Measuring the Disclosure Risk for Multiway Tables with Fixed Marginals Corresponding to Decomposable Loglinear Models

Adrian Dobra
Department of Statistics, Carnegie Mellon University
Pittsburgh, PA 15213-3890

October 11, 2000

Abstract

The underlying connection between disclosure avoidance techniques for categorical data and sampling from the exact conditional distribution associated with a loglinear model is the data swaps necessary to link all the contingency tables having a set of fixed marginal totals. In this paper we show that simple data swaps are the only swaps needed when the set of fixed margins define a decomposable graph. Furthermore, we characterize the set of data swaps associated with more general graphical structures.

Keywords: Contingency tables; Decomposable loglinear models; Reducible loglinear models; Disclosure limitation; Exact distributions; Markov chain Monte Carlo.

---

1. Introduction

Major statistical agencies, such as the U.S. Bureau of the Census or Statistics Canada, conduct extensive surveys to find out the general public opinion about important matters or to measure quantities of interest to policy makers. In most of the situations, the resulting databases can be viewed as multidimensional arrays where all the variables are categorical. To fully exploit the information they collected, the agencies are often forced to make available parts of these arrays (e.g., lower dimensional margins) to various users. The users could be individuals within the agency, independent researchers, the academic community, etc. Publicly releasing survey data could undoubtedly lead to revealing sensitive facts about the respondents since businesses or individuals might be in the position to divulge private information when answering the survey questions.

A potential intruder wants to learn confidential information from the released data, therefore the agencies have to develop and implement methods to protect the confidentiality of the respondents in order to avoid partial or full disclosure. However, the existent procedures for protecting contingency tables typically lead to a significant decrease in the accuracy and, consequently, in the utility of the publicly released datasets. Given that the users might make errornous inferences if the information they are provided with is overly distorted, the quality of the released data has to be a tradeoff between confidentiality and uncertainty.

We let $\mathbf{n} = \{n(i)\}_{i\in\mathcal{I}}$ (see Section 2 for notations) be the cross-classification representing the confidential database. The entries in table $\mathbf{n}$ are frequency counts, i.e., non-negative integers. From an intruder's perspective, these entries are variables corresponding to the unknown values he is trying to learn. The data collecting agency has already made public r marginal tables $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$. Assume that none of these tables, taken on its own, contains sensitive information. Since the released marginals might overlap, the intruder will attempt to find out additional information by matching the records in the public views. The goal of this paper is to present sound techniques for evaluating the disclosure risk associated with the combination of released marginal totals. Without employing suitable disclosure avoidance procedures, the data collecting agency will be completely ad-hoc in its decision-taking process, hence a violation of confidentiality might occur. To simplify our setup, we make the assumption that the intruder does not have access to any external sources of information relating to the confidential database the agency is trying to protect.

The contingency tables $\mathbf{x} = \{x(i)\}_{i\in\mathcal{I}}$ with non-negative integer entries and fixed marginal totals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$, are the lattice points in the convex polytope defined by the linear system of equations:

$$\begin{cases} \mathbf{x}_{D_1} = \mathbf{n}_{D_1}, \\ \mathbf{x}_{D_2} = \mathbf{n}_{D_2}, \\ \qquad \vdots \\ \mathbf{x}_{D_r} = \mathbf{n}_{D_r}. \end{cases} \tag{1}$$

We denote by $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ the set of admissible solutions of Eq. 1. The constraints given by the values in the released marginals induce upper and lower bounds on the interior cells of the initial table $\mathbf{n}$. These bounds are the solutions of the linear programming problems:

$$\min\{\pm x(i) : \mathbf{x} \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)\}. \tag{2}$$

The importance of systematically investigating Eq. 1 and Eq. 2 should be readily apparent. If the number of lattice points in $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ is below a certain threshold, we have significant evidence that a potential disclosure of the entire dataset $\mathbf{n}$ might have occurred. Moreover, if the bounds in Eq. 2 are too tight or too close to the actual sensitive value $n(i)$, the information associated with the individuals classified in cell $n(i)$ is likely to become public knowledge. The same would happen if the intruder could infer strictly positive lower bounds from Eq. 2.

The agency could accurately establish how revealing the released set of marginals might be by studying the linear systems defined above. The most natural method of determining sharp upper and lower bounds for the cell entries subject to some linear constraints is the simplex algorithm. In that case we would have to run the procedure twice for every element in the table, and consequently we would ignore the underlying dependencies among the marginals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ by regarding the maximization/minimization problem associated with some cell $i_0 \in \mathcal{I}$ as unrelated to the maximization/minimization problems associated with the rest of the cells $i \in \mathcal{I} \setminus \{i_0\}$. The computational inefficiency of the simplex approach is further augmented by the fact that the runtime and memory requirements increase exponentially with the size of the table. These considerations show the need for more dedicated, computationally unexpensive algorithms that could fully exploit the special structure of Eq. 1 and Eq. 2.

Another major reason of concern for an agency are population uniques. Individuals or establishments that have an uncommon combination of attributes translate into cell counts of "1" or "2" in the database $\mathbf{n}$. An intruder could effortlessly identify a survey respondent if marginal tables containing such small counts would be released. In order reduce the likelihood of this happening, the agency employs disclosure limitation methods such as cell suppression and data swapping. The object of both methods is to create a replacement table for the genuine unsafe marginal such that the counts smaller than some $s \geq 3$ are eliminated. The easiest approach would be to simply delete these small counts or *primary suppressions* (Fienberg et al., 1998). However, this might not assure adequate protection since one could still recover a deleted count from the unsuppressed cells and the marginal totals that were previously made public. As a consequence, we have to consider suppressing other cell values (*complementary suppressions*). Determining the minimum patterns of candidate cells for suppression is a hard computational problem, hence agencies delete cells in the reported views virtually at random by trial and error. Nevertheless, by deleting cells we could severely impair the utility of the data, thus the users can be led into false inferences.

Data swapping involves moving table entries from one cell to the other. If no more than one individual has been added to or deleted from every cell in the table, the data swap is called *simple*, otherwise it is a *generalized* swap. 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 $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$. The advantage of data swapping over cell suppression is that one could actually determine in real time the minimum number of data swaps that are sufficient to guarantee the confidentiality of the perturbed table $\mathbf{n}'$. Since we make only the changes that are strictly needed, the information conveyed by the table we are planning to release is nearly similar to the information contained in the "true" unsafe marginal.

Data swaps have one additional feature that makes them extremely appealing. Assume we were able to determine a finite set of swaps that could connect the initial table $\mathbf{n}$ to any other table in $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$. Then, starting from $\mathbf{n}$, one could easily generate elements of $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$, and hence determine an estimate of the number of tables with fixed marginals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$.

We notice that both table protection methods we presented so far preserve a given set of marginal totals. Loglinear models are the most usual way of representing and studying contingency tables with fixed marginals, hence it seems natural to attempt to establish a link between loglinear models and disclosure avoidance techniques. For example, calculating the number of degrees of freedom associated with a loglinear model could give us an idea about the cardinality of the class of tables $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$. The loglinear models theory will be exploited throughout the length of this paper.

In Section 2 we introduce generating sets of moves (or data swaps) and present an algebraic approach for obtaining them. As Fienberg (2000) postulated, simple data swaps are all we need to connect the tables in $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$, when $D_1, D_2, \ldots, D_r$ are the minimal sufficient statistics of a decomposable loglinear model. In Section 3 we give a detailed proof of this result and we describe the corresponding set of simple data swaps. In Section 4, by further exploiting the links with loglinear models theory, we extend the results to reducible graphical structures. Section 5 describes how the generating sets we identified could help us sample from the exact conditional distribution associated with a loglinear model, and in the final section we make some concluding remarks.

2. Data Swapping and Gröbner 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 $\mathcal{I} = \mathcal{I}_1 \bigotimes \mathcal{I}_2 \bigotimes \ldots \bigotimes \mathcal{I}_k$, 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))

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

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 $D = \{i_1, i_2, \ldots, i_p\}$ 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* $i_d \in \mathcal{I}_D := \mathcal{I}_{i_1} \bigotimes \mathcal{I}_{i_2} \bigotimes \ldots \bigotimes \mathcal{I}_{i_p}$ 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 $\mathbf{n}_1 = \mathbf{0}$. With these notations, we are ready to formally define the notion of swapping entries in a table while preserving a given set of marginal totals.

Definition 1. *A move or data swap for the set of marginal tables $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ 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 $\mathbf{x} \in \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$.*

From the second requirement in Definition 1, it follows that $\mathbf{f}$ is a move for $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ if and only if $\mathbf{f}$ is an integer solution for the linear system of equations:

$$\left\{ \begin{array}{rcl} \mathbf{f}_{D_1} & = & \mathbf{0}, \\ \mathbf{f}_{D_2} & = & \mathbf{0}, \\ & \vdots & \\ \mathbf{f}_{D_r} & = & \mathbf{0}. \end{array} \right. \tag{3}$$

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. An obvious solution of Eq. 3 is $\mathbf{f} = \mathbf{0}$. This degenerate move that leaves unchanged any table in class $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ will be called the *null move*. If $\mathbf{f} = \{f(i)\}_{i\in\mathcal{I}}$ is a move, then so is $-\mathbf{f} := \{-f(i)\}_{i\in\mathcal{I}}$. Moreover, if $f(i) \in \{-1, 0, 1\}, \forall i \in \mathcal{I}$, we call $\mathbf{f}$ a *simple move*.

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 *generating set* (or Markov basis, c.f. Diaconis and Sturmfels (1998)) for the tables in $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$. A class of moves $\mathcal{F}$ is a *minimal generating set* if, for any $\mathbf{f} \in \mathcal{F}$, the graph $\mathbf{T}_{\mathcal{F}\setminus\{\mathbf{f}\}}$ is no longer connected. For simplicity, we will assume that, in any generating set of moves $\mathcal{F}$, $\mathbf{f} \in \mathcal{F}$ implies $-\mathbf{f} \in \mathcal{F}$.

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 generating set for the class of tables $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$. If $\mathcal{F}$ is such a generating set, it should be easy to see that $\mathcal{F}$ is also a generating set for the class of tables

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

where $\mathbf{x} = \{x(i)\}_{i\in\mathcal{I}}$ is an arbitrary k-way cross-classification. Hence a generating set of moves is determined only by the set of marginal constraints, not by the actual table we started with. The notation we will employ in the rest of the paper will reflect this remark.

Computational algebra offers excellent tools for characterizing the possible solutions of Eq. 3. Diaconis and Sturmfels (1998) show that computing a generating set for $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ is equivalent to finding a special basis of a polynomial ideal. Following Cox, Little and O'Shea (1992), we let $\mathcal{K}$ be a field, and I be an ideal in $\mathcal{K}[x_1, x_2, \ldots, x_n]$. We denote by $LT(g)$ the leading term of a polynomial $g \in \mathcal{K}[x_1, x_2, \ldots, x_n]$ with respect to some monomial order, and by $LT(I)$ the set of leading terms of elements of I. The Hilbert Basis Theorem tells us that I has a finite generating set, i.e., $I = \langle g_1, g_2, \ldots, g_t\rangle$ for some $g_1, g_2, \ldots, g_t \in I$. Furthermore, the basis $G := \{g_1, g_2, \ldots, g_t\}$ of I can be chosen such that

$$\langle LT(I)\rangle = \langle LT(g_1), LT(g_2), \ldots, LT(g_t)\rangle,$$

and in this case G is said to be a Gröbner basis of I. One could 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 or MATHEMATICA implement this algorithm, hence the task of the users interested in finding a generating set for $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ is reduced to specifying the polynomial ideal associated with Eq. 3. Conti and Traverso (1991) showed how to find such a polynomial ideal by introducing a variable for each linear equation in Eq. 3, say $\sigma_1, \sigma_2, \ldots, \sigma_{L_1}$, and a variable for each unknown 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 generating sets for contingency tables with more than three dimensions. For example, even for a $3 \times 3 \times 3$ table, the Buchberger algorithm takes several hours to run.

It turns out, however, that it is straightforward to describe a generating set for a two-way table with fixed one-way marginals (Diaconis and Gangolli, 1995; Diaconis and Sturmfels, 1998; Fienberg et al., 2000).

Proposition 1. *Consider a two-way contingency table* $\mathbf{n} = \{n(i,j) : (i,j) \in \mathcal{I}_1 \otimes \mathcal{I}_2\}$ *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* $1 \leq i_1 < i_2 \leq I_1$ *and* $1 \leq j_1 < j_2 \leq I_2$, *we define a table* $\mathbf{f}^{i_1 i_2; j_1 j_2} = \{f^{i_1 i_2; j_1 j_2}(i,j) : (i,j) \in \mathcal{I}_1 \otimes \mathcal{I}_2\}$ *by*

$$f^{i_1 i_2; j_1 j_2}(i,j) = \begin{cases} 1, & \text{if } (i,j) \in \{(i_1,j_1),(i_2,j_2)\}, \\ -1, & \text{if } (i,j) \in \{(i_1,j_2),(i_2,j_1)\}, \\ 0, & \text{otherwise} . \end{cases} \tag{4}$$

Then

$$\left\{\mathbf{f}^{i_1 i_2; j_1 j_2} : 1 \leq i_1 < i_2 \leq I_1, 1 \leq j_1 < j_2 \leq I_2\right\} \cup \left\{-\mathbf{f}^{i_1 i_2; j_1 j_2} : 1 \leq i_1 < i_2 \leq I_1, 1 \leq j_1 < j_2 \leq I_2\right\} \tag{5}$$

is a generating set 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. 5 is indeed a generating set. For further reference, the number of moves in this generating set is

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

■

The set of simple 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 generating sets for an arbitrary decomposable graphical structure.

3. Generating Sets Associated with Decomposable Loglinear Models

The theory on loglinear 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 $\mathbf{n}_{C_1}, \mathbf{n}_{C_2}, \ldots, \mathbf{n}_{C_p}$ being fixed are the minimal sufficient of statistics of a decomposable loglinear model. In this particular case, the estimated expected values for the table entries can be written in closed form as a function of the marginal totals. Dobra and Fienberg (2000) employed a similar machinery to derive explicit formulas for the tightest upper and lower bounds for the cell counts in the table $\mathbf{n}$ given that $\mathbf{n}_{C_1}, \mathbf{n}_{C_2}, \ldots, \mathbf{n}_{C_p}$ are known. By further exploiting the special structural properties of decomposable graphs, we derive a generating set of simple moves for the class of tables $\mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$. We begin by presenting some results relating to decomposable graphs. We closely follow the notations and definitions introduced in Dobra and Fienberg (2000).

3.1. Decomposable Graphs. Let $X_1, X_2, \ldots, X_k$ be the discrete random variables cross-classified in the k-way table $\mathbf{n} = \{n(i)\}_{i \in \mathcal{I}}$. The vertex set of the independence graph $\mathcal{G}$ defined by the fixed marginals $\mathbf{n}_{C_1}, \mathbf{n}_{C_2}, \ldots, \mathbf{n}_{C_p}$ is $K := \bigcup_{j=1}^{p} C_j$, while its edge set is

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

Every variable X_j corresponds to a vertex $j \in K$. The conditional independence relationships induced among $X_1, X_2, \ldots, X_k$ by the fixed set of marginal totals are embedded in the graph $\mathcal{G}$: if $S \subset K$ is a separator for A_1 and A_2, then X_{A_1} and X_{A_2} are conditionally independent given X_S. A loglinear model with minimal sufficient statistics $C_1, C_2, \ldots, C_p$ is *graphical* if $C_1, C_2, \ldots, C_p$ are the set of cliques of the independence graph $\mathcal{G}$, otherwise the loglinear model is not graphical. Moreover, a graphical loglinear model is *decomposable* if the independence graph induced by its minimal sufficient statistics is decomposable.

Assume that $\mathcal{G}$ is decomposable and $\mathcal{C}(\mathcal{G}) := \{C_1, C_2, \ldots, C_p\}$ is 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. If we denote $H_j := C_1 \cup C_2 \cup \ldots \cup C_j$ and $S_j := H_{j-1} \cap C_j$, it follows that, for every $j = 2, \ldots, p$, $(H_{j-1} \setminus S_j, S_j, C_j \setminus S_j)$ is a decomposition of $\mathcal{G}(H_j)$ (Lauritzen, 1996). We let $\mathcal{S}(\mathcal{G}) := \{S_2, \ldots, S_p\}$ be the set of separators of the graph $\mathcal{G}$.

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, $(C, C') \in \mathcal{E}_{\mathcal{T}}$ implies $(C', C) \notin \mathcal{E}_{\mathcal{T}}$. If $(C, C') \in \mathcal{E}_{\mathcal{T}}$, we will say that C' is the *parent* of C and C is the *child* of C'. A clique C is *terminal* in $\mathcal{T}$ if C is not the parent of any other clique. Moreover, C is the *root* of the tree $\mathcal{T}$ if C 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 $S = C_j \cap C_i$ for some edge $(C_j, C_i) \in \mathcal{E}_{\mathcal{T}}$. The set of separators $\mathcal{S}(\mathcal{G})$ associated with $\mathcal{C}(\mathcal{G})$ will be given by $\mathcal{S}(\mathcal{G}) = \{C_j \cap C_i : (C_j, C_i) \in \mathcal{E}_{\mathcal{T}}\}$.

To determine a perfect sequence of cliques of $\mathcal{G}$, we identify a terminal clique C for which we set $C_1 \leftarrow C$. The next clique C_2 in the resulting perfect sequence of cliques will be a terminal clique in the tree obtained by eliminating C from $\mathcal{T}$. For the sake of completeness we present the clique numbering algorithm in pseudo-code. We assume that we know the number of cliques p of the graph $\mathcal{G}$.

- $q \leftarrow p$;
- $\mathcal{V} \leftarrow \mathcal{C}(\mathcal{G})$; $\mathcal{E} \leftarrow \mathcal{E}_{\mathcal{T}}$;
- **while** $\mathcal{E} \neq \emptyset$ **do**
    1. Identify a terminal clique C;
    2. Let C' be the unique clique in $\mathcal{V}$ with $(C, C') \in \mathcal{E}$;
    3. $\mathcal{V} \leftarrow \mathcal{V} \setminus \{C\}$;
    4. $\mathcal{E} \leftarrow \mathcal{E} \setminus \{(C, C')\}$;
    5. $C_q \leftarrow C$; $S_q \leftarrow C \cap C'$; $q \leftarrow q - 1$;

  end while
- Set C_1 to be the unique clique that is still in $\mathcal{V}$ after exiting the **while** loop;

The tree $\mathcal{T}$ generated by the MCS algorithm has one additional property that is essential for developing our results.

Definition 2 (The Star Property). *Take $C_j \in \mathcal{C}(\mathcal{G})$ and let $S = C_j \cap C_i$ for some $(C_j, C_i) \in \mathcal{E}_{\mathcal{T}}$. 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 (C_j, C_i) from $\mathcal{T}$, with $C_j \in \mathcal{K}_j$ and $C_i \in \mathcal{K}_i$. Consider the vertex sets*

$$V_j := \bigcup_{C \in \mathcal{K}_j} C \quad \text{and} \quad V_i := \bigcup_{C \in \mathcal{K}_i} C. \tag{6}$$

Then $(V_j \setminus S, S, V_i \setminus S)$ is a decomposition of $\mathcal{G}$ (Blair and Barry, 1993).

The maximum cardinality search algorithm, as described by 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

$$\mathcal{E} := \mathcal{E}_1 \cup \ldots \cup \mathcal{E}_L \cup \left\{ (C_1^{l-1}, C_2^l) : l = 2, \ldots, L \right\}, \tag{7}$$

where C_1^{l-1} is a root of $\mathcal{T}_{l-1}$ and C_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. 7 will have the Star Property on the set of cliques of an arbitrary decomposable graph $\mathcal{G}$.

By exploiting the existence of a perfect ordering of the cliques of the decomposable graph $\mathcal{G}$, it is straightforward to create an algorithm for generating random tables with given marginals $\mathbf{n}_{C_1}, \mathbf{n}_{C_2}, \ldots, \mathbf{n}_{C_p}$ based on an algorithm for generating random tables with given row and column totals. Such an algorithm has been described, for example, by Mount (1995) or Patefield (1981). First we consider a decomposable graph with two cliques.

Algorithm 1 (Random table generator for a decomposable model with two cliques).

We want to generate a random table $\mathbf{n} = \{n(i)\}_{i \in C_1 \cup C_2}$ given that we know its marginals $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$. We assume that $S = C_1 \cap C_2$ separates $C_1 \setminus S$ from $C_2 \setminus S$ in $\mathcal{G} = (C_1 \cup C_2, E)$, where $E := \{(u, v) : \{u, v\} \subset C_1 \text{ or } \{u, v\} \subset C_2\}$. We make use of the notations we will introduce in Algorithm 3.

- **for every** $i^0 = (i_q^0, \ldots, i_l^0) \in \mathcal{I}_S$ **do**
  - *Randomly generate a two-way table $\mathbf{n}^{i_q^0, \ldots, i_l^0}$ that cross-classifies the variables $Y_1^{i_q^0, \ldots, i_l^0}$ and $Y_2^{i_q^0, \ldots, i_l^0}$. The table $\mathbf{n}^{i_q^0, \ldots, i_l^0}$ has fixed row sums $\overline{\mathbf{n}}_{1, \ldots, r-1}^{i_q^0, \ldots, i_l^0}$ and column sums $\overline{\mathbf{n}}_{l+1, \ldots, m}^{i_q^0, \ldots, i_l^0}$.*

- **for every** $j \in \mathcal{J}$ **do**

  $n(i^0, j) \leftarrow n^{i_q^0, \ldots, i_l^0}(j)$.

  end for

end for

- *Output the table* $\mathbf{n}$. ■

We can repeatedly apply Algorithm 1 to obtain a random table generator for a decomposable graph with any number of cliques.

Algorithm 2 (Random table generator for a decomposable model).

- *Generate a random table* $\mathbf{x}^{(2)} = \left\{x^{(2)}(i)\right\}_{i \in \mathcal{I}_{H_2}}$ *with given marginals* $\mathbf{n}_{C_1}$ *and* $\mathbf{n}_{C_2}$.
- **for** $q = 3, \ldots, p$ **do**
  - *Generate a random table* $\mathbf{x}^{(q)} = \left\{x^{(q)}(i)\right\}_{i \in \mathcal{I}_{H_q}}$ *with given marginals* $\mathbf{x}^{(q-1)}$ *and* $\mathbf{n}_{C_q}$.

  end for
- *Output the final table* $\mathbf{x}^{(p)} \in \mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$. ■

Despite its simplicity, the random table generator we present below should be used with care since it is not clear what is the underlying distribution we are sampling from when employing this procedure. In Section 5 we give a Markov chain Monte Carlo approach for correctly generating tables from the uniform distribution on $\mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$.

3.2. Generating Sets for Decomposable Loglinear Models. By making use of the same technique we utilized for extending a random table generator for a two-way table to a random table generator for an arbitrary decomposable model, we develop an algorithm for determining a generating set for the class of tables $\mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$, when $C_1, C_2, \ldots, C_p$ are the cliques of a decomposable graph $\mathcal{G}$. It turns out that the generating set for a two-clique model is the union of the generating sets of one or more two-way tables with fixed one-way marginals. In other words, simple moves are the only moves we have to include in the generating set of a two-clique model.

Algorithm 3 (Generating set for a decomposable model with two cliques). *Suppose the independence graph* $\mathcal{G}$ *has only two cliques, namely* $C_1 = \{1, \ldots, l\}$ *and* $C_2 = \{q, \ldots, k\}$. *We distinguish two cases:*

(i) Suppose $l < q$. *In this case the variables in* C_1 *and* C_2 *are (unconditionally) independent. Introduce two new compound variables* Y_1 *and* Y_2 *with level sets* $\mathcal{I}_{C_1} := \bigotimes_{\delta \in C_1} \mathcal{I}_\delta$ *and* $\mathcal{I}_{C_2} := \bigotimes_{\delta \in C_2} \mathcal{I}_\delta$, *respectively. Take the two-way table* $\mathbf{n}'$ *corresponding to* Y_1, Y_2 *with known row sums* $\overline{\mathbf{n}}_{C_1} = \{n(i_{c_1})\}_{i_{c_1} \in \mathcal{I}_{C_1}}$ *and column sums* $\overline{\mathbf{n}}_{C_2} = \{n(i_{c_2})\}_{i_{c_2} \in \mathcal{I}_{C_2}}$. *The generating set* $\mathcal{F}(C_1, C_2)$

for the original table $\mathbf{n}$ *will be given by the generating set of moves for the two-way table* $\mathbf{n}'$ *as described in Proposition 1.*

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

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

Take the set of contingency tables

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

where

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

For a given vector of indices $(i_q^0,\ldots,i_l^0) \in \mathcal{I}_q \bigotimes \ldots \bigotimes \mathcal{I}_l$, *the variables in* $\{1,\ldots,q-1\}$ *and* $\{l+1,\ldots,k\}$ *are independent, hence we can pursue a line of reasoning analogous to (i). First we introduce two compound variables* $Y_1^{i_q^0,\ldots,i_l^0}$ *and* $Y_2^{i_q^0,\ldots,i_l^0}$ *with level sets* $\mathcal{I}_1 \bigotimes \ldots \bigotimes \mathcal{I}_{q-1}$ *and* $\mathcal{I}_{l+1} \bigotimes \ldots \bigotimes \mathcal{I}_k$, *respectively. The generating set* $\mathcal{F}^{i_q^0,\ldots,i_l^0}$ *for the table* $\mathbf{n}^{i_q^0,\ldots,i_l^0}$ *is equivalent to the generating set for the two-way table corresponding to* $Y_1^{i_q^0,\ldots,i_l^0}$ *and* $Y_2^{i_q^0,\ldots,i_l^0}$ *with known row sums* $\overline{\mathbf{n}}_{1,\ldots,q-1}^{i_q^0,\ldots,i_l^0}$ *and column sums* $\overline{\mathbf{n}}_{l+1,\ldots,k}^{i_q^0,\ldots,i_l^0}$. *Then one generating set of moves for the table* $\mathbf{n}$ *that would preserve the marginals* $\mathbf{n}_{C_1}$ *and* $\mathbf{n}_{C_2}$ *is given by*

$$\mathcal{F}(C_1,C_2) = \bigcup \left\{\mathcal{F}^{i_q^0,\ldots,i_l^0} : i_q^0\in\mathcal{I}_q,\ldots,i_l^0\in\mathcal{I}_l\right\}. \tag{8}$$

The generating set of moves $\mathcal{F}(C_1,C_2)$ *is clearly minimal for a model with two cliques* C_1 *and* C_2. *There are, however, reasons why working with such a set of moves might not be a good idea. If we were to select a move from the generating set given in Eq. 8, we would change exactly four cells in one particular "slice"* $\mathbf{n}^{i_q^0,\ldots,i_l^0}$ *of the original table* $\mathbf{n}$, *hence the resulting sequence of tables would be strongly correlated. We could easily overcome this drawback by noticing that modifying cells in one slice will not affect in any way the cells contained in the remaining slices. At every step, we could actually select a move in every slice of the table and obtain a generating set of compound moves. Nevertheless, we do not necessarily have to make changes in every slice of the table at any given step. We could randomly select a subset of slices and make a move only in that subset. Let* $\mathcal{F}_0^{i_q^0,\ldots,i_l^0}$ *be the set of moves obtained by adding the null move to the generating set* $\mathcal{F}^{i_q^0,\ldots,i_l^0}$. *The null move leaves the table* $\mathbf{n}^{i_q^0,\ldots,i_l^0}$ *unchanged. The resulting generating set for the complete table* $\mathbf{n}$ *is*

$$\mathcal{F}_0(C_1,C_2) = \bigotimes \left\{\mathcal{F}_0^{i_q^0,\ldots,i_l^0} : i_q^0\in\mathcal{I}_q,\ldots,i_l^0\in\mathcal{I}_l\right\}, \tag{9}$$

We cannot select the null move in every $\mathbf{n}^{i_q^0,\ldots,i_l^0}$ *because this will not make any changes to the original table. The application of the set of compound moves* $\mathcal{F}_0(C_1,C_2)$ *in the place of the set of simple moves* $\mathcal{F}(C_1,C_2)$ *will result in accelerating the convergence of the Markov chain. However, when proving the validity of the results we present in this paper, we prefer using the generating set of simple moves* $\mathcal{F}(C_1,C_2)$. ∎

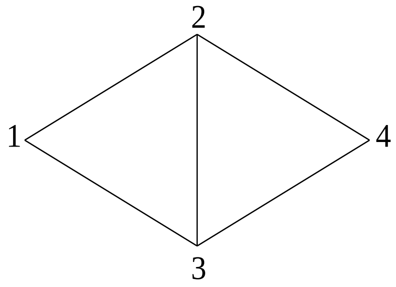

Figure 1: A decomposable graph with two cliques.

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

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

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 $\mathbf{n}_1^{i_2^0,i_3^0} := \{x(i_1,i_2^0,i_3^0) : i_1 \in \mathcal{I}_1\}$ and 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 generating set of moves $\mathcal{F}^{i_2^0,i_3^0}$ that will leave unchanged the one-way marginals of the table $\mathbf{n}^{i_2^0,i_3^0}$ could be obtained as in Proposition 1. Algorithm 3 tells us that the generating set of simple moves $\mathcal{F}(\{1,2,3\},\{2,3,4\})$ that would preserve the marginals $\mathbf{x}$ and $\mathbf{y}$ is just the union of all the sets $\left\{\mathcal{F}^{i_2^0,i_3^0} : (i_2^0,i_3^0) \in \mathcal{I}_2 \otimes \mathcal{I}_3\right\}$. $\blacksquare$

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

Definition 3. *Let $\mathcal{C}(\mathcal{G}) = \{C_1, C_2, \ldots, C_p\}$ 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 $(C_j, C_i) \in \mathcal{E}_{\mathcal{T}}$, we consider the vertex sets V_j and V_i as in Eq. 6. The set of simple moves associated with the decomposable graph $\mathcal{G}$ is given by:*

$$\mathcal{F}(\mathcal{G}) = \mathcal{F}(C_1, C_2, \ldots, C_p) := \bigcup_{(C_j,C_i)\in\mathcal{E}_{\mathcal{T}}} \mathcal{F}(V_j, V_i), \tag{10}$$

where $\mathcal{F}(V_j, V_i)$ was defined in Eq. 8.

By removing an edge (C_j, C_i) from $\mathcal{T}$, we disconnected it and end up with 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 $S_{ij} := C_j \cap C_i$ separates $V_j \setminus S_{ij}$ from $V_i \setminus S_{ij}$ in $\mathcal{G}^{ij}$. As a result, G^{ij} is the independence graph of a decomposable loglinear model with two cliques and, from Algorithm 3, we know that the set of simple moves corresponding to $\mathcal{G}^{ij}$ is $\mathcal{F}(V_j, V_i)$. Eq. 10 essentially says that the set of simple moves for a decomposable loglinear model with independence graph $\mathcal{G}$ is just the union of the sets of simple moves associated with the two-clique loglinear models induced by each minimal vertex separator of $\mathcal{G}$.

There might exist more than one tree $\mathcal{T}$ defined on the set of cliques of $\mathcal{G}$ such that $\mathcal{T}$ satisfies the Star Property. It is not hard to see that the set of simple moves $\mathcal{F}(\mathcal{G})$ in Eq. 10 is not influenced by the particular choice of $\mathcal{T}$ we worked with since $\mathcal{G}$ has a unique multiset of separators (Blair and Barry, 1993). We have to show that Definition 3 is correct.

Proposition 2. *The set of simple moves defined in Eq. 10 is indeed a set of moves for the class of tables* $\mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$.

Proof. Let $\mathbf{f} \in \mathcal{F}(C_1, C_2, \ldots, C_p)$. Then $\mathbf{f} \in \mathcal{F}(V_j, V_i)$ for some $(C_j, C_i) \in \mathcal{E}_{\mathcal{T}}$. For any arbitrary clique $C \in \mathcal{C}(\mathcal{G})$, we have either $C \subset V_j$ or $C \subset V_i$. Since $\mathbf{f}_{V_j} = \mathbf{0}$ and $\mathbf{f}_{V_i} = \mathbf{0}$, it follows that we also have $\mathbf{f}_C = \mathbf{0}$. ■

Next we will announce and justify a series of results that will help us prove the main theorem of our paper. Most of these propositions should be self-explanatory. However, it is worth mentioning the intuition that triggered them: if we delete a simplicial vertex of a decomposable graph along with the edges incident to it, we will obtain a graph that is still decomposable. Consequently, by collapsing across a variable associated with a simplicial vertex, we will preserve all the conditional dependencies existent among the remaining variables.

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 $\mathcal{G} = (V, E)$. Assume that $\mathcal{T}$ has the Star Property for $\mathcal{G}$. Let C be a terminal clique in $\mathcal{T}$ and let C' be the the unique clique in $\mathcal{C}(\mathcal{G})$ such that $(C, C') \in \mathcal{E}_{\mathcal{T}}$. We consider $\mathcal{T}' = (\mathcal{C}(\mathcal{G}) \setminus \{C\}, \mathcal{E}_{\mathcal{T}} \setminus \{(C, C')\})$ to be the tree obtained by removing C 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 $\mathcal{C}(\mathcal{G}) \setminus \{C\}$.*

Proof. Consider an arbitrary edge $(C_j, C_i) \in \mathcal{E}_{\mathcal{T}} \setminus \{(C, C')\}$. 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 (C_j, C_i) from $\mathcal{T}$, with $C_j \in \mathcal{T}_j$ and $C_i \in \mathcal{T}_i$. Let V_j and V_i the vertex sets defined in Eq. 6.

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

$$V_j' := \bigcup_{C'' \in \mathcal{K}_j \setminus \{C\}} C''. \tag{11}$$

Since C is terminal in $\mathcal{E}_{\mathcal{T}}$, we have $C_j \neq C$, hence $V_j' \neq \emptyset$. The vertex set $S := C_j \cap C_i$ 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 set of simple moves associated with a two-clique model induces a set of simple moves for a two-clique model embedded in it. Collapsing across some of the variables not contained in both cliques preserve the structure of the moves in Eq. 10.

Proposition 3. *Let $\mathcal{G}$ be a decomposable graph with two cliques C_1, C_2. The separator of $\mathcal{G}$ is $S := C_1 \cap C_2$. Consider a vertex set C such that $S \subset C \subset C_1$. Define a map ϕ which assigns to every $\mathbf{f} \in \mathcal{F}(C_1, C_2)$ its $(C \cup C_2)$-marginal, i.e.*

$$\phi(\mathbf{f}) = \mathbf{f}_{C \cup C_2}.$$

Then, for any $\mathbf{f} \in \mathcal{F}(C_1, C_2)$, $\phi(\mathbf{f}) \in \mathcal{F}(C, C_2)$ or $\phi(\mathbf{f}) = \mathbf{0}$. Moreover, the map $\phi : \mathcal{F}(C_1, C_2) \to \mathcal{F}(C, C_2)$ is surjective.

Proof. To simplify the notations, we will assume that $S = \emptyset$. We consider the marginals $\mathbf{n}_{C_1}$, $\mathbf{n}_{C_2}$ and $\mathbf{n}_C$, along with their associated vectors $\overline{\mathbf{n}}_{C_1}$, $\overline{\mathbf{n}}_{C_2}$ and $\overline{\mathbf{n}}_C$. The table $\mathbf{n}_C$ can be obtained from $\mathbf{n}_{C_1}$ by collapsing across the variables in $C_1 \setminus C$.

In Proposition 1, we constructed $\mathcal{F}(C_1, C_2)$ by considering the two-way table with row marginal $\overline{\mathbf{n}}_{C_1}$ and column marginal $\overline{\mathbf{n}}_{C_2}$. A simple move $\mathbf{f} \in \mathcal{F}(C_1, C_2)$ was obtained by choosing two row indices $i^1_{C_1}$ and $i^2_{C_1}$, and two column indices $i^1_{C_2}$ and $i^2_{C_2}$. Then the table $\mathbf{f}$ is given by:

$$f(i_{C_1}, i_{C_2}) = \begin{cases} \pm 1, & \text{if } (i_{C_1}, i_{C_2}) \in \left\{ (i^1_{C_1}, i^1_{C_2}), (i^2_{C_1}, i^2_{C_2}) \right\}, \\ \mp 1, & \text{if } (i_{C_1}, i_{C_2}) \in \left\{ (i^1_{C_1}, i^2_{C_2}), (i^2_{C_1}, i^1_{C_2}) \right\}, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

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

$$f_1(i_{C \cup C_2}) = f_1(i_C, i_{C_2}) = \sum_{j_{C_1 \setminus C} \in \mathcal{I}_{C_1 \setminus C}} f(j_{C_1 \setminus C}, i_C, i_{C_2}).$$

We distinguish two cases.

(i) $i^1_C = i^2_C$. Since $i^1_{C_1} \neq i^2_{C_1}$, we need to have $i^1_{C_1 \setminus C} \neq i^2_{C_1 \setminus C}$. It follows that

$$f_1(i^1_C, i^r_{C_2}) = f(i^1_{C_1}, i^r_{C_2}) + f(i^2_{C_1}, i^r_{C_2}) = 0, \text{ for } r = 1, 2.$$

Clearly, $f_1(i^1_C, i_{C_2}) = 0$ if $i_{C_2} \notin \{i^1_{C_2}, i^2_{C_2}\}$. Moreover, for $i_C \neq i^1_C$, we have $f_1(i_C, i_{C_2}) = 0$. Hence $\phi(\mathbf{f}) = \mathbf{f}_1 = \mathbf{0}$.

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

$$f_1(i_C, i_{C_2}) = \begin{cases} f(i^{r_1}_{C_1}, i^{r_2}_{C_2}), & \text{if } (i_C, i_{C_2}) \in \left\{ (i^{r_1}_C, i^{r_2}_{C_2}) : r_1, r_2 \in \{1, 2\} \right\}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus $\phi(\mathbf{f}) = \mathbf{f}_1 \in \mathcal{F}(C, C_2)$. In order to prove that ϕ is surjective, we pick an arbitrary move $\mathbf{f}_1 \in \mathcal{F}(C, C_2)$. We choose an index $i^0_{C_1 \setminus C} \in \mathcal{I}_{C_1 \setminus C}$ and define the move $\mathbf{f} = \{f(i)\}_{i \in \mathcal{I}_{C_1 \cup C_2}}$ by

$$f(i) = f(i_{C_1 \setminus C}, i_{C \cup C_2}) := \begin{cases} f_1(i_{C \cup C_2}), & \text{if } i_{C_1 \setminus C} = i^0_{C_1 \setminus C}, \\ 0, & \text{otherwise .} \end{cases}$$

It is easy to see that $\mathbf{f} \in \mathcal{F}(C_1, C_2)$ and $\phi(\mathbf{f}) = \mathbf{f}_1$. Therefore ϕ is indeed surjective on $\mathcal{F}(C, C_2)$. ∎

Proposition 4 extends Proposition 3 to an arbitrary decomposable model. Again, we are only allowed to collapse across variables that are associated with simplicial vertices in the corresponding independence graph.

Proposition 4. *Let $\mathcal{C}(\mathcal{G}) = \{C_1, C_2, \ldots, C_p\}$ the set of cliques of a decomposable graph $\mathcal{G}$ and let $\mathcal{T} = (\mathcal{C}(\mathcal{G}), \mathcal{E}_\mathcal{T})$ a tree having the Star Property for $\mathcal{G}$. We can assume that the clique C_1 is terminal in $\mathcal{T}$. Let $A := \bigcup_{j=2}^{p} C_j$. Define a map ϕ which assigns to every $\mathbf{f} \in \mathcal{F}(C_1, C_2, \ldots, C_p)$ its A-marginal i.e.*

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

Then, for any $\mathbf{f} \in \mathcal{F}(C_1, C_2, \ldots, C_p)$, $\phi(\mathbf{f}) \in \mathcal{F}(C_2, \ldots, C_p)$ or $\phi(\mathbf{f}) = \mathbf{0}$. In addition, the map ϕ is surjective on $\mathcal{F}(C_2, \ldots, C_p)$.

Proof. Since the clique C_1 is terminal in $\mathcal{T}$, there exists a unique clique in $\mathcal{C}(\mathcal{G})$, say C_2, such that $(C_1, C_2) \in \mathcal{E}_\mathcal{T}$. The set of simple moves corresponding to the edge (C_1, C_2) is $\mathcal{F}(C_1, A)$, and take $\mathbf{f} \in \mathcal{F}(C_1, A)$. By definition, $\mathbf{f}_A = \mathbf{0}$, hence $\phi(\mathbf{f}) = \mathbf{0}$.

The graph $\mathcal{G}'$ defined by the cliques $C_2, \ldots, C_p$ is decomposable and clearly $\mathcal{C}(\mathcal{G}') = \{C_2, \ldots, C_p\}$. Let $\mathcal{T}'$ the subtree obtained by removing C_1 from $\mathcal{T}$, i.e. $\mathcal{T}' = (\mathcal{C}(\mathcal{G}'), \mathcal{E}_\mathcal{T} \setminus \{(C_1, C_2)\})$. Consider an arbitrary edge $(C_j, C_i) \in \mathcal{E}_\mathcal{T} \setminus \{(C_1, C_2)\}$. 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 (C_j, C_i) from $\mathcal{T}$, with $C_j \in \mathcal{K}_j$ and $C_i \in \mathcal{K}_i$. Without restricting the generality, we assume that we always have $C_1 \in \mathcal{K}_j$.

By removing the same edge from the tree $\mathcal{T}'$, we obtain the subtrees $\mathcal{T}_j' = (\mathcal{K}_j \setminus \{C_1\}, \mathcal{E}_j \setminus \{(C_1, C_2)\})$ and $\mathcal{T}_i$. We define the vertex sets V_j, V_j', and V_i by

$$V_j := \bigcup_{C \in \mathcal{K}_j} C, \quad V_j' := \bigcup_{C \in \mathcal{K}_j \setminus \{C_1\}} C, \quad V_i := \bigcup_{C \in \mathcal{K}_i} C. \tag{13}$$

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

$$\mathcal{F}(\mathcal{G}') = \mathcal{F}(C_2, \ldots, C_p) = \bigcup_{(C_j, C_i) \in \mathcal{E}_\mathcal{T} \setminus \{(C_1, C_2)\}} \mathcal{F}(V_j', V_i). \tag{14}$$

Consider an arbitrary move $\mathbf{f} \in \mathcal{F}(C_1, C_2, \ldots, C_p)$ such that $\mathbf{f} \notin \mathcal{F}(C_1, A)$. From Eq. 10, we see that there must exist some edge $(C_j, C_i) \in \mathcal{E}_\mathcal{T} \setminus \{(C_1, C_2)\}$ such that $\mathbf{f} \in \mathcal{F}(V_j, V_i)$. We have $C_j \neq C_1$ and $C_j \subset V_j$, thus $V_j' \neq \emptyset$. In addition, we have $V_j' \subset V_j$ and $A = V_j' \cup V_i$. By employing Proposition 3, we obtain that $\phi(\mathbf{f}) \in \mathcal{F}(V_j', V_i) \subset \mathcal{F}(C_2, \ldots, C_p)$ or $\phi(\mathbf{f}) = \mathbf{0}$.

To prove that ϕ is surjective on $\mathcal{F}(C_2, \ldots, C_p)$, we pick an arbitrary move $\mathbf{f}_1 \in \mathcal{F}(C_2, \ldots, C_p)$. From Eq. 14, we see that there is an edge $(C_j, C_i) \in \mathcal{E}_\mathcal{T} \setminus \{(C_1, C_2)\}$ such that $\mathbf{f}_1 \in \mathcal{F}(V_j', V_i)$. Since $V_j' \subset V_j$, Proposition 3 tells us that there must exist some $\mathbf{f} \in \mathcal{F}(V_j, V_i) \subset \mathcal{F}(C_1, C_2, \ldots, C_p)$ such that $\phi(\mathbf{f}) = \mathbf{f}_1$. Therefore ϕ is indeed surjective. ■

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 $\mathcal{C}(\mathcal{G}) = \{C_1, C_2, \ldots, C_p\}$. Then the set of simple moves $\mathcal{F}(\mathcal{G}) = \mathcal{F}(C_1, C_2, \ldots, C_p)$ defined in Eq. 10 is a generating set for the class of tables $\mathbf{T}^{(\mathbf{n})} := \mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$.*

Proof. By induction. If $\mathcal{G}$ decomposes in $p = 2$ cliques, then we know from Proposition 1 that $\mathcal{F}(C_1, C_2)$ is a generating set for $\mathbf{T}^{(\mathbf{n})}(C_1, C_2)$. Suppose the theorem holds for any decomposable graph with $p-1$ cliques. We want to prove that the theorem is true for a decomposable graph with p cliques.

The original table $\mathbf{n}$ is in the set $\mathbf{T}^{(\mathbf{n})} = \mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$. Take an arbitary table $\mathbf{x} \in \mathbf{T}^{(\mathbf{n})}$. We have to show that there exist $\mathbf{f}^1, \ldots, \mathbf{f}^q \in \mathcal{F}(C_1, C_2, \ldots, C_p)$ such that

$$\mathbf{x} - \mathbf{n} = \sum_{i=1}^{q} \mathbf{f}^i. \tag{15}$$

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 C_1 is terminal in $\mathcal{T}$. Denote $A := \overset{p}{\underset{j=2}{\cup}} C_j$. Consider the map ϕ which assigns to every $\mathbf{f} \in \mathcal{F}(C_1, C_2, \ldots, C_p)$ its A-marginal i.e.

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

The marginals $\mathbf{n}_A$ and $\mathbf{x}_A$ lie in the set $\mathbf{T}^{(\mathbf{n})}(C_2, \ldots, C_p)$. From the induction hypothesis we know that $\mathcal{F}(C_2, \ldots, C_p)$ is a generating set for $\mathbf{T}^{(\mathbf{n})}(C_2, \ldots, C_p)$, so there exists a sequence of moves $\mathbf{g}^1, \ldots, \mathbf{g}^{q_1} \in \mathcal{F}(C_2, \ldots, C_p)$ such that

$$\mathbf{x}_A - \mathbf{n}_A = \sum_{i=1}^{q_1} \mathbf{g}^i.$$

Proposition 4 tells us that the sequence of moves $\mathbf{g}^1, \ldots, \mathbf{g}^{q_1}$ translates into a sequence of moves $\mathbf{f}^1, \ldots, \mathbf{f}^{q_1}$ in $\mathcal{F}(C_1, C_2, \ldots, C_p)$ by choosing $\mathbf{f}^i \in \phi^{-1}(\mathbf{g}^i), 1 \le i \le q_1$. We obtain a table $\mathbf{x}' \in \mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$, given by

$$\mathbf{x}' - \mathbf{n} = \sum_{i=1}^{q_1} \mathbf{f}^i, \tag{16}$$

such that the marginals $\mathbf{x}'_A$ and $\mathbf{x}_A$ are the same. Moreover, since we employed moves in $\mathcal{F}(C_1, C_2, \ldots, C_p)$, the marginals $\mathbf{x}'_{C_1}$ and $\mathbf{n}_{C_1}$ are also equal, and hence $\mathbf{x}' \in \mathbf{T}^{(\mathbf{x})}(C_1, A)$. This implies that we can find a series of moves $\mathbf{f}^{q_1+1}, \ldots, \mathbf{f}^q$ in $\mathcal{F}(C_1, A)$ which transform the table $\mathbf{x}'$ in $\mathbf{x}$ i.e.

$$\mathbf{x} - \mathbf{x}' = \sum_{i=q_1+1}^{q} \mathbf{f}^i. \tag{17}$$

From Eq. 16 and Eq. 17 we obtain Eq. 15, fact that completes the proof. ■

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 should not make a difference.

Corollary 1. *Let $\mathcal{G}$ be an arbitrary graph having connected components $\mathcal{G}(D_1), \mathcal{G}(D_2), \ldots, \mathcal{G}(D_s)$. Then the set of simple moves*

$$\bigcup_{j=2}^{s} \mathcal{F}(D_1 \cup \ldots \cup D_{j-1}, D_j \cup \ldots \cup D_s) \tag{18}$$

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

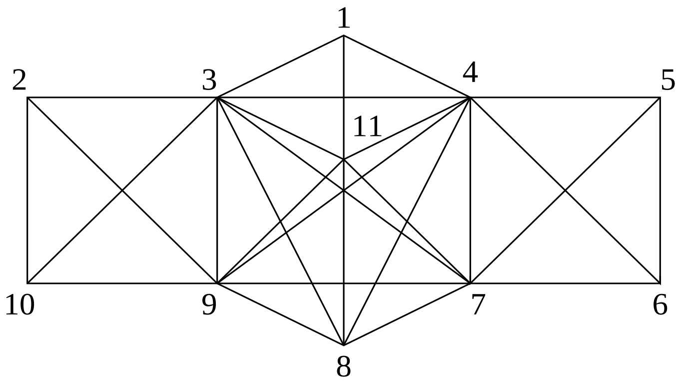

Figure 2: A decomposable graph with four cliques.

Example 2. The graph $\mathcal{G}$ in Fig. 2 has 11 vertices and 28 edges. Consider the vertex sets $C_1 := \{1, 3, 4, 11\}$, $C_2 := \{3, 4, 7, 8, 9, 11\}$, $C_3 := \{2, 3, 9, 10\}$ and $C_4 := \{4, 5, 6, 7\}$. The set of cliques of $\mathcal{G}$ is $\mathcal{C}(\mathcal{G}) = \{C_1, C_2, C_3, C_4\}$. The MCS algorithm constructs a tree $\mathcal{T}$ on $\mathcal{C}(\mathcal{G})$, where the edge set of $\mathcal{T}$ is $\mathcal{E}_{\mathcal{T}} = \{(C_2, C_1), (C_3, C_2), (C_4, C_2)\}$.

We can see that the cliques C_4 and C_3 are terminal in $\mathcal{T}$. After we eliminate them from $\mathcal{T}$, the clique C_2 becomes terminal. C_1 is the root of the tree. A possible perfect ordering of the cliques of $\mathcal{G}$ is C_1, C_2, C_3, C_4. Another perfect ordering could be C_1, C_2, C_4, C_3. Therefore $\mathcal{G}$ is decomposable and its separators are $S_2 := C_2 \cap C_1 = \{3, 4, 11\}$, $S_3 := C_3 \cap C_2 = \{3, 9\}$, and $S_4 := C_4 \cap C_3 = \{4, 7\}$.

The set of simple moves associated with $\mathcal{G}$ is

$$\mathcal{F}(\mathcal{G}) = \mathcal{F}(C_1, C_2 \cup C_3 \cup C_4) \cup \mathcal{F}(C_3, C_1 \cup C_2 \cup C_4) \cup \mathcal{F}(C_4, C_1 \cup C_2 \cup C_3).$$

Assume we are given an eleven-way table $\mathbf{n}$ with fixed marginals $\mathbf{n}_{C_1}$, $\mathbf{n}_{C_2}$, $\mathbf{n}_{C_3}$ and $\mathbf{n}_{C_4}$. The independence graph associated with $\mathbf{n}$ is $\mathcal{G}$. Theorem 1 tells us that $\mathcal{F}(\mathcal{G})$ is a generating set for the class of tables $\mathbf{T}^{(\mathbf{n})}(C_1, C_2, C_3, C_4)$. ■

We can conclude that a generating set of moves for a decomposable model with p cliques can be expressed as a union of the generating sets of $(p-1)$ two-clique models. Because the generating set of a two-clique model is the set of simple moves corresponding to one or more two-way tables, we deduce that the decomposable case essentially reduces to the two-way case. The set of simple moves $\mathcal{F}(C_1, C_2, \ldots, C_p)$ is not, however, a minimal generating set for the class of tables $\mathbf{T}^{(\mathbf{n})}(C_1, C_2, \ldots, C_p)$.

The generating set associated with a decomposable model we proposed is extremely appealing to the potential user since one doesn't even need to actually produce the set of moves $\mathcal{F}(C_1, C_2, \ldots, C_p)$. $\mathcal{F}(C_1, C_2, \ldots, C_p)$ 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 $\mathcal{F}(C_1, C_2, \ldots, C_p)$. The first step consists of computing the number of moves associated with every edge of the tree $\mathcal{T}$. We uniformly generate a simple move in $\mathcal{F}(C_1, C_2, \ldots, C_p)$ by choosing an edge in $\mathcal{E}_{\mathcal{T}}$ with probability proportional with the number of simple moves associated with it, then uniformly selecting a move from the set of simple moves corresponding to the edge we picked.

Algorithm 4. *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 $\mathcal{S}(\mathcal{G}) = \{S_2, \ldots, S_p\}$ associted with $\mathcal{C}(\mathcal{G}) = \{C_1, C_2, \ldots, C_p\}$ will be given by $\mathcal{S}(\mathcal{G}) = \{C_j \cap C_j : (C_j, C_i) \in \mathcal{E}_{\mathcal{T}}\}$.*

- **for every** $S_l \in \mathcal{S}(\mathcal{G})$ **do**
  - *Assume that $S_l = C_j \cap C_i$, where $(C_j, C_i) \in \mathcal{E}_{\mathcal{T}}$. Consider the subtrees $\mathcal{T}_j$ and $\mathcal{T}_i$ obtained by removing the edge (C_j, C_i) from $\mathcal{T}$, and let V_j and V_i be the vertex sets associated with these subtrees, as defined in Eq. 6.*
  - *Calculate the weight*

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

  The weight w_l represents the number of simple moves corresponding to the edge (C_j, C_i).

  end for

- *Normalize the weights $w_2, \ldots, w_p$, i.e.*

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

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

1. *Randomly select an edge $(C_j, C_i) \in \mathcal{E}_{\mathcal{T}}$ with probability $P(S_l) = w_l$, where $S_l = C_j \cap C_i$.*
2. *Consider the decomposable model with two cliques V_j and V_i, where V_j and V_i were defined in Eq. 6. The set of simple moves $\mathcal{F}(V_j, V_i)$ associated with this model was described in Algorithm 3.*
3. *Uniformly pick a move in $\mathcal{F}(V_j, V_i)$.* ■

To reduce the correlation between two consecutive tables, we could replace the simple moves $\mathcal{F}(V_j, V_i)$ with the compound moves $\mathcal{F}_0(V_j, V_i)$ introduced in Eq. 9. The algorithm will remain the same, except that we would have to modify accordingly the weights associated with the edges (C_j, C_i):

$$w_l \leftarrow \left[2 \cdot \prod_{v \in V_j \setminus S_l} \binom{I_v}{2} \cdot \prod_{v \in V_i \setminus S_l} \binom{I_v}{2} + 1\right]^{\prod_{v \in S_l} I_v} - 1.$$

Again, w_l represents the number of compound moves in $\mathcal{F}_0(V_j, V_i)$. In the next section we will study generating sets for a larger class of models which include decomposable models as a special case.

4. Generating Sets Associated with Reducible Loglinear Models

Hitherto we have demonstrated how one could determine a generating set for a class of tables defined by a fixed set of marginal totals when that set of marginals induce a decomposable graph. We were able to completely specify a generating set because we could recursively decompose the independence graph into its cliques by means of complete separators. The natural question to ask is whether we could extend this approach to more general graphical structures that have a proper decomposition but are not decomposable. For example, the triangle $\{3, 4, 11\}$ is a minimal complete separator for the graph in Fig. 3, but the graph itself is not decomposable because it has a four-cycle $\{2, 3, 9, 10\}$. Again, we will follow the notations and definitions introduced in Dobra and Fienberg (2000).

A graph $\mathcal{G}$ that admits a proper decomposition will be called *reducible*, otherwise $\mathcal{G}$ is a *prime* graph. The reducible graph $\mathcal{G}$ can be decomposed in two subgraphs $\mathcal{G}_1$ and $\mathcal{G}_2$, then, if possible, $\mathcal{G}_1$ and $\mathcal{G}_2$ could be decomposed, and so on, until a set of prime subgraphs of $\mathcal{G}$ is generated. We will refer to such a set of prime subgraphs as a *derived system* of $\mathcal{G}$. Since a graph might have more than one minimal separator, there could exist several feasible derived systems, hence we would like to single out one of them that could fully characterize the input graph.

A subgraph is a *maximal prime* (mp-) subgraph if it is prime and its subgraphs are all reducible. If the input graph $\mathcal{G}$ is decomposable, the mp-subgraphs will be complete, i.e., the cliques of $\mathcal{G}$ are maximal prime subgraphs in $\mathcal{G}$. It is not hard to see that the set $\mathcal{V}(\mathcal{G})$ of mp-subgraphs of $\mathcal{G}$ is contained in every derived system of $\mathcal{G}$ and, in addition, it is a derived system by itself (Leimer, 1993). If we could select the appropriate separator at every step of the decomposition procedure, we would prevent the inclusion of nonmaximal prime sugraphs in the derived system we produce. It turns out that one could employ the maximum cardinality search algorithm to generate the set of mp-subgraphs of a reducible graph (Leimer, 1993; Dobra and Fienberg, 2000). The MCS algorithm will construct a tree on $\mathcal{V}(\mathcal{G})$, and this tree will have the Star Property. Therefore, by employing the almost same procedures as in the decomposable case, we could produce a perfect sequence of mp-subgraphs of a reducible graph. Unlike in the decomposable case, it is not possible to describe a generating set of moves for the class of tables with fixed marginals when these marginals define a reducible graph. Nevertheless, the procedure we present will significantly reduce the computational effort required to produce such a generating set. Assuming we were provided with a generating set for every class of tables corresponding to a mp-subgraph of $\mathcal{G}$, we show how we could combine these generating sets into a generating set for the initial table $\mathbf{n}$.

To be more explicit, consider a reducible graph $\mathcal{G} = (K, E)$ with a proper decomposition $(V_1 \setminus S, S, V_2 \setminus S)$. We are given two generating sets $\mathcal{H}_1$ and $\mathcal{H}_2$ corresponding to the loglinear models defined by the independence graphs $\mathcal{G}(V_1)$ and $\mathcal{G}(V_2)$, respectively. Our goal is to find a way to combine $\mathcal{H}_1$ and $\mathcal{H}_2$ into a generating set $\mathcal{H}$ for the loglinear model with independence graph $\mathcal{G}$. In addition, consider the graphs $\mathcal{G}_1 = (K, E_1)$, $\mathcal{G}_2 = (K, E_2)$, and $\mathcal{G}_3 = (K, E_3)$, where

$$\begin{aligned} E_1 &:= \{(u,v) : (u,v) \in E \cap (V_1 \times V_1) \text{ or } \{u,v\} \subset V_2\}, \\ E_2 &:= \{(u,v) : (u,v) \in E \cap (V_2 \times V_2) \text{ or } \{u,v\} \subset V_1\}, \\ E_3 &:= \{(u,v) : \{u,v\} \subset V_1 \text{ or } \{u,v\} \subset V_2\}. \end{aligned}$$

It is easy to notice that $\mathcal{C}(\mathcal{G}_1) = \mathcal{C}\left(\mathcal{G}(V_1)\right) \cup \{V_2\}$, $\mathcal{C}(\mathcal{G}_2) = \mathcal{C}\left(\mathcal{G}(V_2)\right) \cup \{V_1\}$, and $\mathcal{C}(\mathcal{G}_3) = \{V_1, V_2\}$. In Proposition 5 we extend $\mathcal{H}_1$, $\mathcal{H}_2$ to two sets of moves $\mathcal{H}_1'$, $\mathcal{H}_2'$ that will keep unchanged the minimal sufficient statistics of the loglinear model with independence graphs $\mathcal{G}_1$, $\mathcal{G}_2$. We do so by making use of a strategy similar to the one we employed when we specified a generating set for a two-clique model based on a generating set for a two-way table with fixed one-way marginals. The loglinear model with independence graph $\mathcal{G}_3$ is a decomposable model with two cliques, hence its generating set is $\mathcal{F}(V_1, V_2)$. In Theorem 2, we prove that

$$\mathcal{H} = \mathcal{H}_1' \cup \mathcal{H}_2' \cup \mathcal{F}(V_1, V_2).$$

The idea is to apply moves in $\mathcal{H}_1$ to the initial table $\mathbf{n}$ while keeping its V_2-marginal unchanged. Then, apply moves in $\mathcal{H}_2$ to the resulting table $\mathbf{n}'$ such that its V_1-marginal remains fixed. All we are left to do is apply a series of moves in $\mathcal{F}(V_1, V_2)$ to reach any table in the class of tables induced by the minimal sufficient statistics of the loglinear model represented by $\mathcal{G}$.

Proposition 5. *Let $D_1, D_2, \dots, D_r$ be vertex sets of a graph $\mathcal{G} = (V, E)$. Denote $V_1 := \bigcup_{i=1}^{r} D_i$ and take $V_2 \subset V$. We assume that (i) $D_1, D_2, \dots, D_r$ and V_2 are complete in $\mathcal{G}$, (ii) $S = V_1 \cap V_2$ separates $V_1 \setminus S$ from $V_2 \setminus S$ in $\mathcal{G}(V_1 \cup V_2)$, and (iii) there exists some $r_0 \in \{1, \dots, r\}$ such that $S \subseteq D_{r_0}$. In addition, we are given a set of moves $\mathcal{H}(D_1, D_2, \dots, D_r)$ for the class of tables $\mathbf{T}(D_1, D_2, \dots, D_r)$. Consider a map ψ which assigns to every $i^0_{V_2 \setminus S} \in \mathcal{I}_{V_2 \setminus S}$ and every $\mathbf{g} \in \mathcal{H}(D_1, D_2, \dots, D_r)$ a move $\mathbf{f} = \{f(i)\}_{i \in \mathcal{I}_{V_1 \cup V_2}}$ defined by*

$$f(i) = f(i_{V_1}, i_{V_2 \setminus S}) := \begin{cases} g(i_{V_1}), & \text{if } i_{V_2 \setminus S} = i^0_{V_2 \setminus S}, \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Then the image $\mathcal{H}(D_1, D_2, \dots, D_r || V_2 \setminus S)$ of $\mathcal{I}_{V_2 \setminus S} \bigotimes \mathcal{H}(D_1, D_2, \dots, D_r)$ through the map ψ is a set of moves for the class of tables $\mathbf{T}(D_1, D_2, \dots, D_r, V_2)$.

Proof. Let $\mathbf{f} \in \mathcal{H}(D_1, D_2, \dots, D_r || V_2 \setminus S)$, $i^0_{V_2 \setminus S} \in \mathcal{I}_{V_2 \setminus S}$ and $\mathbf{g} \in \mathcal{H}(D_1, D_2, \dots, D_r)$ such that $\psi(i^0_{V_2 \setminus S}, \mathbf{g}) = \mathbf{f}$. The marginal $\mathbf{f}_{V_2}$ of $\mathbf{f}$ has entries

$$\begin{aligned} f_{V_2}(i_{V_2}) &= \sum_{j_{V_1 \setminus S} \in \mathcal{I}_{V_1 \setminus S}} f(j_{V_1 \setminus S}, i_{V_2}) \\ &= \sum_{j_{V_1 \setminus S} \in \mathcal{I}_{V_1 \setminus S}} f(j_{V_1 \setminus S}, i_S, i_{V_2 \setminus S}) \\ &= \begin{cases} \sum\limits_{j_{V_1 \setminus S} \in \mathcal{I}_{V_1 \setminus S}} g(j_{V_1 \setminus S}, i_S) = g_S(i_S), & \text{if } i_{V_2 \setminus S} = i^0_{V_2 \setminus S}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Since $S \subset D_{r_0}$, $\mathbf{g}_{D_{r_0}} = \mathbf{0}$ implies $\mathbf{g}_S = \mathbf{0}$, and hence $\mathbf{f}_{V_2} = \mathbf{0}$. Next we look at the marginal $\mathbf{f}_{V_1}$ of $\mathbf{f}$:

$$\begin{aligned} f_{V_1}(i_{V_1}) &= \sum_{j_{V_2 \setminus S} \in \mathcal{I}_{V_2 \setminus S}} f(i_{V_1}, j_{V_2 \setminus S}) \\ &= f(i_{V_1}, i^0_{V_2 \setminus S}) \\ &= g(i_{V_1}). \end{aligned}$$

We have $D_j \subset V_1$, thus $\mathbf{f}_{D_j} = \mathbf{g}_{D_j}$. But $\mathbf{g}_{D_j} = \mathbf{0}$, which implies $\mathbf{f}_{D_j} = \mathbf{0}$. It follows that we indeed have $\mathbf{f} \in \mathbf{T}(D_1, D_2, \ldots, D_r, V_2)$. ■

To reduce the dependency between two consecutive tables, we could define a set of compound moves associated with $\mathcal{H}(D_1, D_2, \ldots, D_r || V_2 \setminus S)$ in an analogous way we defined the compound moves $\mathcal{F}_0(V_j, V_i)$ associated with $\mathcal{F}(V_j, V_i)$. Instead of making a move in only one slice determined by an index $i^0_{V_2 \setminus S} \in \mathcal{I}_{V_2 \setminus S}$, we could simultaneously make various moves in several slices corresponding to indices in $\mathcal{I}_{V_2 \setminus S}$. By proceeding in this manner we change more cell entries at every given step, and consequently reduce the correlation between two successive draws. We have to make sure, however, that all the moves but one cancel when we collapse across the variables indexed by $V_2 \setminus S$. In the following, $|A|$ is the number of elements in the set A.

Proposition 6. *We use the same notations as in Proposition 5. We partition the set $\mathcal{I}_{V_2 \setminus S}$ in two subsets $\mathcal{I}_0$, $\mathcal{I}_1$ such that the number of elements in $\mathcal{I}_1$ is odd, i.e. $|\mathcal{I}_1| = 2q + 1$. We order the indices in $\mathcal{I}_1$ in lexicographical order, and let $\mathcal{I}_2$ be the set containing the first $q+1$ elements in $\mathcal{I}_1$. Denote $\mathcal{I}_3$ the set containing the remaining q elements. Define a map ϕ that associates an index in $\mathcal{I}_2$ to each index from $\mathcal{I}_3$. The map ϕ is injective, i.e.*

$$\phi(i^1_{V_2 \setminus S}) \neq \phi(i^2_{V_2 \setminus S}), \text{ if } i^1_{V_2 \setminus S} \neq i^2_{V_2 \setminus S}.$$

There will exist an unique index $i^0_{V_2 \setminus S} \in \mathcal{I}_2$ such that $i^0_{V_2 \setminus S} \neq \phi(i_{V_2 \setminus S})$, for every $i_{V_2 \setminus S} \in \mathcal{I}_3$. For every $i_{V_2 \setminus S} \in \mathcal{I}_2$, we pick an arbitrary move $\mathbf{g}^{i_{V_2 \setminus S}} \in \mathcal{H}(D_1, D_2, \ldots, D_r)$. Consider the table $\mathbf{f} = \{f(i)\}_{i \in \mathcal{I}_{V_1 \cup V_2}}$ given by

$$f(i) = f(i_{V_1}, i_{V_2 \setminus S}) := \begin{cases} 0, & \text{if } i_{V_2 \setminus S} \in \mathcal{I}_0, \\ g^{i_{V_2 \setminus S}}(i_{V_1}), & \text{if } i_{V_2 \setminus S} \in \mathcal{I}_2, \\ -g^{\phi(i_{V_2 \setminus S})}(i_{V_1}), & \text{if } i_{V_2 \setminus S} \in \mathcal{I}_3. \end{cases} \tag{20}$$

Let $\mathcal{HC} := \mathcal{HC}(D_1, D_2, \ldots, D_r || V_2 \setminus S)$ the set of moves $\mathbf{f}$ defined as in Eq. 20 corresponding to all possible partitions of $\mathcal{I}_{V_2 \setminus S}$ and all the moves in $\mathcal{H}(D_1, D_2, \ldots, D_r)$. Then $\mathcal{HC}(D_1, D_2, \ldots, D_r || V_2 \setminus S)$ is a set of moves for the class of tables $\mathbf{T}(D_1, D_2, \ldots, D_r, V_2)$.

Proof. Let $\mathbf{f}$ be the move defined in Eq. 20. The marginal $\mathbf{f}_{V_2}$ of $\mathbf{f}$ has entries

$$\begin{aligned}
f_{V_2}(i_{V_2}) &= \sum_{j_{V_1\setminus S}\in\mathcal{I}_{V_1\setminus S}} f(j_{V_1\setminus S}, i_{V_2}) \\
&= \sum_{j_{V_1\setminus S}\in\mathcal{I}_{V_1\setminus S}} f(j_{V_1\setminus S}, i_S, i_{V_2\setminus S}) \\
&= \begin{cases} 0, & \text{if } i_{V_2\setminus S}\in\mathcal{I}_0, \\ \sum\limits_{j_{V_1\setminus S}\in\mathcal{I}_{V_1\setminus S}} g^{i_{V_2\setminus S}}(j_{V_1\setminus S}, i_S) = g_S^{i_{V_2\setminus S}}(i_S), & \text{if } i_{V_2\setminus S}\in\mathcal{I}_2, \\ \sum\limits_{j_{V_1\setminus S}\in\mathcal{I}_{V_1\setminus S}} -g^{\phi(i_{V_2\setminus S})}(j_{V_1\setminus S}, i_S) = -g_S^{\phi(i_{V_2\setminus S})}(i_S), & \text{if } i_{V_2\setminus S}\in\mathcal{I}_3. \end{cases}
\end{aligned}$$

Since $S\subset D_{r_0}$, $\mathbf{g}_{D_{r_0}}=\mathbf{0}$ implies $\mathbf{g}_S=\mathbf{0}$, for any $\mathbf{g}\in\mathcal{H}(D_1,D_2,\ldots,D_r)$. This implies that $\mathbf{f}_{V_2}=\mathbf{0}$. Next we look at the marginal $\mathbf{f}_{V_1}$ of $\mathbf{f}$:

$$\begin{aligned}
f_{V_1}(i_{V_1}) &= \sum_{j_{V_2\setminus S}\in\mathcal{I}_{V_2\setminus S}} f(i_{V_1}, j_{V_2\setminus S}) \\
&= \sum_{j_{V_2\setminus S}\in\mathcal{I}_0} f(i_{V_1}, j_{V_2\setminus S}) + \sum_{j_{V_2\setminus S}\in\mathcal{I}_2} f(i_{V_1}, j_{V_2\setminus S}) + \sum_{j_{V_2\setminus S}\in\mathcal{I}_3} f(i_{V_1}, j_{V_2\setminus S}) \\
&= g^{i^0_{V_2\setminus S}}(i_{V_1}).
\end{aligned}$$

Therefore $\mathbf{f}_{V_1}=\mathbf{g}^{i^0_{V_2\setminus S}}$. We have $D_j\subset V_1$, thus $\mathbf{f}_{D_j}=\mathbf{g}_{D_j}^{i^0_{V_2\setminus S}}$. But $\mathbf{g}_{D_j}^{i^0_{V_2\setminus S}}=\mathbf{0}$, which implies $\mathbf{f}_{D_j}=\mathbf{0}$. It follows that we indeed have $\mathbf{f}\in\mathbf{T}(D_1,D_2,\ldots,D_r,V_2)$. ■

In Proposition 6, if we already included in $\mathcal{HC}$ the moves $\mathbf{f}$ associated with some $\mathbf{g}\in\mathcal{H}(D_1,D_2,\ldots,D_r)$, we will not include in $\mathcal{HC}$ the moves $\mathbf{f}$ associated with the move $-\mathbf{g}$. By constructing the set of compound moves in this manner, we will avoid generating a move $\mathbf{f}\in\mathcal{HC}$ more than once. Theorem 2 shows how to construct a generating set for a reducible graph $\mathcal{G}$ having a proper decomposition $(V_1\setminus S, S, V_2\setminus S)$, provided we are given generating sets for $\mathcal{G}(V_1)$ and $\mathcal{G}(V_2)$.

Theorem 2. *Let $D_1^1,D_2^1,\ldots,D_r^1,D_1^2,D_2^2,\ldots,D_s^2$ be vertex sets of a graph $\mathcal{G}$. Denote $V_1:=\bigcup_{i=1}^r D_i^1$, $V_2:=\bigcup_{j=1}^s D_j^2$ and $S:=V_1\cap V_2$. We assume that (i) $D_1^1,D_2^1,\ldots,D_r^1,D_1^2,D_2^2,\ldots,D_s^2$ are complete in $\mathcal{G}$, (ii) $(V_1\setminus S,S,V_2\setminus S)$ is a decomposition of $\mathcal{G}(V_1\cup V_2)$, and (iii) there exist $r_0\in\{1,\ldots,r\}$ and $s_0\in\{1,\ldots,s\}$ such that $S\subseteq D_{r_0}^1\cap D_{s_0}^2$. Moreover, $\mathcal{H}(D_1^1,D_2^1,\ldots,D_r^1)$ and $\mathcal{H}(D_1^2,D_2^2,\ldots,D_s^2)$ are generating sets for the classes of tables $\mathbf{T}(D_1^1,D_2^1,\ldots,D_r^1)$ and $\mathbf{T}(D_1^2,D_2^2,\ldots,D_s^2)$, respectively. Then*

$$\mathcal{H}:=\mathcal{H}(D_1^1,D_2^1,\ldots,D_r^1||V_2\setminus S)\cup\mathcal{H}(D_1^2,D_2^2,\ldots,D_s^2||V_1\setminus S)\cup\mathcal{F}(V_1,V_2) \tag{21}$$

is a generating set for the class of tables $\mathbf{T}:=\mathbf{T}(D_1^1,D_2^1,\ldots,D_r^1,D_1^2,D_2^2,\ldots,D_s^2)$.

Proof. We want to transform a given table $\mathbf{n}\in\mathbf{T}$ into an arbitrary table $\mathbf{x}\in\mathbf{T}$ using only moves in $\mathcal{H}$. Since $\mathbf{n}_{V_1},\mathbf{x}_{V_1}\in\mathbf{T}(D_1^1,D_2^1,\ldots,D_r^1)$, we can find $\mathbf{g}^1,\ldots,\mathbf{g}^{q_1}\in\mathcal{H}(D_1^1,D_2^1,\ldots,D_r^1)$

such that

$$\mathbf{x}_{V_1} - \mathbf{n}_{V_1} = \sum_{i=1}^{q_1} \mathbf{g}^i.$$

Proposition 5 tells us that $\mathbf{g}^1, \ldots, \mathbf{g}^{q_1}$ translate into a sequence of moves $\mathbf{f}^1, \ldots, \mathbf{f}^{q_1} \in \mathcal{H}(D_1^1, D_2^1, \ldots, D_r^1 || V_2 \setminus S)$. We obtain a table $\mathbf{n}' \in \mathbf{T}^{(\mathbf{n})}(D_1^1, D_2^1, \ldots, D_r^1, V_2)$ given by

$$\mathbf{n}' - \mathbf{n} = \sum_{i=1}^{q_1} \mathbf{f}^i, \tag{22}$$

such that $\mathbf{n}'_{V_1} = \mathbf{x}_{V_1}$. Similarly, since $\mathbf{n}_{V_2}, \mathbf{x}_{V_2} \in \mathbf{T}(D_1^2, D_2^2, \ldots, D_s^2)$, we can find $\mathbf{g}^{q_1+1}, \ldots, \mathbf{g}^{q_2} \in \mathcal{H}(D_1^2, D_2^2, \ldots, D_s^2)$ such that

$$\mathbf{x}_{V_2} - \mathbf{n}_{V_2} = \sum_{i=q_1+1}^{q_2} \mathbf{g}^i.$$

Again, Proposition 5 tells us that $\mathbf{g}^{q_1+1}, \ldots, \mathbf{g}^{q_2}$ translate into a sequence of moves $\mathbf{f}^{q_1+1}, \ldots, \mathbf{f}^{q_2} \in \mathcal{H}(D_1^2, D_2^2, \ldots, D_s^2 || V_1 \setminus S)$. We obtain a table $\mathbf{n}'' \in \mathbf{T}^{(\mathbf{n}')}(V_1, D_1^2, D_2^2, \ldots, D_s^2)$ given by

$$\mathbf{n}'' - \mathbf{n}' = \sum_{i=q_1+1}^{q_2} \mathbf{f}^i, \tag{23}$$

such that $\mathbf{n}''_{V_2} = \mathbf{x}_{V_2}$. We have $\mathbf{n}'' \in \mathbf{T}^{(\mathbf{x})}(V_1, V_2)$, thus there exist $\mathbf{f}^{q_2+1}, \ldots, \mathbf{f}^q \in \mathcal{F}(V_1, V_2)$ such that

$$\mathbf{x} - \mathbf{n}'' = \sum_{i=q_2+1}^{q} \mathbf{f}^i. \tag{24}$$

From Eq. 22, Eq. 23 and Eq. 24, it follows that we can find $\mathbf{f}^1, \mathbf{f}^2, \ldots, \mathbf{f}^q \in \mathcal{H}$ such that

$$\mathbf{x} - \mathbf{n} = \sum_{i=1}^{q} \mathbf{f}^i. \tag{25}$$

Therefore $\mathcal{H}$ is indeed a generating set for the class of tables $\mathbf{T}$. ■

Example 3. Let $\mathbf{n}$ be a $3 \times 3 \times 3 \times 3$ table with fixed marginals $\mathbf{n}_{[1,2]}$, $\mathbf{n}_{[1,3]}$, $\mathbf{n}_{[2,3]}$, $\mathbf{n}_{[2,4]}$, and $\mathbf{n}_{[3,4]}$. The independence graph $\mathcal{G}$ associated with this set of marginals is represented in Fig. 1. As we explained in Example 1, $\mathcal{G}$ is decomposable. However, the corresponding log-linear model is not graphical since the marginal tables associated with the cliques of $\mathcal{G}$, $\mathbf{n}_{[1,2,3]}$ and $\mathbf{n}_{[2,3,4]}$, are not considered fixed. Nevertheless, we could still apply Theorem 2 since $S := \{2,3\}$ is complete in $\mathcal{G}$ and $(\{1\}, S, \{4\})$ is a decomposition of $\mathcal{G}$. Suppose we know the generating set of moves that would leave unchanged the two-way marginals of a $3 \times 3 \times 3$ table. For example, such a generating set is described in Diaconis and Sturmfels (1998). Using the notations we introduced before, in this particular situation we are given the generating sets $\mathcal{H}(\{1,2\},\{2,3\},\{1,3\})$ and $\mathcal{H}(\{2,3\},\{3,4\},\{2,4\})$ for the classes of tables $\mathbf{T}^{(\mathbf{n})}(\{1,2\},\{2,3\},\{1,3\})$ and $\mathbf{T}^{(\mathbf{n})}(\{2,3\},\{3,4\},\{2,4\})$, respectively. From Theorem 2 we learn that

$$\mathcal{H}(\{1,2\},\{2,3\},\{1,3\}||\{4\}) \cup \mathcal{H}(\{2,3\},\{3,4\},\{2,4\}||\{1\}) \cup \mathcal{F}(\{1,2,3\},\{2,3,4\})$$

is a generating set for $\mathbf{T}^{(\mathbf{n})}(\{1,2\},\{2,3\},\{1,3\},\{3,4\},\{2,4\})$. ■

We will not give a proof for the next two results because they follow immediately from Theorem 2.

Corollary 2. *We use the same notations as in Theorem 2. The set of compound moves*

$$\mathcal{HC}(D_1^1, D_2^1, \ldots, D_r^1||V_2 \setminus S) \cup \mathcal{HC}(D_1^2, D_2^2, \ldots, D_s^2||V_1 \setminus S) \cup \mathcal{F}_0(V_1, V_2) \tag{26}$$

is a generating set for the class of tables $\mathbf{T}(D_1^1, D_2^1, \ldots, D_r^1, D_1^2, D_2^2, \ldots, D_s^2)$.

Corollary 3. *We use the same notations as in Theorem 2. If* V_2 *is a clique in* $\mathcal{G}$, *then*

$$\mathcal{H}(D_1^1, D_2^1, \ldots, D_r^1||V_2 \setminus S) \cup \mathcal{F}(V_1, V_2) \tag{27}$$

is a generating set for the class of tables $\mathbf{T}(D_1^1, D_2^1, \ldots, D_r^1, V_2)$, *and so is*

$$\mathcal{HC}(D_1^1, D_2^1, \ldots, D_r^1||V_2 \setminus S) \cup \mathcal{F}_0(V_1, V_2). \tag{28}$$

Example 4. We refer to the situation from Example 3. Suppose we are provided with the marginals $\mathbf{n}_{[1,2]}$, $\mathbf{n}_{[1,3]}$ and $\mathbf{n}_{[2,3,4]}$. The independence graph $\mathcal{G}$ remains the same - see Fig. 1. Corollary 3 tells us that the set of moves

$$\mathcal{H}(\{1,2\},\{2,3\},\{1,3\}||\{4\}) \cup \mathcal{F}(\{1,2,3\},\{2,3,4\})$$

is a generating set for the class of tables $\mathbf{T}^{(\mathbf{n})}(\{1,2\},\{1,3\},\{2,3,4\})$. ■

We are now ready to describe a procedure for determining a generating set associated with the class of tables induced by the minimal sufficient statistics of a reducible loglinear model (Dobra and Fienberg, 2000). Suppose we are provided with a set of marginals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ of a k-way cross-classification $\mathbf{n} = \{n(i)\}_{i \in \mathcal{I}}$. There are no redundant configurations in the previous sequence, that is there are no $r_1, r_2 \in \{1, 2, \ldots, r\}$ such that $D_{r_1} \subset D_{r_2}$. Moreover, the set of marginals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ is consistent, meaning that, for every $r_1, r_2 \in \{1, 2, \ldots, r\}$, the $(D_{r_1} \cap D_{r_2})$-marginal of $\mathbf{n}_{D_{r_1}}$ is equal to the $(D_{r_1} \cap D_{r_2})$-marginal of $\mathbf{n}_{D_{r_2}}$ whenever $D_{r_1} \cap D_{r_2} \neq \emptyset$. The set of fixed marginals define a graph $\mathcal{G} = (V, E)$, where

$$V := \bigcup_{j=1}^{r} D_j \quad \text{and} \quad E := \{(u,v) : \{u,v\} \subset D_j, \text{ for some } j \in \{1, \ldots, r\}\}. \tag{29}$$

Let $\mathcal{V}(\mathcal{G}) := \{V_1, V_2, \ldots, V_q\}$ be the maximal prime subgraphs of $\mathcal{G}$ ordered in a perfect sequence, and let $\mathcal{S}(\mathcal{G}) := \{S_2, \ldots, S_q\}$ be the sequence of separators associated with $V_1, V_2, \ldots, V_q$. With every mp-subgraph V_j, we associate

$$\mathcal{L}(V_j) := \{D_l : D_l \subset V_j\},$$

and let $\{D_1^j, D_2^j, \ldots, D_{r_j}^j\} \subset \{D_1, D_2, \ldots, D_r\}$ be the configurations in $\mathcal{L}(V_j)$. For every $j \in \{1, 2, \ldots, q\}$, we were provided with a generating set

$$\mathcal{H}_j := \mathcal{H}(D_1^j, D_2^j, \ldots, D_{r_j}^j)$$

for the class of tables $\mathbf{T}_j := \mathbf{T}(D_1^j, D_2^j, \ldots, D_{r_j}^j)$. In order to assure the consistency of the generating sets $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_q$, we have to assume that, for every separator $S \in \mathcal{S}(\mathcal{G})$ with $S \subset V_j$, there exists $D \in \mathcal{L}(V_j)$ with $S \subset D$. With these notations, we outline below the algorithm for determining a generating set of moves $\mathcal{H}$ for the set of tables $\mathbf{T} := \mathbf{T}(D_1, D_2, \ldots, D_r)$.

Algorithm 5 (Generating set for a reducible model).

- *Obtain a generating set $\mathcal{H}^{1,2}$ for the class of tables $\mathbf{T}(D : D \in \mathcal{L}(V_1) \cup \mathcal{L}(V_2))$ from the generating sets $\mathcal{H}_1$ and $\mathcal{H}_2$ by making use of Eq. 21.*
- **for** $j = 3, \ldots, q$ **do**
  - *Obtain a generating set $\mathcal{H}^{1,2,\ldots,j}$ for the class of tables $\mathbf{T}(D : D \in \mathcal{L}(V_1) \cup \mathcal{L}(V_1) \cup \ldots \cup \mathcal{L}(V_j))$ from the generating sets $\mathcal{H}^{1,2,\ldots,j-1}$ and $\mathcal{H}_j$ by making use of Eq. 21.*

  end for

■

The resulting set of moves $\mathcal{H}^{1,2,\ldots,q}$ will be a generating set for the class of tables $\mathbf{T} := \mathbf{T}(D_1, D_2, \ldots, D_r)$. Theorem 2 is applicable at every step j of Algorithm 5 because the sequence $V_1, V_2, \ldots, V_q$ is perfect and hence $(H_{j-1} \setminus S_j, S_j, V_j \setminus H_{j-1})$ is a decomposition of $\mathcal{G}(H_j)$, where $H_j := V_1 \cup \ldots \cup V_j$ (Lauritzen, 1996). To avoid including redundant moves in $\mathcal{H}^{1,2,\ldots,q}$ at the j-th step of Algorithm 5, we utilize Eq. 27 instead of Eq. 21 if $r_j = 1$ and D_1^j is a clique in $\mathcal{G}$. Moreover, in order to reduce the correlation between two consecutive tables, we replace Eq. 21 and Eq. 27 with Eq. 26 and Eq. 28, respectively.

It is essential to notice that $D_1, D_2, \ldots, D_r$ are not necessarily cliques in $\mathcal{G}$, i.e. the loglinear model with minimal sufficient statistics $D_1, D_2, \ldots, D_r$ might not be graphical. Two loglinear models might have the same independence graph, but different minimal sufficient statistics if, for example, one of them is graphical and the other one is not. While Algorithm 5 is still applicable in this general framework, we have to be very careful to correctly set up the problem. If V_j is a clique in $\mathcal{G}$ but $r_j > 1$, we should not replace Eq. 21 with Eq. 27 at the j-th step. Similarly, if $\mathcal{G}$ is decomposable, the mp-subgraphs $V_1, V_2, \ldots, V_q$ are the set of cliques of the graph $\mathcal{G}$, hence the set of simple moves $\mathcal{F}(\mathcal{G})$ associated with $\mathcal{G}$ will be a generating set for the class of tables $\mathbf{T}(V_1, V_2, \ldots, V_q)$. Although $\mathcal{F}(\mathcal{G})$ is obviously a set of moves for $\mathbf{T}(D_1, D_2, \ldots, D_r)$, it will not be a generating set for this class of tables. However, if $D_1, D_2, \ldots, D_r$ are the set of cliques of $\mathcal{G}$ and $\mathcal{G}$ is decomposable, then by employing Algorithm 5 we will end up with the set of simple moves $\mathcal{F}(\mathcal{G})$ defined in Eq. 10.

Example 5. Let $\mathcal{G}$ be the graph in Fig. 3 and let $\mathbf{n}$ be an eleven-way table. We make use of the notations we introduced in Example 2. The set of marginal tables we are provided with corresponds to the set of edges of $\mathcal{G}$ from which we take out $\{3, 4\}$, $\{3, 11\}$, $\{4, 11\}$, and then

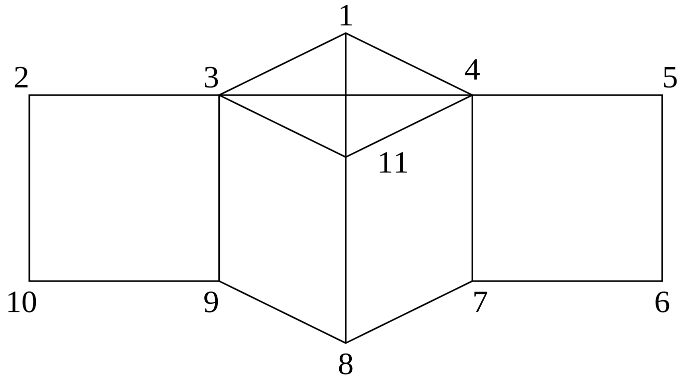

Figure 3: A reducible graph with four mp-subgraphs.

add $\{3, 4, 11\}$. The graph $\mathcal{G}$ is reducible and its maximal prime subgraphs are the cliques of the graph represented in Fig. 2. Therefore $\mathcal{V}(\mathcal{G}) = \{C_1, C_2, C_3, C_4\}$ and, as we explained in Example 2, this is also a perfect ordering of the vertex sets in $\mathcal{V}(\mathcal{G})$.

Assume we somehow managed to compute the generating sets $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$, $\mathcal{H}_4$ associated with the classes of tables $\mathbf{T^{(n)}}(\{1,3\},\{1,11\},\{1,4\},\{3,4,11\})$, $\mathbf{T^{(n)}}(\{3,4,11\},\{3,9\},\{8,11\},\{4,7\},\{8,9\},\{7,8\})$, $\mathbf{T^{(n)}}(\{2,3\},\{3,9\},\{9,10\},\{2,10\})$, and $\mathbf{T^{(n)}}(\{4,5\},\{5,6\},\{6,7\},\{4,7\}))$, respectively. Then Algorithm 5 goes as follows:

- By employing Theorem 2, make use of $\mathcal{H}^1$ and $\mathcal{H}^2$ to obtain a generating set $\mathcal{H}^{1,2}$ for the class of tables $\mathbf{T^{(n)}}(\{1,3\},\{1,11\},\{1,4\},\{3,4,11\},\{3,9\},\{8,11\},\{4,7\},\{8,9\},\{7,8\})$.
- Similarly, use $\mathcal{H}^{1,2}$ and $\mathcal{H}^3$ to determine a generating set $\mathcal{H}^{1,2,3}$ for the class of tables

  $$\mathbf{T^{(n)}}(\{1,3\},\{1,11\},\{1,4\},\{3,4,11\},\{3,9\},\{8,11\},\{4,7\},\{8,9\},\{7,8\},\{2,3\},\{2,10\},\{9,10\}).$$

- Finally, employ Theorem 2 to combine $\mathcal{H}^{1,2,3}$ and $\mathcal{H}^{1,2,3,4}$ into a generating set $\mathcal{H}^{1,2,3,4}$ for the class of tables

  $$\begin{aligned}\mathbf{T^{(n)}}(\{1,3\},\{1,11\},\{1,4\},\{3,4,11\},\{3,9\},\{8,11\},\{4,7\},\{8,9\},\{7,8\},\\ \{2,3\},\{2,10\},\{9,10\},\{4,5\},\{5,6\},\{6,7\}).\end{aligned}$$

■

There is one more remark we have to make. As opposed to the decomposable case, the potential user will have to actually generate an array containing the moves in the generating set obtained from Algorithm 5. Apparently, there does not exist an obvious straightforward solution for dynamically producing moves associated with reducible graphical structures that are not decomposable.

5. Exact Distributions

In this section we will explore additional applications of the generating sets of moves we previously described. In particular, we are interested in making draws from the exact distribution of all contingency tables with a given set of marginals that are the minimal sufficient statistics of a loglinear model. The estimation of a multivariate c.d.f. that satisfies a set of constraints is very appealing in the categorical data framework because the empirical c.d.f. is the contingency table itself (Bishop et al., 1975). Sampling from the exact distribution has obvious benefits, such as the calibration of test statistics. For example, standard asymptotic results tell us to approximate the value of Pearson's chi-squared statistic

$$\chi^2 = \sum_{i \in \mathcal{I}} \frac{[n(i) - \hat{m}(i)]^2}{\hat{m}(i)}.$$

with a chi-squared distribution with the appropriate number of degrees of freedom. Nevertheless, the asymtotic 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 of the correct number of degrees of freedom could be a very difficult task by itself (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). The 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.

In the Poisson sampling scheme, the cell counts $n(i)$ in $\mathbf{n}$ are independent Poisson random variables with mean $m(i) > 0$. Denote $\mu = \log(\mathbf{m})$. Let $D_1, D_2, \ldots, D_r$ be the minimal sufficient statistics of the loglinear model specified by μ. Sampling could be considered conditional on $\mathbf{n} \in \mathbf{T}^{(\mathbf{n})} = \mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$. Since $\mathbf{T}^{(\mathbf{n})}$ has at least one element, namely the original table $\mathbf{n}$, the conditional probabilities will be meaningful. If $\mathbf{n}' \in \mathbf{T}^{(\mathbf{n})}$, then

$$\sum_{i \in \mathcal{I}} n'(i) = \sum_{i_{D_1} \in \mathcal{I}_{D_1}} n_{D_1}(i_{D_1}). \tag{30}$$

The class of tables $\mathbf{T}^{(\mathbf{n})}$ is finite because there are only a finite number of k-way cross-classifications $\mathbf{n}'$ with positive integer cell entries such that Eq. 30 holds. The unconditional probability of $\mathbf{n}$ based on the Poisson model is

$$\begin{aligned} P(\mathbf{n}) &= \prod_{i \in \mathcal{I}} \left[\frac{m(i)^{n(i)}}{n(i)!} e^{-m(i)} \right] \\ &= \frac{1}{\prod_{i \in \mathcal{I}} n(i)!} \cdot e^{\sum_{i \in \mathcal{I}} [n(i)\mu(i) - m(i)]}. \end{aligned} \tag{31}$$

Following Haberman (1974), the conditional probability of $\mathbf{n}$ given the set of marginal constraints

$\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ is

$$\begin{aligned}\sigma(\mathbf{n}) &= \frac{P(\mathbf{n})}{\sum\limits_{\mathbf{n}' \in \mathbf{T}^{(\mathbf{n})}} P(\mathbf{n}')} \\ &= \frac{\left[\prod\limits_{i\in\mathcal{I}} n(i)!\right]^{-1} \cdot e^{\sum\limits_{i\in\mathcal{I}} n(i)\mu(i)}}{\sum\limits_{\mathbf{n}' \in \mathbf{T}^{(\mathbf{n})}} \left[\prod\limits_{i\in\mathcal{I}} n'(i)!\right]^{-1} \cdot e^{\sum\limits_{i\in\mathcal{I}} n'(i)\mu(i)}}. \end{aligned} \tag{32}$$

Under some not too restrictive conditions, Eq. 32 further simplifies to (Fienberg et al., 2000):

$$\sigma(\mathbf{n}) = \frac{\left[\prod\limits_{i\in\mathcal{I}} n(i)!\right]^{-1}}{\sum\limits_{\mathbf{n}' \in \mathbf{T}^{(\mathbf{n})}} \left[\prod\limits_{i\in\mathcal{I}} n'(i)!\right]^{-1}} \tag{33}$$

The denominator in Eq. 33 is a normalizing constant that cannot in general be evaluated. However, if $D_1, D_2, \ldots, D_r$ are the cliques of a decomposable graph arranged in a perfect sequence, the denominator in Eq. 33 will have a closed form expression (Lauritzen, 1996):

$$\sum_{\mathbf{n}' \in \mathbf{T}^{(\mathbf{n})}} \left[\prod_{i\in\mathcal{I}} n'(i)!\right]^{-1} = \frac{\prod\limits_{j=2}^{r} \prod\limits_{i_{S_j} \in \mathcal{I}_{S_j}} n(i_{S_j})!}{\prod\limits_{j=1}^{r} \prod\limits_{i_{D_j} \in \mathcal{I}_{D_j}} n(i_{D_j})!},$$

where $S_j := (D_1 \cup \ldots \cup D_{j-1}) \cap D_j$, for $j = 2, \ldots, r$, is the sequence of separators associated with $D_1, D_2, \ldots, D_r$.

Diaconis and Sturmfels (1998) present a general Markov chain algorithm for generating draws from the conditional distribution $\sigma(\mathbf{n})$. They proved the following result:

Theorem 3. *Let $\sigma(\mathbf{x})$ be a positive function on $\mathbf{T}^{(\mathbf{n})}$. Given a generating set $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_L\}$, generate a Markov chain on $\mathbf{T}^{(\mathbf{n})}$ by choosing l uniformly in $\{1, 2, \ldots, L\}$. If the chain is currently at $\mathbf{x} \in \mathbf{T}^{(\mathbf{n})}$, it moves to $\mathbf{x}' = \mathbf{x} + f_l$ (provided that $\mathbf{x}'(i) \geq 0, \forall i \in \mathcal{I}$) with probability*

$$\min\left\{\frac{\sigma(\mathbf{x}')}{\sigma(\mathbf{x})}, 1\right\}.$$

In all other cases the chain stays at $\mathbf{x}$. This is a connected, reversible, aperiodic Markov chain on $\mathbf{T}^{(\mathbf{n})}$ with stationary distribution proportional to $\sigma(\mathbf{x})$.

We notice that Theorem 3 could still be used even if $\sigma(\mathbf{n})$ is expressed in an intractable form and cannot actually be evaluated. From Eq. 32, it follows that

$$\sigma(\mathbf{n}) \propto \left[\prod_{i\in\mathcal{I}} n(i)!\right]^{-1} \cdot e^{\sum\limits_{i\in\mathcal{I}} n(i)\mu(i)}.$$

If $\mathbf{x}, \mathbf{x}' \in \mathbf{T}^{(\mathbf{n})}$, we have

$$\begin{aligned}
\frac{\sigma(\mathbf{x}')}{\sigma(\mathbf{x})} &= \prod_{i\in\mathcal{I}} \frac{x(i)!}{x'(i)!} e^{[x'(i)\mu(i)-x(i)\mu(i)]} \\
&= \prod_{\{i\in\mathcal{I}:x'(i)\neq x(i)\}} \frac{x(i)!}{x'(i)!} e^{[x'(i)-x(i)]\mu(i)}. \qquad (34)
\end{aligned}$$

Therefore, in order to employ Theorem 3, we only need to know how to calculate the ratio of the conditional densities associated with two tables in $\mathbf{T}^{(\mathbf{n})}$, and, from Eq. 34, we see that this ratio is a function of the cell counts that differ in the two tables.

Theorem 3 essentially tells us how to generate a random walk on a given set of tables $\mathbf{T}^{(\mathbf{n})}$ if we know its generating set. Due to the computational complexity of determining such a generating set of moves, the approach proposed by Diaconis and Sturmfels has been implemented primarily for two- and three-dimensional tables. The results presented in this paper will make possible to implement it for contingency tables of any dimension.

6. Discussion

One could acquire a better understanding of the significance of our findings by envisioning a remote data access system as described in Keller-McNulty and Unger (1998). The users are allowed to sequentially request marginal totals of a confidential database consisting of a large contingency table $\mathbf{n}$. Such a query system should have the capability of accepting or denying the release of some marginal $\mathbf{n}_{D_r}$ given that the marginal tables $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_{r-1}}$ have already been made available to various clients. Clearly, if $D_r \subset D_j$ for some $j \in \{1, 2, \ldots, r-1\}$, we will release $\mathbf{n}_{D_r}$ without any additional testing. Otherwise, we have to measure how vulnerable to disclosure the dataset $\mathbf{n}$ might become if the D_r-marginal will be made public. The problem is further compounded in our context because the system should be designed to be autonomous and should work in real time. To be more explicit, the system should be able to take deterministic decisions on its own without any external (human) intervention, and these decisions should be taken in a reasonable amount of time.

We argued before that the disclosure risk associated with $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_{r-1}}, \mathbf{n}_{D_r}$ could be assessed by examining Eq. 1 and Eq. 2. We could refuse to release the D_r-marginal due to several reasons, such as:

- the number of tables in $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ could drop below a certain threshold;
- the upper and lower bounds on the cell entries in table $\mathbf{n}$ induced by $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ could reveal some cell count $n(i_0)$, $i_0 \in \mathcal{I}$;
- $\mathbf{n}_{D_r}$ could contain counts of "1" or "2" that correspond to small(er) counts in the initial table $\mathbf{n}$.

In order to find a sensible answer to these questions, we should first obtain a picture of the dependency patterns induced by the released marginals by constructing the independence graph $\mathcal{G} = (V, E)$ (see Eq. 29). In the ideal situation when $D_1, D_2, \ldots, D_r$ are the set of cliques of $\mathcal{G}$ and $\mathcal{G}$ is decomposable, the upper and lower bounds in Eq. 2 can be expressed explicitly as linear functions of $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$ (Dobra and Fienberg, 2000) and, in addition, there exists a generating set of simple moves $\mathcal{F}(D_1, D_2, \ldots, D_r)$ for the class of tables $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ (see Eq. 10). Since we know how to dynamically produce moves in $\mathcal{F}(D_1, D_2, \ldots, D_r)$, we could easily generate tables in $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ to estimate the number of elements in this class of tables. Furthermore, it is straightforward to perturb the entries in table $\mathbf{n}$ to eliminate the potentially dangerous counts of "1" or "2" from the D_r-marginal.

It is possible that, while $\mathcal{G}$ is decomposable, some of $D_1, D_2, \ldots, D_r$ are not cliques in $\mathcal{G}$. In that case we will have to determine the set $\mathcal{C}(\mathcal{G})$ of cliques of $\mathcal{G}$. The formulas in Dobra and Fienberg (2000) allow us to quickly find upper and lower bounds for the cell entries in table $\mathbf{n}$ given the marginals $\{\mathbf{n}_C : C \in \mathcal{C}(\mathcal{G})\}$. These bounds will not be sharp for Eq. 2, but based on them we could obtain a reasonable overestimate of the disclosure risk associated with the marginals we made public. While the set of simple moves $\mathcal{F}(C : C \in \mathcal{C}(\mathcal{G}))$ is not a generating set for $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$, it should be obvious that any $\mathbf{f} \in \mathcal{F}(C : C \in \mathcal{C}(\mathcal{G}))$ will leave unchanged the marginals $\mathbf{n}_{D_1}, \mathbf{n}_{D_2}, \ldots, \mathbf{n}_{D_r}$, hence we could still employ the moves in $\mathcal{F}(C : C \in \mathcal{C}(\mathcal{G}))$ to perturb the table $\mathbf{n}$.

In the unfavorable situation when $\mathcal{G}$ is reducible but not decomposable, we could still reduce the computational effort required to determine sharp bounds for Eq. 2 or to produce a generating set for $\mathbf{T}^{(\mathbf{n})}(D_1, D_2, \ldots, D_r)$ by solving the "smaller" problems associated with every maximal prime subgraph of $\mathcal{G}$ and then by combining the solutions to solve the "complete" problem associated with $\mathcal{G}$. To solve the problem associated with an mp-subgraph $\mathcal{G}(V_j)$, $V_j \in \mathcal{V}(\mathcal{G})$, we could employ methods borrowed from linear programming or computational algebra, or we could simply add edges to $\mathcal{G}(V_j)$ to make it reducible or decomposable.

We plan to implement the perturbation algorithms we described in this paper and perhaps to develop a query-based data access system that could dynamically assess the vulnerability of a database in tabular form.

Acknowledgment

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 his advisor Dr. Stephen E. Fienberg for all his support, comments and valuable discussions.

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.

Agresti, A. (1992). "A Survey of Exact Inference for Contingency Tables." *Statistical Science*, 7, 1, 131–177.

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.

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 and New York.

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

— (1998). "Using Algebraic Geometry." In *Graduate Texts in Mathematics*, vol. 185. Springer-Verlag.

Dawid, A. P. and Lauritzen, S. L. (1993). "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models." *The Annals of Statistics*, 21, 3, 1272–1317.

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, 3, 845–874.

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

Diaconis, P. and Sturmfels, B. (1998). "Algebraic Algorithms for Sampling From Conditional Distributions." *The Annals of Statistics*, 26, 1, 363–397.

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*. To appear.

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. IOS Press.

Fienberg, S. E. and Udi, E. M. (1998). "Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data." *Journal of Official Statistics*, 14, 4, 485–502.

Fienberg, S. E., Udi, E. M., Meyer, M. M., and Steele, R. J. (2000). "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., Udi, E. M., and Steele, R. J. (1998). "Disclosure Limitation Using Perturbation and Related Methods for Categorical Data." *Journal of Official Statistics*, 14, 4, 485–502.

Gavril, F. (1977). "Algorithms on Clique Separable Graphs." *Discrete Mathematics*, 19, 159–165.

Guo, S. W. and Thompson, E. A. (1989). "Analysis of Sparse Contingency Tables: Monte Carlo Estimation of Exact p-values." Tech. Rep. 187, University of Washington.

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.

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

Holmes, R. B. and Jones, L. K. (1996). "On Uniform Generation of Two-Way Tables with Fixed Margins and The Conditional Volume Test of Diaconis and Efron." *The Annals of Statitics*, 24, 1, 64–68.

Julian, B. and Clifford, P. (1989). "Generalized Monte Carlo Significance Tests." *Biometrika*, 76, 4, 633–642.

— (1991). "Sequential Monte Carlo p-values." *Biometrika*, 78, 2, 301–304.

Keller-McNulty, S. and Unger, E. A. (1998). "A Database System Prototype for Remote Access to Information Based on Confidential Data." *Journal of Official Statistics*, 14, 4, 347–360.

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

Kreiner, S. (1987). "Analysis of Multidimensional Contingency Tables by Exact Conditional Tests: Techniques and Strategies." *Scandinavian Journal of Statistics*, 14, 97–112.

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

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

Mishra, B. (1993). *Algorithmic Algebra*. Texts and Monographs in Computer Science. Springer-Verlag, New York.

Mount, J. (1995). "Application of Convex Sampling to Optimization and Contingency Table Generation/Counting." Ph.D. thesis, Carnegie Mellon University.

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

Patefield, W. M. (1981). "Algorithm AS 159: An Efficient Method of Generating Random $R \times C$ Tables with Given Row and Column Totals." *Applied Statistics*, 30, 91–97.

Rüschendorf, L. (1991). "Bounds for distributions with multivariate margins." In *Stochastic Orders and Decision under Risk*, eds. K. Mosler and e. M. Scarsini, vol. 19, 285–310. IMS Lecture Notes - Monograph Series.

Schrijver, A. (1986). *Theory of Linear and Integer Programming*. John Wiley & Sons. Oxford.

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." In *University Lecture Series*, vol. 8. American Mathematical Society.

Sturmfels, B., Weismantel, R., and Ziegler, G. M. (1995). "Gröbner Bases of Lattices, Corner Polyhedra, and Integer Programming." *Beitrage zur Algebra und Geometrie*, 36, 2, 291–298.

Tarjan, R. E. (1985). "Decomposition by Clique Separators." *Discrete Mathematics*, 55, 221–232.

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

Ziegler, G. M. (1995). *Lectures on Polytopes*, vol. 152 of *Graduate Texts in Mathematics*. Springer-Verlag. New York.