Faster algorithms for Markov equivalence

Zhongyi Hu
Department of Statistics
University of Oxford
zhongyi.hu@keble.ox.ac.uk

Robin Evans
Department of Statistics
University of Oxford
evans@stats.ox.ac.uk

Abstract

Maximal ancestral graphs (MAGs) have many desirable properties; in particular they can fully describe conditional independences from directed acyclic graphs (DAGs) in the presence of latent and selection variables. However, different MAGs may encode the same conditional independences, and are said to be *Markov equivalent*. Thus identifying necessary and sufficient conditions for equivalence is essential for structure learning. Several criteria for this already exist, but in this paper we give a new non-parametric characterization in terms of the heads and tails that arise in the parameterization for discrete models. We also provide a polynomial time algorithm ($O(ne^2)$, where n and e are the number of vertices and edges respectively) to verify equivalence. Moreover, we extend our criterion to ADMGs and summary graphs and propose an algorithm that converts an ADMG or summary graph to an equivalent MAG in polynomial time ($O(n^2e)$). Hence by combining both algorithms, we can also verify equivalence between two summary graphs or ADMGs.

1 INTRODUCTION

DAG models, also known as Bayesian networks, are popular graphical models that associate a probability distribution $P(X_V)$ with a graph consisting of vertices representing random variables X_V joined by directed edges. In the context of causal inference, a directed edge $a \to b$ can be interpreted as 'a has a direct causal effect on b'. A DAG encodes conditional independence in P by a criterion called d-separation (Pearl, 2009). For example, $1 \to 2 \leftarrow 3$ is a DAG with vertices $1, 2, 3$ and implies one independence: $X_1 \perp\!\!\!\perp X_3$. DAGs are also associated with an elegant factorization of probability distributions, which allows fast statistical inference and fitting. With some additional assumptions they can be used for causal modelling, and thus they are used in many fields such as expert systems, pattern recognition in machine learning, or estimating causal effects in experimental science.

An interesting question is how to learn unknown DAGs from a dataset. Spirtes et al. (2000) provide an algorithm called the PC algorithm; this learns the underlying DAG by testing conditional independences inherited in the data. However, when latent variables are present, conditional independence in the observed variables may imply the wrong underlying causal structure, or even not correspond to any DAGs at all. For example, in Figure 1(i) with latent variable h (this is an example from Richardson and Spirtes (2002)), there is no DAG that describes precisely the independence on the marginal. We say, then, that DAGs are not closed under marginalization. Classes of supermodels have been developed to tackle this problem, one of which is *maximal ancestral graphs* (MAGs) introduced by Richardson and Spirtes (2002). This includes graphs with additional types of edges: bidirected edges ($\leftrightarrow$) and undirected edges. A bidirected edge $1 \leftrightarrow 2$ can be interpreted as saying that there is a latent variable h such that $1 \leftarrow h \to 2$. An undirected edge arises when there are some variables being conditioned upon. Graphical implications for conditional independence are extended from d-separation to m-separation, see definitions in Section 2. Moreover one can project a DAG with latent and selection variables to a Markov equivalent MAG on the observed margin. The projection is described in Section 3.3. The resulting MAG not only captures the exact conditional independence of observed variables in the original graph but also preserves ancestral relations. In addition, Gaussian variables associated with MAGs are curved exponential families (Richardson and Spirtes, 2002), and hence have some desirable statistical properties.

Graphs in this paper are directed and contain no undi-

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

rected edge. Extensions to *summary graphs* and MAGs with undirected edges are straightforward and we have placed them in the supplementary materials. Note that summary graphs defined in Wermuth (2011) are actually the same as ADMGs with undirected components at the top. Graphically, one just needs to change the dashed lines to bidirected edges and they encode the same conditional independence. We include details of this discussion in the supplementary materials.

Learning causal structures via testing only conditional independence leads to another problem. Different MAGs can imply the same set of constraints on variables, for example Figures 3(i) and (ii) both only encode $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$. We say such MAGs are *Markov equivalent*, and are in the same *Markov equivalence class*. Thus equivalent graphs represent the same set of distributions. Although each class can be uniquely described by a *partial ancestral graph* (PAG) (Colombo et al., 2012), non-experimental data cannot distinguish graphs in the same class. Therefore identifying conditions for Markov equivalence is important for modelling and estimating causal effects from data.

There have been three graphical characterizations that give necessary and sufficient conditions for when two MAGs are equivalent. Among those three criteria, only Ali et al. (2009) provide a polynomial time algorithm to verify Markov equivalence. Zhao et al. (2005) characterize MAGs by *minimal collider paths* (MCPs). The criterion of Spirtes and Richardson (1997) uses *discriminating paths*, which we will define in Section 3 (we will employ them in our proofs). This paper gives a new characterization and it lead to a faster algorithm to test equivalence compared to existing ones. Also we show a similar equivalence criterion for wider classes of acyclic graphs, ADMGs.

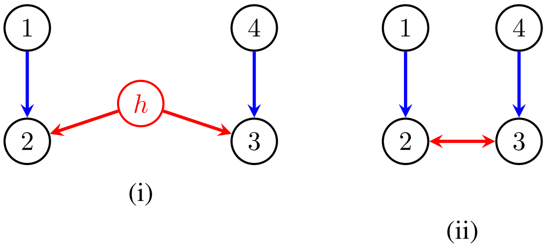


Figure 1: (i) A DAG with latent variable h. (ii) A Markov equivalent MAG of (i) on the margin $\{1, 2, 3, 4\}$.

In Section 2, we give basic definitions and terminology for graphical models. In Section 3 we present the main results, including theorems on the Markov equivalence of MAGs and ADMGs. Algorithms to verify Markov equivalence and their complexities are shown in Section 4. Missing proofs are found in the Supplementary Material.

2 DEFINITIONS

2.1 Graphs

A *graph* $\mathcal{G}$ consists of a vertex set $\mathcal{V}$ and an edge set $\mathcal{E}$ of distinct pairs of vertices. For an edge in $\mathcal{E}$ connecting vertices a and b, we say these two vertices are the *endpoints* of the edge and the two vertices are *adjacent* (if there is no edge between a and b, they are *nonadjacent*).

A *path* is a set of distinct vertices $v_i, 1 \leq i \leq k$ such that v_i and v_{i+1} is connected by some edge for all $i \leq k-1$. A path is *directed* if its edges are all directed and point in the same direction. A graph $\mathcal{G}$ is *acyclic* if there is no directed cycle (any *directed path* such that $v_1 \rightarrow v_2 \rightarrow ... \rightarrow v_k$ and $v_k \rightarrow v_1$). A *graph* $\mathcal{G}$ is called an *acyclic directed mixed graph* (ADMG) if it is *acyclic* and contains only directed and bidirected edges.

For a vertex v in an ADMG $\mathcal{G}$, we define the following sets:

$$\begin{aligned}
\mathrm{pa}_{\mathcal{G}}(v) &= \{w : w \rightarrow v \text{ in } \mathcal{G}\} \\
\mathrm{sib}_{\mathcal{G}}(v) &= \{w : w \leftrightarrow v \text{ in } \mathcal{G}\} \\
\mathrm{an}_{\mathcal{G}}(v) &= \{w : w \rightarrow ... \rightarrow v \text{ in } \mathcal{G} \text{ or } w = v\} \\
\mathrm{de}_{\mathcal{G}}(v) &= \{w : v \rightarrow ... \rightarrow w \text{ in } \mathcal{G} \text{ or } w = v\} \\
\mathrm{dis}_{\mathcal{G}}(v) &= \{w : w \leftrightarrow ... \leftrightarrow v \text{ in } \mathcal{G} \text{ or } w = v\}.
\end{aligned}$$

They are known as the *parents*, *siblings*, *ancestors*, *descendants* and *district* of v, respectively. These sets are also defined disjunctively for a set of vertices $W \subseteq \mathcal{V}$. For example $\mathrm{pa}_{\mathcal{G}}(W) = \bigcup_{w \in W} \mathrm{pa}_{\mathcal{G}}(w)$. Vertices in the same district are connected by a bidirected path and this is an equivalence relation, so we can partition $\mathcal{V}$ and denote the *districts* of a *graph* $\mathcal{G}$ by $\mathcal{D}(\mathcal{G})$. We sometimes ignore the subscript if the graph we refer to is clear, for example $\mathrm{an}(v)$ instead of $\mathrm{an}_{\mathcal{G}}(v)$.

2.2 Separation Criterion

For a path π with vertices v_i, $1 \leq i \leq k$ we call v_1 and v_k the *endpoints* of π and any other vertices the *nonendpoints* of π. For a nonendpoint w in π, it is a *collider* if $?\!\!\rightarrow w \leftarrow\!\!?$ on π and a *noncollider* otherwise (an edge $?\!\!\rightarrow$ is either $\rightarrow$ or $\leftrightarrow$). For two vertices a, b and a disjoint set of vertices C in $\mathcal{G}$ (C might be empty), a path π is *m-connecting* a, b given C if (i) a, b are endpoints of π, (ii) every noncollider is not in C and (iii) every collider is in $\mathrm{an}_{\mathcal{G}}(C)$. A *collider path* is a path where all the nonendpoints vertices are colliders.

Definition 2.1. For three disjoint sets A, B and set C (A, B are non-empty), A and B are *m-separated* by C in

$\mathcal{G}$ if there is no m-connecting path between any $a \in A$ and any $b \in B$ given C. We denote the m-separation by $A \perp_m B \mid C$.

For a triple (a, b, c) in a graph $\mathcal{G}$, we call this an *unshielded triple* if $\{a, b\}$ and $\{b, c\}$ are adjacent but $\{a, c\}$ are not. If b is a also collider in the path $\langle a, b, c\rangle$ then we also call the triple an *unshielded collider*, and an *unshielded noncollider* otherwise.

Definition 2.2. A distribution $P(X_V)$ is said to be in the *Markov model* of an ADMG $\mathcal{G}$ if whenever $A \perp_m B \mid C$ in $\mathcal{G}$, $X_A \perp\!\!\!\perp X_B \mid X_C$ in P.

This definition, known as the global Markov property, associates distributions with a given ADMG via m-separations. There are also other equivalent definitions in terms of the local Markov property or moralization (see Richardson, 2003), but the global Markov property has the advantage of being 'complete': that is, if there is no m-separation then almost every distribution in the model does not satisfy the associated conditional independence.

Remark 1. The model in Definition 2.2 defined by conditional independences is sometimes referred as the *ordinary Markov model*. There is a model called the *nested Markov model* defined by generalized conditional independences which captures all the equality constraints that arise from latent variable model (see Richardson et al., 2017; Evans, 2018).

For an ADMG $\mathcal{G}$, given a subset $W \subseteq \mathcal{V}$, the induced subgraph $\mathcal{G}_W$ is defined as the graph with vertex set W and edges in $\mathcal{G}$ whose endpoints are both in W. Also for the *district* of a vertex v in an induced subgraph $\mathcal{G}_W$, we may denote it by $\text{dis}_W(v)$.

2.3 MAGs

Definition 2.3. An ADMG $\mathcal{G}$ is *maximal* if for every pair of *nonadjacent* vertices a and b, there exists some set C such that a, b are m-separated given C in $\mathcal{G}$.

Definition 2.4. An ADMG $\mathcal{G}$ is *ancestral* if for every $v \in \mathcal{V}$, $\text{sib}_{\mathcal{G}}(v) \cap \text{an}_{\mathcal{G}}(v) = \emptyset$.

Definition 2.5. An ADMG $\mathcal{G}$ is called a *maximal ancestral graph* (MAG) if it is *maximal* and *ancestral*.

Note that in an ancestral graph, there is at most one edge between each pair of vertices.

For example, the graph in Figure 2(i) is not maximal because 1 and 2 are not adjacent, but no subset of $\{3, 4\}$ will m-separate them. (ii) is not ancestral as 1 is a sibling of 3, which is also one of its descendants. (iii) is a MAG in which the only conditional independence is $X_1 \perp\!\!\!\perp X_3 \mid X_4$.

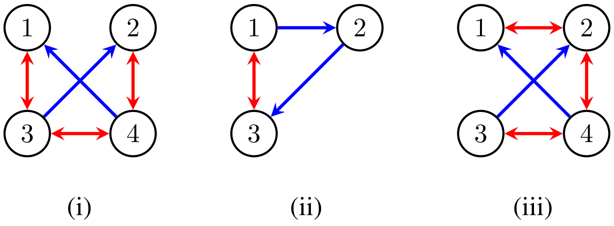


Figure 2: (i) An ancestral graph that is not maximal. (ii) A maximal graph that is not ancestral. (iii) A maximal ancestral graph.

Definition 2.6. Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with the same vertex sets, are said to be *Markov equivalent* if any m-separation holds in $\mathcal{G}_1$ if and only if it holds in $\mathcal{G}_2$.

2.4 Heads and Tails

For a vertex set $W \subseteq \mathcal{V}$, we define the *barren subset* of W as:

$$\text{barren}_{\mathcal{G}}(W) = \{w \in W : \text{de}_{\mathcal{G}}(w) \cap W = \{w\}\}.$$

A vertex set H is called a *head* if (i) $\text{barren}_{\mathcal{G}}(H) = H$ and (ii) H is contained in a single district in $\mathcal{G}_{\text{an}(H)}$. For an ADMG $\mathcal{G}$, we denote the set of all heads in $\mathcal{G}$ by $\mathcal{H}(\mathcal{G})$. A *tail* of a head is defined as:

$$\text{tail}(H) = (\text{dis}_{\text{an}(H)}(H) \setminus H) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\text{an}(H)}(H)).$$

Distributions associated with an Markov model can be factorized in terms of heads and tails (Richardson, 2009).

Definition 2.7. The *parametrizing set* of $\mathcal{G}$, denoted by $\mathcal{S}(\mathcal{G})$ is defined as:

$$\mathcal{S}(\mathcal{G}) = \{H \cup A : H \in \mathcal{H}(\mathcal{G}) \text{ and } \emptyset \subseteq A \subseteq \text{tail}(H)\}.$$

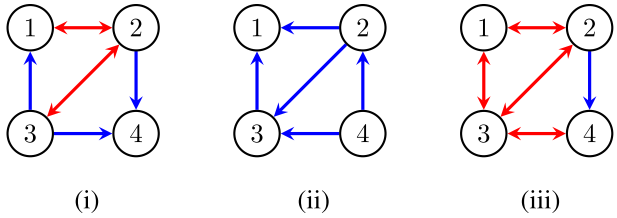
Note that it is called the parametrizing set because it is closely related to the discrete parameterization (Evans and Richardson, 2014). However the theorem developed in this paper is entirely non-parametric. We also define $\mathcal{S}_k(\mathcal{G})$ for $k \geq 2$ as:

$$\mathcal{S}_k(\mathcal{G}) = \{S \in \mathcal{S}(\mathcal{G}) : 2 \leq |S| \leq k\}.$$

In particular, we are interested in:

$$\tilde{\mathcal{S}}_3(\mathcal{G}) = \{S \in \mathcal{S}_3(\mathcal{G}) \mid \text{there are 1 or 2 adjacencies among the vertices in } S\}.$$

We write $\mathcal{S}, \mathcal{S}_k, \tilde{\mathcal{S}}_3$ if the graph $\mathcal{G}$ we are referring to is clear. Note that we are not considering any singleton sets in $\mathcal{S}_k(\mathcal{G})$ or $\tilde{\mathcal{S}}_3(\mathcal{G})$; these are just all vertices because $\{v\}$ is trivially a head. For a MAG $\mathcal{G}$, a pair of vertices are



(i) (ii) (iii)

Figure 3: Three MAGs where (i) and (ii) are Markov equivalent but (iii) is not.

in $\mathcal{S}(\mathcal{G})$ if and only if they are adjacent (This is easy to prove).

We give an example to illustrate what the sets defined above are. Consider the three MAGs in Figure 3, Table 1 lists their heads and tails, Table 2 lists their parametrizing sets $\mathcal{S}$ and Table 3 lists their $\mathcal{S}_3$ and $\tilde{\mathcal{S}}_3$.

Table 1: Heads and tails of graphs in Figure 3

Figure	heads	tails	Figure	heads	tails
3(i)	1	3	3(iii)	1	∅
	2	∅		2	∅
	3	∅		3	∅
	4	2,3		4	2
	1,2	3		1,2	∅
	2,3	∅		1,3	∅
3(ii)	1	2,3		2,3	∅
	2	4		3,4	2
	3	2,4		1,2,3	∅
	4	∅		1,3,4	2

Table 2: Parametrizing set of graphs in Figure 3

Figure	parametrizing sets	missing sets
3(i)(ii)	$\{1\},\{2\},\{3\},\{4\}$ $\{1,2\},\{1,3\},\{2,3\}$ $\{2,4\},\{3,4\}$ $\{1,2,3\},\{2,3,4\}$	$\{1,4\}$ $\{1,2,4\}$ $\{1,3,4\}$ $\{1,2,3,4\}$
3(iii)	$\{1\},\{2\},\{3\},\{4\}$ $\{1,2\},\{1,3\},\{2,3\}$ $\{2,4\},\{3,4\}$ $\{1,2,3\},\{1,3,4\},\{2,3,4\}$ $\{1,2,3,4\}$	$\{1,4\}$ $\{1,2,4\}$

In Figure 3, (i) is Markov equivalent to (ii) and they also have the same parametrizing sets; however, (iii) has a different parametrizing set and is not Markov equivalent to either (i) or (ii). In Figure 3(i) and (ii), $1 \perp_m 4 \mid 2,3$ is the only m-separation while Figure 3(iii) encodes $1 \perp_m 4 \mid 2$. Note that these conditional independences correspond precisely to these missing sets which are in the form $\{a,b\} \cup C'$ where $a \perp_m b \mid C$ and $C' \subseteq C$. Thus it is reasonable to conjecture that equivalent graphs should have the same parametrizing sets. It turns out that not only is this true, but in fact equivalence conditions can be refined even further and it is sufficient to consider $\mathcal{S}_3$ or $\tilde{\mathcal{S}}_3$.

Table 3: $\mathcal{S}_3$ and $\tilde{\mathcal{S}}_3$ graphs in Figure 3

Figure	$\mathcal{S}_3$	$\tilde{\mathcal{S}}_3$
3(i)(ii)	$\{1,2\},\{1,3\},\{2,3\}$ $\{2,4\},\{3,4\}$ $\{1,2,3\},\{2,3,4\}$	$\{1,2\},\{1,3\}$ $\{2,3\},\{2,4\}$ $\{3,4\}$
3(iii)	$\{1,2\},\{1,3\},\{2,3\}$ $\{2,4\},\{3,4\}$ $\{1,2,3\},\{2,3,4\}$ $\{1,3,4\}$	$\{1,2\},\{1,3\}$ $\{2,3\},\{2,4\}$ $\{3,4\}$ $\{1,3,4\}$

3 MARKOV EQUIVALENCE

3.1 Previous Work

The first theorem on Markov equivalence of MAGs is from Spirtes and Richardson (1997).

Theorem 3.1. *Two MAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if (i) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same adjacencies, (ii) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same unshielded colliders and (iii) if π forms a discriminating path for b in $\mathcal{G}_1$ and $\mathcal{G}_2$, then b is a collider on the path π in $\mathcal{G}_1$ if and only it is a collider on the path π in $\mathcal{G}_2$.*

For x and y nonadjacent, a *discriminating path* $\pi = \langle x, q_1, \ldots, q_m, b, y\rangle$, $m \geq 1$ for b, is a subgraph comprised of a collection of paths:

$$\begin{aligned} &x \mathrel{?\!\!\rightarrow} q_1 \leftrightarrow \cdots \leftrightarrow q_i \rightarrow y, \qquad 1 \leq i \leq m;\\ &x \mathrel{?\!\!\rightarrow} q_1 \leftrightarrow \cdots \leftrightarrow q_m \mathrel{\leftarrow\!\!?} b \mathrel{?\!\!\rightarrow} y. \end{aligned}$$

For example, $\langle 1,2,3,4\rangle$ forms a discriminating path for 3 in both Figure 3(i) and (iii), but not (ii). The vertex 3 is a collider on the path in (iii) but not (i), so (i) and (iii) are not equivalent; however (i) and (ii) are equivalent. In general, the cost of identifying all the discriminating paths is not polynomial in the number of vertices and edges. However, we will make use of Theorem 3.1 in later proofs.

3.2 Markov Equivalence Of MAGs

We now present the main result of this paper.

Theorem 3.2. *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two MAGs. Then $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if $\mathcal{S}(\mathcal{G}_1) = \mathcal{S}(\mathcal{G}_2)$.*

Theorem 3.2 already provides a method to find equivalence between two MAGs by searching all the heads and corresponding tails, however, the number of heads is not polynomial in the size of the graph.

Corollary 3.2.1. *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two MAGs. Then $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if $\mathcal{S}_3(\mathcal{G}_1) = \mathcal{S}_3(\mathcal{G}_2)$. This in turn occurs if and only if $\tilde{\mathcal{S}}_3(\mathcal{G}_1) = \tilde{\mathcal{S}}_3(\mathcal{G}_2)$.*

The motivation for defining $\tilde{\mathcal{S}}_3(\mathcal{G})$ is that we cannot obtain the same complexity if we allow triangles to be included, as in $\mathcal{S}_3$. To see this, consider a complete bidirected graph with e edges: this will require $O(e^3)$ operations to list all the triangles (which are all heads). Note we do not care about triples with three or zero adjacencies. Theorem 3.1 tells us that apart from adjacencies between pairs of vertices, and unshielded triples which lack one adjacency, we only need to find that for a discriminating path $\pi = \langle x, q_1, \dots, q_m, b, y\rangle$, whether b is a collider on the path or not. Later we will show that $\{x, b, y\} \in \mathcal{S}(\mathcal{G})$ if and only if b is a collider on π, and note that b, y are adjacent but x, y are not.

Corollary 3.2.1 is particularly important for identifying Markov equivalence. It not only allows the algorithm to run in polynomial time as we only need to check heads with size at most 3, but also accelerates it further as we do not need to find triples with full adjacencies or no adjacencies, nor to store lots of triangles from the dense part of the graph.

To prove Theorem 3.2 and Corollary 3.2.1, we first prove the following propositions.

Proposition 3.3. *Let $\mathcal{G}$ be a MAG with vertex set $\mathcal{V}$. For a set $W \subseteq \mathcal{V}$, $W \notin \mathcal{S}(\mathcal{G})$ if and only if there are two vertices a, b in W such that we can m-separate them by a set C such that $a, b \notin C$ with $W \subseteq C \cup \{a, b\}$.*

Proof. We prove an equivalent statement of this proposition, that is: $W \in \mathcal{S}(\mathcal{G})$ if and only if for any two vertices a, b in W we cannot m-separate them by a set C such that $a, b \notin C$ with $W \subseteq C \cup \{a, b\}$.

To prove $\Rightarrow$: if $W \in \mathcal{S}(\mathcal{G})$, then there is a nonempty subset $W' \subseteq W$ such that W' is a head and $W \subseteq W' \cup \operatorname{tail}(W')$. Because $\operatorname{tail}(W') \subseteq \operatorname{an}(W')$, we have $W' \cup \operatorname{tail}(W') \subseteq \operatorname{an}(W')$. By definition of the heads and tails, any two vertices a, b in $W \subseteq W' \cup \operatorname{tail}(W')$ are connected by a collider path where all the colliders are in $\operatorname{an}(W') \subseteq \operatorname{an}(C \cup \{a, b\})$. Let d_i, $1 \leq i \leq n$ be intermediate vertices in the path. Now if all of d_i are ancestors of C then this path m-connects a and b. So some of d_i are only ancestors of a, b.

Suppose there exists some $d_i \in \operatorname{an}_{\mathcal{G}}(a) \setminus \operatorname{an}_{\mathcal{G}}(C)$, let d_j be the furthest one on path π from a, so there exists a directed path $\pi' : a \leftarrow \cdots \leftarrow d_j$ such that none of vertices in π' after a is an ancestor of C and hence not in C. If all d_k after d_j belong to $\operatorname{an}_{\mathcal{G}}(C)$ then we find a m-connecting path between a and b: $a \leftarrow \cdots \leftarrow d_j \leftrightarrow \cdots \leftrightarrow d_n \leftarrow ? b$. If not, let d_m be the first one after d_j such that $d_m \in \operatorname{an}_{\mathcal{G}}(b) \setminus \operatorname{an}_{\mathcal{G}}(C)$ then again we find a m-connecting path between a and b: $a \leftarrow \cdots \leftarrow d_j \leftrightarrow \cdots \leftrightarrow d_m \rightarrow \cdots \rightarrow b$.

If all $d_i \notin \operatorname{an}_{\mathcal{G}}(C)$ are ancestors of b then let d_j be the closest one to a in path π which also leads to a m-connecting path between a and b: $a \,?\!\!\rightarrow d_1 \leftrightarrow \cdots \leftrightarrow d_j \rightarrow \cdots \rightarrow b$. Hence in all cases any a, b in W are not m-separated given any $C \supseteq W \setminus \{a, b\}$.

To prove $\Leftarrow$: define W' = $\operatorname{barren}(W)$. We claim that it is a head. Suppose it is not a head, by the definitions of a barren set and a head, W' does not lie in a single district in $\mathcal{G}_{\operatorname{an}(W')}$. Let $D_i \subset W'$ index bidirected-connected components of W' in $\operatorname{an}_{\mathcal{G}}(W')$ where $1 \leq i \leq m$. Clearly by assumption $m > 1$, and now consider D_1 and D_2. For any edge in $\mathcal{G}_{\operatorname{an}(W')}$ which has an endpoint $a \in W'$, it is of the form $a \!\leftarrow\! ?$ by definition of a barren set, so if there is a collider path between D_1 and D_2, it would be a bidirected path which is a contradiction to the definition of D_1 and D_2. This means that any path in $\operatorname{an}_{\mathcal{G}}(W')$ between D_1 and D_2 contains at least one non-collider which is not in W' and hence it is in $\operatorname{an}_{\mathcal{G}}(W') \setminus W'$. Thus for any two vertices in D_1 and D_2, given $\operatorname{an}_{\mathcal{G}}(W') \setminus W'$, they are m-separated in $\operatorname{an}_{\mathcal{G}}(W')$. Since $\operatorname{an}_{\mathcal{G}}(W')$ is ancestral, the m-separation also holds in the whole graph. Thus W' is a head.

By Remark 4.14 in Evans and Richardson (2014), for any head H we have $H \perp_m \operatorname{an}_{\mathcal{G}}(H) \setminus (H \cup \operatorname{tail}(H)) \mid \operatorname{tail}(H)$. Thus if $(W \setminus W')$ is not in $\operatorname{tail}(W')$, we can m-separate a vertex in $(W \setminus W') \setminus \operatorname{tail}(W')$ and a vertex in W' given the remaining vertices in $\operatorname{an}_{\mathcal{G}}(W')$, which is a contradiction. □

Proposition 3.4. *For a MAG $\mathcal{G}$, we have (i) any two vertices a and b are adjacent in $\mathcal{G}$ if and only if $\{a, b\} \in \mathcal{S}(\mathcal{G})$; (ii) for any unshielded triple (a, b, c) in $\mathcal{G}$, $\{a, b, c\} \in \mathcal{S}(\mathcal{G})$ if and only if b is a collider on the triple (a, b, c); (iii) if π forms a discriminating path for b with two end vertices x and y in $\mathcal{G}$ then $\{x, b, y\} \in \mathcal{S}(\mathcal{G})$ if and only if b is a collider on the path π.*

Proof. For (i), by maximality, any two vertices a and b are adjacent in a MAG if and only if we can not m-separate them by a set C, hence by Proposition 3.3 if and only if $\{a, b\} \in \mathcal{S}(\mathcal{G})$.

For (ii), the only nonadjacent pair of vertices are a, c, for any set C that m-seperates them, $b \notin C$ if and only if b is a collider on the triple (a, b, c), hence by Proposition 3.3 if and only if $\{a, b, c\} \in \mathcal{S}(\mathcal{G})$.

For (iii), if x, b are not adjacent, then for any set that m-separates them, y is not in the set, as the path $x \mathrel{?\!\!\rightarrow} q_1 \leftrightarrow \cdots \leftrightarrow q_m \mathrel{\leftarrow\!\!?} b$ would be m-connecting x and b. Since x, y are not adjacent, there exists some set C such that $x \perp_m y \mid C$. From page 11 in Ali et al. (2009), we know that for any such C, $q_i \in C$ for all $i \leq n$ and b is a collider if and only if $b \notin C$, hence by Proposition 3.3 if and only if $\{x, b, y\} \in \mathcal{S}(\mathcal{G})$. □

Now we are able to prove Theorem 3.2 and Corollary 3.2.1

Proof of Theorem 3.2. (⇒) Proposition 3.3 ensures that missing sets in $\mathcal{S}(\mathcal{G})$ are only due to m-separations in graphs. But as Markov equivalence is characterized by m-separations, $\mathcal{S}(\mathcal{G}_1)$ and $\mathcal{S}(\mathcal{G}_2)$ in two equivalent MAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are the same. (⇐) Proposition 3.4 implies that any violation of conditions in Theorem 3.1 result in different $\mathcal{S}(\mathcal{G}_1)$ and $\mathcal{S}(\mathcal{G}_2)$. Hence if $\mathcal{S}(\mathcal{G}_1) = \mathcal{S}(\mathcal{G}_2)$, $\mathcal{G}_1$ is Markov equivalent to $\mathcal{G}_2$. □

Proof of Corollary 3.2.1. (⇒) This follows from Theorem 3.2 and the fact that Markov equivalent MAGs have the same adjacencies. (⇐) This follows from the fact that in the proof the 'if' part of Theorem 3.2, we only consider sets in $\tilde{\mathcal{S}}_3(\mathcal{G}_1)$ and $\tilde{\mathcal{S}}_3(\mathcal{G}_2)$. □

Frydenberg (1990) gives conditions for when two DAGs are equivalent, i.e. if and only if they have the same adjacencies and unshielded colliders. DAGs are a subclass of MAGs so Corollary 3.2.1 also applies to them. When $\mathcal{G}$ is just a DAG, $\tilde{\mathcal{S}}_3(\mathcal{G})$ (and indeed $\mathcal{S}_3(\mathcal{G})$) contains the exact information of $\mathcal{G}$'s adjacencies and unshielded colliders. By Proposition 3.4, $\{a, b\} \in \tilde{\mathcal{S}}_3(\mathcal{G})$ if and only if a, b are adjacent. And a triple is in $\tilde{\mathcal{S}}_3(\mathcal{G})$ if and only if it is an unshielded collider; this is because in DAGs, heads are precisely the individual vertices, and the corresponding tails are their parent sets.

3.3 Projection From ADMGs To MAGs

Richardson and Spirtes (2002) give a projection that projects a DAG $\mathcal{G}$ with latent variables L to a Markov equivalent MAG $\mathcal{G}^m$: (i) every pair of vertices $a, b \in \mathcal{V}$ in $\mathcal{G}$ that are connected by an *inducing path* becomes adjacent in $\mathcal{G}^m$; (ii) an edge connecting a, b in $\mathcal{G}^m$ is oriented as follows: if $a \in \text{an}_\mathcal{G}(b)$ then $a \rightarrow b$; if $b \in \text{an}_\mathcal{G}(a)$ then $b \rightarrow a$; if neither is the case, then $a \leftrightarrow b$. An *inducing path* between a, b is a path such that every collider in the path is in $\text{an}(\{a, b\})$, and every noncollider is in L. Note if we already have an ADMG $\mathcal{G}$, we can apply the projection to $\mathcal{G}$ with no latent variable to construct the corresponding $\mathcal{G}^m$, so an inducing path in this case is just a collider path with every collider in $\text{an}(\{a, b\})$. In addition, the projection preserves ancestral relations from the original graph.

To extend previous theorems to $\mathcal{G}$ we need following lemmas to link $\mathcal{G}$ and $\mathcal{G}^m$.

Lemma 3.5. *If v, w are connected by a collider path π_1 in an ADMG $\mathcal{G}$ then they are connected by a collider path π_2 in $\mathcal{G}^m$ where π_2 uses a subset of the internal vertices of π_1. Also, if π_1 starts with $v \rightarrow$, so does π_2.*

Lemma 3.5 is in analogue to Lemma 23 in Shpitser et al. (2018). Now we show heads and tails are preserved through the projection.

Proposition 3.6. *If $\mathcal{G}$ is an ADMG, $\mathcal{H}(\mathcal{G}) = \mathcal{H}(\mathcal{G}^m)$ and for every $H \in \mathcal{H}(\mathcal{G})$, $\text{tail}_\mathcal{G}(H) = \text{tail}_{\mathcal{G}^m}(H)$.*

Proof. Suppose H is a head in $\mathcal{G}$. Then it is bidirected-connected in $\mathcal{G}_{\text{an}(H)}$, so by Lemma 3.5 each bidirected path connecting vertices in H is preserved as a collider path in $\mathcal{G}^m_{\text{an}(H)}$. Further as the projection preserves ancestral relation and $H = \text{barren}(\text{an}(H))$, each path is bidirected. Hence any head H in $\mathcal{G}$ is a head in $\mathcal{G}^m$. By similar argument, we can see that for a head H in $\mathcal{G}$, any $w \in \text{tail}_\mathcal{G}(H)$ is in $\text{tail}_{\mathcal{G}^m}(H)$.

Suppose H is a head in $\mathcal{G}^m$ so it is bidirected-connected in $\text{an}(H)$ in $\mathcal{G}^m$. But each bidirected edge in $\mathcal{G}^m$ corresponds to a collider path in $\mathcal{G}$ with intermediate colliders in ancestors of endpoints; hence as the projection preserves ancestral relations, the path is bidirected. Therefore H is also a head in $\mathcal{G}$. Note in general for any $v \leftrightarrow w$ in $\mathcal{G}^m$, there is a bidirected path between them in $\mathcal{G}$.

Let $z \in \text{tail}_{\mathcal{G}^m}(H)$ so there is a collider path π between z and $h \in H$ in $\mathcal{G}^m$ ending $\cdots \leftrightarrow h$. We know every bidirected edge in the path π corresponds to a bidirected path in $\text{an}(H)$ in $\mathcal{G}$. If the path π begins with $z \leftrightarrow$ then z is bidirected-connected to h in $\text{an}(H)$ so $z \in \text{tail}_\mathcal{G}(H)$. If the path π begins with $z \rightarrow w_1$ then in $\mathcal{G}$ we have a collider path between z and w_1 in $\text{an}(H)$, which ends with $\leftrightarrow w_1$. Thus z is also in $\text{tail}_\mathcal{G}(H)$. □

Definitions of heads and tails are closely related to the projection of ADMGs. The next lemma allows us to project an ADMG to a Markov equivalent MAG in polynomial time. The algorithm is shown in next section. Let $\mathcal{G}$ be a ADMG and $\mathcal{G}^m$ be its projected MAG.

Lemma 3.7. *Let v, w be two vertices then (i) $v \rightarrow w$ in $\mathcal{G}^m$ if and only if $v \in \text{tail}_\mathcal{G}(w)$ and (ii) $v \leftrightarrow w$ in $\mathcal{G}^m$ if and only if $\{v, w\} \in \mathcal{H}(\mathcal{G})$.*

Since there is at most one edge between any two vertices in a MAG, if we know the tails of every vertex in $\mathcal{G}^d$ and every head of size 2, this is sufficient to construct $\mathcal{G}^m$.

Consider Figure 2(i), this is an ADMG but not a MAG. Tails of 1, 2, 3, 4 are $\{4\}$, $\{3\}$, $\emptyset$, $\emptyset$, respectively. Heads of size 2 are $\{1,2\}$, $\{1,3\}$, $\{2,4\}$, $\{3,4\}$, hence a Markov equivalent MAG of Figure 2(i) preserves all the original edges and adds one edge $1 \leftrightarrow 2$.

3.4 Markov Equivalence Of ADMGs

We now show that Theorem 3.2 and Corollary 3.2.1 can be extended to ADMGs. Note that in general two Markov equivalent ADMGs do not necessarily have the same adjacencies defined with respect to edges; thus we need to redefine adjacencies in terms of m-separations.

Definition 3.1. For a ADMG $\mathcal{G}$ and two vertices v, w in $\mathcal{G}$, v and w are *adjacent* if and only if there is no set C such that $v \perp_m w \mid C$ with $v, w \notin C$.

Two vertices that are connected by an edge are clearly adjacent, we are excluding pairs that do not share any edges and yet have no conditional independence. In maximal graphs, these two definitions are equivalent.

Theorem 3.8. *For two ADMGs $\mathcal{G}_1$ and $\mathcal{G}_2$, they are Markov equivalent if and only $\mathcal{S}(\mathcal{G}_1) = \mathcal{S}(\mathcal{G}_2)$.*

Corollary 3.8.1. *Two ADMGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if $\mathcal{S}_3(\mathcal{G}_1) = \mathcal{S}_3(\mathcal{G}_2)$, and this occurs if and only if $\tilde{\mathcal{S}}_3(\mathcal{G}_1) = \tilde{\mathcal{S}}_3(\mathcal{G}_2)$.*

4 ALGORITHM

In this section, n, e denote number of vertices and total edges, respectively.

4.1 MAGs

We assume that $n = O(e)$, since otherwise the graph will be disconnected. Firstly, we propose an algorithm to identify $\tilde{\mathcal{S}}_3(\mathcal{G})$ of a given MAG $\mathcal{G}$ and show that it runs in polynomial time ($O(ne^2)$). To test equivalence of two MAGs, it is sufficient to compare their $\tilde{\mathcal{S}}_3$, by Corollary 3.2.1. Vertices are assumed to be in topological order. If not, this can be achieved with an $O(n+e)$ sort. We assume we have access to $\mathrm{pa}_{\mathcal{G}}(v)$ and $\mathrm{sib}_{\mathcal{G}}(v)$ for each $v \in \mathcal{V}$.

Let $A_1(\mathcal{G})$ denote the output of Algorithm 1 when applied to a MAG, $\mathcal{G}$.

Proposition 4.1. *For a MAG $\mathcal{G}$, $A_1(\mathcal{G}) = \tilde{\mathcal{S}}_3(\mathcal{G})$.*

4.2 Complexity Of Algorithm 1

The first loop from line 2 to line 7 runs at most $O(e^2)$ times as the worst case is that one vertex have all others as its parents. There are at most e bidirected edges so the second loop from line 8 to line 17 repeats at most e times. There are three esrial tasks inside the second loop. The first one is line 10 which obtains the tails of $\{v, w\}$. The computation of obtaining tails given parents is $O(n+e)$. The second task, i.e. the first subloop from line 11 to line 12, is carried at most $n-2$ times as the size of each tail is at most $n-2$. For the third task from line 13 to line 17, there are at most $n-2$ potential candidates for the third member, and obtaining the district costs $O(n+e)$. Thus the overall complexity of Algorithm 1 is $O(e^2 + e((n+e) + n + n(n+e))) = O(ne^2)$.

Note that the number of potential candidates for third member of heads of size 3 depends on sizes of districts. If the number is high then it means districts are large so there are at least as many bidirected edges as potential candidates, so if the graph is sparse we can use e to represent the number of candidates instead of n when computing complexity. There are most $O(e^2)$ sets in $\tilde{\mathcal{S}}_3(\mathcal{G})$, and some graphs achieve this bound, for example, a DAG where one vertex have all others as its parents.

To test ordinary Markov equivalence of two MAGs, it is sufficient to compare their output of Algorithm 1 after a sort of order $O(e^2 \log e^2) = O(e^2 \log e)$. Note that $\log e = O(\log n)$, therefore the complexity of verifying Markov equivalence between two MAGs is still $O(ne^2)$. Thus our algorithm is faster than the one proposed by Ali et al. (2009), which is only $O(ne^4)$.

4.3 ADMGs

Algorithm 2 converts an ADMG $\mathcal{G}$ to a Markov equivalent MAG $\mathcal{G}^m$, as proven by Lemma 3.7. To test Markov equivalence between two ADMGs, it is sufficient to put their equivalent MAGs in Algorithm 1 to obtain the corresponding sets $\tilde{\mathcal{S}}_3$ and compare the sets.

4.4 Complexity Of Algorithm 2

For the first loop from line 3 to 5, it costs $O(n(n+e))$ since there are n vertices and it takes $O(n+e)$ to obtain a district. The second loop from line 6 to 9 is at $O(n^2(n+e))$. Thus the overall complexity is $O(n(n+e) + n^2(n+e)) = O(n^3 + n^2 e) = O(n^2 e)$. The total cost for identifying Markov equivalence between two ADMGs is therefore $O(ne^2)$.

```
Input:   A MAG $\mathcal{G}(\mathcal{V}, \mathcal{E})$
Output:  $\tilde{\mathcal{S}}_3(\mathcal{G})$
$S = \emptyset$;
for each $v \in \mathcal{V}$:
    obtain $\mathrm{an}_{\mathcal{G}}(v) = \{v\} \cup \mathrm{an}_{\mathcal{G}}(\mathrm{pa}_{\mathcal{G}}(v))$
    for each $w \in \mathrm{pa}_{\mathcal{G}}(v)$:
        $S = S \cup \{v, w\}$;
    for each $z, w \in \mathrm{pa}_{\mathcal{G}}(v)$ with $z \neq w$ and $z$ is not adjacent to $w$:
        $S = S \cup \{v, w, z\}$;
for each $v \leftrightarrow w$:
    $S = S \cup \{v, w\}$;
    $\mathrm{tail}(\{v, w\}) = \mathrm{dis}_{\mathrm{an}(\{v,w\})}(v) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_{\mathrm{an}(\{v,w\})}(v)) \setminus \{v, w\}$;
    for each $z \in \mathrm{tail}(\{v, w\})$ with $z$ not adjacent to both $v$ and $w$:
        $S = S \cup \{v, w, z\}$;
    for each $z \in \mathrm{sib}_{\mathcal{G}}(\mathrm{an}_{\mathcal{G}}(\{v, w\})) \cap \mathrm{dis}_{\mathcal{G}}(v) \setminus (\mathrm{an}_{\mathcal{G}}(\{v, w\}) \cup \mathrm{de}_{\mathcal{G}}(\{v, w\})$
            and not adjacent to both $v$ and $w$:
        obtain $\mathrm{dis}_{\mathrm{an}(\{v,w,z\})}(v)$;
        if $z \in \mathrm{dis}_{\mathrm{an}(\{v,w,z\})}(v)$:
        $S = S \cup \{v, w, z\}$;
return $S$
```

Table 4: Algorithm 1: obtain $\tilde{\mathcal{S}}_3(\mathcal{G})$ for a MAG $\mathcal{G}$

```
Input:   An ADMG $\mathcal{G}(\mathcal{V}, \mathcal{E})$
Output:  A Markov equivalent MAG $\mathcal{G}^m(\mathcal{V}, \mathcal{E}^m)$
Start with $\mathcal{G}^m$ that have the same vertices as $\mathcal{G}$ but no adjacencies;
for each $v \in \mathcal{V}$:
    obtain $\mathrm{an}_{\mathcal{G}}(v) = \{v\} \cup \mathrm{an}_{\mathcal{G}}(\mathrm{pa}_{\mathcal{G}}(v))$
    $\mathrm{tail}(v) = \mathrm{dis}_{\mathrm{an}(v)}(v) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_{\mathrm{an}(v)}(v)) \setminus \{v\}$
    add $w \rightarrow v \in \mathcal{E}^m$ for each $w \in \mathrm{tail}(v)$;
for each $v, w \in \mathcal{V}$ with no ancestral relation and in the same district:
    obtain $\mathrm{dis}_{\mathrm{an}(\{v,w\})}(v)$;
    if $w \in \mathrm{dis}_{\mathrm{an}(\{v,w\})}(v)$:
        add $v \leftrightarrow w \in \mathcal{E}^m$;
return $\mathcal{G}^m$
```

Table 5: Algorithm 2: obtain a MAG $\mathcal{G}^m$ for an ADMG $\mathcal{G}$

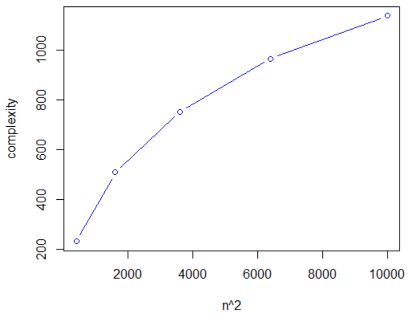


Figure 4: Empirical complexity against n^2

4.5 Comparison To Previous Algorithms

Among previous characterizations of MAGs, only Ali et al. (2009) provide a polynomial time algorithm to verify Markov equivalence. They consider all triples in a discriminating path; in order to do this, they iterate through (up to) $n-2$ levels; at each level they consider all remaining colliders ($O(e^2)$) and then check each set of reachable edges ($O(e^2)$). Conversely, we ignore any triples for which all three adjacencies are present (since they will trivially always be present).

In addition to the reduction in complexity, if we modify Algorithm 1 to compute $\mathcal{S}_3(\mathcal{G})$, the output contains more information. By Proposition 3.3, a set $\{a_1, a_2, a_3\}$ is missing from $\mathcal{S}_3$ if and only if there is a corresponding m-separation between (say) a_1, a_2 conditional on a set that includes a_3. Thus we can view the parametrizing set as a summary of independence information in the graph. This is a novel perspective compared to previous theorems, which characterize graphs by structures like minimal collider paths or colliders with order, and do not have a straightforward connection to conditional independence.

4.6 Empirical Complexity

An experiment on random graphs shows that empirical complexity of Algorithm 1 is at $O(e^2)$ for many sparse graphs ($e = O(n)$). One random graph (ADMG) is generated in the following way. We first fix a topological ordering and the total number of edges ($e = 3n$). Then two vertices become adjacent with uniform probability. Once skeleton is determined, an edge is independently either directed or bidirected with $p = 0.5$. For each $n = 20, 40, 60, 80, 100$, we generate $N = 250$ random graphs then average the empirical complexity. Figure 4 is the empirical complexity against n^2.

Suppose directed edges are added independently with probability r/n according to a predetermined topological order, where n is the number of vertices and $r \in \mathbb{R}^+$ is constant. The following proposition bounds the size of the ancestor sets in our sparse random graphs. In particular, the largest average number of ancestors is at most e^r.

Proposition 4.2. *Let A_i be the number of ancestors of the vertex i. Then*

$$\mathbb{E}A_i = \left(1 + \frac{r}{n}\right)^{i-1}.$$

In particular,

$$\mathbb{E}A_n = \left(1 + \frac{r}{n}\right)^{n-1} \longrightarrow e^r.$$

Markov's inequality gives us an easy corollary.

Corollary 4.2.1. $\mathbb{P}(A_i \geq k) \leq e^r/k$ *for any* $1 \leq i \leq n$ *and* $k \geq 1$.

Now it is straightforward to show that for sparse graphs, the complexity will be $O(e^2)$. This is because the main contribution of the complexity comes from counting heads of size 3. By bounding the sizes of ancestor sets, line 15 will run in constant time $O(1)$ instead of $O(n+e)$. Thus the overall complexity for sparse graphs is at $O(e^2 + e((n+e)+n+n)) = O(e^2)$.

Here is an example for which the upper bound of complexity of Algorithm 1 is reached. Consider the graph in Figure 5. For every i and j, $\{v_i, w, z_j\}$ forms a head of size 3. If N, M, L are at $O(n)$ then the cost for identifying all these heads is at $O(ne^2)$.

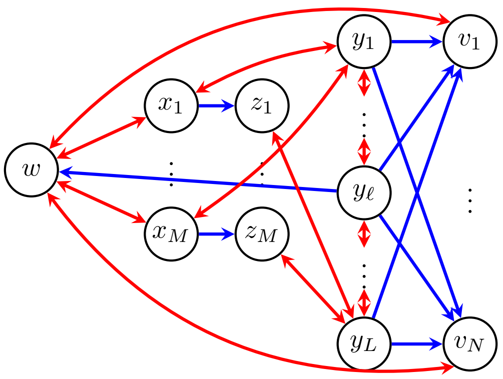


Figure 5: A sequence of graphs in which the maximum complexity is achieved by Algorithm 1. Note that y_1 is connected by a bidirected edge to every x_i, and y_L to every z_i.

References

R. A. Ali, T. S. Richardson, and P. Spirtes. Markov equivalence for ancestral graphs. *Annals of Statistics*, 37(5B):2808–2837, 10 2009.

D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, pages 294–321, 2012.

R. J. Evans. Margins of discrete Bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, Dec 2018. ISSN 0090-5364. doi: 10.1214/17-aos1631.

R. J. Evans and T. S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, 42(4):1452–1482, 2014.

M. Frydenberg. The chain graph markov property. *Scandinavian Journal of Statistics*, pages 333–353, 1990.

J. Pearl. *Causality*. Cambridge University Press, second edition, 2009.

T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30 (1):145–157, 2003.

T. S. Richardson. A factorization criterion for acyclic directed mixed graphs. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 462–470, 01 2009.

T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 08 2002.

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2017.

I. Shpitser, R. J. Evans, and T. S. Richardson. Acyclic linear SEMs obey the nested Markov property. *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

P. Spirtes and T. S. Richardson. A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias, 1997.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

N. Wermuth. Probability distributions with summary graph structure. *Bernoulli*, 17(3):845–879, 08 2011.

H. Zhao, Z. Zheng, and B. Liu. On the Markov equivalence of maximal ancestral graphs. *Science in China Series A: Mathematics*, 48(4):548–562, Apr 2005.