Towards standard imsets for maximal ancestral graphs

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

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

August 21, 2023

Abstract

The imsets of Studený (2005) are an algebraic method for representing conditional independence models. They have many attractive properties when applied to such models, and they are particularly nice for working with directed acyclic graph (DAG) models. In particular, the 'standard' imset for a DAG is in one-to-one correspondence with the independences it induces, and hence is a label for its Markov equivalence class. We first present a proposed extension to standard imsets for maximal ancestral graph (MAG) models, using the parameterizing set representation of Hu and Evans (2020). We show that for many such graphs our proposed imset is *perfectly Markovian* with respect to the graph, including a class of graphs we refer to as *simple* MAGs, which includes DAGs as a special case. In these cases the imset provides a scoring criteria by measuring the discrepancy for a list of independences that define the model; this gives an alternative to the usual BIC score that is also consistent, and much easier to compute. We also show that, of independence models that do represent the MAG, the imset we give is minimal. Unfortunately, for some graphs the representation does not represent all the independences in the model, and in certain cases does not represent any at all. For these general MAGs, we refine the reduced ordered local Markov property (Richardson, 2003) by a novel graphical tool called *power DAGs*, and this results in an imset that induces the correct model and which, under a mild condition, can be constructed in polynomial time.

1 Introduction

Maximal ancestral graphs (MAGs) (Richardson and Spirtes, 2002) are used to model distributions via *conditional independence* (CI) relations. They are an extension of models based on *directed acyclic graphs* (DAGs), as MAG models remove the assumption of no unobserved common parents and allow data arising from distributions with a more general independence structure. These graphs have been proven to be useful in various scenarios, for example, to infer causal effects from observational data. Therefore learning the best graph associated with a given dataset is a crucially important task.

There are three classical types of learning algorithm for graphical models: as well as constraint-based and score-based methods, there are hybrid methods that combine the first two approaches. The canonical constraint-based method for learning DAGs/MAGs is the PC/FCI algorithm (Spirtes et al., 2000). Briefly speaking, this type of method tests for conditional independences in the empirical distribution, and uses the results to reconstruct

the graph. The problem of constraint-based methods is that when the group of variables is large, it is likely that a mistake in testing a conditional independence will be made; this error will be propagated through the algorithm, and the resulting graph will not reflect the true independence structure generating the data (see, e.g. Ramsey et al., 2006; Evans, 2020).

Score-based methods tend to be more accurate. The main idea is that they go through many graphs and select the graph with the highest score, which asymptotically gives the correct graph. There now exist several score-based methods for MAGs (Triantafillou and Tsamardinos, 2016; Rantanen et al., 2021; Chen et al., 2021; Claassen and Bucur, 2022). There are two main problems with these approaches. First, the score used for DAGs is generally the Bayesian information criterion (BIC, Chickering, 2002; Jaakkola et al., 2010), which is also used by the above methods for MAGs. This scoring criteria is known to be *consistent*, that is, in the limit of infinite samples, models with lower complexity that contain the true model always have better BIC compared to higher complexity models. However, although methods for fitting Gaussian or discrete MAG models via maximum likelihood have been given by Drton et al. (2009) and Evans and Richardson (2010, 2014) (so the corresponding BIC score can be obtained), MLEs are not available in closed-form, and therefore they need to be computed iteratively using a numerical method. The functions being maximized are also not generally convex if the model is not a DAG, and so these iterative algorithms may converge to a point which is not globally optimal. Moreover, the factorisation of distributions in MAG models is complicated (Richardson, 2009), and the scores are only decomposable with respected to the components connected by bidirected paths, also known as districts or c-components. In contrast, for DAGs the MLE is available in closed-form and the BIC score can be decomposed in terms of individual variables and their parent sets.

Another problem is how to reduce the number of graphs visited. Logically we could score every graph, but this is obviously hopelessly inefficient. Graphs that represent the same CI relations are said to be in the same *Markov equivalence class* (MEC). Graphs in the same MEC generally have the same score, and so it is a waste of time to score other graphs in the same class. Among the previous works mentioned, only Claassen and Bucur (2022) explore graphs in the space of MECs. However their method uses the BIC score, which is computationally expensive. We will see that the framework of imsets (Studený, 2005) combined with another representation of MECs (Hu and Evans, 2020), provides a solution to address the above two issues at the same time.

1.1 Imsets

Imsets are 'integer valued vectors indexed by subsets of the variables' that can be used to encode arbitrary CI models. Imsets are very useful in the context of graphical models, because DAG models fit into this representation very elegantly. The imset used to represent a DAG model is usually the 'standard imset', meaning that it is the simplest imset that correctly and precisely represents the conditional independences implied by the corresponding graph. For DAGs, there are several advantages to using standard imsets. Firstly, when two DAGs represent the same CI model, in other words they are Markov equivalent, their standard imsets also agree. Moreover, the BIC score of a DAG (defined in Appendix 3.6) is an affine function of the standard imset. So we can see that for DAGs, imsets not only provide a representation of the MEC but also lead to a consistent scoring criterion.

In this paper, we propose a standard imset formula for MAGs which, though it does not always define the same model as the graph, and sometimes does not define a model at all, works for large sub-classes of MAGs. We use the *parametrizing set* that arises in the discrete parametrization and a factorization theorem of MAG models (Richardson, 2009; Hu and Evans, 2020). One of the motivations is that for DAGs, the *characteristic* imset $c_\mathcal{G}$ introduced by Studený et al. (2010), agrees with the *parametrizing set*; since there is a bijection between the characteristic and standard imsets, we use the parameterizing set to define the characteristic imset for MAGs, and then work backwards to deduce an expression for the standard imset. Another motivation is that Hu and Evans (2020) show that two MAGs are Markov equivalent if and only if they have the same *parametrizing set*. Thus by construction two MAGs agree on their standard imsets if and only if they are Markov equivalent.

For general MAGs, we provide an imset from a new Markov property. The imset always defines the same model as the graph it is derived from, and can be constructed in polynomial time under mild assumptions on the graph structure.

The paper is organised as follows. The definitions for basic terminology in graphical models and imsets are given in Section 2. In Section 3, we present our main results on:

- the formula of the 'standard' imset $u_\mathcal{G}$ and its properties;

- when the MAGs are *simple* (no head of size three or more), the imset does define the right model;

- a proof that $\mathcal{I}_{u_\mathcal{G}} \subseteq \mathcal{I}_\mathcal{G}$ for all MAGs $\mathcal{G}$ (provided that $\mathcal{I}_{u_\mathcal{G}}$ is well-defined), where $\mathcal{I}_{u_\mathcal{G}}$ and $\mathcal{I}_\mathcal{G}$ are lists of conditional independences implied by the imset $u_\mathcal{G}$ and graph $\mathcal{G}$ respectively.

If $\mathcal{I}_{u_\mathcal{G}} \subsetneq \mathcal{I}_\mathcal{G}$, this means that the imset we define does not include all the conditional independences implied by the graph. Full definitions will be provided in Section 2.

In Section 4, we introduce the concept of *power DAGs*, which are DAGs on certain subsets of vertices. This is inspired by the decomposition of $u_\mathcal{G}$ given in Section 4.5 and it helps us to give a list of independences, which is equivalent to, but shorter than, the list of independences given by the ordered local Markov property. This results in an imset for general MAGs.

In Section 5, we report our experimental results on which graphs are perfectly Markovian ($\mathcal{I}_\mathcal{G} = \mathcal{I}_{u_\mathcal{G}}$) when there are at most seven nodes. Then we discuss related and future work in Section 6.

We also study our imsets of bidirected graphs and show that for a large class of bidirected graphs, the imset defines the right model but due to limit of space, we present our results in Appendix F. The condition we give is proved to be sufficient and empirically we checked that it is also necessary for graphs with at most seven vertices. The condition, however, is complicated, thus we conjecture that it is combinatorially difficult to obtain a minimal list of independences that define the model. We also give a list of forbidden induced subgraphs; that is, a motif that graphs cannot contain if $\mathcal{I}_{u_\mathcal{G}}$ is to be valid and be perfectly Markovian with respect to $\mathcal{G}$.

Lastly, the MAGs considered in this paper are all *directed* and contain no undirected edges. Such MAGs are sometimes referred as *directed MAGs* by other authors. To obtain imsets for MAGs an with undirected component, one can first obtain an imset for the directed

part as though the undirected component were complete, then obtain another imset for the undirected component (Kashimura and Takemura, 2015), and add two imsets together. If the directed part alone has an imset that defines the model for the subgraph, then this combination will also define the model for the whole graph.

1.2 Related work

Similar work has been done by Andrews (2021); Andrews et al. (2022), who call the parametrizing set of MAGs the *m-connecting* set. In particular, we have the same initial motivation as them, which is the similarity between the 0-1 characteristic imset (Studený et al., 2010) and parametrizing sets (Hu and Evans, 2020) of DAGs. Our work contains results distinct from those works; we give a more detailed comparison in Section 6.

2 Definitions

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

A *path* of length k is an alternating sequence of $k+1$ distinct vertices v_i and edges connecting v_i and v_{i+1} for $i = 0, \ldots, k-1$. Note that if there is more than one edge between a pair of vertices, it is ambiguous which edge is used in a path; however, we will only ever use paths in graphs which are *simple* (i.e. there is at most one edge between any pair of vertices). A path is *directed* if its edges are all directed and point from v_i to v_{i+1}. A *directed cycle* is a directed path of length $k \geq 1$ plus the edge $v_k \rightarrow v_0$, and a graph $\mathcal{G}$ is *acyclic* if it has no directed cycle. A *graph* $\mathcal{G}$ is called an *acyclic directed mixed graph* (ADMG) if it is *acyclic* and contains only directed and bidirected edges.

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

$$\mathrm{pa}_{\mathcal{G}}(v) = \{w : w \rightarrow v \text{ in } \mathcal{G}\} \qquad \mathrm{sib}_{\mathcal{G}}(v) = \{w : w \leftrightarrow v \text{ in } \mathcal{G}\}$$
$$\mathrm{an}_{\mathcal{G}}(v) = \{w : w \rightarrow \cdots \rightarrow v \text{ in } \mathcal{G} \text{ or } w = v\} \quad \mathrm{de}_{\mathcal{G}}(v) = \{w : v \rightarrow \cdots \rightarrow w \text{ in } \mathcal{G} \text{ or } w = v\}$$
$$\mathrm{dis}_{\mathcal{G}}(v) = \{w : w \leftrightarrow \cdots \leftrightarrow v \text{ in } \mathcal{G} \text{ or } w = v\}.$$

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

A *topological ordering* is an ordering on the vertices such that if $w \in \mathrm{an}_{\mathcal{G}}(v)$ then w precedes v in the ordering. There might be several topological orderings for any single graph. For an ADMG $\mathcal{G}$, given a subset $W \subseteq \mathcal{V}$, the *induced subgraph* $\mathcal{G}_W$ is defined as the graph with vertex set W and edges in $\mathcal{G}$ whose endpoints are both in W. Also for operators on an induced subgraph $\mathcal{G}_W$, we may denote them by the shorthand (e.g.) $\mathrm{dis}_W(v) := \mathrm{dis}_{\mathcal{G}_W}(v)$ if $\mathcal{G}$ is clear.
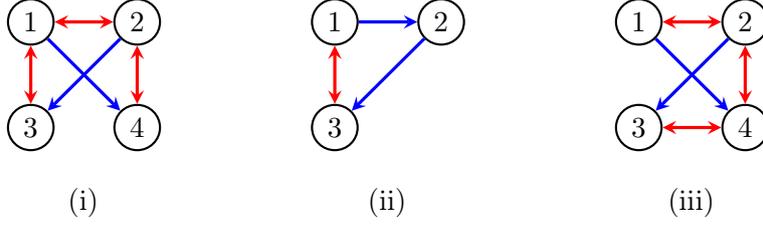
4

Figure 1: (i) An ancestral graph that is not maximal. (ii) A maximal graph that is not ancestral. (iii) A maximal ancestral graph.

For the set of vertices $\mathcal{V}$ in an ADMG $\mathcal{G}$, we associate them with random variables $(X_v)_{v \in \mathcal{V}}$ which are defined in real finite dimensional probability spaces. Let $X_A = (X_v)_{v \in A}$; we will sometimes abuse notation by using A to denote both a set of vertices and the corresponding collection of variables X_A.

Graphs are associated with conditional independence relations via a separation criterion; in the case of mixed graphs we use *m-separation*, which is defined in Appendix A.1. Distributions are associated with graphs via different but equivalent Markov properties (lists of independences). We will use the (reduced) *ordered local Markov property*.

Definition 2.1. If a set A satisfies $\mathrm{an}_{\mathcal{G}}(A) = A$, then it is an *ancestral set*.

If A is an ancestral set in an ADMG $\mathcal{G}$, and v is a vertex in A such that $\mathrm{ch}_{\mathcal{G}}(v) \cap A = \emptyset$, then the *Markov blanket* of v with respect to A is defined as:

$$\mathrm{mb}_{\mathcal{G}}(v, A) = \mathrm{pa}_A(\mathrm{dis}_A(v)) \cup \mathrm{dis}_A(v) \setminus \{v\}.$$

Definition 2.2. A distribution $P(X_{\mathcal{V}})$ is said to satisfy the *ordered local Markov property* with respect to an ADMG $\mathcal{G}$ if for any ancestral set A and a childless vertex v in A,

$$X_v \perp\!\!\!\perp X_{A \setminus (\mathrm{mb}(v, A) \cup \{v\})} \mid X_{\mathrm{mb}(v, A)} \quad [P].$$

It is possible to reduce this ordered local Markov property to a smaller collection of ancestral sets, as first shown by Richardson (2003).

Definition 2.3. For an ancestral set A and a childless vertex $v \in A$, we say that A is *maximal* with respect to $\mathrm{mb}_{\mathcal{G}}(v, A)$ if, for every ancestral set B such that $A \subseteq B$ and $\mathrm{mb}_{\mathcal{G}}(v, A) = \mathrm{mb}_{\mathcal{G}}(v, B)$, we have $A = B$.

The *reduced ordered local Markov property* is then the ordered local Markov property but restricted to only maximal ancestral sets.

2.1 MAGs

Definition 2.4. An ADMG $\mathcal{G}$ is called a *maximal ancestral graph* (MAG), if:

(i) for every pair of *nonadjacent* vertices a and b, there exists some set C such that a, b are m-separated given C in $\mathcal{G}$ (*maximality*);

(ii) for every $v \in \mathcal{V}$, $\mathrm{sib}_{\mathcal{G}}(v) \cap \mathrm{an}_{\mathcal{G}}(v) = \emptyset$ (*ancestrality*).

For example, the graph in Figure 1(i) is not maximal because 1 and 2 are not adjacent, but no subset of $\{3, 4\}$ will m-separate them; (ii) is not ancestral as 1 is a sibling of 3, which is

also one of its descendants; and (iii) is a MAG, in which the only conditional independence is $X_1 \perp\!\!\!\perp X_3 \mid X_2$.

Definition 2.5. Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with the same vertex sets, are said to be *Markov equivalent* if any m-separation holds in $\mathcal{G}_1$ if and only if it holds in $\mathcal{G}_2$.

For every ADMG $\mathcal{G}$, we can project it to a MAG $\mathcal{G}^m$ such that $\mathcal{G}$ is Markov equivalent to $\mathcal{G}^m$, and $\mathcal{G}^m$ preserves the ancestral relations in $\mathcal{G}$ (Richardson and Spirtes, 2002). Moreover, Hu and Evans (2020) show that the heads and tails defined below (and so the parametrizing set) are preserved through the projection. Hence in this paper, we will only consider MAGs.

2.2   Heads and tails

A head is a subset of vertices with a corresponding tail. The concept of heads and tails originates from Richardson (2009), who provides a factorization theorem for ADMGs. The heads and tails are analogous to vertices and parents set in DAGs, but are generalized to set of vertices; in particular, the tail of a head H is the Markov blanket of H within the ancestral set $\mathrm{an}_{\mathcal{G}}(H)$.

Definition 2.6. For a vertex set $W \subseteq \mathcal{V}$, we define the *barren subset* of W as:

$$\mathrm{barren}_{\mathcal{G}}(W) = \{w \in W : \deg_{\mathcal{G}}(w) \cap W = \{w\}\},$$

which is the subset of W such that each node has no non-trivial descendants in W.

Definition 2.7. A vertex set H is called a *head* if:

(i) $\mathrm{barren}_{\mathcal{G}}(H) = H$;

(ii) H is contained in a single district in $\mathcal{G}_{\mathrm{an}(H)}$.

For an ADMG $\mathcal{G}$, we denote the set of all heads in $\mathcal{G}$ by $\mathcal{H}(\mathcal{G})$. The *tail* of a head is defined as:

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

Remark 1. Note that for a head H and each $h \in H$, the set $(H \setminus \{h\}) \cup \mathrm{tail}_{\mathcal{G}}(H)$ is the Markov blanket of h within $\mathrm{an}_{\mathcal{G}}(H)$.

Definition 2.8. The *parametrizing set* of $\mathcal{G}$, denoted by $\mathcal{S}(\mathcal{G})$ is defined as:

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

In Figure 2, we give an example for heads and tails, as well as the sets not contained in the parametrizing set.

For a conditional independence $I = \langle A \perp\!\!\!\perp B \mid C \rangle$, let

$$\bar{\mathcal{S}}(I) = \{A' \cup B' \cup C' : \emptyset \subset A' \subseteq A, \ \emptyset \subset B' \subseteq B, \ \emptyset \subseteq C' \subseteq C\}$$

be the (*constrained*) sets associated with I. These can be thought of as the set of interactions that are constrained to be zero by I, within the margin over $A \cup B \cup C$. Proposition 3.6 shows that a set S is *not* in $\mathcal{S}(\mathcal{G})$ if and only if $S \in \bar{\mathcal{S}}(I)$ for some conditional independence I entailed by the graph.

Hu and Evans (2020) has a detailed introduction to the concept of heads and tails, and we recommend it for background reading. They prove that Markov equivalent MAGs have the same parametrizing set. We will explain this in more detail as we will use this result.
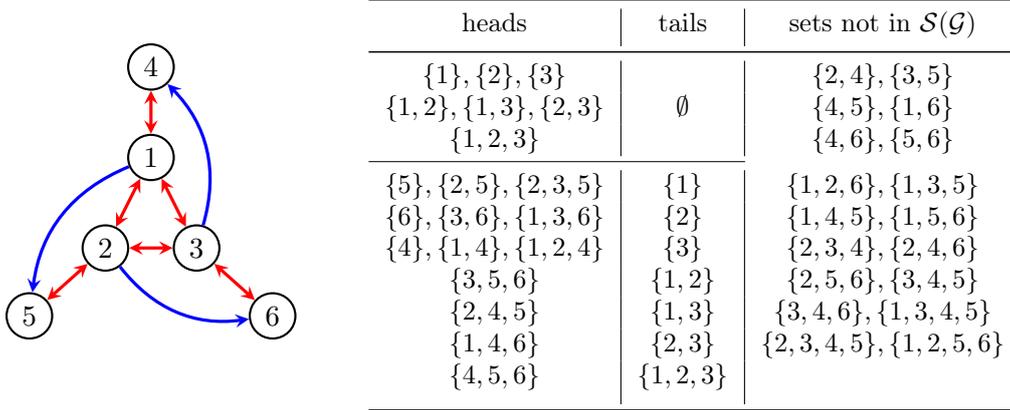
heads	tails	sets not in $\mathcal{S}(\mathcal{G})$
$\{1\},\{2\},\{3\}$ $\{1,2\},\{1,3\},\{2,3\}$ $\{1,2,3\}$	$\emptyset$	$\{2,4\},\{3,5\}$ $\{4,5\},\{1,6\}$ $\{4,6\},\{5,6\}$
$\{5\},\{2,5\},\{2,3,5\}$ $\{6\},\{3,6\},\{1,3,6\}$ $\{4\},\{1,4\},\{1,2,4\}$ $\{3,5,6\}$ $\{2,4,5\}$ $\{1,4,6\}$ $\{4,5,6\}$	$\{1\}$ $\{2\}$ $\{3\}$ $\{1,2\}$ $\{1,3\}$ $\{2,3\}$ $\{1,2,3\}$	$\{1,2,6\},\{1,3,5\}$ $\{1,4,5\},\{1,5,6\}$ $\{2,3,4\},\{2,4,6\}$ $\{2,5,6\},\{3,4,5\}$ $\{3,4,6\},\{1,3,4,5\}$ $\{2,3,4,5\},\{1,2,5,6\}$

Figure 2: A MAG $\mathcal{G}$ and a table for its heads and tails.

2.3 Imsets

We now introduce the framework of imsets and conditional independence of Studený (2005). Let $\mathcal{P}(\mathcal{V})$ be the power set of a finite set of variables $\mathcal{V}$. For any three disjoint sets, $A, B, C \subseteq \mathcal{V}$, we write the triple as $\langle A, B \,|\, C \rangle$ and denote the set of all such triples by $\mathcal{T}(\mathcal{V})$. The set of conditional independence statements under P is denoted as:

$$\mathcal{I}_P = \{\langle A, B \,|\, C \rangle \in \mathcal{T}(\mathcal{V}) : A \perp\!\!\!\perp B \mid C \,[P]\},$$

and[1] this is called the *conditional independence model* induced by P.

Definition 2.9. An *imset* is an integer-valued function $u\colon \mathcal{P}(\mathcal{V}) \to \mathbb{Z}$. The *identifier* function δ_A of a set $A \subseteq \mathcal{V}$ is an imset, defined as $\delta_A(B) = 1$ if $B = A$ and otherwise $\delta_A(B) = 0$.

A *semi-elementary imset* $u_{\langle A, B | C \rangle}$ associated with any triple $\langle A, B \,|\, C \rangle \in \mathcal{T}(\mathcal{V})$ is defined as: $u_{\langle A, B | C \rangle} = \delta_{A \cup B \cup C} - \delta_{A \cup C} - \delta_{B \cup C} + \delta_C$. An *elementary* imset corresponds to the case when both A and B are singletons.

An imset u is *combinatorial* if it can be written as a non-negative integer combination of elementary imsets. We call an imset u *structural* if there exists $k \in \mathbb{N}$ such that $k \cdot u$ is combinatorial.

We also define the *degree* of a structural imset as the number of elementary imsets in the sum for the minimum n that makes $n \cdot u$ combinatorial.

Note that in the case of $|\mathcal{V}| \leq 4$, every structural imset is also combinatorial, but for $|\mathcal{V}| \geq 5$ this is not true (Hemmecke et al., 2008). We will also give an example in Section 3.5.

A triple $\langle A, B \,|\, C \rangle$ is represented in a structural imset u over $\mathcal{V}$, written as $A \perp\!\!\!\perp B \mid C \,[u]$ if there exists $k \in \mathbb{N}$ such that $k \cdot u - u_{\langle A, B | C \rangle}$ is a combinatorial imset over $\mathcal{V}$. The *model induced by* u then is defined as:

$$\mathcal{I}_u = \{\langle A, B \mid C \rangle \in \mathcal{T}(\mathcal{V}) : A \perp\!\!\!\perp B \mid C \,[u]\}.$$

We say that an imset u is *Markovian* with respect to an independence model $\mathcal{I}$ if for every $\langle A, B \mid C \rangle \in \mathcal{I}$, we have that $\langle A, B \mid C \rangle \in \mathcal{I}_u$. If the converse holds, we say it is *faithful*, and if it is both Markovian and faithful we say it is *perfectly Markovian* with respect to $\mathcal{I}$.

---

[1]Here for simplicity we use $A \perp\!\!\!\perp B \mid C \,[P]$ to represent $X_A \perp\!\!\!\perp X_B \mid X_C$ under P.
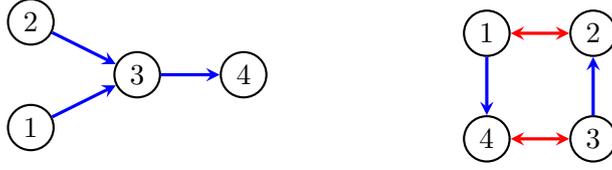
Figure 3: (i) A DAG with 4 nodes; (ii) a MAG $\mathcal{G}$ in which there is no topological ordering such that the tail of a head precedes any vertex in the head.

For a MAG $\mathcal{G}$, we write $A \perp_m B \mid C \, [\mathcal{G}]$ if A and B are m-separated by C in $\mathcal{G}$. The *model induced by* $\mathcal{G}$ is then defined as:

$$\mathcal{I}_\mathcal{G} = \{\langle A, B \mid C \rangle \in \mathcal{T}(\mathcal{V}) : A \perp_m B \mid C \, [\mathcal{G}]\}.$$

Example 2.1. Consider the DAG in Figure 3; a combinatorial imset u such that $\mathcal{I}_u = \mathcal{I}_\mathcal{G}$ is: $u = u_{\langle 1,2 \rangle} + u_{\langle 4,12 \mid 3 \rangle}$, which has entries:

C	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$	$\{3\}$	$\{3,4\}$	$\{1,2,3\}$	$\{1,2,3,4\}$
$u(C)$	$+1$	-1	-1	$+1$	$+1$	-1	-1	$+1$

.

Remark 2. Independence models (whether represented by imsets or not) always obey the semi-graphoid axioms, which are listed in Appendix A.2. In the previous example, the conditional independences $4 \perp\!\!\!\perp 2 \mid 3$ and $4 \perp\!\!\!\perp 1 \mid 2, 3$ can both be deduced from $4 \perp\!\!\!\perp 1, 2 \mid 3$, and indeed $\mathcal{I}_u$ represents these constraints as well.

Appendix A.2 also lists some other rules, but these only apply to probabilistic independence models under some additional assumptions. Note that we restrict to using the semi-graphoids, these additional rules are for discussion with related work in Section 1.2.

3 Main results

In this section we will define the standard imset of MAGs. We begin by reviewing existing results on imsets for DAGs, and then draw a link between them and the parametrizing set of MAGs. This provides the motivation for our definition for the standard imsets of MAGs.

3.1 Previous work

For a DAG $\mathcal{G}$ over $\mathcal{V}$, its standard imset is defined as:

$$u_\mathcal{G} = \delta_\mathcal{V} - \delta_\emptyset - \sum_{i \in \mathcal{V}} \left\{ \delta_{\{i\} \cup \mathrm{pa}_\mathcal{G}(i)} - \delta_{\mathrm{pa}_\mathcal{G}(i)} \right\}. \tag{1}$$

Studený (2005) shows that for a DAG $\mathcal{G}$, $\mathcal{I}_\mathcal{G} = \mathcal{I}_{u_\mathcal{G}}$. That is, $u_\mathcal{G}$ is perfectly Markov with respect to $\mathcal{I}_\mathcal{G}$. We will often just say that $u_\mathcal{G}$ is perfectly Markovian with respect to $\mathcal{G}$, rather than explicitly invoking the list of independences.

Example 3.1. Consider the DAG $\mathcal{G}$ in Figure 3(i), by definition, its standard imset is:

$$u_\mathcal{G} = \delta_{1234} - \delta_\emptyset - (\delta_{34} - \delta_3) - (\delta_{123} - \delta_{12}) - (\delta_2 - \delta_\emptyset) - (\delta_1 - \delta_\emptyset)$$

$$= (\delta_{1234} - \delta_{123} - \delta_{34} + \delta_3) + (\delta_{12} - \delta_1 - \delta_2 + \delta_\emptyset)$$
$$= u_{\langle 4,12 \,|\, 3\rangle} + u_{\langle 1,2\rangle}.$$

In the last line, the conditional independences of the semi-elementary imsets are the independences required for the ordered local Markov property of $\mathcal{G}$, that is, given the numerical (and also topological) ordering

$$i \perp\!\!\!\perp [i-1] \setminus \mathrm{pa}_\mathcal{G}(i) \mid \mathrm{pa}_\mathcal{G}(i),$$

where $[i] = \{1, 2, \ldots, i\}$.

For a DAG $\mathcal{G}$, Studený et al. (2010) introduce the *characteristic imset* $c_\mathcal{G}$, which can be obtained from the standard imset $u_\mathcal{G}$ by a one-to-one linear transformation called the Möbius transform (Lauritzen, 1996):

$$c_\mathcal{G}(S) = 1 - \sum_{T : S \subseteq T \subseteq \mathcal{V}} u_\mathcal{G}(T) \quad \text{and} \quad u_\mathcal{G}(T) = \sum_{S : T \subseteq S \subseteq \mathcal{V}} (-1)^{|V \setminus S|} \{1 - c_\mathcal{G}(S)\}.$$

These transforms are just inclusion and exclusion formulae. This gives another representation of the equivalence classes of DAGs, and (Studený et al., 2010), provide the following theorem.

Theorem 3.1. *For a DAG $\mathcal{G}$, we have:*

(i) $c_\mathcal{G}(S) \in \{0, 1\}$ *for each $S \subseteq \mathcal{V}$, and*

(ii) $c_\mathcal{G}(S) = 1$ *if and only if either $S = \emptyset$ or there exists some $i \in S$ with $S \setminus \{i\} \subseteq \mathrm{pa}_\mathcal{G}(i)$.*

It can be shown that for a DAG $\mathcal{G}$, a nonempty set S is in the *parametrizing set* $\mathcal{S}(\mathcal{G})$ if and only if $c_\mathcal{G}(S) = 1$. We also have the following result about parameterizing sets (Hu and Evans, 2020).

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

Inspired by the relation in DAGs between standard imsets and characteristic imsets, and the consistency between characteristic imsets and parametrizing sets, we extend the definition of the characteristic imset to MAGs using the parametrizing set. The previous Möbius transform used on $u_\mathcal{G}$ to obtain $c_\mathcal{G}$ is linear and thus has an inverse form (we will see later), we then apply this inverse formulae on $1 - c_\mathcal{G}$, to obtain the 'standard' imsets, $u_\mathcal{G}$. We will show that for many MAGs, these 'standard' imsets are combinatorial and perfectly Markovian with respect to the original graph $\mathcal{G}$.

Example 3.2. Consider the graph $\mathcal{G}$ in Figure 3(ii), which is an example from Richardson (2009). There are two absent edges, each of them corresponding to a conditional independence; specifically $1 \perp\!\!\!\perp 3$ and $2 \perp\!\!\!\perp 4 \mid 1, 3$. Clearly we want a standard imset $u_\mathcal{G}$ such that it is the sum of the elementary imsets for these two independences. So we want

$$u_\mathcal{G} = u_{\langle 1,3\rangle} + u_{\langle 2,4|13\rangle}$$
$$= \delta_{1234} - \delta_{134} - \delta_{123} + 2\delta_{13} - \delta_3 - \delta_1 + \delta_\emptyset.$$

The fourth term δ_{13} with coefficient 2 is quite interesting. Graphically these vertices are nonadjacent; however, $\{1\}$ is the tail for the head $\{3, 4\}$ that contains 3, and conversely $\{3\}$ is the tail for the head $\{1, 2\}$ that contains 1.

3.2   Standard imsets of MAGs

The fact that the parametrizing set and the characteristic imset of DAGs agree motivates our definition for the characteristic imset of MAGs, then we work backwards to deduce the form of the standard imset of MAGs.

Definition 3.1. For a MAG $\mathcal{G}$, define its characteristic imset $c_{\mathcal{G}}$ as $c_{\mathcal{G}}(S) = 1$ if $S \in \mathcal{S}(\mathcal{G})$ and $c_{\mathcal{G}}(S) = 0$ otherwise. Moreover, we define $c_{\mathcal{G}}(\emptyset) = 1$ by convention and in order to preserve the characteristic imset for DAG models.

Theorem 3.3. *For a MAG $\mathcal{G}$ with characteristic imset $c_{\mathcal{G}}$, let*

$$u_{\mathcal{G}}(T) = \sum_{S: T \subseteq S \subseteq \mathcal{V}} (-1)^{|S \setminus T|}(1 - c_{\mathcal{G}}(S)).$$

Then

$$u_{\mathcal{G}} = \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \mathrm{tail}(H)}.$$

We will refer to $u_{\mathcal{G}}$ as the 'standard' imset of the MAG $\mathcal{G}$, the quotes being in recognition of the fact that it does not always define the model (see Example 3.4).

Proof. We can obtain $c_{\mathcal{G}}$ via the following transformation from $u_{\mathcal{G}}$:

$$c_{\mathcal{G}}(S) = 1 - \sum_{T: S \subseteq T \subseteq \mathcal{V}} u_{\mathcal{G}}(T).$$

Then we can prove the theorem by showing that, after substituting

$$u_{\mathcal{G}} = \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \mathrm{tail}(H)}$$

to the RHS of the previous equality, the result (say $c_{\mathcal{G}}^{*}$) is the same as $c_{\mathcal{G}}$.

Note that $c_{\mathcal{G}}^{*}(\emptyset) = 1$, as the sum of coefficients in $u_{\mathcal{G}}$ is 0. Suppose $S \neq \emptyset$ and let $\mathbb{1}_{P}$ denote an indicator function, taking value 1 if P is true and 0 otherwise. Then

$$
\begin{aligned}
c_{\mathcal{G}}^{*}(S) &= 1 - \sum_{T: S \subseteq T \subseteq \mathcal{V}} \left\{ \delta_{\mathcal{V}}(T) - \delta_{\emptyset}(T) - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \mathrm{tail}(H)}(T) \right\} \\
&= \sum_{T: S \subseteq T \subseteq \mathcal{V}} \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \mathrm{tail}(H)}(T) \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{T: S \subseteq T \subseteq \mathcal{V}} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \mathrm{tail}(H)}(T) \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{T: S \subseteq T \subseteq \mathcal{V}} (-1)^{|H \setminus T|} \mathbb{1}_{\mathrm{tail}(H) \subseteq T \subseteq H \cup \mathrm{tail}(H)} \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \mathbb{1}_{S \subseteq H \cup \mathrm{tail}(H)} \sum_{K \subseteq H \setminus S} (-1)^{|K| + |H \setminus S|} \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \mathbb{1}_{H \subseteq S \subseteq H \cup \mathrm{tail}(H)}.
\end{aligned}
$$

The fifth equality can be seen in the following way: for each S and H we are counting the number of supersets $T \supseteq S$ such that $\mathrm{tail}_\mathcal{G}(H) \subseteq T \subseteq H \cup \mathrm{tail}_\mathcal{G}(H)$, multiplied by some constant $(-1)^{|H \setminus T|}$. The indicator function comes from the fact that if S is not a subset of $H \cup \mathrm{tail}_\mathcal{G}(H)$ then there is no such T. Now S can be partitioned into two sets $S = S_1 \dot\cup S_2$ which are subsets of H and $\mathrm{tail}_\mathcal{G}(H)$, respectively. For any set $T \supseteq S$ such that $\mathrm{tail}_\mathcal{G}(H) \subseteq T \subseteq H \cup \mathrm{tail}_\mathcal{G}(H)$, It can be partitioned into three sets: $\mathrm{tail}_\mathcal{G}(H) \supseteq S_2$, S_1 and K. The last one K is (any) subset of $H \setminus S$ and the first two sets are deterministic. Moreover, $(-1)^{|H \setminus T|}$ is then equal to $(-1)^{|(H \setminus S) \setminus K|} = (-1)^{|K| + |H \setminus S|}$.

Now the result follows from the proof of Lemma 4.3 in Evans and Richardson (2013) which shows that for each set A there is at most one head H such that $H \subseteq A \subseteq H \cup \mathrm{tail}_\mathcal{G}(H)$. $\qquad\square$

If $\mathcal{G}$ is a DAG, then $u_\mathcal{G}$ in Theorem 3.3 agrees with (1).

Remark 3. Notice that the form of standard imsets in Theorem 3.3 considers a tail with subsets of its head, in an opposite manner to the parametrizing set where we consider a head with subsets of its tail. One can check that this is how $2\delta_{13}$ is obtained in Example 3.2.

Corollary 3.3.1. *For two MAGs $\mathcal{G}$ and $\mathcal{H}$, they are Markov equivalent if and only if $u_\mathcal{G} = u_\mathcal{H}$.*

Proof. This follows from Theorem 3.2 and the fact that the transformation between the standard imset and the characteristic imset is one-to-one. $\qquad\square$

We also have the following result for induced subgraphs, proven in Appendix B.

Proposition 3.4. *Let $\mathcal{G}$ be a MAG, and suppose that for some ancestral subset $A \subset V$ we have that $u_{\mathcal{G}_A}$ is not Markovian with respect to $\mathcal{G}_A$. Then the model $u_\mathcal{G}$ is not perfectly Markovian with respect to $\mathcal{G}$.*

Before we proceed, we would like to clarify some of the terms and notations used in this paper. Originally, the standard imset of a DAG refers to the fact that it is the simplest imset that is perfectly Markovian w.r.t. the graph (Studený, 2005). However for MAGs, the 'standard' imset we defined is not necessarily perfectly Markovian, and we use the quotes to refer to the imset obtained by applying the Möbius inversion formula to the 0-1 imset defined by the parametrizing set.

If we use c with some subscript to denote an imset then it is in the characteristic form, i.e. obtained by applying the Möbius transform to some imset u. If u is structural then it induces a model, and we associate the same model with both c and u.

3.3 Choice of the characteristic imset

Obviously, there are many imset representations for an independence model. The reason we choose to firstly work with a 0-1 characteristic imset is because if its corresponding 'standard' imset *is* perfectly Markovian w.r.t. the graph, then it is also the minimal representation. We proceed to prove this fact by studying the characteristic imset form c of any structural imset u that represent the model.

Example 3.5 gives a graph for which the 'standard' imset is not structural. Fortunately, such graphs seem to be comparatively unusual.

We begin with a simple observation on the linear relationship between structural imset and its characteristic imset.

Lemma 3.5. *For a structural imset u and its characteristic imset c, we have that $c(S) = 1$ if and only if u contains no independence with S as a constrained set, and otherwise $c(S) \leq 0$.*

The following result from Hu and Evans (2020) will be useful.

Proposition 3.6. *Let $\mathcal{G}$ be a MAG with vertex set $\mathcal{V}$. For a set $S \subseteq \mathcal{V}$, we have $S \notin \mathcal{S}(\mathcal{G})$ if and only if there are two vertices a, b in S such that we can m-separate them by a set C such that $a, b \notin C$ with $S \subseteq C \cup \{a, b\}$.*

Corollary 3.6.1. *Consider a MAG $\mathcal{G}$. For any structural imset u such that $\mathcal{I}_u = \mathcal{I}_\mathcal{G}$, its characteristic imset c must be an integer valued vector and satisfy the following:*

(i) if $S \in \mathcal{S}(\mathcal{G})$, then $c_\mathcal{G}(S) = 1$;

(ii) if $S \notin \mathcal{S}(\mathcal{G})$, then $c_\mathcal{G}(S) \leq 0$.

We now argue for the choice of the 0-1 characteristic imset, which Corollary 3.6.1 shows is unique for any given graph. There is no way to totally order combinatorial imsets, though the degree does provide a partial ordering. It is clear that, if our imset is combinatorial and perfectly Markovian, then it is also an *imset of the smallest degree* (Studený, 2005), because there is exactly one independence for each missing edge. We will see in Example 3.4 that some MAG models cannot be represented in this way.

We can define a partial order on imsets defining the same model based on the coefficients of their corresponding characteristic imsets. That is, given two structural imsets u and u' with corresponding characteristic imsets c and c', we say that u *is smaller than* u' if $c(S) \geq c'(S)$ for every $S \subseteq \mathcal{P}(V)$. Lemma 3.5 makes clear that this is a useful definition. By Corollary 3.6.1, if the 0-1 imset's standard imset induces the same model as the graph, it is both an imset of smallest degree and the unique minimal such imset according to this partial order.

3.4 Standard imsets of simple MAGs

Unlike the standard imset of a DAG, it is often very hard to tell how to decompose this 'standard' imset as a sum of semi-elementary imsets, because the size and number of heads seems arbitrary. However, if we restrict the size of heads to two, we can show that for any such MAG $\mathcal{G}$, its standard imset is always both combinatorial and perfectly Markovian with respect to $\mathcal{G}$.

Definition 3.2. A MAG $\mathcal{G}$ is said to be *simple* if $\mathcal{G}$ contains no head with size more than two.

In Appendix B.1, we consider how dense simple MAGs are as a subset of all MAGs by: (i) listing how many Markov equivalence classes contain at least one simple MAG; and (ii) simulating MAGs and counting proportions of them being Markov equivalent to some simple MAGs. Note that simple MAGs can have large districts and not be Markov equivalent to a graph with smaller districts; see Figure 11.
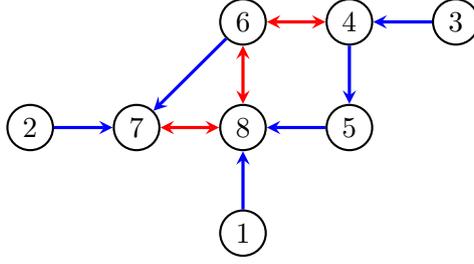
Figure 4: A simple MAG used in Example 3.3

It is not hard to show that for a MAG, $\{i, j\}$ is a head if and only if $i \leftrightarrow j$. Moreover, consider any $j \leftrightarrow i \leftrightarrow k$. If the MAG is simple, then there must be ancestral relations between j and k so that $\{i, j, k\}$ does not form a head (this is necessary and sufficient as shown by Lemma 7.3 of Evans and Richardson, 2013). Hence for each vertex i, there is a total ordering on heads that contain i, and we now show that their tails are nested within one another.

Lemma 3.7. *Suppose $\mathcal{G}$ is a simple MAG with a given topological ordering. For every vertex i and all heads of size two, $\{i, j_s\}$, with $j_1 < \cdots < j_k < i$, we have $\mathrm{pa}_\mathcal{G}(i) = \mathrm{tail}_\mathcal{G}(i) \subseteq \mathrm{tail}_\mathcal{G}(i, j_1) \subseteq \cdots \subseteq \mathrm{tail}_\mathcal{G}(i, j_k)$.*

Example 3.3. Consider the simple MAG $\mathcal{G}$ in Figure 4. The sets $\{7, 8\}$ and $\{6, 8\}$ are the heads of size two associated with the vertex 8. Their tails are $\{1, 2, 3, 4, 5, 6\}$ and $\{1, 3, 4, 5\}$ respectively. Note that $\mathrm{tail}_\mathcal{G}(\{6, 8\}) \subset \mathrm{tail}_\mathcal{G}(\{7, 8\})$. Also, we have $\mathrm{pa}_\mathcal{G}(8) = \{1, 5\}$, which is a subset of both the other tails.

One can check that its standard imset may be written as:

$$u_\mathcal{G} = u_{\langle 8,2|13456\rangle} + u_{\langle 8,34|15\rangle} + u_{\langle 7,1345|26\rangle} + u_{\langle 6,125|34\rangle}$$
$$+ u_{\langle 6,3\rangle} + u_{\langle 5,123|4\rangle} + u_{\langle 4,12|3\rangle} + u_{\langle 3,12\rangle} + u_{\langle 1,2\rangle}.$$

Let us focus on the semi-elementary imsets that contain 8. The conditional independence between vertex 8 and vertex 2 can be seen in the following way: after marginalizing $\{7\}$ (the remaining seven vertices still form an ancestral set), $\{2\}$ would be outside of the Markov blanket of vertex 8 which is $\mathrm{tail}_\mathcal{G}(\{6, 8\}) = \{1, 3, 4, 5, 6\}$, and this independence is implied by the ordered local Markov property. Similarly, if one further marginalizes $\{6\}$, then $\{3, 4\}$ are not in the Markov blanket of vertex 8, which is just $\mathrm{pa}_\mathcal{G}(8) = \{1, 5\}$, and this corresponds to $8 \perp\!\!\!\perp 3, 4 \mid 1, 5$.

Thus for simple MAGs with a given topological ordering, we can obtain conditional independences for each vertex v by sequentially marginalizing the lesser member of each head of size two that contains v, and using the ordered local Markov property. See details in the proof of the following theorem.

Theorem 3.8. *For a MAG $\mathcal{G}$ and its standard imset $u_\mathcal{G}$, if $\mathcal{G}$ is simple then $\mathcal{I}_{u_\mathcal{G}} = \mathcal{I}_\mathcal{G}$.*

Proof. See Appendix B.2. □

Corollary 3.8.1. *For a MAG $\mathcal{G}$, if there exists a simple MAG $\mathcal{H}$ which is Markov equivalent to $\mathcal{G}$, then $\mathcal{I}_{u_\mathcal{G}} = \mathcal{I}_\mathcal{G}$.*
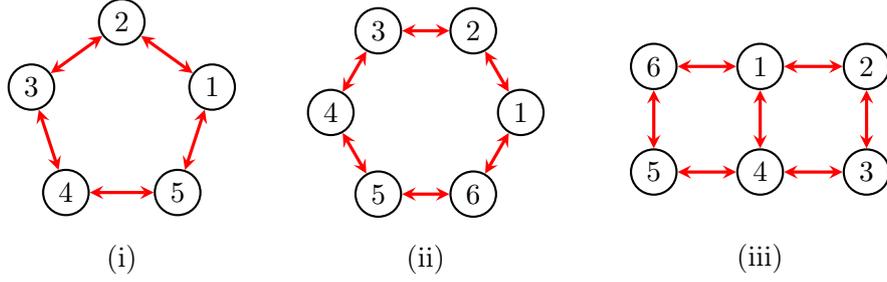
Figure 5: (i) a bidirected 5-cycle; (ii) a bidirected 6-cycle; (iii) a bidirected 6-cycle with an additional edge.

3.5 Examples where the imset is not perfectly Markovian

In general—as the following examples show—the 'standard' imset we have defined is not structural, and even if it is structural, $u_{\mathcal{G}}$ may not be perfectly Markovian with respect to the model induced by the graph.

Example 3.4. Consider the 5-cycle $\mathcal{G}$ with bidirected edges in Figure 5. Its 'standard' imset is given by

$$
\begin{aligned}
u_{\mathcal{G}} &= u_{\langle 1,3|4\rangle} + u_{\langle 2,4|5\rangle} + u_{\langle 3,5|1\rangle} + u_{\langle 4,1|2\rangle} + u_{\langle 5,2|3\rangle} \\
&= u_{\langle 1,3|5\rangle} + u_{\langle 2,4|1\rangle} + u_{\langle 3,5|2\rangle} + u_{\langle 4,1|3\rangle} + u_{\langle 5,2|4\rangle}.
\end{aligned}
$$

Any (strictly) conditional independence that holds in the graph is contained in $\mathcal{I}_{u_{\mathcal{G}}}$; however, marginal independences such as $1 \perp\!\!\!\perp 3$ are not in the imset.

Note, however, that if we assume the underlying distribution is strictly positive then we are able to deduce that any marginal independence in the graph is in $\mathcal{I}_{u_{\mathcal{G}}}$; by noting that (for example) $1 \perp\!\!\!\perp 3 \mid 4$ and $1 \perp\!\!\!\perp 4 \mid 3$ are both in $\mathcal{I}_{u_{\mathcal{G}}}$, and can be combined by using the intersection graphoid provided that the density is positive (See Appendix A), to obtain $1 \perp\!\!\!\perp 3, 4$. It turns out that for MAGs on at most 5 vertices, this graph is the only one such that $u_{\mathcal{G}}$ is not perfectly Markovian with respect to $\mathcal{G}$.

As no marginal independences are represented in $\mathcal{I}_{u_{\mathcal{G}}}$ for the bidirected 5-cycle, this shows that for some MAGs, if we want a structural imset $u_{\mathcal{G}}$ that is perfectly Markovian with respect to the graph, it is inevitable that there is some overlap between the sets associated with the independence decomposition of $u_{\mathcal{G}}$. This also shows that the minimum degree of a model defining the 5-cycle is 6, even though there are only 5 pairs of vertices that are not adjacent; this in turn demonstrates the existence of MAG models that can only be defined by having a list of independences that involves a repeated pair of vertices.

Note also that this model does not have a unique minimal representation using the partial order from Section 3.3. Indeed, there are at least five different such representations, obtained by adding $u_{\langle i,i+2\rangle}$ to our 'standard' imset for some $i \in \{1,\ldots,5\}$ (and using arithmetic modulo 5).

Example 3.5. Consider the bidirected 6-cycle $\mathcal{G}$ shown in Figure 5(ii), and $u_{\mathcal{G}}$, which is too long to list here. Write $u_{\mathcal{G}} = \sum_{A \in \mathcal{P}(\mathcal{V})} \alpha_A \delta_A$, where α_A are coefficients for each identifier imset. The largest disconnected sets are of size 4, and $\sum_{|A|=4} \alpha_A = 9$ and $\sum_{|A|=3} \alpha_A = -22$; It is easy to show that when we add semi-elementary imset to an imset, if the semi-elementary imset contributes to the term with largest set size, we will subtract

14

at most two to the coefficient of the terms with set size exactly one smaller than the largest one. But 22 is larger than $18 = 2 \cdot 9$. Hence this imset cannot be neither combinatorial or structural. Indeed, we have used a linear program to show that it is neither of these things.

Example 3.6. Consider Figure 5 (iii). This graph has a 'standard' imset that is not perfectly Markovian with respect to $\mathcal{G}$, and it is also not combinatorial. However, it *is* structural, because:

$$2u_\mathcal{G} = u_{\langle 1,3 \rangle} + u_{\langle 1,3|56 \rangle} + u_{\langle 1,5 \rangle} + u_{\langle 1,5|23 \rangle} + u_{\langle 2,4 \rangle} + u_{\langle 2,4|56 \rangle} +$$
$$+ u_{\langle 2,5|13 \rangle} + u_{\langle 2,5|46 \rangle} + u_{\langle 2,6 \rangle} + u_{\langle 2,6|35 \rangle} + u_{\langle 3,5 \rangle} + u_{\langle 3,5|26 \rangle} +$$
$$+ u_{\langle 3,6|24 \rangle} + u_{\langle 3,6|15 \rangle} + u_{\langle 4,6 \rangle} + u_{\langle 4,6|23 \rangle}.$$

This constitutes another example of an imset that is structural but not combinatorial, and it arises in a much more natural way that the one given by Hemmecke et al. (2008).

3.6 Relating to scoring criteria

Here we explain how to use imsets to provide a consistent scoring criteria. For DAGs, the maximum likelihood part of the BIC (defined in Appendix E) can be shown to be the empirical entropy of $X_\mathcal{V}$, minus the inner product between the standard imset and the empirical entropy vector (we demonstrate this in Appendix E for discrete variables). This inner product is explained in the following.

For random variables $X_\mathcal{V}$ with density function p, define the entropy $\mathsf{H}(X_\mathcal{V})$ as the expectation of $-\log p(X_\mathcal{V})$, i.e. $\mathbb{E}[-\log p(X_\mathcal{V})]$. For three random variables X_A, X_B and X_C, the inner product between the semi-elementary imset $u_{\langle A,B|C \rangle}$ and the entropy vector, whose entries correspond to the entropy H of every subset of $\mathcal{V}$, is the mutual information between X_A, X_B given X_C; that is, $\mathsf{H}(X_{ABC}) - \mathsf{H}(X_{AC}) - \mathsf{H}(X_{BC}) + \mathsf{H}(X_C)$. This quantity is always non-negative, and is zero if and only if the independence $A \perp\!\!\!\perp B \mid C$ holds under p.

Hence for DAGs, BIC scores can be interpreted as the discrepancy for a list of independences from the ordered local Markov property plus penalty terms for model complexity. It follows that we can do something similar for simple MAGs, since if $u_\mathcal{G}$ is perfectly Markovian w.r.t. the graph, this inner product (suitably penalized) provides a valid score. In fact, we prove in Appendix E that if $\mathcal{I}_{u_\mathcal{G}} = \mathcal{I}_\mathcal{G}$ then, this score is consistent as it approximates the BIC. However, for simple MAGs this score can be obtained much faster as the imsets can be constructed in quadratic time in the number of vertices; in contrast, there is no guarantee on computation time for BIC. Note that consistency of the BIC score only requires that the data are generated from one of the models being scored (and do not coincidentally lie on any other models with at most the same number of parameters). For this score, we also require that the graph that generates the data has a perfectly Markovian standard imset.

Moreover, as we have shown the decomposition (with positive integer coefficients) of simple MAGs, if the CI relations in a distribution can be described by a simple MAG, we always have that this inner product is zero. Empirically we observed that this is true even for general MAGs, in spite of previous examples where our imsets are not perfectly Markovian w.r.t. MAGs or not even structural. This suggests that the standard imsets we defined *can* be expressed as the sum of semi-elementary imsets corresponding to conditional independences advertised by the graph, but some of the coefficients are negative, this will indeed turn out to be the case (Theorem 4.6).

3.7 Motivations to simplify Markov property

Next we discuss different choices of imsets. Obviously there are many imsets that represent the same model induced by graphs, but they are different in terms of computational complexity and statistical performance. Both the pairwise Markov property (Sadeghi and Lauritzen, 2014) and the (reduced) ordered local Markov property (Richardson, 2003) can be used to construct imsets by summing semi-elementary imsets corresponding to list of independences[2].

However, the pairwise Markov property in general have more conditioning variables if graphs have complicated ancestral relations, and thus require to estimate entropy of more variables, which is hard problem. Even if it has the same degree as the standard imsets for simple MAGs, in practice we found it having worse performance in model search algorithms. For imsets corresponding to the (reduced) ordered local Markov property, since it in general contains redundant conditional independences (even for simple MAGs, see the example on next section), it has higher degree compared to the standard imsets, hence are less useful. Therefore for simple MAGs, our standard imsets are faster to compute (polynomial time) and more reliable.

Now for general MAGs, we have seen that the 'standard' imsets obtained from the 0-1 characteristic imsets in general are invalid. To construct valid imsets one can of course use the imsets from the pairwise/reduced ordered local Markov property. However, we aim imsets that have fewer degree and fewer conditioning variables, hence simplifying the Markov property is essential to find the standard (simplest) imsets for general MAGs. In the next section we use a graphical tool called the Power DAGs to achieve this, and we show that our *refined Markov property* is strictly simpler than the ordered local Markov property. The result is optimal for simple MAGs as it gives the list of independences that decomposes the standard imsets, but not optimal in general.

Our refined Markov property results in an imset that is perfectly Markovian to a given MAG and have fewer degree/conditioning variables compared to pairwise/ordered local Markov property. We also show that fixing the maximal head size, the computational time for this imset is polynomial.

4 Power DAGs

We aim to give a simpler representation of the conditional independence model induced by MAGs. To describe these relations, we use *power DAGs* of these graphs, which are DAGs over the set of heads.

Our power DAG is completely analogous to the *intrinsic power DAG* in (Richardson et al., 2023), as there is a one-to-one correspondence between the collections of head and *intrinsic sets* used in that paper; indeed, for most MAGs, the two graphs are isomorphic. The approach we give later to simplifying the power DAG is not suitable for intrinsic power DAGs for the *nested Markov* model studied in that paper, because the order in which vertices are *fixed* (marginalized in our case) is important for deriving nested constraints.

---

[2]For the pairwise Markov property to be equivalent to the global Markov property it requires that the independence model is a 'compositional graphoid' (see Appendix A.2). One can show the inner product with entropy is zero if and only if the distribution obeys the pairwise Markov property.
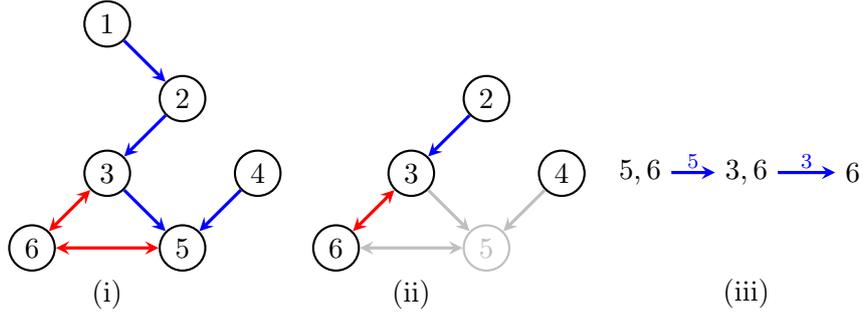
Figure 6: (i) A simple MAG $\mathcal{G}$; (ii) the graph after removing 1 and marginalizing 5; (iii) a DAG on heads in $\mathcal{G}$ that contain the vertex 6.

4.1 Motivations and examples

Example 4.1. Consider the simple MAG in Figure 6(i), given the numerical ordering, the reduced ordered local Markov property for the vertex 6 would require:

$$6 \perp\!\!\!\perp 1 \mid 2,3,4,5 \qquad 6 \perp\!\!\!\perp 1,4 \mid 2,3 \qquad 6 \perp\!\!\!\perp 1,2,4.$$

However, we only actually need:

$$6 \perp\!\!\!\perp 1 \mid 2,3,4,5 \qquad 6 \perp\!\!\!\perp 4 \mid 2,3 \qquad 6 \perp\!\!\!\perp 2.$$

This list is the simplest as it comes from decomposition of standard imset $u_{\mathcal{G}}$ of the simple MAG $\mathcal{G}$.

We start with a topological ordering, in this case the numeric ordering of the vertices. Then, for each vertex we consider the graph of its predecessors; for the vertex 6 this is just the graph itself. We see from this that 1 is not in the Markov blanket of 6, deduce that $6 \perp\!\!\!\perp 1 \mid 2,3,4$ holds, so we remove 1 from the graph. Then we must marginalize something other than 6 in the head $\{5,6\}$; for simple MAGs there is only ever one choice, so we obtain the graph in Figure 6(ii). The maximal head in this graph is now $\{3,6\}$, and we now see that $\{4\}$ is no longer in the Markov blanket of 6; hence we obtain $6 \perp\!\!\!\perp 4 \mid 2,3$ and remove it from further consideration. Finally we marginalize 3 and see that $6 \perp\!\!\!\perp 2$.

Alongside these operations, in Figure 6(iii) we construct the corresponding component of the power DAG for heads that contain 6. We (potentially) associate a single independence with the initial node (so $6 \perp\!\!\!\perp 1 \mid 2,3,4,5$ in our case) and another with each of the transitions in the graph ($6 \perp\!\!\!\perp 4 \mid 2,3$ and $6 \perp\!\!\!\perp 2$).

Specifically, in this example, we reach the head $\{3,6\}$ from the head $\{5,6\}$ by marginalizing $\{5\}$, as illustrated in . Now $\{4\}$ is no longer in the Markov blanket of 6, so we obtain $6 \perp\!\!\!\perp 4 \mid 2,3$ for it and remove it from consideration. Next by marginalizing $\{3\}$, we reach the head $\{6\}$ and now $\{2\}$ is not in the Markov blanket of $\{6\}$ anymore, so we add the independence $6 \perp\!\!\!\perp 2$. For better illustration, we can draw a DAG on heads if one head can reach another by marginalizing vertices in the head, see Figure 6 (iii). Then the second and third independences in our list are represented by this DAG.

Now we give formal definitions of power DAGs and a more complicated example on non-simple MAGs will be given.
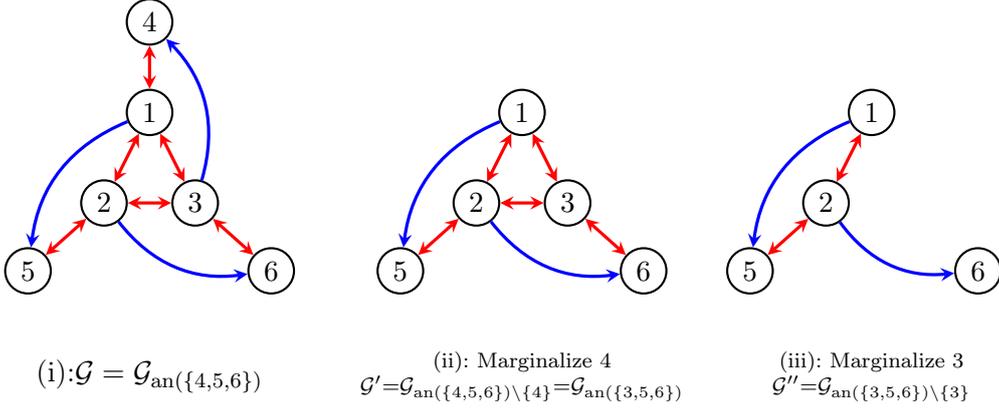
(i):$\mathcal{G} = \mathcal{G}_{\mathrm{an}(\{4,5,6\})}$

(ii): Marginalize 4
$\mathcal{G}'=\mathcal{G}_{\mathrm{an}(\{4,5,6\})\setminus\{4\}}=\mathcal{G}_{\mathrm{an}(\{3,5,6\})}$

(iii): Marginalize 3
$\mathcal{G}''=\mathcal{G}_{\mathrm{an}(\{3,5,6\})\setminus\{3\}}$

Figure 7: An example for Definition 4.1.

4.2 Definition of power DAGs

We assume throughout that the numerical ordering of vertices is also topological. To properly formulate the definition of power DAGs, we need a few more notations and definitions. For a vertex set A and a vertex i, we write $A \leq i$ if i is the *maximal* vertex in A w.r.t. a given topological ordering.

Definition 4.1. For a MAG $\mathcal{G}$ and two heads $H, H' \leq i$, we write $H \to^K H'$ ($\emptyset \subset K \subseteq H \setminus \{i\}$) if $\mathrm{barren}_{\mathcal{G}'}(\mathrm{dis}_{\mathcal{G}'}(i)) = H'$, where $\mathcal{G}' = \mathcal{G}_{\mathrm{an}(H)\setminus K}$. We will refer to K as a marginalization set.

Graphically, $H \to^K H'$ means that in the subgraph $\mathcal{G}_{\mathrm{an}(H)}$, the maximal head (i.e. the barren subset of the district) that contains i after marginalizing K (subset of the barren subset) is H'. Moreover, to save space we eschew set notation and union signs and write (e.g.) k for $\{k\}$ and HT for $H \cup T$.

Take the graph in Figure 7(i) (see Figure 2), under the numerical topological ordering. Consider the final vertex 6; other barren vertices in its Markov blanket are 4 and 5, which together with 6 form a head. By marginalizing 4, we reach the head $\{3,5,6\}$, hence by definition, $\{4,5,6\} \to^4 \{3,5,6\}$. Then if we marginalize 3, we have $\{3,5,6\} \to^3 \{6\}$. The above two marginalization steps are shown in Figure 7 (ii) and (iii).

With each $H \to^K H'$ we associate a conditional independence. By Definition 4.1, if $H \to^K H'$ then $\mathrm{an}_{\mathcal{G}}(H) \setminus K = B$, where B is an ancestral set, $T' = \mathrm{tail}_{\mathcal{G}}(H')$ and $\{i\} \cup \mathrm{mb}_{\mathcal{G}}(i, B) = H' \cup T'$. Hence by the ordered local Markov property and marginalizing vertices that lie outside of the Markov blanket of i in $\mathcal{G}_B$, we have

$$i \perp\!\!\!\perp (H \cup T) \setminus (H' \cup T' \cup K) \mid (H' \cup T') \setminus \{i\}.$$

For instance, in Figure 7, the conditional independence associated with the edge $\{3,5,6\} \to^3 \{6\}$ is obtained as the following: we have $H = \{3,5,6\}, T = \mathrm{tail}(H) = \{1,2\}, H' = \{6\}, T' = \mathrm{tail}(H') = \{2\}$ and $K = \{3\}$, hence

$$6 \perp\!\!\!\perp (\{3,5,6\} \cup \{1,2\}) \setminus (\{6\} \cup \{2\} \cup \{3\}) \mid \{6\} \cup \{2\} \setminus \{6\} = 6 \perp\!\!\!\perp 1, 5 \mid 2.$$

The following definitions will help us to characterize the marginalization sets.

Definition 4.2. For a MAG $\mathcal{G}$ and a set of vertices W, define the *ceiling* of W as

$$\mathrm{ceil}_{\mathcal{G}}(W) = \{w \in W : W \cap \mathrm{an}_{\mathcal{G}}(w) = w\}.$$

Given a head H we define its *Hamlet*[3] as

$$\mathrm{ham}_{\mathcal{G}}(H) = \mathrm{sib}_{\mathcal{G}}(\mathrm{dis}_{\mathrm{an}(H)}(H)) \setminus \mathrm{dis}_{\mathrm{an}(H)}(H).$$

Lemma 4.1. *For a MAG $\mathcal{G}$ with heads $H, H' \leq i$, if $H \to^k H'$, then $k \in \mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H'))$.*

Proof. This is a direct consequence of Lemma C.6. $\qquad\square$

Intuitively, $\mathrm{ham}_{\mathcal{G}}(H)$ serves as the bidirected boundary of H and so must be contained in the marginalization set to reach a graph in which H is the maximal head. Also clearly the last marginalization vertex must be at ceiling of the hamlet, otherwise the barren subset of the district will contain some vertices not in H.

For every i, we can draw a DAG on all the heads $H \leq i$ with edges that are precisely of the form $H \to^k H'$. The list of independences associated with the resulting graph, which we refer to as the *complete* power DAG $\mathfrak{I}^{\mathcal{G}}$ is sufficient to deduce the ordered local Markov property (see Appendix D.1 and Theorem D.4). However, this leads to many redundant independences, and we will show that it is sufficient to only include a single independence for each head. Consequently we will call these simpler power DAGs *refined*. An example of the power DAGs of the graph in Figure 7 is given in Figure 8.

Our definition for a refined power DAG will require a partial ordering on the heads.

Definition 4.3. Define a partial ordering on heads by setting $H < H'$ if and only if H and H' share the same maximal vertex, and $\mathrm{an}_{\mathcal{G}}(H) \subseteq \mathrm{an}_{\mathcal{G}}(H')$.

Lemma 4.2. *For any MAG $\mathcal{G}$ and topological ordering, the relation $<$ is a partial ordering on $\mathcal{H}(\mathcal{G})$.*

Proof. This is a direct consequence of Lemma 4.8 in Evans and Richardson (2014) as the ordering is a sub-order (in that heads are only comparable if they have the same maximal vertex) of the partial orders defined in that paper. $\qquad\square$

Definition 4.4. For a MAG $\mathcal{G}$ and a topological order $<$, the *refined power DAG* $\widetilde{\mathfrak{I}}^{\mathcal{G}}_{<}$ for $\mathcal{G}, <$ consists of a component for each vertex i. Denote this by $\widetilde{\mathfrak{I}}^{\mathcal{G}}_{i}$; it has vertices $\{H : H \leq i\}$, and an edge $H' \to^k H''$ where

$$k = \min \mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H')), \text{ and}$$
$$H'' = \max\{H' : H' \in \mathrm{pa}_{\mathfrak{I}_i(\mathcal{G})}(H'') \text{ and } H' \to^k H''\}.$$

That is, for each head, we only take at most one edge and therefore at most one independence into it.

Taking the maximal set among parent heads of H requires some justification. First we note that it is well defined, since a direct consequence of Proposition C.7 is that if $H_1 \to^k H'$

---

[3]This nomenclature makes sense on understanding that the *Claudius* of H, within a set such that H is barren, is the subset of vertices after strict siblings of H and their descendants are removed. Note that this set that has been removed is precisely the Hamlet of H.
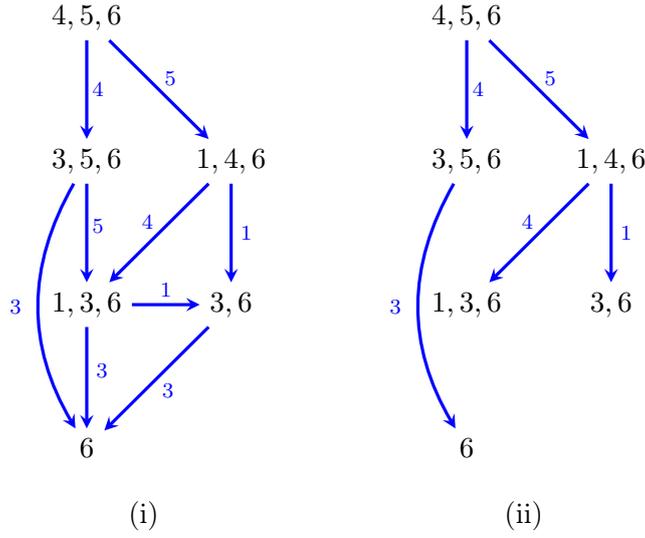
Figure 8: (i) The component of the power DAG for the graph in Figure 7(i); and (ii) the refined version of the same component. Both are on the heads of $\mathcal{G}$ with maximal vertex 6.

and $H_2 \to^k H'$, then $H_3 := \text{barren}_{\mathcal{G}}(H_1 \cup H_2)$ satisfies $H_3 \to^k H'$. Thus there always exists a maximal parent head in any non-empty $\{H : H \in \text{pa}_{\mathfrak{I}_i(\mathcal{G})}(H') \text{ and } H \to^k H'\}$.

For simple MAGs, the complete and refined power DAGs are identical and are chains, in the sense that any node has at most one parent head and at most one child. An important motivation inspired by simple MAGs is that the list of conditional independences associated with every edge in the power DAG of a simple MAG decomposes is $u_{\mathcal{G}}$, and $u_{\mathcal{G}}$ is the simplest imset that defines the same model as the graph.

4.3   The refined Markov property from the refined power DAGs

Next we define the list of independences associated with the refined power DAGs $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$. Recall that $[n]$ denotes the set $\{1, \ldots, n\}$.

Definition 4.5. For a MAG $\mathcal{G}$ and each i, let $\widetilde{\mathbb{L}}_i^{\mathcal{G}}$ be a list of independences, such that:

  (a) it contains $i \perp\!\!\!\perp [i-1] \setminus \text{mb}_{\mathcal{G}}(i, [i]) \mid \text{mb}_{\mathcal{G}}(i, [i])$, and

  (b) for every head H' other than the maximal one, it contains the independence associated with the unique edge into it in $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$

We will refer to the collection $\widetilde{\mathbb{L}}^{\mathcal{G}} = \bigcup_i \widetilde{\mathbb{L}}_i^{\mathcal{G}}$ as the *refined (ordered) Markov property*.

Proposition 4.3. *For a MAG $\mathcal{G}$, the refined Markov property is equivalent to the ordered local Markov property. Further, it contains fewer and smaller independences compared to the reduced ordered local Markov property.*

By 'smaller', we mean any independence in the refined Markov property can be deduced from the reduced ordered local Markov property by simply marginalizing some vertices.

20

Now, since the number of independences is bounded by the number of heads, this will greatly reduce the number of independences arising from the reduced ordered local Markov property.

Remark 4. For simple MAGs, the refined Markov property is the simplest possible description of the model and cannot be further reduced; however, for some graphs even less complicated descriptions exist. For example, for the bidirected 5-cycle, adding the semi-elementary imsets corresponding to these independences would give an imset of degree 7. However, in fact one can build an imset that represents the model of 5-cycle of only degree 6, simply by adding the elementary imset corresponding to any valid marginal independence to our 'standard' imset.

Another example is the 5-chain with bidirected edges (see Figure 16(i)); its standard imset *is* perfectly Markovian for the graph, but the list of independences that defines the imset model is smaller than our refined Markov property (see Example F.1). In Appendix D.3, we give one more interesting example where there are still redundant independences arising in refined power DAGs.

Remark 5. We use the concept of the 'Hamlet' to characterize minimal marginalization sets and maximal parent heads. This concept originates in Richardson (2003), though he did not use this term. He uses the Hamlet to reduce the ordered local Markov property by only visiting maximal ancestral sets. This is similar to our approach in that one visits each head only once (one can easily show there is a one-to-one correspondence between heads and maximal ancestral sets), however, Richardson only reduces the number of ancestral sets visited, while we also marginalize some unnecessary vertices in the independence statements. In Example 3.3, for instance, $A = \{1,2,3,4,5,8\}$ is a maximal ancestral set with $\mathrm{mb}_{\mathcal{G}}(8; A) = \mathrm{pa}_{\mathcal{G}}(8) = \{1,5\}$ for the head $\{8\}$. The reduced ordered local Markov property gives $8 \perp\!\!\!\perp 2,3,4 \mid 1,5$, but we only need $8 \perp\!\!\!\perp 3,4 \mid 1,5$; this is because $8 \perp\!\!\!\perp 2 \mid 1,3,4,5,6$ has already been obtained for the head $\{6,8\}$, and we already know that $6 \perp\!\!\!\perp 2 \mid 1,3,4,5$ from variables earlier in the graph. Reducing the number of variables in a independence statement is crucial in practice when trying to obtain a combinatorial imset by adding the semi-elementary imsets corresponding to the list of independences.

4.4 Computing refined power DAGs

In this section, we show that $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ can be obtained without computing the complete power DAG. We use Algorithm 1 to achieve this and it contains two important ideas:

(i) only marginalizing vertices that are smaller (w.r.t. a topological ordering) than those already marginalized, and

(ii) for each head, only keep one parent of it, which is on the shortest path from the maximal head.

Proposition 4.4. *Given a MAG $\mathcal{G}$, Algorithm 1 computes $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ for each i.*

Proposition 4.5. *Given a MAG $\mathcal{G}$ with n vertices, e edges, and maximal head size k, then the complexity of Algorithm 1 and 2 are $O(kn^k(n+e))$ and $O(n^k(n+e))$, respectively.*

Remark 6. If the input MAG is given to be simple, then the refined power DAG, which is the same as the complete power DAG, can be obtained very fast as shown in Algorithm 3 in Appendix. In fact, the complexity of Algorithm 3 is linear in the number of edges and vertices, as the structure of the power DAG is inherited completely given the topological ordering.

Algorithm 1: Obtain the refined power DAGs for a general MAG

Input: A MAG $\mathcal{G}([n], \mathcal{E})$
Result: The refined power DAGs $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ for each i

1 **define** $M(H)$ *the minimum vertex marginalized to reach* $H \in \mathcal{H}_i(\mathcal{G})$;
2 **define** $SD(H)$ *the shortest path to* $H \in \mathcal{H}_i(\mathcal{G})$;
3 **for** $i \in [n]$ **do**
4   **define** VS_i *the set of visited heads for* $H \in \mathcal{H}_i(\mathcal{G})$;
5   **define** UVS_i *the set of not visited heads for* $H \in \mathcal{H}_i(\mathcal{G})$;
6   Compute the maximal head $H^* = \text{barren}(\text{mb}(i, [i]))$ and set
       $VS_i = \{\{i\}\}, UVS_i = \{H^*\}$;
7   **if** $H^* = \{i\}$ **then**
8   Next
9   **end**
10   Set $M(H^*) = i$, $SD(H^*) = 0, M(\{i\}) = SD(\{i\}) = \infty$;
11   Start with a graph $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ with vertex H^*;
12   **while** UVS_i *is not empty* **do**
13   Take any $H \in UVS_i$ and **move** H to VS_i;
14   **for** $k \in H \setminus \{i\}$ *and* $k < M(H)$ **do**
15   Let $H' : H \to^k H'$;
16   **if** $H' \notin VS_i \cup UVS_i$ **then**
17   **add** H' to $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ and $H \to^k H'$; $UVS_i = UVS_i \cup \{H'\}$;
                 $SD(H') = SD(H) + 1; M(H') = k$;
18   **else**
19   **if** $SD(H') > SD(H) + 1$ **then**
20   **delete** any edges into H';
21   **add** $H \to^k H'$;
22 $SD(H') = SD(H) + 1$;
23 $M(H') = k$;
24   **end**
25   **end**
26   **end**
27   **end**
28 **end**
29 **return** $(\widetilde{\mathfrak{I}}_1^{\mathcal{G}}, \widetilde{\mathfrak{I}}_2^{\mathcal{G}}, \ldots, \widetilde{\mathfrak{I}}_n^{\mathcal{G}})$

Algorithm 2: Obtain the imset $u_{\mathcal{G}}^r$ from refined Markov property

---

Input: The refined power DAG $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ of a MAG $\mathcal{G}$ for each i

Result: The imset $u_{\mathcal{G}}^r$ that is perfectly Markovian w.r.t. $\mathcal{G}$

1 Let $u_{\mathcal{G}}^r$ be an empty imset;

2 for $i \in [n]$ **do**

3   **for** H *in* $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ *along the partial ordering* **do**

4   **Compute** $T = \mathrm{tail}(H)$;

5   **if** $\mathrm{pa}(H) = \emptyset$ **then**

6 $u_{\mathcal{G}}^r := u_{\mathcal{G}}^r + u_{\langle i, |[i]\setminus HT | HT \rangle}$

7   **else**

8   Let H' be the *only* parent of H and $T' = \mathrm{tail}(H')$;

9 $u_{\mathcal{G}}^r := u_{\mathcal{G}}^r + u_{\langle i, H'T'\setminus HT | HT \rangle}$

10   **end**

11   **end**

12 end

13 return $u_{\mathcal{G}}^r$

4.5 Decomposition of the 'standard' imset for general MAGs

In this section, we give a decomposition of $u_{\mathcal{G}}$, using all independences arising from $H \to^K H'$ for every possible pair of heads H, H' and every possible marginalization set K. Note that this theorem does not have practical use but it proves why empirically the inner product between the standard imset and entropy is zero for any MAGs. Moreover, it shows that if the standard imset is structural then $\mathcal{I}_{u_{\mathcal{G}}} \subseteq \mathcal{I}_{\mathcal{G}}$. See discussion after the sketch proof.

Theorem 4.6. *For a MAG $\mathcal{G}$, with vertices $[n]$ (topologically ordered), we have*

$$u_{\mathcal{G}} = \sum_{i=1}^{n} \Bigg\{ u_{\langle i, [i-1]\setminus \mathrm{mb}(i,[i]) | \mathrm{mb}(i,[i]) \rangle}$$
$$+ \sum_{\substack{H \in \mathcal{H}(\mathcal{G})\setminus\{i\}: \\ H \leq i}} \sum_{\substack{\emptyset \subset K \subseteq H\setminus\{i\}: \\ H \to^K H'}} (-1)^{|K|+1} u_{\langle i, HT\setminus H'T'K | H'T'\setminus i \rangle} \Bigg\}.$$

We provide a sketch; the full proof is in Appendix D.4.

Sketch Proof. We can rewrite $u_{\mathcal{G}}$ as the following:

$$u_{\mathcal{G}} = \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H\setminus W|} \delta_{W \cup \mathrm{tail}(H)};$$

$$= \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{K \subseteq H} (-1)^{|K|} \delta_{HT\setminus K};$$

$$= \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{i=1}^{n} \sum_{\substack{H \in \mathcal{H}(\mathcal{G}): \\ H \leq i}} \sum_{K \subseteq H} (-1)^{|K|} \delta_{HT\setminus K};$$

$$= \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{i=1}^{n} \sum_{\substack{H \in \mathcal{H}(\mathcal{G}): \\ H \leq i}} \sum_{K \subseteq H\setminus i} (-1)^{|K|} (\delta_{HT\setminus K} - \delta_{HT\setminus Ki});$$

23

$$= \sum_{i=1}^{n} \delta_{[i]} - \delta_{[i-1]} - \sum_{\substack{H \in \mathcal{H}(\mathcal{G}): \\ H \leq i}} \left\{ \sum_{\emptyset \subset K \subseteq H \setminus i} (-1)^{|K|} (\delta_{HT \setminus K} - \delta_{HT \setminus Ki}) + \delta_{HT} - \delta_{HT \setminus i} \right\}.$$

For any i and $H \leq i$, a marginalization set can be any set K such that $\emptyset \subset K \subseteq H \setminus i$. Suppose $H \to^{K} H'$ where $H, H' \leq i$. We know that $i \perp\!\!\!\perp HT \setminus H'T'K \mid H'T' \setminus i$, and this corresponds to the semi-elementary imset $u_{\langle i, HT \setminus H'T'K \mid H'T' \setminus i \rangle} = \delta_{HT \setminus K} - \delta_{HT \setminus Ki} - \delta_{H'T' \setminus i} + \delta_{H'T'}$.

The last equality already has two terms in this semi-elementary imset. It remains to show that for each i, where $1 \leq i \leq n$ and each head $H' \leq i$,

$$\sum_{\substack{H, K: \\ H \to^{K} H', H \leq i}} (-1)^{|K|} = 1.$$

This corresponds to the extra terms $\delta_{HT} - \delta_{HT \setminus i}$. Informally this is saying that for each head, if we count the number of pairs of heads and marginalization sets K that could reach it, and assign a ± 1 to each such head depending on parity of K's size, the sum of these is 1.

Note that given a vertex i, there is no head that will marginalize to the maximal head whose maximal vertex is i; that is $\text{barren}_{\mathcal{G}}(\text{dis}_{[i]}(i))$. Instead for this maximal head H, $\delta_{HT} - \delta_{HT \setminus i}$ appears in $u_{\langle i, [i-1] \setminus \text{mb}(i, [i]) \mid \text{mb}(i, [i]) \rangle}$ which contains $\delta_{[i]} - \delta_{[i-1]}$. $\qquad \square$

Corollary 4.6.1. *For a MAG $\mathcal{G}$, if $u_{\mathcal{G}}$ is structural, then $\mathcal{I}_{u_{\mathcal{G}}} \subseteq \mathcal{I}_{\mathcal{G}}$.*

Remark 7. Note also that we can separate the independences out by district, by replacing the first sum in Theorem 4.6 by a sum over districts, and then pushing in the summations over vertices in that district.

5 Experimental results

We went through graphs with $|\mathcal{V}| \leq 7$ nodes, checking their standard imset from Theorem 3.3. For $|\mathcal{V}| \leq 4$, all MAGs have combinatorial standard imsets that are perfectly Markovian with respect to the graph. We summarize the information in Table 1 for sizes of graph between 4 and 7. For $|\mathcal{V}| = 5$, the only failure is the bidirected 5-cycle mentioned in Section 3.5. For $|\mathcal{V}| = 6$, all the imsets which are perfectly Markovian with respect to the graph are also combinatorial. For those models where the imset is not perfectly Markovian, only two of them are not combinatorial—one of these imsets *is* structural, and both graphs are shown in Figures 5(ii–iii) in Section 3.

6 Discussion

6.1 Relation to the work in Andrews

The main result of Andrews et al. (2022) is an algorithm that computes an imset that is guaranteed to define the model induced by an ADMG $\mathcal{G}$, but its complexity is exponential in the number of vertices even restricting the maximum head size. In comparison, our imset $u_{\mathcal{G}}^{r}$ in Algorithm 2 can be constructed in polynomial time if the maximum head size

$|V|$	equiv. classes	PM	SNPM	NS
4	19	19	0	0
5	285	284	1	0
6	13,303	13,248	54	1
7*	1,161,461	1,146,501	14,562	8

Table 1: Number of equivalence classes of connected maximal ancestral graphs for various numbers of nodes (for 7 nodes we only include graphs having at most 13 or at least 18 edges.) PM represents models that are Perfectly Markovian, SNPM those where the imset is Structural but represents a strict subset of the independences (so is Not Perfectly Markovian), and NS where the imset is Not Structural.

is bounded. In addition, the imset that Andrews et al. gives has much higher degree than $u_{\mathcal{G}}^r$ for most graphs.

A focus of both works is on consistent scoring and searching for an optimal model, and they are the first to introduce BIC_{MF}, the inner product between empirical entropy and the standard imset from the 0-1 characteristic imset, to approximate the actual BIC. They conduct a brute force search for Gaussian MAG models with five variables by scoring all MECs and selecting the best one. The result is better than FCI (Spirtes et al., 2000) and GFCI (Ogarrio et al., 2016), and comparable to scoring by the true BIC (Drton et al., 2009).

Since, as we have seen, for any MAG with at most five vertices, $u_{\mathcal{G}}$ is perfectly Markovian w.r.t. the graph, this application of BIC_{MF} is valid. (Note that Andrews et al. (2022) finesses this by assuming the model is a graphoid, and that therefore the imset for the bidirected 5-cycle is also perfectly Markovian.) They also express this slightly differently by saying the imset is valid if every set in the parametrizing set has cardinality at most five (see their Theorem 3.1).

For general MAGs, our algorithmic advantage in constructing $u_{\mathcal{G}}^r$ comes from the refined power DAGs, which consider only one parent of each head and marginalize one vertex at each time. The approach of Andrews et al. is to take intersections between all subsets of every Markov blanket which gives the exponential complexity. As well as looking all graphs up to a certain size, we present theoretical proofs that for simple MAGs and a class of bidirected graphs, the standard imset $u_{\mathcal{G}}$ is valid.

Some extra assumptions are made by Andrews in his Theorem 3.1. In particular he makes an assumption that one can use the *intersection axiom* (see Section A.2), which does not generally hold for conditional independence. This means that our 'standard' imsets work for every MAG with size less than or equal to five. We emphasize that we make no such assumption, and use only properties that hold for all distributions, i.e. the semi-graphoid axioms. Stronger rules can change the simplest representation (standard imset) of an independence models. For example, if one assumes positivity of the distribution, then we have already seen that the simplest imset representation for the 5-cycles has degree five, whereas the minimum degree needed without this condition is six. We have verified that using compositional graphoids is sufficient for 51 of the 54 graphs with five or six vertices such that $\mathcal{I}_{u_{\mathcal{G}}} \neq \mathcal{I}_{\mathcal{G}}$, to define the model. However, there are three other graphs that require additional axioms. Specifically, the 'standard' imsets for these graphs will define the model if we can assume *ordered downward stability* (see Section A.2). For further

details, see Example A.1 in Appendix A.2.

6.2 Future work

This paper leaves open several interesting questions. First, if we are to use an algorithm to search for the optimal MAG using the BIC_{MF}-score, what moves should it propose and make? The 0-1 imset formulation makes it easy to search, as—provided that Meek's conjecture (Meek, 1997) for MAGs is true—we only need to search by first adding, and later deleting sets from the parametrizing set.

Another still open problem is to provide a complete solution to obtaining standard imsets of MAGs. This is analogous to undirected graphs, where Kashimura and Takemura (2015) provide a solution for the standard imset of non-decomposable undirected graphs; previously graphs with chordless cycles were not covered by the theory. It seems clear from the example of the bidirected 5-cycle that we will have to choose between having the lowest possible degree and symmetry of the standard imset. We hope that this can expand the class of graphs that can be scored further, and that it may lead to algorithms that are more accurate or more efficient than the current state-of-the-art.

Acknowledgements

References

B. J. Andrews. *Inducing Sets: A New Perspective for Ancestral Graph Markov Models*. PhD thesis, University of Pittburgh, 2021. URL `http://d-scholarship.pitt.edu/42158/`.

B. J. Andrews, G. F. Cooper, T. S. Richardson, and P. Spirtes. The m-connecting imset and factorization for ADMG models. *arXiv preprint:2207.08963*, 2022.

R. Chen, S. Dash, and T. Gao. Integer programming for causal structure learning in the presence of latent variables. In *Proc. 38th Int. Conf. Machine Learning*, pages 1550–1560. PMLR, 2021.

D. M. Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(Nov):507–554, 2002.

T. Claassen and I. G. Bucur. Greedy equivalence search in the presence of latent confounders. In *Proc. 38th Conf. Uncertainty in Artificial Intelligence*. PMLR, 2022.

M. Drton and T. S. Richardson. Binary models for marginal independence. *J. R. Stat. Soc., Ser. B*, 70(2):287–309, 2008.

M. Drton, M. Eichler, and T. S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *J. Mach. Learn. Res.*, 10(10), 2009.

R. J. Evans. Model selection and local geometry. *Ann. Statist.*, 48(6):3513–3544, 2020.

R. J. Evans and T. S. Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proc. 26th Conf. Uncertainty in Artificial Intelligence*. PMLR, 2010.

R. J. Evans and T. S. Richardson. Marginal log-linear parameters for graphical Markov models. *J. R. Stat. Soc., Ser. B*, 75(4):743–768, Sep 2013.

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

R. Hemmecke, J. Morton, A. Shiu, B. Sturmfels, and O. Wienand. Three counter-examples on semi-graphoids. *Comb. Probab. Comput.*, 17(2):239–257, 2008.

Z. Hu and R. Evans. Faster algorithms for Markov equivalence. In *Proc. 36th Conf. Uncertainty in Artificial Intelligence*. PMLR, 2020.

T. Jaakkola, D. Sontag, A. Globerson, and M. Meila. Learning Bayesian network structure using LP relaxations. In *Proc. 13th Int. Conf. Artificial Intelligence and Statistics*, pages 358–365. JMLR Workshop and Conference Proceedings, 2010.

T. Kashimura and A. Takemura. Standard imsets for undirected and chain graphical models. *Bernoulli*, 21(3):1467–1493, 2015.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

C. Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University, 1997.

J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. In *Proc. 8th Int. Conf. Probabilistic Graph. Models*, pages 368–379. PMLR, 2016.

J. Ramsey, P. Spirtes, and J. Zhang. Adjacency-faithfulness and conservative causal inference. In *Proc. 22nd Conf. Uncertainty in Artificial Intelligence*. PMLR, 2006.

K. Rantanen, A. Hyttinen, and M. Järvisalo. Maximal ancestral graph structure learning via exact search. In *Proc. 37th Conf. Uncertainty in Artificial Intelligence*. PMLR, 2021.

T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.*, 30 (1):145–157, 2003.

T. S. Richardson. A factorization criterion for acyclic directed mixed graphs. In *Proc. 25th Conf. Uncertainty in Artificial Intelligence*, 2009.

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

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *Ann. Statist.*, 51(1):334–361, 2023.

K. Sadeghi. Faithfulness of probability distributions and graphs. *J. Mach. Learn. Res.*, 18 (148):1–29, 2017.

K. Sadeghi and S. Lauritzen. Markov properties for mixed graphs. *Bernoulli*, 20(2):676–696, 2014.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

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

M. Studeny. Conditional independence relations have no finite complete characterization. *Inf. Theory Statist. Decis. Funct. Random Process. Trans. 11th Prague Conf.*, 1992.

M. Studený. *On Probabilistic Conditional Independence Structures*. Springer Science & Business Media, 2005.

M. Studený, R. Hemmecke, and S. Lindner. Characteristic imset: a simple algebraic representative of a Bayesian network structure. In *Proc. 5th Eur. WS. Probabilistic Graphical Models*, pages 257–264. HIIT Publications, 2010.

S. Triantafillou and I. Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *Proc. 32rd Conf. Uncertainty in Artificial Intelligence*. PMLR, 2016.

J. Zhang. A characterization of Markov equivalence classes for directed acyclic graphs with latent variables. In *Proc. 23rd Conf. Uncertainty in Artificial Intelligence*, pages 450–457, 2007.

A More definitions

A.1 Separation criterion

For a path π with vertices v_i, $0 \leq i \leq k$ we call v_0 and v_k the *endpoints* of π and any other vertices the *nonendpoints* of π. For a nonendpoint w in π, it is a *collider* if $? \rightarrow w \leftarrow ?$ on π and a *noncollider* otherwise (an edge $? \rightarrow$ is either $\rightarrow$ or $\leftrightarrow$). For two vertices a, b and a disjoint set of vertices C in $\mathcal{G}$ (C might be empty), a path π is *m-connecting* a, b given C if (i) a, b are endpoints of π, (ii) every noncollider is not in C and (iii) every collider is in $\text{an}_{\mathcal{G}}(C)$. A *collider path* is a path where all the nonendpoints are colliders.

Definition A.1. For three disjoint sets A, B and C (A, B are non-empty), A and B are *m-separated* by C in $\mathcal{G}$ if there is no m-connecting path between any $a \in A$ and any $b \in B$ given C. We denote m-separation by $A \perp_m B \mid C$.

Definition A.2. A distribution $P(X_{\mathcal{V}})$ is said to satisfy the *global Markov property* with respect to an ADMG $\mathcal{G}$ if whenever $A \perp_m B \mid C$ in $\mathcal{G}$, we have $X_A \perp\!\!\!\perp X_B \mid X_C$ under P.

A.2 Graphoids

Graphoids are rules for logical implications between conditional independences. There is no finite axiomatization of conditional independences (Studeny, 1992), but there are some rules that hold for any distribution, for example, the *semi-graphoids*:

(1) *symmetry*: $X \perp\!\!\!\perp Y \mid Z \iff Y \perp\!\!\!\perp X \mid Z$;

(2) *decomposition*: $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z$;

(3) *weak union*: $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp W \mid Y, Z$;

(4) *contraction*: $(X \perp\!\!\!\perp Y \mid Z)$ and $(X \perp\!\!\!\perp W \mid Y, Z) \implies X \perp\!\!\!\perp Y, W \mid Z$.

Under the assumption that the distribution is positive, there is an additional property that is guaranteed to hold:

(5) *intersection*: $(X \perp\!\!\!\perp Y \mid W, Z)$ and $(X \perp\!\!\!\perp W \mid Y, Z) \implies X \perp\!\!\!\perp Y, W \mid Z$.

The semi-graphoids with (5) are called *graphoids*. Also if one assumes that the distribution is Markov w.r.t. a graph, then the following axiom also holds:

(6) *composition*: $(X \perp\!\!\!\perp Y \mid Z)$ and $(X \perp\!\!\!\perp W \mid Z) \implies X \perp\!\!\!\perp Y, W \mid Z$.

Graphoids with (6) are called *compositional graphoids*,

Sadeghi (2017) defines three mode graphoid-like rules; two of these require a 'pre-order' $\prec$, which for directed mixed graphs is essentially the partial order implied by the directed edges.

(7) *singleton transitivity*: $(X_i \perp\!\!\!\perp X_j \mid Y)$ and $(X_i \perp\!\!\!\perp X_j \mid Y, X_k) \implies (X_i \perp\!\!\!\perp X_k \mid Y)$ or $(X_j \perp\!\!\!\perp X_k \mid Y)$.
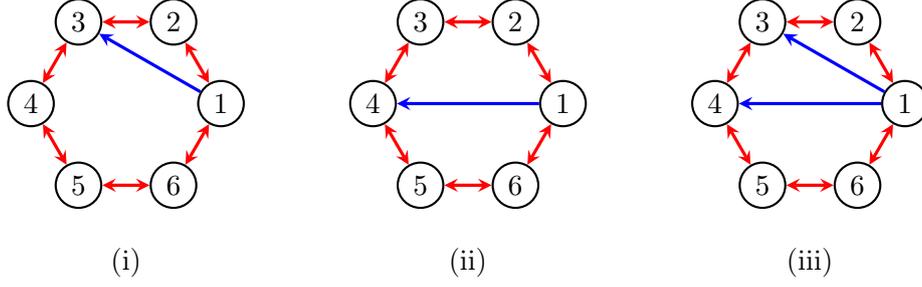
Figure 9: Three graphs that do not define the model without assuming rules beyond composition and intersection.

(8) *ordered upward stability*: $(X_i \perp\!\!\!\perp X_j \mid Y) \implies (X_i \perp\!\!\!\perp X_j \mid Y, X_k)$ for any $k \prec i$ or $k \prec j$.

(9) *ordered downward stability*: $(X_i \perp\!\!\!\perp X_j \mid Y, X_k) \implies (X_i \perp\!\!\!\perp X_j \mid Y)$ for any k that is either larger than or incomparable to i, j and every element of Y.

Here is an example where ordered downward stability is needed to deduce all the independences in the graph.

Example A.1. Consider the bidirected 6-cycle with one additional directed edge in Figure 9(i). The imset $u_{\mathcal{G}}$ of this graph $\mathcal{G}$ is structural and $\mathcal{I}_{u_{\mathcal{G}}}$ contains the following independences:

$$
\begin{array}{llll}
6 \perp\!\!\!\perp 2 & 6 \perp\!\!\!\perp 4 \mid 1,3 & 2 \perp\!\!\!\perp 4 \mid 1,6 & 2 \perp\!\!\!\perp 5 \mid 4,6 \\
1 \perp\!\!\!\perp 4 & 1 \perp\!\!\!\perp 5 \mid 2,4 & 6 \perp\!\!\!\perp 3 \mid 1,5 & 3 \perp\!\!\!\perp 5 \mid 1,2.
\end{array}
$$

In particular, even with the intersection and compositional graphoids, we do not get any joint independences, when the graph says we should have (e.g.) $2 \perp\!\!\!\perp 4, 5, 6$. In the list of the local Markov property, it does not satisfy $4 \perp\!\!\!\perp 1, 2$ or $5 \perp\!\!\!\perp 1, 2$ or $4 \perp\!\!\!\perp 2, 6$. However, ordered downward stability does help. The only pair that is ordered is $1 \to 3$, with all other pairs being incomparable. Then we can deduce $2 \perp\!\!\!\perp 5$ by removing 4 and 6 from $2 \perp\!\!\!\perp 5 \mid 4, 6$, and use composition to obtain $2 \perp\!\!\!\perp 5, 6$; we can then obtain $2 \perp\!\!\!\perp 4, 5, 6$. Similarly we can remove first 3 and then 1 from $6 \perp\!\!\!\perp 4 \mid 1, 3$ to obtain the marginal independence, and get $4 \perp\!\!\!\perp 2, 6$ using composition (or just contraction). Also, we can remove 2, 4 from $5 \perp\!\!\!\perp 1 \mid 2, 4$ to get $5 \perp\!\!\!\perp 1$, then combined with $5 \perp\!\!\!\perp 2$, we have $5 \perp\!\!\!\perp 1, 2$. On the other hand, we can obtain $4 \perp\!\!\!\perp 2$ by removing 6 and 1 from $4 \perp\!\!\!\perp 2 \mid 1, 6$, then use composition to get $4 \perp\!\!\!\perp 1, 2$, as $4 \perp\!\!\!\perp 1$ is already in $\mathcal{I}_{u_{\mathcal{G}}}$.

B   Missing materials from section 3

A useful fact about MAGs is that any induced subgraph of a MAG is itself a MAG.

Lemma B.1. *Let $\mathcal{G}$ be a graph that is maximal and ancestral. Then for any subset of the vertices $W \subseteq V$, so is the induced subgraph $\mathcal{G}_W$.*

Proof. The ancestrality follows from Proposition 3.5 of Richardson and Spirtes (2002). For maximality, note that there are no more paths in an induced subgraph than in the original graph, so in particular there cannot be any more inducing paths. Since all bidirected edges between vertices are preserved, this implies that the graph remains maximal. □

30

Another useful fact will concern conditional independences in induced subgraphs.

Proposition B.2. *Let $\mathcal{G}$ be a MAG and $\mathcal{G}_W$ an induced subgraph. Then for $a, b \in W$, any m-separation $a \perp_m b \mid C$ holding in $\mathcal{G}$ implies that $a \perp_m b \mid E$ holds in $\mathcal{G}_W$, where $E = C \cap W$. Additionally, if W is an ancestral set, then $a \perp_m b \mid E$ in $\mathcal{G}_W$ if and only if it also holds in $\mathcal{G}$.*

Proof. Removing vertices that are not on the path from the conditioning set can only block a path, so the removal of vertices in $C \setminus E$ will not affect the status of any of the paths through W. Hence the result holds. The result for ancestral subgraphs follows, for example, from the results of Richardson (2003). $\qquad\square$

The above result is also useful for considering the parametrizing set. Let $\mathcal{P}(W)$ denote the *power set* of W, i.e. the collection of subsets of W.

Corollary B.2.1. *We have $\mathcal{S}(\mathcal{G}_W) \subseteq \mathcal{S}(\mathcal{G}) \cap \mathcal{P}(W)$.*

Proof. This follows immediately from Propositions 3.6 and B.2. $\qquad\square$

Note that, of course, the sets of size at most two are always identical in the original MAG and any induced subgraph, since these are just the adjacencies.

The next result shows that the certain graphs must not appear as induced subgraphs within a MAG, if we want the corresponding 'standard' imset to be perfectly Markovian with respect to it.

Proposition 3.4 *Let $\mathcal{G}$ be a MAG, and suppose that for some ancestral subset $A \subset V$ we have that $u_{\mathcal{G}_A}$ is not Markovian with respect to $\mathcal{G}_A$. Then the model $u_{\mathcal{G}}$ is not perfectly Markovian with respect to $\mathcal{G}$.*

Proof. Since the subgraph is ancestral, the independences it encodes are just those from the larger graph, with potentially smaller conditioning sets. Now, suppose that it is not possible to write the independences implied by the smaller graph in such a way as to avoid repeating a set. Then the same problem will clearly arise in the larger graph, since (by Proposition B.2) we must specify isomorphic independences with potentially more restrictions on the conditioning set. $\qquad\square$

Remark 8. We know from Theorem 4.6 that if we take an ancestral subset $W = A$, the structure of the imset is preserved. Hence, if we marginalize the graph then the imset will just be the induced subimset over the entries that are subsets of A, and so by Proposition 9.3 of Studený (2005), the imset will match the model for the ancestral subgraph. Hence, in this case the result is clear.

The following example shows that Proposition 3.4 does not work if a subset W is not ancestral. This was pointed out by one of the anonymous reviewers.

Example B.1. One can check that $u_{\langle 4,12|6\rangle} + u_{\langle 6,35|12\rangle} + u_{\langle 1,3|5\rangle} + u_{\langle 2,5|3\rangle} + u_{\langle 3,5\rangle}$ is the standard imset of the MAG in Figure 10 and is perfectly Markovian with respect to it. The subgraph induced by $\{1, 2, 3, 4, 5\}$ is the bidirected 5-cycle, however, and hence the 'standard' imset does not work.
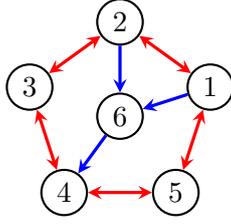
Figure 10: A counter example for Prop 3.4 when W is not ancestral

For the graph, if we were to interpret it using the nested property then, after fixing 6, we would have two additional constraints: $4 \perp\!\!\!\perp 1 \mid 3$ and $4 \perp\!\!\!\perp 2 \mid 5$. This model *cannot* be rewritten in a manner that avoids overlap, and therefore if the imset represented the nested model, this graph would not have a perfectly Markovian 'standard' imset. Of course, the imset *does not* represent the nested model, and so this is merely an academic point.

Lemma 3.5 *For a structural imset u and its characteristic imset c, we have that $c(S) = 1$ if and only if u contains no independence with S as a constrained set, and otherwise $c(S) \leq 0$.*

Proof. As u is structural, there exists an positive integer k and some conditional independences I such that

$$k \cdot u = \sum_I u_I.$$

The transformation from standard imsets to characteristic imsets is bijective and linear. Then

$$
\begin{aligned}
c(S) &= 1 - \sum_{T : S \subseteq T \subseteq \mathcal{V}} u(T) \\
&= 1 - \sum_{T : S \subseteq T \subseteq \mathcal{V}} k^{-1} \sum_I u_{\langle a,b \mid C \rangle}(T) \\
&= 1 - k^{-1} \sum_I \sum_{C' \subseteq C} \delta_{abC'}(T).
\end{aligned}
$$

Hence we see that for any elementary imset that is represented in u the corresponding constrained sets are all strictly less than one, and hence at most zero; conversely any set not so represented will have value 1. Notice that for an empty standard imset, its characteristic imset is just a vector of 1s. $\qquad\square$

Corollary 3.6.1 *Consider a MAG $\mathcal{G}$. For any structural imset u such that $\mathcal{I}_u = \mathcal{I}_\mathcal{G}$, its characteristic imset c must be an integer valued vector and satisfy the following:*

(i) if $S \in \mathcal{S}(\mathcal{G})$, then $c_\mathcal{G}(S) = 1$;

(ii) if $S \notin \mathcal{S}(\mathcal{G})$, then $c_\mathcal{G}(S) \leq 0$.

Proof. This is a direct consequence of Lemma 3.5 and Proposition 3.6. If $S \notin \mathcal{S}(\mathcal{G})$ then there exists $k \in \mathbb{N}$ such that $k \cdot u - u_{\langle ab|C \rangle}$ is combinatorial and $\{a, b\} \subseteq S \subseteq \{a, b\} \cup C$, hence by the same argument used in Lemma 3.5 we have $c(S) \leq 0$.

If $S \in \mathcal{S}(\mathcal{G})$ then there is no independence represented by $\mathcal{G}$ that has S as a constrained set. By Lemma 3.5, $c(S) = 1$. $\qquad\square$
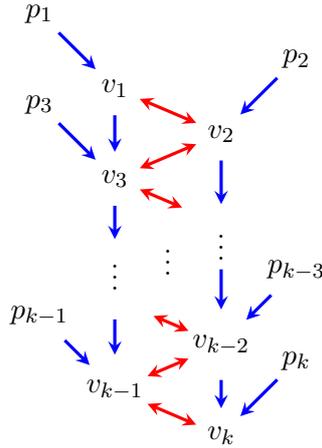
Figure 11: Simple MAGs with arbitrarily large districts

B.1 Simple MAGs

Example B.2. We first start by providing an example of a simple MAG with an arbitrarily large district. This illustrates that a search algorithm over simple MAGs is potentially very useful, since it includes many more causal models than one would obtain by restricting the maximum district size to two or three. This is in contrast with other methods for score-based learning, which generally make this kind of restriction (e.g. Chen et al., 2021).

The graph in Figure 11 has a district of size k and that district also has k parents. However, note that the only heads are just the bidirected edges, so this is a simple MAG. In addition, the independences for each of the V_j are of the form

$$V_j \perp\!\!\!\perp P_1, V_1, \ldots, V_{j-3}, P_{j-2}, P_{k-1} \mid P_j, V_{j-2},$$

and (unsurprisingly) there is no way to order the variables so that the independences are nested within one another as would be required by a DAG.

We demonstrate how common simple MAGs are by using the following results. The first column in Table 2 is the number vertices. Then the second column list the number of equivalence classes of MAGs with the corresponding number of vertices. The next two columns further count how many equivalence classes that contain at least one simple MAG and one DAG, respectively. In particular, the proportion of equivalence classes that contain simple MAGs decreases but not as sharply as that for DAGs.

Then Figure 12, we simulate 1000 random MAGs for number of vertices ranging from five to forty, and plot empirical probabilities that the simulated graphs are Markov equivalent to some simple MAGs. The method we use to simulate MAGs is the same as Claassen and Bucur (2022). We fix the average and maximal degree of each vertex to three and ten respectively. For each edge, the probability of being bidirected is 0.2. We simulate an ADMG first and project it into a Markov equivalent MAG. Then by converting the MAG to a *partially ancestral graph* (PAG), we finally select a representative MAG from the PAG. The last two steps are from Zhang (2007). If the representative MAG is simple, we consider the original graph as being Markov equivalent to some simple MAGs.
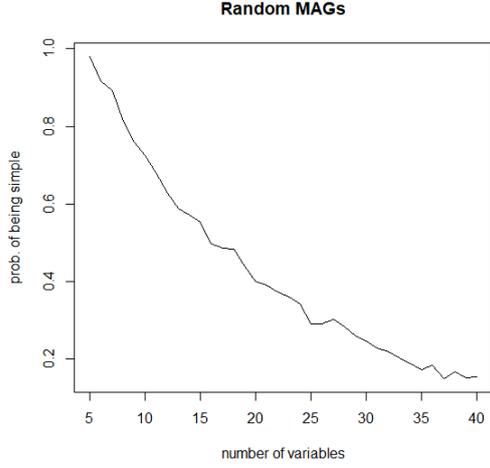
Random MAGs

Figure 12: A plot for probability of random graphs being Markov equivalent to some simple MAGs

$|V|$	equiv. classes	simple MAGs	DAGs
5	285	205	119
6	13,303	6,278	2,025
7*	1,161,461	331,310	57,661

*having at most 13 or at least 18 edges.

Table 2: Number of equivalence classes

B.2 Proof of Theorem 3.8

Proof of Theorem 3.8. Given a topological ordering, we can write the standard imset $u_{\mathcal{G}}$ of a simple MAG $\mathcal{G}$ in the following way:

$$
\begin{aligned}
u_{\mathcal{G}} = {}& \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{i \in \mathcal{V}} \left\{ \delta_{\{i\} \cup \mathrm{pa}(i)} - \delta_{\mathrm{pa}(i)} \right\} \\
& - \sum_{i \leftrightarrow j} \left\{ \delta_{\{i,j\} \cup \mathrm{tail}(i,j)} - \delta_{\{i\} \cup \mathrm{tail}(i,j)} - \delta_{\{j\} \cup \mathrm{tail}(i,j)} + \delta_{\mathrm{tail}(i,j)} \right\} \\
= {}& \sum_{i \in \mathcal{V}} \left\{ \delta_{[i]} - \delta_{[i-1]} - \delta_{\{i\} \cup \mathrm{pa}(i)} + \delta_{\mathrm{pa}(i)} \right\} \\
& - \sum_{i \leftrightarrow j} \left\{ \delta_{\{i,j\} \cup \mathrm{tail}(i,j)} - \delta_{\{i\} \cup \mathrm{tail}(i,j)} - \delta_{\{j\} \cup \mathrm{tail}(i,j)} + \delta_{\mathrm{tail}(i,j)} \right\} \\
= {}& \sum_{i \in \mathcal{V}} \left\{ \delta_{[i]} - \delta_{[i-1]} - \delta_{\{i\} \cup \mathrm{pa}(i)} + \delta_{\mathrm{pa}(i)} \right. \\
& \left. + \sum_{i \leftrightarrow j, i > j} -\delta_{\{i,j\} \cup \mathrm{tail}(i,j)} + \delta_{\{i\} \cup \mathrm{tail}(i,j)} + \delta_{\{j\} \cup \mathrm{tail}(i,j)} - \delta_{\mathrm{tail}(i,j)} \right\}
\end{aligned}
$$

For each vertex i, consider the topological ordering on all the j such that $j < i$ and $i \leftrightarrow j$, where $j_1 < \ldots < j_k < i$. Hence we have:

$$
\begin{aligned}
u_{\mathcal{G}} = {}& \sum_{i \in \mathcal{V}} \left\{ \delta_{[i]} - \delta_{[i-1]} - \delta_{\{i\} \cup \mathrm{pa}(i)} + \delta_{\mathrm{pa}(i)} \right. \\
& \left. + \sum_{l=1}^{k} -\delta_{\{i,j_l\} \cup \mathrm{tail}(i,j_l)} + \delta_{\{i\} \cup \mathrm{tail}(i,j_l)} + \delta_{\{j_l\} \cup \mathrm{tail}(i,j_l)} - \delta_{\mathrm{tail}(i,j_l)} \right\}
\end{aligned}
$$

34

$$= \sum_{i \in \mathcal{V}} \left\{ \delta_{[i]} - \delta_{[i-1]} - \delta_{\{i,j_k\} \cup \text{tail}(i,j_k)} + \delta_{\{j_k\} \cup \text{tail}(i,j_k)} \right.$$
$$+ \sum_{l=1}^{k-1} -\delta_{\{i,j_l\} \cup \text{tail}(i,j_l)} + \delta_{\{i\} \cup \text{tail}(i,j_{l+1})} + \delta_{\{j_l\} \cup \text{tail}(i,j_l)} - \delta_{\text{tail}(i,j_{l+1})}$$
$$\left. + \delta_{\{i\} \cup \text{tail}(i,j_1)} - \delta_{\text{tail}(i,j_1)} - \delta_{\{i\} \cup \text{pa}(i)} + \delta_{\text{pa}(i)} \right\}.$$

Now for each vertex i, consider the following list of conditional independence, denoted by $\mathbb{L}_i$:

$$i \perp\!\!\!\perp [i-1] \setminus (\text{tail}(i,j_k) \cup \{j_k\}) \mid \text{tail}(i,j_k) \cup \{j_k\}$$
$$i \perp\!\!\!\perp \text{tail}(i,j_k) \setminus (\text{tail}(i,j_{k-1}) \cup \{j_{k-1}\}) \mid \text{tail}(i,j_{k-1}) \cup \{j_{k-1}\}$$
$$\vdots$$
$$i \perp\!\!\!\perp \text{tail}(i,j_2) \setminus (\text{tail}(i,j_1) \cup \{j_1\}) \mid \text{tail}(i,j_1) \cup \{j_1\}$$
$$i \perp\!\!\!\perp \text{tail}(i,j_1) \setminus \text{pa}(i) \mid \text{pa}(i).$$

It is straightforward to check that $u_{\mathcal{G}}$ is a sum of the semi-elementary imsets corresponding to the conditional independence list $\mathbb{L}^{\mathcal{G}} = \bigcup_{i \in \mathcal{V}} \mathbb{L}_i$. Notice that if there is no such j for i ($k = 0$) then it reduces to the ordered local Markov property of DAGs:

$$i \perp\!\!\!\perp [i-1] \setminus \text{pa}(i) \mid \text{pa}(i).$$

Thus $u_{\mathcal{G}}$ is a combinatorial imset and every independence in $\mathbb{L}$ is in $\mathcal{I}_{u_{\mathcal{G}}}$.

By Theorem D.4, any ordered local Markov property can be deduced from $\mathbb{L}^{\mathcal{G}} = \bigcup_i \mathbb{L}_i$ using semi-graphoid thus $\mathcal{I}_{\mathcal{G}} \subseteq \mathcal{I}_{u_{\mathcal{G}}}$. Now the faithfulness result in Richardson and Spirtes (2002) shows that for every MAG $\mathcal{G}$ there exists a distribution P such that $\mathcal{I}_P = \mathcal{I}_{\mathcal{G}}$ and Theorem 5.2 in Studený (2005) implies the existence of a structural imset u such that $I_u = I_P$. Now every independence in $\mathbb{L}^{\mathcal{G}}$ is in $\mathcal{I}_u$ by Theorem D.4, and $u_{\mathcal{G}}$ is sum of the semi-elementary imsets corresponding to the independence in $\mathbb{L}^{\mathcal{G}}$, so by Lemma 6.1 in Studený (2005), we have $I_{u_{\mathcal{G}}} \subseteq \mathcal{I}_u = \mathcal{I}_{\mathcal{G}}$. Note the idea of this proof is similar to the proof of Lemma 7.1 in Studený (2005). $\square$

C   Some useful results on heads

C.1   Characterization of the marginalization sets

We first study some properties of the marginalization sets K.

Lemma C.1. *For a MAG $\mathcal{G}$ and two heads H, H' whose maximal vertex is i, if $H \to^K H'$ and $H \to^L H'$ then $H \to^{K \cap L} H'$.*

Proof. Let $T = K \cap L$ and $\dot{K} = K \setminus T$. Suppose $H \to^T H''$. Now after marginalizing T, the vertices in $\dot{K}$ are either not in the same district as i or lie in H''. If all of them are not in the same district with i, then $H' = H''$ because then marginalizing $\dot{K}$ would not change the barren subset of the district containing i, which are H'' and H' before and after the marginalization, respectively.

Let $\tilde{K}$ be those lie in H_3 and similarly we define $\tilde{T}$ for L, note that they are disjoint by definition, because T is the intersection of K and L, and $\dot{K}, \dot{K}_2$ are the complement of T in K and L respectively. Also $H \to^{T \cup \tilde{K}} H'$ and $H \to^{T \cup \tilde{T}} H'$.

We firstly consider $H \to^{T \cup \tilde{K}} H'$. This implies that for any bidirected path from any vertex $i \in \tilde{T}$ to any $w \in H' \cup T'$, there is a vertex from $\tilde{K}$ on the path, otherwise some vertices of $\tilde{T}$ would be preserved, which is a contradiction. The equivalent statement by swapping $\tilde{K}$ and $\tilde{T}$ also holds. If one considers the first vertex in either $\tilde{K}$ or $\tilde{T}$ on any such path, one of the two statements would be false. Hence the lemma is true. $\qquad\square$

An implication of this result is that for any $H, H' \leq i$, if H is a parent head of H' on the power DAG for i, there exists a unique *minimal marginalization set of vertices* that leads from H to H'.

Definition C.1. For a MAG $\mathcal{G}$ and two heads $H, H' \leq i$, let K^m be the minimal marginalization set of vertices such that $H \to^{K^m} H'$.

Lemma C.2. *For a MAG $\mathcal{G}$ and two heads $H, H' \leq i$, and $H \to^{K^m} H'$, then $H \to^K H'$ if and only if $K = K^m \dot{\cup} B$ where $B \subseteq (H \setminus (H' \cup K^m))$.*

Proof. ($\Rightarrow$): By definition of K^m, $K^m \subseteq K$, and $H \setminus (H' \cup K^m)$ are simply all the vertices we can marginalize if we want to reach the head H'.

($\Leftarrow$): This is because after marginalization of some vertices in the barren subset, the remaining vertices are either outside of the Markov blanket of i or it stays in the barren subset of the district, i.e. the head. Then by the definition of K^m, we may marginalize any other irrelevant vertices in the barren subset after marginalization of K^m as they would be outside of the Markov blanket. $\qquad\square$

We only require one more lemma to validate our decomposition of u_G. One key step is to find situations when there is only one marginalization set, which is also the minimal marginalization set. We propose the following definitions.

Lemma C.3. *Consider $H' \to^K H$ where $K = H' \setminus H$ is the minimal marginalization set of vertices, i.e. K is the only marginalization set to reach H from H'. Then this happens if and only if $H' = \text{barren}(K \cup H)$ where $K \subseteq \text{ceil}(\text{ham}(H))$.*

Proof. ($\Rightarrow$): After marginalizing vertices in the barren subset, the remaining vertices of H' either stay in the barren subset or outside of the Markov blanket. Since K is the only marginalization set of vertices, we know the remaining vertices stay in the barren subset (also in H). Hence $H' = \text{barren}(K \cup H)$. We first show that $H' \setminus H \subseteq \text{ham}(H)$. Suppose, for a contradiction, that it is not true; let $K^* = K \setminus (\text{sib}(\text{dis}_{\text{an}(H)}(i)) \setminus \text{dis}_{\text{an}(H)}(i))$. Consider any bidirected path between any $i \in K^*$ and any $k \in \text{dis}_{\text{an}(H)}(i)$, the first vertex x from k that is not in $\text{dis}_{\text{an}(H)}(i)$ lies in $\text{an}(K) \setminus \text{an}(H)$.

If x is in $\text{an}(K) \setminus (\text{an}(H) \cup K)$ we have after marginalizing K then $\text{barren}(\text{dis}_{H \setminus K}(i)) \neq H$, because it would include a descendant of x (possibly x itself) that is *not* an ancestor of H. This is a contradiction to our assumption.

Hence x must lie in K, so by assumption $x \in K \setminus K^*$. Since the choice of the path and vertices are free, it means that if we marginalize $K \setminus K^*$ then K^* would be outside the district of i, so K is not minimal. Hence we reach another contradiction.

Then $K \subseteq \text{ceil}(\text{ham}(H))$ follows, as if it has some ancestors that are also siblings of $\text{dis}_{\text{an}(H)}(i)$ then after marginalizing K, the barren subset of the district would not be H.

($\Leftarrow$): Let K be any subset of $\text{ceil}(\text{ham}(H))$ (but not $\emptyset$) and consider $H' = \text{barren}(K \cup H)$. Clearly H' is a head. Moreover $K \subseteq H'$ as K either has no ancestral relation with H or K are descendent of H; in particular this means that K is a valid marginalization set for H'. We still need to show that (1) after marginalizing K, we reach the head H and (2) K is minimal.

For (1), suppose there is a vertex $t \notin K$ stays in the barren subset of the district but t is not in H. Consider any bidirected path from t to $\text{dis}_{\text{an}(H)}(i)$ in $\mathcal{G}_{\text{an}(H')}$. By definition, on this path there is a vertex in K (not just $\text{an}(K)$) next to some vertex in $\text{dis}_{\text{an}(H)}(i)$, then this path is removed after marginalizing K. Also note that all vertices in $\text{dis}_{\text{an}(H)}(i)$ stay in the graph.

For (2), If we marginalize some subset of K, then the remaining vertices of K stay in the Markov blanket and hence in the barren subset of the district, which then would not be H. $\qquad\square$

C.2 Existence of maximal parents in the power DAG

Lemma C.4. *Suppose that for two heads $i \geq H, H'$, we have $H \to^K H'$. Then $H \to^L H'$ for $L = H \setminus H'$.*

Proof. Clearly, any marginalization set from H to H' must not contain any element from H'. By existence of the minimal marginalization set, we can just adding all irrelevant vertices to the set. $\qquad\square$

Proposition C.5. *For a MAG $\mathcal{G}$, suppose that for three heads $i \geq H_1, H_2, H$, we have $H_1 \to^K H$ and $H_2 \to^L H$. Then $H_3 = \text{barren}(H_1 \cup H_2)$ is a head and $H_3 \to^{K'} H$ for $K' = H_3 \setminus H$. This means that in the power DAG for i, if a head has a parent head, then there exists a maximal parent head.*

Proof. First of all, we show that H_3 is a head. Since i is the maximal vertex, $i \in H_3$, then for any $j \in H_3$, it is either in H_1 or H_2, which means that j lies in the same district as i in either $\mathcal{G}_{\text{an}(H_1)}$ or $\mathcal{G}_{\text{an}(H_2)}$. This graph is a subgraph of $\mathcal{G}_{\text{an}(H_3)}$, hence j lies in the same district as i in $\mathcal{G}_{\text{an}(H_3)}$.

Now let $K_3 = H_3 \setminus H$. It does not contain i and also it is not empty since $H_3 \neq H$, thus it is a valid marginalization set for H_3. Suppose also that $H_3 \to^{K_3} H'$. We know that $\text{an}(H_3) \supseteq \text{an}(H_1) \supseteq \text{an}(H)$, so $\text{an}(H_3) \setminus K_3 = B = \text{an}(H') \supseteq \text{an}(H)$. Now suppose $H' \neq H$, this means that $\text{dis}_{\text{an}(H')}(i)$ contains some vertices that are not in $\text{dis}_{\text{an}(H)}(i)$. Among those vertices, there are the strict siblings of $\text{dis}_{\text{an}(H)}(i)$; there must exist such vertices because H' is bidirected-connected. Now select one of these siblings, say, j. WLOG, j belongs to $\text{an}(H_1)$. Then to go from H_1 to H, j must be marginalized, thus $j \in H_1$. If j is not in H_3, then this means that there are some descendants of j that are in H_2, but then we cannot go from H_2 to H since j stays in the districts, there is some extra vertex in the barren subset of the district of i, other than H. Hence $j \in H_3$. But then as j is in $\text{an}(H')$, j does not lie in K_3, so $j \in H$, which is a contradiction to the definition of j. $\qquad\square$

Lemma C.6. *For a MAG $\mathcal{G}$, if head H is a parent head of H' in the power DAG, then the minimal marginalization set K is $H \cap \text{ceil}_{\mathcal{G}}(\text{ham}_{\mathcal{G}}(H'))$.*

Proposition C.7. *Consider the four heads H, H_k, $k = 1, 2, 3$ with the same setting in Proposition C.5, then the minimal marginalization set for H_3 (to H) is the union of the minimal marginalization sets of H_1, H_2 (to H).*

Proof. Lemma C.6 shows that the minimal marginalization set is the intersection between the parent head and $\mathrm{ceil}(\mathrm{ham}(H))$, thus since $\mathrm{an}_{\mathcal{G}}(H_3)$ is just an union of $\mathrm{an}_{\mathcal{G}}(H_1)$ and $\mathrm{an}_{\mathcal{G}}(H_2)$, it will not introduce extra vertices in $\mathrm{ceil}(\mathrm{ham}(H))$. $\qquad\square$

In Lemma C.6, by construction, $H_3 \geq H_1, H_2$. One can show that if $H \geq H'$ then $H \cup T \supseteq H' \cup T'$, therefore by Lemma C.6, the maximal independence from the H_3 always implies the maximal independences from H_1 and H_2.

Proposition C.8. *Suppose in a power DAG for i, H is a parent head of H', then the independence associated with the minimal marginalization set K^m,*

$$i \perp\!\!\!\perp H \cup T \setminus H' \cup T' \cup K^m \mid H' \cup T' \setminus \{i\},$$

implies all the independences associated with any $H \to^K H'$ and any other marginalization set K.

Proof. The conditioning set is fixed, and the more elements K has, the fewer i is independent from. $\qquad\square$

We call this the *maximal independence* associated with $H \to^K H'$

D   Proofs and supplementary materials for section 4

D.1   Complete power DAGs

Definition D.1. Consider a MAG $\mathcal{G}$ with a topological ordering. Given a set $S \subseteq \mathcal{V}$ we say that $s \in S$ is a *marginalization vertex* if it is in $\mathrm{barren}_{\mathcal{G}}(S)$ and is not maximal in S.

Define the *complete power DAG* $\mathfrak{I}(\mathcal{G})$ as a graph with vertices $\mathcal{H}(\mathcal{G})$. An edge is added from $H \to H'$ if there is a marginalization vertex $k \in H$ such that $H \to^k H'$. In this case we call H a *parent head* of H'. There is a unique component for each vertex i, which we denote $\mathfrak{I}_i(\mathcal{G})$.

Next we present a few results that justify the nomenclature of power DAGs.

Lemma D.1. *For a MAG $\mathcal{G}$ and any i, there is at most one edge between any two heads in $\mathfrak{I}_i(\mathcal{G})$.*

Proof. This is a direct consequence of Lemma C.1, which proves that for a MAG $\mathcal{G}$ and two heads H, H' whose maximal vertex is i, if $H \to^K H'$ and $H \to^L H'$ then $H \to^{K \cap L} H'$. $\quad\square$

Lemma D.2. *For two heads $H, H' \leq i$, we have $H > H'$ if and only if H is an ancestor of H' in $\mathfrak{I}_i(\mathcal{G})$*

Proof. This can be proved by marginalizing vertices in $\mathrm{an}(H) \setminus \mathrm{an}(H)$ step by step. $\quad\square$

A useful fact is that for any i and $\mathfrak{I}_i^{\mathcal{G}}$, there exists a (maximal) head H such that $H \geq H'$ for any $H' \leq i$ and this head is the barren subset of the district of i in $\mathcal{G}_{[i]}$.

Lemma D.3. *For a MAG $\mathcal{G}$ and any i, $\mathfrak{I}_i(\mathcal{G})$ is a DAG*

Proof. This is a direct consequence of Lemma D.1 and D.2. $\qquad\qquad\square$

Here we prove that the list of independences associated with edges in $\mathfrak{I}_i^{\mathcal{G}}$ for every i, combined with the independences $i \perp\!\!\!\perp [i-1] \setminus \mathrm{mb}_{\mathcal{G}}(i, [i]) \mid \mathrm{mb}_{\mathcal{G}}(i, [i])$, are sufficient to deduce the ordered local Markov property.

Definition D.2. For a MAG $\mathcal{G}$ and any i, we associate $\mathfrak{I}_i^{\mathcal{G}}$ with a collection of independences $\mathbb{L}_i^{\mathcal{G}}$ that contains:

(a) $i \perp\!\!\!\perp [i-1] \setminus \mathrm{mb}_{\mathcal{G}}(i, [i]) \mid \mathrm{mb}_{\mathcal{G}}(i, [i])$, and

(b) for every head H (except $\{i\}$) whose maximal element is i:

$$i \perp\!\!\!\perp (H \cup T) \setminus (H' \cup T' \cup k) \mid H' \cup T' \setminus \{i\} \qquad \text{for } k \in H \setminus \{i\},$$

where $H \to^k H'$, and $T = \mathrm{tail}_{\mathcal{G}}(H)$ and $T' = \mathrm{tail}_{\mathcal{G}}(H')$.

Theorem D.4. *For a MAG $\mathcal{G}$, the collection $\mathbb{L}^{\mathcal{G}} = \bigcup_i \mathbb{L}_i^{\mathcal{G}}$ is equivalent to the list of independences implied by the ordered local Markov property for $\mathcal{G}$.*

Proof. ($\Longleftarrow$): notice that $H' \cup T' \setminus \{i\}$ is the Markov blanket of i in the ancestral set $\mathrm{an}(H \cup T \setminus K)$, thus it follows from the ordered local Markov property by marginalizing irrelevant vertices.

($\Longrightarrow$): let $\mathcal{A}^x$ denote the set of all ancestral sets whose maximal elements are x. Further let $\mathcal{A}_r^x = \{A \in \mathcal{A}^x : |A| = r\}$ (note $1 \leq r \leq x$). We will proceed by induction on x from $x = 1$ to $x = n$. To show for every $A \in \mathcal{A}^x$, the corresponding independence implied by the ordered local Markov property is implied by $\mathbb{L}_i$, we will apply a further induction on $\mathcal{A}_r^x$ from $r = x$ to $r = 1$.

For $A \in \mathcal{A}^x$, $A \subseteq [x]$. The base case $x = 1$ is trivial. Suppose the induction hypothesis is true, i.e. for any ancestral set $A \in \mathcal{A}^x, x \leq i-1$ the corresponding conditional independence implied by the ordered local Markov property is in $\mathcal{I}_{u_{\mathcal{G}}}$.

Now consider $\mathcal{A}^i$, we will then apply induction on $\mathcal{A}_r^i$. For the base case $r = i$, i.e. $A = [i]$, the independence is in $\mathbb{L}_i$, that is, (a). Now suppose the induction hypothesis is true, that is: for any $A \in \mathcal{A}_r^i, s + 1 \leq r \leq i$, the corresponding conditional independence from the ordered local Markov property is implied by $\mathbb{L}_i$ and we can use the ordered local Markov property on $\mathcal{G}_{[i-1]}$.

Now consider any $A \in \mathcal{A}_s^i$. There is at least one vertex v in $[i] \setminus A$ such that v is parentless in $[i] \setminus A$, in other words, $A \cup \{v\} = A'$ is ancestral and $A' \in \mathcal{A}_{s+1}^i$. Hence by the second induction hypothesis we have

$$i \perp\!\!\!\perp A' \setminus (\mathrm{mb}(i, A') \cup \{i\}) \mid \mathrm{mb}(i, A'). \tag{2}$$

If $\mathrm{mb}(i, A) = \mathrm{mb}(i, A')$, then we can get required conditional independence by marginalizing $v \in A' \setminus (\mathrm{mb}(i, A') \cup \{i\})$.

So now assume $\{v\} \cup \mathrm{mb}(i, A) \subseteq \mathrm{mb}(i, A')$ (v, i are in the same district). Let $A'' = A \cup \{v\} \setminus \{i\} = A' \setminus \{i\}$. We know A'' is ancestral and is in $\mathcal{A}^{i-1}$, thus by the firt induction hypothesis, we can apply the ordered local Markov property to any vertex in $\mathrm{barren}(A'')$ (changing the topological order on $[i-1]$). Moreover, $v \in \mathrm{barren}(A'')$ because if v has any child in A then A is not ancestral. So we have:

$$v \perp\!\!\!\perp A'' \setminus (\mathrm{mb}(v, A'') \cup \{v\}) \mid \mathrm{mb}(v, A'').$$

Next notice that by assumption, $\mathrm{mb}(v, A') \cup \{v\} = \mathrm{mb}(i, A') \cup \{i\}$ (i, v are in the same district in A'). As $A'' \subset A'$, we have $\mathrm{mb}(v, A'') \subset \mathrm{mb}(v, A') = \mathrm{mb}(i, A') \cup \{i\} \setminus \{v\}$. Then because $\mathrm{mb}(v, A'') \subset \mathrm{mb}(i, A') \cup \{i\} \setminus \{v\}$ and $v \in \mathrm{mb}(i, A')$, by semi-graphoids, moving $\mathrm{mb}(i, A') \setminus (\mathrm{mb}(v, A'') \cup \{v\})$ (which is a subset of A'') to the conditioning set, we have:

$$v \perp\!\!\!\perp A'' \setminus \mathrm{mb}(i, A') \mid \mathrm{mb}(i, A') \setminus \{v\}. \tag{3}$$

Now (2) is equivalent to:

$$i \perp\!\!\!\perp A'' \setminus \mathrm{mb}(i, A') \mid \mathrm{mb}(i, A'). \tag{4}$$

Thus by semi-graphoids, we can deduce the following from (3) and (4):

$$\{i, v\} \perp\!\!\!\perp A'' \setminus \mathrm{mb}(i, A') \mid \mathrm{mb}(i, A') \setminus \{v\}.$$

The next step is to marginalize v so that:

$$\{i\} \perp\!\!\!\perp A'' \setminus \mathrm{mb}(i, A') \mid \mathrm{mb}(i, A') \setminus \{v\}.$$

which is equivalent to:

$$\{i\} \perp\!\!\!\perp A \setminus ((\mathrm{mb}(i, A') \setminus \{v\}) \cup \{i\}) \mid \mathrm{mb}(i, A') \setminus \{v\}. \tag{5}$$

Note that $(\mathrm{mb}(i, A') \setminus \{v\}) \subseteq A$. To extract vertices in $(\mathrm{mb}(i, A') \setminus \{v\}) \setminus \mathrm{mb}(i, A)$, the key point is to notice that $\{v, i\} \subseteq \mathrm{barren}(\mathrm{dis}_{A'}(i)) = H$, where H is a head that contains the maximal vertex i. Thus $\{v\}$ can be a marginalizing set K for H with the tail T, and $H \cup T = \mathrm{mb}(i, A')$. Now by marginalizing $\{v\}$ we reach $B = A' \setminus \{v\} = A$, thus the following conditional independence is in $\mathbb{L}_i$:

$$\{i\} \perp\!\!\!\perp \mathrm{mb}(i, A') \setminus ((\mathrm{mb}(i, A) \cup \{v\}) \cup \{i\}) \mid \mathrm{mb}(i, A). \tag{6}$$

From (5) and (6), we can deduce that:

$$i \perp\!\!\!\perp A \setminus (\mathrm{mb}(i, A) \cup \{i\}) \mid \mathrm{mb}(i, A).$$

$\square$

D.2 Missing proofs

Proof of Proposition 4.3. In Theorem D.4, we prove that the list of independences associated with the complete power DAGs, denoted as $\mathbb{L}^{\mathcal{G}}$, is equivalent to the ordered local Markov property. Because $\widetilde{\mathbb{L}}^{\mathcal{G}} \subseteq \mathbb{L}^{\mathcal{G}}$, by Theorem D.4, it is sufficient to prove that $\widetilde{\mathbb{L}}^{\mathcal{G}}$ implies $\mathbb{L}^{\mathcal{G}}$.

We proceed by three inductions. The first induction is on the topological ordering of vertices and it is sufficient to show that given $\mathbb{L}_k^{\mathcal{G}}$ is true for $1 \leq k \leq i-1$, $\mathbb{L}_i^{\mathcal{G}}$ is implied by $\widetilde{\mathbb{L}}_i^{\mathcal{G}}$. The base case is trivial.

The second induction is on the topological ordering on heads. We start from the maximal head in $\mathfrak{I}_i(\mathcal{G})$ and proceed downwards to show that the independences in $\mathbb{L}_i^{\mathcal{G}}$ associated with each head and its parent heads are true. Again, the base case is trivial. Now suppose that for a head H', every independence associated with any heads preceding H' is true, in particular, independences in $\mathbb{L}^{\mathcal{G}}$ associated with any parent head of H' hold. We need to show that independences associated with edges $H \to^k H'$ for any parent head H of H' hold.

Now consider the topological ordering on the vertices in $\mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H'))$, which is made of all the marginalization vertices and we proceed to the third induction on this ordering. The base case is for those parent heads of H' with marginalization vertex $k = \min \mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H'))$. Clearly the independence from the maximal parent head is in $\widetilde{\mathbb{L}}_i^{\mathcal{G}}$, and this implies all independences from other parent heads H such that $H \to^k H'$.

For the inductive step: consider some parent head H_j of H' with marginalization vertex $l \in \mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H'))$ and $l > k$. Take some parent head H_1 of H' with marginalization vertex k, then clearly $k \notin H_j$ and $l \notin H_1$. Let

$$H^m = \mathrm{barren}_{\mathcal{G}}(H_j \cup H_1),$$

then by Proposition C.7, $H^m \to^{kl} H'$. Therefore, we have $H^m \to^l H^{m'}$ for some parent head $H^{m'}$ of H' and also we have $H^{m'} \to^k H'$.

By the hypothesis of the second induction, we have

$$i \perp\!\!\!\perp H^m T^m \setminus H^{m'} T_t^m l \mid H^{m'} T^{m'} \setminus \{i\}.$$

By the hypothesis of the third induction, we have

$$i \perp\!\!\!\perp H^{m'} T^{m'} \setminus H'T'k \mid H'T' \setminus \{i\}.$$

Hence, by the contraction semi-graphoid, we have

$$i \perp\!\!\!\perp H^m T^m \setminus H'T'kl \mid H'T' \setminus \{i\}.$$

Now we have $k \notin H_j$ and also, by construction, $H^m T^m \supseteq H_j T_j$. Therefore, we can marginalize irrelevant vertices to get

$$i \perp\!\!\!\perp H_j T_j \setminus H'T'l \mid H'T' \setminus \{i\}.$$

Next we show it contains fewer and smaller independences than the reduced ordered local Markov property. For each maximal ancestral set A and its maximal vertex i, there is a corresponding head by taking barren of the district of i and this relation is one-to-one. Hence the number of independences in the refined Markov property is not more than the number of independences in the refined Markov property. Further every independence in the refined Markov property can be deduced by an independence from the reduced ordered local Markov property by simply marginalizing some vertices, hence the statement is true. □

Proof for Proposition 4.4. At line 12, the algorithm proceeds by the partial ordering on heads. Thus for any head H, once we arrive at line 12, there will be no edge added that is into it. So it is sufficient to show that for any head H, once we arrive it at line 12, there is only one edge into it, which corresponds to the edge in $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$. We prove this by induction

on the partial ordering on heads. Clearly the algorithm does not add any edge into the maximal head. The base case is for the children of the maximal head and this clearly holds.

For the inductive step, consider all the parent heads of a head H', which exceed H' in the partial ordering and consider H^* and k defined in Definition 4.4. Suppose $H^* \to^k H'$ does not appear and instead we have $H \to^{k'} H'$ for some other parent head of H'.

Firstly if $k' \neq k$ then by Lemma 4.1 and definition of k, we must have $k' > k$. Now it is clear that $k \in \mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H))$, so k must be marginalized at some point to reach H, therefore $k \leq M(H)$ and then line 14 prevents marginalizing of k'.

Then we have $k' = k$, so $H \to^k H'$. Then by definition of H^* we have $H^* > H$, thus from the maximal head, H^* can be reached within fewer steps than H, and line 16 prevents adding $H \to^k H'$. $\qquad\square$

Proof of Proposition 4.5. For Algorithm 1: there are at most $\binom{n}{k}$ heads at line 3 and 12. Then line 14 is of order k. At line 15, it takes $O(n+e)$ to check which new head is reached. Remaining lines all takes constant time or does not contain any loop.

In Algorithm 2, there are at most $\binom{n}{k}$ heads. Then we need to compute the tail for each head and add the corresponding semi-elementary imset. The latter is of constant time and the tail of each head can be computed simultaneously when we visit the head from its parent head, therefore is also of $O(n+e)$ time. $\qquad\square$

Algorithm 3: Obtain the refined power DAGs for a simple MAG

    Input: A simple MAG $\mathcal{G}([n], \mathcal{E})$ ($\mathcal{E}$ stored as adjacencies)
    Result: The refined power DAGs $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ for each i
1 **for** $i \in [n]$ **do**
2   Let smaller siblings of i be $j_1, \ldots, j_k$
3   Set $\widetilde{\mathfrak{I}}_i = \{\{j_k, i\} \to \ldots \to \{j_1, i\} \to \{i\}\}$
4 **end**
5 **return** $(\widetilde{\mathfrak{I}}_1^{\mathcal{G}}, \widetilde{\mathfrak{I}}_2^{\mathcal{G}}, \cdots, \widetilde{\mathfrak{I}}_n^{\mathcal{G}})$

D.3 An example for redundant independences in the refined power DAGs

Example D.1. Consider Figure 13(i). The independences associated with the component of its refined power DAG for 6 under a numerical topological ordering are:

$$6 \perp\!\!\!\perp 3 \mid 1, 2, 4 \qquad\qquad 6 \perp\!\!\!\perp 2 \mid 1, 3, 5 \qquad\qquad 6 \perp\!\!\!\perp 2 \mid 1.$$

Whatever the topological order over the other vertices, there is always a third independence that is redundant. For example, with this numerical ordering, one can deduce $6 \perp\!\!\!\perp 2 \mid 1$ from $6 \perp\!\!\!\perp 2 \mid 1, 3, 5$ and $2 \perp\!\!\!\perp 3, 5 \mid 1$. If one adds an edge between 2 and 3 as in Figure 13(ii), then only semi-graphoids are insufficient to deduce $6 \perp\!\!\!\perp 2 \mid 1$ from only first two independences and independences associated with earlier vertices; hence, in this case, we need the third independence.

This example suggests that a minimal list of independences required to define the model sometimes depends on structures not local to the vertex being considered. We are investigating
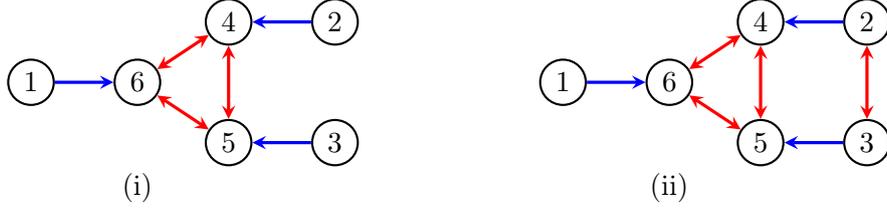
Figure 13: Example where refined Markov property is still redundant

how to obtain such a list, but conjecture that it may be computationally difficult to do so in general.

D.4   Proof of Theorem 4.6 in Section 4.5

Proof of Theorem 4.6. Let $T = \text{tail}_{\mathcal{G}}(H)$. Note that $u_{\mathcal{G}}$ from Theorem 3.3 can be rewritten as:

$$\sum_{i=1}^{n} \left\{ \delta_{[i]} - \delta_{[i-1]} - \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ i \geq H}} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup T} \right\}.$$

By induction, if we restrict to the final vertex of a topological ordering, say n, then all we need to prove is that:

$$\delta_{[n]} - \delta_{[n-1]} - \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ n \geq H}} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup T}$$

$$= u_{\langle n, [n-1] \setminus \text{mb}(n,[n]) | \text{mb}(n,[n]) \rangle} + \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \setminus \{n\} \\ n \geq H}} \sum_{\substack{\emptyset \subset K \subseteq H \setminus \{i\}: \\ H \to^K H'}} (-1)^{|K|+1} u_{\langle i, HT \setminus H'T'K | H'T' \setminus i \rangle}.$$

Note that $u_{\langle n, [n-1] \setminus \text{mb}(n,[n]) | \text{mb}(n,[n]) \rangle} = \delta_{[n]} - \delta_{[n-1]} - \delta_{\{n\} \cup \text{mb}(n)} + \delta_{\text{mb}(n)}$, so we can reduce the equivalence to

$$- \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ n \geq H}} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup T} \tag{7}$$

$$= \delta_{\text{mb}(n,[n])} - \delta_{\{n\} \cup \text{mb}(n,[n])}$$
$$+ \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \setminus \{n\} \\ n \geq H}} \sum_{\substack{\emptyset \subset K \subseteq H \setminus \{i\}: \\ H \to^K H'}} (-1)^{|K|+1} (\delta_{HT \setminus K} - \delta_{(HT \setminus Kn)} - \delta_{H'T'} + \delta_{H'T' \setminus n}). \tag{8}$$

We can rewrite (7) as:

$$- \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ n \geq H}} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup T}$$

$$= - \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ n \geq H}} \sum_{K \subseteq H} (-1)^{|K|} \delta_{HT \setminus K}$$

$$= \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ n \geq H}} \sum_{K \subseteq H \setminus \{n\}} (-1)^{|K|+1} (\delta_{HT \setminus K} - \delta_{HT \setminus Kn}). \tag{9}$$

43

The repeating part is only $-\delta_{HT} + \delta_{HT\backslash n}$ for each H (when $K = \emptyset$). For each H, the two terms do not appear in the summation for H (as we rule out the case when $K = \emptyset$ in Theorem 4.6) but they appear when other heads marginalized to H and the two terms are multiplied by some constants.

Our objective is to show that (9) and (8) are equivalent, for which it is sufficient to prove that for every head H, the coefficient of $-\delta_{HT} + \delta_{HT\backslash i}$ is 1 in (8). For the largest head, which is barren($\mathrm{dis}_{[n]}(n)$), it is clearly true because there is no head that is 'larger' than it and for this head $-\delta_{HT} + \delta_{HT\backslash i}$ is simply $\delta_{\mathrm{mb}(n,[n])} - \delta_{\{n\}\cup\mathrm{mb}(n,[n])}$, which never appears in the summation.

For any other head H with maximal vertex n, we need to prove:

$$\sum_{K,H':H'\to^K H} (-1)^{|K|+1} = 1.$$

Note that the summation is over both H' and K because for a pair of heads there might exist different marginalization sets that lead one to another.

Now by Lemma C.2, any head H' that can reach H with multiple marginalizing sets and minimal marginalizing set K' contributes $\sum_{B:B\subseteq(H'\backslash(H\cup K'))}(-1)^{|K'|+|B|+1} = 0$ to the coefficient of $-\delta_{H\cup T} + \delta_{(H\backslash\{n\}\cup T)}$.

Thus it is sufficient to consider any head H' that can reach H with the only (minimal) marginalizing set $K = H' \backslash H$. By Lemma C.3, we find all these heads and in total they contribute

$$\sum_{K:\emptyset\subset K\subseteq\mathrm{ceil}(\mathrm{sib}(\mathrm{dis}_{\mathrm{an}(H)}(i))\backslash\mathrm{dis}_{\mathrm{an}(H)}(i))} (-1)^{|K|+1} = 1$$

to the coefficient of $-\delta_{H\cup T} + \delta_{(H\backslash\{n\}\cup T)}$. $\qquad\square$

Proof for Corollary 4.6.1. By Theorem 4.6, the 'standard' imset $u_{\mathcal{G}}$ can be expressed as one combinatorial imset u_p subtracted by another combinatorial imset u_n where u_p are obtained by adding all semi-elementary imsets with positive coefficients (where $|K|$ is odd), and similarly for u_n but with negative coefficients (where $|K|$ is even).

Let u_p^1 be sum of semi-elementary imsets corresponding to marginalize only one vertex, which are all in u_p. By Theorem D.4, $I_{\mathcal{G}} \subseteq \mathcal{I}_{u_p^1}$. The faithfulness result in Richardson and Spirtes (2002) shows that for every MAG $\mathcal{G}$ there exists a distribution P such that $\mathcal{I}_P = \mathcal{I}_{\mathcal{G}}$ and Theorem 5.2 in Studený (2005) implies the existence of a structural imset u such that $I_u = I_P$. Now every independence in the list of independences that we used to construct u_p^1 is in $\mathcal{I}_u$ by Theorem D.4, so by Lemma 6.1 in Studený (2005), we have $I_{u_p^1} \subseteq \mathcal{I}_u = \mathcal{I}_{\mathcal{G}}$. Hence $\mathcal{I}_{u_p^1} = \mathcal{I}_{\mathcal{G}}$. Now the remaining semi-elementary imsets that we use to construct u_p are those corresponding to marginalize odd number of vertices (more than one), and the independences they correspond to are all in $\mathcal{I}_{\mathcal{G}} = \mathcal{I}_{u_p^1}$. Therefore $\mathcal{I}_{u_p^1} = \mathcal{I}_{u_p}$.

Suppose $u_{\mathcal{G}}$ is structural and take any distribution P that is not Markov to it. Studený (2005) shows that a distribution is Markov to a structural imset if and only if the inner product between the imset and entropy vector of the distribution is zero. Moreover this inner product is non-negative. Therefore, for P and $u_{\mathcal{G}}$, this inner product is positive. Further, as $u_{\mathcal{G}}$ is also $u_p - u_n$ where u_p and u_n are both combinatorial, this means that the inner product between P's entropy vector and u_p are also positive, and hence P is not Markov to u_p. Thus P is not Markov to $I_{\mathcal{G}}$. As a result, $\mathcal{I}_{u_{\mathcal{G}}} \subseteq \mathcal{I}_{\mathcal{G}}$. $\qquad\square$
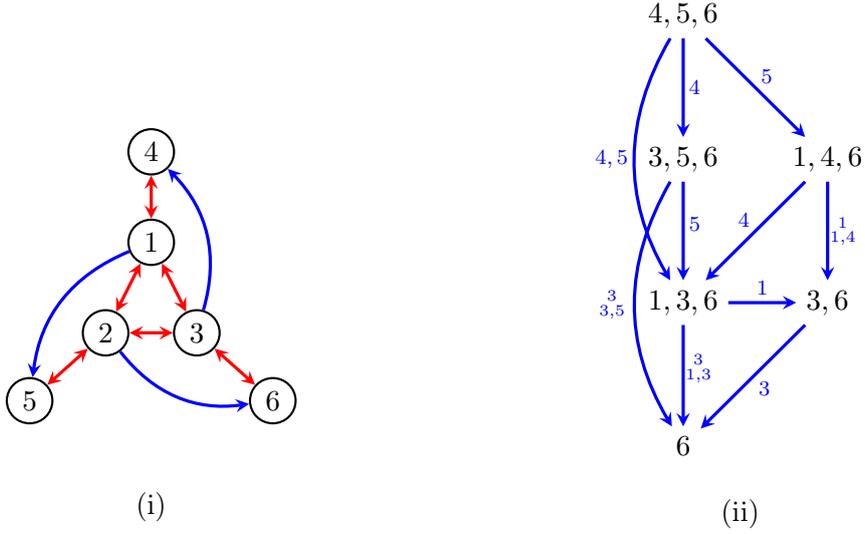
Figure 14: (i) A MAG $\mathcal{G}$ (ii) A complete power DAG on the heads of $\mathcal{G}$ with maximal vertex 6, under the numerical topological ordering.

Definition D.3. Let $\mathcal{G}$ be a MAG containing a district D. Then by $\mathcal{G}^{|D}$ denote the induced subgraph over $D \cup \mathrm{pa}_{\mathcal{G}}(D)$, but where all vertices within $\mathrm{pa}_{\mathcal{G}}(D) \setminus D$ have been joined by bidirected edges.

Corollary D.4.1. *Let $\mathcal{G}$ be a MAG. The 'standard' imset $u_{\mathcal{G}}$ is perfectly Markovian with respect to $\mathcal{G}$ if the 'standard' imsets $u_{\mathcal{G}^{|D}}$ are perfectly Markovian with respect to $\mathcal{G}^{|D}$.*

Proof. Take the expression for the standard imset given in Then note that, for a district D, each summand other than the very first term only contains independences between an element of the district, and other elements of the district and its parents, and the collection of all first terms gives a DAG model. Furthermore, the subimset defined by the second inner summation represents the model in which all the vertices in $\mathrm{pa}_{\mathcal{G}}(D) \setminus D$ have been joined by an edge. Since we know that any MAG can be defined by independences only within the district (plus the initial independence in the expression above), this shows that a 'standard' imset for a MAG $\mathcal{G}$ will be perfectly Markovian with respect to the graph if and only if the associated standard imsets for each district and its parents are perfectly Markovian with respect to the induced subgraph obtained after filling in any missing edges between parents. □

Example D.2. Consider Figure 14. The multiple labels on the edges in (ii) means that there are different sets of vertices that can be marginalized to lead from one head to another.

The edge $\{4,5,6\} \to^{45} \{1,3,6\}$ indicates that when we compute the semi-elementary imset for $\{4,5,6\}$ by marginalizing vertices in $\{4,5\}$, we only reach the head $\{1,3,6\}$ once. Hence the head $\{4,5,6\}$ contributes $(-1)^{2+1} = -1$ to the coefficient of $-\delta_{1236} + \delta_{123}$ for the head $\{1,3,6\}$. Similarly the edge from $\{3,5,6\}$ to $\{6\}$ means that we may reach to the head $\{6\}$ from the head $\{3,5,6\}$ by marginalizing either $\{3\}$ or $\{3,5\}$. So the head $\{3,5,6\}$ contributes $(-1)^{1+1} + (-1)^{2+1} = 0$ to the coefficient of $-\delta_{26} + \delta_2$ for the head $\{6\}$.

E Scoring criteria

The BIC (Schwarz, 1978) is a consistent (defined below) scoring criterion. For discrete models, Evans and Richardson (2014, 2010) provide procedures for fitting ADMGs by maximum likelihood, and thus we can compute the BIC.

Let ℓ be the log-likelihood and q^θ denote the family of distributions that are Markov to a fitted graph $\mathcal{G}$, with parameter θ; denote the maximum likelihood estimator for θ by $\hat{\theta}$. Let $d = |\mathcal{S}(\mathcal{G})|$ be the dimension of the binary model (Evans and Richardson, 2014), and N and $N(x_\mathcal{V})$ be the number of samples and the number of samples such that $X_\mathcal{V} = x_\mathcal{V}$, respectively. Then the BIC for the model defined by $\mathcal{G}$ is:

$$-2\hat{\ell} + d \log N,$$

where $\hat{\ell} = \sum_{x_\mathcal{V}} N(x_\mathcal{V}) \log q_{\hat{\theta}}(x_\mathcal{V})$.

However this score is not very suitable for a greedy learning algorithm as we need to re-fit the whole graph when we consider new models. Thus in this chapter we aim to study a scoring criterion proposed by Andrews (2021) that is decomposable with respect to the parametrizing set $\mathcal{S}(\mathcal{G})$. The motivation is that equivalent MAGs have the same BIC and the parametrizing set, and every time we move between Markov equivalence classes of MAGs, we simply change the 0-1 vector of the characteristic imset, and it would save a lot computations if for those sets remaining in the parametrizing set, we do not need to compute their corresponding scores again.

The proposing scoring criterion is $-2N \langle c_\mathcal{G}, \hat{\mathsf{I}} \rangle + d \log N$, where $\hat{\mathsf{I}}$ is the plug-in estimate of *interaction information*, defined below. Thus we are approximating $\hat{\ell}$ with $\langle c_\mathcal{G}, \hat{\mathsf{I}} \rangle$.

E.1 Definitions

Definition E.1. The *entropy* of a set of discrete variables $X_\mathcal{V}$, is defined as:

$$\mathsf{H}(X_\mathcal{V}) = \sum_{x_\mathcal{V}} P(X_\mathcal{V} = x_\mathcal{V}) \log P(X_\mathcal{V} = x_\mathcal{V}).$$

Definition E.2. The interaction information of a set of discrete variables $X_\mathcal{V}$, is defined as:

$$\mathsf{I}(X_\mathcal{V}) = \sum_{S \subseteq V} (-1)^{|V \setminus S|} \mathsf{H}(X_S).$$

We have the following identity:

$$\mathsf{H}(X_\mathcal{V}) = \sum_{T \subseteq V} \sum_{S \subseteq T} (-1)^{|T \setminus S|} \mathsf{H}(X_S) = \sum_{T \subseteq V} \mathsf{I}(X_T).$$

The plug-in estimate of entropy of $X_\mathcal{V}$, is

$$\hat{\mathsf{H}}(X_\mathcal{V}) = \sum_{x_\mathcal{V}} \frac{N(x_\mathcal{V})}{N} \log \frac{N(x_\mathcal{V})}{N}.$$

The plug-in estimate of interaction information of $X_\mathcal{V}$, is

$$\hat{\mathsf{I}}(X_\mathcal{V}) = \sum_{S \subseteq V} (-1)^{|V \setminus S|} \hat{\mathsf{H}}(X_S).$$

Theorem 4.6 allows us to decompose the entropy in terms of the interaction information over the parametrizing set. We first need a definition to simplify the expression.

Definition E.3. Given a function f which takes X_A for any $A \subseteq V$ as input, and an imset u over V, we define

$$\langle u, f \rangle = \sum_{A \subseteq V} u(A) f(x_A).$$

Let H, I be the entropy function and the interaction information function, respectively.

Proposition E.1. *For a MAG $\mathcal{G}$ with vertex set V and a distribution p that is Markov to $\mathcal{G}$, we have*

$$\mathsf{H}(X_{\mathcal{V}}) = \sum_{T \in \mathcal{S}(\mathcal{G})} \mathsf{I}(X_T) = \langle c_{\mathcal{G}}, \mathsf{I} \rangle = \langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \mathsf{H} \rangle.$$

Proof. Recall that $\mathcal{P}(S)$ denotes the power-set of S. By Theorem 4.6, it is sufficient to show that $\langle u_{\mathcal{G}}, \mathsf{H} \rangle = \sum_{T \notin \mathcal{S}(\mathcal{G})} \mathsf{I}(X_T) = 0$.

$$
\begin{aligned}
\langle u_{\mathcal{G}}, \mathsf{H} \rangle &= \sum_{S \in \mathcal{P}(\mathcal{V})} \sum_{\substack{T \in \mathcal{P}(\mathcal{V}) \\ S \subseteq T \subseteq \mathcal{V}}} (-1)^{|T \setminus S|} (1 - c_{\mathcal{G}}(T)) \mathsf{H}(X_S) \\
&= \sum_{S \in \mathcal{P}(\mathcal{V})} \sum_{\substack{T \in \mathcal{P}(\mathcal{V}) \\ S \subseteq T \subseteq \mathcal{V}}} (-1)^{|T \setminus S|} \delta_{T \notin \mathcal{S}(\mathcal{G})} \mathsf{H}(X_S) \\
&= \sum_{S \in \mathcal{P}(\mathcal{V})} \sum_{\substack{T \notin \mathcal{S}(\mathcal{G}) \\ S \subseteq T \subseteq \mathcal{V}}} (-1)^{|T \setminus S|} \mathsf{H}(X_S) \\
&= \sum_{T \notin \mathcal{S}(\mathcal{G})} \sum_{S \subseteq T} (-1)^{|T \setminus S|} \mathsf{H}(X_S) \\
&= \sum_{T \notin \mathcal{S}(\mathcal{G})} \mathsf{I}(X_T).
\end{aligned}
$$

The last equality in the statement then holds, as $\langle \delta_{\mathcal{V}}, \mathsf{H} \rangle = \mathsf{H}(X_{\mathcal{V}})$. $\qquad \square$

A consistent scoring criterion requires the following property.

Definition E.4. Let $\mathcal{D}$ be a set of data that are i.i.d. samples from a distribution P. A scoring criterion S is asymptotically consistent if in the limit when the number of sample grows, the following holds:

1. If $\mathcal{G}$ contains P and $\mathcal{G}'$ does not, then $S(\mathcal{G}, \mathcal{D}) < S(\mathcal{G}', \mathcal{D})$.

2. If both $\mathcal{G}$ and $\mathcal{G}'$ contain P but $\mathcal{G}'$ has higher dimension then $\mathcal{G}$, then $S(\mathcal{G}, \mathcal{D}) < S(\mathcal{G}', \mathcal{D})$.

E.2   Bayesian networks

We firstly show that if we are fitting DAG models, then $\hat{\ell}$ is equal to $N \langle c_{\mathcal{G}}, \hat{\mathsf{I}} \rangle$. Recall that for DAGs, the maximum likelihood estimate (MLE) of $P(x_v \mid x_{\mathrm{pa}_v})$ is $N(x_v, x_{\mathrm{pa}_v})/N(x_{\mathrm{pa}_v})$.

Given N i.i.d. samples, the log-likelihood can be expressed as:

$$\ell(P; N) = \sum_{x_\mathcal{V}} N(x_\mathcal{V}) \log P(x_\mathcal{V})$$

$$= \sum_{x_\mathcal{V}} N(x_\mathcal{V}) \sum_v \log P(x_v \mid x_{\mathrm{pa}(v)})$$

$$= \sum_v \sum_{x_v, x_{\mathrm{pa}(v)}} N(x_v, x_{\mathrm{pa}(v)}) \log P(x_v \mid x_{\mathrm{pa}(v)})$$

$$= \sum_v \sum_{x_{\mathrm{pa}(v)}} \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log P(x_v \mid x_{\mathrm{pa}(v)}).$$

Replacing $P(x_v \mid x_{\mathrm{pa}(v)})$ by the MLE $N(x_v, x_{\mathrm{pa}_v})/N(x_{\mathrm{pa}_v})$, the log-likelihood is then

$$\hat{\ell} = \sum_v \sum_{\mathrm{pa}(v)} \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log \frac{N(x_v, x_{\mathrm{pa}(v)})}{N(x_{\mathrm{pa}(v)})}$$

$$= \sum_v \sum_{\mathrm{pa}(v)} \left\{ \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log \frac{N(x_v, x_{\mathrm{pa}(v)})}{N} + \right.$$

$$\left. - \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log \frac{N(x_{\mathrm{pa}(v)})}{N} \right\}$$

$$= \sum_v \sum_{\mathrm{pa}(v)} \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log \frac{N(x_v, x_{\mathrm{pa}(v)})}{N} +$$

$$- \sum_v \sum_{\mathrm{pa}(v)} N(x_{\mathrm{pa}(v)}) \log \frac{N(x_{\mathrm{pa}(v)})}{N}$$

$$= N \langle \delta_\mathcal{V} - u_\mathcal{G}, \hat{\mathsf{H}} \rangle$$

$$= N \langle c_\mathcal{G}, \hat{\mathsf{I}} \rangle.$$

The above calculations have appeared in the literature before (Studený, 2005) but we demonstrate it in our notation to make other proofs clear.

E.3 Consistency for MAGs when imset is perfectly Markovian

In this section, we show that when fitted MAGs have imsets that are perfectly Markovian w.r.t. the graphs, the scoring criterion is consistent. First, we introduce the well-studied Kullback-Leibler divergence.

Definition E.5. For two distributions P and Q defined over $X_\mathcal{V}$, and two disjoint subsets $A, C \subseteq V$, define the *Kullback-Leibler* (KL) divergence between $P_{A|C}$ and $Q_{A|C}$ as

$$KL(P_{A|C} \parallel Q_{A|C}) = \mathbf{E}_{P_{A|C}}[\log \frac{P_{A|C}}{Q_{A|C}}].$$

Two nice properties of KL divergence are that it is non-negative, and that it is zero if and only if $P = Q$ almost surely.

Proposition E.2. *For any two distributions P and Q and any semi-elementary imset $u_{\langle A,B|C\rangle}$, provided that Q satisfies $X_A \perp\!\!\!\perp X_B \mid X_C$, we have*

$$\langle u_{\langle A,B|C\rangle}, KL(P \parallel Q)\rangle \geq 0,$$

with equality holding if and only if P also satisfies $X_A \perp\!\!\!\perp X_B \mid X_C$.

Proof.
$$
\begin{aligned}
\langle u_{\langle A,B|C\rangle}, KL(P \parallel Q)\rangle &= \mathbf{E}_{P_C}[KL(P_{AB|C} \parallel Q_{AB|C}) - KL(P_{A|C} \parallel Q_{A|C}) - KL(P_{B|C} \parallel Q_{B|C})] \\
&= \mathbf{E}_{P_C}\left[\mathsf{I}(A;B \mid C) - \mathbf{E}_{P_{AB|C}} \log \frac{Q_{AB|C}}{Q_{A|C}Q_{B|C}} \right] \\
&= \mathbf{E}_{P_C}[\mathsf{I}(A;B \mid C)] \qquad \text{(provided } Q \text{ satisfies } A \perp\!\!\!\perp B \mid C) \\
&\geq 0,
\end{aligned}
$$

where the last inequality comes from the fact that mutual information is always non-negative. $\qquad\square$

Next we show that the score is consistent.

Proposition E.3. *The score $-2N\langle c_{\mathcal{G}}, \hat{\mathsf{I}}\rangle + d\log N$, is consistent when $\mathcal{I}_{u_{\mathcal{G}}} = \mathcal{I}_{\mathcal{G}}$, where $\mathcal{G}$ is the fitting MAG.*

Proof. Let us first consider any general MAG $\mathcal{G}$ and q^θ with the parameters $\theta = \{q(x_H \mid x_{\text{tail}(H)}) : H \in \mathcal{H}(\mathcal{G})\}$ (Evans and Richardson, 2014). Note that q^θ can be factorized by Theorem 4.6 and the log-likelihood can be rewritten as the inner product between $\delta_{\mathcal{V}} - u_{\mathcal{G}}$ and $\log q^\theta$ where the inner product is taken over all subsets of V:

$$
\begin{aligned}
\ell &= \sum_{x_{\mathcal{V}}} N(x_{\mathcal{V}})\langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \log q^\theta\rangle \\
&= N\langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \sum_{x_{\mathcal{V}}} \frac{N(x_{\mathcal{V}})}{N} \log q^\theta\rangle.
\end{aligned}
$$

The difference between $\hat{\ell}$ and $N\langle c_{\mathcal{G}}, \hat{\mathsf{I}}\rangle$ is then

$$
\begin{aligned}
\hat{\ell} - N\langle c_{\mathcal{G}}, \hat{\mathsf{I}}\rangle &= N\langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \sum_{x_{\mathcal{V}}} \frac{N(x_{\mathcal{V}})}{N} \log q^{\hat{\theta}} - \hat{\mathsf{H}}\rangle \\
&= -N\langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, KL(\hat{P} \parallel q^{\hat{\theta}})\rangle.
\end{aligned}
$$

Hence the difference from the BIC of the fitted model $q^{\hat{\theta}}$ with graph $\mathcal{G}$ and the new score is

$$D_1 = 2N \times \langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, KL(\hat{P} \parallel q^{\hat{\theta}})\rangle.$$

Moreover the difference between the fitted model $q^{\hat{\theta}}$ with graph $\mathcal{G}$ and the BIC of the true model $p^{\hat{\theta}'}$ is:

$$
\begin{aligned}
D_2 &= -2\sum_{x_{\mathcal{V}}} N(x_{\mathcal{V}}) \log q^{\hat{\theta}}(x_{\mathcal{V}}) + d_2 \log N + 2\sum_{x_{\mathcal{V}}} N(x_{\mathcal{V}}) \log p^{\hat{\theta}'}(x_{\mathcal{V}}) - d_1 \log N \\
&= 2N \sum_{x_{\mathcal{V}}} \hat{P}(x_{\mathcal{V}}) \log \frac{p^{\hat{\theta}'}(x_{\mathcal{V}})}{q^{\hat{\theta}}(x_{\mathcal{V}})} + (d_2 - d_1) \log N
\end{aligned}
$$

$$= 2N \sum_{x_{\mathcal{V}}} \hat{P}(x_{\mathcal{V}}) \log \left[\frac{\hat{P}(x_{\mathcal{V}})}{q^{\hat{\theta}}(x_{\mathcal{V}})} \frac{p^{\hat{\theta}'}(x_{\mathcal{V}})}{\hat{P}(x_{\mathcal{V}})} \right] + (d_2 - d_1) \log N$$

$$= 2N \times KL(\hat{P} \parallel q^{\hat{\theta}}) - 2N \times KL(\hat{P} \parallel p^{\hat{\theta}'}) + (d_2 - d_1) \log N,$$

where d_1, d_2 denote the dimension of the true model and fitted model, respectively.

To show that the score is consistent, it is sufficient to show that the following conditions are satisfied:

(i) when $D_2 = 0$, i.e. the fitted model is the true model, $D_1 = O_p(1)$;

(ii) when $D_2 > 0$, i.e. the fitted model is not the true model, we have $D_2 - D_1$ diverges at rate at least $O_p(\log N)$.

Consider condition (i) when we are fitting the true model. For any subset $A \subseteq V$, the term

$$2N \times KL(\hat{P}(X_A) \parallel q^{\hat{\theta}}(X_A))$$

has χ^2 distribution with some degree of freedom. Since $\delta_{\mathcal{V}} - u_{\mathcal{G}}$ has fixed number of terms, $D_1 = O_p(1)$ on all these as N tends to infinity.

Consider condition (ii), we have:

$$D_2 - D_1 = 2N \times \langle u_{\mathcal{G}}, KL(\hat{P} \parallel q^{\hat{\theta}}) \rangle - 2N \times KL(\hat{P} \parallel p^{\hat{\theta}'}) + (d_2 - d_1) \log N.$$

The second term $2N \times KL(\hat{P} \parallel p^{\hat{\theta}'})$ has $\chi^2_{d_1}$ distribution, so it is $O_p(1)$. Moreover, by Proposition E.2 and given that $u = \sum u_{\langle A, B | C \rangle}$ is combinatorial, we have:

$$2N \times \langle u_{\mathcal{G}}, KL(\hat{P} \parallel q^{\hat{\theta}}) \rangle = 2N \sum \mathbb{E}_{\hat{P}_C}[\mathsf{I}(A; B \mid C)]$$

Now we need to split into two cases. The first case is if the fitted model contains the true model, but has higher dimension and contains more parameters then $d_2 > d_1$ so the third term $(d_2 - d_1) \log N$ is at $O(\log N)$. For the first term, because the fitted model contains the true model, it satisfies all the conditional independence corresponding to the semi-elementary imsets in u_G. Treating each $2N \times \mathbb{E}_{\hat{P}_C}[\mathsf{I}(A; B \mid C)]$ as a likelihood ratio test, it has chi-squared distribution, so in total the first two terms are at $O_p(1)$. Thus in this case, $D_2 - D_1 = O_p(\log N)$.

For the second case when the fitted model is wrong and does not contain the true model, given that $u_{\mathcal{G}}$ is combinatorial and is perfectly Markovian with respect to the graph, there exists at least one conditional independence in $u_{\mathcal{G}} = \sum u_{\langle A, B | C \rangle}$ such that $\mathbb{E}_{P_C}[\mathsf{I}(A; B \mid C)] = \lambda > 0$. Empirically, when N tends to infinity, $\mathbb{E}_{\hat{P}_C}[\mathsf{I}(A; B \mid C)]$ will be close to λ, and hence the first term in $D_2 - D_1$ grows at $O_p(N)$. Even if the third term $(d_2 - d_1) \log N$ is negative, it grows at $O(\log N)$, which is slower than the first term, hence in the second case, $D_2 - D_1 = O_p(N)$. $\square$

F   Bidirected graphs

For general MAGs, the problem of characterizing the conditions for when $\mathcal{I}_{u_{\mathcal{G}}}$ is well defined and $\mathcal{I}_{u_{\mathcal{G}}} = \mathcal{I}_{\mathcal{G}}$, seems hard in general. To reduce the difficulty, we focus on bidirected graphs

in this section. We will give a condition such that our proposed 'standard' imsets $u_\mathcal{G}$ for *bidirected graphs* are always combinatorial and perfectly Markovian with respect to $\mathcal{G}$. We have computationally verified that this condition is also necessary for $|\mathcal{V}| \leq 7$, and not found any graphs for which it is not.

Definition F.1. Let $\mathcal{G}$ be a bidirected graph, and define its undirected *dual graph*, $\overline{\mathcal{G}}$, by $i - j \in \overline{\mathcal{G}}$ if and only if $i \not\leftrightarrow j$ in $\mathcal{G}$.

The dual graph is a powerful tool when we analyse the bidirected graphs; see Example F.1.

F.1   How does the characteristic imset help?

There are many MAGs that are not simple, but still have combinatorial standard imsets that are perfectly Markovian with respect to the graph; one example is the bidirected 4-cycle, which has a standard imset consisting of the elementary imsets for its two marginal independences. For bidirected graphs, we found that since the vertices can be given any topological order and there are many heads (any connected subgraph), it is difficult to directly decompose the standard imset or prove the validity of decomposition; however, it turns out to be easier if we work with the characteristic imset. In this section, we consider the relationship between the semi-elementary imset decomposition of combinatorial standard imsets and the characteristic imset.

Recall that two conditional independences I_1, I_2 overlap if $\bar{\mathcal{S}}(I_1) \cap \bar{\mathcal{S}}(I_2) \neq \emptyset$.

Proposition F.1. *A 'standard' imset $u_\mathcal{G}$ is combinatorial if and only if there is a list of non-overlapping conditional independences $\mathbb{L}$ such that $\cup_{I \in \mathbb{L}} \bar{\mathcal{S}}(I) = \mathcal{P}(\mathcal{V}) \setminus \mathcal{S}(\mathcal{G})$.*

Proof. This follows from Lemma 3.5, Corollary 3.6.1 and the arguments used in their proofs. $\qquad\square$

For example, if $\mathcal{G}$ is the bidirected 4-cycle $1 \leftrightarrow 2 \leftrightarrow 3 \leftrightarrow 4 \leftrightarrow 1$ the independences are $1 \perp\!\!\!\perp 3$ and $2 \perp\!\!\!\perp 4$. The corresponding sets to be constrained are $\{1, 3\}$ and $\{2, 4\}$, and these are clearly disjoint. Therefore the standard imset is perfectly Markovian and is simply $u_\mathcal{G} = u_{\langle 1,3 \rangle} + u_{\langle 2,4 \rangle}$.

Proposition 3.4 allows us to quickly determine if a standard imset is not perfectly Markovian w.r.t. the graph by checking if the graph contains certain ancestral structures.

Remark 9. In Figure 15, we display all dual graphs with at most 6 vertices that must not appear as induced subgraphs to the dual of a bidirected graph for the standard imset to be perfectly Markovian with respect to the graph. This is primarily a consequence of Proposition 3.4.

F.2   For which bidirected graphs do we get standard imsets that are perfectly Markovian?

We will present a theorem which gives sufficient criteria for bidirected graphs to have combinatorial standard imsets and induce the same model as the graph. For $|\mathcal{V}| \leq 7$ we have empirically verified that this is also necessary. Before we present the theorem, we give a motivating example.
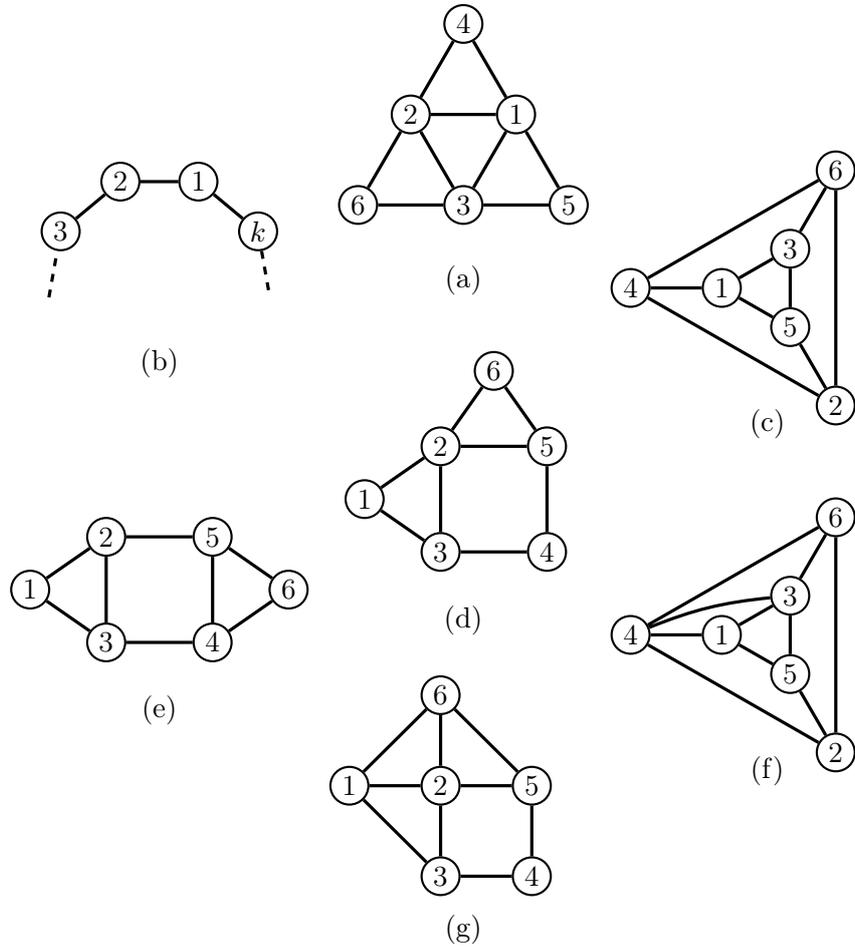
Figure 15: Forbidden dual subgraphs: (a) triangles; (b) a k-cycle, for $k \geq 5$; (c) dual to the 6-cycle; (f) dual to the 6-chain; (d), (e), (g) other graphs with at least one chordless 4-cycle.
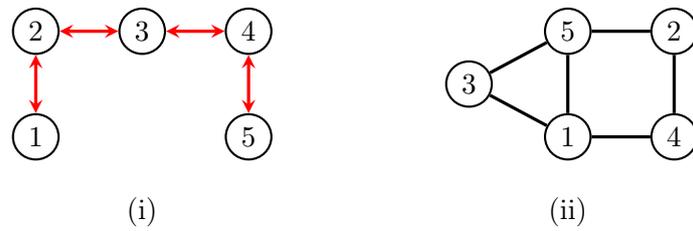

Figure 16: (i) The bidirected 5-chain and (ii) its dual graph.

Example F.1. Consider the bidirected 5-chain and its dual graph in Figure 16. Consider the following list of conditional independences:

$$\left\{ \begin{array}{l} 4 \perp\!\!\!\perp 2 \,|\, 1, 5 \\ 4 \perp\!\!\!\perp 1 \,|\, 3, 5 \end{array} \right\} \qquad\qquad 2 \perp\!\!\!\perp 5 \,|\, 1, 3 \qquad\qquad 5 \perp\!\!\!\perp 1, 3 \qquad\qquad 1 \perp\!\!\!\perp 3.$$

One can check that the sum of semi-elementary imsets corresponding to the above list of independences is the same as the standard imset for the bidirected 5-chain. In addition the standard imset is perfectly Markovian with respect to the graph.

This decomposition starts with 4, and by symmetry it could also start with 2; however, none of the other vertices will work, and we cannot obtain a decomposition in this manner if we try to do so. In particular, if we take the imset for the graph where all edges for 1, 3, 5 are added and subtract from it the standard imset for the 5-chain, what remains is not a structural imset. This shows that for bidirected graphs, even though the topological order can be arbitrary, in order to properly decompose the standard imset further restrictions are required.

Here are some observations. In the dual graph, 4 has neighbours 1 and 2, these vertices share one common neighbour 5, and 1 has one more neighbour 3. However for 3, two of its neighbours have distinct neighbours. This suggests that perhaps we need the neighbours of neighbours to be nested within one another.

For a vertex v, we may also want to treat its neighbours which have the same neighbours as a block, since any path to any one of them also lead to any other vertex in the block. Then any edge within one block does not have any effect on any path to v.

Suppose we have a vertex v and that it has neighbours A in the dual graph. Now partition its neighbours $A^j, 1 \leq j \leq m$ such that the neighbours of each $w \in A^j$ (outside $\{v\} \cup A$) are precisely the set N^j, and such that $N^1 \subseteq N^2 \subseteq \cdots \subseteq N^m$.

Now the question is how the edges connects the blocks of neighbours of v so that v is a valid ending vertex. One observation from the 5-chain is that for two blocks A_i and A_j, if there is one edge between A_i and A_j, then we want A_i and A_j to be fully connected. For example, consider vertex 1 which has neighbours $\{3, 4, 5\}$. Vertices 4 and 5 share the same neighbours outside A, so they are in the same block, while 3 is connected only to 5.

Another question is how A_i and A_j are connected, i.e. the cross-block edges. Suppose v has three blocks of neighbours, A_1, A_2 and A_3. There are 8 possible ways to connect the three blocks. We cannot write out a proper decomposition for three of them. They are: $A_1 - A_3$, $A_2 - A_3$ and $A_1 - A_2 - A_3$. Inspired by the 5-chain example, we have the following definition. For convenience, we introduce the notation $A^{[i,j]}$ to denote $\bigcup_{k=i}^{j} \{A^k\}$ and each A^k is referred as a block (of vertices).

Definition F.2. For a vertex v and an ascending partition of its neighbours $A^{[j_l, j_h]}$, we say it is *rooted* if:

(i) it is empty; or

(ii) there exists a block A^j (the root) for some $j_l \leq j \leq j_h$ and $H^j := A^{[j+1, j_h]}$ and $L^j := A^{[j_l, j-1]}$ (respectively empty if $j = j_h$ or j_l) such that A^j and each set in H^j are fully connected to each set in L^j, and for every set in H^j it is entirely disconnected from A^j. In addition, H^j and L^j must themselves be rooted.

One can check that for three blocks and 8 possible ways to connect them, the three ways that do not lead to a proper decomposition also do not satisfy the above definition, but that the other five ways do.

Definition F.3. For a sequence of blocks $A^{[j_l, j_h]}$ rooted at A^j, we define T^j to be the subset of $A^{[1, j_l - 1]}$ such that A^i is in T^j if and only if it is connected to A^j.

Lemma F.2. *Let $A^{[j_l, j_h]}$ be a collection of neighbours of v that is rooted. Then any subset of it is also rooted.*

Lemma F.3. *For a vertex v and a collection of its consecutive neighbours, if it is rooted then the root is unique.*

Note that H^j or L^j and $A^{[i,j]}$ are written as sets of sets, but when we use them in certain set expressions or in any independence, we just think of it as the union of the blocks contained in that set, i.e. sets of variables. The context should prevent any confusion.

Lemma F.4. *Suppose a block $A^{[j_l, j_h]} = H^j \cup L^j \cup \{A^j\}$ is rooted at A^j, where H^j and L^j are rooted at A^k and A^t respectively (non-empty), then $T^k = T^j \cup L^j$ and $T^t = T^j$. Moreover $k_l = j + 1$ and $t_l = j_l$.*

Theorem F.5. *For a bidirected graph, if there exists an ordering of the vertices such that for each i in $\overline{\mathcal{G}}_{[i]}$, its neighbours A can be partitioned into blocks $\{A^j, 1 \leq j \leq m\}$ such that*

(i) *the neighbours of each $w \in A^j$ (outside $\{v\} \cup A$) in $\overline{\mathcal{G}}$ are the same set N^j;*

(ii) $N^1 \subseteq N^2 \subseteq \cdots \subseteq N^m$;

(iii) *for any two blocks, either there is no edge between them or they are fully connected;*

(iv) $\{A^j, 1 \leq j \leq m\}$ *are rooted.*

Then $u_{\mathcal{G}}$ is combinatorial and $\mathcal{I}_{u_{\mathcal{G}}} = \mathcal{I}_{\mathcal{G}}$.

The key to the proof of Theorem F.5 is to construct a proper independence decomposition of $u_{\mathcal{G}}$ with the aid of Proposition F.1, and show that it is equivalent to the global Markov property. The following example will give some intuition on how the independences are constructed.

Example F.2. Consider a representation of the dual graph $\overline{\mathcal{G}}$ in Figure 17. The vertex i has neighbours partitioned into A^1, A^2, A^3 which have neighbours $N^1 \subseteq N^2 \subseteq N^3$ respectively.

To construct a proper decomposition of $u_{\mathcal{G}}$, we first look at the following conditional independences:

$$i \perp\!\!\!\perp \widetilde{A} \,|\, (\cap_{w \in \widetilde{A}} \mathrm{nb}(w)) \setminus \{i\}, \tag{10}$$

for every $\widetilde{A} \subseteq A = \mathrm{nb}(i)$. These independences are definitely implied by the graph as none of the conditioning variables are siblings of $\widetilde{A}$, so no paths are open. Moreover, one can see that each disconnected set containing i is associated with one of the independences. However, the independences also overlap. Proposition F.1 suggests that we do not need all of them, but just a subset of independences (that may require further modification), such that:
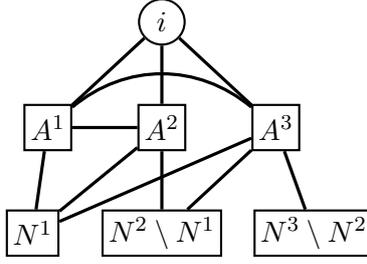
Figure 17: Example for Theorem F.5

(i) the associated constrained sets do not overlap; and

(ii) the independences associated with any disconnected set can be deduced from them using the semi-graphoid axioms.

Let's consider the following independences from (10):

$$i \perp\!\!\!\perp A^3, A^2, A^1 \mid N^1 \tag{11}$$

$$i \perp\!\!\!\perp A^3, A^2 \mid N^2, A^1 \tag{12}$$

$$i \perp\!\!\!\perp A^3, A^1 \mid N^1 \tag{13}$$

$$i \perp\!\!\!\perp A^3 \mid N^3, A^1 \tag{14}$$

$$i \perp\!\!\!\perp A^2 \mid N^2, A^1 \tag{15}$$

$$i \perp\!\!\!\perp A^1 \mid N^1, A^2, A^3. \tag{16}$$

(13), (15) and (16) are each implied by (11), (12) and (11) respectively, so we can ignore them. Then looking at (11), (12) and (14) we see that there are some overlaps. The set $\{i\} \cup A^3 \cup A^1$ is associated with both (11) and (14), but the latter cannot be further simplified so we marginalize A^3 in (11).

The set $\{i\} \cup A^3 \cup A^2 \cup A^1$ only appears in (12) so we want to keep that, but $\{i\} \cup A^3 \cup A^1$ is associated with both (12) and (14). Move A^3 into the conditioning set for (12) to resolve this. Also $\{i\} \cup A^2 \cup A^1$ is associated with both (11) and (12), so we marginalize A^2 in (11). This gives the following:

$$
\begin{aligned}
(11) &\rightarrow & i \perp\!\!\!\perp A^1 \mid N^1 \\
(12) &\rightarrow & i \perp\!\!\!\perp A^2 \mid A^3, A^1, N^2 \\
(14) &\rightarrow & \text{keep (14)}.
\end{aligned}
$$

One can check that any independence involving i can be deduced from the above three independences, together with other independences that do not involve i from the induction hypothesis. In addition, these cover all the disconnected sets containing i.

Proof of Theorem F.5. The main idea of this proof is to construct a list of independences $\mathbb{L}$, where the sum of semi-elementary imsets corresponding to the independences in $\mathbb{L}$ is $u_{\mathcal{G}}$, then using the similar ideas in the proof of Theorem 3.8, we just need to show that $\mathbb{L}$ holds in $\mathcal{G}$ so $I_{u_{\mathcal{G}}} \subseteq I_{\mathcal{G}}$, and $\mathbb{L}$ implies the global Markov property of $\mathcal{G}$ so $I_{\mathcal{G}} \subseteq I_{u_{\mathcal{G}}}$ or the disconnected Markov property from Drton and Richardson (2008).

We will proceed by induction on the ordering of vertices. For a vertex v, suppose its neighbours in $\overline{\mathcal{G}}$ are partitioned into $A^j, 1 \leq j \leq m$, each of which have neighbours N^j (outside of $\{v\} \cup \mathrm{nb}_{\overline{\mathcal{G}}}(v)$) such that $N^1 \subset N^2 \subset \cdots \subset N^m$.

For each A^j, it will be a root for exactly one collection of consecutive sets $\hat{A}^j = A^{[j_l, j_h]}$, and we consider the following list of independences $\mathbb{L}$:

$$v \perp\!\!\!\perp A^j \mid H^j, L^j, N^j, T^j, \qquad j \in 1, \ldots, m.$$

We will prove that for each j, the following independence holds for each j from the above list of independences:

$$v \perp\!\!\!\perp \hat{A}^j \mid N^{j_l}, T^j.$$

We will proceed by two inductions, an outer induction on the number of vertices in $\mathcal{G}$, and an inner induction on the lengths of $\hat{A}^j$. The base case for the outer induction is trivial, since a graph with one vertex has no independences. For the inner induction, the base case is $|\hat{A}^j| = 1$, so $\hat{A}^j = \{A^j\}$, $j_l = j$ and $H^j = L^j = \emptyset$, hence it is in the given list of independences.

For the inner induction step, suppose $\hat{A}^j = H^j \cup L^j \cup A^j$, where H^j and L^j are rooted at k and t (non-empty), respectively. Then by the induction hypothesis, we have:

$$v \perp\!\!\!\perp H^j \mid N^{k_l}, T^k \qquad\qquad v \perp\!\!\!\perp L^j \mid N^{t_l}, T^t,$$

but by Lemma F.4, this is equivalent to:

$$v \perp\!\!\!\perp H^j \mid N^{j+1}, L^j, T^j \tag{17}$$

$$v \perp\!\!\!\perp L^j \mid N^{j_l}, T^j. \tag{18}$$

Since v is the last vertex, any independence not involving v will hold by the induction hypothesis, hence we have:

$$H^j \perp\!\!\!\perp L^j, N^{j+1} \mid T^j \tag{19}$$

$$H^j, A^j \perp\!\!\!\perp L^j, N^j \mid T^j. \tag{20}$$

Combining (19) and (17), we have:

$$H^j \perp\!\!\!\perp \{v\}, L^j, N^{j+1} \mid T^j. \tag{21}$$

Then marginalizing N^{j+1} to N^j and moving $L^j \cup N^j$ to conditioning variables, we have:

$$v \perp\!\!\!\perp H^j \mid L^j, N^j, T^j. \tag{22}$$

Now from the given list of independences we have:

$$v \perp\!\!\!\perp A^j \mid H^j, L^j, N^j, T^j.$$

Putting this with (22) we obtain:

$$v \perp\!\!\!\perp H^j, A^j \mid L^j, N^j, T^j. \tag{23}$$

Using (20), (23) then changes to:

$$\{v\}, L^j, N^j \perp\!\!\!\perp H^j, A^j \mid T^j. \tag{24}$$

marginalizing N^{j_l+1} to N^j and moving $L^j \cup N^{j_l}$ to the conditioning variables, we have:

$$v \perp\!\!\!\perp H^j, A^j \mid L^j, N^{j_l}, T^j. \tag{25}$$

Finally, combining (25) and (18), we obtained the required independence:

$$v \perp\!\!\!\perp H^j, A^j, L^j \mid N^{j_l}, T^j. \tag{26}$$

Next we show that for any disconnected sets involving v, we can deduce the associated independence, that is, the global Markov property. We use the independence (24) that appears in the previous deduction and combine it with an independence from the induction hypothesis, $T^j \perp\!\!\!\perp H^j, A^j$, to get:

$$\{v\}, L^j, T^j, N^j \perp\!\!\!\perp H^j, A^j. \tag{27}$$

Based on this, we prove by induction that for every j, we have:

$$\{v\}, L^j, T^j, N^j \perp\!\!\!\perp A^{[j,m]}. \tag{28}$$

Before the induction, one should notice that for any $i > j$, $L^i \cup T^i \supseteq L^j \cup T^j$.

An order can be given to blocks based on the collection of blocks in which they are rooted. If H^j, L^j are rooted at A^k, A^t respectively, then we say that j is precedes both k and t, and apply the induction on this order. The base case is for the root of $A^{[1,m]}$, say A^j. In this case, $H^j \cup A^j = A^{[j,m]}$, and it is true.

Suppose for a block A^j, any other block A^i such that i precedes j in the above order so we have $\{v\} \cup L^i \cup T^i \cup N^i \perp\!\!\!\perp A^{[i,m]}$. If $H^j \cup A^j \neq A^{[j,m]}$, then A^{j_h+1} must be a block that also precedes A^j in the order, and so we also have $\{v\} \cup L^{j_h+1} \cup T^{j_h+1} \cup N^{j_h+1} \perp\!\!\!\perp A^{[j_h+1,m]}$ and (more importantly) $L^{j_h+1} \cup T^{j_h+1} \supseteq A^{[j,j_h]} \cup L^j \cup T^j$ (it is possible these sets are equal). now put $A^{[j,j_h]} = H^j \cup A^j$ into the conditioning variables and marginalize N^{j_h+1} to N^{j-1} and other irrelevant variables, to obtain:

$$\{v\} \cup L^j \cup T^j \cup N^j \perp\!\!\!\perp A^{[j_h,m]} \mid H^j \cup A^j. \tag{29}$$

Then combining this with (24), the result holds for all j.

Now we prove that for every disconnected set C involving v, the associated independence $\{v\} \cup D \perp\!\!\!\perp A$ is true, where $\{v\} \cup D$ is the district for $\{v\}$ in the subgraph induced by C. We only need to consider vertices in $\{v\} \cup A^{[1,m]} \cup N^m \cup \cdots \cup N^1$, as other vertices will definitely link v and any of its neighbours. This A must be a subset of $A^{[1,m]}$. Suppose A^j is the least block that have non-empty intersection with A, so $A \subseteq A^{[j,m]}$.

Consider the vertices in D. Firstly notice that it cannot contain any vertices in N^i or A^i where $i > j$. The first one is clear, suppose the second is not true, consider the bidirected path from A^i to v, the last vertex before v is a sibling of v and it must belong to some of the N^k. Now if $k > j$, then as A^j is connected to N^k, this contradicts to the assumption of j. so $k < j$ but then as A^i is a neighbour of N^k in the dual graph, they are disconnected, so there must be other blocks lower than A^j involved. however as we have $L^i \cup T^i \supseteq L^j \cup T^j$, this means that any block connect to A^i must also connect to A^j in the bidirected graphs and then A^j lies in the same districts as v, contradiction. Thus D can only contain vertices in $L^j \cup T^j \cup N^j$, and we have the required independence.

We are left to prove that the list of independence $v \perp\!\!\!\perp A^j \mid H^j \cup L^j \cup T^j \cup N^j$ are non-overlapping and associated with every disconnected set. For the disconnected set M, suppose $D \cup \{v\} = \mathrm{dis}_M(i)$ and we have this independence $\{v\} \cup D \perp\!\!\!\perp A$ with A^j the least block. Then consider the root of $A^m, \ldots, A^j$ (by Lemma F.2), say A^i, and also consider $\hat{A}^j$

then it is clear that the sets associates with the independence $v \perp\!\!\!\perp A^i \mid H^i \cup L^i \cup T^i \cup N^i$ contains this disconnected set.

To show there is no overlap, assume that $i < j$ (in the numerical sense). Then A^i will appear in the conditioning set for j only if it is in $L^j \cup T^j$, which by definition implies that A^j and A^i are fully connected. However A^j will appear in the conditioning set for A^i only if $A^j \subseteq H^i$, which implies that they are completely disconnected. Since at most one of these conditions can be true, there is no overlap in the sets generated by these independences. $\qquad\square$