Markov Bases for Decomposable Graphical Models

Adrian Dobra
National Institute of Statistical Sciences
Research Triangle Park, NC 27709-4006

February 13, 2002

Abstract

The underlying connection between disclosure avoidance techniques for categorical data and sampling from the exact conditional distribution of a table of counts given a set of fixed marginals is the Markov basis that links all the contingency tables having that set of marginals. In this paper we show that primitive data swaps or moves are the only moves that have to be included in a Markov basis that preserve a set of fixed marginals, when these marginals are the minimal sufficient statistics of a decomposable log-linear model.

Keywords: Contingency tables; Decomposable log-linear models; Disclosure limitation; Exact distributions; Gröbner bases; Markov chain Monte Carlo; Markov bases.

1 Introduction

Statistical agencies collect information from survey respondents in the form of microdata. Every variable recorded in the database is assumed to be categorical, i.e. it can take a relatively small number of values or categories. By aggregating the records associated with the same combination of categories, one obtains a *frequency count table*. A *cell entry* in a frequency count table is a positive integer representing the number of units or individuals sharing the same attributes. To fully exploit the information they collected, the agencies are often forced to make available parts of these cross-classifications (e.g., lower dimensional marginals) to various users. A *marginal* is obtained from the original cross-classification by aggregating over one or more variables.

Any frequency count table $\mathbf{x} = \{x(i)\}_{i\in\mathcal{I}}$ with a fixed set of marginal totals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$, is a lattice point in the convex polytope defined by the linear system of equations:

$$\{\mathbf{x}_{D_j} = \mathbf{n}_{D_j} : j = 1, 2, \ldots, r\} \cap \{x(i) \in \{0, 1, 2, \ldots\} : i \in \mathcal{I}\}. \tag{1}$$

We denote by $\mathbf{T}^{(\mathbf{n})} = \mathbf{T}^{(\mathbf{n})}(D_1, \ldots, D_r)$ the set of admissible solutions of Eq. 1. Thus $\mathbf{T}^{(\mathbf{n})}$ is the set of all integer tables having the same margins as the observed data and this is the support of the exact conditional distribution $P(\mathbf{x}|\mathbf{x} \in \mathbf{T}^{(\mathbf{n})})$ of a table of counts $\mathbf{x}$ with a fixed set of marginals induced by the index sets $D_1, D_2, \ldots, D_r$. Since the reference set $\mathbf{T}^{(\mathbf{n})}$ is finite and has at least one element, namely the original table $\mathbf{n}$, the conditional probabilities will be meaningful.

Exact conditional distributions play an important role in the disclosure limitation context (Fienberg et al., 1998; Fienberg et al., 2001) and, in particular, in developing perturbation methods for categorical data (Cox, 1999; Duncan and Fienberg, 1999). One also wants to sample from the exact distribution for other purposes, such as the calibration of test statistics. Standard asymptotic results tell us how to approximate the value of Pearson's chi-squared statistic with a chi-squared distribution with the appropriate number of degrees of freedom. Nevertheless, the asymptotic approximation might not work well due to several reasons relating to the sample size and to the sparseness of the table. In some situations, calculating the correct number of degrees of freedom is itself a very difficult task (Haslett, 1990; Stirling, 1986; Mukerjee, 1987; Baker et al., 1985). An alternate approach to the asymptotic approximation was suggested, among others, by Diaconis and Efron (1985). They argue that calibration problem could be solved by sampling from the exact distribution on the space of all possible tables with fixed marginals and calculating the test statistic for the tables in the sample.

By considering conditional sampling schemes for cross-classifications of arbitrary dimension, Haberman (1974) proves that, in general, the exact conditional distribution cannot be expressed in a tractable form. If the marginals being fixed are the minimal sufficient statistics of a decomposable log-linear model $\mathcal{A}$, the exact conditional distribution reduces to the generalized hypergeometric distribution corresponding to $\mathcal{A}$ (Lauritzen, 1996; Fienberg et al., 2001). Many papers deal with the problem of generating from the exact conditional distribution of a two-way table given its row and column totals. However, the existent algorithms for sampling from the hypergeometric distribution associated with the model of unconditionally independence of two categorical variables do not readily generalize to more complicated decomposable log-linear models. Even for three-way tables, there do not exist standard sampling procedures that are guaranteed to be correct.

Diaconis and Sturmfels (1998) present a very general procedure for generating draws from the conditional distribution $P(\mathbf{x}|\mathbf{x} \in \mathbf{T}^{(\mathbf{n})})$. Their approach relies on the existence of a Markov basis, a collection of moves or data swaps that link all the tables in $\mathbf{T}^{(\mathbf{n})}$. Swapping data inside a table of counts involves moving individuals/units from one cell to another (Dalenius and Reiss, 1982). If exactly two individuals have been moved, the corresponding data swap is called *primitive*. The table $\mathbf{n}'$ created by repeatedly applying data swaps to the original table $\mathbf{n}$ has to be consistent with the set of released marginals, i.e., the data swaps should preserve $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$. Unfortunately, the sampling algorithm proposed by Diaconis and Sturmfels (1998) has not been largely used because a Markov basis could very difficult to generate even for problems of modest size.

Log-linear models are the most usual way of representing and studying contingency tables with fixed marginals, and Fienberg et al. (1998) and Fienberg (1999) have demonstrated the clear links between log-linear models and disclosure limitation techniques. Our attention, however, will be focused on *graphical* log-linear models. A graphical model is a statistical model corresponding to a number of simultaneous conditional independence relationships which can be summarized by means of an independence graph - see, for example Madigan and York (1995), Whittaker (1990) and Lauritzen (1996). When all the random variables embedded in the graphical model $\mathcal{M}$ are discrete and the independence graph $\mathcal{G}$ associated with $\mathcal{M}$ is undirected, $\mathcal{M}$ is said to be a graphical log-linear model. If $\mathcal{G}$ is decomposable (Lauritzen, 1996), then $\mathcal{M}$ is said to be decomposable. The class of decomposable log-linear models have closed form structure and special properties that will be exploited throughout this paper.

Our aim is to show how graphical models could help us identify special settings in which we could develop efficient techniques for considerably reducing and possibly eliminating the amount of computations needed to identify a Markov basis. After presenting some notation and definitions in Section 2, in Section 3 we formally introduce decomposable graphical models as well as results that characterize decomposable graphs. Section 4 proves a result postulated by Fienberg (1999) which says that primitive data swaps or moves are the only moves that have to be included in a Markov basis when the index sets defining the marginals that have to remain unchanged are the minimal sufficient statistics of a decomposable log-linear model. In Section 5 we give an example of using Markov bases to generate a "replacement" for a table having a set of fixed marginals. In the last section we make some concluding remarks.

2 Data Swapping and Markov Bases

Let $X = (X_1, X_2, \ldots, X_k)$ be a vector of discrete random variables. Denote $K = \{1,2,...,k\}$ the index set associated with $X_1, X_2, \ldots, X_k$. The random variable X_j can take the values $x_j \in \mathcal{I}_j := \{1,2,\ldots,I_j\}$, for $j = 1,2,\ldots,k$. Let $\mathcal{I} = \mathcal{I}_1 \times \mathcal{I}_2 \times \ldots \times \mathcal{I}_k$, and consider the k-way contingency table $\mathbf{n} := \{n(i)\}_{i\in\mathcal{I}}$. By assigning to every $i = (i_1, i_2, \ldots, i_k) \in \mathcal{I}$ an index (see Knuth (1973))

$$IND(i_1, i_2, \ldots, i_k) := \sum_{l=1}^{k} \left[\prod_{s=l+1}^{k} I_s \right] (i_l - 1) + 1 \in \{1, 2, \ldots, I_1 \cdot I_2 \cdot \ldots \cdot I_k\},$$

we could arrange the cells in the contingency table $\mathbf{n}$ in a linear list of objects, in which case we will write $\overline{\mathbf{n}}$ instead of $\mathbf{n}$. We let $D = \{i_1, i_2, \ldots, i_l\}$ denote an arbitrary subset of K, and we define X_D as the ordered tuple $X_D = (X_i; i \in D)$. The D-marginal $\mathbf{n}_D$ of $\mathbf{n}$ is the contingency table with *marginal cells* $i_D \in \mathcal{I}_D := \mathcal{I}_{i_1} \times \mathcal{I}_{i_2} \times \ldots \times \mathcal{I}_{i_l}$ and entries given by

$$n_D(i_D) = \sum_{j_{K\setminus D} \in \mathcal{I}_{K\setminus D}} n(i_D, j_{K\setminus D}).$$

Two k-way tables $\mathbf{n}_1 = \{n_1(i)\}_{i\in\mathcal{I}}$ and $\mathbf{n}_2 = \{n_2(i)\}_{i\in\mathcal{I}}$ are *equal* if $n_1(i) = n_2(i)$ for all $i \in \mathcal{I}$, and in this case we write $\mathbf{n}_1 = \mathbf{n}_2$. If all the counts in table $\mathbf{n}_1$ are zero, i.e. $n_1(i) = 0$, $\forall i \in \mathcal{I}$, we write $\mathbf{n}_1 = \mathbf{0}$. We define the notion of swapping entries in a table while preserving a given set of marginal totals.

Definition 1. *A move or data swap that does not modify the set of marginal tables* $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ *is a cross-classification* $\mathbf{f} = \{f(i)\}_{i\in\mathcal{I}}$ *with the following two properties:*

1. $f(i) \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, *for all* $i \in \mathcal{I}$.
2. $(\mathbf{x} \pm \mathbf{f})_{D_j} = \mathbf{x}_{D_j}$, *for all* $j = 1, 2, \ldots, r$ *and* $\mathbf{x} \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$.

Definition 1 says that $\mathbf{f}$ is a move for $\mathbf{T}^{(\mathbf{n})}(D_1, \ldots, D_r)$ if and only if $\mathbf{f}$ is an integer solution for the linear system of equations:

$$\left\{\mathbf{f}_{D_j} = \mathbf{0} : j = 1, 2, \ldots, r\right\}. \tag{2}$$

We observe that the above system has $L_1 := \sum_{s=1}^{r} \prod_{j\in D_s} I_j$ equations, and $L_2 := I_1 \cdot I_2 \cdot \ldots \cdot I_k$ unknowns. If $\mathbf{f} = \{f(i)\}_{i\in\mathcal{I}}$ is a move, then so is $-\mathbf{f} := \{-f(i)\}_{i\in\mathcal{I}}$. A move $\mathbf{f}$ is *primitive* if there exist four indices i_1, i_2, i_3, i_4 in $\mathcal{I}$ such that

$$f(i_1) = f(i_3) = 1 = -f(i_2) = -f(i_4), \tag{3}$$

and $f(i) = 0$, if $i \in \mathcal{I} \setminus \{i_1, i_2, i_3, i_4\}$.

Let $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_L\}$ be a set of moves. We define a graph denoted $\mathbf{T}_{\mathcal{F}}$ as follows. The nodes in this graph are the elements in $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$, and two nodes $\mathbf{x}$ and $\mathbf{x}'$ are connected by an edge if $\mathbf{x} - \mathbf{x}' \in \mathcal{F}$ or $\mathbf{x}' - \mathbf{x} \in \mathcal{F}$. If $\mathcal{F}$ is chosen to be large enough, the graph $\mathbf{T}_{\mathcal{F}}$ will be connected, in which case $\mathcal{F}$ forms a *Markov basis* (Diaconis and Sturmfels, 1998) for the tables in $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$. For simplicity, we will assume that, in any Markov basis of moves $\mathcal{F}$, $\mathbf{f} \in \mathcal{F}$ implies $-\mathbf{f} \in \mathcal{F}$. For any two tables $\mathbf{x}$, $\mathbf{x}'$ in $\mathbf{T}^{(\mathbf{n})}$, there exists a sequence of moves $\mathbf{f}^1, \mathbf{f}^2, \ldots, \mathbf{f}^s$ such that

$$\mathbf{x}' - \mathbf{x} = \sum_{j=1}^{s} f^j,$$

and

$$\mathbf{x} + \sum_{j=1}^{s'} \mathbf{f}^j \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r), \tag{4}$$

for $1 \leq s' \leq s$. Determining a set of data swaps needed to connect the initial table $\mathbf{n}$ to any other k-way table having marginals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ is equivalent to determining a Markov basis $\mathcal{F}$ for the class of tables $\mathbf{T}^{(\mathbf{n})}(D_1, \ldots, D_r)$. If $\mathbf{x} = \{x(i)\}_{i \in \mathcal{I}}$ is an arbitrary k-way cross-classification, $\mathcal{F}$ is also a Markov basis for the class of tables

$$\mathbf{T}^{(\mathbf{x})}(D_1, D_2, \ldots, D_r) := \{\mathbf{y} : y(i) \geq 0, \forall i \in \mathcal{I}, \mathbf{y}_{D_j} = \mathbf{x}_{D_j}, j = 1, 2, \ldots, r\},$$

(c.f., Diaconis and Sturmfels (1998); Conti and Traverso (1991)). Therefore a Markov basis of moves is determined only by the set of marginal constraints, not by the actual table we started with. Consequently, we will write $\mathbf{T}(D_1, \ldots, D_r)$ instead of $\mathbf{T}^{(\mathbf{n})}(D_1, \ldots, D_r)$ when the original table for which the marginals were calculated is not necessarily relevant.

Computational algebra offers excellent tools for characterizing the possible solutions of Eq. 2. Diaconis and Sturmfels (1998) and Dinwoodie (1998) show that computing a Markov basis for $\mathbf{T}(D_1, \ldots, D_r)$ is equivalent to finding a Gröbner basis of a special polynomial ideal. One can construct a Gröbner basis of a polynomial ideal by employing the Buchberger algorithm (Cox et al., 1992) or one of its more computationally efficient variants. Computer algebra systems such as MACAULAY, MAPLE, COCOA and MATHEMATICA implement this algorithm, hence the task of the users interested in finding a Markov basis for $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ is reduced to specifying the polynomial ideal associated with Eq. 2. Conti and Traverso (1991) showed how to find such a polynomial ideal by introducing a variable for each linear equation in Eq. 2, say $\sigma_1, \sigma_2, \ldots, \sigma_{L_1}$, and a variable for each unknown $\overline{f}_j$, say $\theta_1, \theta_2, \ldots, \theta_{L_2}$. The desired polynomial ideal is

$$\langle \theta_j - g_j : j = 1, 2, \ldots, L_2 \rangle,$$

where g_j is a monomial in $\sigma_1, \sigma_2, \ldots, \sigma_{L_1}$. Although simple and attractive, the algebraic approach is not feasible for finding Markov bases for contingency tables with more than three dimensions because of the huge amount of computing time it requires. The computational complexity of the Buchberger algorithm increases double exponentially with the number of variables as well as with the number of categories per variable.

It turns out, however, that it is straightforward to describe a Markov basis for a two-way table with fixed one-way marginals (Diaconis and Gangolli, 1995; Diaconis and Sturmfels, 1998).

Proposition 1. *Consider a two-way contingency table* $\mathbf{n} = \{n(i,j) : (i,j) \in \mathcal{I}_1 \times \mathcal{I}_2\}$ *with fixed row sums* $\mathbf{n}_1 := \{n_{i+} : i \in \mathcal{I}_1\}$ *and column sums* $\mathbf{n}_2 := \{n_{+j} : j \in \mathcal{I}_2\}$*. For some indices* i_1*,* i_2*,* j_1*,* j_2 *chosen such that* $1 \le i_1 < i_2 \le I_1$ *and* $1 \le j_1 < j_2 \le I_2$*, we define a table* $\mathbf{f}^{i_1 i_2; j_1 j_2} = \{f^{i_1 i_2; j_1 j_2}(i,j) : (i,j) \in \mathcal{I}_1 \times \mathcal{I}_2\}$ *by*

$$f^{i_1 i_2; j_1 j_2}(i,j) = \begin{cases} 1, & \textit{if } (i,j) \in \{(i_1,j_1),(i_2,j_2)\}, \\ -1, & \textit{if } (i,j) \in \{(i_1,j_2),(i_2,j_1)\}, \\ 0, & \textit{otherwise} . \end{cases} \tag{5}$$

Then

$$\{\pm \mathbf{f}^{i_1 i_2; j_1 j_2} : 1 \le i_1 < i_2 \le I_1, 1 \le j_1 < j_2 \le I_2\} \tag{6}$$

is a Markov basis for the class of tables with fixed row sums $\mathbf{n}_1$ *and column sums* $\mathbf{n}_2$*.*

Proof. The one-way marginals of $\mathbf{f}^{i_1 i_2; j_1 j_2}$ are zero, hence $\mathbf{f}^{i_1 i_2; j_1 j_2}$ will leave unchanged $\mathbf{n}_1$ and $\mathbf{n}_2$. By making use of computational algebra techniques, Sturmfels (1995) gives a complete proof of the fact that the set of moves described in Eq. 6 is indeed a Markov basis. For further reference, the number of moves in this Markov basis is

$$\left[2 \cdot \binom{I_1}{2} \cdot \binom{I_2}{2}\right].$$

■

The set of primitive moves we described above allows one to transform a given two-way table in any other two-way table with the same row and column totals. Proposition 1 is the starting point for developing Markov bases for an arbitrary decomposable graphical structure.

3 Decomposable Graphical Models

Let X_1, X_2, ..., X_k be the discrete random variables cross-classified in a k-way table $\mathbf{n} = \{n(i)\}_{i \in \mathcal{I}}$. We visualize the dependency patterns induced by the released marginals by constructing an independence graph $\mathcal{G}$ for the variables in the underlying cross-classification. The vertex set of $\mathcal{G}$ defined by the fixed marginals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ is $\bigcup_{j=1}^{r} D_j = K = \{1, 2, \ldots, k\}$, while its edge set is

$$E := \{(u,v) : \{u,v\} \subset D_j, \text{ for some } j \in \{1, \ldots, r\}\}.$$

Each variable X_j cross-classified in the table is associated with a vertex $j \in K$. The conditional independence relationships induced among X_1, X_2, ..., X_k by the fixed set of marginal totals are embedded in the graph $\mathcal{G}$ in the following way: if two variables are not connected, they are conditionally independent given the remainder. Lauritzen (1996) shows that this property is equivalent to:

If S *is a separator for* A_1 *and* A_2*, then* X_{A_1} *and* X_{A_2} *are conditionally independent given* X_S*.*

A log-linear model with minimal sufficient statistics $D_1, D_2, \ldots, D_r$ is *graphical* if $D_1, D_2, \ldots, D_r$ are the set of cliques of the independence graph $\mathcal{G}$, otherwise the log-linear model is not graphical. Moreover, a graphical log-linear model is *decomposable* if the independence graph induced by its minimal sufficient statistics is decomposable.

Assume that $\mathcal{G}$ is decomposable and let $\mathcal{C}(\mathcal{G}) := \{D_1, D_2, \ldots, D_r\}$ be the set of cliques of $\mathcal{G}$. Since $\mathcal{G}$ is decomposable, it is possible to order the vertex sets in $\mathcal{C}(\mathcal{G})$ in a perfect sequence (Blair and Barry, 1993). If we denote $H_j := D_1 \cup D_2 \cup \ldots \cup D_j$ and $S_j := H_{j-1} \cap D_j$, it follows that, for every $j = 2, \ldots, r$, $(H_{j-1} \setminus S_j, S_j, D_j \setminus S_j)$ is a decomposition of $\mathcal{G}(H_j)$ (Lauritzen, 1996). We let $\mathcal{S}(\mathcal{G}) := \{S_2, \ldots, S_r\}$ be the set of separators of the graph $\mathcal{G}$ associated with $\mathcal{C}(\mathcal{G})$.

By employing an expanded version of the maximum cardinality search algorithm (Blair and Barry, 1993), one can order the set $\mathcal{C}(\mathcal{G})$ of cliques so that they form a perfect sequence by constructing a tree $\mathcal{T} = (\mathcal{C}(\mathcal{G}), \mathcal{E}_{\mathcal{T}})$. The edges in $\mathcal{E}_{\mathcal{T}}$ are oriented, namely, $(D, D') \in \mathcal{E}_{\mathcal{T}}$ implies $(D', D) \notin \mathcal{E}_{\mathcal{T}}$. If $(D, D') \in \mathcal{E}_{\mathcal{T}}$, we will say that D' is the *parent* of D and D is the *child* of D'. A clique D is *terminal* in $\mathcal{T}$ if D is not the parent of any other clique. Moreover, D is the *root* of the tree $\mathcal{T}$ if D is not the child of any other clique. The tree $\mathcal{T}$ has the property that $S \subset V$ is a minimal separator of $\mathcal{G}$ if and only if $S = D_j \cap D_i$ for some edge $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}}$. The set of separators $\mathcal{S}(\mathcal{G})$ associated with $\mathcal{C}(\mathcal{G})$ will be given by $\mathcal{S}(\mathcal{G}) = \{D_j \cap D_i : (D_j, D_i) \in \mathcal{E}_{\mathcal{T}}\}$.

Definition 2 (The Star Property). *Take $D_j \in \mathcal{C}(\mathcal{G})$ and let $S = D_j \cap D_i$ for some $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}}$. Let $\mathcal{T}_j = (\mathcal{K}_j, \mathcal{E}_j)$ and $\mathcal{T}_i = (\mathcal{K}_i, \mathcal{E}_i)$ be the two subtrees obtained by removing the edge (D_j, D_i) from $\mathcal{T}$, with $D_j \in \mathcal{K}_j$ and $D_i \in \mathcal{K}_i$. Consider the vertex sets*

$$V_j := \bigcup_{D \in \mathcal{K}_j} D \quad \textit{and} \quad V_i := \bigcup_{D \in \mathcal{K}_i} D. \tag{7}$$

The tree $\mathcal{T}$ is said to have the Star Property for $\mathcal{G}$ if, for every edge $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}}$, $(V_j \setminus S, S, V_i \setminus S)$ is a decomposition of $\mathcal{G}$.

Blair and Barry (1993) prove that any tree $\mathcal{T}$ generated by the MCS algorithm has the Star Property. By removing a terminal clique from such a tree, the Star Property is preserved.

Lemma 1. *Let $\mathcal{T} = (\mathcal{C}(\mathcal{G}), \mathcal{E}_{\mathcal{T}})$ be a tree defined on the set of cliques of a decomposable graph $\mathcal{G}$. Assume that $\mathcal{T}$ has the Star Property for $\mathcal{G}$. Let D be a terminal clique in $\mathcal{T}$ and let D' be the the unique clique in $\mathcal{C}(\mathcal{G})$ such that $(D, D') \in \mathcal{E}_{\mathcal{T}}$. We consider $\mathcal{T}' = (\mathcal{C}(\mathcal{G}) \setminus \{D\}, \mathcal{E}_{\mathcal{T}} \setminus \{(D, D')\})$ to be the tree obtained by removing D from $\mathcal{T}$. Then $\mathcal{T}'$ is a tree with the Star Property for the decomposable graph $\mathcal{G}'$ defined by the set of cliques $\mathcal{C}(\mathcal{G}) \setminus \{D\}$.*

Proof. Consider an arbitrary edge $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}} \setminus \{(D, D')\}$. As before, we let $\mathcal{T}_j = (\mathcal{K}_j, \mathcal{E}_j)$ and $\mathcal{T}_i = (\mathcal{K}_i, \mathcal{E}_i)$ be the two subtrees obtained by removing the edge (D_j, D_i) from $\mathcal{T}$, with $D_j \in \mathcal{T}_j$ and $D_i \in \mathcal{T}_i$. Let V_j and V_i the vertex sets defined in Equation 7.

We can assume that $D \in \mathcal{K}_j$. If we were to remove the edge (D_j, D_i) from $\mathcal{T}'$, we would obtain the subtrees $\mathcal{T}_j' = (\mathcal{K}_j \setminus \{D\}, \mathcal{E}_j \setminus \{(D, D')\})$ and $\mathcal{T}_i = (\mathcal{K}_i, \mathcal{E}_i)$. The vertex set associated with $\mathcal{T}_j'$ is

$$V_j' := \bigcup_{D'' \in \mathcal{K}_j \setminus \{D\}} D''. \tag{8}$$

Since D is terminal in $\mathcal{E}_{\mathcal{T}}$, we have $D_j \neq D$, hence $V_j' \neq \emptyset$. The vertex set $S := D_j \cap D_i$ which is a separator for $\mathcal{G}$, is also a separator for $\mathcal{G}'$. Moreover, $(V_j \setminus S, S, V_i \setminus S)$ is a decomposition of $\mathcal{G}$. From

$V_j' \subset V_j$, it follows that $(V_j' \setminus S, S, V_i \setminus S)$ will be a decomposition of $\mathcal{G}'$. Therefore the tree $\mathcal{T}'$ has the Star Property for $\mathcal{G}'$. ■

The maximum cardinality search algorithm (Blair and Barry, 1993), can be employed only if the decomposable graph $\mathcal{G}$ is connected. Assume $\mathcal{G}$ is disconnected and let $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_L$ be its connected components. We apply the MCS algorithm for every connected component and obtain the collection of trees $\mathcal{T}_l = (\mathcal{C}(\mathcal{G}_l), \mathcal{E}_l)$, for $l = 1, 2, \ldots, L$. The set of cliques of $\mathcal{G}$ is the union of the sets of cliques associated with the connected components $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_L$, i.e.,

$$\mathcal{C}(\mathcal{G}) = \mathcal{C}(\mathcal{G}_1) \cup \mathcal{C}(\mathcal{G}_2) \cup \ldots \cup \mathcal{C}(\mathcal{G}_L).$$

We define the tree $\mathcal{T} = (\mathcal{C}(\mathcal{G}), \mathcal{E})$ by

$$\mathcal{E} := \mathcal{E}_1 \cup \ldots \cup \mathcal{E}_L \cup \left\{(D_1^{l-1}, D_2^l) : l = 2, \ldots, L\right\}, \tag{9}$$

where D_1^{l-1} is a root of $\mathcal{T}_{l-1}$ and D_2^l is terminal in $\mathcal{T}_l$. The tree generated on the set of cliques of a connected decomposable graph by the MCS algorithm will always have the Star Property. In addition, it is not hard to see that the tree $\mathcal{T}$ specified by Eq. 9 will have the Star Property on the set of cliques of an arbitrary decomposable graph $\mathcal{G}$.

3.1 Markov Bases for Decomposable Graphical Models

The theory on log-linear models for cross-classified counts could provide a real insight into the structure of the constraints induced by fixing a set of marginal totals of a k-dimensional contingency table $\mathbf{n}$. Assume the margins $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ being fixed are the minimal sufficient of statistics of a decomposable log-linear model. In this situation, the estimated expected values for the table entries can be written in closed form as a function of the marginal totals. Moreover, Dobra and Fienberg (2000) derived explicit formulas for the tightest upper and lower bounds for the cell counts in the table $\mathbf{n}$ given that $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ are known. The special structural properties of decomposable graphs can be further exploited to derive a Markov basis of primitive moves for the class of tables $\mathbf{T}^{(\mathbf{n})}(D_1, \ldots, D_r)$.

We start with the simplest case, $r = 2$, and show that the Markov basis for $\mathbf{T}^{(\mathbf{n})}(D_1, D_2)$ is the union of the Markov bases of one or more two-way tables with fixed one-way marginals.

Proposition 2. *Primitive moves are the only moves we have to include in the Markov basis of* $\mathbf{T}(D_1, D_2)$.

Proof. The independence graph $\mathcal{G} = (K, E)$ associated with marginals $\mathbf{n}_{D_1}$ and $\mathbf{n}_{D_2}$ of $\mathbf{n}$ has vertex set $K = D_1 \cup D_2$ and edge set

$$E := \{(j^1, j^2) : \{j^1, j^2\} \subset D_1 \text{ or } \{j^1, j^2\} \subset D_2\}. \tag{10}$$

Without restricting the generality, we can assume that $D_1 = \{1, \ldots, l\}$ and $D_2 = \{q, \ldots, k\}$. We distinguish two cases:

(i) If $l < q$, X_{D_1} and X_{D_2} are unconditionally independent. Introduce two new compound variables Y_1 and Y_2 with level sets $\mathcal{I}_{D_1} := \times_{\delta \in D_1} \mathcal{I}_\delta$ and $\mathcal{I}_{D_2} := \times_{\delta \in D_2} \mathcal{I}_\delta$, respectively. Take the two-way table $\mathbf{n}'$ corresponding to Y_1, Y_2 with known row sums $\overline{\mathbf{n}}_{D_1} = \{n_{D_1}(i_{D_1})\}_{i_{D_1} \in \mathcal{I}_{D_1}}$ and column sums $\overline{\mathbf{n}}_{D_2} = \{n_{D_2}(i_{D_2})\}_{i_{D_2} \in \mathcal{I}_{D_2}}$. The basis $\mathcal{F}(D_1, D_2)$ for the original table $\mathbf{n}$ will be given by the Markov basis of moves for the two-way table $\mathbf{n}'$ as described in Proposition 1.

(ii) Suppose $l \geq q$. The variables indexed $\{1,\dots,q-1\}$ and $\{l+1,\dots,k\}$ are conditionally independent given the variables indexed $\{q,\dots,l\}$. Denote

$$\mathcal{J} := \times_{\delta\in\{1,\dots,q-1,l+1,\dots,k\}} \mathcal{I}_\delta.$$

Take the set of contingency tables

$$\left\{\mathbf{n}^{i_q^0,\dots,i_l^0} = \left\{n^{i_q^0,\dots,i_l^0}(i)\right\}_{i\in\mathcal{J}} : i_q^0 \in \mathcal{I}_q,\dots,i_l^0 \in \mathcal{I}_l\right\},$$

where

$$n^{i_q^0,\dots,i_l^0}(i) = n^{i_q^0,\dots,i_l^0}(i_1,\dots,i_{q-1},i_{l+1},\dots,i_k) = n(i_1,\dots,i_{q-1},i_q^0,\dots,i_l^0,i_{l+1},\dots,i_k).$$

For a given vector of indices $(i_q^0,\dots,i_l^0) \in \mathcal{I}_q\times\dots\times\mathcal{I}_l$, the variables indexed $\{1,\dots,q-1\}$ and $\{l+1,\dots,k\}$ are independent, hence we can pursue a line of reasoning analogous to (i). First we introduce two compound variables $Y_1^{i_q^0,\dots,i_l^0}$ and $Y_2^{i_q^0,\dots,i_l^0}$ with level sets $\mathcal{I}_1\times\dots\times\mathcal{I}_{q-1}$ and $\mathcal{I}_{l+1}\times\dots\times\mathcal{I}_k$, respectively. The Markov basis $\mathcal{F}^{i_q^0,\dots,i_l^0}$ for the table $\mathbf{n}^{i_q^0,\dots,i_l^0}$ is equivalent to the Markov basis for the two-way table corresponding to $Y_1^{i_q^0,\dots,i_l^0}$ and $Y_2^{i_q^0,\dots,i_l^0}$ with known row sums $\overline{\mathbf{n}}_{1,\dots,q-1}^{i_q^0,\dots,i_l^0}$ and column sums $\overline{\mathbf{n}}_{l+1,\dots,k}^{i_q^0,\dots,i_l^0}$. Then a Markov basis of moves for the table $\mathbf{n}$ that preserves the marginals $\mathbf{n}_{D_1}$ and $\mathbf{n}_{D_2}$ is given by

$$\mathcal{F}(D_1,D_2) = \bigcup\left\{\mathcal{F}^{i_q^0,\dots,i_l^0} : i_q^0 \in \mathcal{I}_q,\dots,i_l^0 \in \mathcal{I}_l\right\}. \tag{11}$$

■

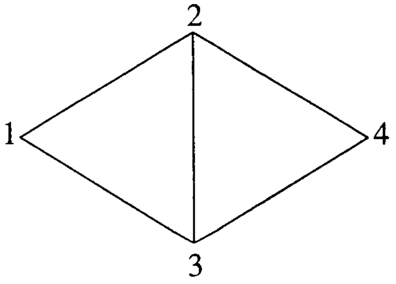

Figure 1: A decomposable graph with two cliques.

Example 1. Consider a four-way table $\mathbf{n}$ with fixed three-way marginals $\mathbf{x} := \mathbf{n}_{\{1,2,3\}}$ and $\mathbf{y} := \mathbf{n}_{\{2,3,4\}}$. The corresponding independence graph $\mathcal{G}$ is represented in Fig. 1. The edge $\{2,3\}$ is a separator for $\{1,2,3\}$ and $\{2,3,4\}$. In addition, $\{1,2,3\}$ and $\{2,3,4\}$ are complete in $\mathcal{G}$, hence $\mathcal{G}$ is a decomposable graph with two cliques. Since X_1 is conditionally independent of X_4 given (X_2,X_3), we consider the set of contingency tables

$$\left\{\mathbf{n}^{i_2^0,i_3^0} = \left\{n^{i_2^0,i_3^0}(i_1,i_4) : (i_1,i_4)\in\mathcal{I}_1\times\mathcal{I}_4\right\} : i_2^0\in\mathcal{I}_2, i_3^0\in\mathcal{I}_3\right\},$$

where $n^{i_2^0,i_3^0}(i_1,i_4) = n(i_1,i_2^0,i_3^0,i_4)$. For every table $\mathbf{n}^{i_2^0,i_3^0}$, we know the row sums $\mathbf{n}_1^{i_2^0,i_3^0} := \{x(i_1,i_2^0,i_3^0) : i_1 \in \mathcal{I}_1\}$ as well as the column sums $\mathbf{n}_2^{i_2^0,i_3^0} := \{y(i_2^0,i_3^0,i_4) : i_4 \in \mathcal{I}_4\}$. The Markov basis of moves $\mathcal{F}^{i_2^0,i_3^0}$ that will leave unchanged the one-way marginals of the table $\mathbf{n}^{i_2^0,i_3^0}$ can be obtained as in Proposition 1. Proposition 2 tells us that the Markov basis of primitive moves $\mathcal{F}(\{1,2,3\},\{2,3,4\})$ that preserves the marginals $\mathbf{x}$ and $\mathbf{y}$ is just the union $\left\{\mathcal{F}^{i_2^0,i_3^0} : (i_2^0,i_3^0) \in \mathcal{I}_2 \times \mathcal{I}_3\right\}$. ■

We introduce the set of primitive moves associated with an arbitrary decomposable graph $\mathcal{G}$.

Definition 3. *Let $\mathcal{C}(\mathcal{G}) = \{D_1, D_2, \ldots, D_r\}$ the set of cliques of a decomposable graph $\mathcal{G}$. We let $\mathcal{T} = (\mathcal{C}(\mathcal{G}), \mathcal{E}_{\mathcal{T}})$ be a tree having the Star Property on the set of cliques of $\mathcal{G}$. For every edge $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}}$, we consider the vertex sets V_j and V_i as in Eq. 7. The set of primitive moves associated with the decomposable graph $\mathcal{G}$ is given by:*

$$\mathcal{F}(\mathcal{G}) = \mathcal{F}(D_1, D_2, \ldots, D_r) := \bigcup_{(D_j,D_i)\in\mathcal{E}_{\mathcal{T}}} \mathcal{F}(V_j, V_i), \tag{12}$$

where $\mathcal{F}(V_j, V_i)$ was defined in Eq. 11.

By removing an edge (D_j, D_i) from $\mathcal{T}$, we create two connected components $\mathcal{T}(V_j)$ and $\mathcal{T}(V_i)$. We think about V_j and V_i as being the cliques of some graph $\mathcal{G}^{ij}$ with vertices $V_j \cup V_i = K$ and edges

$$E_{ij} := \{(u,v) : \{u,v\} \subset V_j \text{ or } \{u,v\} \subset V_i\}.$$

The tree $\mathcal{T}$ has the Star Property, hence $S_{ij} := D_j \cap D_i$ separates $V_j \setminus S_{ij}$ from $V_i \setminus S_{ij}$ in $\mathcal{G}^{ij}$. As a result, $\mathcal{G}^{ij}$ is the independence graph of a decomposable graphical model with two cliques and, from Proposition 2, we know that the set of primitive moves corresponding to $\mathcal{G}^{ij}$ is $\mathcal{F}(V_j, V_i)$. Eq. 12 essentially says that the set of primitive moves for a decomposable graphical model with independence graph $\mathcal{G}$ is just the union of the sets of primitive moves associated with the two-clique models induced by each minimal vertex separator of $\mathcal{G}$. We have to show that Definition 3 is correct.

Proposition 3. *The set of primitive moves defined in Eq. 12 is indeed a set of moves for the class of tables $\mathbf{T}(D_1, D_2, \ldots, D_r)$.*

Proof. Let $\mathbf{f} \in \mathcal{F}(D_1, D_2, \ldots, D_r)$. Then $\mathbf{f} \in \mathcal{F}(V_j, V_i)$ for some $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}}$. For any arbitrary clique $D \in \mathcal{C}(\mathcal{G})$, we have either $D \subset V_j$ or $D \subset V_i$. Since $\mathbf{f}_{V_j} = \mathbf{0}$ and $\mathbf{f}_{V_i} = \mathbf{0}$, it follows that we also have $\mathbf{f}_D = \mathbf{0}$. ■

Next we will state and prove a series of results that will help us prove the main theorem of the paper. Most of these propositions should be self-explanatory. However, it is worth mentioning the intuition that triggered them: if we delete a vertex that belongs to exactly one clique from a decomposable graph, along with the edges incident to it, we obtain a graph that is still decomposable (Blair and Barry, 1993). Consequently, by collapsing across a variable associated with such a vertex, all the conditional dependencies existent among the remaining variables are preserved.

The set of primitive moves associated with a two-clique model induces a set of primitive moves for a two-clique model embedded in it. Collapsing across some of the variables not contained in both cliques preserves the structure of the moves in Eq. 12.

Proposition 4. *Let* $\mathbf{n}$ *be a table with two fixed marginals* $\mathbf{n}_{D_1}$ *and* $\mathbf{n}_{D_2}$. *The corresponding independence graph* $\mathcal{G}$ *is decomposable and has two cliques* D_1, D_2. *The separator of* $\mathcal{G}$ *is* $S := D_1 \cap D_2$. *Consider a vertex set* D *such that* $S \subset D \subset D_1$. *Define a map* ϕ *which assigns to every* $\mathbf{f} \in \mathcal{F}(D_1, D_2)$ *its* $(D \cup D_2)$*-marginal, i.e.*

$$\phi(\mathbf{f}) = \mathbf{f}_{D \cup D_2}.$$

Then the following are true:

(a) for any $\mathbf{f} \in \mathcal{F}(D_1, D_2)$, $\phi(\mathbf{f}) \in \mathcal{F}(D, D_2)$ *or* $\phi(\mathbf{f}) = \mathbf{0}$.

(b) the map $\phi : \mathcal{F}(D_1, D_2) \to \mathcal{F}(D, D_2)$ *is surjective.*

(c) for every table $\mathbf{x} \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2)$ *and every move* $\mathbf{g} \in \mathcal{F}(D, D_2)$ *such that*

$$\mathbf{x}_{D \cup D_2} + \mathbf{g} \in \mathbf{T}^{(\mathbf{n})}(D, D_2), \tag{13}$$

there exists $\mathbf{f} \in \mathcal{F}(D_1, D_2)$ *with* $\phi(\mathbf{f}) = \mathbf{g}$ *and*

$$\mathbf{x} + \mathbf{f} \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2). \tag{14}$$

Proof. To simplify the notation, assume that $S = \emptyset$. We consider the marginals $\mathbf{n}_{D_1}$, $\mathbf{n}_{D_2}$ and $\mathbf{n}_D$, along with their associated vectors $\overline{\mathbf{n}}_{D_1}$, $\overline{\mathbf{n}}_{D_2}$ and $\overline{\mathbf{n}}_D$. The table $\mathbf{n}_D$ can be obtained from $\mathbf{n}_{D_1}$ by collapsing across the variables in $D_1 \setminus D$.

(a) In Proposition 1, we constructed $\mathcal{F}(D_1, D_2)$ by considering the two-way table with row marginal $\overline{\mathbf{n}}_{D_1}$ and column marginal $\overline{\mathbf{n}}_{D_2}$. A primitive move $\mathbf{f} \in \mathcal{F}(D_1, D_2)$ was obtained by choosing two "row" indices $i^1_{D_1}$ and $i^2_{D_1}$, and two "column" indices $i^1_{D_2}$ and $i^2_{D_2}$. Then the table $\mathbf{f}$ is given by:

$$f(i_{D_1}, i_{D_2}) = \begin{cases} \pm 1, & \text{if } (i_{D_1}, i_{D_2}) \in \left\{(i^1_{D_1}, i^1_{D_2}), (i^2_{D_1}, i^2_{D_2})\right\}, \\ \mp 1, & \text{if } (i_{D_1}, i_{D_2}) \in \left\{(i^1_{D_1}, i^2_{D_2}), (i^2_{D_1}, i^1_{D_2})\right\}, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Let $\mathbf{f}_1 = \phi(\mathbf{f})$. We have:

$$f_1(i_{D \cup D_2}) = f_1(i_D, i_{D_2}) = \sum_{j_{D_1 \setminus D} \in \mathcal{I}_{D_1 \setminus D}} f(j_{D_1 \setminus D}, i_D, i_{D_2}).$$

We distinguish two cases.

(i) $i^1_D = i^2_D$. Since $i^1_{D_1} \neq i^2_{D_1}$, we need to have $i^1_{D_1 \setminus D} \neq i^2_{D_1 \setminus D}$. It follows that

$$f_1(i^1_D, i^r_{D_2}) = f(i^1_{D_1}, i^r_{D_2}) + f(i^2_{D_1}, i^r_{D_2}) = 0, \text{ for } r = 1, 2.$$

Clearly, $f_1(i^1_D, i_{D_2}) = 0$ if $i_{D_2} \notin \{i^1_{D_2}, i^2_{D_2}\}$. Moreover, for $i_D \neq i^1_D$, $f_1(i_D, i_{D_2}) = 0$. Hence $\phi(\mathbf{f}) = \mathbf{f}_1 = \mathbf{0}$.

(ii) $i^1_D \neq i^2_D$. In this case it is not hard to see that

$$f_1(i_D, i_{D_2}) = \begin{cases} f(i^{r_1}_{D_1}, i^{r_2}_{D_2}), & \text{if } (i_D, i_{D_2}) = (i^{r_1}_D, i^{r_2}_{D_2}), \text{ where } r_1, r_2 \in \{1, 2\}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus $\phi(\mathbf{f}) = \mathbf{f}_1 \in \mathcal{F}(D, D_2)$.

(b) In order to prove that ϕ is surjective, we pick an arbitrary move $\mathbf{g} \in \mathcal{F}(D, D_2)$. We choose an index $i^0_{D_1\setminus D} \in \mathcal{I}_{D_1\setminus D}$ and define the move $\mathbf{f} = \{f(i)\}_{i\in\mathcal{I}_{D_1\cup D_2}}$ by

$$f(i) = f(i_{D_1\setminus D}, i_{D\cup D_2}) := \begin{cases} g(i_{D\cup D_2}), & \text{if } i_{D_1\setminus D} = i^0_{D_1\setminus D}, \\ 0, & \text{otherwise .} \end{cases}$$

It is easy to see that $\mathbf{f} \in \mathcal{F}(D_1, D_2)$ and $\phi(\mathbf{f}) = \mathbf{g}$.

(c) The move $\mathbf{g} \in \mathcal{F}(D, D_2)$ is given by

$$g(i_D, i_{D_2}) = \begin{cases} 1, & \text{if } (i_D, i_{D_2}) \in \left\{(i^1_D, i^1_{D_2}), (i^2_D, i^2_{D_2})\right\}, \\ -1, & \text{if } (i_D, i_{D_2}) \in \left\{(i^1_D, i^2_{D_2}), (i^2_D, i^1_{D_2})\right\}, \\ 0, & \text{otherwise,} \end{cases} \tag{16}$$

where $i^1_D, i^2_D \in \mathcal{I}_D$ and $i^1_{D_2}, i^2_{D_2} \in \mathcal{I}_{D_2}$. A move $\mathbf{f} \in \mathcal{F}(D_1, D_2)$ such that $\mathbf{f}_{D\cup D_2} = \mathbf{g}$ is obtained by choosing two indices $i^1_{D_1\setminus D}$ and $i^2_{D_1\setminus D}$ in $\mathcal{I}_{D_1\setminus D}$. Then

$$f(i) = f(i_{D_1}, i_{D_2}) = \begin{cases} 1, & \text{if } i \in \left\{(i^1_{D_1\setminus D}, i^1_D, i^1_{D_2}), (i^2_{D_1\setminus D}, i^2_D, i^2_{D_2})\right\}, \\ -1, & \text{if } i \in \left\{(i^1_{D_1\setminus D}, i^1_D, i^2_{D_2}), (i^2_{D_1\setminus D}, i^2_D, i^1_{D_2})\right\}, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

For any $i^1_{D_1\setminus D}, i^2_{D_1\setminus D}$ in $\mathcal{I}_{D_1\setminus D}$, the corresponding move $\mathbf{f}$ defined in Eq. 17 is such that

$$\begin{aligned} (\mathbf{x}+\mathbf{f})_{D_1} &= \mathbf{x}_{D_1} = \mathbf{n}_{D_1}, \\ (\mathbf{x}+\mathbf{f})_{D_2} &= \mathbf{x}_{D_2} = \mathbf{n}_{D_2}, \end{aligned} \tag{18}$$

and

$$(x+f)(i) \geq 0, \tag{19}$$

for every $i \in \mathcal{I} \setminus \left\{(i^1_{D_1\setminus D}, i^1_D, i^2_{D_2}), (i^2_{D_1\setminus D}, i^2_D, i^1_{D_2})\right\}$. Therefore we have to choose $i^1_{D_1\setminus D}, i^2_{D_1\setminus D}$ such that

$$\begin{aligned} (x+f)(i^1_{D_1\setminus D}, i^1_D, i^2_{D_2}) &= x(i^1_{D_1\setminus D}, i^1_D, i^2_{D_2}) - 1 \geq 0, \text{ and} \\ (x+f)(i^2_{D_1\setminus D}, i^2_D, i^1_{D_2}) &= x(i^2_{D_1\setminus D}, i^2_D, i^1_{D_2}) - 1 \geq 0. \end{aligned} \tag{20}$$

In this case, Eq. 14 holds. From Eq. 13, we obtain that

$$\begin{aligned} (x_{D\cup D_2} + g)(i^1_D, i^2_{D_2}) &\geq 0, \text{ and} \\ (x_{D\cup D_2} + g)(i^2_D, i^1_{D_2}) &\geq 0. \end{aligned} \tag{21}$$

But

$$\begin{aligned} (x_{D\cup D_2} + g)(i^1_D, i^2_{D_2}) &= x_{D\cup D_2}(i^1_D, i^2_{D_2}) - 1, \\ &= \sum_{j_{D_1\setminus D}\in\mathcal{I}_{D_1\setminus D}} x(j_{D_1\setminus D}, i^1_D, i^2_{D_2}) - 1. \end{aligned} \tag{22}$$

Eq. 21 and Eq. 22 imply that

$$\sum_{j_{D_1\setminus D}\in\mathcal{I}_{D_1\setminus D}} x(j_{D_1\setminus D}, i^1_D, i^2_{D_2}) \geq 1, \tag{23}$$

hence there has to exist an index $i^1_{D_1\setminus D} \in \mathcal{I}_{D_1\setminus D}$ with $x(i^1_{D_1\setminus D}, i^1_D, i^2_{D_2}) \geq 1$. Similarly, there has to exist another index $i^2_{D_1\setminus D} \in \mathcal{I}_{D_1\setminus D}$ with $x(i^2_{D_1\setminus D}, i^2_D, i^1_{D_2}) \geq 1$. With this choice, Eq. 20 holds. ■

Proposition 5 extends Proposition 4 to an arbitrary decomposable model.

Proposition 5. *Let* $\mathbf{n}$ *be a table with fixed marginals* $\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r}$ *such that* $\mathcal{C}(\mathcal{G}) = \{D_1, \ldots, D_r\}$ *the set of cliques of a decomposable graph* $\mathcal{G}$. *Consider a tree* $\mathcal{T} = (\mathcal{C}(\mathcal{G}), \mathcal{E}_\mathcal{T})$ *having the Star Property for* $\mathcal{G}$. *Assume that the clique* D_r *is terminal in* $\mathcal{T}$ *and let* $A := \bigcup_{j=1}^{r-1} D_j$. *Define a map* ϕ *which assigns to every* $\mathbf{f} \in \mathcal{F}(D_1, D_2, \ldots, D_r)$ *its* A*-marginal, i.e.*

$$\phi(\mathbf{f}) = \mathbf{f}_A.$$

Then the following are true:

(a) for any $\mathbf{f} \in \mathcal{F}(D_1, D_2, \ldots, D_r)$, $\phi(\mathbf{f}) \in \mathcal{F}(D_1, \ldots, D_{r-1})$ *or* $\phi(\mathbf{f}) = \mathbf{0}$.

(b) the map ϕ *is surjective on* $\mathcal{F}(D_1, \ldots, D_{r-1})$.

(c) for every table $\mathbf{x} \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ *and every move* $\mathbf{g} \in \mathcal{F}(D_1, \ldots, D_{r-1})$ *such that*

$$\mathbf{x}_A + \mathbf{g} \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_{r-1}), \tag{24}$$

there exists $\mathbf{f} \in \mathcal{F}(D_1, D_2, \ldots, D_r)$ *with* $\phi(\mathbf{f}) = \mathbf{g}$ *and*

$$\mathbf{x} + \mathbf{f} \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r). \tag{25}$$

Proof. (a) Since the clique D_r is terminal in $\mathcal{T}$, there exists a unique clique in $\mathcal{C}(\mathcal{G})$, say D', such that $(D_r, D') \in \mathcal{E}_\mathcal{T}$. The set of primitive moves corresponding to the edge (D_r, D') is $\mathcal{F}(A, D_r)$, and take $\mathbf{f} \in \mathcal{F}(A, D_r)$. By definition, $\mathbf{f}_A = \mathbf{0}$, hence $\phi(\mathbf{f}) = \mathbf{0}$.

The subgraph $\mathcal{G}' = \mathcal{G}(D_1 \cup \ldots \cup D_{r-1})$ is decomposable and $\mathcal{C}(\mathcal{G}') = \{D_1, \ldots, D_{r-1}\}$. Let $\mathcal{T}'$ the subtree obtained by removing D_r from $\mathcal{T}$, i.e. $\mathcal{T}' = (\mathcal{C}(\mathcal{G}'), \mathcal{E}_\mathcal{T} \setminus \{(D_r, D')\})$. Consider an arbitrary edge $(D_j, D_i) \in \mathcal{E}_\mathcal{T} \setminus \{(D_r, D')\}$. Let $\mathcal{T}_j = (\mathcal{K}_j, \mathcal{E}_j)$ and $\mathcal{T}_i = (\mathcal{K}_i, \mathcal{E}_i)$ be the two subtrees obtained by removing the edge (D_j, D_i) from $\mathcal{T}$, with $D_j \in \mathcal{K}_j$ and $D_i \in \mathcal{K}_i$. Without restricting the generality, we assume that we always have $D_r \in \mathcal{K}_j$.

By removing the same edge from the tree $\mathcal{T}'$, we obtain the subtrees $\mathcal{T}'_j = (\mathcal{K}_j \setminus \{D_r\}, \mathcal{E}_j \setminus \{(D_r, D')\})$ and $\mathcal{T}_i$. We define the vertex sets V_j, V'_j and V_i by

$$V_j := \bigcup_{D\in\mathcal{K}_j} D, \quad V'_j := \bigcup_{D\in\mathcal{K}_j\setminus\{D_r\}} D, \quad V_i := \bigcup_{D\in\mathcal{K}_i} D. \tag{26}$$

With this notation, according to Lemma 1, the tree $\mathcal{T}'$ will have the Star Property for the graph $\mathcal{G}'$, and consequently the set of primitive primitive associated with $\mathcal{G}$ is

$$\mathcal{F}(\mathcal{G}') = \mathcal{F}(D_1, \ldots, D_{r-1}) = \bigcup_{(D_j, D_i)\in\mathcal{E}_\mathcal{T}\setminus\{(D_r, D')\}} \mathcal{F}(V'_j, V_i). \tag{27}$$

Consider an arbitrary move $\mathbf{f} \in \mathcal{F}(D_1, D_2, \ldots, D_r)$ such that $\mathbf{f} \notin \mathcal{F}(A, D_r)$. From Eq. 12, we see that there must exist some edge $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}} \setminus \{(D_r, D')\}$ such that $\mathbf{f} \in \mathcal{F}(V_j, V_i)$. We have $D_j \neq D_r$ and $D_j \subset V_j$, thus $V_j' \neq \emptyset$. In addition, we have $V_j' \subset V_j$ and $A = V_j' \cup V_i$. By employing Proposition 4, we obtain that $\phi(\mathbf{f}) \in \mathcal{F}(V_j', V_i) \subset \mathcal{F}(D_1, \ldots, D_{r-1})$ or $\phi(\mathbf{f}) = \mathbf{0}$.

(b) In order to prove that ϕ is surjective on $\mathcal{F}(D_1, \ldots, D_{r-1})$, we pick an arbitrary move $\mathbf{g}$ in $\mathcal{F}(D_1, \ldots, D_{r-1})$. From Eq. 27, we see that there is an edge $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}} \setminus \{(D_r, D')\}$ such that $\mathbf{g} \in \mathcal{F}(V_j', V_i)$. Since $V_j' \subset V_j$, Proposition 4 tells us that there must exist some $\mathbf{f} \in \mathcal{F}(V_j, V_i) \subset \mathcal{F}(D_1, \ldots, D_r)$ such that $\phi(\mathbf{f}) = \mathbf{g}$.

(c) Again, Eq. 27 tells us that we can find an edge $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}} \setminus \{(D_r, D')\}$ such that $\mathbf{g} \in \mathcal{F}(V_j', V_i)$. This means that

$$\mathbf{x}_A + \mathbf{g} \in \mathbf{T}^{(\mathbf{x})}(V_j', V_i). \tag{28}$$

We have $V_j' \subset V_i$ and $V_j' \cap V_i = V_j \cap V_i$. From Proposition 4, we learn that there exists a move

$$\mathbf{f} \in \mathcal{F}(V_j, V_i) \subset \mathcal{F}(D_1, D_2, \ldots, D_r), \tag{29}$$

such that

$$\mathbf{x} + \mathbf{f} \in \mathbf{T}^{(\mathbf{x})}(V_j, V_i). \tag{30}$$

But we also have

$$\mathbf{T}^{(\mathbf{x})}(V_j, V_i) \subset \mathbf{T}^{(\mathbf{x})}(D_1, D_2, \ldots, D_r) = \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r), \tag{31}$$

hence Eq. 25 is true. ■

We are now ready to present and prove the main theorem of the paper.

Theorem 1. *Let $\mathcal{G}$ be a decomposable graph with cliques $\mathcal{C}(\mathcal{G}) = \{D_1, D_2, \ldots, D_r\}$. Then the set of primitive moves $\mathcal{F}(\mathcal{G}) = \mathcal{F}(D_1, D_2, \ldots, D_r)$ defined in Eq. 12 is a Markov basis for the class of tables $\mathbf{T}(D_1, D_2, \ldots, D_r)$.*

Proof. By induction. If $\mathcal{G}$ decomposes in $r = 2$ cliques, then we know from Proposition 1 that $\mathcal{F}(D_1, D_2)$ is a Markov basis for $\mathbf{T}^{(\mathbf{n})}(D_1, D_2)$. Suppose the theorem holds for any decomposable graph with $r - 1$ cliques. We want to prove that the theorem is true for a decomposable graph with r cliques.

The original table $\mathbf{n}$ is in the set $\mathbf{T}^{(\mathbf{n})} = \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$. Take an arbitrary table $\mathbf{x} \in \mathbf{T}^{(\mathbf{n})}$. We have to show that there exist $\mathbf{f}^1, \ldots, \mathbf{f}^l \in \mathcal{F}(D_1, D_2, \ldots, D_r)$ such that

$$\begin{aligned} \mathbf{x} - \mathbf{n} &= \sum_{i=1}^{l} \mathbf{f}^i, \text{ and} \\ \mathbf{n} + \sum_{i=1}^{l'} \mathbf{f}^i &\in \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r), \end{aligned} \tag{32}$$

for $1 \leq l' \leq l$. Let $\mathcal{T} = (\mathcal{C}(\mathcal{G}), \mathcal{E}_{\mathcal{T}})$ a tree having the Star Property for $\mathcal{G}$ and assume that the clique D_r is terminal in $\mathcal{T}$. Denote $A := \bigcup_{j=1}^{r-1} D_j$. Consider the map ϕ which assigns to every $\mathbf{f} \in \mathcal{F}(D_1, D_2, \ldots, D_r)$ its A-marginal, i.e.

$$\phi(\mathbf{f}) = \mathbf{f}_A.$$

The marginals $\mathbf{n}_A$ and $\mathbf{x}_A$ lie in the set $\mathbf{T}^{(\mathbf{n})}(D_1,\ldots,D_{r-1})$. From the induction hypothesis we know that $\mathcal{F}(D_1,\ldots,D_{r-1})$ is a Markov basis for $\mathbf{T}^{(\mathbf{n})}(D_1,\ldots,D_{r-1})$, so there exists a sequence of moves $\mathbf{g}^1,\ldots,\mathbf{g}^{l_1} \in \mathcal{F}(D_2,\ldots,D_r)$ such that

$$\begin{aligned}\mathbf{x}_A - \mathbf{n}_A &= \sum_{i=1}^{l_1} \mathbf{g}^i, \text{ and}\\ \mathbf{n}_A + \sum_{i=1}^{l'_1} \mathbf{g}^i &\in \mathbf{T}^{(\mathbf{n})}(D_1,D_2,\ldots,D_{r-1}),\end{aligned} \tag{33}$$

for $1 \leq l'_1 \leq l_1$. Proposition 5 tells us that the sequence of moves $\mathbf{g}^1,\ldots,\mathbf{g}^{l_1}$ translates into another sequence of moves $\mathbf{f}^1,\ldots,\mathbf{f}^{l_1}$ in $\mathcal{F}(D_1,D_2,\ldots,D_r)$ such that, for every $1 \leq l'_1 \leq l_1$, we have

$$\begin{aligned}\mathbf{f}^{l'_1}_A &= \mathbf{g}^{l'_1}, \text{ and}\\ \mathbf{n} + \sum_{i=1}^{l'_1} \mathbf{f}^i &\in \mathbf{T}^{(\mathbf{n})}(D_1,D_2,\ldots,D_r).\end{aligned} \tag{34}$$

We obtain a table $\mathbf{x}' \in \mathbf{T}^{(\mathbf{n})}(D_1,D_2,\ldots,D_r)$, given by

$$\mathbf{x}' - \mathbf{n} = \sum_{i=1}^{l_1} \mathbf{f}^i, \tag{35}$$

such that the marginals $\mathbf{x}'_A$ and $\mathbf{x}_A$ are the same. Moreover, since we employed moves in $\mathcal{F}(D_1,\ldots,D_r)$, the marginals $\mathbf{x}'_{D_r}$ and $\mathbf{n}_{D_r}$ are also equal, and hence $\mathbf{x}' \in \mathbf{T}^{(\mathbf{x})}(A,D_r)$. This implies that we can find a series of moves $\mathbf{f}^{l_1+1},\ldots,\mathbf{f}^l$ in $\mathcal{F}(A,D_r)$ which transform the table $\mathbf{x}'$ in $\mathbf{x}$ i.e.

$$\begin{aligned}\mathbf{x} - \mathbf{x}' &= \sum_{i=l_1+1}^{l} \mathbf{f}^i, \text{ and}\\ \mathbf{x}' + \sum_{i=l_1+1}^{l'} \mathbf{f}^i &\in \mathbf{T}^{(\mathbf{x})}(A,D_r) \subset \mathbf{T}^{(\mathbf{n})}(D_1,D_2,\ldots,D_r),\end{aligned} \tag{36}$$

for $1 \leq l' \leq l$. From Eq. 34, Eq. 35 and Eq. 36 we obtain Eq. 32, which completes the proof. ■

Example 2. The graph $\mathcal{G}$ in Fig. 2 has 11 vertices and 28 edges. This is a decomposable graph with the set of cliques $\mathcal{C}(\mathcal{G}) = \{D_1,D_2,D_3,D_4\}$, where $D_1 := \{1,3,4,11\}$, $D_2 := \{3,4,7,8,9,11\}$, $D_3 := \{2,3,9,10\}$ and $D_4 := \{4,5,6,7\}$. The MCS algorithm constructs a tree $\mathcal{T}$ on $\mathcal{C}(\mathcal{G})$, where the edge set of $\mathcal{T}$ is $\mathcal{E}_{\mathcal{T}} = \{(D_2,D_1),(D_3,D_2),(D_4,D_2)\}$.

Therefore the separators of $\mathcal{G}$ are $S_2 := D_2 \cap D_1 = \{3,4,11\}$, $S_3 := D_3 \cap D_2 = \{3,9\}$, and $S_4 := D_4 \cap D_3 = \{4,7\}$. The set of primitive moves associated with $\mathcal{G}$ is

$$\mathcal{F}(\mathcal{G}) = \mathcal{F}(D_1, D_2 \cup D_3 \cup D_4) \cup \mathcal{F}(D_3, D_1 \cup D_2 \cup D_4) \cup \mathcal{F}(D_4, D_1 \cup D_2 \cup D_3).$$

Assume we are given an eleven-way table $\mathbf{n}$ with fixed marginals $\mathbf{n}_{D_1}$, $\mathbf{n}_{D_2}$, $\mathbf{n}_{D_3}$ and $\mathbf{n}_{D_4}$. The independence graph associated with $\mathbf{n}$ is $\mathcal{G}$. Theorem 1 tells us that $\mathcal{F}(\mathcal{G})$ is a Markov basis for the class of tables $\mathbf{T}^{(\mathbf{n})}(D_1,D_2,D_3,D_4)$. ■

Corollary 1 follows immediately from Theorem 1. If the marginals we are given are non-overlapping, then Corollary 1 provides us with the set of moves that will leave these marginals unchanged. Since two vertex sets included in two different connected components are separated by the empty set, the set of variables corresponding to each marginal will be unconditionally independent of the rest of the variables. Thus the order in which we consider the non-overlapping marginals does not make a difference.

Corollary 1. *Let $\mathcal{G}$ be an arbitrary graph having connected components $\mathcal{G}(D_1), \ldots, \mathcal{G}(D_s)$. Then the set of primitive moves*

$$\bigcup_{j=2}^{s} \mathcal{F}(D_1 \cup \ldots \cup D_{j-1}, D_j \cup \ldots \cup D_s) \tag{37}$$

is a Markov basis for the class of tables $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_s)$.

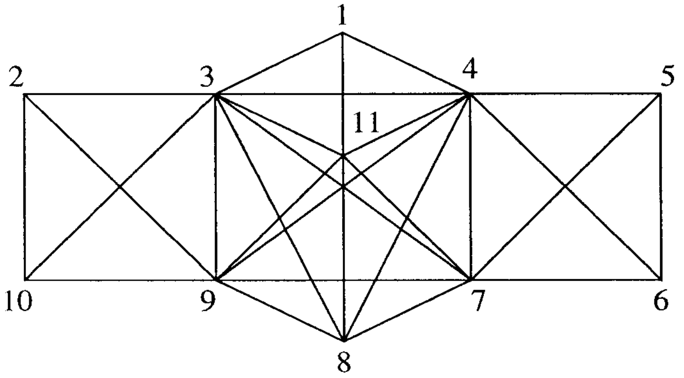

Figure 2: A decomposable graph with four cliques.

We can conclude that a Markov basis for a decomposable model with r cliques can be expressed as a union of the Markov bases of $(r-1)$ two-clique models. Because the Markov basis of a two-clique model is the set of primitive moves corresponding to one or more two-way tables, we deduce that the decomposable case essentially reduces to the two-way case.

The family of Markov bases we identified is extremely appealing to the potential user since one doesn't even need to actually produce the set of moves $\mathcal{F}(D_1, D_2, \ldots, D_r)$. This Markov basis could grow extremely large due to the size of the original table $\mathbf{n}$, hence handling it might become quite problematic. The procedure we outline below gets around this obstacle by dynamically generating moves in $\mathcal{F}(D_1, D_2, \ldots, D_r)$. The first step consists of computing the number of moves associated with every edge of the tree $\mathcal{T}$. We uniformly generate a primitive move in $\mathcal{F}(D_1, D_2, \ldots, D_r)$ by choosing an edge in $\mathcal{E}_{\mathcal{T}}$ with probability proportional to the number of primitive moves associated with it, then uniformly selecting a move from the set of primitive moves corresponding to the edge we picked.

Algorithm 1. *Let $\mathcal{T} = (\mathcal{C}(\mathcal{G}), \mathcal{E}_{\mathcal{T}})$ be the tree generated by the maximum cardinality search algorithm. The set of separators $\mathcal{S}(\mathcal{G}) = \{S_2, \ldots, S_r\}$ associated with $\mathcal{C}(\mathcal{G}) = \{D_1, \ldots, D_r\}$ will be given by $\mathcal{S}(\mathcal{G}) = \{D_j \cap D_j : (D_j, D_i) \in \mathcal{E}_{\mathcal{T}}\}$.*

- **for every** $S_l \in \mathcal{S}(\mathcal{G})$ **do**

- *There exists $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}}$ with $S_l = D_j \cap D_i$. Consider the subtrees $\mathcal{T}_j$ and $\mathcal{T}_i$ obtained by removing the edge (D_j, D_i) from $\mathcal{T}$, and let V_j and V_i be the vertex sets associated with these subtrees, as defined in Eq. 7.*
- *Calculate the weight w_l representing the number of primitive moves corresponding to the edge (D_j, D_i):*

$$w_l \leftarrow \left[2 \cdot \prod_{v \in V_j \setminus S_l} \binom{I_v}{2} \cdot \prod_{v \in V_i \setminus S_l} \binom{I_v}{2} \right]^{\prod\limits_{v \in S_l} I_v}.$$

end for

- *Normalize the weights $w_2, \ldots, w_r$:*

$$w_l \leftarrow \frac{w_l}{w_2 + \ldots + w_r}, \text{ for } l = 2, \ldots, r.$$

To uniformly select a move in $\mathcal{F}(\mathcal{G})$, follow the steps below:

1. *Randomly select an edge $(D_j, D_i) \in \mathcal{E}_{\mathcal{T}}$ with probability $P(S_l) = w_l$, where $S_l = D_j \cap D_i$.*
2. *Uniformly pick a move in $\mathcal{F}(V_j, V_i)$, where V_j and V_i were defined in Eq. 7. The set of primitive moves $\mathcal{F}(V_j, V_i)$ associated the decomposable model with two cliques V_j and V_i was described in Proposition 2.* ■

4 Example

In this section we present a straightforward technique for creating a replacement for a table having a fixed set of marginals. We refer to the data in Table 1 that comes 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 the 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 an agency has released three marginals, namely $\mathbf{n}_{BF}$, $\mathbf{n}_{ABCE}$ and $\mathbf{n}_{ADE}$, and now considers releasing the entire dataset $\mathbf{n}$. In Table 1, there are three entries containing "small" counts of "1" or "2". For various reasons, the agency believes that the identity of the individuals corresponding to these entries is not adequately protected and, consequently, the agency needs to find a replacement $\mathbf{n}'$ for $\mathbf{n}$ such that the cell counts in these three cells are modified. The replacement table has to be consistent with the marginals that were already made public, so that a possible intruder will not realize that the "real" table $\mathbf{n}$ was substituted with $\mathbf{n}'$.

The marginals $\mathbf{n}_{BF}$, $\mathbf{n}_{ABCE}$ and $\mathbf{n}_{ADE}$ define a decomposable independence graph $\mathcal{G}$, therefore we know a Markov basis

$$\mathcal{F}(\mathcal{G}) \quad = \quad \mathcal{F}(\{F\}, \{A, B, C, D, E\}) \cup \mathcal{F}(\{A, B, C, E, F\}, \{A, E, D\}), \tag{38}$$

for the class of tables

				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	$\boxed{1}$	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	$\boxed{2}$		[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	$\boxed{2}$	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: Autoworkers data (left-hand panel) and bounds for Autoworkers data (right-hand panel) given the marginals associated with the index sets $\{B, F\}$, $\{A, B, C, E\}$, and $\{A, D, E\}$.

$$\mathbf{T} := \mathbf{T}^{(\mathbf{n})}(\{B,F\},\{A,B,C,E\},\{A,D,E\}). \tag{39}$$

In this context, it is possible to construct a Markov chain on the space $\mathbf{T}$ by employing the methods of Diaconis and Sturmfels (1998). We assume that all the tables in $\mathbf{T}$ are equally probable. If the chain is currently in $\mathbf{x} \in \mathbf{T}$, we uniformly select a move $\mathbf{f}$ in $\mathcal{F}(\mathcal{G})$. If $\mathbf{x} + \mathbf{f}$ does not have a negative entry, the chain moves to $\mathbf{x} + \mathbf{f}$, otherwise it stays in $\mathbf{x}$. Since the moves in $\mathcal{F}(\mathcal{G})$ connect all the tables in $\mathbf{T}$, we will eventually find a table that satisfies our requirements provided that such a table exists in $\mathbf{T}$. It is important to notice that no burn-in period is needed. In our case, we started with Table 1 and, after making 52 primitive moves, we found Table 2. This table has the same marginals $\mathbf{n}_{BF}$, $\mathbf{n}_{ABCE}$, $\mathbf{n}_{ADE}$ as Table 1, but the three "small" counts of "1" and "2" are replaced by two counts of "0" and one count of "3".

5 Conclusions

The results described in this paper are relevant not only in the disclosure limitation context, but also in the general framework of log-linear models theory. Several fixed marginals induce a set of tables $\mathcal{W}$. When the index sets defining these fixed marginals are the cliques of a decomposable graph, we were able to fully characterize the set $\mathcal{W}$: we gave *formulas* for dynamically generating a Markov basis that allows one to reach any table in $\mathcal{W}$ starting from any other feasible table.

Techniques that worked well for low-dimensional examples are almost impossible to use for high-dimensional problems that arise in practice due to the huge computational effort they usually require. This paper demonstrates that graphical modeling is a very powerful tool for effectively overcoming the

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		44	41	112	64
			yes		126	149	14	23
		≥ 140	no		34	11	82	35
			yes		111	65	4	9
	≥ 3	< 140	no		24	32	70	68
			yes		50	78	5	14
		≥ 140	no		23	24	73	57
			yes		51	66	7	15
pos	< 3	< 140	no		7	7	20	10
			yes		10	15	$\boxed{0}$	3
		≥ 140	no		3	3	10	8
			yes		14	17	7	$\boxed{3}$
	≥ 3	< 140	no		9	2	16	12
			yes		8	17	$\boxed{0}$	4
		≥ 140	no		2	2	11	11
			yes		6	12	8	3

Table 2: Table obtained from Autoworkers data by preserving the marginals $\mathbf{n}_{BF}$, $\mathbf{n}_{ABCE}$ and $\mathbf{n}_{ADE}$.

curse of dimensionality. We were able to model the dependency patterns induced by a number of fixed marginals by means of graphs and, by doing so, we identified Markov bases for an entire family of sets of tables.

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. The author is extremely grateful to Dr. Stephen E. Fienberg and Dr. Stephen F. Roehrig for all their support, comments and valuable discussions. The author would like to thank Seth Sullivant who provided comments on an earlier version of the paper.

References

Adams, W. W. and Loustaunau, P. (1994). *An Introduction to Gröbner Bases*, Vol. 3 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, R.I.

Baker, R. J., Clarke, M. R. B., and Lane, P. W. (1985). "Zero Entries in Contingency Tables." *Computational Statistics & Data Analysis*, 3, 33–45.

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

Blair, J. R. S. and Barry, P. (1993). "An Introduction to Chordal Graphs and Clique Trees." In *Graph Theory and Sparse Matrix Computation*, ed. IMA, Vol. 56, 1–30. Springer-Verlag, New York.

Conti, P. and Traverso, C. (1991). "Buchberger Algorithm and Integer Programming." In *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes AAECC'9*, eds. Mattson, H. F., Mora, T., and Rao, T. R. N., Vol. 539 of *Lecture Notes in Computer Science*, 130–139. Springer-Verlag, Berlin.

Cox, D., Little, J., and O'Shea, D. (1992). "Ideals, Varieties and Algorithms." In *Undergraduate Texts in Mathematics*. Springer-Verlag.

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

Dalenius, T. and Reiss, S. P. (1982). "Data-swapping: a Technique for Disclosure Control." *Journal of Statistical Planning and Inference*, 6, 73–85.

Diaconis, P. and Efron, B. (1985). "Testing for Independence in a Two-Way Table: New Interpretations of the Chi-Square Statistic." *The Annals of Statistics*, 13, 845–874.

Diaconis, P. and Gangolli, A. (1995). "Rectangular Arrays with Fixed Margins." In *Discrete Probability and Algorithms*, 15–41. Springer-Verlag, New York.

Diaconis, P. and Sturmfels, B. (1998). "Algebraic Algorithms for Sampling From Conditional Distributions." *The Annals of Statistics*, 26, 363–397.

Dinwoodie, I. H. (1998). "The Diaconis-Sturmfels Algorithm and Rules of Succession." *Bernoulli*, 4, 401–410.

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.

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.

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.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

Haslett, S. (1990). "Degrees of Freedom and Parameter Estimability in Hierarchical Models for Sparse Complete Contingency Tables." *Computational Statistics & Data Analysis*, 9, 179–195.

Knuth, D. (1973). *The Art of Computer Programming*. Addison–Wesley, Reading, MA. Vol. 3.

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

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

Madigan, D. and York, J. (1995). "Bayesian Graphical Models for Discrete Data." *International Statistical Review*, 63, 215–232.

Mukerjee, R. (1987). "On Zero Cells in Log-linear Models." *Sankhyā: The Indian Journal of Statistics*, 49, 97–102.

Stirling, W. D. (1986). "A Note on Degrees of Freedom in Sparse Contingency Tables." *Computational Statistics & Data Analysis*, 4, 67–70.

Sturmfels, B. (1995). *Gröbner Bases and Convex Polytopes*. University Lecture Series, American Mathematical Society. Vol. 8.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons. New York.