NISS

Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs

Adrian Dobra and Stephen E. Fienberg

Technical Report Number 108
October, 2000

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Bounds for cell entries in contingency tables given marginal totals and decomposable graphs

Adrian Dobra and Stephen E. Fienberg*

Department of Statistics and Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213-3890

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected on April 27, 1999.

Contributed by Stephen E. Fienberg, August 2, 2000

Upper and lower bounds on cell counts in cross-classifications of nonnegative counts play important roles in a number of practical problems, including statistical disclosure limitation, computer tomography, mass transportation, cell suppression, and data swapping. Some features of the Fréchet bounds are well known, intuitive, and regularly used by those working on disclosure limitation methods, especially those for two-dimensional tables. We previously have described a series of results relating these bounds to theory on loglinear models for cross-classified counts. This paper provides the actual theory and proofs for the special case of decomposable loglinear models and their related independence graphs. It also includes an extension linked to the structure of reducible graphs and a discussion of the relevance of other results linked to nongraphical loglinear models.

Fréchet bounds | loglinear models | reducible graphs | disclosure limitation

1. Introduction

Upper and lower bounds on cell counts in cross-classifications of positive counts given certain marginal totals play important roles in a number of the disclosure limitation procedures, e.g., see the various papers in the 1998 special issue of *The Journal of Official Statistics* (1). In that context, if a cell count is small and the upper bound is "close" to the lower bound, the intruder knows with certainty that there is only a small number of individuals possessing the characteristics corresponding to the cell and this may pose an undo risk of disclosure of the identity of these individuals. Similarly, such bounds also arise in a variety of other contexts including mass transportation problems (2), computer tomography (3), ecological inference in the social sciences (4), causal inference in imperfect experiments (5), and are the focus of the probabilistic literature on copulas (6). Much of the work on this problem has been focused on bounds in the case when the marginal totals are nonoverlapping.

The class of bounds we describe is a generalization of bounds usually attributed to Fréchet (7), whose original presentation was in terms of cumulative distribution functions (c.d.f.) for a random vector $(D_1, D_2, \ldots, D_m)$ in R^m:

$$F_{1,2,\ldots,m}(x_1, x_2, \ldots, x_m) = \Pr(D_1 \leq x_1, D_2 \leq x_2, \ldots, D_m \leq x_m), \quad \textbf{[1]}$$

which are essentially equivalent to contingency tables when the underlying variables are categorical. For example, suppose we have a two-dimensional table of counts, $\{n_{ij}\}$ adding up to the total $n_{++} = n$. If we normalize each entry by dividing by n and then create a table of partial sums, by cumulating the proportions from the first row and first column to the present ones, we have a set of values of the form **[1]**. Thus, Fréchet bound results for distribution functions correspond to bounds for the cell counts where the values $\{x_i\}$ in **[1]** represent "cut-points" between categories for the ith categorical variable. Bonferroni (8) and Hoeffding (9) independently developed related results on bounds.

We are interested in the following generalization of the Bonferroni–Fréchet–Hoeffding bounds. Consider a k-dimensional contingency table n_K arranged as a linear list of m counts. The random variable assigned to the ith cell will be denoted Y_i. Let $\mathcal{S}$ be a system of nonempty subsets of $\{1, 2, \ldots, m\}$, such that $\cup_{S \in \mathcal{S}} S = \{1, 2, \ldots, m\}$. The *Fréchet class* $\mathcal{F}(\mathcal{S})$ (6) is the class of m-variate distributions with fixed marginals $\{F_S : S \in \mathcal{S}\}$, where F_S is the joint c.d.f. of random variables $\{Y_i : i \in \mathcal{S}\}$. Because the indices of the margins being fixed might be overlapping, we have to impose a *consistency* constraint, namely

$$\pi_{S_1 \cap S_2} F_{S_1} = \pi_{S_1 \cap S_2} F_{S_2}, \text{ whenever } S_1, S_2 \in \mathcal{S}, S_1 \cap S_2 \neq \emptyset,$$

where π_S means integrating out the variables that do not appear in S. Following Rüschendorf (10), for a measurable function $\phi: \mathbf{R}^m \to \mathbf{R}$, we define $M(\phi) = \sup\{\int \phi \mathrm{d}F : F \in \mathcal{F}(\mathcal{S})\}$ and $m(\phi) = \inf\{\int \phi \mathrm{d}F : F \in \mathcal{F}(\mathcal{S})\}$. Our goal is to determine $M(\phi)$ and $m(\phi)$ in the particular case when ϕ is the identity function on the set $\mathbf{R} \times \ldots \times (-\infty, y_i] \times \ldots \times \mathbf{R}$. This is equivalent to determining sharp upper and lower bounds for the ith cell in the cross-classification n_K, given the marginals $\{n_S : S \in \mathcal{S}\}$.

Fienberg (11) noted that there is an intimate link between bounds for non-negative cell entries in a cross-classification subject to marginal constraints, and maximum likelihood estimates for the same cell entries under the loglinear model whose minimal sufficient statistics are the margins. This link seems especially clear in the special case of cross-classifications of non-negative counts and loglinear models for their expectations that are *decomposable*, i.e., for tables where estimated expected values can be *explicitly* written as a function of the marginal totals (e.g., see refs. 12–14). Such models are a special subclass of the graphical loglinear models (e.g., see refs. 14 and 15), and these models are representable in terms of graphs that display conditional independence relationships. We present the results here in terms of graphs and explain how they apply to the more general situation. In the next section, we introduce some basic notation for the corresponding theory of *decomposable graphs*. Then, in Section 3, we give results on Fréchet bounds when the margins correspond to those that characterize decomposable loglinear models. Sections 4 and 5 extend the approach to reducible graphs and provide some explicit examples. In the final section, we present some conjectures on how these bound results can be extended to cases corresponding to bounds for cross-classifications that are not quite representable in graphical form but that utilize our results for reducible graphs.

Abbreviations: c.d.f., cumulative distribution function; PEO, perfect elimination ordering; MCS, maximum cardinality search; MLE, maximum likelihood estimate; mp-, maximal prime.

*To whom reprint requests should be addressed. E-mail: fienberg@stat.cmu.edu.

2. Basic Graph Theory Results

In this section, we begin with some basic definitions and notations for graphs and then define decomposable graphs and present some results that characterize them.

2.1. Graph Terminology. A *graph* is a pair $\mathcal{G} = (V, E)$, where V is a finite set of vertices and $E \subseteq V \times V$ is a set of edges linking the vertices. Our interest is in *undirected* graphs, for which $(u, v) \in E$ implies $(v, u) \in E$. For any vertex set $A \subseteq V$, we define the edge set associated with it as

$$E(A) := \{(u,v) \in E | u,v \in A\}.$$

Let $\mathcal{G}(A) = (A, E(A))$ denote the subgraph of $\mathcal{G}$ induced by A. The *section graph* $\mathcal{G}\backslash A := \mathcal{G}(V\backslash A)$ is the subgraph of $\mathcal{G}$ obtained by removing a set of vertices $A \subset V$ from the graph. Two vertices $u, v \in V$ are *adjacent (neighbors)* if $(u, v) \in E$. A set of vertices of $\mathcal{G}$ is *independent* if no two of its elements are adjacent. The boundary $\mathrm{bd}(A)$ of a subset of vertices $A \subset V$ is the set of vertices in $V\backslash A$ adjacent to at least one vertex in A:

$$\mathrm{bd}(A) := \{v \in V | v \notin A \text{ and } (u,v) \in E \text{ for some vertex } u \in A\}.$$

The *closure* of $A \subset V$ is $\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$. An induced subgraph $\mathcal{G}(A)$ is *complete* if the vertices in A are pairwise adjacent in $\mathcal{G}$. We also say that A is *complete* in $\mathcal{G}$. A complete vertex set A in $\mathcal{G}$ that is maximal is a *clique*.

Let $u, v \in V$. A *path* (or *chain*) from u to v is a sequence $u = v_0, \ldots, v_n = v$ of distinct vertices such that $(v_{i-1}, v_i) \in E$ for all $i = 1, 2, \ldots, n$. The path is a *cycle* if the end points are allowed to be the same, $u = v$. If there is a path from u to v we say that u and v are *connected*. The sets $A, B \subset V$ are *disconnected* if u and v are not connected for all $u \in A$, $v \in B$. The *connected component* of a vertex $u \in V$ is the set of all vertices connected with u. A graph is *connected* if all the pairs of vertices are connected.

The set $C \subset V$ is an *uv-separator* if all paths from u to v intersect C. The set $C \subset V$ *separates* A from B if it is an *uv*-separator for every $u \in A$, $v \in B$. C is a *separator* (*cut-set*) of $\mathcal{G}$ if two vertices in the same connected component of $\mathcal{G}$ are in two distinct connected components of $\mathcal{G}\backslash C$ or, equivalently, if $\mathcal{G}\backslash C$ is disconnected. In addition, C is a *minimal* separator of $\mathcal{G}$ if C is a separator and no proper subset of C separates the graph. Unless otherwise stated, the separators we work with will be complete.

Consider a connected graph $\mathcal{G} = (V, E)$ having a clique separator C, and let $V_1, \ldots, V_s$ be the vertex sets of the connected components of $\mathcal{G}\backslash C$. The subgraphs $\mathcal{G}(V_1 \cup C), \ldots, \mathcal{G}(V_s \cup C)$ are the *leaves* of $\mathcal{G}$ produced by C. A graph is *bipartite* if its set of vertices can be partitioned into two disjoint subsets V_1 and V_2 such that every edge of the graph connects between a vertex of V_1 and a vertex of V_2, i.e. V_1 and V_2 are independent sets. A *tree* is a connected graph with no cycles. It has n vertices and $n - 1$ edges. In a tree, there is a unique path between any two vertices.

2.2. Decomposable Graphs. *Decomposable* graphs possess the special property that allows us to "decompose" them into components or subgraphs and work directly with these components. They also allow us to make use of divide-and-conquer techniques to solve any type of problem associated with such a graphical structure. The idea is to decompose the graph $\mathcal{G}$ in two possibly overlapping subgraphs $\mathcal{G}'$ and $\mathcal{G}''$ so that no structural information of the graph is lost when transforming $\mathcal{G}$ into $\mathcal{G}'$ and $\mathcal{G}''$. Furthermore, by "correctly" decomposing $\mathcal{G}'$ and $\mathcal{G}''$, and so on, one ends up with a set of subgraphs of $\mathcal{G}$ that allow for no further decompositions. A set of subgraphs of $\mathcal{G}$ generated in this way is called a *derived system* of $\mathcal{G}$, while its elements are called *atoms* (16). If one does not lose any information along the way in the decomposition, then one can solve problems for each atom and then put together the component solutions to solve a combined problem for the initial graph $\mathcal{G}$. But first we need to define what we mean by "correct" decomposition.

Definition 1: The partition (A_1, A_2, A_3) of V is said to form a *decomposition* of $\mathcal{G}$ if A_2 is a minimal separator of A_1 and A_3.

In this case (A_1, A_2, A_3) *decomposes* $\mathcal{G}$ into the *components* $\mathcal{G}(A_1 \cup A_2)$ and $\mathcal{G}(A_2 \cup A_3)$. The decomposition is *proper* if A_1 and A_3 are not empty. If A_2 is empty, A_1 and A_3 form two nonoverlapping connected components.

Throughout the remainder of this section, we will assume that the graphs we work with are connected. No loss of generality is incurred because all the results can be applied to a disconnected graph by applying them successively to each connected component. We follow closely Blair and Barry (17) and Lauritzen (18).

Definition 2: The graph $\mathcal{G}$ is *decomposable* if it is complete or if there exists a proper decomposition (A_1, A_2, A_3) into decomposable graphs $\mathcal{G}(A_1 \cup A_2)$ and $\mathcal{G}(A_2 \cup A_3)$.

Because we require a proper decomposition of the graph at every step, the components $\mathcal{G}(A_1 \cup A_2)$ and $\mathcal{G}(A_2 \cup A_3)$ have fewer vertices than the original graph $\mathcal{G}$, hence the procedure will stop after a finite number of steps. The smallest nondecomposable graph is a cycle with four vertices.

Definition 3: A vertex $v \in V$ is *simplicial* in $\mathcal{G} = (V, E)$ if $\mathrm{bd}(v)$ is a clique.

If $v \in V$ is simplicial in $\mathcal{G}$ and $\mathcal{G}$ is not complete, $(\{v\}, \mathrm{bd}(v), V\backslash\mathrm{cl}(v))$ is a proper decomposition of $\mathcal{G}$. Simplicial vertices have very nice and useful properties:

LEMMA 1. (*i*) *A vertex is simplicial if and only if it belongs to precisely one clique.* (*ii*) *Any decomposable graph has at least one simplicial vertex.*

The importance of simplicial vertices in describing the structure of decomposable graphs will soon become apparent. Assume that the graph $\mathcal{G}$ has n vertices. An *ordering* of $\mathcal{G}$ is a bijection from the vertex set V to a set of labels $\{1, 2, \ldots, n\}$. Let $v_1, v_2, \ldots, v_n$ be an ordering of the vertex set V. The *monotone adjacency set* of v_i is given by:

$$\mathrm{madj}(v_i) = \mathrm{bd}(v_i) \cap \{v_{i+1}, v_{i+2}, \ldots, v_n\}. \qquad [2]$$

There is a special class of orderings of $\mathcal{G}$ that plays a central role in the characterization of decomposable graphs.

Definition 4: The ordering $v_1, v_2, \ldots, v_n$ is a perfect elimination ordering (PEO) if v_i is simplicial in the graph $\mathcal{G}(\{v_i, v_{i+1}, \ldots, v_n\})$ for every $i = 1, 2, \ldots, n$.

Any decomposable graph is characterized by the possession of a PEO, as the next result shows.

THEOREM 1. *A graph $\mathcal{G}$ is decomposable if and only if $\mathcal{G}$ has a perfect elimination ordering.*

The *maximum cardinality search* algorithm (MCS) is a linear-time procedure for generating a perfect elimination ordering. It starts with an arbitrary vertex $v \in V$ for which it sets $v = v_n$. The next vertex will be labeled $n - 1$ and will be one of the unlabeled vertices with the maximum number of labeled neighbors. The ordering $v_1, v_2, \ldots, v_n$ generated by continuing in this way will always be a PEO if the input graph is decomposable.

Let $\mathcal{C}(\mathcal{G}) = \{C_1, C_2, \ldots, C_p\}$ be the set of cliques of a decomposable graph $\mathcal{G}$ and $v_1, v_2, \ldots, v_n$ be a PEO obtained by applying the MCS algorithm. We will refer to v_{i_q} as the *representative vertex* of C_q whenever $C_q = \{v_{i_q}\} \cup \mathrm{madj}(v_{i_q})$. The following result shows how MCS can efficiently generate the cliques in $\mathcal{C}(\mathcal{G})$ by identifying their representative vertices.

THEOREM 2. [*Blair and Barry* (17).] *Let $v_1, v_2, \ldots, v_n$ be a PEO obtained by applying the MCS algorithm to a connected decomposable graph $\mathcal{G}$. Then $\mathcal{C}(\mathcal{G})$ contains precisely the following*

sets: $\{v_1\} \cup \text{madj}(v_1)$ *and* $\{v_{i+1}\} \cup \text{madj}(v_{i+1})$, $1 \le i \le n-1$, *for which* $|\text{madj}(v_i)| \le |\text{madj}(v_{i+1})|$.

Because MCS labels the vertices of $\mathcal{G}$ in decreasing order, the cliques also will be generated in a decreasing order with respect to the labels of their representative vertices. More explicitly, assume that $v_{i_1}, v_{i_2}, \ldots, v_{i_p}$ are the representative vertices of the cliques $C_1, C_2, \ldots, C_p$, respectively, where $i_1 > i_2 > \ldots > i_p$. The MCS algorithm finds the cliques in $\mathcal{C}(\mathcal{G})$ in the order $C_1, C_2, \ldots, C_p$. We need to introduce one additional class of sets.

Definition 5: Let $V_1, \ldots, V_k$ be a sequence of subsets of the vertex set of a graph $\mathcal{G} = (V, E)$. Let $H_j = V_1 \cup \ldots \cup V_j$, $S_j = H_{j-1} \cap V_j$, and $R_j = V_j \backslash H_{j-1}$. The sequence is said to be *perfect* if (*i*) for all $j > 1$, there is an $i < j$ such that $S_j \subseteq V_i$, and (*ii*) the sets S_j are complete for all j.

The first condition in *Definition 5* is known as the *running intersection property*. The sets S_j are called the *separators* of the sequence.

THEOREM 3. [*Lauritzen (14).*] *Let $V_1, \ldots, V_k$ be a perfect sequence of sets that contains all cliques of a graph $\mathcal{G}$. Then for every j, S_j separates $H_{j-1} \backslash S_j$ from R_j in $\mathcal{G}(H_j)$ and hence $(H_{j-1} \backslash S_j, S_j, R_j)$ decomposes $\mathcal{G}(H_j)$.*

A total ordering $C_1, C_2, \ldots, C_p$ of the cliques in $\mathcal{C}(\mathcal{G})$ generated by the MCS algorithm will always have the running intersection property (17). Because $C_1, C_2, \ldots, C_p$ are complete in $\mathcal{G}$, the vertex sets $S_j = (C_1 \cup \ldots \cup C_{j-1}) \cap C_j$ will also be complete, and consequently $C_1, C_2, \ldots, C_p$ is a perfect sequence of sets. By recursively applying *Theorem 3*, we obtain that $\mathcal{C}(\mathcal{G})$ is a derived system of $\mathcal{G}$, whereas S_j ($j = 2, \ldots, p$) is the corresponding sequence of separators [c.f. the recursive result described by Rüschendorf (10)]. We note that, although a clique can appear only once in $\mathcal{C}(\mathcal{G})$, a separator can appear more than once in $\mathcal{C}(\mathcal{G})$. Therefore, $\mathcal{S}(\mathcal{G})$ is not really a set, but a "multiset" of separators (17).

3. Generalized Fréchet Bounds for Decomposable Loglinear Models

Let $X = (X_1, X_2, \ldots, X_k)$ be a vector of discrete random variables. Denote $K = \{1, 2, \ldots, k\}$ the index set associated with $X_1, X_2, \ldots, X_k$. The random variable X_j can take the values $x_j \in \{1, 2, \ldots, I_j\}$, for $j = 1, 2, \ldots, k$. Let $J_K = I_1 \times I_2 \times \ldots \times I_k$ and $x = (x_1, x_2, \ldots, x_k) \in J_K$.

Consider the k-way contingency table $n_K := \{n_K(x)\}_{x \in J_K}$. We let $a = \{i_1, i_2, \ldots, i_p\}$ denote an arbitrary subset of K, and we define X_a as the ordered tuple $X_a = (X_i; i \in a)$. Similarly, we denote $J_a = J_{i_1} \times J_{i_2} \times \ldots \times J_{i_p}$. The marginal table of counts $n_a := \{n_a(x_a)\}_{x_a \in J_a}$ corresponding to X_a is given by

$$n_a(x_a) = \sum_{x_{K \backslash a} \in J_{K \backslash a}} n_K(x_a, x_{K \backslash a})$$

We write n_{ab} instead of $n_{a \cup b}$, where $a, b \subseteq K$. The grand total of the complete table is $n_\emptyset$.

Assume we are given m possibly overlapping marginal tables $n_{C_1}, n_{C_2}, \ldots, n_{C_p}$ such that $C_1 \cup C_2 \cup \ldots \cup C_p = K$. Moreover, $C_1, C_2, \ldots, C_p$ are the cliques of a decomposable graph $\mathcal{G} = (K, E)$. Let $S_2, \ldots, S_p$ be the separators associated with $(C_j)_j$. Every S_j is included in some clique C_i, hence the marginals $n_{S_2}, \ldots, n_{S_p}$ will also be fixed.

The class of Fréchet bounds we present is linked with the theory of decomposable loglinear models. We think of every vertex $i \in K$ of $\mathcal{G}$ as being associated with a variable X_i. The structural information embedded in $\mathcal{G}$ might be interpreted in the following way: If S separates A_1 and A_2 in $\mathcal{G}$, then X_{A_1} is conditionally independent of X_{A_2} given X_S. The loglinear model with minimal sufficient statistics $C_1, C_2, \ldots, C_p$ will be decomposable because its independence graph $\mathcal{G}$ is decomposable, and consequently the maximum likelihood estimates (MLEs) will exist and can be expressed in a closed form (14, 15). We develop explicit formulas for the tightest upper and lower bounds for the cell counts in the cross-classification n_K provided that the marginals $n_{C_1}, n_{C_2}, \ldots, n_{C_p}$ are known by employing a similar machinery to the one used for developing formulas for MLEs for a decomposable loglinear model. This machinery provides us with the tools we need for extending the usual Fréchet bounds to more complicated graphical structures.

We begin with a slightly more general statement of the original Fréchet bound result (2, 11).

THEOREM 4. (FRÉCHET). (*i*) *Let $a_1, a_2 \subseteq K$ such that $(a_1 \backslash a_2, a_1 \cap a_2, a_2 \backslash a_1)$ is a proper decomposition of the graph $\mathcal{G}$ $(a_1 \cup a_2)$. Then the following inequality holds:*

$$\min\{n_{a_1}, n_{a_2}\} \ge n_{a_1 a_2} \ge \max\{n_{a_1} + n_{a_2} - n_{a_1 \cap a_2}, 0\}. \quad \textbf{[3]}$$

(*ii*) *The above inequality provides sharp bounds for the cells in the contingency table $n_{a_1 \cup a_2}$ given the marginals n_{a_1} and n_{a_2}.*

If two vertex sets are in two distinct connected components, they are separated by the empty set. It is not hard to see that *Theorem 4* implies the following result.

COROLLARY 1. (*i*) *If a_1 and a_2 are two disjoint subsets of K, we have*

$$\min\{n_{a_1}, n_{a_2}\} \ge n_{a_1 a_2} \ge \max\{n_{a_1} + n_{a_2} - n_\emptyset, 0\}.$$

(*ii*) *The above inequality provides sharp bounds for the cells in the contingency table $n_{a_1 \cup a_2}$ given the marginals n_{a_1} and n_{a_2}.*

This immediately generalizes to a graph with any number of connected components.

THEOREM 5. (*i*) *Let $\{a_1, a_2, \ldots, a_m\}$ denote the set of connected components of the graph $\mathcal{G}(\cup_{i=1}^m a_i)$. Then the following is true*:

$$\min\{n_{a_1}, n_{a_2}, \ldots, n_{a_m}\} \ge n_{a_1 a_2 \cdots a_m} \ge \max\left\{\sum_{i=1}^m n_{a_i} - (m-1) \cdot n_\emptyset, 0\right\}. \quad \textbf{[4]}$$

(*ii*) *The above inequality provides sharp bounds for the cells in the contingency table $n_{\cup_{i=1}^m a_i}$ given the marginals $n_{a_1}, n_{a_2}, \ldots, n_{a_m}$.*

We are now ready to explore the situation when the minimal sufficient statistics of a decomposable loglinear model define a connected graph.

THEOREM 6. *Suppose $\mathcal{G} = (K, E)$ is connected and decomposable. Let $\mathcal{C}(\mathcal{G}) = \{C_1, C_2, \ldots, C_p\}$ the set of cliques of $\mathcal{G}$ ordered in a perfect sequence and $\mathcal{S}(\mathcal{G}) = \{S_2, \ldots, S_p\}$ the corresponding set of separators. Then*

$$\min\{n_{C_1}, n_{C_2}, \ldots, n_{C_p}\} \ge n_K \ge \max\left\{\sum_{i=1}^p n_{C_i} - \sum_{i=2}^p n_{S_i}, 0\right\}, \quad \textbf{[5]}$$

and these are sharp bounds for the cells in the contingency table n_K given the marginals $n_{C_1}, \ldots, n_{C_p}$.

Proof: By induction. If $\mathcal{G}$ decomposes in $p = 2$ cliques, then Eq. **5** is a direct consequence of *Theorem 4*. Suppose we know that Eq. **5** holds for any connected decomposable graph with $p - 1$ cliques. We want to prove Eq. **5** for a graph with p cliques.

Theorem 3 tells us that $(H_{p-1} \backslash S_p, S_p, R_p)$ is a decomposition of the graph $\mathcal{G}(H_p) = \mathcal{G}$. By using *Theorem 4*, we obtain

$$\min\{n_{H_{p-1}}, n_{C_p}\} \ge n_K \ge \max\{n_{H_{p-1}} + n_{C_p} - n_{S_p}, 0\}. \quad \textbf{[6]}$$

The cliques of $\mathcal{G}(H_{p-1})$ are $C_1, C_2, \ldots, C_{p-1}$, and this is a perfect sequence in $\mathcal{G}(H_{p-1})$. From the induction assumption that we made, we have

$$\min\{n_{C_1}, n_{C_2}, \ldots, n_{C_{p-1}}\} \geq n_{H_{p-1}} \geq \max\left\{\sum_{i=1}^{p-1} n_{C_i} - \sum_{i=2}^{p-1} n_{S_i}, 0\right\}. \quad [7]$$

By combining Eqs. **6** and **7**, we obtain the desired Eq. **5**. Again, because the bounds in Eq. **6** are the tightest possible for the counts in table n_K, and the same is true for the bounds in Eq. **7** for the cell counts in table $n_{H_{p-1}}$, we conclude that the bounds in Eq. **5** are also the tightest bounds for the counts in table n_K.

Buzzigoli and Giusti (18) proposed an algorithm, which they call *the shuttle algorithm*, that alternates iteratively between upper and lower bounds, and that when applied to decomposable structures appears to indirectly exploit the structure implicit in *Theorem 6*. But it does not achieve the sharp bounds in as computationally efficient fashion as we can by using the formula directly.

At this point we succeeded in developing formulas for the sharpest bounds when the sets of indices defining the known marginals define a connected decomposable graph. However, the connectivity assumption is not by any means essential. We can extend the definition of decomposable graphs to include disconnected graphs with all their connected components decomposable. By employing the maximum cardinality search algorithm sequentially for every connected component, we can determine the set of cliques of such a disconnected decomposable graph as the union of the sets of cliques associated with the connected components. The corresponding set of separators can be obtained in the same way.

The next result provides an explicit formula for the generalized Fréchet bounds associated with an arbitrary decomposable graphical structure. We emphasize that the generalized Fréchet bounds are sharp bounds given the information that we assumed we have.

THEOREM 7. (*i*) *Let* $\mathcal{G} = (K, E)$ *be a decomposable graph. Then the following inequality is true*:

$$\min\{n_C | C \in \mathcal{C}(\mathcal{G})\} \geq n_K \geq \max\left\{\sum_{C \in \mathcal{C}(\mathcal{G})} n_C - (m-1)\cdot n_\emptyset - \sum_{S \in \mathcal{S}(\mathcal{G})} n_S, 0\right\}, \quad [8]$$

where $\mathcal{C}(\mathcal{G})$ *is the set of cliques of* $\mathcal{G}$, $\mathcal{S}(\mathcal{G})$ *is the set of separators associated with* $\mathcal{C}(\mathcal{G})$, *and m is the number of connected components of the graph* $\mathcal{G}$. (*ii*) *The above inequality provides sharp bounds for the cells in the contingency table* n_K *given the marginals* $\{n_C | C \in \mathcal{C}(\mathcal{G})\}$.

Proof: We apply *Lemma 6* for each connected component of $\mathcal{G}$, then *Theorem 5* to combine the resulting inequalities. All the bounds for the marginal tables involved are tight, hence the bounds in Eq. **8** will also be tight.

4. Reducible Graphs

By exploiting decomposability in an appropriate manner, we have been able to find sharp bounds for cell counts when some special sets of marginals characterizing decomposable loglinear models are given. It is natural to ask ourselves whether we could develop similar results for reducible graphs, as described in refs. 16 and 19.

Definition 6: A graph $\mathcal{G}$ is *reducible* if $\mathcal{G}$ admits a proper decomposition, otherwise $\mathcal{G}$ is a *prime* graph.

Any complete graph is prime, whereas any disconnected graph is reducible. By definition, the atoms contained in a derived system of a graph are all prime. Given that every reducible graph $\mathcal{G}$ might have several derived systems (16), we would like to be able to isolate one of them that could fully characterize the input graph $\mathcal{G}$.

Definition 7: A subgraph $\mathcal{G}(A)$ is a *maximal prime* (mp-) subgraph of $\mathcal{G}$, if $\mathcal{G}(A)$ is prime and $\mathcal{G}(B)$ is reducible for all B with $A \subset B \subseteq V$.

The set of mp-subgraphs of $\mathcal{G}$ is contained in every derived system of $\mathcal{G}$. Moreover, the set of mp-subgraphs of $\mathcal{G}$ is always a derived system of $\mathcal{G}$ (19), and consequently it is the *unique minimal derived system*. If $\mathcal{G}$ is decomposable, the mp-subgraphs of $\mathcal{G}$ are complete, hence the unique minimal derived system of a decomposable graph contains only its cliques (19).

Section 2 describes a procedure for finding the mp-subgraphs of a decomposable graph. The order in which the MCS algorithm identifies the mp-subgraphs along with the set of separators are needed to reconstruct the original graph from its minimal derived system. We would like to devise a similar decomposition algorithm for the more general case when the input graph is reducible, not necessarily decomposable.

It is easy to see that any decomposable graph is reducible, but the converse is not true, as we will prove next. Gavril (20) introduced the family of *clique separable graphs* in the following recursive manner.

Definition 8: $\mathcal{G} = (V, E)$ is a *clique-separable graph* if (*i*) $\mathcal{G}$ is a Type 1 or Type 2 graph, or (*ii*) $\mathcal{G}$ has a separator C, and the leaves of $\mathcal{G}$ produced by C are clique-separable graphs.

A graph $\mathcal{G}$ is a Type 1 graph if its vertex set can be partitioned in two subsets V_1, V_2, such that $|V_1| \geq 3$, $\mathcal{G}(V_1)$ is a connected bipartite graph, V_2 is complete, and every vertex of V_1 is adjacent to every vertex of V_2. In addition, $\mathcal{G} = (V, E)$ is a Type 2 graph if there exists a partition $V_1, \ldots, V_k$ of V, such that $V_1, \ldots, V_k$ are independent sets in $\mathcal{G}$, and every vertex of V_i is adjacent to every vertex of V_j, for $i \neq j$.

By definition, any decomposable graph is also clique-separable, and any clique-separable graph is reducible. However, Type 2 graphs are clique-separable but obviously they are not necessarily decomposable, hence the class of reducible graphs is much richer than the class of decomposable graphs.

Tarjan (16) has proposed an O(nm)-time method for decomposing a reducible graph with n vertices and m edges. The downside of Tarjan's algorithm is that it generates an arbitrary derived system of prime graphs. Leimer (19) has adapted this algorithm so that the input graph is decomposed exactly into its mp-subgraphs. A reducible graph $\mathcal{G}$ might have several separators that would induce a proper decomposition of $\mathcal{G}$. If we could select the "right" separator at every step of the decomposition procedure, then we would manage to avoid including nonmaximal prime subgraphs in the final derived system.

Definition 9: [Leimer (19).] Let (A_1, A_2, A_3) be a decomposition of $\mathcal{G}$ into the subgraphs $\mathcal{G}' = \mathcal{G}(A_1 \cup A_2)$ and $\mathcal{G}'' = \mathcal{G}(A_2 \cup A_3)$. If the mp-subgraphs of $\mathcal{G}'$ and $\mathcal{G}''$ are pairwise different and if they are all mp-subgraphs of $\mathcal{G}$, then (A_1, A_2, A_3) is called a *P-decomposition* and A_2 is called a *P-separator.*

Moreover, a decomposition (A_1, A_2, A_3) is a P-decomposition if and only if $\mathcal{G}(A_2)$ is not an mp-subgraph of any of the graphs $\mathcal{G}(A_1 \cup A_2)$ and $\mathcal{G}(A_2 \cup A_3)$ (19). If a graph has a decomposition, then it also has a P-decomposition. Therefore it is possible to decompose a reducible graph by means of P-separators, and in this case we are guaranteed to obtain the minimal derived system of maximal prime subgraphs.

Assume that we somehow managed to order the vertex sets of the mp-subgraphs $\mathcal{G}(V_1), \ldots, \mathcal{G}(V_k)$ of a graph $\mathcal{G}$ in a perfect sequence. By using the same notations as before, we have the following result.

THEOREM 8. [*Leimer* (19).] *$(H_{k-1}\backslash S_k, S_k, R_k)$ is a P-decomposition of $\mathcal{G}$ into $\mathcal{G}' = \mathcal{G}(H_{k-1})$ and the prime graph $\mathcal{G}'' = \mathcal{G}(V_k)$. $\mathcal{G}(V_1), \ldots, \mathcal{G}(V_{k-1})$ are the mp-subgraphs of $\mathcal{G}'$ and $V_1, \ldots, V_{k-1}$ is a perfect sequence of sets in $\mathcal{G}'$.*

Theorem 8 can be applied recursively to generate a derived system of $\mathcal{G}$. Because the decompositions performed along the way are P-decompositions, the minimal derived system of $\mathcal{G}$ will be generated.

We are interested in the existence of a perfect sequence of the mp-subgraphs of a graph only for proving the correctness of our results. The ordering of the mp-subgraphs is not relevant when computing the generalized Fréchet bounds, and consequently, in an actual implementation of our algorithms, we would only have to obtain the set $\mathcal{V}(\mathcal{G})$ of mp-subgraphs along with the corresponding sequence $\mathcal{S}(\mathcal{G})$ of separators.

Leimer (19) has suggested an alternative approach that would allow us accomplish this task by taking advantage of the MCS algorithm we previously presented. The first step would be to transform a connected reducible graph $\mathcal{G} = (V, E)$ in a closely related decomposable graph by adding extra edges in E. We would like to keep the number of edges added to a minimum, so that a minimal decomposable graph is derived.

Definition 10: [Tarjan (16).] Let π be an ordering of the vertex set of a graph $\mathcal{G} = (V, E)$. The fill-in F_π caused by the ordering π is the set of edges:

$$F_\pi = \{(u, v) \mid \quad u \neq v,\ (u, v) \notin E, \text{ and there is a path}$$
$$u = v_0, v_1, \ldots, v_k = v \text{ in } \mathcal{G} \text{ such that}$$
$$\pi(v_i) < \min\{\pi(u), \pi(v)\} \text{ for } i = 1, \ldots, k-1\}. \quad \textbf{[9]}$$

The graph $\mathcal{G}_\pi = (V, E \cup F_\pi)$ is called the *minimal fill-in graph* if there does not exist a numbering π' of $\mathcal{G}$ with $F_{\pi'} \subset F_\pi$. It can be shown that the fill-in graph $\mathcal{G}_\pi$ is decomposable for any numbering π of $\mathcal{G}$. Algorithms for generating a minimal fill-in graph can be found in Ohtsuki and Cheung (21).

The second step consists of applying the maximum cardinality search algorithm to the minimal fill-in graph $\mathcal{G}_\pi$ associated with the input graph $\mathcal{G}$. However, we will not employ the "original" maximum cardinality search algorithm. We will make use instead of an expanded version (17) that can find the set $\mathcal{C}(\mathcal{G}_\pi) = \{C_1, C_2, \ldots, C_r\}$ of cliques of $\mathcal{G}_\pi$ along with the associated system $\mathcal{S}(\mathcal{G}_\pi) = \{S_2, \ldots, S_r\}$ of separators by constructing a tree $\mathcal{T}_\pi = (\mathcal{C}(\mathcal{G}_\pi), \mathcal{E}_{\tau_\pi})$. We assume that the sequence $C_1, C_2, \ldots, C_r$ is perfect. For every clique $C_j, j > 1$, we choose a "parent" clique $C_i, i < j$ such that $S_j \subset C_i$, and include the edge (C_j, C_i) in $\mathcal{E}_{\tau_\pi}$. Because the parent of a clique might not be unique, more than one tree could be constructed on $\mathcal{C}(\mathcal{G}_\pi)$. Moreover, C_1 cannot have a parent and will be called the *root* of the tree. This is certainly not a restriction because every clique can be C_1 in some perfect sequence. The tree $\mathcal{T}_\pi$ generated by the MCS algorithm has the additional property that $S \subset V$ is a minimal vertex separator of $\mathcal{G}_\pi$ if and only if $S = C_j \cap C_i$ for some edge $(C_j, C_i) \in \mathcal{E}_{\tau_\pi}$. Consequently, the set of separators associated with $\mathcal{C}(\mathcal{G}_\pi)$ will be given by $\mathcal{S}(\mathcal{G}_\pi) = \{C_i \cap C_j : (C_i, C_j) \in \mathcal{E}_{\tau_\pi}\}$. Then $S \in \mathcal{S}(\mathcal{G}_\pi)$ will also be a minimal separator in $\mathcal{G}$ if S is complete in $\mathcal{G}$.

The last step of the algorithm is presented below in pseudocode. With every clique $C \in \mathcal{C}(\mathcal{G}_\pi)$, we associate a vertex set $\Delta(C)$. Initially we set $\Delta(C) \leftarrow C$ for all $C \in \mathcal{C}(\mathcal{G}_\pi)$. A clique C is *terminal* in $\mathcal{T}_\pi$ if C is not the parent of any other clique, i.e., if there is no such C' with $(C', C) \in \mathcal{E}_{\tau_\pi}$.

- $\mathcal{V}(\mathcal{G}) \leftarrow \emptyset$; $\mathcal{S}(\mathcal{G}) \leftarrow \emptyset$;
- **while** $\mathcal{E}_{\tau_\pi} \neq \emptyset$ do

  1. Identify a terminal clique C_j;
  2. $\mathcal{C}(\mathcal{G}_\pi) \leftarrow \mathcal{C}(\mathcal{G}_\pi)\backslash\{C_j\}$;
  3. $\mathcal{E}_{\tau_\pi} \leftarrow \mathcal{E}_{\tau_\pi}\backslash\{(C_j, C_i)\}$;
  4. **if** $C_j \cap C_i$ is complete in $\mathcal{G}$ then
     $\mathcal{V}(\mathcal{G}) \leftarrow \mathcal{V}(\mathcal{G}) \cup \{\Delta(C_j)\}$;
     $\mathcal{S}(\mathcal{G}) \leftarrow \mathcal{S}(\mathcal{G}) \cup \{C_j \cap C_i\}$;
     else
     $\Delta(C_i) \leftarrow \Delta(C_i) \cup \Delta(C_j)$;
     end while

- $\mathcal{V}(\mathcal{G}) \leftarrow \mathcal{V}(\mathcal{G}) \cup \{\Delta(C_1)\}$.

This algorithm provides a computational approach for identifying the maximal prime subgraphs $\mathcal{V}(\mathcal{G})$ of an arbitrary connected reducible graph $\mathcal{G}$, along with its associated system $\mathcal{S}(\mathcal{G})$ of separators. We utilize it in the following section.

5. Generalized Fréchet Bounds for Reducible Loglinear Models

In Section 3, we showed that we can explicitly determine the tightest bounds for the cells in a table of counts n_K given a set of marginals when that set of marginals define a decomposable graph $\mathcal{G} = (K, E)$. When the graph associated with some set of marginals is not decomposable, we have no choice but to employ iterative methods such as the simplex algorithm. Generally speaking, linear programming methods are computationally expensive and might yield results that are very difficult to interpret, so they should be used with care. The natural question to ask is 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.

To be more specific, assume we want to determine the bounds for a contingency table n_K given the marginals $n_{C_1}, n_{C_2}, \ldots, n_{C_p}$. In addition, $C_1, C_2, \ldots, C_p$ are the cliques of the graph $\mathcal{G} = (K, E)$. $\mathcal{G}$ is assumed to be reducible, not necessarily decomposable. Let $V_1, V_2, \ldots, V_q$ be the maximal prime subgraphs of $\mathcal{G}$ ordered in a perfect sequence, and let $S_2, S_3, \ldots, S_q$ be the sequence of separators associated with $V_1, V_2, \ldots, V_q$. Suppose we could compute tight bounds for the marginals $n_{V_1}, n_{V_2}, \ldots, n_{V_q}$ given $n_{C_1}, n_{C_2}, \ldots, n_{C_p}$, i.e., we know $n^U_{V_1}, n^U_{V_2}, \ldots, n^U_{V_q}$ and $n^L_{V_1}, n^L_{V_2}, \ldots, n^L_{V_q}$ such that

$$n^U_{V_j} \geq n_{V_j} \geq n^L_{V_j}, \text{ for all } j = 1,2, \ldots, q. \quad \textbf{[10]}$$

Because S_j is complete in $\mathcal{G}$, there will exist an $i \in \{1, 2, \ldots, p\}$ such that $S_j \subseteq C_i$. Hence n_{S_j} is a marginal table of n_{C_i}. Therefore, once we fixed $n_{C_1}, n_{C_2}, \ldots, n_{C_p}$, the marginals $n_{S_2}, \ldots, n_{S_q}$ will also be fixed. With the notations introduced above, we develop explicit formulas for sharp bounds for the cells counts in table n_K.

THEOREM 9. *Suppose $\mathcal{G} = (K, E)$ is connected and reducible. The tightest bounds for the cell counts in the contingency table n_K given the marginals $n_{C_1}, n_{C_2}, \ldots, n_{C_p}$ are given by*

$$\min\{n^U_{V_1}, n^U_{V_2}, \ldots, n^U_{V_q}\} \geq n_K \geq \max\left\{\sum_{i=1}^{q} n^L_{V_i} - \sum_{i=2}^{q} n_{S_i}, 0\right\}. \quad \textbf{[11]}$$

Proof: Because $V_1, V_2, \ldots, V_q$ is a derived system of $\mathcal{G}$, we could think about the subgraphs $\mathcal{G}(V_1), \ldots, \mathcal{G}(V_q)$ as being the cliques of a connected decomposable graph $\mathcal{G}'$. Moreover, $S_2, \ldots, S_q$ will be the system of separators associated with $V_1, V_2, \ldots, V_q$ in $\mathcal{G}'$. By employing *Theorem 6*, we obtain

$$\min\{n_{V_1}, n_{V_2}, \ldots, n_{V_q}\} \geq n_K \geq \max\left\{\sum_{i=1}^{q} n_{V_i} - \sum_{i=2}^{q} n_{S_i}, 0\right\}. \quad \textbf{[12]}$$

Then Eq. **11** follows immediately from Eqs. **12** and **10**. The bounds for the marginal tables involved are all sharp, hence the bounds in Eq. **11** will also be tight.

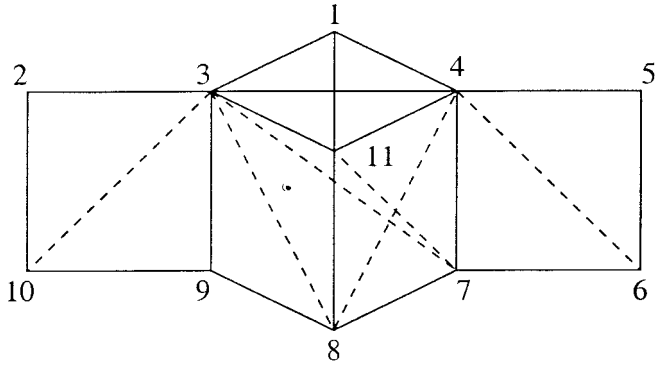

Fig. 1. A graph and its minimal fill-in set of edges.

Once again, we will point out the link with maximum likelihood estimation in loglinear models. We define a *reducible loglinear model* as one for which the corresponding minimal sufficient statistics are margins that characterize the maximal prime subgraphs of a reducible graph. Assuming that one has calculated maximum likelihood estimates for the loglinear models determined by the independence graphs $\mathcal{G}(V_1),\mathcal{G}(V_2), \ldots ,\mathcal{G}(V_q)$, then one can easily derive explicit formulae for the maximum likelihood estimates in the reducible loglinear model with independence graph $\mathcal{G}$. By employing results of Lauritzen (14), we find that

$$\hat{m}(i) = \frac{\prod_{j=1}^{q} m(i_{V_j})}{\prod_{j=2}^{q} n(i_{S_j})}, \qquad [13]$$

(c.f. the special cases given in ref. 12).

We continue the analogy with the decomposable case we previously discussed by considering a reducible disconnected graph. We know how to find the maximal prime subgraphs (along with the corresponding sequence of separators) successively for every connected component. The set of mp-subgraphs for the complete graph is defined as the union of the sets of mp-subgraphs of every connected component. The set of separators can be determined in a similar way. We are now ready for the main result of the paper. However, we are going to postpone presenting it for the moment.

5.1. Example. To clarify the concepts and the results presented so far, we use an example similar to the one proposed by Tarjan (16). The graph $\mathcal{G}$ in Fig. 1 has 11 vertices and 17 edges represented by continuous lines. We want to determine the mp-subgraphs of $\mathcal{G}$. The edge {3, 9} is a separator for {1, 3, 4, 5, 6, 7, 8, 9, 11} and {2, 3, 9, 10}. The latter is a four-cycle, hence cannot be further decomposed, and because it is not a complete, $\mathcal{G}$ cannot be decomposable. Similarly, {4, 7} separates {1, 3, 4, 7, 8, 9, 11} and {4, 5, 6, 7}. Again, the latter is a four-cycle, hence it is a prime subgraph. Now the clique {1, 3, 4, 11} is separated from {3, 4, 7, 8, 9, 11} by the triangle {3, 4, 11}. The subgraph $\mathcal{G}(\{3, 4, 7, 8, 9, 11\})$ does not have a separator, therefore we have finished decomposing $\mathcal{G}$. The set of mp-subgraphs is $\mathcal{V}(\mathcal{G}) = \{\{2, 3, 9, 10\}, \{4, 5, 6, 7\}, \{1, 3, 4, 11\}, \{3, 4, 7, 8, 9, 11\}\}$, whereas the sequence of separators is $\mathcal{S}(\mathcal{G}) = \{\{3, 9\}, \{4, 7\}, \{3, 4, 11\}\}$.

Next we illustrate how to obtain $\mathcal{V}(\mathcal{G})$ and $\mathcal{S}(\mathcal{G})$ by using the decomposition algorithm from Section 4. The minimal fill-in graph $\mathcal{G}_\pi$ is obtained by adding six new edges to $\mathcal{G}$. These edges are represented with dotted lines in Fig. 1. The cliques of $\mathcal{G}_\pi$ are $C_1 = \{1, 3, 4, 11\}$, $C_2 = \{3, 4, 7, 11\}$, $C_3 = \{3, 7, 8, 11\}$, $C_4 = \{4, 6, 7\}$, $C_5 = \{4, 5, 6\}$, $C_6 = \{3, 8, 9\}$, $C_7 = \{3, 9, 10\}$, and $C_8 = \{2, 3, 10\}$. The tree $\mathcal{T}_\pi$ constructed by the MCS algorithm on $\mathcal{C}(\mathcal{G}_\pi) = \{C_1, C_2, \ldots , C_8\}$ has edges

$$\mathcal{E}_\pi = \{(C_2,C_1), (C_3,C_2), (C_4,C_2), (C_5,C_2), (C_6,C_3), (C_7,C_6), (C_8,C_7)\}.$$

We proceed to the last step of the algorithm. The clique C_5 is terminal, but $C_5 \cap C_4 = \{4, 6\}$ is not complete in $\mathcal{G}$, hence we set $\Delta(C_4) = \{4, 5, 6, 7\}$. After eliminating C_5 from the clique tree, C_4 becomes terminal. Because $S_1 = C_4 \cap C_2 = \{4, 7\}$ is complete in $\mathcal{G}$, we identified the first mp-subgraph $V_1 = \Delta(C_4)$ and its associated separator S_1. We eliminate C_4 from $\mathcal{T}_\pi$, and the algorithm proceeds in a similar manner.

The set $\mathcal{C}(\mathcal{G})$ of cliques is essentially the set of edges of $\mathcal{G}$ from which we take out {1, 3}, {1, 4}, {1, 11}, {4, 11}, {3, 4}, {3, 11}, and then add {1, 3, 4, 11}. Assume we want to determine upper and lower bounds for a cross-classification n_K with 11 dimensions. Given the marginal tables $\{n_C : C \in \mathcal{C}(\mathcal{G})\}$, it is possible to compute sharp bounds for the marginal tables corresponding to the mp-subgraph of $\mathcal{G}$. Because the separators in $\mathcal{S}(\mathcal{G})$ are subsets of some cliques, they will define marginals of some tables in $\{n_C : C \in \mathcal{C}(\mathcal{G})\}$, hence it is possible to make use of Theorem 9 to calculate sharp bounds for the cell counts in table n_K.

5.2. Bounds for Reducible Loglinear Models. The foregoing example indicates that *Theorem 9* is applicable in a more general setting than the one we previously suggested. Determining bounds for cell counts in a cross-classification given the marginals defined by the set of cliques is equivalent to the problem of calculating the MLEs of a graphical loglinear model. The minimal sufficient statistics of a graphical log-linear model define a graph, and the cliques of this graph are exactly the minimal sufficient statistics. If the minimal sufficient statistics are not cliques in the associated graph, the model is not graphical.

For example, suppose we have a table n_K corresponding to the graph in Fig. 1. Now assume we don't have access to the marginal $n_{[1,3,4,11]}$, but instead we do know $n_{[1,3]}$, $n_{[1,4]}$, $n_{[1,11]}$ and also $n_{[3,4,11]}$. These marginals no longer correspond to the cliques of a graph. Yet it is still possible to compute sharp bounds for the marginals determined by the mp-subgraphs of $\mathcal{G}$, and then to combine these bounds using Theorem 9 to obtain tight bounds for the complete table n_K.

To be more explicit, suppose we are provided with a set of marginals $n_{D_1}, n_{D_2}, \ldots , n_{D_r}$ that define a graph $\mathcal{G} = (K, E)$. We have $K = \cup_{i=1}^{r} D_i$ and $E = \{(u, v) : \{u, v\} \subset D_j$, for some $j = 1, \ldots , r\}$. The mp-subgraphs of $\mathcal{G}$ are $\mathcal{V}(\mathcal{G}) = \{V_1, V_2, \ldots , V_q\}$, whereas the corresponding sequence of separators is $\mathcal{S}(\mathcal{G}) = \{S_2, \ldots , S_q\}$. We emphasize that $(D_j)_j$ do not have to be the set of cliques of $\mathcal{G}$ and that $\mathcal{G}$ is not necessarily connected. However, we need to impose one additional constraint, namely for every S_i, there is a $j \in \{1, 2, \ldots , r\}$ such that $S_i \subset D_j$. This implies that the marginals $n_{S_2}, \ldots , n_{S_q}$ will be fixed once $n_{D_1}, n_{D_2}, \ldots , n_{D_r}$ are fixed. With these notations, we announce a more general version of *Theorem 9*.

THEOREM 10. *Let $\mathcal{G} = (K, E)$ be a reducible graph. Then the following inequality is true*:

$$\min\{n_V^U | V \in \mathcal{V}(\mathcal{G})\} \geq n_K \geq \max\left\{ \sum_{V \in \mathcal{V}(\mathcal{G})} n_V^L - (m-1)\cdot n_\emptyset - \sum_{S \in \mathcal{S}(\mathcal{G})} n_S, 0 \right\}, \qquad [14]$$

where $\mathcal{V}(\mathcal{G})$ is the set of maximal prime subgraphs of $\mathcal{G}$, $\mathcal{S}(\mathcal{G})$ is the set of separators associated with $\mathcal{V}(\mathcal{G})$, and m is the

number of connected components of the graph $\mathcal{G}$. In addition, $\{n^U_V | V \in \mathcal{V}(\mathcal{G})\}$ and $\{n^L_V | V \in \mathcal{V}(\mathcal{G})\}$ are the tightest upper and lower bounds for the marginal tables $\{n_V | V \in \mathcal{V}(\mathcal{G})\}$, respectively.

Proof: Because $\mathcal{V}(\mathcal{G})$ is a derived system of the graph $\mathcal{G}$, we can think about the subgraphs $\{\mathcal{G}(V) | V \in \mathcal{V}(\mathcal{G})\}$ as being the derived system of cliques of a graph $\mathcal{G}'$. In this case $\mathcal{S}(\mathcal{G})$ will be the set of separators associated with $\mathcal{V}(\mathcal{G})$ in $\mathcal{G}'$, hence Eq. **14** follows immediately from Eq. **8**.

6. Conclusions

The results described in this paper are part of a programmatic effort to understand and operationalize the computation of upper and lower bounds for non-negative entries in cross-classifications subject to a set of marginal constraints. From research on mass transportation and other versions of this problem, we know that the computational problem is typically characterized as being *NP*-complete, and thus we cannot expect to find a simple approach that will deal effectively with the bound calculation problem, especially in high dimensions. Thus, instead of attempting to utilize a general computational approach such as linear programming or the simplex algorithm (22) or network methods (23, 24), we have opted to exploit the structure of the underlying probability structures based on statistical and mathematical theory.

In particular, we have worked with the graphical representation of probability distributions subject to conditional independence relationships and utilized existing results on decomposable graphs to derive explicit bounds for cell entries when the given marginals correspond to the maximal cliques of a decomposable graph. Our approach was motivated by the more specialized results for decomposable loglinear models for tables of counts where the minimal sufficient statistics are marginals and the expected cell values are explicit functions of them.

We also have extended the bound results from the decomposable to the reducible case, and this allows us to exploit other results and computational approaches for bounds applied to subtables corresponding to the reducible components that are not cliques. The results of Section 5 focus on the cases where tables still have a graphical representation representing conditional independence relationships. But there are many other probability structures where we would like to be able to calculate bounds but are not graphical in this sense. For example, a k-dimensional probability distribution given all $(k-1)$-dimensional marginals is not graphical, but we are still able to exploit statistical theory to compute upper and lower bounds in this case. Fienberg (11) outlines an approach for doing this in the $k = 3$ case, and Dobra and Fienberg (25) provide detailed algorithms for $k > 3$. Suppose that one wants to compute bounds for a cross-classification that has a structure similar to that in the reducible case, except that we replace a d-dimensional nonclique by a d-dimensional probability distribution given all $(d-1)$-dimensional marginals. Then we can combine the bounds computed for this nongraphical distribution using the reducible representation of Section 5.

Cox (26) raised a very interesting question, namely, whether one can actually construct a feasible table with a prescribed set of possibly overlapping margins. The solution to the feasibility problem is straightforward if the marginal tables constitute the maximal cliques of a decomposable graph. In this case, the explicit formulas for calculating the MLEs of the associated loglinear model provide us with a feasible table. In addition, if the set of margins are the minimal sufficient statistics of a reducible loglinear model, Eq. **13** tells us how to construct a feasible table given a consistent set of marginals associated with the maximal prime subgraphs of the induced independence graph. Therefore, the results substantially reduce the computational effort needed to solve the feasibility problem by reducing it to a number of smaller and hopefully easier-to-solve problems.

These results represent only a small part of those needed to allow the computation of upper and lower bounds for high-dimensional cross-classifications of the sort that arise in disclosure limitation and other practical problems.

Preparation of this paper was supported in part by the U.S. Bureau of the Census and the National Science Foundation under Grant EIA-9876619 to the National Institute of Statistical Sciences.

Appendix A

We give below a proof of *Theorem 4*.

Proof: Let $a_1 = \{1, \ldots, p\}$ and $a_2 = \{q, \ldots, m\}$ where $1 \leq q \leq p \leq m$. Let $n_{i_1^0 \ldots i_m^0}$ be an arbitrary cell in the table $n_{a_1 \cup a_2}$. To avoid confusion, the marginals $n_{a_1}, n_{a_2}, n_{a_1 \cap a_2}$ will be denoted by A^1, A^2, A^{12}, respectively. The following equalities should hold:

$$\sum_{i_{p+1}, \ldots, i_m} n_{i_1^0 \ldots i_p^0 i_{p+1} \ldots i_m} = A^1_{i_1^0 \ldots i_p^0},$$

$$\sum_{i_1, \ldots i_{q-1}} n_{i_1 \ldots i_{q-1} i_q^0 \ldots i_m^0} = A^2_{i_q^0 \ldots i_m^0},$$

$$\sum_{i_1, \ldots, i_{q-1}, i_{p+1}, \ldots, i_m} n_{i_1 \ldots i_{q-1} i_q^0 \ldots i_p^0 i_{p+1} \ldots i_m} = A^{12}_{i_q^0 \ldots i_p^0}.$$

Consider the sums:

$$\sum_{\{i_{p+1}, \ldots, i_m\} \neq \{i_{p+1}^0, \ldots, i_m^0\}} n_{i_1^0 \ldots i_p^0 i_{p+1} \ldots i_m} = D_1,$$

$$\sum_{\{i_1, \ldots, i_{q-1}\} \neq \{i_1^0, \ldots, i_{q-1}^0\}} n_{i_1 \ldots i_{q-1} i_q^0 \ldots i_m^0} = D_2,$$

$$\sum_{\substack{\{i_1, \ldots, i_{q-1}, i_{p+1}, \ldots, i_m\} \\ \neq \{i_1^0, \ldots, i_{q-1}^0, i_{p+1}^0, \ldots, i_m^0\}}} n_{i_1 \ldots i_{q-1} i_q^0 \ldots i_p^0 i_{p+1} \ldots i_m} = D_{12}.$$

With these notations, we have

$$n_{i_1^0 \ldots i_m^0} = A^1_{i_1^0 \ldots i_p^0} - D_1 = A^2_{i_q^0 \ldots i_m^0} - D_2 = A^{12}_{i_q^0 \ldots i_p^0} - D_{12}. \quad \textbf{[15]}$$

We can write $D_{12} = D_1 + D_2 + D_3$, where

$$\sum_{\substack{\{i_{p+1}, \ldots, i_m\} \neq \{i_{p+1}^0, \ldots i_m^0\} \\ \{i_1, \ldots, i_{q-1}\} \neq \{i_1^0, \ldots, i_{q-1}^0\}}} n_{i_1 \ldots i_{q-1} i_q^0 \ldots i_p^0 i_{p+1} \ldots i_m} = D_3.$$

It follows that

$$n_{i_1^0 \ldots i_m^0} = A^1_{i_1^0 \ldots i_p^0} + A^2_{i_q^0 \ldots i_m^0} - A^{12}_{i_q^0 \ldots i_p^0} + D_3. \quad \textbf{[16]}$$

Clearly $D_1, D_2, D_3 \geq 0$. From Eqs. **15** and **16**, we deduce

$$A^1_{i_1^0 \ldots i_p^0} + A^2_{i_q^0 \ldots i_m^0} - A^{12}_{i_q^0 \ldots i_p^0} \leq n_{i_1^0 \ldots i_m^0} \leq \min\{A^1_{i_1^0 \ldots i_p^0}, A^2_{i_q^0 \ldots i_m^0}\},$$

which concludes the proof.

1. Various authors (1998) *J. Off. Stat.* **14,** Special Issue 4.
2. Rachev, S. T. & Rüschendorf, L. (1998) *Mass Transportation Problems* (Springer, New York), Vols. 1 and 2.
3. Gutmann, S., Kemperman, J. H. B., Reeds, J. A. & Shepp, L. A. (1991) *Ann. Probabil.* **19,** 1781–1797.
4. King, G. (1997) *A Solution to the Ecological Inference Problem* (Princeton Univ. Press, Princeton).
5. Balke, A. & Pearl, J. (1997) *J. Am. Stat. Assoc.* **92,** 1172–1176.
6. Joe, H. (1997) *Multivariate Models and Dependence Concepts* (Chapman & Hall, New York).

7. Fréchet, M. (1940) *Les Probabilitiés, Associées a un Système d'Événments Compatibles et Dépendants* (Hermann & Cie, Paris), Vol. Premiere Partie.
8. Bonferroni, C. E. (1936) *Teoria Statistica delle Classi e Calcolo delle Probabilitá* (Publicazioni del R. Instituto Superiore di Scienze Economiche e Commerciali di Firenze, Florence, Italy), Vol. 8, pp. 1–62.
9. Hoeffding, W. (1940) *Scale-Invariant Correlation Theory* (Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin, Berlin), Vol. 5, pp. 181–233.
10. Rüschendorf, L. (1991) in *Stochastic Orders and Decision Under Risk* (Institute of Mathematical Statistics Lecture Notes—Monograph Series), Vol. 19, pp. 285–310.
11. Fienberg, S. E. (1999) in *Statistical Data Protection, Proceedings of the Conference (Lisbon, 25 to 27 March 1998)* (Eurostat, Luxembourg), pp. 115–129.
12. Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice* (MIT Press, Cambridge, MA).
13. Haberman, S. J. (1974) *The Analysis of Frequency Data* (Univ. of Chicago Press, Chicago).
14. Lauritzen, S. L. (1996) *Graphical Models* (Clarendon, Oxford).
15. Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics* (Wiley, New York).
16. Tarjan, R. E. (1985) *Discrete Math.* **55,** 221–232.
17. Blair, J. R. S. & Barry, P. (1993) in *Graph Theory and Sparse Matrix Computation* (Springer, New York), Vol. 56, pp. 1–30.
18. Buzzigoli, L. & Giusti, A. (1999) in *Statistical Data Protection, Proceedings of the Conference (Lisbon, 25 to 27 March 1998)* (Eurostat, Luxembourg), pp. 131–147.
19. Leimer, H. G. (1993) *Discrete Math.* **113,** 99–123.
20. Gavril, F. (1977) *Discrete Math.* **19,** 159–165.
21. Ohtsuki, T., Cheung, L. K. & Fujisawa, T. (1976) *J. Math. Anal. Appl.* **54,** 622–633.
22. Roehrig, S. F., Padman, R., Duncan, G. T. & Krishnan, R. (1999) in *Statistical Data Protection, Proceedings of the Conference (Lisbon, 25 to 27 March 1998)* (Eurostat, Luxembourg), pp. 149–162.
23. Cox, L. H. (1980) *J. Am. Stat. Assoc.* **75,** 377–385.
24. Cox, L. H. (1995) *J. Am. Stat. Assoc.* **90,** 1453–1462.
25. Dobra, A. & Fienberg, S. E. (2000) *Computing Bounds for Entries in k-dimensional Cross-Classifications Given All (k − 1)-Dimensional Marginals* (Department of Statistics Technical Report, Carnegie Mellon University, New York).
26. Cox, L. H. (1999) in *Statistical Data Protection, Proceedings of the Conference (Lisbon, 25 to 27 March 1998)* (Eurostat, Luxembourg), pp. 163–176.