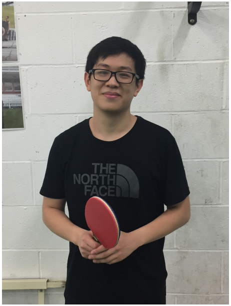
Towards Standard Imsets for Maximal Ancestral Graphs

Robin Evans, University of Oxford

Miniworkshop on Graphical Models and Causality (with beer)
TU Munich
28th September 2023

Collaborators



Zhongyi Hu, University of Oxford

Outline

1. Imsets

2. DAG Models

3. MAG Models

4. Power DAGs

5. Model Search

Imsets

Imsets were introduced by Studený (1995), as a method for representing arbitrary conditional independence models.

Let $\mathcal{P}(V)$ be the **power-set** of a finite set V.

Definition

An **imset** is an **i**nteger-valued **m**ulti**set**, or in other words a function

$$u : \mathcal{P}(V) \to \mathbb{Z}.$$

Since they are often sparse, we tend to represent them with combinations of identity functions:

$$\delta_A(X) = \begin{cases} 1 & \text{if } X = A, \\ 0 & \text{otherwise.} \end{cases}$$

Conditional Independence Models

Definition

We identify a **semi-elementary imset** with a triple (A, B, C) where

$$u_{\langle A,B|C\rangle} = \delta_C - \delta_{A\cup C} - \delta_{B\cup C} + \delta_{A\cup B\cup C}.$$

$u_{\langle A,B|C\rangle}$ represents the conditional independence $X_A \perp\!\!\!\perp X_B \mid X_C$.

Notice this conditional independence is equivalent to:

$$p(x_{ABC}) \cdot p(x_C) = p(x_{AC}) \cdot p(x_{BC})$$
$$\log p(x_C) - \log p(x_{AC}) - \log p(x_{BC}) + \log p(x_{ABC}) = 0.$$

Now we can see the analogy to the log-factorization.

Indeed, one can **test** a conditional independence by using the **entropy operator** $\mathrm{H}_p : \mathcal{P}(V) \to \mathbb{R}$, and we have that $X_A \perp\!\!\!\perp X_B \mid X_C$ if and only if

$$\langle \mathrm{H}_P, u_{\langle A,B|C\rangle}\rangle = \mathrm{H}(p(x_C)) - \mathrm{H}(p(x_{AC})) - \mathrm{H}(p(x_{BC})) + \mathrm{H}(p(x_{ABC})) = 0.$$

Structural Imsets

Definition

An imset u is said to be **structural** if there exists some natural number k such that we can write

$$k \cdot u = \sum_{v \in \mathcal{I}(V)} k_v \cdot v, \qquad k_v \in \mathbb{N} \cup \{0\},$$

where $\mathcal{I}(V)$ is the collection of (semi-)elementary imsets over the variables in the set V.

Structural imsets can be said to represent a model.

Models

Definition

Given an independence $X_A \perp\!\!\!\perp X_B \mid X_C$, we say that it is **represented in a structural imset** u over V (and write $A \perp\!\!\!\perp B \mid C\,[u]$) if there exists $k \in \mathbb{N}$ such that

$$k \cdot u - u_{\langle A,B|C \rangle}$$

is also structural.

Can be tested with an integer linear program (Bouckaert et al., 2010).

Imsets are useful because they can be used to **score** models consistently, and in particular can select the optimal **directed acyclic graph** model.

Example

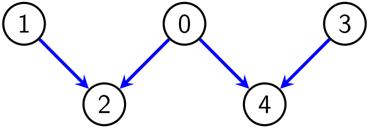
Consider the following imset:

$$\begin{aligned} u &= \langle +1, -1, +1, -1, -1, +1, \quad 0, \quad 0, \\ &\qquad 0, \quad 0, -1, +1, \quad 0, \quad 0, \quad 0, \quad 0\rangle \\ &= \langle +1, -1, \quad 0, \quad 0, -1, +1, \quad 0, \quad 0, \\ &\qquad 0, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0\rangle \\ &\quad + \langle \quad 0, \quad 0, +1, -1, \quad 0, \quad 0, \quad 0, \quad 0, \\ &\qquad\quad 0, \quad 0, -1, +1, \quad 0, \quad 0, \quad 0, \quad 0\rangle \\ &= u_{\langle 1,3\rangle} + u_{\langle 1,4|2\rangle}. \end{aligned}$$

Hence u is a structural imset, and represents $X_1 \perp\!\!\!\perp X_3$ and $X_1 \perp\!\!\!\perp X_4 \mid X_2$.

DAG Models

Directed acyclic graphs (DAGs) can represent comparatively simple independence models.



We can use a **local Markov property** to completely define the model.

- pick a topological order;
- then each variable is conditionally independent of its predecessors in the ordering given its parents;

$$X_i \perp\!\!\!\perp X_{\mathrm{pre}(i)\setminus\mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)}, \qquad \forall i \in V.$$

Imsets for DAG Models

Correspondingly, we can define the **standard imset** for a DAG $\mathcal{G}$ as:

$$\begin{aligned} u_{\mathcal{G}} &:= \sum_{i \in V} u_{\langle i, \mathrm{pre}(i) \mid \mathrm{pa}(i) \rangle} \\ &= \delta_V - \delta_\emptyset + \sum_{i \in V} \left(\delta_{\mathrm{pa}(i)} - \delta_{\{i\} \cup \mathrm{pa}(i)} \right). \end{aligned}$$

This has several nice properties:

- it is clearly a structural imset;
- P is Markov with respect to $\mathcal{G}$ if and only if $\langle \mathrm{I}_P, u_{\mathcal{G}} \rangle = 0$;
- $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if $u_{\mathcal{G}} = u_{\mathcal{G}'}$;
- it is sparse (at most $2|V|$ terms).

Characteristic Imsets for DAG Models

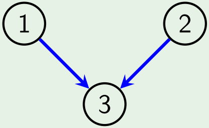
There is a bijective (Möbius) transformation we can make to obtain the **characteristic imset** (Studený et al., 2010) for a DAG:

$$c_{\mathcal{G}}(A) = 1 - \sum_{B \supseteq A} u_{\mathcal{G}}(B).$$

One can then show that

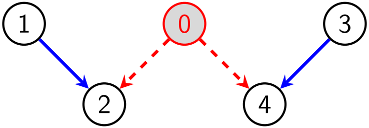
$$c_{\mathcal{G}}(A) = \begin{cases} 1 & \text{if } \exists v : \{v\} \subseteq A \subseteq \{v\} \cup \mathrm{pa}_{\mathcal{G}}(v) \\ 0 & \text{otherwise.} \end{cases}$$

Example. Consider the graph on the right. Then the non-zero sets are: $\emptyset$, $\{1\}$, $\{2\}$, $\{3\}$, $\{1,3\}$, $\{2,3\}$, $\{1,2,3\}$.



MAG Models

A (directed) **maximal ancestral graph** (MAG) model is just a collection of independences that can be represented by a DAG with hidden variables. (Richardson and Spirtes, 2002)



This MAG implies the independences

$$X_1 \perp\!\!\!\perp X_3, X_4 \qquad\qquad X_3 \perp\!\!\!\perp X_2 \mid X_1,$$

which cannot be faithfully represented by any DAG.

Markov Equivalence

In Hu and Evans (2020), we gave a criterion for two MAGs to be Markov equivalent based on collections of subsets.

Parametrizing Sets

The **parametrizing sets** for a MAG $\mathcal{G}$ are

$$\mathcal{S}(\mathcal{G}) = \{H \cup A : H \in \mathcal{H}(\mathcal{G}),\, A \subseteq \mathrm{tail}_{\mathcal{G}}(H)\},$$

where $\mathcal{H}(\mathcal{G})$ is the collection of **heads** in $\mathcal{G}$.

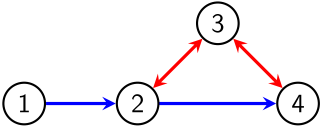
Given a vertex v in a head H, if we condition on $X_{H \setminus \{v\}}$, then the distribution cannot be m-separated from any $t \in \mathrm{tail}_{\mathcal{G}}(H)$.

As an analogy, for DAGs heads = vertices and tails = parent sets.

Parametrizing Sets Example

Consider this MAG, which implies

$$X_3 \perp\!\!\!\perp X_1 \quad \text{and} \quad X_4 \perp\!\!\!\perp X_1 \mid X_2 :$$



head	tail	parametrizing sets
$\{1\}$	$\emptyset$	$\{1\}$
$\{2\}$	$\{1\}$	$\{2\}, \{1,2\}$
$\{3\}$	$\emptyset$	$\{3\}$
$\{2,3\}$	$\{1\}$	$\{2,3\}, \{1,2,3\}$
$\{4\}$	$\{2\}$	$\{4\}, \{2,4\}$
$\{3,4\}$	$\{1,2\}$	$\{3,4\}, \{1,3,4\},$ $\{2,3,4\}, \{1,2,3,4\}$

Parametrizing set is missing only subsets $\{1,3\}$, $\{1,4\}$ and $\{1,2,4\}$.

Markov Equivalence Class and Characteristic Imsets

The parametrizing sets also give a representation of the Markov equivalence class of a MAG.

Theorem (Hu and Evans, 2020)

Two MAGs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if $\mathcal{S}(\mathcal{G}) = \mathcal{S}(\mathcal{G}')$.

Now note that the **characteristic imset** for a DAG takes the same form:

$$\begin{aligned} \mathcal{S}(\mathcal{G}) &= \{\{v\} \cup A : A \subseteq \mathrm{pa}_{\mathcal{G}}(v)\} \\ &= \{A : c_{\mathcal{G}}(A) = 1\}. \end{aligned}$$

So let's try using the parametrizing set to build the characteristic imset for a MAG!

Definition

Define the **characteristic imset** for a MAG $\mathcal{G}$ as

$$c_{\mathcal{G}}(A) = \begin{cases} 1 & \text{if } A \in \mathcal{S}(\mathcal{G}) \\ 0 & \text{otherwise.} \end{cases}$$

Retrofit for MAGs

Then define the **'standard' imset** as the inverse transformation of this.

$$u_{\mathcal{G}}(A) = \sum_{B \supseteq A} (-1)^{|B \setminus A|}(1 - c_{\mathcal{G}}(B)).$$

Proposition

Given a MAG $\mathcal{G}$, the 'standard' imset is the same as:

$$u_{\mathcal{G}} = \delta_V - \delta_\emptyset - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup T},$$

where $T = \mathrm{tail}_{\mathcal{G}}(H)$.

Note that this is consistent with the definition for DAGs.

Defining the Model

We used ILPs to check which 'standard' imsets define the model.

There are three cases, based on whether the 'standard' imset $u_{\mathcal{G}}$:

(i) does define the **same model** as $\mathcal{G}$;

(ii) defines a model with a (strict) **subset** of the independence restrictions of $\mathcal{G}$;

(iii) is **not structural** (so does not define any model).

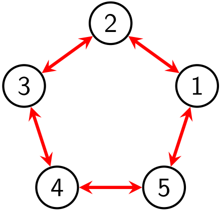
For small graphs, we find that they *usually* fall into category (i).

For all MAGs with 5 or 6 nodes, and 7 nodes and ≤ 13 or ≥ 18 edges:

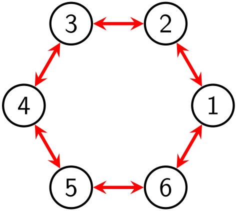
n	equiv. classes	(i)	(ii)	(iii)
5	285	284	1	0
6	13,303	13,248	54	1
7^*	1,161,461	1,146,501	14,562	8

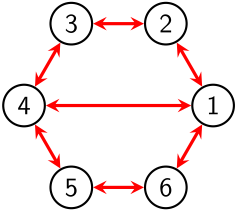
Defining the Model

For $n = 5$ 'standard' imsets **all** define the model, **except** for the bidirected 5-cycle.



The bidirected 6-cycle is not even **structural**.





'Simple' MAGs

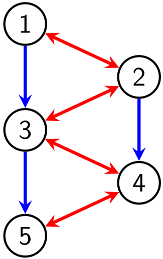
Definition

We say that a MAG is **simple** if its maximal head size is at most two.

$|V|$	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.

Example.



Proposition

For every simple MAG $\mathcal{G}$, the standard imset **does** define the model implied by the graph. In addition, it contains at most $2(|V| + |E|)$ terms.

Local Markov Property for MAGs

The **(ordered) local Markov property** for MAGs is more complicated than that for DAGs.

We consider every **ancestral set** (closed under taking parents) A, and the maximal vertex in that set v.

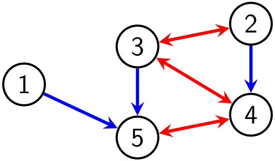
Then P satisfies the ordered local Markov property w.r.t. $\mathcal{G}$ if

$$X_v \perp\!\!\!\perp X_{A \setminus (H \cup T)} \mid X_{(H \cup T) \setminus \{v\}},$$

where H is the 'maximal' head in A, and T its tail.
That is, the set of vertices joined to v by paths of colliders within A.

Clearly we can restrict to 'maximal' sets A for each head.
This is the **reduced** OLMP for MAGs (Richardson, 2003).

Reduced Ordered Local Markov Property



Use the numerical order, which is topological, and consider 4.
For the reduced OLMP, we have:

$$X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3 \qquad\qquad X_4 \perp\!\!\!\perp X_1 \mid X_2.$$

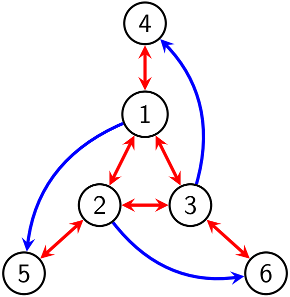
But! Notice that we already knew that $X_3 \perp\!\!\!\perp X_1 \mid X_2$, so:

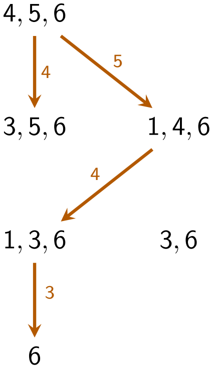
$$\begin{aligned} X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3 \quad \wedge \quad X_3 \perp\!\!\!\perp X_1 \mid X_2 \quad &\Longrightarrow \quad X_3, X_4 \perp\!\!\!\perp X_1 \mid X_2 \\ &\Longrightarrow \quad X_4 \perp\!\!\!\perp X_1 \mid X_2. \end{aligned}$$

So the second independence is **redundant**.

Power DAGs

Draw a **power DAG** whose vertices represent heads. Draw edge from H to H' if we can marginalize a **non-maximal** vertex in H to obtain H'.





$4, 5, 6$

4

5

$3, 5, 6$

$1, 4, 6$

4

$1, 3, 6$

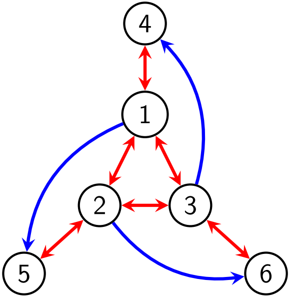
$3, 6$

3

6

Refined OLMP

We can automate this using a **refined** ordered local Markov property.



$$\begin{tikzcd}
4,5,6 \arrow[d, "4"] \arrow[dr, "5"] & \\
3,5,6 \arrow[d, "5"] \arrow[dd, bend right, "3"'] & 1,4,6 \arrow[dl, "4"'] \arrow[d, "1"] \\
1,3,6 \arrow[r, "1"] \arrow[d, "3"] & 3,6 \arrow[dl, "3"] \\
6 &
\end{tikzcd}$$

We have a separate component for each maximal vertex.

In fact, we only need one edge into each vertex (the one from the 'maximal' head).

Refined OLMP

In the case of the previous graph, our refined OLMP gives:

$$X_2 \perp\!\!\!\perp X_4 \mid X_3 \qquad X_5 \perp\!\!\!\perp X_3, X_4 \mid X_1$$

$$X_6 \perp\!\!\!\perp X_1 \mid X_2, X_3, X_4, X_5 \qquad X_6 \perp\!\!\!\perp X_4 \mid X_2, X_3$$

$$X_6 \perp\!\!\!\perp X_2.$$

Notice that the 'standard' imset **will** define the model in this case.

The reduced OLMP gives

$$X_2 \perp\!\!\!\perp X_4 \mid X_3 \qquad X_5 \perp\!\!\!\perp X_3, X_4 \mid X_1$$

$$X_6 \perp\!\!\!\perp X_1 \mid X_2, X_3, X_4, X_5 \qquad X_6 \perp\!\!\!\perp X_1, X_4 \mid X_2, X_3$$

$$X_6 \perp\!\!\!\perp X_1, X_2, X_4.$$

Model Scoring

Usual consistent score for model scoring is the BIC. This requires us to find the maximum likelihood for each model we score.

We have a proposal for a scoring models (Andrews, 2022):

$$h(\mathcal{G}) := 2n\langle \mathrm{H}_P, u_{\mathcal{G}} \rangle - k \log n,$$

where n is the number of samples, k is the number of parameters, and I_P is the **interaction information operator** (see appendix).

If $u_{\mathcal{G}}$ defines the model, we have

$$n\langle \mathrm{H}_P, u_{\mathcal{G}} \rangle \approx \ell_{\mathcal{G}}(P; \boldsymbol{X}_V),$$

so our score approximates the BIC.

Hence the score is consistent over this set of MAGs
(i.e. the highest score is given asymptotically to the true model).

We restrict our search to simple MAGs.

Simulation

We randomly simulate 100 ADMGs with

- $n \in \{5, 10, 15, 20\}$ nodes;
- average node degree 3;
- edges are directed with probability 0.8.

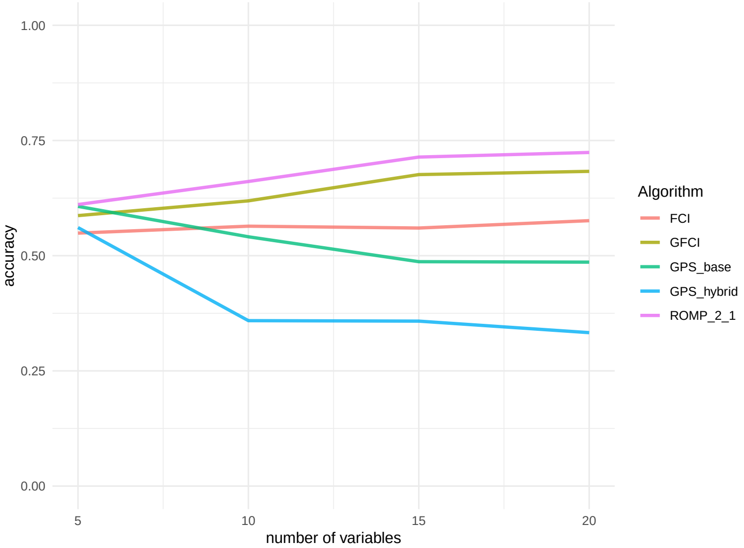
These are then projected to a Markov equivalent MAG, and then a random SEM is generated.

Edge strengths are drawn uniformly from $\pm[0.1, 1]$.

We compare our algorithm to one of Claassen and Bucur (2022), **greedy PAG search** (GPS), which uses the exact BIC.

We also compare to the constraint-based FCI and GFCI algorithms.

Results (Edge mark accuracy)



Summary

- Imsets can be used to define arbitrary conditional independence models;
- they have particularly nice properties when applied to DAGs.

- *Some* of those properties are replicated in MAGs, but (unfortunately) not all of them!
- Problem is that (for some graphs) it is not possible to describe the model without using conditional independences that repeat sets.

- Imsets that **do** represent the models can be used to give a new consistent score, which is easier to compute than the BIC.
- We have also developed a greedy algorithm for learning MAGs.

Open Problems

- How do we see if a graph has a perfectly Markovian 'standard' imset?
- Can we develop a Markov property that does not include overlapping sets?

Thank you!

References I

Andrews—*Inducing Sets: A New Perspective for Ancestral Graph Markov Models*, PhD Thesis, U. Pitt, 2022.

Andrews et al.—The m-connecting imset and factorization for ADMG models, *arXiv:2207.08963*, 2022.

Bouckaert et al.—Efficient algorithms for conditional independence inference, *JMLR*, 2010.

Claassen and Bucur—Greedy equivalence search in the presence of latent confounders, *UAI*, 2022.

Hemmecke et al.—Characteristic imsets for learning Bayesian network structure. *IJAR*, 2012.

Hu and Evans— Faster algorithms for Markov equivalence, *UAI-20*, 2020.

Hu and Evans—Towards standard imsets for maximal ancestral graphs. *Bernoulli (accepted), arXiv:2208.10436*, 2023.

References II

Richardson—Markov properties for acyclic directed mixed graphs, *Scandinavian Journal of Statistics*, 2003.

Studený—Description of structures of stochastic conditional independence by means of faces and imsets, 1st part, *Int. J. General Systems*, 1995.

Studený—*Probabilistic Conditional Independence Structures*, Springer, 2005.

Studený et al.—Characteristic imset: a simple algebraic representative of a Bayesian network structure. *European PGMs*, 2010.

Multi-information

We define the **entropy** $H(P)$ of a distribution P as

$$\mathrm{H}(P) = \int P(x) \log P(x) \, d\mu(x).$$

The **relative entropy** (or KL-divergence) of P with respect to Q is

$$\mathrm{H}(P \mid Q) = \mathrm{H}(P) - \int P(x) \log Q(x) \, d\mu(x).$$

The **multi-information** over a set of variables X_S is the relative entropy between $P(X_S)$ and $\prod_{i \in S} P(X_i)$. That is:

$$\mathrm{m}(P_S) = \mathrm{H}(P_S) - \sum_{s \in S} \mathrm{H}(P_s).$$

The **interaction information** for a set S is

$$\mathrm{I}(P_S) := \sum_{T \subseteq S} (-1)^{|S \setminus T|} \mathrm{H}(P_T).$$

The **multi-information** and **interaction information functions for** P are operators

$$\mathrm{m}_P : S \mapsto \mathrm{m}(P_S) \qquad \text{and} \qquad \mathrm{I}_P : S \mapsto \mathrm{I}(P_S).$$

'Standard' imset

Theorem

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

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

Simplest Model

How do we know that $u_{\mathcal{G}}$ is the ‘standard’ imset?

If $u_{\mathcal{G}}$ does define the model, it is the simplest possible imset that does so.

Why? Well, we know that $c_{\mathcal{G}} \leq 1$ and that if we add any additional semi-elementary imset to get $u'_{\mathcal{G}} = u_{\mathcal{G}} + u_{\langle a,b|C\rangle}$, then:

$$c'_{\mathcal{G}}(S) = c_{\mathcal{G}}(S) - \sum_{C' \subseteq C} \mathbb{1}_{\{S=\{a,b\}\cup C'\}}.$$

Hence $c'_{\mathcal{G}} \leq 1$ but any additional independences will lead to sets with a coefficient of -1.

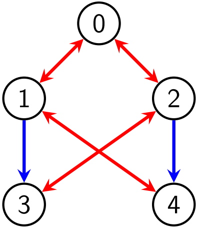
Heads and Tails

Let $\mathcal{G}$ be an **acyclic directed mixed graph** (ADMG). This includes all MAGs as a special case.

A **head** H is a subset of vertices such that the vertices are all in the same district of $\mathcal{G}_{\mathrm{an}(H)}$, and all **barren** (have no children) within that set.

The associated **tail** is $\mathrm{tail}_{\mathcal{G}}(H) = \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_{\mathrm{an}(H)}(H))$.

Example.



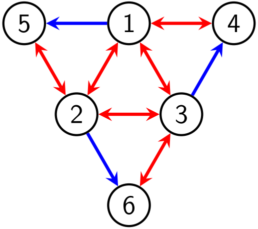
H	T
$\{0\}, \{1\}, \{2\},$ $\{0,1\}, \{0,2\}, \{0,1,2\}$	$\emptyset$
$\{2\}, \{2,3\}, \{0,2,3\}$	$\{1\}$
$\{4\}, \{1,4\}, \{0,1,4\}$	$\{2\}$
$\{0,3,4\}$	$\{1,2\}$

Missing sets correspond to constraints:

$X_1 \perp\!\!\!\perp X_2$; $\quad X_0 \perp\!\!\!\perp X_3 \mid X_1$; $\quad X_0 \perp\!\!\!\perp X_4 \mid X_2$; $\quad X_3 \perp\!\!\!\perp X_4 \mid X_1, X_2$.

Heads and Tails Example

This MAG has the head $\{4, 5, 6\}$, but no subset of size two is a head!



H	T
$\{1\}, \{2\}, \{3\}, \{1, 2\},$ $\{1, 3\}, \{2, 3\}, \{1, 2, 3\}$	$\emptyset$
$\{4\}, \{1, 4\}, \{1, 2, 4\}$	$\{3\}$
$\{5\}, \{2, 5\}, \{2, 3, 5\}$	$\{1\}$
$\{6\}, \{3, 6\}, \{1, 3, 6\}$	$\{2\}$
$\{2, 4, 5\}$	$\{1, 3\}$
$\{1, 4, 6\}$	$\{2, 3\}$
$\{3, 5, 6\}$	$\{1, 2\}$
$\{4, 5, 6\}$	$\{1, 2, 3\}$

Moves

Parameters: t number of colliders/non-colliders to consider.

Initialize: $\mathcal{P}$ is empty PAG.

Start from a PAG $\mathcal{P}$.

For each missing edge i, j in the skeleton of $\mathcal{P}$, add in $i \ast\!\!-\!\!\ast j$:

- Suppose $i \ast\!\!-\!\!\ast j \ast\!\!-\!\!\ast k$ is a (new) unshielded triple. Use invariant edge marks from $\mathcal{P}$ to reduce the search space.
- If a new discriminating path is created, then consider both possible orientations of the relevant edges.
- Finally, consider exchanging each collider for a non-collider, and vice versa.

If we find graph with lower score, record the score of new optimal PAG $\mathcal{P}'$, then go back to the start.

Otherwise, return $\mathcal{P}$.