# STRUCTURAL MARKOV GRAPH LAWS FOR BAYESIAN MODEL UNCERTAINTY

BY SIMON BYRNE[1] AND A. PHILIP DAWID

*University College London and University of Cambridge*

This paper considers the problem of defining distributions over graphical structures. We propose an extension of the hyper Markov properties of Dawid and Lauritzen [*Ann. Statist.* **21** (1993) 1272–1317], which we term *structural Markov properties*, for both undirected decomposable and directed acyclic graphs, which requires that the structure of distinct components of the graph be conditionally independent given the existence of a separating component. This allows the analysis and comparison of multiple graphical structures, while being able to take advantage of the common conditional independence constraints. Moreover, we show that these properties characterise exponential families, which form conjugate priors under sampling from compatible Markov distributions.

**1. Introduction.** A graphical model consists of a graph and a probability distribution that satisfies a *Markov property* of the graph, being a set of conditional independence constraints encoded by the graph. Such models arise naturally in many statistical problems, such as contingency table analysis and covariance estimation.

Dawid and Lauritzen (1993) consider distributions over these distributions, which they term *laws* to emphasise the distinction from the underlying sampling distribution. Laws arise primarily in two contexts: as sampling distributions of estimators and as prior and posterior distributions in Bayesian analyses. Specifically, Dawid and Lauritzen (1993) focus on hyper Markov laws that exhibit conditional independence properties analogous to those of the distributions of the model. By exploiting such laws, it is possible to perform certain inferential tasks locally; for instance, posterior laws can be calculated from subsets of the data pertaining to the parameters of interest.

Although other types of graphical model exist, we restrict ourselves to undirected decomposable graphs and directed acyclic graphs, which exhibit the special property that their Markov distributions can be constructed in a recursive fashion by taking *Markov combinations* of smaller components. In the case of undirected decomposable graphs, for any decomposition $(A, B)$ of the graph $\mathcal{G}$, a Markov

---

*Key words and phrases.* Graphical models, structural estimation, hyper Markov laws, structural Markov laws.

distribution is uniquely determined by the marginal distributions over *A* and *B* [Dawid and Lauritzen (1993), Lemma 2.5]. By a recursion argument, this is equivalent to specifying marginal distributions on cliques. A similar construction can be derived for directed acyclic graphs: the distribution of each vertex conditional on its parent set can be chosen arbitrarily, and the set of such distributions determines the joint distribution. As we demonstrate in Section 5, this property can also be characterised in terms of a partitioning based on *ancestral sets*.

It is this partitioning that makes the notion of hyper Markov laws possible. In essence, these are laws for which the partitioned distributions exhibit conditional independence properties analogous to those of the underlying distributions. In the case of undirected decomposable graphs, a law £ for $\tilde{\theta}$ over $\mathfrak{P}(\mathcal{G})$, the set of Markov distributions with respect to $\mathcal{G}$, is (*weak*) *hyper Markov* if for any decomposition $(A, B)$,

$$(1.1) \qquad\qquad \tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_B | \tilde{\theta}_{A \cap B} \qquad [\text{£}].$$

Weak hyper Markov laws arise naturally as sampling distributions of maximum likelihood estimators of graphical models [Dawid and Lauritzen (1993), Theorem 4.22]. A more specific class of laws are those that satisfy the *strong hyper Markov property*, where for any decomposition $(A, B)$,

$$(1.2) \qquad\qquad \tilde{\theta}_{A|B} \perp\!\!\!\perp \tilde{\theta}_B \qquad [\text{£}].$$

When used as prior laws in a Bayesian analysis, strong hyper Markov laws allow for local posterior updating, in that the posterior law of clique marginal distributions only depends on the data in the clique [Dawid and Lauritzen (1993), Corollary 5.5].

However, hyper Markov laws only apply to individual graphs: if the structure of the graph itself is unknown, then a full Bayesian analysis requires a prior distribution over graphical structures, which we term a *graph law*. Very little information is available to guide the choice of such priors, with a typical choice being a simple uniform or constrained Erdős–Rényi prior.

The aim of this paper is to extend the hyper Markov concept to the structure of the graph itself. We study graph laws that exhibit similar conditional independence structure, termed *structural Markov properties*. These properties exhibit analogous local inference properties, and under minor assumptions, characterise exponential families, which serve as conjugate families to families of compatible Markov distributions and hyper Markov laws.

The outline of the paper is as follows. In Section 2 we introduce the terms and notation used in the paper, in particular the notion of a semi-graphoid to define what we mean by structure. Section 3 develops the notion of a structural Markov property and characterises such laws for undirected decomposable graphs. Section 4 briefly develops a similar notion for directed graphs consistent with a fixed ordering. In Section 5 we consider the notion of Markov equivalence of directed

acyclic graphs, and extend the structural Markov property to these equivalence classes. Finally, in Section 6 we discuss some properties, computational considerations, and future directions.

**2. Background.** Much of the terminology in this paper is standard in the graphical modelling literature. For this we refer the reader to texts such as Lauritzen (1996) or Cowell et al. (2007). For clarity and consistency, the following presents some specific terms and notation used in this paper.

2.1. *Graphs.* A *graph* $\mathcal{G}$ consists of a set of *vertices* $\mathcal{V}(\mathcal{G})$ and a set of *edges* $\mathcal{E}(\mathcal{G})$ of pairs of vertices. In the case of undirected graphs, $\mathcal{E}(\mathcal{G})$ will be a set of unordered pairs of vertices $\{u, v\}$; in the directed case it will be a set of ordered pairs $(u, v)$, denoting an arrow from $u$ to $v$, of which $v$ is termed the *head*. For any subset $A \subseteq \mathcal{V}(\mathcal{G})$, $\mathcal{G}_A$ will denote the induced subgraph with vertex set $A$. A graph is *complete* if there exists an edge between every pair of vertices, and *sparse* if no edges are present (the graph with empty vertex set is both complete and sparse).

We focus on two particular classes of graphs.

2.1.1. *Undirected decomposable graphs.* A *path* in an undirected graph $\mathcal{G}$ is a sequence of vertices $v_0, v_1, \ldots, v_k$ such that $\{v_i, v_{i+1}\} \in \mathcal{E}(\mathcal{G})$, in which case we can say $v_0$ is *connected* to $v_k$. Sets $A, B \subseteq \mathcal{V}(\mathcal{G})$ are *separated* by $S \subseteq \mathcal{V}(\mathcal{G})$ if every path starting at an element of $A$ and ending at an element of $B$ contains an element of $S$.

A pair of sets $(A, B)$ is a *covering pair* of $\mathcal{G}$ if $A \cup B = \mathcal{V}(\mathcal{G})$. A covering pair is a *decomposition* if $\mathcal{G}_{A \cap B}$ is complete and $A$ and $B$ are separated by $A \cap B$ in $\mathcal{G}$. A decomposition is *proper* if both $A$ and $B$ are strict subsets of $\mathcal{V}(\mathcal{G})$. For any set of undirected graphs $\mathfrak{F}$, define $\mathfrak{F}(A, B)$ to be the set of $\mathcal{G} \in \mathfrak{F}$ for which $(A, B)$ is a decomposition.

A graph is *decomposable* if it can be recursively decomposed into complete subgraphs. An equivalent condition is that the graph is *chordal*, in that there exists no set which induces a cycle graph of length 4 or greater. Throughout the paper we will take $V$ to be a fixed, finite set, and define $\mathfrak{U}$ to be the set of undirected decomposable graphs (UDGs) with vertex set $V$.

The maximal sets inducing complete subgraphs are termed *cliques*, the set of which is denoted by $\mathrm{cl}(\mathcal{G})$. For any decomposable graph it is possible to construct a *junction tree* of the cliques. The intersections of neighbouring cliques in a junction tree are termed (*clique*) *separators*, the set of which is denoted by $\mathrm{sep}(\mathcal{G})$. The *multiplicity* of a separator is the number of times it appears in the junction tree. The cliques, separators, and their multiplicities are invariants of the graph.

An undirected graph $\mathcal{G}$ is *collapsible* onto $A \subseteq \mathcal{V}(\mathcal{G})$ if each connected component $C$ of $\mathcal{G}_{V \setminus A}$ has a boundary $B = \{u : \{u, v\} \in \mathcal{E}(\mathcal{G}), v \in C, u \notin C\}$ which induces a complete subgraph. Note that if $(A, B)$ is a decomposition of $\mathcal{G}$, then $\mathcal{G}$ is collapsible onto both $A$ and $B$.

2.1.2. *Directed acyclic graphs.* A directed graph $\mathcal{G}$ is *acyclic* if there exists a *compatible well-ordering* $\prec$ on $\mathcal{V}(\mathcal{G})$, that is, such that $u \prec v$ for all $(u, v) \in \mathcal{E}(\mathcal{G})$. For any such $\prec$, the *predecessors* of a vertex $v$ is the set $\mathrm{pr}_{\prec}(v) = \{u \in \mathcal{V}(\mathcal{G}) : u \prec v\}$. The set of directed acyclic graphs (DAGs) on $V$ will be denoted by $\mathfrak{D}$, and the subset for which $\prec$ is a compatible well-ordering is denoted by $\mathfrak{D}^{\prec}$.

A vertex $u$ is a *parent* of $v$ if $(u, v) \in \mathcal{E}(\mathcal{G})$. The set of parents of $v$ is denoted by $\mathrm{pa}_{\mathcal{G}}(v)$. Conversely, $u$ is a *child* of $v$. A set $A \subseteq \mathcal{V}(\mathcal{G})$ is *ancestral* in $\mathcal{G}$ if $v \in A \Rightarrow \mathrm{pa}_{\mathcal{G}}(v) \subseteq A$. The minimal ancestral set containing $B \subseteq \mathcal{V}(\mathcal{G})$ is denoted by $\mathrm{an}_{\mathcal{G}}(B)$.

The *skeleton* of a directed graph $\mathcal{G}$ is the undirected graph obtained by replacing all the directed edges with undirected edges. The *moral graph* of $\mathcal{G}$, denoted by $\mathcal{G}^{\mathrm{M}}$, is the skeleton of the graph obtained by adding (if necessary) an edge between each pair of vertices having a common child.

2.2. *Distributions and laws.* Let $X = (X_v)_{v \in V}$ be a random vector on some product space $\prod_{v \in V} \mathcal{X}_v$, with distribution denoted by $P$ or $\theta$. A *model* is a family of distributions $\Theta$ for $X$.

Following Dawid and Lauritzen (1993), a distribution over $\Theta$ will be termed a *law* and denoted by £. A random distribution following such a law will be denoted by $\tilde{\theta}$.

For any $A \subseteq V$, $X_A$ will denote the subvector $(X_v)_{v \in V}$, with $P_A$ or $\theta_A$ denoting its marginal distribution. The marginal law of $\tilde{\theta}_A$ will be denoted by £$_A$. Furthermore, for any pair $A, B \subseteq V$, we can denote by $\theta_{A|B}$ the collection of conditional distributions of $X_A | X_B$ under $\theta$, and by £$_{A|B}$ the induced law of $\tilde{\theta}_{A|B}$ under £. We will use $\simeq$ to indicate the existence of a bijective function; for instance, we can write $(\theta_A, \theta_{V|A}) \simeq \theta$ for any $A \subseteq V$.

2.3. *Semi-graphoids.* When discussing the "structure" of a graphical model, many authors use this term to refer to the graph itself. In particular, when they talk of "estimating the structure," they mean inferring the presence or absence of individual edges of the graph.

In this paper, we take the view that "structure" refers to a set of conditional independence properties, and that a graph is merely a representation of this structure. This distinction is an important one: it implies that graphs that encode the same set of conditional independence statements must be treated as identical, leading to the notion of *Markov equivalence*. A more subtle but even more important point is that when investigating properties such as decompositions or ancestral sets, we are, effectively, looking at properties of sets of conditional independencies.

To make this more concrete, we use the notion of a *semi-graphoid*, a special case of a *separoid* [Dawid (2001a)], to describe the abstract properties of conditional independence.

DEFINITION 2.1. Given a finite set $V$, a *semi-graphoid* is a set $M$ of triples of the form $\langle A, B | C \rangle$, where $A, B, C \subseteq V$, satisfying the properties:

S0  for all $A, B \subseteq V$, $\langle A, B|A \rangle \in M$;
S1  if $\langle A, B|C \rangle \in M$, then $\langle B, A|C \rangle \in M$;
S2  if $\langle A, B|C \rangle \in M$ and $D \subseteq A$, then $\langle D, B|C \rangle \in M$;
S3  if $\langle A, B|C \rangle \in M$ and $D \subseteq A$, then $\langle A, B|C \cup D \rangle \in M$;
S4  if $\langle A, B|C \rangle \in M$ and $\langle A, D|B \cup C \rangle \in M$, then $\langle A, B \cup D|C \rangle \in M$.

These properties match the well-established properties of conditional independence [Dawid (1979)].

We can define the semi-graphoid of a graph as the set of triples encoding its global Markov property: the semi-graphoid of an undirected graph $\mathcal{G}$ is

$$(2.1) \qquad \mathfrak{M}(\mathcal{G}) = \big\{ \langle A, B|C \rangle : A \text{ and } B \text{ are separated by } C \text{ in } \mathcal{G} \big\},$$

and the semi-graphoid of a directed acyclic graph $\mathcal{G}$ is the set

$$(2.2) \quad \mathfrak{M}(\mathcal{G}) = \big\{ \langle A, B|C \rangle : A \text{ and } B \text{ are separated by } C \text{ in } \mathcal{G}^{\mathrm{M}}_{\mathrm{an}(A \cup B \cup C)} \big\}.$$

We say that a joint distribution $P$ for $X = (X_v)_{v \in V}$ is *Markov* with respect to a semi-graphoid $M$ if

$$\langle A, B|C \rangle \in M \quad \Rightarrow \quad X_A \perp\!\!\!\perp X_B | X_C \qquad [P].$$

That is, a distribution is Markov with respect to a graph if it is Markov with respect to the semi-graphoid of the graph. We write $\mathfrak{P}(\mathcal{G})$ or $\mathfrak{P}(M)$ to be the set of distributions that are Markov with respect to $\mathcal{G}$ or $M$.

Similarly, a law £ is *weak hyper Markov* with respect to the semi-graphoid if

$$\langle A, B|C \rangle \in M \quad \Rightarrow \quad \tilde{\theta}_{A \cup C} \perp\!\!\!\perp \tilde{\theta}_{B \cup C} | \tilde{\theta}_C \qquad [£].$$

However, the strong hyper Markov laws cannot be directly characterised in terms of the semi graphoid.

Semi-graphoids have a natural projection operation: for any set $U \subseteq V$, we can define the projection onto $U$ of a semi-graphoid $M$ on $V$ to be

$$M_U = \big\{ \langle A, B|C \rangle \in M : A, B, C \subseteq U \big\}.$$

Under certain conditions, this can match the natural projection operation, the induced subgraph, of the underlying graph. For undirected graphs, $[\mathfrak{M}(\mathcal{G})]_U = \mathfrak{M}(G_U)$ if and only if $\mathcal{G}$ is collapsible onto $U$ [Asmussen and Edwards (1983), Corollary 2.5]. For directed acyclic graphs, we have the weaker sufficient condition that if $A$ is ancestral in $\mathcal{G}$, then $[\mathfrak{M}(\mathcal{G})]_A = \mathfrak{M}(\mathcal{G}_A)$.

**3. Undirected structural Markov property.** We now extend the hyper Markov framework to the case where the graph itself is regarded as a random object $\tilde{\mathcal{G}}$, taking values in the set of undirected decomposable graphs with vertex set $V$; equivalently, $\tilde{\mathcal{G}}$ can be thought of as a random vector of length $\binom{|V|}{2}$ indicating the presence or absence of individual edges. As the graph is a parameter of

the model, we term its distribution a *graph law*, denoted by $\mathfrak{G}(\tilde{\mathcal{G}})$. Our aim is to identify and characterise hyper Markov-type properties for $\tilde{\mathcal{G}}$.

Hyper Markov laws are motivated by the property that graph decompositions allow one to decompose Markov distributions into separate components. For a fixed graph $\mathcal{G} \in \mathfrak{U}(A, B)$, then any Markov distribution $\theta \in \mathfrak{P}(\mathcal{G})$ is uniquely characterised by its marginals $\theta_A$ and $\theta_B$, taking values in $\mathfrak{P}(\mathcal{G}_A)$ and $\mathfrak{P}(\mathcal{G}_B)$, respectively [Dawid and Lauritzen (1993), Lemma 2.5]. Moreover, these can be chosen arbitrarily, subject only to the constraint $(\theta_A)_{A \cap B} = (\theta_B)_{A \cap B}$. Hyper Markov laws are derived by imposing probabilistic conditional independence on this natural separation.

In a similar manner, graphs themselves can be characterised by their projections onto each part of a decomposition.

PROPOSITION 3.1. *Let $\mathcal{H}$ and $\mathcal{J}$ be decomposable graphs with vertex set $A$ and $B$, respectively, such that both $\mathcal{H}_{A \cap B}$ and $\mathcal{J}_{A \cap B}$ are complete. Then the graph $\mathcal{G}$ with $\mathcal{E}(\mathcal{G}) = \mathcal{E}(\mathcal{H}) \cup \mathcal{E}(\mathcal{J})$ is the unique decomposable graph on $A \cup B$ such that*:

   (i)  $\mathcal{G}_A = \mathcal{H}$,
  (ii)  $\mathcal{G}_B = \mathcal{J}$, *and*
 (iii)  $(A, B)$ *is a decomposition of* $\mathcal{G}$.

PROOF. To satisfy (i) and (ii), the edge set must contain $\mathcal{E}(\mathcal{H}) \cup \mathcal{E}(\mathcal{J})$. It cannot contain any additional edges $\{u, v\}$, as this would violate: (i), if $\{u, v\} \subseteq A$; (ii), if $\{u, v\} \subseteq B$; or (iii), if $u \in A \setminus B$ and $v \in B \setminus A$. □

In other words, a graph $\mathcal{G} \in \mathfrak{U}(A, B)$ is characterised by $\mathcal{G}_A$ and $\mathcal{G}_B$, and $\mathcal{G}_A$ and $\mathcal{G}_B$ can be chosen independently. Moreover, this also decomposes the semi-graphoid, as $\mathcal{G}$ is collapsible onto both $A$ and $B$.

We define the graph $G$ resulting from Proposition 3.1 to be the *graph product* of $\mathcal{H}$ and $\mathcal{J}$, denoted by

$$\mathcal{G} = \mathcal{H} \otimes \mathcal{J}.$$

REMARK. Although we only use the graph product when $\mathcal{G}_{A \cap B}$ is complete, the definition can be extended to the case where $\mathcal{H}$ and $\mathcal{J}$ are collapsible onto $A \cap B$.

For a graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ over $\mathfrak{U}(A, B)$, a straightforward way to extend the hyper Markov property in this case would be to require that

$$(3.1) \qquad\qquad \tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B | \tilde{\mathcal{G}}_{A \cap B} \qquad [\mathfrak{G}].$$

Note that in this case the term $\tilde{\mathcal{G}}_{A \cap B}$ is redundant: if $(A, B)$ is a decomposition of $\mathcal{G}$, then $\mathcal{G}_{A \cap B}$ must be complete, and so we are left with a statement of marginal independence $\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B$.
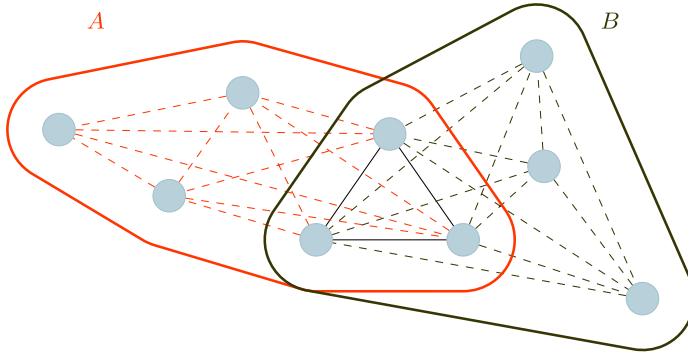
FIG. 1.   *A representation of the structural Markov property for undirected graphs. Conditional on* $(A, B)$ *being a decomposition, the existence of the remaining edges in* $\tilde{\mathcal{G}}_A$ (- - -) *is independent of those in* $\tilde{\mathcal{G}}_B$ (- - -).

A more general question remains: how might this property be extended to a graph law over all undirected graphs? A seemingly simple requirement is that (3.1) should hold whenever a decomposition exists. This motivates the following definition.

DEFINITION 3.1 (Structural Markov property).    A graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ over $\mathfrak{U}$ is *structurally Markov* if for any covering pair $(A, B)$ where $\mathfrak{G}(\mathfrak{U}(A, B)) > 0$, then $\tilde{\mathcal{G}}_A$ is independent of $\tilde{\mathcal{G}}_B$, conditional on $(A, B)$ being a decomposition of $\tilde{\mathcal{G}}$. This is written as

$$(3.2) \qquad\qquad \tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B | \{ \tilde{\mathcal{G}} \in \mathfrak{U}(A, B) \} \qquad [\mathfrak{G}].$$

In essence, the structural Markov property states that the structures of different induced subgraphs are conditionally independent given that they are in separate parts of a decomposition. See Figure 1 for a depiction.

The use of braces on the right-hand side of (3.2) is to emphasise that the conditional independence is defined with respect to the *event* $\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$, and not a random variable as in the Markov and hyper Markov properties. In other words, we do not assume $\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B | \{ \tilde{\mathcal{G}} \notin \mathfrak{U}(A, B) \}$.

3.1. *Products and projections.*   The graph product operation provides a very useful characterisation of the structural Markov property.

PROPOSITION 3.2.    *A graph law* $\mathfrak{G}$ *is structurally Markov if and only if for every covering pair* $(A, B)$, *and every* $\mathcal{G}, \mathcal{G}' \in \mathfrak{U}(A, B)$,

$$(3.3) \qquad\qquad \pi(\mathcal{G})\pi(\mathcal{G}') = \pi(\mathcal{G}_A \otimes \mathcal{G}'_B)\pi(\mathcal{G}'_A \otimes \mathcal{G}_B),$$

*where* $\pi$ *is the density of* $\mathfrak{G}$ *with respect to the counting measure on* $\mathfrak{U}$.

PROOF.   By Proposition 3.1, both $\mathcal{G}_A \otimes \mathcal{G}'_B, \mathcal{G}'_A \otimes \mathcal{G}_B \in \mathfrak{U}(A, B)$, and so if $\mathfrak{G}(\mathfrak{U}(A, B)) = 0$, the statement is trivial. Otherwise, the conditional density of a structural Markov law is of the form

$$\pi\big(\mathcal{G}|\{\mathcal{G} \in \mathfrak{U}(A, B)\}\big) = \pi\big(\mathcal{G}_A|\mathfrak{U}(A, B)\big)\pi\big(\mathcal{G}_B|\mathfrak{U}(A, B)\big).$$

The result follows by substitution into (3.3).   □

The structural Markov property has an inherent divisibility property that arises on subgraphs induced by decompositions. First we require the following lemma.

LEMMA 3.3.   *Let $(A, B)$ be a decomposition of a graph $\mathcal{G}$, and $(S, T)$ a covering pair of $A$ with $A \cap B \subseteq T$. Then $(S, T)$ is a decomposition of $\mathcal{G}_A$ if and only if $(S, T \cup B)$ is a decomposition of $\mathcal{G}$.*

PROOF.   Recall that $W$ separates $U$ and $V$ in $\mathcal{G}$ if and only if $\langle U, V|W \rangle \in \mathfrak{M}(\mathcal{G})$. Since $(S, T)$ is a covering pair of $A$, $\langle S \cup T, B|S \cap B \rangle \in \mathfrak{M}(\mathcal{G})$, and hence $\langle S, B|T \rangle \in \mathfrak{M}(\mathcal{G})$. If $(S, T)$ is a decomposition of $G_A$, then $\langle S, T|S \cap T \rangle \in \mathfrak{M}(\mathcal{G}_A)$, which implies that $\langle S, B \cup T|T \cap S \rangle \in \mathfrak{M}(\mathcal{G})$. Since $\mathcal{G}_{(S \cup B) \cap T} = \mathcal{G}_{T \cap S}$ is complete, $(S \cup B, T)$ is a decomposition of $\mathcal{G}$.

The converse result follows by the reverse argument.   □

THEOREM 3.4.   *Let $\mathfrak{G}(\tilde{\mathcal{G}})$ be a structurally Markov graph law. Then the conditional law for $\tilde{\mathcal{G}}_A|\{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}$ is also structurally Markov.*

PROOF.   Let $(S, T)$ be a covering pair of $A$: If we restrict $\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$, then $\tilde{\mathcal{G}}_{A \cap B}$ must be complete. As we are only interested in the case where $(S, T)$ is a decomposition of $\tilde{\mathcal{G}}_A$, then $A \cap B$ must be a subset of either $S$ or $T$: without loss of generality, we may assume $A \cap B \subseteq T$.

Since $(S, T \cup B)$ is a covering pair of $V$, by the structural Markov property,

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_{T \cup B}|\{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

If $\mathbb{1}_E$ is the indicator variable of an event $E$, we can write

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp (\tilde{\mathcal{G}}_T, \mathbb{1}_{\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)})|\{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

By the properties of conditional independence [Dawid (1979)], the term $\mathbb{1}_{\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)}$ may be moved to the right-hand side. Furthermore, we are only interested in the case where it equals 1. Hence we can write

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T|\{\mathcal{G}_{T \cup B} \in \mathfrak{U}(T, B)\}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

By Lemma 3.3, $\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)$ if and only if $\tilde{\mathcal{G}} \in \mathfrak{U}(S \cup T, B) = \mathfrak{U}(A, B)$. So

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T|\{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

Again, by Lemma 3.3, $\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)$ if and only if $\tilde{\mathcal{G}}_A \in \mathfrak{U}(S, T)$, hence:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T|\{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \{\tilde{\mathcal{G}}_A \in \mathfrak{U}(S, T)\}.   □$$

3.2. *Structural meta Markov property.* Dawid and Lauritzen (1993) define a *meta Markov model* as a set of Markov distributions that exhibits *conditional variation independence*, denoted by the ternary relation $(\cdot \ddagger \cdot | \cdot)$, in place of the conditional probabilistic independence of hyper Markov laws; see also Dawid (2001b). Analogous structural properties can be defined for families of graphs.

DEFINITION 3.2 (Structural meta Markov property). Let $\mathfrak{F}$ be a family of undirected decomposable graphs on $V$. Then $\mathfrak{F}$ is *structurally meta Markov* if for every covering pair $(A, B)$, the set $\{\mathcal{G}_A : \mathcal{G} \in \mathfrak{F}(A, B), \mathcal{G}_B = \mathcal{J}\}$ is the same for all $\mathcal{J} \in [\mathfrak{F}(A, B)]_B]$. That is,

$$\mathcal{G}_A \ddagger \mathcal{G}_B | \{\mathcal{G} \in \mathfrak{F}(A, B)\}.$$

In other words, this property requires that the set of pairs $(\mathcal{G}_A, \mathcal{G}_B)$ of $\mathcal{G} \in F(A, B)$ be a product set. Clearly the set $\mathfrak{U}$ of all decomposable graphs on $V$ is structurally meta Markov.

As with probabilistic independence, we can characterise it in terms of the graph product operation.

THEOREM 3.5. *A family of undirected decomposable graphs $\mathfrak{F}$ is structurally meta Markov if and only if $\mathcal{G}_A \otimes \mathcal{G}'_B \in \mathfrak{F}$ for all $\mathcal{G}, \mathcal{G}' \in \mathfrak{F}(A, B)$.*

PROOF. This follows directly from Proposition 3.1. $\square$

Theorem 3.5 is particularly useful in that if a family of graphs is characterised by a specific property, we can show that it is structurally meta Markov if this property is preserved under the graph product operation.

EXAMPLE 3.1. The set of undirected decomposable graphs whose clique size is bounded above by some $n$ and whose separator size is bounded below by $m$ is structurally meta Markov. To see this, note that a clique of $\mathcal{G}_A \otimes \mathcal{G}'_B$ must be a clique of either $\mathcal{G}_A$ or $\mathcal{G}'_B$ (and hence of either $\mathcal{G}$ or $\mathcal{G}'$), and therefore the graph product operation cannot increase the size of the largest clique. Similarly, it is not possible for a graph product to decrease the size of the smallest separator: a separator of $\mathcal{G}_A \otimes \mathcal{G}'_B$ must either be a separator of $\mathcal{G}$ or $\mathcal{G}'$, or be $A \cap B$ (this is a consequence of Lemma 3.12).

In the case $n = 2$ and $m = 0$, this is the set of forests on $V$, and when $n = 2$ and $m = 1$, this is the set of trees on $V$.

EXAMPLE 3.2. Consider two graphs $\mathcal{G}^L, \mathcal{G}^U \in \mathfrak{U}$ such that $\mathcal{E}(\mathcal{G}^L) \subseteq \mathcal{E}(\mathcal{G}^U)$. Then the "sandwich" set between the two graphs,

$$\{\mathcal{G} \in \mathfrak{U} : \mathcal{E}(\mathcal{G}^L) \subseteq \mathcal{E}(\mathcal{G}) \subseteq \mathcal{E}(\mathcal{G}^U)\},$$

is structurally meta Markov. This follows from the fact that an edge can only appear in a graph product if it is in one of the elements of the product.

As with hyper Markov laws, being a structural meta Markov family is a necessary condition for the existence of a structural Markov law.

THEOREM 3.6.   *The support of a structurally Markov graph law is a structurally meta Markov family.*

PROOF.    Let $\mathfrak{F}$ be the support of the structurally Markov graph law $\mathfrak{G}$ with density $\pi$. By Proposition 3.2, if $\mathcal{G}, \mathcal{G}' \in \mathfrak{F}(A, B)$ and both $\pi(\mathcal{G})$ and $\pi(\mathcal{G}')$ are nonzero, then $\pi(\mathcal{G}_A \otimes \mathcal{G}'_B)$ must also be nonzero, and hence in $\mathfrak{F}(A, B)$. Therefore, by Theorem 3.5, $\mathfrak{F}$ is structurally meta Markov.   □

3.3. *Compatible distributions and laws.*   We now investigate how the structural Markov property interacts with the Markov and hyper Markov properties. In order to do this, we need to define families of distributions and laws for every graph.

DEFINITION 3.3.    For $\mathfrak{F} \subseteq \mathfrak{U}$, let $\vartheta = \{\theta^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{F}\}$ be a family of probability distributions for $X$. We write $X \sim \vartheta | \tilde{\mathcal{G}}$ if, given $\tilde{\mathcal{G}} = \mathcal{G}$, $X \sim \theta^{(\mathcal{G})}$. Then $\vartheta$ is *compatible* if:

   (i)  for each $\mathcal{G} \in \mathfrak{F}$, $X$ is Markov with respect to $\mathcal{G}$ under $\theta^{(\mathcal{G})}$, and
   (ii) $\theta_A^{(\mathcal{G})} = \theta_A^{(\mathcal{G}')}$ whenever $\mathcal{G}, \mathcal{G}' \in \mathfrak{F}$ are collapsible onto $A$ and $\mathcal{G}_A = \mathcal{G}'_A$.

Similar properties can be defined for laws.

DEFINITION 3.4.    For $\mathfrak{F} \subseteq \mathfrak{U}$, let $\mathfrak{L} = \{\pounds^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{F}\}$ be a family of laws for the parameters $\tilde{\theta}$ of a family of distributions on $X$. Again, we can write $\tilde{\theta} \sim \mathfrak{L} | \tilde{\mathcal{G}}$ if, given $\tilde{\mathcal{G}} = \mathcal{G}$, $\tilde{\theta} \sim \pounds^{(\mathcal{G})}$. Then $\mathfrak{L}$ is *hyper compatible* if:

   (i)  for all $\mathcal{G} \in \mathfrak{F}$, $\pounds^{(\mathcal{G})}$ is weak hyper Markov with respect to $\mathcal{G}$, and
   (ii) $\pounds_A^{(\mathcal{G})} = \pounds_A^{(\mathcal{G}')}$ whenever $\mathcal{G}, \mathcal{G}' \in \mathfrak{F}$ are collapsible onto $A$ and $\mathcal{G}_A = \mathcal{G}'_A$.

REMARK.    Dawid and Lauritzen (1993), Section 6.2, originally used the term compatible to refer to what we term the hyper compatible case: we introduce the distinction so as to extend the terminology to the distributional (nonhyper) case.

As Markov distributions and hyper Markov laws are characterised by their clique-marginal distributions [Dawid and Lauritzen (1993), Theorems 2.6 and 3.9], it is sufficient for condition (ii) in Definitions 3.3 and 3.4 to hold when $\mathcal{G}_A$ and $\mathcal{G}'_A$ are complete. Moreover, if the complete graph $\mathcal{G}^{(V)}$ is contained in $\mathfrak{F}$, then the compatible and hyper compatible families are characterised entirely by $\theta^{(\mathcal{G}^{(V)})}$ and $\pounds^{(\mathcal{G}^{(V)})}$, respectively.

EXAMPLE 3.3. The *inverse Wishart law* for the covariance selection model $\theta(X) = \mathcal{N}(0, \Sigma)$ assigns $\pounds(\Sigma) = \mathcal{IW}(\delta; \Phi)$. This law is strong hyper Markov with respect to the complete graph on $V$, and the hyper compatible family generated by $\pounds$ are the *hyper inverse Wishart laws* $\pounds^{(\mathcal{G})}(\Sigma) = \mathcal{HIW}_{\mathcal{G}}(\delta; \Phi)$ [Dawid and Lauritzen (1993), Example 7.3].

A law induces marginal distribution $\theta_{\pounds}$ for $X$ such that $\theta_{\pounds}(A) = \mathbb{E}_{\pounds}[\tilde{\theta}(A)]$, referred to as the *predictive distribution* in Bayesian problems. Therefore a family of laws will also induce a family of distributions. Although in general hyper compatibility will not imply compatibility, there is one important special case.

PROPOSITION 3.7. *Let $\mathfrak{L}$ be a family of laws such that each law $\pounds^{(\mathcal{G})} \in \mathfrak{L}$ is strong hyper Markov. Then the family of marginal distributions*

$$\{\theta_{\pounds} : \pounds \in \mathfrak{L}\}$$

*is hyper compatible.*

PROOF. By Dawid and Lauritzen (1993), Proposition 5.6, the marginal distribution of a strong hyper Markov law is Markov with respect to the same graph. The result follows by noting that the marginal distribution on a complete subgraph is a function of the marginal law. □

A graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ combined with a compatible set of distributions $\vartheta$ defines a joint distribution $(\mathfrak{G}, \vartheta)$ for $(\tilde{\mathcal{G}}, X)$ under which $X|\tilde{\mathcal{G}} = \mathcal{G} \sim \theta^{(\mathcal{G})}$. Likewise, $\mathfrak{G}$ combined with a set of hyper compatible laws $\mathfrak{L}$ defines a joint law $(\mathfrak{G}, \mathfrak{L})$ for $(\tilde{\mathcal{G}}, \tilde{\theta})$, and so a joint distribution on $(\tilde{\mathcal{G}}, \tilde{\theta}, X)$.

The key conditional independence property of any such joint distribution or law can be characterised as follows.

PROPOSITION 3.8. *For any graph law $\mathfrak{G}$ over $\mathfrak{F} \subseteq \mathfrak{U}$ for $\tilde{\mathcal{G}}$, and $X \sim \vartheta$ for a compatible family $\vartheta$ indexed by $\mathfrak{F}$,*

$$X_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B | \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \qquad [\mathfrak{G}, \vartheta].$$

*Similarly, if $\tilde{\theta} \sim \mathfrak{L}$ for a hyper compatible family $\mathfrak{L}$ indexed by $\mathfrak{F}$, then*

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B | \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \qquad [\mathfrak{G}, \mathfrak{L}].$$

PROOF. Let $\mathcal{G}, \mathcal{G}' \in \mathfrak{U}(A, B)$ such that $\mathcal{G}_A = \mathcal{G}'_A$. As $\mathcal{G}$ and $\mathcal{G}'$ are both collapsible onto $A$, then $\theta_A^{(\mathcal{G})} = \theta_A^{(\mathcal{G}')}$ in a compatible family, and $\pounds_A^{(\mathcal{G})} = \pounds_A^{(\mathcal{G}')}$ in a hyper compatible family. □

When combined with the structural Markov property, we obtain some useful results.

THEOREM 3.9. *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law $\mathfrak{G}$, and $X$ has a distribution from a compatible set $\vartheta$, then*

$$(X_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) | X_{A \cap B}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \qquad [\mathfrak{G}, \vartheta].$$

PROOF.    See Appendix B.    □

COROLLARY 3.10.    *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law, and $X$ has a distribution from a compatible set $\vartheta$, then the posterior graph law for $\tilde{\mathcal{G}}$ is structurally Markov.*

PROOF.    By Theorem 3.9 and the axioms of conditional independence, we easily obtain

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B | X, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}. \qquad\qquad □$$

We can also apply similar arguments at the hyper level.

THEOREM 3.11.    *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law $\mathfrak{G}$, and $\theta$ has a law from a hyper compatible set $\mathfrak{L}$, then*

$$(\tilde{\theta}_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (\tilde{\theta}_B, \tilde{\mathcal{G}}_B) | \tilde{\theta}_{A \cap B}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \qquad [\mathfrak{G}, \mathfrak{L}].$$

*Furthermore, if each law $\mathfrak{t}^{(\mathcal{G})} \in \mathfrak{L}$ is strong hyper Markov with respect to $\mathcal{G}$, then*

$$(\tilde{\theta}_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (\tilde{\theta}_{B|A}, \tilde{\mathcal{G}}_B) | \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \qquad [\mathfrak{G}, \mathfrak{L}].$$

PROOF.    The proof for the first case is the same as in Theorem 3.9. The proof for the strong case follows similar steps, except starting with the strong hyper Markov property

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_{B|A} | \tilde{\mathcal{G}}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}. \qquad\qquad □$$

Hyper compatible sets of strong hyper Markov laws have the additional advantage that the posterior graph law will also be structurally Markov: this follows from Theorem 3.9 and Dawid and Lauritzen (1993), Proposition 5.6, which states that the marginal distribution of the data under a strong hyper Markov law is Markov. Furthermore, the posterior family of graph laws $\{\mathfrak{t}^{(\mathcal{G})}(\cdot | X) : \mathcal{G} \in \mathfrak{U}\}$ will maintain hyper compatibility.

3.4. *Clique vector.* We show that the family of structural Markov laws forms an exponential family of conjugate distributions for Bayesian updating under compatible sampling.

DEFINITION 3.5. Define the *completeness vector* of a graph to be the function $c : \mathfrak{U} \to \{0, 1\}^{2^V}$ such that, for each $A \subseteq V$,

$$c_A(\mathcal{G}) = \begin{cases} 1, & \text{if } \mathcal{G}_A \text{ is complete,} \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, define the *clique vector* of a graph $t : \mathfrak{U} \to \mathbb{Z}^{2^V}$ to be the Möbius inverse of $c$ by *superset* inclusion

$$(3.4) \qquad t_B(\mathcal{G}) = \sum_{A \supseteq B} (-1)^{|A \setminus B|} c_A(\mathcal{G}).$$

In the language of Studený (2005b), $c$ and $t$ are both *imsets*.
The decomposition of $c$ and $t$ mirrors that of the graph.

LEMMA 3.12. *If $\mathcal{G} \in \mathfrak{U}(A, B)$, then*

$$(3.5) \qquad c(\mathcal{G}) = [c(\mathcal{G}_A)]^0 + [c(\mathcal{G}_B)]^0 - [c(\mathcal{G}_{A \cap B})]^0 \quad \text{and}$$

$$(3.6) \qquad t(\mathcal{G}) = [t(\mathcal{G}_A)]^0 + [t(\mathcal{G}_B)]^0 - [t(\mathcal{G}_{A \cap B})]^0,$$

*where $[\cdot]^0$ denotes the expansion of a vector with zeroes to the required coordinates.*

PROOF. A subset $U \subseteq V$ induces a complete subgraph of $\mathcal{G} \in \mathfrak{U}(A, B)$ if and only if it induces a complete subgraph of $\mathcal{G}_A$ or of $\mathcal{G}_B$ (or of both). (3.5) follows by the inclusion-exclusion principle. (3.6) may then be obtained by substitution into (3.4). □

THEOREM 3.13. *For any decomposable graph $\mathcal{G} \in \mathfrak{U}$ and $A \subseteq V$,*

$$t_A(\mathcal{G}) = \begin{cases} 1, & \text{if } A \in \mathrm{cl}(\mathcal{G}), \\ -\nu_{\mathcal{G}}(A), & \text{if } A \in \mathrm{sep}(\mathcal{G}), \text{ and} \\ 0, & \text{otherwise,} \end{cases}$$

*where $\mathrm{cl}(\mathcal{G})$ are the cliques of $\mathcal{G}$, and $\mathrm{sep}(\mathcal{G})$ are the clique separators, and each separator $S$ has multiplicity $\nu_{\mathcal{G}}(S)$.*

PROOF. For any $C \subseteq V$, let $\mathcal{G}^{(C)}$ be the graph on $V$ whose edges are the set of all pairs $\{u, v\} \subseteq C$ (i.e., complete on $C$ and sparse elsewhere). Then it is straightforward to see that

$$t_A(\mathcal{G}_C^{(C)}) = \begin{cases} 1, & \text{if } A = C, \\ 0, & \text{otherwise.} \end{cases}$$

Now let $C_1, \ldots, C_k$ be a perfect ordering of the cliques of $G$, and $S_2, \ldots, S_k$ be the corresponding separators. By Lemma 3.12, it follows that

$$t(\mathcal{G}) = \sum_{i=1}^{k} t(\mathcal{G}_{C_i}^{(C_i)}) - \sum_{i=2}^{k} t(\mathcal{G}_{S_i}^{(S_i)}).$$

$\square$

Objects similar to the clique vector have arisen in several contexts. Notably, it appears to be equivalent to the index $v$ of Lauritzen, Speed and Vijayan [(1984), Definition 5], which is characterised in a combinatorial manner. It is also closely related to the *standard imset* of Studený (2005b), which is equal to

$$t(\mathcal{G}^{(V)}) - t(\mathcal{G}),$$

where $\mathcal{G}^{(V)}$ is the complete graph.

The algorithm of Wormald (1985) for the enumeration of decomposable graphs is based on a generating function for the vector $\mathbb{R}^{|V|}$ that he termed the "maximal clique vector," and is equivalent to

$$\mathrm{mcv}_k(\mathcal{G}) = \sum_{A \subseteq V\,:\,|A|=k} t_A(\mathcal{G}), \qquad k = 1, \ldots, |V|.$$

PROPOSITION 3.14. *For any $\mathcal{G} \in \mathfrak{U}$, the vector $t(\mathcal{G})$ has the following properties*:

(i)
$$\sum_{A \subseteq V} t_A(\mathcal{G}) = 1,$$

(ii) *for each $v \in V$*
$$\sum_{A \ni v} t_A(\mathcal{G}) = 1,$$

(iii)
$$\sum_{A \subseteq V} |A| t_A(\mathcal{G}) = |V| \quad and$$

(iv)
$$\sum_{A \subseteq V} \binom{|A|}{2} t_A(\mathcal{G}) = |\mathcal{E}(\mathcal{G})|.$$

PROOF. By the Möbius inversion theorem [see, e.g., Lauritzen (1996), Lemma A.2], $c$ can also be expressed in terms of $t$,

$$c_A(\mathcal{G}) = \sum_{B \supseteq A} t_B(\mathcal{G}), \qquad A \subseteq V.$$

(i) and (ii) are $c_A(\mathcal{G})$ at $A = \varnothing$ and $A = \{v\}$, respectively, both of which induce complete subgraphs. (iii) is obtained from (ii) by summation over $v \in V$, and (iv) is obtained from (ii) by double counting each edge via summation over both elements $\{u, v\} \in \mathcal{E}(\mathcal{G})$.  $\square$

3.5. *Clique exponential family.*

DEFINITION 3.6. The *clique exponential family* is the exponential family of graph laws over $\mathfrak{F} \subseteq \mathfrak{U}$, with $t$ as a natural statistic (with respect to the uniform measure on $\mathfrak{U}$). That is, laws in the family have densities of the form

$$\pi_\omega(\mathcal{G}) = \frac{1}{Z(\omega)} \exp\{\omega \cdot t(\mathcal{G})\}, \qquad \mathcal{G} \in \mathfrak{F}, \omega \in \mathbb{R}^{2^V},$$

where $Z(\omega)$ is the normalisation constant, which will generally be hard to compute.

Equivalently, the distribution can be parameterised in terms of $c$,

$$\pi_\omega(\mathcal{G}) = \frac{1}{Z(\omega)} \exp\left\{\left(\sum_{B \subseteq A} (-1)^{|A \setminus B|} \omega_A\right)_{A \subseteq V} \cdot c(\mathcal{G})\right\},$$

but $t$ is more useful due to the fact that it is sparse (by Theorem 3.13) and, as we shall see, is the natural statistic for posterior updating.

Note that this distribution is over-parametrised. By Proposition 3.14(i) and (ii), there are $|V| + 1$ linear constraints in the set of possible $t(G)$, adding multiples of $\alpha = (1)_{S \subseteq V}$, or $\beta_v = (\mathbb{1}_{v \in S})_{S \subseteq V}$ to $\omega$ will leave the resulting $\pi$ unchanged. For the purpose of identifiability, we could define a standardised vector $\omega^*$ as

$$\omega^* = \omega + \omega_\varnothing \alpha + \sum_{v \in V} (\omega_{\{v\}} - \omega_\varnothing) \beta_v = \left(\omega_A + (|A| - 1)\omega_\varnothing - \sum_{v \in A} \omega_{\{v\}}\right)_{A \subseteq V}$$

such that $\pi_\omega = \pi_{\omega^*}$, and $\omega_{\{v\}}^* = \omega_\varnothing^* = 0$ for all $v \in V$.

THEOREM 3.15. *Let $\mathfrak{G}$ be a graph law whose support is $\mathfrak{U}$. Then $\mathfrak{G}$ is structurally Markov if and only if it is a member of the clique exponential family.*

PROOF. See Appendix B. □

REMARK. It is possible to weaken the condition of full support; for example, the same argument applies to any family $\mathfrak{F}$ with the property that if $\mathcal{G} \in \mathfrak{F}$ and $C$ is a clique of $\mathcal{G}$, then $\mathcal{G}^{(C)} \in \mathfrak{F}$.

A very similar family was proposed by Bornn and Caron (2011); however, their family allows the use of different parameters for cliques and separators, which will generally not be structurally Markov.

EXAMPLE 3.4 [Giudici and Green (1999); Brooks, Giudici and Roberts (2003), Section 8]. The simplest example of such a distribution is the uniform distribution over $\mathfrak{U}$, which by Proposition 3.14(i), corresponds to $\omega_A$ being constant for all $A$.

EXAMPLE 3.5 [Jones et al. (2005), Madigan and Raftery (1994)].    Another common approach is to use a set of $\binom{|V|}{2}$ independent Bernoulli variables with probability $\psi$ to indicate edge inclusion (i.e., an Erdős–Rényi random graph), conditional on $\tilde{\mathcal{G}}$ being decomposable. The density of such a law is of the form

$$\pi(\mathcal{G}) \propto \psi^{|\mathcal{E}(\mathcal{G})|}(1-\psi)^{\binom{p}{2}-|\mathcal{E}(\mathcal{G})|} \propto \left(\frac{\psi}{1-\psi}\right)^{|\mathcal{E}(\mathcal{G})|}.$$

By Proposition 3.14(iv), it follows that this distribution is a member of the exponential family with parameter

$$\omega_A = \binom{|A|}{2}\log\left(\frac{\psi}{1-\psi}\right).$$

More generally, the family with parameter

$$\omega_A = \sum_{e\in\binom{A}{2}} \log\left(\frac{\psi_e}{1-\psi_e}\right)$$

would correspond to the extension where each edge $e$ has its own probability $\psi_e$.

EXAMPLE 3.6.    By adjusting parameters of the family, particular graphical features can be emphasised. For example, a family of the form

$$\omega_A = \binom{|A|}{2}\rho - \kappa\max(0, |A| - 2),$$

with $\kappa > 0$, will penalise clique sizes greater than 2, placing a higher probability on forest structures.

EXAMPLE 3.7 [Armstrong et al. (2009)].    For comparison, it is useful to consider a nonstructurally Markov graph law. Define the distribution over the number of edges to be uniform, and the conditional distribution over the set of graphs with a fixed number of edges to be uniform. This has density of the form

$$\pi(\mathcal{G}) = \frac{1}{\binom{p}{2}+1}\frac{1}{|\{\mathcal{G}'\in\mathfrak{U}:|\mathcal{E}(\mathcal{G}')|=|\mathcal{E}(\mathcal{G})|\}|}.$$

Specifically for graphs on three vertices, we have that

$$\pi(\begin{smallmatrix}\bullet\!\frown\\ \bullet\end{smallmatrix}) = \tfrac{1}{12}, \qquad \pi(\begin{smallmatrix}\bullet\!\searrow\\ \bullet\end{smallmatrix}) = \tfrac{1}{12}, \qquad \pi(\bullet\ \begin{smallmatrix}\bullet\\ \bullet\end{smallmatrix}) = \tfrac{1}{4} \quad\text{and}\quad \pi(\begin{smallmatrix}\bullet\!<\\ \bullet\end{smallmatrix}) = \tfrac{1}{12}.$$

Therefore by Proposition 3.2 the law cannot be structurally Markov.

3.6. *Posterior updating.* We saw in Corollary 3.10 that if the sampling distributions are compatible, then posterior updating will preserve the structural Markov property. In this section we show that this updating may be performed locally, with the exponential clique family forming a conjugate prior for a family of compatible models.

THEOREM 3.16. *Let $\vartheta$ be a family of compatible distributions for $X$, where each $\theta^{(\mathcal{G})}$ has density $\pi^{(\mathcal{G})}$ with respect to some product measure. Then*

$$\pi^{(\mathcal{G})}(x) = \prod_{A \subseteq V} p_A(x_A)^{[t(\mathcal{G})]_A},$$

*for all $x$ such that $\pi^{(\mathcal{G})}(x) > 0$, where $p_A$ is the marginal density of $X_A$ whenever $\mathcal{G}_A$ is complete, and $p_\varnothing(x_\varnothing) = 1$.*

PROOF. For any decomposition $(A, B)$ of $\mathcal{G}$, then for any $x$ such that $\pi^{(G)}(x) > 0$,

$$\pi^{(\mathcal{G})}(x) = \pi_A^{(\mathcal{G})}(x_A)\pi_{B|A}^{(\mathcal{G})}(x_{B\setminus A}|x_A) = \pi_A^{(\mathcal{G})}(x_A)\pi_{B|A\cap B}^{(\mathcal{G})}(x_{B\setminus A}|x_{A\cap B})$$

$$= \pi_A^{(\mathcal{G})}(x_A)\frac{\pi_B^{(\mathcal{G})}(x_B)}{\pi_{A\cap B}^{(\mathcal{G})}(x_{A\cap B})}.$$

The result follows by recursive decomposition over the clique tree. $\square$

Therefore if the prior law for $\tilde{\mathcal{G}}$ is a clique exponential with parameter $\omega$, then under sampling from a compatible family the resulting posterior law is of the same family,

$$\pi(\mathcal{G}|X = x) \propto \exp\{[\omega + (\log p_A(x_A))_{A\subseteq V}] \cdot t(\mathcal{G})\}.$$

A key benefit of this conjugate formation is that we can describe the posterior law with a parameter of dimension $2^{|V|}$ (strictly speaking, we only need $2^{|V|} - |V| - 1$, due to the over-parametrisation). This is much smaller than for an arbitrary law over the set of undirected decomposable graphs, which would require a parameter of length approximately $2^{\binom{|V|}{2}}$.

**4. Ordered directed structural Markov property.** We now investigate the first of two different methods by which the structural Markov property might be extended to directed acyclic graphical models (DAGs). In this section, we consider a law for a random graph $\tilde{\mathcal{G}}$ over the set $\mathfrak{D}^\prec$: the set of directed acyclic graphs that respect a fixed well ordering $\prec$ on $V$.

The set $\mathfrak{D}^\prec$ is straightforward to characterise, as $\prec$ determines the directionality of an edge between a pair of vertices. Therefore, as in the undirected case, a random graph $\tilde{\mathcal{G}}$ on $\mathfrak{D}^\prec$ can also be interpreted as a random vector of length $\binom{|V|}{2}$.

In order to develop a structural Markov graph law over $\mathfrak{D}^{\prec}$, recall that the strong directed hyper Markov property can be expressed as

$$(4.1) \qquad\qquad \tilde{\theta}_{v|\operatorname{pr}(v)} \perp\!\!\!\perp \tilde{\theta}_{\operatorname{pr}(v)},$$

for all $v \in V$. This in turn implies mutual independence of the collection $(\tilde{\theta}_{v|\operatorname{pr}(v)})_{v \in V}$. Each element $\tilde{\theta}_{v|\operatorname{pr}(v)}$ is constrained by $\mathcal{G}$ only through the parent set $\operatorname{pa}_{\mathcal{G}}(v)$, as we require that $X_v \perp\!\!\!\perp X_{\operatorname{pr}(v)}|X_{\operatorname{pa}_{\mathcal{G}}(v)}$. This motivates the following definitions.

DEFINITION 4.1. The *ordered remainder graph of $\mathcal{G}$ of $v \in V$ with respect to $\prec$*, denoted by $\mathcal{G}^{\prec}_{v|\operatorname{pr}(v)}$ is the graph on $\{v\} \cup \operatorname{pr}(v)$, and edge set $\mathcal{E}(\mathcal{G}_{\{v\}\cup\operatorname{pr}(v)}) \cup \{(w, u) : w, u \in \operatorname{pr}(v), w \prec u\}$, that is, the subgraph $\mathcal{G}_{\{v\}\cup\operatorname{pr}(v)}$ with the addition of all possible edges between elements of $\operatorname{pr}(v)$ respecting $\prec$.

The ordered remainder graph directly corresponds to the parent set of the vertex, or equivalently, the set of vertices with a common head,

$$\mathcal{G}^{\prec}_{v|\operatorname{pr}(v)} \simeq \operatorname{pa}_{\mathcal{G}}(v) \simeq \{(u, w) \in \mathcal{E}(\mathcal{G}) : w = v\}.$$

The advantage of the remainder graph is that it allows the partitioning of the semi-graphoid into its constituent components.

PROPOSITION 4.1. *Let $\mathcal{G}$ be a directed acyclic graph compatible with the ordering $\prec$. Then a distribution $P$ is Markov with respect to $\mathcal{G}$ if and only if for each $v \in V$, $P_{\{v\}\cup\operatorname{pr}(v)}$ is Markov with respect to $\mathcal{G}^{\prec}_{v|\operatorname{pr}(v)}$.*

*Similarly, a law £ is weak/strong hyper Markov if and only if for each $v \in V$, $£_{\{v\}\cup\operatorname{pr}(v)}$ is weak/strong hyper Markov with respect to $\mathcal{G}^{\prec}_{v|\operatorname{pr}(v)}$.*

PROOF. These follow from the ordered directed Markov property. □

The motivation of the term "remainder" is that $\mathcal{G}^{\prec}_{v|\operatorname{pr}(v)}$ encodes the remainder of the semi-graphoid of $\mathcal{G}_{\{v\}\cup\operatorname{pr}(v)}$ that is not determined by $\mathcal{G}_{\operatorname{pr}(v)}$.

DEFINITION 4.2 (Ordered directed structural Markov property). The graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ over $\mathfrak{D}^{\prec}$ is *ordered directed structurally Markov* with respect to the ordering $\prec$ if for each $v \in V$,

$$\tilde{\mathcal{G}}^{\prec}_{v|\operatorname{pr}(v)} \perp\!\!\!\perp \tilde{\mathcal{G}}_{\operatorname{pr}(v)}.$$

As $\tilde{\mathcal{G}}_{\operatorname{pr}(v)} \simeq (\tilde{\mathcal{G}}^{\prec}_{u|\operatorname{pr}(v)})_{u \in \operatorname{pr}(v)}$, this implies that the set of all ordered remainder graphs, or equivalently, the set of all parent sets, are mutually independent; see Figure 2.

Admittedly this construction is not very complicated, but it does demonstrate how structure can be "decomposed" in directed graphs, which will be used in the next section.
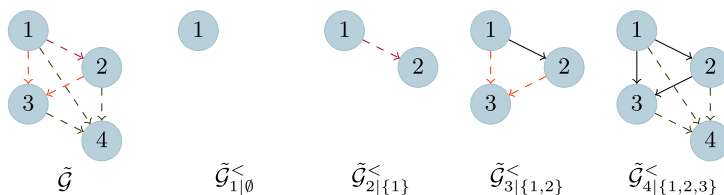
FIG. 2. *A random directed acyclic graph $\tilde{\mathcal{G}}$ on $V = \{1, 2, 3, 4\}$, subject to the ordering $<$, and its corresponding ordered remainder graphs. Under an ordered directed structural Markov graph law, the ordered remainder graphs—or equivalently, the collections of like-coloured edges—are independent.*

## 5. Markov equivalence and the dagoid structural Markov property.

The approach in Section 4 cannot be applied directly to distributions over the $\mathfrak{D}$, the set of all directed acyclic graphs on $V$. For instance, parent sets of individual vertices cannot be independent: if $u$ is a parent of $v$, then $v$ is precluded from being a parent of $u$.

A bigger problem is that there is no longer a one-to-one correspondence between a graph and its semi-graphoid. That is, two or more distinct DAGs may have identical conditional independence properties, for example, $\overset{\cdot\cdot\cdot}{\nearrow}$, $\overset{\cdot\cdot\cdot}{\nwarrow}$, and $\overset{\cdot\cdot\cdot}{\nwarrow}$.

DEFINITION 5.1. Let $\mathcal{G}$ and $\mathcal{G}'$ be directed acyclic graphs such that $\mathfrak{M}(\mathcal{G}) = \mathfrak{M}(\mathcal{G}')$. Then $\mathcal{G}$ and $\mathcal{G}'$ are termed *Markov equivalent*, and we write

$$\mathcal{G} \overset{\mathfrak{M}}{\sim} \mathcal{G}'.$$

A *dagoid* is a Markov equivalence class of directed acyclic graphs. We define the *complete* and *sparse* dagoids to be the Markov equivalence classes of complete and sparse DAGs, respectively. We use $\mathfrak{D}^{\mathfrak{M}}$ to denote the set of dagoids on $V$.

There are various methods of characterising Markov equivalence, several of which are mentioned in the Appendix.

So when specifying a law for directed acyclic graphs, we are left with the question of whether or not we should treat Markov equivalent graphs as the same model. In other words, whether the model is defined by the graph or the set of conditional independence statements which it encodes. As noted earlier, we take the latter view.

A further advantage of working with equivalence classes is that a smaller number of models needs be considered. Unfortunately this may not be as beneficial as one may initially hope: Castelo and Kočka (2004) observed empirically that the ratio of the number DAGs to the number of equivalence classes appears to converge to approximately 3.7 as the number of vertices increases.

5.1. *Ancestral sets and remainder dagoids.* Although ancestral sets are used in the definition of the global directed Markov property, ancestral sets themselves are not preserved under Markov equivalence. However, as noted in Section 2.3, subgraphs induced by ancestral sets preserve the projection of the semi-graphoid. A somewhat trivial consequence is the following.

PROPOSITION 5.1. *Let* $\mathcal{G} \overset{\mathfrak{M}}{\sim} \mathcal{G}'$ *and* $A \subseteq V$ *be ancestral in both* $\mathcal{G}$ *and* $\mathcal{G}'$. *Then* $\mathcal{G}_A \overset{\mathfrak{M}}{\sim} \mathcal{G}'_A$.

This motivates the following definition.

DEFINITION 5.2. A set $A \subseteq V$ is *ancestral* in a dagoid $\mathcal{D}$ if it is ancestral for some graph $\mathcal{G} \in \mathcal{D}$. For any such $A$, define the *subdagoid induced by* $A$ to be the Markov equivalence class of $\mathcal{G}_A$, and denote it by $\mathcal{D}_A$.

For any $A \subseteq V$, let $\mathfrak{D}(A)$ denote the set of dagoids on $V$ in which $A$ is an ancestral set.

Note that the dagoid ancestral property is not as strong as the collapsibility property in undirected graphs, in that there can exist nonancestral sets that also preserve the semi-graphoid of the induced subgraph.

However, ancestral sets are still quite powerful, in that they can be used to decompose the semi-graphoid.

DEFINITION 5.3. Let $\mathcal{G}$ be a directed acyclic graph on $V$, of which $A$ is an ancestral set, and let $\mathcal{H}$ be a directed acyclic graph on $A$. Then the *insertion of* $\mathcal{H}$ *into* $\mathcal{G}$, written

$$\mathcal{H} \ltimes \mathcal{G},$$

is the directed acyclic graph on $V$ with edge set

$$\mathcal{E}(\mathcal{H}) \cup \left[ \mathcal{E}(\mathcal{G}) \setminus A^2 \right].$$

In other words, the edges between elements of $A$ are determined by $\mathcal{H}$, and all other edges are determined by $\mathcal{G}$. This operation preserves Markov equivalence.

LEMMA 5.2. *Let* $\mathcal{G}$ *and* $\mathcal{G}'$ *be Markov equivalent graphs in which* $A$ *is an ancestral set, and* $\mathcal{H}$ *and* $\mathcal{H}'$ *be Markov equivalent graphs on* $A$. *Then*

$$\mathcal{H} \ltimes \mathcal{G} \overset{\mathfrak{M}}{\sim} \mathcal{H}' \ltimes \mathcal{G}'.$$

PROOF. We use the notation and results of Appendix A. Both graphs must have the same skeleton. Let $(a, b, c)$ be an immorality in $\mathcal{H} \ltimes \mathcal{G}$. Then if $b \in A$,
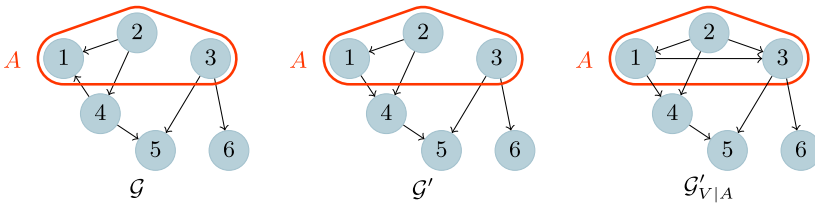
FIG. 3. $A = \{1, 2, 3\}$ *is an ancestral set of the dagoid containing* $\mathcal{G}$, *as it is ancestral in the graph* $\mathcal{G}'$ *obtained by reversing the covered edge* $(4, 1)$. $\mathcal{G}'_{V|A}$ *is obtained by replacing the edges between elements of A with those of a complete graph on A.*

then $(a, b, c)$ must be an immorality of $\mathcal{H}$, and hence also an immorality of $\mathcal{H}'$, and so also of $\mathcal{H}' \ltimes \mathcal{G}'$.

Otherwise if $b \notin A$, and at least one of $a$ or $c$ is not in $A$, then $(a, b, c)$ must be an immorality of $\mathcal{G}$, and hence an immorality of $\mathcal{G}'$ and $\mathcal{H}' \ltimes \mathcal{G}'$.

Finally, if $b \notin A$ and $a, c \in A$, then $\{a, c\}$ must not be an edge in the skeleton $\mathcal{H}$, nor an edge in the skeleton of $\mathcal{H}'$. Hence it must also be an immorality of $\mathcal{H}' \ltimes \mathcal{G}'$. □

Consequently for a dagoid $\mathcal{D}$ with ancestral set $A$, we can define the *ancestral insertion* of a dagoid $\mathcal{K}$ on $A$ into $\mathcal{D}$ as

$$\mathcal{K} \ltimes \mathcal{D} = [\mathcal{H} \ltimes \mathcal{G}],$$

where $\mathcal{G} \in \mathcal{D}$ is a directed acyclic graph with an ancestral set $A$, $\mathcal{H} \in \mathcal{K}$, and $[\cdot]$ denotes the Markov equivalence class.

We use this approach to extend the notion of a remainder graph from the previous section without the use of a fixed well-ordering.

DEFINITION 5.4. Let $A$ be an ancestral set of a directed acyclic graph $\mathcal{G}$. A directed acyclic graph $\mathcal{G}_{V|A}$ is a *remainder graph of $\mathcal{G}$ given A* if

$$\mathcal{G}_{V|A} = \mathcal{C}^{(A)} \ltimes \mathcal{G},$$

where $\mathcal{C}^{(A)}$ is a complete dagoid on $A$.

By Lemma 5.2, the remainder graph must be unique up to Markov equivalence. Hence for a dagoid $\mathcal{D} \in \mathfrak{D}(A)$, we can uniquely define the *remainder dagoid of $\mathcal{D}$ given A*, denoted by $\mathcal{D}_{V|A}$; see Figure 3.

Analogous with the ordered case, the induced and remainder dagoids $\mathcal{D}_A$ and $\mathcal{D}_{V|A}$ characterise the complete dagoid (via the ancestral insertion). Moreover, they can be chosen independently.

THEOREM 5.3. *For any $A \subseteq V$, we have*

$$\mathcal{D}_A \ddagger \mathcal{D}_{V|A} | \{\mathcal{D} \in \mathfrak{D}(A)\}.$$

PROOF.    For any $\mathcal{D}, \mathcal{D}' \in \mathfrak{D}(A)$, we can construct $\mathcal{D}^* = \mathcal{D}_A \ltimes \mathcal{D}'_{V|A}$. This will have the required properties that $\mathcal{D}^*_A = \mathcal{D}_A$ and $\mathcal{D}^*_{V|A} = \mathcal{D}'_{V|A}$.    $\square$

5.2. *Dagoid structural Markov property.*    This motivates the following construction for the structural Markov property.

DEFINITION 5.5 (Dagoid structural Markov property).    We say a graph law $\mathfrak{G}(\tilde{\mathcal{D}})$ is *structurally Markov* if for any $A \subseteq V$, we have

$$\tilde{\mathcal{D}}_{V|A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A | \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}    \qquad [\mathfrak{G}].$$

As in the undirected case, we can characterise this property via the odds ratio of the density.

PROPOSITION 5.4.    *A graph law is structurally Markov if and only if for any* $\mathcal{D}, \mathcal{D}' \in \mathfrak{D}(A)$, *we have*

$$(5.1)    \qquad \pi(\mathcal{D})\pi(\mathcal{D}') = \pi(\mathcal{D}_A \ltimes \mathcal{D}'_{V|A})\pi(\mathcal{D}'_A \ltimes \mathcal{D}_{V|A}).$$

PROOF.    As in Proposition 3.2, we may write the density

$$\pi(\mathcal{D}|\mathfrak{D}(A)) = \pi(\mathcal{D}_A|\mathfrak{D}(A))\pi(\mathcal{D}_{V|A}|\mathfrak{D}(A)).    \qquad \square$$

5.3. *d-Clique vector.*    The equivalence class formulation of a dagoid is difficult to work with, both algebraically and computationally. Instead we propose a characteristic vector similar to the clique vector of Section 3.4.

DEFINITION 5.6.    The *d-clique vector* of a directed acyclic graph $\mathcal{G}$ is

$$(5.2)    \qquad t(\mathcal{G}) = \sum_{v \in V} \left[ \delta(\{v\} \cup \mathrm{pa}_{\mathcal{G}}(v)) - \delta(\mathrm{pa}_{\mathcal{G}}(v)) \right] + \delta(\varnothing) \in \mathbb{Z}^{2^V},$$

where $\delta(A) = (\mathbb{1}_{S=A})_{S \subseteq V}$.

Again, we note the relationship to the imsets of Studený (2005b), specifically the structural imset $u_{\mathcal{G}} = \delta(V) - t(\mathcal{G})$ in Section A.4. For our purposes, the d-clique vector is a more convenient object with which to work. This exhibits analogous properties to those of the clique vector of Section 3.4.

PROPOSITION 5.5.    *The properties of Proposition 3.14 apply to all directed graphs* $\mathcal{G} \in \mathfrak{D}$.

PROOF.    (i) follows directly from the definition. (ii) is obtained by noting that each term of (5.2) contributes 1 if the summand is $v$, and 0 otherwise. For (iii),

each term of (5.2) contributes 1, and (iv) is due to each term of (5.2) counting the number of edges whose head is $v$. $\square$

In a similar manner to the undirected case, we can define the *d-completeness vector* to be the Möbius transform of the d-clique vector,

$$(5.3) \qquad c_A(\mathcal{G}) = \sum_{B \supseteq A} t_B(\mathcal{G}),$$

and say that a set $A \subseteq B$ is *d-complete* if $c_A(\mathcal{G}) = 1$. This corresponds to the definition of the characteristic imset of Hemmecke, Lindner and Studený (2012).

LEMMA 5.6 (Hemmecke, Lindner and Studený (2012), Theorem 1). *Let $\prec$ be a well-ordering of a directed acyclic graph $\mathcal{G}$. For any nonempty set $A \subseteq V$, with maximal element $a$ under $\prec$,*

$$c_A(\mathcal{G}) = \begin{cases} 1, & \text{if } A \setminus \{a\} \subseteq \mathrm{pa}_{\mathcal{G}}(a), \\ 0, & \text{otherwise.} \end{cases}$$

This provides the link to the completeness and clique vectors of undirected graphs from Section 3.4.

COROLLARY 5.7. *If $\mathcal{G}$ is a perfect directed acyclic graph, and $\mathcal{G}^s$ is its skeleton, then $c_{\mathcal{G}} = c_{\mathcal{G}^s}$, and hence $t(\mathcal{G}) = t(\mathcal{G}^s)$.*

Most important, the d-clique vector is a unique representation of the dagoid.

THEOREM 5.8. *Let $\mathcal{G}, \mathcal{G}'$ be directed acyclic graphs on $V$. Then $\mathcal{G} \overset{\mathfrak{M}}{\sim} \mathcal{G}'$ if and only if $t(\mathcal{G}) = t(\mathcal{G}')$.*

PROOF. To show that the d-clique vector is preserved under Markov equivalence, by Theorem A.3 it is sufficient to show that it is preserved under a covered edge reversal. If $(a, b)$ is a covered edge of $\mathcal{G}$, then the contribution of these vertices to the sum (5.2) is

$$t(\mathcal{G}) = \left[ \delta(\{a\} \cup \mathrm{pa}_{\mathcal{G}}(a)) - \delta(\mathrm{pa}_{\mathcal{G}}(a)) \right] + \left[ \delta(\{b\} \cup \mathrm{pa}_{\mathcal{G}}(b)) - \delta(\mathrm{pa}_{\mathcal{G}}(b)) \right]$$
$$+ \sum_{v \neq a,b} \left[ \delta(\{b\} \cup \mathrm{pa}_{\mathcal{G}}(b)) - \delta(\mathrm{pa}_{\mathcal{G}}(b)) \right] + \delta(\varnothing).$$

By definition, $\mathrm{pa}_{\mathcal{G}}(a) \cup \{a\} = \mathrm{pa}_{\mathcal{G}}(b)$, and so the corresponding terms will cancel. If $\mathcal{G}^*$ is obtained from $\mathcal{G}$ by reversing $(a, b)$, note that

$$\mathrm{pa}_{\mathcal{G}}(a) = \mathrm{pa}_{\mathcal{G}^*}(b) \quad \text{and} \quad \mathrm{pa}_{\mathcal{G}}(b) \cup \{b\} = \mathrm{pa}_{\mathcal{G}^*}(a) \cup \{a\},$$

and the remaining terms will be unchanged. Hence $t(\mathcal{G}) = t(\mathcal{G}^*)$.
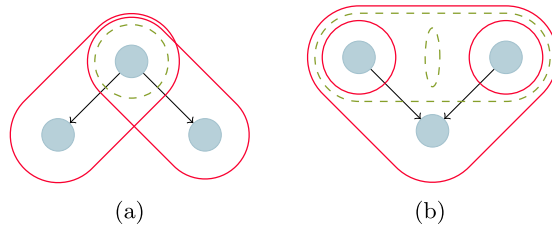
FIG. 4.   *The d-cliques (——) and d-separators (- - -) of different directed acyclic graphs. Note that in the perfect DAG* (a), *the d-cliques and d-separators are the cliques and separators of the skeleton. However, as in* (b), *d-separators may contain d-cliques.*

To show that the d-completeness vector (and hence, also the d-clique vector) is unique to the equivalence class, by Theorem A.1 we can show that it determines the skeleton and immoralities. By Lemma 5.6, there is an edge between $u$ and $v$ in $\mathcal{G}$ if and only if $c_{\{u,v\}}(\mathcal{G}) = 1$. Likewise, $(u, v, w)$ is an immorality if and only if $c_{\{u,v,w\}}(\mathcal{G}) = 1$ and $c_{\{u,w\}}(\mathcal{G}) = 0$.   □

This cancellation of terms involving covered edges is very useful: as a consequence, the d-clique vector will generally be quite sparse. In line with the clique vector, we term a set $A \subseteq V$ such that $t_A(\mathcal{D}) = 1$ a *d-clique*, and set $A$ such that $t_A(\mathcal{D}) < 0$ a *d-separator*: See examples in Figure 4.

THEOREM 5.9.   *Let A be an ancestral set of a dagoid $\mathcal{D}$. Then*

$$t(\mathcal{D}) = \left[t(\mathcal{D}_A)\right]^0 + t(\mathcal{D}_{V|A}) - \delta(A),$$

*where $[\cdot]^0$ denotes the expansion of the vector with zeroes to the required coordinates.*

PROOF.   Let $\mathcal{G} \in \mathcal{D}$ in which $A$ is ancestral, and $\prec$ be a well-ordering of $\mathcal{G}$ in which elements of $A$ precede those of $V \setminus A$. Then

$$\mathrm{pa}_{\mathcal{G}}(v) = \begin{cases} \mathrm{pa}_{\mathcal{G}_A}(v), & v \in A, \\ \mathrm{pa}_{\mathcal{G}_{V|A}}(v), & v \notin A. \end{cases}$$

The result follows after noting that

$$\sum_{v \in A} \left[\delta\left(\mathrm{pa}_{\mathcal{G}_{V|A}}(v) \cup \{v\}\right) - \delta\left(\mathrm{pa}_{\mathcal{G}_{V|A}}(v)\right)\right] = \delta(A).$$

   □

We now arrive at the key result of this section: the dagoid structural Markov property characterises an exponential family of graph laws.

THEOREM 5.10.   *Let $\mathfrak{G}$ be a graph law whose support is $\mathfrak{D}^{\mathfrak{M}}$. Then $\mathfrak{G}$ is structurally Markov if and only if it is a member of the exponential family with the d-clique vector as natural sufficient statistic, that is, if $\mathfrak{G}$ has density of the form*

(5.4)                                $\pi_\omega(\mathcal{D}) \propto \exp\{\omega \cdot t(\mathcal{D})\}.$

PROOF.   See Appendix B.   □

EXAMPLE 5.1.   As in the undirected case, the simplest example of a structurally Markov graph law is the uniform law over $\mathfrak{D}^{\mathfrak{M}}$, on taking $\omega_A = 0$.

EXAMPLE 5.2.   For any directed graph $\mathcal{G} \in \mathcal{D}$, let $e(\mathcal{D})$ denote $|\mathcal{E}(\mathcal{G})|$. By Proposition 5.5, $e(\mathcal{D}) = \sum_A \binom{|A|}{2} t_A(\mathcal{D})$. So, for any $\rho > 0$, the graph law specified by

$$\pi(\mathcal{D}) \propto \rho^{e(\mathcal{D})}$$

is structurally Markov, on taking $\omega_A = \binom{|A|}{2} \log \rho$.

However, we note that some simple laws are *not* structurally Markov.

EXAMPLE 5.3.   Consider the law in which $\pi(\mathcal{D})$ is proportional to $|\mathcal{D}|$, in other words, the uniform law on $\mathfrak{D}$ projected onto $\mathfrak{D}^{\mathfrak{M}}$. Then using $[\cdot]$ to denote Markov equivalence class, we note the size of the following dagoids:

$$[\overset{\bullet\ \ \bullet}{\bullet}] = \{\overset{\bullet\ \ \bullet}{\bullet}\},$$

$$[\overset{\bullet\rightarrow\bullet}{\bullet}] = \{\overset{\bullet\rightarrow\bullet}{\bullet}, \overset{\bullet\leftarrow\bullet}{\bullet}\},$$

$$[\overset{}{\underset{}{\vee}}] = \{\overset{}{\underset{}{\vee}}\},$$

$$[\overset{}{\underset{}{\vee}}] = \{\overset{}{\underset{}{\vee}}, \overset{}{\underset{}{\vee}}, \overset{}{\underset{}{\vee}}, \overset{}{\underset{}{\vee}}, \overset{}{\underset{}{\vee}}, \overset{}{\underset{}{\vee}}\}.$$

As a consequence, this law does not satisfy the property $\pi([\overset{}{\underset{}{\vee}}])\pi([\overset{\bullet\ \ \bullet}{\bullet}]) = \pi([\overset{\bullet\rightarrow\bullet}{\bullet}])\pi([\overset{}{\underset{}{\vee}}])$ required by Proposition 5.4.

We note that similar exponential families were proposed by Mukherjee and Speed (2008). However, they treat Markov equivalent graphs as distinct, and allow them to have different probabilities.

5.4. *Compatible distributions and laws.*   As with the undirected case, a graph law is only part of the story. For each dagoid $\mathcal{D}$, we also require a method to specify a Markov sampling distribution and a law over such sampling distributions.

DEFINITION 5.7.   Distributions $\theta$ and $\theta'$, Markov with respect to directed acyclic graphs $\mathcal{G}$ and $\mathcal{G}'$ respectively, are termed *graph compatible* if, for every vertex $v$ such that $\mathrm{pa}_{\mathcal{G}}(v) = \mathrm{pa}_{\mathcal{G}'}(v)$, there exist versions of the conditional probability distributions for $X_v | X_{\mathrm{pa}(v)}$ such that

$$\theta(X_v | X_{\mathrm{pa}(v)}) = \theta'(X_v | X_{\mathrm{pa}(v)}).$$

Distributions $\theta$ and $\theta'$, Markov with respect to dagoids $\mathcal{D}$ and $\mathcal{D}'$, respectively, are termed (*dagoid*) *compatible* if they are graph compatible for every pair of graphs $\mathcal{G} \in \mathcal{D}, \mathcal{G}' \in \mathcal{D}'$.

Likewise, laws $\pounds(\tilde{\theta})$ and $\pounds'(\tilde{\theta})$, hyper Markov with respect to $\mathcal{G}$ and $\mathcal{G}'$, respectively, are termed *graph hyper compatible* if for every vertex $v$ such that $\mathrm{pa}_{\mathcal{G}}(v) = \mathrm{pa}_{\mathcal{G}'}(v)$, there exist versions of the conditional laws for $\tilde{\theta}_{v|\mathrm{pa}(v)}|\tilde{\theta}_{\mathrm{pa}(v)}$ such that

$$\pounds(\tilde{\theta}_{v|\mathrm{pa}(v)}|\tilde{\theta}_{\mathrm{pa}(v)}) = \pounds'(\tilde{\theta}_{v|\mathrm{pa}(v)}|\tilde{\theta}_{\mathrm{pa}(v)}).$$

By Dawid (2001a), Section 8.2, the weak hyper Markov property may be characterised in terms of $\mathfrak{M}(\mathcal{G})$, and so the weak hyper Markov property can be defined with respect to a dagoid. Laws $\pounds(\tilde{\theta})$ and $\pounds'(\tilde{\theta})$, that are hyper Markov with respect to $\mathcal{D}$ and $\mathcal{D}'$, respectively, are (*dagoid*) *hyper compatible* if they are graph compatible for every pair of graphs $\mathcal{G} \in \mathcal{D}, \mathcal{G}' \in \mathcal{D}'$.

As in the undirected case, we can define a family of compatible distributions $\vartheta = \{\theta^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{U}\}$ and a family of hyper compatible laws $\mathfrak{L} = \{\pounds^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{U}\}$ if they are pairwise compatible or hyper compatible with respect to the relevant graphs.

PROPOSITION 5.11.   *Suppose* $\mathfrak{G}(\tilde{\mathcal{D}})$ *is a graph law over* $\mathfrak{D}^{\mathfrak{M}}$ *and* $\vartheta$ *is a family of compatible distributions. Then*

$$(5.5) \qquad X_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A}|\tilde{\mathcal{D}}_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \qquad [\vartheta, \mathfrak{G}]$$

*and*

$$(5.6) \qquad X_{V\setminus A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A|X_A, \tilde{\mathcal{D}}_{V|A}, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \qquad [\vartheta, \mathfrak{G}].$$

*Likewise, if* $\mathfrak{G}(\tilde{\mathcal{D}})$ *is a graph law over* $\mathfrak{D}^{\mathfrak{M}}$ *and* $\mathfrak{L}$ *is a hyper compatible family of laws, then*

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A}|\tilde{\mathcal{D}}_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \qquad [\mathfrak{L}, \mathfrak{G}]$$

*and*

$$\tilde{\theta}_{V\setminus A|A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A|\tilde{\theta}_A, \tilde{\mathcal{D}}_{V|A}, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \qquad [\mathfrak{L}, \mathfrak{G}].$$

PROOF.   This is much the same as Proposition 3.8: for (5.5), the distribution of $X_A$ is determined by the parent sets of the vertices in $A$ in some $\mathcal{G} \in \mathcal{D}$ in which $A$ is ancestral. Likewise, in (5.6), the conditional distribution for $X_{V\setminus A}|X_A$ is determined by the parent sets of vertices in $V \setminus A$. The same argument applies at the hyper level.   □

Note that in the definition of compatibility and hyper compatibility we specifically refer to *versions* of conditional probabilities and laws, as in some cases the

conditional distributions/laws will not be uniquely defined, due to conditioning on null sets.

As the weak hyper Markov property is defined on the separoid, the weak directed hyper Markov property is well-defined for any dagoid. However, the strong form requires further conditions.

DEFINITION 5.8. A law $£(\tilde{\theta})$ over $\mathfrak{P}(\mathcal{D})$ is *strong hyper Markov* with respect to $\mathcal{D}$ if it is strong directed hyper Markov with respect to every $\mathcal{G} \in \mathcal{D}$.

If $\mathcal{G} \in \mathcal{D}$ is perfect, then the strong dagoid hyper Markov property is equivalent to the undirected strong hyper Markov property on the skeleton of $\mathcal{G}$; see Dawid and Lauritzen [(1993), Proposition 3.15]. The notion of hyper compatibility is equivalent to the "parameter modularity" property of Heckerman, Geiger and Chickering (1995). Likewise, the strong hyper Markov property is equivalent to their "parameter independence."

EXAMPLE 5.4. For each vertex $v$ of a directed acyclic graph $\mathcal{G}$, we define the law for the conditional parameter $£(\tilde{\theta}_{v|\operatorname{pa}_\mathcal{G}(v)})$ to be the same as that of the inverse Wishart $\mathscr{IW}(v; \Phi)$. That is, using the notation of Dawid (1981), we have

$$\theta_{v|\operatorname{pa}_\mathcal{G}(v)} = \mathscr{N}(\Gamma_{v|\operatorname{pa}_\mathcal{G}(v)}, \Sigma_{v|\operatorname{pa}_\mathcal{G}(v)}),$$

where

$$£(\tilde{\Sigma}_{v|\operatorname{pa}_\mathcal{G}(v)}) = \mathscr{IW}(v + |\operatorname{pa}_\mathcal{G}(v)|; \Phi_{v|\operatorname{pa}_\mathcal{G}(v)}),$$

$$£(\tilde{\Gamma}_{v|\operatorname{pa}_\mathcal{G}(v)}|\tilde{\Sigma}_{v|\operatorname{pa}_\mathcal{G}(v)}) = \Phi_{\{v\},\operatorname{pa}_\mathcal{G}(v)}\Phi^{-1}_{\operatorname{pa}_\mathcal{G}(v)} + \mathscr{N}_{\{v\}\times\operatorname{pa}_\mathcal{G}(v)}(\tilde{\Sigma}_{v|\operatorname{pa}_\mathcal{G}(v)}, \Phi^{-1}_{\operatorname{pa}_\mathcal{G}(v)}).$$

By the properties of the inverse Wishart law, it follows that the law is preserved under covered edge reversals. Therefore by Theorem A.3, it is well defined for a dagoid, and so may be termed the *dagoid hyper inverse Wishart law*. Note that this property is not satisfied by the more general inverse type-II Wishart family of Letac and Massam (2007).

THEOREM 5.12. *If $\mathfrak{L}$ is a family of strong hyper Markov hyper compatible laws, then the family of marginal data distributions is compatible.*

PROOF. The hyper compatibility and the strong hyper Markov property imply that, for any two dagoids $\mathcal{D}, \mathcal{D}'$ and any $\mathcal{G} \in \mathcal{D}, \mathcal{G}' \in \mathcal{D}'$, if $\operatorname{pa}_\mathcal{G}(v) = \operatorname{pa}_{\mathcal{G}'}(v)$ for some $v \in V$, then

$$£^{(\mathcal{D})}(\tilde{\theta}_{v|\operatorname{pa}}) = £^{(\mathcal{D}')}(\tilde{\theta}_{v|\operatorname{pa}}).$$

Therefore, the family of marginal data distributions $\bar{\vartheta} = \{\bar{\theta}^{(\mathcal{D})} : \mathcal{D} \in \mathfrak{D}^{\mathfrak{M}}\}$ will have

$$\bar{\theta}^{(\mathcal{D})}(X_v|X_{\operatorname{pa}_\mathcal{G}}) = \mathbb{E}^{(\mathcal{D})}_£[\tilde{\theta}_{v|\operatorname{pa}_\mathcal{G}}] = \bar{\theta}^{(\mathcal{D}')}(X_v|X_{\operatorname{pa}_\mathcal{G}}) = \mathbb{E}^{(\mathcal{D}')}_£[\tilde{\theta}_{v|\operatorname{pa}_\mathcal{G}}]. \qquad \square$$

This is particularly useful because, as in the undirected case, the structural Markov property will be preserved in the posterior under compatible sampling.

THEOREM 5.13. *Suppose $\mathfrak{G}(\tilde{\mathcal{D}})$ is a structurally Markov graph law over $\mathfrak{D}^{\mathfrak{M}}$ and $\vartheta$ is a family of compatible distributions. Then the posterior graph law for $\tilde{\mathcal{D}}$ is structurally Markov.*

PROOF. By the structural Markov property and (5.5), we have

$$(X_A, \tilde{\mathcal{D}}_A) \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} | \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\},$$

and hence

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} | X_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}.$$

Combining this with (5.6), we get

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp (\tilde{\mathcal{D}}_{V|A}, X_{V \setminus A}) | X_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\},$$

and hence

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} | X, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}. \qquad \square$$

5.5. *Posterior updating.* If it is possible to avoid the problem of conditioning on null sets, then as in the undirected case, a compatible family can be characterised by a distribution on the complete dagoid.

THEOREM 5.14. *If the distribution on the complete dagoid has positive density p with respect to some product measure, then the compatible distribution for any dagoid $\mathcal{D}$ has density*

$$(5.7) \qquad \pi^{(\mathcal{D})}(x) = \prod_{A \subseteq V} p(x_A)^{[t(\mathcal{D})]_A}.$$

PROOF. Let $\mathcal{G}$ be an arbitrary graph in $\mathcal{D}$. Then by compatibility,

$$p^{(\mathcal{D})}(x) = \prod_{v \in V} p(x_v | x_{\mathrm{pa}(v)}) = \frac{\prod_{i=1}^{p} p(x_{\{v_i\} \cup \mathrm{pa}(v_i)})}{\prod_{i=2}^{p} p(x_{\mathrm{pa}(v_i)})} = \prod_{A \subseteq V} [p(x_A)]^{t(\mathcal{D})_A}. \qquad \square$$

As a consequence, if the graph law has a d-clique exponential family of the form (5.4), and the sampling distributions are compatible with density of the form (5.7), then the posterior graph law will have density

$$\pi(\mathcal{D}|X) \propto \exp\{[\omega + (\log p_A(X_A))_{A \subseteq V}] \cdot t(\mathcal{D})\}.$$

That is, the d-clique exponential family is a conjugate prior under sampling from a compatible family.

**6. Discussion.** We have demonstrated how conditional independence can be used to characterise families of distributions over undirected graphs, ordered directed graphs, and equivalence classes of directed graphs.

One point to emphasise is that all three structural Markov properties are distinct, in that no one property can be derived as a special case of another; for example, the undirected structural Markov property does not arise from the dagoid structural Markov property restricted to equivalence classes of perfect DAGs.

6.1. *Open questions.* One significant open question is how the full support requirements of Theorems 3.15 and 5.10 might be weakened. Obviously these theorems would not hold for all subsets of graphs/dagoids, though we conjecture that they will hold for any structurally meta Markov subsets. A related problem is characterising structurally meta Markov subsets of graphs.

6.2. *Computation.* One problem which we have not broached is the numerical calculation of such graph laws. Except for the ordered directed case, where the computations can be done in parallel, for even small numbers of vertices it can quickly become infeasible to enumerate all graphs, and hence some sort of numerical approximation will usually be required. Markov chain Monte Carlo (MCMC) methods are commonly utilised for this purpose.

For undirected decomposable graphs, Giudici and Green (1999) proposed a method in which each iteration proposes adding or removing a single edge. They consider the problem of sampling from the posterior of a uniform prior with a compatible family sampling distributions, though this procedure can be applied to any structural Markov graph law. This requires computing the Metropolis–Hastings acceptance ratio,

$$\min\left(\frac{\pi(\mathcal{G}')}{\pi(\mathcal{G})}, 1\right) = \begin{cases} \min(\exp\{\omega \cdot [t(\mathcal{G}') - t(\mathcal{G})]\}, 1), & \mathcal{G}, \mathcal{G}' \in \mathfrak{U}, \\ 0, & \text{otherwise.} \end{cases}$$

The results of Frydenberg and Lauritzen [(1989), Lemma 3] and Giudici and Green [(1999), Theorem 2] characterise such so-called neighbouring graphs, and also imply that for any two such graphs $\mathcal{G}, \mathcal{G}'$, the vector $t(\mathcal{G}') - t(\mathcal{G})$ has only 4 nonzero elements. Consequently, for any structurally Markov graph law over $\mathfrak{U}$, the parameter $\omega$ need only be evaluated on 4 such places: this is particularly beneficial for posterior graph laws where each element of $\omega$ requires the evaluation of the marginal density of the model.

Unfortunately, such algorithms often exhibit poor mixing properties [Kijima et al. (2008)], resulting in unreliable estimates. Green and Thomas (2013) develop an extension for making proposals which add or remove multiple edges, resulting in faster mixing: this algorithm is also able to take advantage of local computations in computing the acceptance ratio.

For dagoids, the problem is considerably more difficult. Chickering (2003), Auvray and Wehenkel (2002) and Studený (2005a) have developed methods for

characterising the neighbouring dagoids (i.e., dagoids obtained by adding or removing an edge to a graph in the current dagoid). He, Jia and Yu (2013) recently developed an MCMC scheme based on this approach: as in the undirected case, the acceptance ratio will also depend on a sparse vector, and so can be computed efficiently.

More generally, the problem of finding the most probable graph under a structural Markov law, which for posterior laws is known as maximum *a posteriori* (MAP) estimation, is an example of a *strong decomposable search criterion* [Studený (2005b), Section 8.2.3]. As suggested by Hemmecke, Lindner and Studený (2012), linear and integer programming techniques based on the (d-)clique or (d-)completeness vectors may provide elegant solutions to this problem.

6.3. *Extensions.* A further open question is how structural Markov properties might be defined for other classes of graphical models, such as nondecomposable undirected graphs, ancestral graphs, and marginal independence (bidirected) graphs. The identification of such properties would rely on establishing constructions for partitioning the structure, analogous to decompositions and ancestral graphs.

### APPENDIX A: CHARACTERISING MARKOV EQUIVALENCE OF DIRECTED ACYCLIC GRAPHS

Numerous techniques have been developed for determining whether two graphs are Markov equivalent.

**A.1. Skeleton and immoralities.** The *skeleton* of a DAG is the undirected graph obtained by substituting the directed edges for undirected ones. A triplet $(a, b, c)$ of vertices is an *immorality* of a DAG $\mathcal{G}$ if the induced graph $\mathcal{G}_{\{a,b,c\}}$ is of the form $a \to b \leftarrow c$.

THEOREM A.1 [Frydenberg (1990), Theorem 5.6; Verma and Pearl (1990), Theorem 1]. *Directed acyclic graphs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if they have the same skeleton and the same immoralities.*

**A.2. Essential graphs.** An edge of a DAG $\mathcal{G}$ is *essential* if it has the same direction in all Markov equivalent DAGs. The *essential graph* of $\mathcal{G}$ is the graph in which all nonessential edges are replaced by undirected edges.

Although not explored further in this work, the essential graph is a type of *chain graph*, a class of graphs that may have both directed and undirected edges. For further details on chain graphs, in particular their Markov properties and how they relate to undirected and directed acyclic graphs, see Frydenberg (1990) and Andersson, Madigan and Perlman (1997b).

THEOREM A.2 [Andersson, Madigan and Perlman (1997a), Proposition 4.3]. *Directed acyclic graphs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if they have the same essential graph.*

Unfortunately, there is no simple criterion for determining whether or not an edge of a given DAG is essential, although Andersson, Madigan and Perlman (1997a) developed an iterative algorithm. This limits their usefulness.

**A.3. Covered edge reversals.** A convenient characterisation of Markov equivalence can be given in terms of edge reversals. An edge $a \to b$ of a DAG $\mathcal{G}$ is *covered* if $\text{pa}(b) = \text{pa}(a) \cup \{a\}$.

THEOREM A.3 [Chickering (1995), Theorem 2]. *Directed acyclic graphs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if there exists a sequence of DAGs*

$$\mathcal{G} = \mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_{k-1}, \mathcal{G}_k = \mathcal{G}'$$

*such that each $(\mathcal{G}_{i-1}, \mathcal{G}_i)$ differ only by the reversal of one covered edge.*

This result is particularly useful for identifying properties that are preserved under Markov equivalence, as it is only necessary to show that the property is preserved under a covered edge reversal.

**A.4. Standard imset.** Imsets for undirected decomposable graphs were briefly mentioned in Section 3.4. This formalism can be extended to directed acyclic graphs. The *standard imset* of a directed acyclic graph $\mathcal{G}$ is [Studený (2005b), page 135]

$$u_{\mathcal{G}} = \delta(V) - \delta(\varnothing) + \sum_{v \in V} \big[ \delta(\text{pa}_{\mathcal{G}}(v)) - \delta(\text{pa}_{\mathcal{G}}(v) \cup \{v\}) \big],$$

where $\delta(A) = (\mathbb{1}_{S=A})_{S \subseteq V}$.

THEOREM A.4 [Studený (2005b), Corollary 7.1]. *Directed acyclic graphs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if $u_{\mathcal{G}} = u_{\mathcal{G}'}$.*

Studený and Vomlel (2009) give details of the relationship between the imset and the essential graph of a DAG, and how one may be obtained from the other.

## APPENDIX B: PROOFS

PROOF OF THEOREM 3.9. The Markov property states that under $[\mathfrak{G}, \vartheta]$,

(B.1) $$X_A \perp\!\!\!\perp X_B | X_{A \cap B}, \tilde{\mathcal{G}}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}.$$

Since if $\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$, then $\tilde{\mathcal{G}} \simeq (\tilde{\mathcal{G}}_A, \tilde{\mathcal{G}}_B)$, we can rewrite (B.1) as

(B.2)  $$X_A \perp\!\!\!\perp X_B | X_{A\cap B}, \tilde{\mathcal{G}}_A, \tilde{\mathcal{G}}_B, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}.$$

As a consequence of Proposition 3.8,

(B.3)  $$X_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B | X_{A\cap B}, \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\},$$

and combined with (B.2),

(B.4)  $$X_A \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) | X_{A\cap B}, \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}.$$

Furthermore, by the structural Markov property and Proposition 3.8,

(B.5)  $$\tilde{\mathcal{G}}_A \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) | \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\},$$

and we can further condition on $X_{A\cap B}$. The result follows from this and (B.4). $\quad\square$

PROOF OF THEOREM 3.15.   For any $C \subseteq V$, define $\mathcal{G}^{(C)}$ as in the proof of Theorem 3.13, and let $\mathfrak{G}$ have density $\pi$.

Suppose that $\mathfrak{G}$ is structurally Markov. For any $\mathcal{G} \in \mathfrak{U}$, let $C_1, \ldots, C_k$ be a perfect ordering of the cliques, and let $S_2, \ldots, S_k$ be the corresponding separators, and $H_i = C_1 \cup \cdots \cup C_i$. Furthermore, recursively define the graphs

$$\mathcal{G}^{*(j)} = \begin{cases} \mathcal{G}^{(C_1)}, & \text{if } j = 1, \\ \mathcal{G}^{*(j-1)}_{H_{j-1}} \otimes \mathcal{G}^{(C_j)}_{(V\setminus H_{j-1})\cup S_j}, & \text{if } j = 2, \ldots, k. \end{cases}$$

By Proposition 3.2, for each $j = 2, \ldots, k$,

$$\pi\big(\mathcal{G}^{*(j)}\big)\pi\big(\mathcal{G}^{(S_j)}\big) = \pi\big(\mathcal{G}^{*(j-1)}\big)\pi\big(\mathcal{G}^{(C_j)}\big).$$

Note that $\mathcal{G}^{*(k)} = \mathcal{G}$. Then, by induction,

$$\pi(\mathcal{G}) = \frac{\prod_{j=1}^{k}\pi(\mathcal{G}^{(C_j)})}{\prod_{j=2}^{k}\pi(\mathcal{G}^{(S_j)})} \propto \exp\{\omega \cdot t(G)\}$$

by Theorem 3.13, where $\omega_C = \log\pi(\mathcal{G}^{(C)})$.

To show the converse let $(\omega)_A = (\omega_S)_{S\subseteq A}$. By Lemma 3.12,

$$\pi\big(\mathcal{G}_A | \mathcal{G}_B, \{\mathcal{G} \in \mathfrak{U}(A, B)\}\big)$$

$$\propto \exp\{(\omega)_A \cdot t(\mathcal{G}_A) + (\omega)_B \cdot t(\mathcal{G}_B) - (\omega)_{A\cap B} \cdot t(\mathcal{G}_{A\cap B})\}$$

$$\propto \exp\{(\omega)_A \cdot t(\mathcal{G}_A) - (\omega)_{A\cap B} \cdot t(\mathcal{G}_{A\cap B})\}$$

$$\propto \pi\big(\mathcal{G}_A | \{\mathcal{G} \in \mathfrak{U}(A, B)\}\big). \qquad\qquad \square$$

PROOF OF THEOREM 5.10.   If the law is in the exponential family (5.4), then, by Theorem 5.9,

$$\pi\big(\mathcal{D} | \mathfrak{D}(A)\big) \propto \exp\{\omega \cdot [t(\mathcal{D}_A) + t(\mathcal{D}_{V|A})] - \omega_A\} \propto p\big(\mathcal{D}_A | \mathfrak{D}(A)\big)p\big(\mathcal{D}_{V|A} | \mathfrak{D}(A)\big),$$

and hence the law must be structurally Markov.

For the converse, define $\mathcal{D}^{(A)}$ to be the dagoid in which the induced dagoid on $A \subseteq V$ is complete, but otherwise sparse (in other words, the remainder dagoid, $\mathcal{D}_{V|A}^{(\varnothing)}$, of the sparse dagoid $\mathcal{D}^{(\varnothing)}$ corresponding to complete independence).

Select some $\mathcal{G} \in \mathcal{D}$, and let $v_1, \ldots, v_d$ be a well-ordering of $V$. Recursively define the dagoids

$$\mathcal{D}^{*(i)} = \begin{cases} \mathcal{D}^{(\{v_1\})}, & \text{if } i = 1, \\ \mathcal{D}_{\mathrm{pr}(v_i)}^{*(i-1)} \ltimes \mathcal{D}_{v_i \mid \mathrm{pr}(v_i)}^{(\{v_i\} \cup \mathrm{pa}(v_i))}, & \text{otherwise.} \end{cases}$$

By Proposition 5.4, for $i = 2, \ldots, d$,

$$\pi\big(\mathcal{D}^{*(i-1)}\big) \pi\big(\mathcal{D}^{(\{v_i\} \cup \mathrm{pa}(v_i))}\big) = \pi\big(\mathcal{D}^{*(i)}\big) \pi\big(\mathcal{D}_{\mathrm{pr}(v_i)}^{(\{v_i\} \cup \mathrm{pa}(v_i))} \ltimes \mathcal{D}_{v_i \mid \mathrm{pr}(v_i)}^{*(i-1)}\big).$$

However,

$$\mathcal{D}_{\mathrm{pr}(v_i)}^{(\{v_i\} \cup \mathrm{pa}(v_i))} \ltimes \mathcal{D}_{v_i \mid \mathrm{pr}(v_i)}^{*(i-1)} = \mathcal{D}^{(\mathrm{pa}(v_i))}.$$

Therefore, since $\mathcal{D}^{*(d)} = \mathcal{D}$,

$$\pi(\mathcal{D}) = \left[\prod_{i=1}^{d} \pi\big(\mathcal{D}^{(\{v_i\} \cup \mathrm{pa}(v_i))}\big)\right] \Big/ \left[\prod_{i=2}^{d} \pi\big(\mathcal{D}^{(\mathrm{pa}(v_i))}\big)\right],$$

which is of the form in (5.4) with

$$\omega_A = \log \pi\big(\mathcal{D}^{(A)}\big). \qquad \square$$

## REFERENCES

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997a). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997b). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scand. J. Stat.* **24** 81–102. MR1436624

ARMSTRONG, H., CARTER, C. K., WONG, K. F. K. and KOHN, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Stat. Comput.* **19** 303–316. MR2516221

ASMUSSEN, S. and EDWARDS, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70** 567–578. MR0725370

AUVRAY, V. and WEHENKEL, L. (2002). On the construction of the inclusion boundary neighbourhood for Markov equivalence classes of Bayesian network structures. In *Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence* (A. Darwiche and N. Friedman, eds.) 26–35. Morgan Kaufmann, San Francisco, CA.

BORNN, L. and CARON, F. (2011). Bayesian clustering in decomposable graphs. *Bayesian Anal.* **6** 829–845. MR2869966

BROOKS, S. P., GIUDICI, P. and ROBERTS, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 3–55. MR1959092

CASTELO, R. and KOČKA, T. (2004). On inclusion-driven learning of Bayesian networks. *J. Mach. Learn. Res.* **4** 527–574. MR2072261

CHICKERING, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence* (*Montreal*, *PQ*, 1995) (P. Besnard and S. Hanks, eds.) 87–98. Morgan Kaufmann, San Francisco, CA. MR1615012

CHICKERING, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554. MR1991085

COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (2007). *Probabilistic Networks and Expert Systems*. Springer, New York.

DAWID, A. P. (1979). Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **41** 1–31.

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963

DAWID, A. P. (2001a). Separoids: A mathematical framework for conditional independence and irrelevance. *Ann. Math. Artif. Intell.* **32** 335–372. MR1859870

DAWID, A. P. (2001b). Some variations on variation independence. In *Artificial Intelligence and Statistics* 2001 (T. Jaakkola and T. Richardson, eds.) 187–191. Morgan Kaufmann, San Francisco, CA.

DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267

FRYDENBERG, M. (1990). The chain graph Markov property. *Scand. J. Stat.* **17** 333–353. MR1096723

FRYDENBERG, M. and LAURITZEN, S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* **76** 539–555. MR1040647

GIUDICI, P. and GREEN, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86** 785–801. MR1741977

GREEN, P. J. and THOMAS, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika* **100** 91–110. MR3034326

HE, Y., JIA, J. and YU, B. (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *Ann. Statist.* **41** 1742–1779. MR3127848

HECKERMAN, D., GEIGER, D. and CHICKERING, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.* **20** 197–243.

HEMMECKE, R., LINDNER, S. and STUDENÝ, M. (2012). Characteristic imsets for learning Bayesian network structure. *Internat. J. Approx. Reason.* **53** 1336–1349. MR2994270

JONES, B., CARVALHO, C., DOBRA, A., HANS, C., CARTER, C. and WEST, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.* **20** 388–400. MR2210226

KIJIMA, S., KIYOMI, M., OKAMOTO, Y. and UNO, T. (2008). On listing, sampling, and counting the chordal graphs with edge constraints. In *Computing and Combinatorics*. *Lecture Notes in Computer Science* **5092** 458–467. Springer, Berlin. MR2473446

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

LAURITZEN, S. L., SPEED, T. P. and VIJAYAN, K. (1984). Decomposable graphs and hypergraphs. *Austral. Math. Soc. Lect. Ser.* **36** 12–29. MR0719998

LETAC, G. and MASSAM, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35** 1278–1323. MR2341706

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MUKHERJEE, S. and SPEED, T. P. (2008). Network inference using informative priors. *Proc. Natl. Acad. Sci. USA* **105** 14313–14318.

STUDENÝ, M. (2005a). Characterization of inclusion neighbourhood in terms of the essential graph. *Internat. J. Approx. Reason.* **38** 283–309. MR2116940

STUDENÝ, M. (2005b). *Probabilistic Conditional Independence Structures*. Springer, London. MR3183760

STUDENÝ, M. and VOMLEL, J. (2009). A reconstruction algorithm for the essential graph. *Internat. J. Approx. Reason.* **50** 385–413. MR2514506

VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence* (P. Bonissone, M. Henrion, L. Kanal and J. Lemmer, eds.) 220–227. Elsevier Science, New York, NY.

WORMALD, N. C. (1985). Counting labelled chordal graphs. *Graphs Combin.* **1** 193–200. MR0951781

DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
GOWER STREET, LONDON WC1E 6BT
UNITED KINGDOM
E-MAIL: simon.byrne@ucl.ac.uk

STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD, CAMBRIDGE CB3 0WB
UNITED KINGDOM
E-MAIL: apd@statslab.cam.ac.uk