		115	173	20	36		[0,314]	[0,344]	[0,119]	[0,119]

Table 2: Marginal [ABCE] from Table 1 and bounds for this marginal given all 2-way totals.

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

Table 3: Marginal [AED] from Table 1 and bounds for this marginal given all 2-way totals.

Since we have upper and lower bounds for each of the components of a reducible graph, Theorem 2 allows us to piece together the bounds for the components [BF], [ABCE] and [ADE] to obtain sharp integer bounds for the original 6-way table - see Table 4. We emphasize that Theorem 2 is a sound technique for replacing the original problem, namely, computing bounds for a 6-way table, by two smaller ones, i.e., computing bounds for a 4-way and a 3-way table. The computational effort required for implement-

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

Table 4: Bounds for Czech auto-workers data from Table 1 given the marginals [BF], [BC], [BE],[AB], [AC], [AE], [CE], [DE], [AD].

ing and using Theorem 2 is minimal once bounds for the components are available, and thus exploiting it in this fashion could lead to appreciable computational savings.

4 A General Bounds Algorithm

In Section 3, we took advantage of the special structure of the conditional independencies "induced" among the variables cross-classified in a table of counts by the set of fixed marginals. However, if all $(k-1)$-way marginal tables are given, the corresponding independence graph is complete, hence there are no conditional independence relationships to exploit. Fienberg (1999) noted that, if the table is dichotomous, the log-linear model of no kth-order interaction has only one degree of freedom and consequently the counts in any cell can be uniquely expressed as a function of one single fixed cell alone. By imposing the non-negativity constraints for every cell in our contingency table, we are then able to derive sharp upper and lower bounds. It turns out that dichotomous tables are the key to derive sharp bounds for a k-way table given an arbitrary set of fixed marginals.

4.1 Terminology and Notation

Let $\mathbf{T}$ denote the set of cells of all possible tables that could be formed by collapsing the original k-way table $\mathbf{n}$ not only across variables, but also across categories. The elements in $\mathbf{T}$ are essentially blocks formed by joining table entries in $\mathbf{n}$. If the set of cell entries in $\mathbf{n}$ that define a "super-cell" $t_1 \in \mathbf{T}$ is included in the set of cells defining another "super-cell" $t_2 \in \mathbf{T}$, we write $t_1 \prec t_2$. With this partial ordering, $(\mathbf{T}, \prec)$ has a maximal element, namely the grand total of $\mathbf{n}$ and several minimal elements, i.e., the cell entries in the initial table $\mathbf{n}$. The grand total of $\mathbf{n}$ is maximal because all the cells in $\mathbf{n}$ "contribute" to it. On the other hand, a cell entry in $\mathbf{n}$ is minimal in $\mathbf{T}$ since any block of cells in $\mathbf{T}$ is constructed from one single cell in $\mathbf{n}$ or by joining at least two other blocks. One can represent $\mathbf{T}$ as a hierarchy of cells induced by the ordering "$\prec$", with the grand total at the top level and the cells in $\mathbf{n}$ at the bottom level of the hierarchy.

Consider three blocks of cells t_1, t_2 and t_3. If t_2 can be formed by joining t_1 and t_3, we write

$$t_1 \oplus t_3 = t_2. \tag{2}$$

The operator "$\oplus$" is equivalent to joining two blocks of cells in $\mathbf{T}$ to form a third block. The blocks to be joined have to be composed from the same categories in $(k-1)$ dimensions and they are also required not to share any categories in the remaining dimension. If either of these conditions does not hold, their union is not going to be a block of cells in $\mathbf{T}$.

Denote by $L(t)$ and $U(t)$ the current upper and lower bounds for the "super-cell" $t \in \mathbf{T}$. Let

$$L(\mathbf{T}) := \{L(t) : t \in \mathbf{T}\} \text{ and } U(\mathbf{T}) := \{U(t) : t \in \mathbf{T}\}. \tag{3}$$

$L(\mathbf{T})$ and $U(\mathbf{T})$ are the bounds arrays we are trying to determine. Every $t \in \mathbf{T}$ could have a value $V(t)$ assigned to it. If t corresponds to an entry in a fixed marginal, we actually "know" the value $V(t)$ of that entry, hence we set the current lower bound and the current upper bound of t to be the known value $V(t)$.

Let $\mathbf{T}_0$ be the set of cells in $\mathbf{T}$ for which the lower bound is currently equal to the upper bound. These are the cells that have a value assigned to them:

$$V(t) = L(t) = U(t) \Leftrightarrow t \in \mathbf{T}_0. \tag{4}$$

When the iterative procedure described below starts, $\mathbf{T}_0$ will contain only the cells in the fixed marginals. For the remaining cells in $\mathbf{T}$, we could set $L(t)$ and $U(t)$ to be the bounds $L_0(t)$, $U_0(t)$ induced by fixing the one-dimensional marginals of $\mathbf{n}$. These bounds are looser than the bounds we are trying to calculate since it is reasonable to assume that the one-dimensional marginals can be obtained by collapsing the marginals we consider to be fixed. In addition, the log-linear model induced by the one-dimensional marginals is decomposable, hence $L_0(t)$ and $U_0(t)$ can be easily calculated by employing Theorem 1. The intervals $[L(t), V(t)]$, $t \in \mathbf{T}$, are the initial feasibility intervals for the iterative procedure we will describe below.

As the algorithm progresses, the bounds for the cells in $\mathbf{T}$ are improved and more and more cells are added to $\mathbf{T}_0$. To be more precise, "improving" the bounds means decreasing the upper bounds and increasing the lower bounds. When the bounds associated with a cell t become equal, the cell is included in $\mathbf{T}_0$ and is assigned a value $V(t) := L(t) = V(t)$. We are now able to state the bounds problem in a new equivalent form:

> *"Find sharp integer bounds for the cells in* $\mathbf{T}$ *if the values of some cells* $\mathbf{T}_0 \subset \mathbf{T}$ *are fixed."*

4.2 The Generalized Shuttle Algorithm

The fundamental idea behind the "shuttle" algorithm is that the upper and lower bounds for the cells in $\mathbf{T}$ are interlinked. Although Buzzigoli and Giusti (1999) sketched this innovative idea for the 3-way table problem given the three 2-way marginals, they did not accurately identify and exploit the full hierarchical structure of the cells contained in the marginals of a frequency count table $\mathbf{n}$. The method we outline here builds on their approach and sequentially improves the bounds for all the cells we are interested in until no further adjustment can be made.

As before, we assume that, for every cell $t \in \mathbf{T}$, we know a valid lower bound $L(t)$ and a valid upper bound $U(t)$. With these notations, the initial set of fixed cells is

$$\mathbf{T}_0 := \{t \in \mathbf{T} : L(t) = U(t)\}. \tag{5}$$

For all the cells t in $\mathbf{T}_0$, we assign a value $V(t) := L(t) = V(t)$. We let $\mathcal{Q} = \mathcal{Q}(\mathbf{T})$ denote the triplets of cells

$$\mathcal{Q}(\mathbf{T}) := \{(t_1, t_2, t_3) \in \mathbf{T} \times \mathbf{T} \times \mathbf{T} : t_1 \oplus t_3 = t_2\}, \tag{6}$$

which represent the cell dependencies we are trying to satisfy. We sequentially go through all these dependencies and update the upper and lower bounds in the following way. Consider a triplet $(t_1, t_2, t_3) \in \mathcal{Q}$. We have $t_1 \prec t_2$ and $t_3 \prec t_2$. If all three cells have fixed values, i.e., $t_1, t_2, t_3 \in \mathbf{T}_0$, we check whether we came across an inconsistency. The procedure stops if

$$V(t_1) + V(t_3) \neq V(t_2). \tag{7}$$

Assume that $t_1, t_3 \in \mathbf{T}_0$, and $t_2 \notin \mathbf{T}_0$. Then t_2 can only take one value, namely $V(t_1) + V(t_3)$. If $V(t_1) + V(t_3) \notin [L(t_2), U(t_2)]$, we encountered an inconsistency and exit the procedure. Otherwise we set

$$V(t_2) = L(t_2) = U(t_2) := V(t_1) + V(t_3), \tag{8}$$

and include t_2 in the set $\mathbf{T}_0$ of cells having a fixed value. Similarly, if $t_1, t_2 \in \mathbf{T}_0$ and $t_3 \notin \mathbf{T}_0$, t_3 can only be equal to $V(t_2) - V(t_1)$. If $V(t_2) - V(t_1) \notin [L(t_3), U(t_3)]$, we again discovered an inconsistency. If this is not true, we set

$$V(t_3) = L(t_3) = U(t_3) := V(t_2) - V(t_1) \text{ and } \mathbf{T}_0 := \mathbf{T}_0 \cup \{t_3\}. \tag{9}$$

In the case when $t_2, t_3 \in \mathbf{T}_0$ and $t_1 \notin \mathbf{T}_0$, we proceed in an analogous manner. Now we examine the situation when at least two of the cells t_1, t_2, t_3 do not have a fixed value. For each of the three cells not having a fixed value, we update its upper and lower bounds so that the new bounds satisfy the dependency $t_1 \oplus t_3 = t_2$. Suppose $t_1 \notin \mathbf{T}_0$. Then the updated bounds for t_1 will be

$$U(t_1) := \min\{U(t_1), U(t_2) - L(t_3)\} \text{ and } L(t_1) := \max\{L(t_1), L(t_2) - U(t_3)\}. \tag{10}$$

If $t_3 \notin \mathbf{T}_0$, we update $L(t_3)$ and $U(t_3)$ in the same way. Finally, if $t_2 \notin \mathbf{T}_0$, we set

$$U(t_2) := \min\{U(t_2), U(t_1) + U(t_3)\} \text{ and } L(t_2) := \max\{L(t_2), L(t_1) + L(t_3)\}. \tag{11}$$

After updating the bounds of some cell $t \in \mathbf{T}$, we check whether the new upper bound is equal to the new lower bound. If this is true, i.e. $L(t) = U(t)$, we include t in the list of cells having a fixed value:

$$\mathbf{T}_0 := \mathbf{T}_0 \cup \{t\}, \tag{12}$$

and set $V(t) := L(t) = U(t)$. We continue going through all the dependencies in $\mathcal{Q}$ until the upper bounds no longer decrease, the lower bounds no longer increase and no new cells are added to $\mathbf{T}_0$. The procedure will come to an end if and only if an inconsistency is detected or if the upper and lower bounds cannot be subsequently improved. Either one of these two events will eventually occur, hence the procedure we described stops after a finite number of steps.

Unfortunately, the bounds we end up with are not necessarily sharp, except in: (i) the decomposable case, and (ii) the case of a dichotomous k-way table with all $(k-1)$-way marginals fixed. To be more explicit, if the marginals we fix are the MSSs of a decomposable log-linear model, the bounds calculated by the generalized shuttle algorithm will coincide with the bounds obtained by making use of Theorem 1, whereas in case (ii), the generalized shuttle algorithm will successfully determine the best integer bounds by expressing any cell as a function of any other cell, and then imposing the non-negativity conditions on these constraints.

For the general k-way bounds problem with an arbitrary set of fixed marginals, we need to "correct" the bounds by constructing feasible integer tables for which those bounds are actually attained. We explore the space $\mathbf{Q}$ by repeatedly assigning values to the cells in the original table. We do not perform an exhaustive search of $\mathbf{Q}$ since we immediately adjust the upper and lower bounds for the remaining cells in $\mathbf{T}$ once we pick a value for a cell entry, and consequently the values we attempt to assign to a particular cell are chosen from the current feasibility interval associated with that entry. Additional technical details can be found in Dobra (2000a).

We note that each bound can be checked independently of any other bound, hence adjusting the bounds can be done in parallel on a multi-processor machine. The computation time could be further decreased by using the following artifice: once a feasible integer table containing a count equal to a bound for some cell entry is constructed, we check to see whether other upper or lower bounds can also be found in that table. This way, we will not have to attempt to construct another table for these bounds. This simple trick proves to be very efficient in the case of large sparse contingency tables.

4.3 Example 1 Revisited

We have already applied this general algorithm to the separable components of the 6-way Czech auto-worker data in Table 1, to get sharp bounds for a separable table. Here we note what happens in the other special case when no "correction" is required for feasible tables: when all 5-way margins are released. The space of tables $\mathbf{Q}$ in this case contains only two integer tables: the original table $\mathbf{n}$ itself and a second table whose entries are found by adding or subtracting one unit from the corresponding entries in $\mathbf{n}$. Consequently, the feasibility intervals $[L(t), U(t)]$ for all the cells in $\mathbf{n}$ have length one. This means that releasing all 5-way margins could well compromise the confidentiality of the individuals corresponding to the entries containing counts of "1" and "2" and perhaps even the entries containing the count of "3".

4.4 Example 2: The National Long Term Care Survey

Our second example involves a 2^{16} contingency table $\mathbf{n}$ extracted from the "analytic" data file for National Long-Term Care Survey created by the Center of Demographic Studies at Duke University. Each dimension corresponds to a measure of disability defined by an activity of daily leaving, and the table contains information cross-classifying individuals aged 65 and above. For a detailed description of this extract see Erosheva (2000).

We have applied the generalized shuttle algorithm of Section 4.2 to compute sharp upper and lower bounds for the entries in this table corresponding to a number of different sets of fixed marginals. Here we describe one complex calculation for the set involving three fixed 15-way marginals obtained by collapsing $\mathbf{n}$ across the variables "managing money", "taking medicine" and "telephoning".

Of the $2^{16} = 65,536$ cells in the table, $62,384$ contain zero entries. Since the target table is so sparse, releasing three marginals of dimension fifteen will lead to the exact disclosure of most of the cell entries. To be more exact, only 128 cells have the upper bounds strictly bigger than the lower bounds! The difference between the upper and lower bounds is equal to 1 for 96 cells, 2 for 16 cells, 6 for 8 cells, and 10 for 8 cells.

We take a closer look to the bounds associated with "small" counts of "1" or "2". A number of $1,729$ cells contain a count of "1". From these, $1,698$ cells have the upper bounds equal to the lower bounds. The difference between the bounds is 1 for 28 of the remaining counts of "1", is 2 for two other cells and is equal to 6 for only one entry. As for the 499 cells with a count of "2", the difference between the bounds is zero for 485 cells, is 1 for 10 cells and is 2 for 4 other cells.

The generalized shuttle algorithm converged in approximately twenty iterations to the "correct" sharp bounds and it took less than six hours to complete on a single-processor machine at the Department of Statistics, Carnegie Mellon University. We re-checked these bounds by determining the feasible integer tables for which they are attained on the Terascale Computing System at the Pittsburgh Supercomputing Center. We used a parallel implementation of the shuttle algorithm and the computations took almost one hour to complete on fifty-six processors. We are currently exploring ways to speed up the calculations as well as approximations that will allow us to apply our results to larger tables.

5 Conclusions

In this paper we have explained how log-linear model statistical theory can help identify situations when explicit formulas exist for computing the best integer bounds on the entries of a cross-classification of arbitrary dimension given a set of marginal totals (the decomposable case). When such formulas do not exist, we illustrated how to derive similar formulas that help to reduce the computational effort (the reducible case). In addition, we explained how log-linear models provide the basis for correcting the shuttle algorithm originally proposed by Buzzigoli and Giusti, and transform it into a general procedure for computing sharp integer bounds given any set of marginals. The generalized shuttle algorithm described here simultaneously computes sharp integer bounds for all the cells by fully exploiting the structure of the bounds problem for multi-way contingency tables and, in addition, it can update the bounds, as more marginals are being released.

Acknowledgments

Preparation of this paper was supported in part by the National Science Foundation under Grant EIA-9876619 to the National Institute of Statistical Sciences and by the National Institute of Aging under Grant 1090019 to Carnegie Mellon University. The computations were performed in part on the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputing Center.

References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* M.I.T. Press, Cambridge, MA.

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

Buzzigoli, L. and Giusti, A. (1999). "An Algorithm to Calculate the Lower and Upper Bounds of the Elements of an Array Given its Marginals." In *Statistical Data Protection (SDP'98) Proceedings*, 131–147. Eurostat, Luxembourg.

Cox, L. H. (1999). "Some Remarks on Research Directions in Statistical Data Protection." In *Statistical Data Protection (SDP'98) Proceedings*, 163–176. Eurostat, Luxembourg.

Dobra, A. (2000a). "Computing Sharp Integer Bounds for Entries in Contingency Tables Given a Set of Fixed Marginals." Tech. Rep., Department of Statistics, Carnegie Mellon University.

— (2000b). "Measuring the Disclosure Risk for Multi-way Tables with Fixed Marginals Corresponding to Decomposable Log-linear Models." Tech. Rep., Department of Statistics, Carnegie Mellon University.

Dobra, A. and Fienberg, S. E. (2000). "Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs." *Proceedings of the National Academy of Sciences*, 97, 11885–11892.

— (2001). "Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals with Applications to Disclosure Limitation." *Statistical Journal of the United Nations ECE*, 17. Paper presented at the 2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, 14-16 March 2001, Skopje, Macedonia.

Duncan, G. T. and Fienberg, S. E. (1999). "Obtaining Information While Preserving Privacy: a Markov Perturbation Method for Tabular Data." In *Statistical Data Protection (SDP'98) Proceedings*, 351–362. Eurostat, Luxembourg.

Edwards, D. E. and Havranek, T. (1985). "A Fast Procedure for Model Search in Multidimensional Contingency Tables." *Biometrika*, 72, 339–351.

Erosheva, E. (2000). "Grade of Membership and Latent Structure Models with Application to Longitudinal Disability Survey Data." Department of Statistics, Carnegie Mellon University, Unpublished manuscript.

Fienberg, S. E. (1999). "Fréchet and Bonferroni Bounds for Multi-way Tables of Counts with Applications to Disclosure Limitation." In *Statistical Data Protection (SDP'98) Proceedings*, 115–129. Eurostat, Luxembourg.

Fienberg, S. E. and Makov, U. E. (1998). "Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data." *Journal of Official Statistics*, 14, 485–502.

Fienberg, S. E., Makov, U. E., Meyer, M. M., and Steele, R. J. (2001). "Computing the Exact Distribution for a Multi-way Contingency Table Conditional on its Marginals Totals." In *Data Analysis from Statistical Foundations: Papers in Honor of D.A.S. Fraser*, ed. A, Saleh. Nova Science Publishing. In press.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). “Disclosure Limitation Using Perturbation and Related Methods for Categorical Data.” *Journal of Official Statistics*, 14, 485–502.

Fréchet, M. (1940). *Les Probabilitiés, Associées a un Système d’Événments Compatibles et Dépendants,* Vol. Premiere Partie. Hermann & Cie, Paris.

Hoeffding, W. (1940). *Scale-invariant correlation theory*, Vol. 5(3), 181–233. Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman & Hall, New York.

Keller-McNulty, S. and Unger, E. A. (1998). “A Database System Prototype for Remote Access to Information Based on Confidential Data.” *Journal of Official Statistics*, 14, 347–360.

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford.

Leimer, H. G. (1993). “Optimal Decomposition by Clique Separators.” *Discrete Mathematics*, 113, 99–123.

Roehrig, S. F., Padman, R., Duncan, G. T., and Krishnan, R. (1999). “Disclosure Detection in Multiple Linked Categorical Datafiles: a Unified Network Approach.” In *Statistical Data Protection (SDP’98) Proceedings*, 149–162. Eurostat, Luxembourg.

Samuels, S. M. (1998). “A Bayesian, Species-sampling-inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment.” *Journal of Official Statistics*, 14, 373–383.

Skinner, C. J. and Holmes, D. J. (1998). “Estimating the Re-identification Risk Per Record in Microdata.” *Journal of Official Statistics*, 14, 373–383.

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

— (2000). *Elements of Statistical Disclosure Control.* Vol. 155, Lecture Notes in Statistics. Springer-Verlag, New York.