PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 139, Number 1, January 2011, Pages 361–373
S 0002-9939(2010)10509-9
Article electronically published on July 21, 2010

FINDING THE MINIMAL SET FOR COLLAPSIBLE GRAPHICAL MODELS

XIAOFEI WANG, JIANHUA GUO, AND XUMING HE

(Communicated by Edward C. Waymire)

Abstract. A graphical model is said to be collapsible onto a set of variables if the implied model for the marginal distribution of those variables is the same as that given by the induced subgraph. We discuss the notion of collapsibility under multinomial, Gaussian, and mixed graphical models for undirected graphs, and we show that there exists a unique minimal set of variables onto which a graphical model can be collapsed. We also provide a useful algorithm for finding the minimal set and give examples to illustrate the utility of using collapsibility.

1. Introduction and preliminaries

Graphical models are increasingly valuable in problems of higher dimension and greater complexity. For example, graphical models have been introduced into systems in biology to explore gene expression data and describe gene association networks with thousands of variables (see Dobra et al. [3]; Rich et al. [8]). The use of collapsibility is an effective model reduction method, as it provides a lower dimensional submodel without loss of relevant information.

Graphical models represent conditional independence among variables by graphs; separation in a graph implies the conditional independence between sets of variables, which can be of discrete as well as of continuous types. A good review can be found in Wermuth and Lauritzen [12], in which graphical models are shown to provide a unifying concept for many statistical techniques that have proven to be useful in data analysis.

A given graphical model is collapsible onto a set of variables if the implied model for the marginal distributions for these variables is equal to the graphical model given by the induced subgraph. Collapsibility plays an important role for structural reductions and model selections. Asmussen and Edwards [1] study the collapsibility in log-linear models for contingency tables and provide an equivalent condition in the language of graph theory. Their results can be applied to graphical models for contingency tables or multivariate Gaussian distributions. Frydenberg [4] describes an equivalent condition for collapsibility for mixed graphical models, again based

Received by the editors April 4, 2009 and, in revised form, March 19, 2010.
2000 *Mathematics Subject Classification.* Primary 62-09, 05E05; Secondary 05C85.
Key words and phrases. Collapsibility, decomposition, graphical models.
This research was supported by the National Natural Science Foundation of China (Grants No. 10701022, 10871038, 10828102 and 10926186), the National 973 Key Project of China (2007CB311002), and the U.S. National Science Foundation award DMS-0630950.

on graph properties. The primary goal of the present paper is to find the minimal set of variables onto which the graphical model can be collapsed. In other words, we aim to find the smallest submodel from a graph that contains all the information about variables of interest.

The SAHR algorithm, developed by Madigan and Mosurski [7], can be used to find the minimal variable set for decomposable graphical models in contingency tables, and for decomposable Gaussian graphical models, but not for more general undirected graphs such as mixed graphical models with both discrete and continuous variables. The algorithm proposed in this paper applies to general undirected graphical models, as shown in Sections 3 and 5 below.

We now provide some technical terms for easy reference. Readers who are familiar with graphical models may proceed directly to Section 2.

Let $G = (V, E)$ be a simple, undirected graph with vertices V and edges E. For $G = (V, E)$ and $x, y \in V$, x, y are adjacent (denoted by $xy \in E$) if there is an edge between them in G. Given a subset $B \subseteq V$ of the vertices, we define the subgraph induced by B to be $G(B) = (B, E(B))$, where $E(B) = \{xy \in E | x, y \in B\}$. The boundary of $B \subseteq V$ in $G = (V, E)$ is denoted by $\partial_G(B) = \{x \in V \setminus B | \exists\, y \in B$, such that $xy \in E\}$. A subset B of V is complete if any two different vertices $x, y \in B$ are adjacent in G. A vertex x is simplicial if $\partial_G(x)$ is complete.

We define a set of distinct vertices $[x_0, x_1, \cdots, x_k]$ as a path L of length k in G between x_0 and x_k if $x_{i-1}x_i \in E$ for all $i = 1, \cdots, k$, and we call x_i the interior of L for $i = 1, \cdots, k-1$ and call x_i the endvertex of L for $i = 0, k$. For a graph $G = (V, E)$ and $C \subseteq V$, C is said to be a connected set if there is a path between x and y in $G(C)$ for any $x, y \in C$, and C is a connected component if C is a maximal connected set. A subset S is called a separator for $A, B \subseteq V$ in $G = (V, E)$ if there are two nonempty sets $A, B \subseteq V \setminus S$ such that every path in G between $\alpha \in A$ and $\beta \in B$ contains a vertex in S. We also say that A and B are separated by S in G.

A distribution P is said to be Markov for an undirected graph $G = (V, E)$ if A is conditionally independent of B given S under P whenever A and B are separated by S in G. The graphical models $\mathcal{P}_G$ are the set of all the Markov distributions for $G = (V, E)$. Let Δ be the set of all the discrete variables, and let Γ be the set of all the continuous variables. If $\mathcal{P}_G$ is a class of multinomial distributions and is a graphical model for $G = (\Delta, E)$, then we call it a multinomial graphical model. If $\mathcal{P}_G$ is a class of multivariate Gaussian distributions and is a graphical model for $G = (\Gamma, E)$, then we call it a Gaussian graphical model. If $\mathcal{P}_G$ is a class of conditional Gaussian distributions and is a graphical model for $G = (\Delta \cup \Gamma, E)$, then we call it a mixed graphical model.

A graphical model $\mathcal{P}_G$ for $G = (V, E)$ is said to be collapsible onto $B \subseteq V$ (or $G(B)$) if $P(x_B) \in \mathcal{P}_{G(B)}$ for any $P(x) \in \mathcal{P}_G$, which is equivalent to the consistency between the marginality $\hat{P}(x_B)$ of the maximum-likelihood estimate on the whole model for G and the maximum-likelihood estimate $\hat{P}_B(x_B)$ on the marginal model for $G(B)$ (Asmussen and Edwards [1]; Frydenberg [4]). The following proposition, as an important illustration on collapsibility, can be deduced from Theorem 2·3 of Asmussen and Edwards [1] for multinomial graphical models and from Frydenberg [4] for Gaussian graphical models. These authors have built a bridge between statistics and graph theory.

Proposition 1.1 ([1], [4])**.** *Let $\mathcal{P}_G$ be a multinomial or Gaussian graphical model for an undirected graph $G = (V, E)$, where $V = \Delta$ or Γ. For a subset $B \subseteq V$, $\mathcal{P}_G$ is*

collapsible onto $G(B)$ if and only if $\partial_G(C)$ is complete for any connected component C of $G(V \setminus B)$.

For an undirected graph $G = (V, E)$ and $D \subseteq V$, we say that D can be collapsed over in G if $\partial_G(C)$ is complete for any connected component C of $G(D)$. In this case, we also say that G is collapsible onto $V \setminus D$. Let $\mathcal{P}_G$ be a multinomial or Gaussian graphical model for $G = (V, E)$, where $V = \Delta$ or Γ. From Proposition 2.1, D can be collapsed over in G if and only if $\mathcal{P}_G$ is collapsible onto $G(V \setminus D)$ if and only if G is collapsible onto $V \setminus D$.

2. More on collapsibility for multinomial and Gaussian graphical models

A chord of a path is an edge joining two nonconsecutive vertices on the path. A path is called an induced path if there is no chord of this path. A subgraph H of G is called convex in G if H contains every induced path in G whose two endvertices are in H. Any convex subgraph of G is an induced subgraph. For an induced graph H of $G = (V, E)$, we denote the vertex set of H as $V(H)$. For $X, Y \subseteq V$, a path L is called an $X - Y$ path if its two endvertices are in X and Y, respectively, and its interiors lie in $V \setminus (X \cup Y)$. The following Proposition 2.1 and Corollaries 2.2 and 2.3 can actually be found in Diestel (1990, Chapter 1.0) in graph theory languages, so we forego the proofs here.

Proposition 2.1 ([2])**.** *The following statements are equivalent for an undirected graph $G = (V, E)$ and an induced subgraph H of G.*

(i) *This induced graph H is a convex subgraph of G.*

(ii) *If $x, y \in V(H)$ are the two endvertices of a $V(H) - V(H)$ path, they are adjacent.*

(iii) *The multinomial or Gaussian graphical model $\mathcal{P}_G$ is collapsible onto H.*

(iv) *If $T \subseteq V(H)$ and $U, W \subseteq V(H) \setminus T$, then T separates U from W in H if and only if T separates U from W in G.*

In particular, the proposition shows the equivalence between the collapsibility of graphical models and the convex subgraphs. Furthermore, the inheritance of convex subgraphs is illustrated by the following corollary.

Corollary 2.2 ([2])**.** *If H is a convex subgraph of G and H' is a subgraph of H, then H' is a convex subgraph of H if and only if H' is a convex subgraph of G.*

Assume $\mathcal{P}_G$ is a multinomial or Gaussian graphical model for $G = (V, E)$. Let A, B be two subsets of V and $A \subseteq B$. Proposition 2.1 and Corollary 2.2 imply that if $V \setminus B$ can be collapsed over in G, then $B \setminus A$ can be collapsed over in $G(B)$ if and only if $\mathcal{P}_G$ is collapsible onto $G(A)$.

Corollary 2.3 ([2])**.** *If Λ is an index set and H_λ is a convex subgraph of $G = (V, E)$ for any $\lambda \in \Lambda$, then $G(\bigcap_{\lambda \in \Lambda} V(H_\lambda))$ is a convex subgraph of G.*

Madigan and Mosurski [7] showed that if $\mathcal{P}_G$ is collapsible onto both $G(V(H_1))$ and $G(V(H_2))$, then $\mathcal{P}_G$ is collapsible onto $G(V(H_1) \cap V(H_2))$, which is only a special case of Corollary 2.3. We now state our main result of this section.

Theorem 2.4. *Given an undirected graph $G = (V, E)$ and any subset $A \subseteq V$ of interest, there exists a unique set B, with $A \subseteq B \subseteq V$, such that (i) the multinomial*

or Gaussian graphical model $\mathcal{P}_G$ is collapsible onto $G(B)$, and (ii) for any set N with $A \subseteq N \subseteq V$, where $\mathcal{P}_G$ is collapsible onto $G(N)$, we have $B \subseteq N$. Furthermore, $G(B)$ is the minimal convex subgraph in G such that $A \subseteq B$.

Proof. We choose all the convex subgraphs of G whose vertex set contains A. The intersection graph $G(B)$ of all the convex subgraphs is also a convex subgraph of G from Corollary 2.3 and $B \supseteq A$. Thus the multinomial or Gaussian graphical model $\mathcal{P}_G$ is collapsible onto $G(B)$ by Proposition 2.1, and every convex subgraph of G containing A must contain this intersection. If $\mathcal{P}_G$ is collapsible onto N and $A \subseteq N \subseteq V$, then $G(N)$ is a convex subgraph of G by Proposition 2.1, and thus we have $B \subseteq N$. □

3. An algorithm for finding the minimal set by graph decomposition

If (A, B, S) is a partition of V with A and B nonempty and S is a complete separator for A, B in $G = (V, E)$, we call this partition a decomposition for G and S a decomposer for G. A graph G is decomposable if G is a complete graph or there is a decomposition (A, B, S) for G such that $G(A \cup S)$ and $G(B \cup S)$ are decomposable. A graph is prime if its vertex set does not contain any decomposer. A prime block is a maximal prime subgraph of G with respect to inclusion. For convenience, we call $U \subseteq V$ a prime block if $G(U)$ is a prime block of $G = (V, E)$.

To illustrate, the graph in Figure 1 has five prime blocks which are $U_1 = \{a, b, m\}, U_2 = \{b, c, i, m\}, U_3 = \{i, j, k, l, m\}, U_4 = \{c, d, e, f, h, i\}, U_5 = \{f, g\}$, and $\{b, m\}, \{i, m\}, \{c, i\}, \{f\}$ are four decomposers for G.

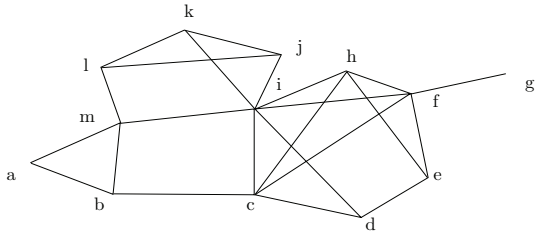

Figure 1. Connected graph with thirteen vertices and five prime blocks.

Leimer (1993) described an O(nm)-time algorithm, which is a modification of Tarjan (1985), to find the unique set $\mathcal{U}_G$ of all the prime blocks in any undirected graph G with n vertices and m edges. We denote Leimer's algorithm as LT(G), whose input is G and the output is $\mathcal{U}_G$. Then our algorithm to find the minimal set B is outlined as follows:

Algorithm Minimal_ Set_ Finder
Input: An undirected graph $G = (V, E)$ and a subset $A \subseteq V$ of interest.
Output: The minimal set B with $A \subseteq B$ onto which the multinomial or Gaussian graphical model for G is collapsible.
Begin
//Call Leimer's algorithm LT(G)
$\mathcal{U}_G \leftarrow LT$(G);

$\mathcal{U} \leftarrow \mathcal{U}_G$; $\mathcal{B} \leftarrow \emptyset$; $m \leftarrow |\mathcal{U}_G|$;
// Pruning procedure
for $U \in \mathcal{U}$ **do**
$K_U \leftarrow U \bigcap (\bigcup_{U' \in \mathcal{U} \setminus \{U\}} U')$;
end for
while There exists a U in $\mathcal{U}$ satisfying the following condition $(C0)$:
$(U \setminus K_U) \bigcap A = \emptyset$ and there exists a U' in $\mathcal{U} \setminus \{U\}$ such that $K_U \subseteq U \bigcap U'$ **do**
$\mathcal{U} \leftarrow \mathcal{U} \setminus \{U\}$; $m \leftarrow m - 1$;
for $U \in \mathcal{U}$ **do**
$K_U \leftarrow U \bigcap (\bigcup_{U' \in \mathcal{U} \setminus \{U\}} U')$;
end for
end
// Slimming procedure
for t = m **to** 1 **step** -1 **do**
Choose an unlabeled prime block $U \in \mathcal{U}$ and label it;
$A_U \leftarrow A \bigcap U$;
if U satisfies the following condition $(C1)$:
$A_U \subseteq K_U$ and K_U is complete
then $\mathcal{B} \leftarrow \{K_U\} \bigcup \mathcal{B}$;
else if U satisfies the following condition $(C2)$:
$A_U \subseteq K_U$ and K_U is not complete
then $\mathcal{B} \leftarrow \{U\} \bigcup \mathcal{B}$;
else if U satisfies the following condition $(C3)$:
$A_U \not\subseteq K_U$ and $K_U \bigcup A_U$ is complete
then $\mathcal{B} \leftarrow \{K_U \bigcup A_U\} \bigcup \mathcal{B}$;
else $\mathcal{B} \leftarrow \{U\} \bigcup \mathcal{B}$;
end for
$B \leftarrow \bigcup_{U' \in \mathcal{B}} U'$;
end

We prove the validity of our algorithm later in the paper. We note here that the motivation of the proposed algorithm is the relationship between decomposition and collapsibility. Actually, for our set of interest A, if a prime block U indicates some decomposition $(U \setminus S, V \setminus U, S)$ for G and $(U \setminus S) \cap A = \emptyset$, then $U \setminus S$ should be collapsed over in G. The decomposition property of $U \setminus S$, where U and S can be found by checking conditions $(C0), (C1)$ and $(C3)$ in the proposed algorithm, is similar to that of the simplicial vertex in the SHAR algorithm of Madigan and Mosurski [7].

In the first part of our algorithm, the prime block U with condition $(C0)$ is always a leaf of some junction tree of $\mathcal{U}$ (Wang and Guo [11]), and $U \setminus K_U$ can be collapsed over. Since K_U is contained in some $U' \in \mathcal{U} \setminus \{U\}$, U can be deleted from $\mathcal{U}$, and we call this the *pruning procedure*. In the second part of our algorithm, if a prime block U satisfies the condition $(C1)$ or $(C3)$, $U \setminus K_U$ or $U \setminus (K_U \cup A_U)$ can be accordingly collapsed over. Since this part only changes the size of prime blocks, we call it

the *slimming procedure*. In the pruning procedure, junction trees can be utilized to reduce the computational complexity in finding prime blocks with $(C0)$ (Wang and Guo [11]), and the structure of $\mathcal{U}$ may be changed because of the elimination of prime blocks with the condition $(C0)$. But after the pruning procedure, the structure of the remaining sets stays invariant in the slimming procedure, and any element of $\mathcal{B}$ after the slimming procedure is a prime block in $G(B)$. Thus, a parallel algorithm can be applied to collapsing operations on each prime block simultaneously, which can efficiently reduce computation time.

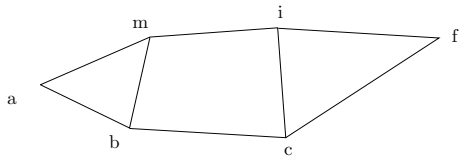

FIGURE 2. Subgraph induced by the minimal set B containing $A = \{a, c, f\}$ of interest which $\mathcal{P}_G$ is collapsible onto.

Let us consider the graph G in Figure 1 again. If $A = \{a, c, f\}$, then we can collapse $\mathcal{P}_G$ onto $G(\{a, b, c, f, i, m\})$ in Figure 2, because $\{j, k, l\}$ and $\{g\}$ are collapsed over in the pruning procedure and $\{d, e, h\}$ is collapsed over in the slimming procedure in our algorithm. Any element of $\mathcal{B} = \{\{a, b, m\}, \{b, c, i, m\}, \{c, f, i\}\}$ is a prime block in $G(B)$. Any statistical analysis on A in the original model with thirteen variables can be carried out in a smaller model with six variables. In the following section, an example for gene association networks is given to further illustrate that this algorithm works efficiently for dimension reduction.

4. An example on gene association networks

Graphical models are frequently used to describe gene association networks and to detect conditionally dependent genes. They provide convenient statistical models for complicated interaction patterns among genes due to biochemical interactions and other regulatory activities.

Figure 3, consisting of 96 genes, was obtained by Schäfer and Strimmer [9] from a global graph with 3883 genes reconstructed from the breast cancer data of West et al. [13] under Gaussian graphical models. We assume that this gene graph represents the covariance matrix of 96 variables in a marginal Gaussian graphical model.

We now focus on three genes: $ESR2$, known to be associated with increased risk for breast cancer, $LAF4$, responsible for lymphocyte differentiation, and $SSX2$. Inferences on these three genes can be considered under a much smaller Gaussian graphical model over the subgraph in Figure 4 induced by genes $ESR2$, $ELK3$, $MLL3$, $LAF4$, $OR3A3$, $MUC3A$, $REG1B$, $SSX2$ and $CD3E$ from our MINIMAL_SET_FINDER algorithm. Thus the original model with 96-dimensional space is collapsed into a submodel in a 9-dimensional space, which reduces the computational complexity in further research. It took us 0.5 seconds on an Intel Pentium 4 PC computer to find the minimal set, i.e., the subgraph shown in Figure 4.

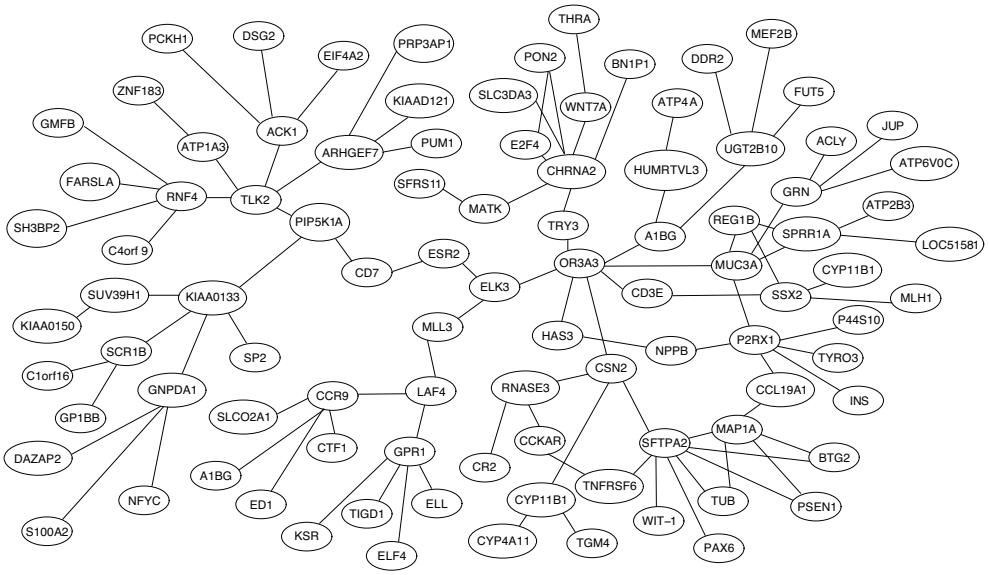

FIGURE 3. Graph consisting of 96 genes centered around the $ERS2$ gene.

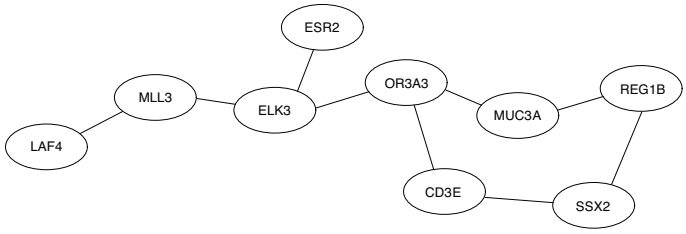

FIGURE 4. Subgraph induced by the minimal set containing genes of interest after collapsing.

5. Collapsibility for mixed graphical models

Mixed graphical models are represented by graphs with two types of vertices, which are often denoted by dots for discrete or qualitative variables and circles for continuous or quantitative variables. Let $G = (V, E)$ be an undirected graph with $V = \Gamma \cup \Delta$, where the vertices in Γ are denoted by circles, and the vertices in Δ are denoted by dots. Figure 5 shows a graphical representation of an emission problem considered in Lauritzen [5].

Definition. If $V = B \cup D$ and $B \cap D = \emptyset$, we say that D can be m-collapsed over in G if (i) $\partial_G(C)$ is complete, and (ii) either $C \subseteq \Gamma$ or $\partial_G(C) \subseteq \Delta$ for any connected component C of $G(D)$.

The notion of m-collapsibility is introduced here for mixed graphical models. Under the above definition, we also say that G can be m-collapsed onto B.

From Frydenberg [4], it is known that for a subset $B \subseteq V$, a mixed graphical model $\mathcal{P}_G$ for $G = (V = \Gamma \cup \Delta, E)$ is collapsible onto $G(B)$ if and only if $\partial_G(C)$

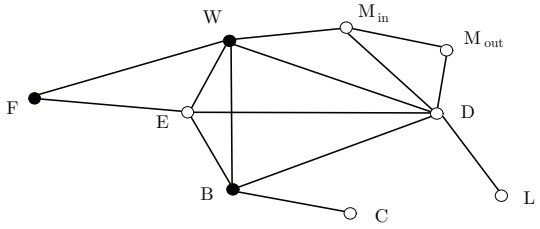

FIGURE 5. The variables, Filter State (F), Waste Type (W) and Burning Regimen (B), corresponding to filled circles are conceived as qualitative variables with states {intact, defect}, {industrial, household}, and {stable, unstable}, respectively. The remaining variables are measured on a quantitative scale: Metals in Waste (M_{in}), Metals Emission (M_{out}), Filter Efficiency (E), Dust Emission (D), CO_2 Concentration in Emission (C), and Light Penetrability (L).

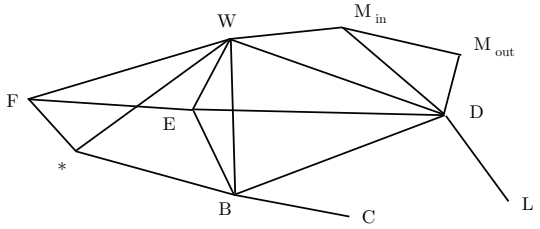

FIGURE 6. Star graph constructed from Figure 5.

is complete and either $C \subseteq \Gamma$ or $\partial_G(C) \subseteq \Delta$ for any connected component C of $G(V \setminus B)$. Thus G can be m-collapsed onto B if and only if the mixed graphical model $\mathcal{P}_G$ is collapsible onto $G(B)$.

For a graph $G = (V = \Gamma \cup \Delta, E)$, we can construct a star graph $G^\star$ from it for further consideration. We add $\star$ into its vertex set and connect this $\star$ with every discrete variable in Δ, and denote the final graph as $G^\star = (V \cup \{\star\}, E^\star)$, where $E^\star = \{(\delta, \star) | \delta \in \Delta\} \cup E$. Figure 6, in which $\star$ is connected with all the discrete variables, is the star graph constructed from Figure 5. The following theorem characterizes a relationship on collapsibility between G and $G^\star$.

Theorem 5.1. *For a graph $G = (V = \Gamma \cup \Delta, E)$ and a subset $D \subseteq V$, D can be m-collapsed over in G if and only if D can be collapsed over in $G^\star$.*

Proof. By the definition of $G^\star$, the following two statements are equivalent:

(1) $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are all the connected components of D in G with $X_i \cap \Delta \neq \emptyset$, $Y_j \subseteq \Gamma$, for $1 \leq i \leq m$, $1 \leq j \leq n$;

(2) $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are all the connected components of D in $G^\star$ with $X_i \cap \Delta \neq \emptyset$, $Y_j \subseteq \Gamma$ for $1 \leq i \leq m$, $1 \leq j \leq n$.

Since $\partial_G Y_i = \partial_{G^\star} Y_i$, then $\partial_G Y_i$ is complete in G if and only if $\partial_{G^\star} Y_i$ is complete in $G^\star$. Because $\partial_G X_i \cup \{\star\} = \partial_{G^\star} X_i$, $\partial_G X_i$ is complete in G and $\partial_G X_i \subseteq \Delta$ if and only if $\partial_{G^\star} X_i$ is complete in $G^\star$. Therefore D can be m-collapsed over in G if and only if D can be collapsed over in $G^\star$. □

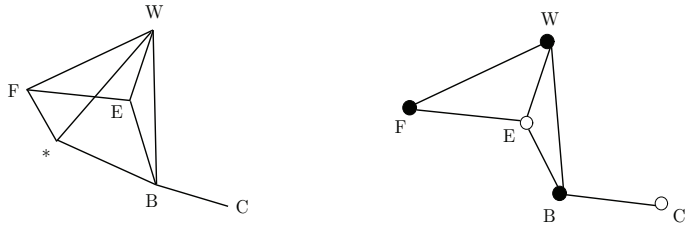

FIGURE 7. Minimal convex subgraph containing $\{\star, W, C\}$ of $G^\star$ and minimal subgraph containing $\{W, C\}$ of G onto which the mixed graphical model is collapsible.

From Proposition 2.1 and Theorem 5.1, we know that G can be m-collapsed onto B if and only if $G^\star(B\cup\{\star\})$ is convex in $G^\star$. We now have the following two results:

Theorem 5.2. *Given an undirected graph $G = (V = \Gamma \cup \Delta, E)$ and $B \subseteq V$, the mixed graphical model $\mathcal{P}_G$ is collapsible onto $G(B)$ if and only if G can be m-collapsed onto B or, equivalently, $G^\star(B \cup \{\star\})$ is convex in $G^\star$.*

Theorem 5.3. *Given an undirected graph $G = (V = \Gamma \cup \Delta, E)$ and any subset $A \subseteq V$ of interest, there exists a minimal set B with $A \subseteq B \subseteq V$, such that (i) the mixed model $\mathcal{P}_G$ is collapsible onto $G(B)$, and (ii) for any set N with $A \subseteq N \subseteq V$, where $\mathcal{P}_G$ is collapsible onto $G(N)$, we have $B \subseteq N$. Furthermore, $G^\star(B \cup \{\star\})$ is the minimal convex subgraph in $G^\star$ such that $A \subseteq B$.*

If G can be m-collapsed onto B, then $G^\star(B\cup\{\star\})$ is convex in $G^\star$. Furthermore, we have the following two conclusions. First, if $G(B)$ can be m-collapsed onto A and $A \subseteq B$, then $G^\star(A\cup\{\star\})$ is convex in $(G(B))^\star$. Because $(G(B))^\star = (G^\star)(B \cup \{\star\})$, $G^\star(A\cup\{\star\})$ is also convex in $G^\star$. Thus G can be m-collapsed onto A. Second, if G can be m-collapsed onto A, then $G^\star(A\cup\{\star\})$ is convex in $G^\star$. Thus $G^\star((A\cap B)\cup\{\star\})$ is also convex in $G^\star$, and then G can be m-collapsed onto $A \cap B$.

The Minimal_ Set_ Finder algorithm of Section 3 can actually give the minimal convex subgraph containing variables of interest, and thus it can also be used to find the minimal set onto which a mixed graphical model is collapsible. Indeed, if we replace G and A by $G^\star$ and $A \cup \{\star\}$, respectively, in the input of the Minimal_ Set_ Finder algorithm, the output $B\cup\{\star\}$ contains B as the minimal set containing A such that the mixed graphical model $\mathcal{P}_G$ is collapsible onto B.

To illustrate the use of collapsibility in mixed graphical models, let us consider the example in Figure 5 again. If we are concerned with the waste type (W) and the CO_2 concentration in mission (C), then we use $A = \{\star, W, C\}$ and $G^\star$ as input to obtain the minimal convex subgraph in the left panel of Figure 7. The right panel of Figure 7 is the subgraph induced by the minimal set containing $\{W, C\}$, onto which the whole mixed graphical model is collapsible. Inference on the relationship between the waste type and the CO_2 concentration in mission can then be carried out on the subgraph induced by $\{B, C, E, F, W\}$.

6. Validity of the minimal_set_finder algorithm

In this section, we first provide some simple lemmas for easy reference, and then prove the validity of the Minimal_Set_Finder Algorithm in steps.

For a set class $\mathcal{U}$, an ordered permutation $\{U_1, \cdots, U_m\}$ of $\mathcal{U}$ is D-ordered if for any $2 \leq t \leq m$, there exists some $q < t$ such that $S_t := (\bigcup_{k=1}^{t-1} U_k) \cap U_t$ is contained in U_q with $S_1 = \emptyset$. This D-ordered permutation is denoted as $\mathcal{U}^o$, and $\{S_k\}_{1 \leq k \leq m}$ is called as the S-system of $\mathcal{U}^o$. For an undirected graph G, let $\mathcal{U}_G$ be the set of all the prime blocks in G. From Theorem 2.5 of Leimer [6], there always exists a D-ordered permutation $\mathcal{U}_G^o$ of $\mathcal{U}_G$.

Given a set class $\mathcal{U}$ and $U \in \mathcal{U}$, let $A_U := A \cap U$, $K_U := U \cap (\bigcup_{U' \in \mathcal{U} \setminus \{U\}} U')$ and $\mathcal{U}_l := \{U \in \mathcal{U} | \exists U' \in \mathcal{U} \setminus \{U\}$ such that $K_U \subseteq U \cap U'\}$. If $\mathcal{U}^o = \{U_t\}_{1 \leq t \leq m}$ is a D-ordered permutation of $\mathcal{U}$, we define $\tau(a) := \min_{1 \leq q < a} \{q | S_a \subset U_q\}$ and $\Lambda_t := \{a | \tau(a) = t\}$ for $1 \leq t \leq m$, and we call $\{\Lambda_t\}_{1 \leq t \leq m}$ the C-system of $\mathcal{U}^o$. For any $1 \leq t \leq m$, we have

$$K_{U_t} = U_t \cap (\bigcup_{s \neq t} U_s) = (\bigcup_{a \in \Lambda_t} S_a) \cup S_t.$$

Lemma 6.1. *If U is a prime block of G and $S \subset U$ is complete, then $G(U \setminus S)$ is connected in $G(U)$ and $\partial_G(U \setminus S) \cap U = S$.*

Proof. This follows from the prime property. □

Lemma 6.2. *If $\mathcal{U}_G^o = \{U_k\}_{1 \leq k \leq n}$, then $G(\bigcup_{k=1}^{l} U_k)$ $(1 \leq l \leq n)$ is convex in G.*

Proof. We prove this lemma by induction. When $l = n$, the result is obvious. Suppose that the result is true for $l = s$. Then we prove that it is also true for $l = s - 1$. By Corollary 2.2, it suffices to prove that $G(\bigcup_{k=0}^{s-1} U_k)$ is convex in $G(\bigcup_{k=0}^{s} U_k)$. Actually, $U_s \setminus S_s$ is connected in U_s from Lemma 6.1. Then $G(\bigcup_{k=0}^{s-1} U_k)$ is convex in $G(\bigcup_{k=0}^{s} U_k)$ from Proposition 2.1 (ii). □

Lemma 6.3. *If $\mathcal{U}_G^o = \{U_1, \cdots, U_t = U, \cdots, U_n\}$ is a D-ordered permutation of $\mathcal{U}_G$, then for any connected set $C \subseteq U$, we have:*

(i) *If $C \cap S_{a_1} \neq \emptyset, C \cap S_{a_2} \neq \emptyset$ and $a_2 > a_1$ for $a_1, a_2 \in \{t\} \cup \Lambda_t$, then $\partial_G(C)$ is not complete.*

(ii) *If there is some $a \in \{t\} \cup \Lambda_t$ such that $C \cap S_a \neq \emptyset$ and $\partial_G(C) \cap U \not\subseteq S_a$, then $\partial_G(C)$ is not complete.*

Proof. (i) If $C \cap S_{a_1} \neq \emptyset, C \cap S_{a_2} \neq \emptyset, a_2 > a_1$, then $U_{a_2} \setminus S_{a_2}$ and $(\bigcup_{k=1}^{a_2 - 1} U_k) \setminus S_{a_2}$ are separated by S_{a_2} in $G(\bigcup_{k=1}^{a_2} U_k)$, and thus in G by Lemma 6.2 and Proposition 2.1. The conclusion then follows from Lemma 6.1.

The proof for (ii) is similar. □

We now prove the validity of our algorithm in steps.

Step 1: Validity of the pruning procedure. In the pruning procedure, if U is the first prime block satisfying the condition $(C0)$, then $U \setminus K_U$ and $V \setminus U$ are separated by K_U in G. Because K_U is complete, Lemma 6.1 and Proposition 1.1 imply that $U \setminus K_U$ can be collapsed over in G. By Corollaries 2.2 and 2.3, we can keep performing the collapsing operation in each step of the pruning procedure while condition $(C0)$ is true.

Step 2: Properties of prime blocks after pruning procedure. A subset $\mathcal{U}$ of $\mathcal{U}_G$ is produced after the execution of the pruning procedure. From Theorem 4.1 of Wang and Guo [11], there is a D-ordered permutation $\{U_1, \cdots, U_m\}$ of $\mathcal{U}$. If we denote the S-system and C-system of $\{U_1, \cdots, U_m\}$ as $\{S_t\}_{1 \leq t \leq m}$ and $\{\Lambda_t\}_{1 \leq t \leq m}$,

respectively, then for any $1 \le t \le m$, we have the following properties, to be used in Step 4 of the proof.

1. If $U_t \in \mathcal{U}_l$, then $A_{U_t} \not\subseteq K_{U_t}$.

2. If $U_t \notin \mathcal{U}_l$, then K_{U_t} contains at least two maximal intersections S_{a_1}, S_{a_2} $(a_1, a_2 \in \{t\} \cup \Lambda_t)$ with regard to inclusion.

Step 3: Resulting set class from the slimming procedure. After the slimming procedure, U_t turns into B_t for $1 \le t \le m$, where $B_t = K_{U_t}$ if B_t satisfies $(C1)$, $B_t = K_{U_t} \cup A_{U_t}$ if B_t satisfies $(C3)$, and $B_t = U_t$ if B_t satisfies neither $(C1)$ nor $(C3)$. Let $\mathcal{B}$ be the set of all B_t for $1 \le t \le m$, and let B be the union of all B_t for $1 \le t \le m$. Since any vertex in K_{U_t} for any $1 \le t \le m$ given $\mathcal{U}$ is not deleted in the slimming procedure, K_{B_t} given $\mathcal{B}$ is the same as K_{U_t} given $\mathcal{U}$ for $1 \le t \le m$. Note that $\{B_1, \cdots, B_m\}$ is a D-ordered permutation of $\mathcal{B}$, and $B_t \in \mathcal{B}_l$ if and only if $U_t \in \mathcal{U}_l$ for $1 \le t \le m$. The S-system and C-system of $\{U_1, \cdots, U_m\}$ are the same as those of $\{B_1, \cdots, B_m\}$; we also denote them as $\{S_t\}_{1 \le t \le m}$ and $\{\Lambda_t\}_{1 \le t \le m}$, respectively. For any $1 \le t \le m$, if $B_t \in \mathcal{B}_l$, then $A_{B_t} \not\subseteq K_{B_t}$; otherwise, K_{B_t} contains at least two maximal intersections S_{a_1}, S_{a_2} $(a_1, a_2 \in \{t\} \cup \Lambda_t)$. Since $G(B_t)$ is prime for $1 \le t \le m$, $\mathcal{B}$ is just the set of all the prime blocks of $G(B)$ by Theorem 2.10 of Leimer [6].

Step 4: Validity of the slimming procedure. We denote as $(C4)$ the complement of $(C1)$, $(C2)$ and $(C3)$ given in the algorithm. First we consider the simplest case where there is only one prime block U in $\mathcal{U}$ after the pruning procedure. In this case we can do the collapsing operation only on U. Because $K_U = \emptyset$, U satisfies condition $(C3)$ or $(C4)$. If U satisfies $(C3)$, $U \setminus A$ can be collapsed over by Lemma 6.1, and $B = A$. If U satisfies $(C4)$, no connected set of $U \setminus A$ can be collapsed over; otherwise, there would be a decomposition for U contrary to the prime property of U. In this case $B = U$.

Next we consider the case when there are at least two prime blocks in $\mathcal{U}$ after the pruning procedure. In this case, any of the four conditions $(C1)$–$(C4)$ is possible. Let $\{U_1, \cdots, U_m\}$ be a D-ordered permutation of $\mathcal{U}$ and $G' = G(\bigcup_{t=1}^{m} U_t)$. We will go through each condition below.

(i) If U_t satisfies condition $(C1)$, then $U_t \setminus K_{U_t}$, $(\bigcup_{k \ne t} U_k) \setminus U_t$ are separated by K_{U_t} in G'. Thus $\partial_{G'}(U \setminus K_{U_t}) = K_{U_t}$, and by Lemma 6.1 and the fact that K_{U_t} is complete, $U_t \setminus K_{U_t}$ is a connected set, and thus $U_t \setminus K_{U_t}$ can be collapsed over in G' by Proposition 1.1, and $B_t = K_{U_t}$.

For any connected set C of $K_{U_t} \setminus A_{U_t}$ in $G(B)$, we shall show that it cannot be collapsed over in $G(B)$. Since $A_{U_t} \subseteq K_{U_t}$, we have $U_t \notin \mathcal{U}_l$; otherwise, U will be dropped in the pruning procedure. As a result, $B_t \notin \mathcal{B}_l$, and K_{B_t} contains at least two maximal S_{a_1} and S_{a_2}, $a_1, a_2 \in \{t\} \cup \Lambda_t$, such that C intersects one or both of them. If C intersects both S_{a_1} and S_{a_2}, then $\partial_{G(B)}(C)$ is not complete due to Lemma 6.3 (i). If C intersects only S_{a_1}, then $\partial_{G(B)}(C) \cap K_U \not\subseteq S_{a_1}$, because K_U is complete. By Lemma 6.3 (ii), we know that $\partial_{G(B)}(C)$ is not complete, and therefore, C cannot be collapsed over in $G(B)$. The same conclusion holds if C intersects only S_{a_2}.

(ii) If U_t satisfies condition $(C2)$, then $B_t = U_t$. Similar arguments used in (i) show that no connected set C of $U_t \setminus A_{U_t}$ in $G(B)$ can be collapsed over in $G(B)$.

(iii) If U_t satisfies condition $(C3)$, then $U_t \setminus (K_{U_t} \cup A_{U_t})$ can be collapsed over in G' by Proposition 1.1 and Lemma 6.1, and $B_t = K_{U_t} \cup A_{U_t}$. If C is any connected set of $K_{U_t} \setminus A_{U_t}$ in $G(B)$, it cannot be collapsed over in $G(B)$ by Lemma 6.3 (ii).

(iv) If U_t satisfies condition $(C4)$, then $B_t = U_t$. To show that a connected set C of $U_t \setminus A_{U_t}$ cannot be collapsed over in $G(B)$, we use a contrapositive argument. If C can be collapsed over in $G(B)$, then $\partial_{G(B)}(C)$ is complete by Proposition 1.1, and thus $C \cap K_{U_t} \neq \emptyset$ by the prime property of U. Assume without loss of generality that $C \cap S_{a_1} \neq \emptyset$ for some $a_1 \in \{t\} \cup \Lambda_t$. Then we have $(\partial_{G(B)}(C) \cap U) \not\subseteq S_{a_1}$; otherwise, there would be a decomposition for U_t, because $A_{U_t} \not\subseteq K_{U_t}$. By Lemma 6.3 (ii), $\partial_{G(B)}(C)$ is not complete, which contradicts the earlier conclusion about its completeness.

So far we have shown that $\mathcal{P}_G$ is collapsible onto $G(B)$ from Corollaries 2.2 and 2.3, and any connected set of $B \setminus A$, which is contained in one prime block of $G(B)$, cannot be collapsed over in $G(B)$. It remains to show that no connected set of $B \setminus A$ in $G(B)$ can be further collapsed over in $G(B)$. Otherwise, suppose that C is a connected set of $B \setminus A$, and that $\partial_{G(B)}(C)$ is complete. We note that $G'' = G(C \cup \partial_{G(B)}(C))$ is a convex subgraph in $G(B)$ and that any prime block of G'' is a prime block of $G(B)$ due to Lemma 3.1 of Wang and Guo [11]. If G'' is a prime graph, then C is contained in a prime block of $G(B)$, and thus $\partial_{G(B)}(C)$ is not complete from the proof above, leading to a contradiction. If G'' has at least two prime blocks, then there is a prime block B' of G'' and also $G(B)$ such that $B' \in \mathcal{B}_l$. Because $(B' \setminus K_{B'}) \cap A \neq \emptyset$ and $B' \setminus K_{B'} \subseteq C$, we have $C \cap A \neq \emptyset$, which is also a contradiction. Therefore any connected set of $B \setminus A$ cannot be further collapsed over in $G(B)$. From Corollary 2.2, we have proven that B is the minimal set containing A such that $\mathcal{P}_G$ is collapsible onto $G(B)$, and thus our algorithm is valid.

Acknowledgments

The authors thank the editor and an anonymous referee for valuable comments and suggestions on an earlier draft of the paper.

References

1. Asmussen, S., Edwards, D. (1983). *Collapsibility and response variables in contingency tables*. Biometrika, **70**, 567-578. MR725370 (85k:62125)
2. Diestel, R. (1990). *Graph decompositions. A study in infinite graph theory*. Clarendon Press, Oxford University Press, New York. MR1078627 (92a:05038)
3. Dobra, A., Hans, C., Jones, B., Nevins, J., and West, M. (2004). *Sparse graphical models for exploring gene expression data*. J. Multivariate. Anal., **90**, 196-212. MR2064941
4. Frydenberg, M. (1990). *Marginalization and collapsibility in graphical interaction models*. Ann. Statist., **18**, 790-805. MR1056337 (91i:62077)
5. Laurizen, S. L. (1992). *Propagation of probabilities, means, and variances in mixed graphical association models*. J. Amer. Statist. Assoc., **87**, 1098-1108. MR1209568 (93j:62040)
6. Leimer, H. G. (1993). *Optimal decomposition by clique separators*. Discrete Math., **113**, 99-123. MR1212872 (94e:05237)
7. Madigan, D. and Mosurski, K. (1990). *An extension of the results of Asmussen and Edwards on collapsibility in contingency tables*. Biometrika, **77**, 315-19. [Amendments and Corrections (1999), Biometrika, **86**, 973-74]. MR1064803 (91i:62075); MR1741994 (2001e:62055)
8. Rich, J., Hans, C., Jones, B., Inversen, E., McClendon, R., Rashed, B., Dobra, A., Dressman, H., Bigner, D., Nevins, J. and West, M. (2005). *Gene expression profiling and analysis in graphical association studies in glioblastoma survival*. Cancer Res., **65**, 4059-66.

9. Schäfer, J. and Strimmer, K. (2005). *An empirical Bayes approach to inferring large-scale gene association networks*. Bioinformatics, **21**, 754-64.
10. Tarjan, R. E. (1985). *Decomposition by clique separators*. Discrete Math., **55**, 221-32. MR798539 (87i:05134)
11. Wang, X. F. and Guo, J. H. (2008). *Some properties of junction trees of general graphs*. Front. Math. China, **3**, 399-413. MR2425162 (2009k:05153)
12. Wermuth, N. and Lauritzen, S. L. (1990). *On the interpretation and analysis of data using conditional independence graphs and graphical chain models*, J. Roy. Statist. Soc. Ser. B., **52**, 21-50. MR1049302 (92e:62106)
13. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R. and Newins, J. R. (2001). *Predicting the clinical status of human breast cancer by using gene expression profiles*. Proc. National Acad. Sci. USA, **98**, 11462-11467.

Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, Jilin Province, People's Republic of China

E-mail address: mathswangxiaofei@yahoo.com.cn

Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, Jilin Province, People's Republic of China

E-mail address: jhguo@nenu.edu.cn

Department of Statistics, University of Illinois, 725 S. Wright Street, Champaign, Illinois 61820

E-mail address: x-he@uiuc.edu