Revisiting Madigan and Mosurski: Collapsibility via Minimal Separators

By Pei Heng

KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun, China

peiheng@nenu.edu.cn

Yi Sun

College of Mathematics and System Science, Xinjiang University, Urumqi, China

brian@xju.edu.cn

Shiyuan He, and Jianhua Guo

School of Mathematics and Statistics, Beijing Technology and Business University, Beijing, China

heshiyuan@btbu.edu.cn; jhguo@btbu.edu.cn

Summary

Collapsibility provides a principled approach for dimension reduction in contingency tables and graphical models. Madigan & Mosurski (1990) pioneered the study of minimal collapsible sets in decomposable models, but existing algorithms for general graphs remain computationally demanding. We show that a model is collapsible onto a target set precisely when that set contains all minimal separators between its non-adjacent vertices. This insight motivates the Close Minimal Separator Absorption (CMSA) algorithm, which constructs minimal collapsible sets using only local separator searches at very low costs. Simulations confirm substantial efficiency gains, making collapsibility analysis practical in high-dimensional settings.

Some key words: Collapsibility; Minimal collapsible set; Minimal separator; Graphical model.

1. Introduction

Collapsibility is a classical idea in multivariate statistical analysis. First introduced by Yule (1903) and Simpson (1951), it provides a principled way to simplify statistical analysis by removing variables without distorting marginal associations. Within the framework of log-linear models, Asmussen & Edwards (1983), in a seminal *Biometrika* article, gave the first rigorous account, showing that collapsibility guarantees that marginal inferences coincide with those drawn from the full model. This property makes collapsibility an attractive tool for dimension reduction in contingency tables, graphical models, and related methods.

Also in *Biometrika*, a sharper problem was posed by Madigan & Mosurski (1990): for a target set of variables, what is the smallest superset onto which the model can be collapsed without loss of inferential validity? Such a superset is called a minimal collapsible set. Madigan & Mosurski (1990) proposed the selective acyclic hypergraph reduction (SAHR) algorithm for searching the minimal collapsible set. Their algorithm is effective but only works for decomposable graphical models. Later work sought to extend this idea to more general graphical models: Wang et al. (2011) developed a convex-hull based procedure, while Heng & Sun (2023) introduced path-absorption methods. These approaches, however, often require global graph operations and become computationally expensive for high-dimensional models.

In this note, we revisit minimal collapsibility from a new perspective. We show that a graphical model is collapsible onto a target set if and only if that set contains all minimal separators between non-adjacent vertices in the set. This characterization, simple yet powerful, shifts focus from global graph operations to local separation structure searching. Building on this observation, we introduce the Close Minimal Separator Absorption (CMSA) algorithm. CMSA identifies minimal collapsible sets by iteratively absorbing close minimal separators, which can be found efficiently within the neighbours of the target vertices at very low cost.

The benefits of CMSA are twofold. Conceptually, it reveals that collapsibility is governed by purely local graph properties. Computationally, it yields dramatic improvements in efficiency: in decomposable models, CMSA consistently outpaces SAHR, while in general graphical models, it scales far better than the existing algorithms. A simple worked example illustrates the mechanism, and simulations on large random graphs confirm its speed and robustness.

Our contribution is therefore both theoretical and practical: we provide a clean separator-based characterization of collapsibility, and a localized algorithm that makes collapsibility analysis feasible in high-dimensional graphical models.

2. Hierarchical log-linear model and Graph Representation

Asmussen & Edwards (1983) analysed hierarchical log-linear models for multidimensional contingency table N, where each cell i is determined by a set V of classifying factors. Specifically, each factor in V is a categorical variable for classifying observations. Let $p(i)$ be the probability of an observation falling into a cell i, where p is usually assumed to be multinomial. A hierarchical log-linear model L consists of a class of probabilities ($p \in L$) characterized by its generating class. The generating class has generators represented with square brackets. For example, for a 4-way table with $V = \{\alpha, \beta, \gamma, \theta\}$, the log-linear model with generator $[\alpha\beta][\beta\gamma\theta]$ is,

$$\log(m_{hjkl}) = \lambda + \lambda_h^{\alpha} + \lambda_j^{\beta} + \lambda_k^{\gamma} + \lambda_l^{\theta} + \lambda_{hj}^{\alpha\beta} + \lambda_{jk}^{\beta\gamma} + \lambda_{kl}^{\gamma\theta} + \lambda_{jl}^{\beta\theta} + \lambda_{jkl}^{\beta\gamma\theta}. \quad (1)$$

In the above, m_{ijkl} represents the expected number of observations in the cell, whose classifying factors $(\alpha, \beta, \gamma, \theta)$ equal (h, j, k, l). Besides, λ_h^{α} denotes the marginal effect of the variable α, $\lambda_{hj}^{\alpha\beta}$ represents the first-order interaction effect between α and β, and so on.

Log-linear models can be analysed from the graphical perspective (Asmussen & Edwards, 1983). An undirected graph $G = (V, E)$ consists of a vertex set V and a edge set $E \subseteq V \times V = \{(x, y) : x, y \in V\}$. Two vertices are connected by an edge if they are in some generator. For the model L in Eqn. (1) and the model $L' = [\alpha\beta][\beta\gamma][\gamma\theta][\beta\theta]$, they share the same interaction graph in Fig. 1 with the vertex set $V = \{\alpha, \beta, \theta, \gamma\}$. A model that is compatible with a given interaction graph and contains all highest-order interactions permitted by the graph is called a *graphical model*. Note that L is a graphical model, whereas L' is not. In this work, we assume that L is a graphical model with its interaction graph $G = (V, E)$.

Remark 1. As in Asmussen & Edwards (1983); Madigan & Mosurski (1990), we let L be a hierarchical log-linear model. However, it is worth noting that the proposed algorithm is equally applicable to general multinomial and Gaussian graphical models. For mixed graphical models containing both discrete and continuous variables, our algorithm can also correctly identify the

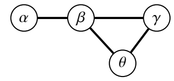

Fig. 1. The interaction graph of $[\alpha\beta][\beta\gamma\theta]$ and $[\alpha\beta][\beta\gamma][\gamma\theta][\beta\theta]$.

minimal collapsible sets by applying it to the constructed star graph (see Frydenberg, 1990; Wang et al., 2011, for more details).

We lastly review some graph terminologies. In G, two vertices x and y are adjacent if connected by an edge; then, x is a neighbour of y, and the set of all neighbours of x is denoted as $N_G(x)$. For a subset $A \subseteq V$, its neighbour set is $N_G(A) = \bigcup_{v \in A} N_G(v) \setminus A$. A path connecting u and v in G, denoted by l_{uv}, is a sequence of distinct vertices and edges $(u = v_0, e_0, v_1, e_1, \ldots, v_{k-1}, e_{k-1}, v_k = v)$ such that $e_i = (v_i, v_{i+1}) \in E$. We set $V(l_{uv}) = \{v_0, v_1, \ldots, v_k\}$ to be the set of vertices on l_{uv}. The subgraph induced by A is $G_A = (A, E_A)$ with $E_A = E \cap \{(x, y) : x, y \in A\}$. We refer to the maximal connected subgraph of G as a connected component.

3. Minimal collapsible set identifying algorithms

3.1. *Minimal collapsible set*

Recall, for a contingency table N, $p(i)$ is the probability for an observation falling into a cell i, and we set $\hat{p}(i)$ as its maximum likelihood estimate (MLE). For $A \subseteq V$, let $\hat{p}(i_A)$ be the probability obtained by marginalizing (summing) $\hat{p}(i)$ over the remaining factors $V \setminus A$. This marginalized probability may change the independence structure as specified by L between variables in A. Alternatively, to keep the independence structure, we can set L_A as the marginal model obtained by removing all factors in $V \setminus A$ and then deleting all redundant generators (those are contained within another remaining generator). For example, for the model $L = [\alpha\beta][\beta\gamma\theta]$ in Equation (1), assuming $A = \{\beta, \theta\}$, the marginal model is specified by $L_A = [\beta][\beta\theta] = [\beta\theta]$. Let p_{L_A} be a probability specified by L_A, and $\hat{p}_{L_A}$ as the corresponding MLE.

Asmussen & Edwards (1983) formalized the notion of *collapsibility*, which means the two MLEs coincide, i.e. $\hat{p}(i_A) = \hat{p}_{L_A}(i_A)$. Collapsibility ensures that marginal inferences are consistent with those drawn from the full model. In practice, such property can yield substantial savings in data collection and computation efforts, while enhancing robustness to unobserved variables.

Definition 1. *A hierarchical log-linear model L is said to be* collapsible onto $A \subseteq V$ *if the two MLEs coincide:* $\hat{p}(i_A) = \hat{p}_{L_A}(i_A)$.

Given a decomposable graphical model L and a vertex subset $A \subseteq V$, Madigan & Mosurski (1990) introduced the problem of identifying a minimal collapsible set: what is the smallest set B with $A \subseteq B \subseteq V$ such that L is collapsible onto B? They established the existence and uniqueness of such B, and proposed a simple and efficient procedure for its searching: iteratively remove simplicial vertices not in A until no further removal is possible. For general graphical models, the existence and uniqueness of a minimal collapsible set were established by Wang et al. (2011).

3.2. *Collapsibility characterization via separability*

The approach of Madigan & Mosurski (1990) only works for decomposable graphical models. For general graphical models, we present a novel characterization of collapsibility based on pairwise separability on a graph. For pairwise disjoint subsets $A, B, S \subseteq V$ in G, if $S \cap V(l_{ab}) \neq \emptyset$ for every path l_{ab} connecting any $a \in A$ and $b \in B$, then S separates A from B in G, denoted by $A \perp\!\!\!\perp B \mid S[G]$. The set S is called an AB-separator in G, and it is minimal if no proper subset of

S separates A from B in G. When $A = \{x\}$ and $B = \{y\}$, S is also called a minimal xy-separator. The proposed method is motivated by the following key lemma.

LEMMA 1 (ASMUSSEN & EDWARDS (1983)). *The graphical model L is collapsible onto subset $A \subseteq V$ if and only if $X \perp\!\!\!\perp Y \mid Z[G]$ implies $X \perp\!\!\!\perp Y \mid Z \cap A[G]$ for any $X, Y \subseteq A$.*

Lemma 1 implies that for a collapsible subset A, any minimal xy-separator Z for non-adjacent vertices $x, y \in A$ is contained within A. Otherwise, $Z \cap A$ would form a smaller separator, contradicting the minimality of Z. This observation motivates the following theorem, which formally characterizes the relationship between collapsibility and minimal separators.

THEOREM 1. *The graphical model L is collapsible onto a subset $A \subseteq V$ if and only if A contains every minimal xy-separator for every pair of non-adjacent vertices $x, y \in A$.*

Proof. The necessity follows by taking Z as a minimal separator in Lemma 1. We now prove sufficiency by contradiction. Suppose L cannot be collapsed onto A. Then there exists a connected component M of $G_{V \setminus A}$ and non-adjacent vertices $x, y \in N_G(M) \subseteq A$, which implies that there is no minimal xy-separator contained in A, leading to a contradiction. □

According to Theorem 1, a minimal collapsible set containing A can be constructed by iteratively absorbing all minimal separators between every pair of non-adjacent vertices in A until no further absorption is possible. While enumerating all such separators may appear computationally prohibitive in high-dimensional graphical models, we show that the task can in fact be carried out efficiently through localized searches.

3.3. *Close minimal separator absorption algorithm*

Although identifying a minimal collapsible set appears to be a "global" task — requiring examination of the entire set A with respect to the graph G — it can, in fact, be accomplished efficiently through localized operations at very low cost. It suffices at each iteration to absorb only one minimal separator for each pair of non-adjacent vertices. The following corollary, which follows directly from the proof of Theorem 1, formalizes this observation; for brevity, the proof is omitted.

COROLLARY 1. *The graphical model L is collapsible onto a subset $A \subseteq V$ if and only if A contains at least one minimal xy-separator for every pair of non-adjacent vertices $x, y \in A$.*

More importantly, the minimal separator required for each pair is what is known as a *close minimal separator*, defined as one that lies entirely within the neighbourhood of a vertex. The importance of this notion is that it reduces the problem to purely localized searches.

DEFINITION 2 (TAKATA (2010)). *For any two non-adjacent vertices $x, y \in V$, a minimal xy-separator S is said to be close to x if $S \subseteq N_G(x)$, denoted by $S_G(x, y)$. Similarly, $S_G(y, x)$ denotes a minimal xy-separator close to y.*

For example, in Fig. 2 (A), $\{e, l\}$ is the minimal bt-separator close to t, whereas $\{e, s\}$ is the minimal bt-separator close to b. Takata (2010) also proposed the **CloseSeparator** algorithm, which identifies close minimal separators with a time complexity of $O(m)$, where m denotes the number of edges in the graph G. The algorithm first finds the neighbour $N_G(x)$ of x, and then identifies the connected component M containing y in the subgraph $G_{V \setminus N_G(x)}$. The neighbour set $N_G(M)$ then forms the minimal xy-separator that is close to x.

We now formally introduce the *Close Minimal Separator Absorption (CMSA)* algorithm (Algorithm 1) for the efficient identification of minimal collapsible sets. Given a hierarchical log-linear

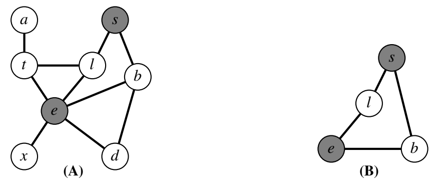

Fig. 2. (A) A graph $G = (V, E)$. (B) The minimal collapsible subgraph G_B containing $\{e, s\}$.

model L with its interaction graph $G = (V, E)$ and a target variable set A, our proposed algorithm consists of the following steps:

(i) Identify all connected components $M_1, \ldots, M_K$ of $G_{V \setminus A}$.

(ii) For each connected component M_i, initialize $B_i := A$ and consider the subgraph $G_i = G_{B_i \cup M_i}$. Iteratively identify non-adjacent vertex pairs within the neighbourhoods of the connected components of G_{M_i}, and absorb their close minimal separators in G_i into B_i. The iteration stops when, for all connected components in the updated G_{M_i}, their neighbours in G_i form complete subsets. At this point, according to Theorem 2.3 in Asmussen & Edwards (1983), G_i can be collapsed onto B_i, and therefore the iteration is terminated.

(iii) Merge the sets B_i obtained from all connected components to form the final minimal collapsible set $B = \bigcup_i B_i$, which contains A.

Algorithm 1 Close Minimal Separator Absorption Algorithm (CMSA)

Input: A graphical model L with its interaction graph $G = (V, E)$ and a target variable set A.
Output: The minimal collapsible set B containing the subset A.

Identify all connected components $M_1, M_2, \ldots, M_K$ of $G_{V \setminus A}$;
for each connected component G_{M_i} of $G_{V \setminus A}$ **do**
  Initialize: $B_i := A$, $G_i = G_{B_i \cup M_i}$;
  repeat
    If there exists a connected component C of G_{M_i} whose neighbour in G_i contains a pair of non-adjacent vertices, select any such pair $\{u, v\} \subseteq N_{G_i}(C) \subseteq B_i$;
    Find $S_{G_i}(u, v)$ and $S_{G_i}(v, u)$ using **CloseSeparator**;
    Update $B_i := B_i \cup S_{G_i}(u, v) \cup S_{G_i}(v, u)$ and $M_i := M_i \setminus (S_{G_i}(u, v) \cup S_{G_i}(v, u))$;
  until In G_i, the neighbour of every connected component of G_{M_i} is complete.
end for
return $\bigcup_i B_i$, where B_i is the set obtained in the i-th connected component M_i.

The following example illustrates the execution of Algorithm 1.

Example 1. Let $G = (V, E)$ be the graph shown in Fig. 2(A), and let the target variable set be $A = \{e, s\}$. At the start of the CMSA algorithm, the subgraph $G_{V \setminus A}$ consists of three connected components: $M_1 = \{x\}$, $M_2 = \{b, d\}$, and $M_3 = \{a, l, t\}$, with B_i initialized as A for $i = 1, 2, 3$. The algorithm then processes each connected component as follows:

1. For $M_1 = \{x\}$, the boundary $N_{G_{\{e,s,x\}}}(\{x\}) = \{e\}$. There are no non-adjacent vertices in $N_G(\{x\})$, so no update of B_1 is needed.
2. For $M_2 = \{b, d\}$, with $G_2 = B_2 \cup M_2 = \{b, d, e, s\}$, the boundary $N_{G_2}(\{b, d\}) = \{e, s\}$, where s and e form a non-adjacent pair. In G_2, the close minimal es-separators

are $\{b\}$ for both vertices; thus, we update $B_2 = \{b, e, s\}$ and $M_2 = \{d\}$. At this point, $N_{G_2}(M_2) = N_{G_2}(\{d\}) = \{b, e\}$. Since vertices b and e are adjacent, processing for this component terminates.

3. For $M_3 = \{a, l, t\}$, the boundary also contains the non-adjacent pair $\{e, s\}$. Both close minimal es-separators are $\{l\}$; hence, we update $B_3 = \{e, l, s\}$ and $M_3 = \{a, t\}$. No further absorption is required for this component.

Finally, the CMSA algorithm merges the three updated sets B_i, and the union $\bigcup B_i = \{b, e, l, s\}$ forms the minimal collapsible set containing $\{e, s\}$. The induced subgraph is shown in Fig. 2(B).

The next two theorems establish the correctness and computational complexity of Algorithm 1.

THEOREM 2. *The subset B obtained by CMSA is the minimal collapsible set containing A.*

Proof. Let B' be the minimal collapsible set that contains A. We aim to show that $B = B'$.

Since the CMSA algorithm absorbs only close minimal separators, and by Theorem 1 the set B' already contains all minimal separators between non-adjacent vertices in A, it follows from the iterative construction that $B \subseteq B'$ holds trivially. Therefore, it suffices to prove that the graphical model L is collapsible onto B.

Toward a contradiction, assume that L is not collapsible onto B. Then there exists a connected component M of $G_{V\setminus B}$ and non-adjacent vertices $x, y \in N_G(M) \subseteq B$. Let M_i denote the connected component of $G_{V\setminus A}$ containing M, and let B_i be the subset returned from processing M_i. By construction, M forms a connected component of $G_{A\cup M_i\setminus B_i}$, and its neighbourhood contains two non-adjacent vertices x and y. This contradicts the termination condition of the CMSA algorithm. Hence, L is collapsible onto B, and we conclude that $B = B'$, completing the proof.□

THEOREM 3. *The CMSA algorithm has a time complexity of $O(nm)$ and a space complexity of at most $O(n)$, where n and m denote the number of vertices and edges in the graph, respectively.*

Proof. When the graph is represented in adjacency-list form, each close minimal separator can be identified in $O(m)$ time with $O(n)$ space (Takata, 2010), and at most $n - 2$ absorptions are required. Consequently, the overall complexity is $O(nm)$ time and $O(n)$ space. □

Remark 2. Existing algorithms for general models achieve the same $O(nm)$ time complexity but typically require $O(n + m)$ space (Wang et al. (2011); Heng & Sun (2023)). In contrast, the CMSA algorithm attains a more favourable space complexity of $O(n)$.

4. EXPERIMENTAL STUDIES

4.1. *Implementation of the algorithms*

We evaluated the proposed algorithm through experiments. All algorithms were implemented in C and interfaced through Python. The complete source code is publicly accessible at https://github.com/Balance-H/Algorithms. All experiments were conducted on a system equipped with an Intel Xeon Silver 4215R CPU and 128 GB of RAM.

4.2. *Efficiency of identifying minimal collapsible sets in decomposable graphical models*

We first evaluate the efficiency of CMSA and SAHR (Madigan & Mosurski (1990)) in identifying minimal collapsible sets within decomposable graphical models. For each combination of graph size $n \in \{250, 500, 750, 1000\}$ and edge probability $p \in \{0.1, 0.01\}$, we generate 100 random chordal graphs with approximately $n - 1 + 0.5\, n(n - 1)p$ edges. In each graph, we randomly select 10 target vertices and apply both CMSA and SAHR to identify the minimal collapsible sets containing these vertices. We record the runtimes and compute the mean runtime over the 100 graphs for each (n, p) configuration.

Table 1: Average running times (s) of CMSA and SAHR in decomposable graphical models

Nodes	250		500		750		1000	
Average edges	529	3334	1812	12912	3567	28652	6062	52959
CMSA	0.0007	0.0012	0.0021	0.0047	0.0044	0.0112	0.0072	0.0248
SAHR	0.0113	0.0611	0.0681	0.5455	0.1876	2.1648	0.3808	6.6983

The experimental results in Table 1 indicate that CMSA consistently outperforms SAHR across all graph sizes and densities. The advantage of CMSA becomes increasingly pronounced as the number of nodes and edges grows. This efficiency gain is primarily due to CMSA's localized separator absorption strategy, which avoids the repeated boundary checks and global updates that dominate SAHR's runtime in high-dimensional graphs.

4.3. *Efficiency of identifying minimal collapsible sets in general graphical models*

We next evaluate the performance of the CMSA algorithm on general graphical models. Since SAHR is limited to decomposable graphs, we adopt the Induced Path Absorption (IPA) algorithm (Heng & Sun, 2023) as a baseline for comparison, as it has been shown to be the most effective method in Heng & Sun (2023). For each combination of graph size $n \in \{2500, 5000, 7500, 10000\}$ and edge probability $p \in \{0.1, 0.01, 0.005, 0.001\}$, we generate 100 random trees with n vertices and independently add edges with probability p to obtain the corresponding random graphs. In each graph, we randomly select 10 target vertices and apply the corresponding algorithm to identify the minimal collapsible sets containing these vertices. We record the runtimes and compute the average over the 100 graphs for each (n, p) configuration.

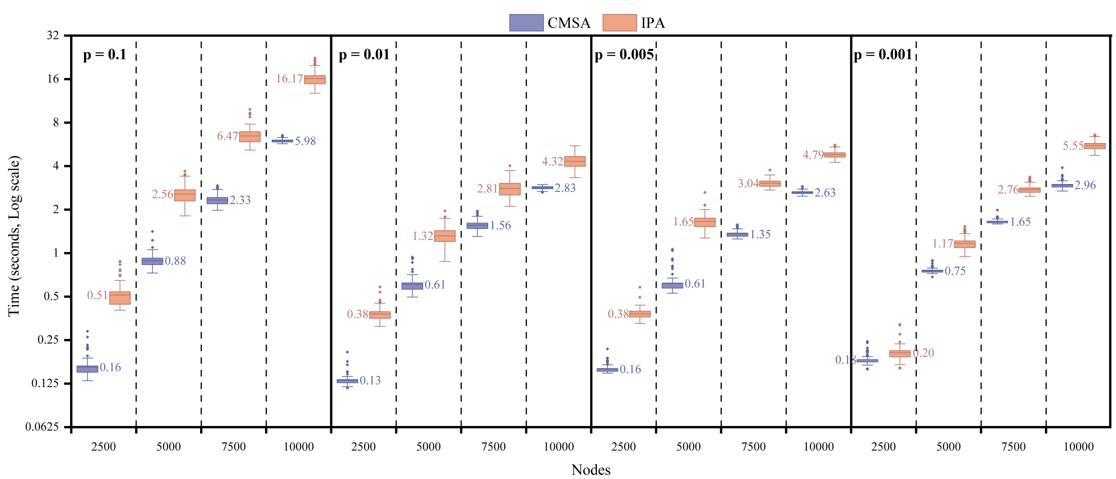

Fig. 3. Average runtimes (log-scaled, in seconds) of CMSA and IPA on networks of varying sizes.

The experimental results are presented in Fig. 3. The horizontal axis represents the number of nodes n (separated by dashed lines), and the vertical axis shows the runtime in seconds on a logarithmic scale. The four subplots correspond to different edge densities p (separated by solid lines). The blue boxplots represent the CMSA algorithm, and the orange ones represent the IPA algorithm, with the average runtime annotated beside each box. The results show that CMSA is substantially more efficient than IPA on dense graphs, with the performance advantage increasing as the number of nodes grows. As the graphs become sparser, the runtime of both algorithms decreases significantly, while CMSA consistently maintains superior efficiency.

5. DISCUSSION

This note revisits a line of research tracing back to seminal works in *Biometrika*: Asmussen & Edwards (1983) first formalized collapsibility in log–linear models, and Madigan & Mosurski (1990) introduced an elegant reduction algorithm for finding minimal collapsible sets of decomposable graphs. We extend this line of work by showing that collapsibility admits a simple separator–based characterization: a model is collapsible onto a target set if and only if the set contains all minimal separators between its non-adjacent vertices.

This perspective reveals that collapsibility is fundamentally linked to local separation structure. On this basis, we developed the Close Minimal Separator Absorption (CMSA) algorithm, which constructs minimal collapsible sets through localized absorption of close minimal separators. The proposed procedure enjoys conceptual clarity and substantial computational gains, consistently outperforming existing methods and scaling to high-dimensional graphs.

Our findings highlight both theoretical and practical implications. Theoretically, they deepen the connection between collapsibility and local separator structure, offering a novel view that advances earlier *Biometrika* contributions. Practically, they provide a tool that makes collapsibility analysis tractable in modern large-scale applications. We anticipate that the separator–based view may prove useful not only for contingency tables but also for broader classes of graphical models.

REFERENCES

ASMUSSEN, S. & EDWARDS, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70**, 567–578.

FRYDENBERG, M. (1990). Marginalization and collapsibility in graphical interaction models. *The Annals of Statistics* , 790–805.

HENG, P. & SUN, Y. (2023). Algorithms for convex hull finding in undirected graphical models. *Applied Mathematics and Computation* **445**, 127852.

MADIGAN, D. & MOSURSKI, K. (1990). An extension of the results of asmussen and edwards on collapsibility in contingency tables. *Biometrika* **77**, 315–319.

SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **13**, 238–241.

TAKATA, K. (2010). Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph. *Discrete Applied Mathematics* **158**, 1660–1667.

WANG, X., GUO, J. & HE, X. (2011). Finding the minimal set for collapsible graphical models. *Proceedings of the American Mathematical Society* **139**, 361–373.

YULE, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika* **2**, 121–134.