Greedy Relaxations of the Sparsest Permutation Algorithm (Supplementary material)

Wai-Yin Lam[1] **Bryan Andrews**[1] **Joseph Ramsey**[1]

[1] Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

A BACKGROUND MATERIALS

A.1 GRAPHICAL DEFINITIONS

A *directed graph* $\mathcal{G}$ over a set of measured variables $\mathbf{V} = \{X_1, ..., X_m\}$ consists of m vertices $\mathbf{v} = \{1, ..., m\}$ where each vertex $i \in \mathbf{v}$ associates to a variable $X_i \in \mathbf{V}$, and each edge in $\mathcal{G}$ is directed with the form $j \to k$ and no vertex has a directed edge to itself. A *path* $\mathfrak{p}$ is a sequence of vertices $\langle \mathbf{i}_1, \mathbf{i}_2, ..., \mathbf{i}_k \rangle$ for some $k \geq 2$ where $\mathbf{i}_j \in \mathbf{v}$ for each $1 \leq j \leq k$, and $\mathbf{i}_j$ and $\mathbf{i}_{j+1}$ are connected by a directed edge (i.e., $\mathbf{i}_j \to \mathbf{i}_{j+1}$ or $\mathbf{i}_{j+1} \to \mathbf{i}_j$). A path $\mathfrak{p}$ is *directed* if $\mathbf{i}_j \to \mathbf{i}_{j+1}$ for each $1 \leq j < k$. A *directed acyclic graph* (DAG) is a directed graph where no vertex can have a directed path to itself.

Denote $\mathtt{E}(\mathcal{G})$ as the set of directed edges in $\mathcal{G}$. A pair of DAGs $\mathcal{G}_1, \mathcal{G}_2$ over the same set of variables $\mathbf{V}$ are equivalent if and only if $\mathtt{E}(\mathcal{G}_1) = \mathtt{E}(\mathcal{G}_2)$. Let $\mathtt{Pa}(j, \mathcal{G}) = \{k \in \mathbf{v} : (k \to j) \in \mathtt{E}(\mathcal{G})\}$ be the set of *parents* of j in $\mathcal{G}$, and $\mathtt{Ch}(j, \mathcal{G}) = \{k \in \mathbf{v} : (j \to k) \in \mathtt{E}(\mathcal{G})\}$ be the set of *children* of j in $\mathcal{G}$. $\mathtt{An}(j, \mathcal{G})$, the *ancestors* of j in $\mathcal{G}$, is defined by the transitive closure of $\mathtt{Pa}(j, \mathcal{G})$. Similarly, $\mathtt{De}(j, \mathcal{G})$, the *descendants* of j in $\mathcal{G}$, is defined by the transitive closure of $\mathtt{Ch}(j, \mathcal{G})$ and union with $\{j\}$ itself (i.e., j is its own descendant). Further let $\mathtt{Nd}(j, \mathcal{G}) = \mathbf{v} \setminus \mathtt{De}(j, \mathcal{G})$ be the set of j's *non-descendants*.

A pair of vertices $j, k \in \mathbf{v}$ are said to be *adjacent* in $\mathcal{G}$ if $(j \to k) \in \mathtt{E}(\mathcal{G})$ or $(k \to j) \in \mathtt{E}(\mathcal{G})$. For any triple of pairwise distinct vertices $i, j, k \in \mathbf{v}$, we say that (i, j, k) is *unshielded* if (i, j) and (j, k) are adjacent pairs in $\mathcal{G}$, but not (i, k). (i, j, k) forms a *triangle* if they are pairwise adjacent. If (i, j, k) is an unshielded triple or is a triangle, j is a *collider* (on the path $\langle i, j, k \rangle$) if $(i \to j), (k \to j) \in \mathtt{E}(\mathcal{G})$, and a *non-collider* otherwise. A path $\mathfrak{p}$ is a *trek* if it contains no collider.

For any $j, k \in \mathbf{v}$ and any $\mathbf{i} \subseteq \mathbf{v} \setminus \{j, k\}$, j and k are *d-connected* given $\mathbf{i}$ in $\mathcal{G}$ if there exists a path $\mathfrak{p}$ between j and k in $\mathcal{G}$ such that no non-collider on $\mathfrak{p}$ is in $\mathbf{i}$, and each collider l on $\mathfrak{p}$ or a l's descendant is in $\mathbf{i}$. j and k are *d-separated* given $\mathbf{i}$ in $\mathcal{G}$ if j and k are not d-connected given $\mathbf{i}$. For any disjoint subsets of vertices $\mathbf{j}, \mathbf{k}, \mathbf{i} \subseteq \mathbf{v}$, $\mathbf{j}$ and $\mathbf{k}$ are d-separated given $\mathbf{i}$ in $\mathcal{G}$ if j and k are d-separated by $\mathbf{i}$ in $\mathcal{G}$ for every $j \in \mathbf{j}$ and every $k \in \mathbf{k}$.

Given a model $(\mathcal{G}, \mathcal{P})$ over $\mathbf{V}$, $\mathcal{G}$ is said to be *local Markov* to $\mathcal{P}$ if $X_j \perp\!\!\!\perp_{\mathcal{P}} \mathbf{X}_{\mathtt{Nd}(j,\mathcal{G})} \setminus \mathbf{X}_{\mathtt{Pa}(j,\mathcal{G})} \,|\, \mathbf{X}_{\mathtt{Pa}(j,\mathcal{G})}$ for every $j \in \mathbf{v}$. It is a well-known fact that $\mathcal{G}$ is local Markov to $\mathcal{P}$ if and only if $\mathtt{I}(\mathcal{G}) \subseteq \mathtt{I}(\mathcal{P})$ (i.e., global Markov as defined by d-separation).

A.2 GRAPHOID AXIOMS

For any pairwise disjoint sets of variables $\mathbf{W}, \mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$,

$$\begin{aligned}
\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z} &\Rightarrow \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{Z} && (\textit{symmetry}) \\
\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \,|\, \mathbf{Z} &\Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \,|\, \mathbf{Z}) && (\textit{decomposition}) \\
\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \,|\, \mathbf{Z} &\Rightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z} \cup \mathbf{W} && (\textit{weak union}) \\
(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \,|\, \mathbf{Z} \cup \mathbf{Y}) &\Rightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \,|\, \mathbf{Z} && (\textit{contraction}) \\
(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z} \cup \mathbf{W}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \,|\, \mathbf{Z} \cup \mathbf{Y}) &\Rightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \,|\, \mathbf{Z} && (\textit{intersection}) \\
(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \,|\, \mathbf{Z}) &\Rightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \,|\, \mathbf{Z} && (\textit{composition})
\end{aligned}$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

A distribution $\mathcal{P}$ is a *semigraphoid* if $\mathtt{I}(\mathcal{P})$ is closed under *symmetry*, *decomposition*, *weak union*, and *contraction*. A semigraphoid $\mathcal{P}$ is a *graphoid* if $\mathtt{I}(\mathcal{P})$ is closed under *intersection*. A graphoid $\mathcal{P}$ is *compositional* if $\mathtt{I}(\mathcal{P})$ is closed under *composition*. See Chapter 2 of [Studený, 2005] for a more comprehensive study of graphoid axioms. In addition, applications of *symmetry* in our upcoming proofs will be done implicitly for the sake of simplicity.

Additionally, Spohn [1994] notes that the following property necessarily holds in the independence models induced by positive discrete probability distributions. For any pairwise disjoint sets of variables $\mathbf{W}, \mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$,

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{W} \cup \mathbf{Z}) \wedge (\mathbf{W} \perp\!\!\!\perp \mathbf{Z} \,|\, \mathbf{X} \cup \mathbf{Y}) \wedge (\mathbf{W} \perp\!\!\!\perp \mathbf{Z} \,|\, \mathbf{X}) \;\Rightarrow\; [(\mathbf{W} \perp\!\!\!\perp \mathbf{Z} \,|\, \mathbf{Y}) \Leftrightarrow (\mathbf{W} \perp\!\!\!\perp \mathbf{Z} \,|\, \varnothing)] \qquad (\textit{Spohn condition})$$

A.3 DAG INDUCED FROM A PERMUTATION

Definition A.1 *Given a semigraphoid $\mathcal{P}$ over $\mathbf{V}$, for every $X \in \mathbf{V}$, we say that $\mathbf{M} \subseteq \mathbf{V}$ is a Markov blanket of X relative to $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X\}$ if*

(i) $\mathbf{M} \subseteq \mathbf{Z}$*;*

(ii) $X \perp\!\!\!\perp_{\mathcal{P}} (\mathbf{Z} \setminus \mathbf{M}) \,|\, \mathbf{M}$.

Such a Markov blanket $\mathbf{M}$ is said to be a Markov boundary if it further satisfies the following condition:

(iii) there does not exist $\mathbf{M}' \subset \mathbf{M}$ s.t. $X \perp\!\!\!\perp_{\mathcal{P}} (\mathbf{Z} \setminus \mathbf{M}') \,|\, \mathbf{M}'$.

Lemma A.2 *[Verma and Pearl, 1988] Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, for every $X \in \mathbf{V}$ and every $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X\}$, there is a unique Markov boundary of X relative to $\mathbf{Z}$.*

In the following, we use $\mathtt{MB}_{\mathcal{P}}(X, \mathbf{Z})$ to refer to the unique Markov boundary of X relative to $\mathbf{Z}$. The subscript $\mathcal{P}$ will be suppressed if the underlying graphoid is clear from context.

Lemma A.3 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, for every $X \in \mathbf{V}$ and every $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X\}$, if $\mathbf{M}$ is a Markov blanket of X relative to $\mathbf{Z}$, then $\mathtt{MB}(X, \mathbf{Z}) \subseteq \mathbf{M}$.*

Proof. Immediate from **Definition A.1** and **Lemma A.2**. □

Next, we revisit the two methods of inducing a DAG from a permutation. Given a semigraphoid $\mathcal{P}$ over $\mathbf{V}$, each $\pi \in \Pi(\mathbf{v})$ induces a DAG satisfying the following condition:

$$X_j \in \mathbf{M} \Leftrightarrow (j \to k) \in \mathtt{E}(\mathcal{G}_\pi) \qquad \text{(VP)}$$

where $\mathbf{M}$ is a Markov boundary of X_k relative to $\mathbf{X}_{\mathtt{Pre}(k,\pi)}$. (VP) is the construction of a *boundary DAG* in [Verma and Pearl, 1988]. On the other hand, given a graphoid $\mathcal{P}$ over $\mathbf{V}$, each $\pi \in \Pi(\mathbf{v})$ induces a DAG satisfying the following condition:

$$j \in \mathtt{Pre}(k, \pi) \text{ and } X_j \not\perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\mathtt{Pre}(k,\pi)\setminus\{j\}} \Leftrightarrow (j \to k) \in \mathtt{E}(\mathcal{G}_\pi). \qquad \text{(RU)}$$

We want to show that the two DAG-inducing methods are equivalent when the underlying distribution is a graphoid.

Lemma A.4 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any $\pi \in \Pi(\mathbf{v})$. Let $\mathcal{G}_\pi$ be the DAG induced from π by (VP), and $\mathcal{H}_\pi$ be the DAG induced from π by (RU). Then $\mathcal{G}_\pi = \mathcal{H}_\pi$.*

Proof. We divide the proof into two directions: (VP) $\Rightarrow$ (RU), and (VP) $\Leftarrow$ (RU). Consider any $j, k \in \mathbf{v}$ where $\pi[j] < \pi[k]$ such that $j \in \mathtt{Pre}(k, \pi)$. Let $\mathbf{M}$ be the unique Markov boundary $\mathtt{MB}(X_k, \mathbf{X}_{\mathtt{Pre}(k,\pi)})$.

[$\Rightarrow$] Suppose that $(j \to k) \notin \mathtt{E}(\mathcal{G}_\pi)$. We have $X_j \notin \mathbf{M}$. By **Definition A.1** (ii), we then have,

$$X_k \perp\!\!\!\perp_{\mathcal{P}} ((\mathbf{X}_{\mathtt{Pre}(k,\pi)} \setminus \mathbf{M}) \setminus \{X_j\}) \cup \{X_j\} \,|\, \mathbf{M} \qquad \because X_k \perp\!\!\!\perp_{\mathcal{P}} \mathbf{X}_{\mathtt{Pre}(k,\pi)} \setminus \mathbf{M} \,|\, \mathbf{M} \qquad (1)$$

$$X_k \perp\!\!\!\perp_{\mathcal{P}} X_j \,|\, \mathbf{M} \cup ((\mathbf{X}_{\mathtt{Pre}(k,\pi)} \setminus \mathbf{M}) \setminus \{X_j\}) \qquad \because (1), \text{weak union} \qquad (2)$$

$$X_k \perp\!\!\!\perp_{\mathcal{P}} X_j \,|\, \mathbf{X}_{\mathtt{Pre}(k,\pi)\setminus\{j\}} \qquad \because (2) \qquad (3)$$

where the last formula amounts to $(j \to k) \notin \mathtt{E}(\mathcal{H}_\pi)$ by (RU).

[$\Leftarrow$] Suppose that $(j \to k) \notin \mathtt{E}(\mathcal{H}_\pi)$. We have $X_k \perp\!\!\!\perp_{\mathcal{P}} X_j \,|\, \mathbf{X}_{\mathtt{Pre}(k,\pi)\setminus\{j\}}$. Let $\mathbf{M}'$ be $\mathbf{X}_{\mathtt{Pre}(k,\pi)\setminus\{j\}}$. We have $X_k \perp\!\!\!\perp_{\mathcal{P}} (\mathbf{X}_{\mathtt{Pre}(k,\pi)} \setminus \mathbf{M}') \,|\, \mathbf{M}'$ such that $\mathbf{M}'$ is a Markov blanket of X_k relative to $\mathbf{X}_{\mathtt{Pre}(k,\pi)}$. By **Lemma A.3**, $X_j \notin \mathbf{M} \subseteq \mathbf{M}'$ and therefore $(j \to k) \notin \mathtt{E}(\mathcal{G}_\pi)$ by (VP). □

Theorem A.5 *[Pearl, 1988] Given a semigraphoid $\mathcal{P}$ over $\mathbf{V}$, $\mathcal{G}_\pi$ induced by π using (VP) is Markovian and SGS-minimal for any $\pi \in \Pi(\mathbf{v})$.*

Theorem 3.5 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, $\mathcal{G}_\pi$ induced by π using (RU) is Markovian and SGS-minimal for any $\pi \in \Pi(\mathbf{v})$.*

Proof. Immediate from **Lemma A.4** and **Theorem A.5**.[1] □

B CORRECTNESS RESULTS

First, we introduce some permutation-based notations to facilitate our coming proofs. In this section, we use $\mathcal{G}_\pi$ to denote the DAG induced by π from a graphoid $\mathcal{P}$ using (RU) unless specified otherwise.

Given a set of variables $\mathbf{V}$, consider any $\pi \in \Pi(\mathbf{v})$ and any pair $j, k \in \mathbf{v}$ where $\pi[j] < \pi[k]$. π can be written as $\langle \boldsymbol{\delta}_{<j}, j, \boldsymbol{\delta}_{j\sim k}, k, \boldsymbol{\delta}_{>k} \rangle$ such that $\boldsymbol{\delta}_{<j} = \langle \pi_i : 1 \leq i < \pi[j] \rangle$, $\boldsymbol{\delta}_{j\sim k} = \langle \pi_i : \pi[j] < i < \pi[k] \rangle$, and $\boldsymbol{\delta}_{>k} = \langle \pi_i : \pi[k] < i \leq |\pi| \rangle$. When $\boldsymbol{\delta}_{j\sim k} = \varnothing$, we say that j and k are π-*adjacent*. In that case, π can be written as $\langle \boldsymbol{\delta}_{<j}, j, k, \boldsymbol{\delta}_{>k} \rangle$ instead.

Definition B.1 *Given a set of variables $\mathbf{V}$, for any $\pi, \tau \in \Pi(\mathbf{v})$,*

(a) τ is said to be (j,k)-different from π for some $j, k \in \mathbf{v}$ if j and k are π-adjacent (i.e., $\pi = \langle \boldsymbol{\delta}_{<j}, j, k, \boldsymbol{\delta}_{>k} \rangle$) and $\tau = \langle \boldsymbol{\delta}_{<j}, k, j, \boldsymbol{\delta}_{>k} \rangle$;

(b) π and τ are said to be in adjacent transposition (AT) if they are (j,k)-different for some $j, k \in \mathbf{v}$.

Lemma B.2 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any $\mathcal{H} \in \mathtt{CMC}(\mathcal{P})$. If $\pi \in \Pi(\mathbf{v})$ is a causal order of $\mathcal{G}$, then $\mathtt{E}(\mathcal{G}_\pi) \subseteq \mathtt{E}(\mathcal{H})$. Also, $\mathcal{G}_\pi = \mathcal{H}$ if $\mathcal{H} \in \mathtt{SGS}(\mathcal{P})$.*

Proof. Consider any $k \in \mathbf{v}$ and $\mathtt{Nd}(k, \mathcal{H})$ (i.e., the set of k's non-descendants in $\mathcal{H}$). Since $\mathcal{H} \in \mathtt{CMC}(\mathcal{P})$, it follows that $X_k \perp\!\!\!\perp_{\mathcal{P}} \mathbf{X}_{\mathtt{Nd}(k,\mathcal{H})} \setminus \mathbf{X}_{\mathtt{Pa}(k,\mathcal{H})} \,|\, \mathbf{X}_{\mathtt{Pa}(k,\mathcal{H})}$. Also, we have $\mathtt{Pa}(k, \mathcal{H}) \subseteq \mathtt{Pre}(k, \pi) \subseteq \mathtt{Nd}(k, \mathcal{H})$ from π's being a causal order of $\mathcal{H}$. By *decomposition*, we have $X_k \perp\!\!\!\perp_{\mathcal{P}} \mathbf{X}_{\mathtt{Pre}(k,\pi)} \setminus \mathbf{X}_{\mathtt{Pa}(k,\mathcal{H})} \,|\, \mathbf{X}_{\mathtt{Pa}(k,\mathcal{H})}$ such that $\mathbf{X}_{\mathtt{Pa}(k,\mathcal{H})}$ is a Markov blanket of X_k relative to $\mathbf{X}_{\mathtt{Pre}(k,\pi)}$. By **Lemma A.3**, we have $\mathtt{MB}(X_k, \mathbf{X}_{\mathtt{Pre}(k,\pi)}) \subseteq \mathtt{Pa}(k, \mathcal{H})$. Consider $\mathcal{G}_\pi$ induced by (VP). The above entails that $\mathtt{E}(\mathcal{G}_\pi) \subseteq \mathtt{E}(\mathcal{H})$ since $\mathtt{Pa}(k, \mathcal{G}_\pi) = \mathtt{MB}(X_k, \mathbf{X}_{\mathtt{Pre}(k,\pi)}) \subseteq \mathtt{Pa}(k, \mathcal{H})$ for each $k \in \mathbf{v}$. Due to **Lemma A.4**, $\mathtt{E}(\mathcal{G}_\pi) \subseteq \mathtt{E}(\mathcal{H})$ still holds even if $\mathcal{G}_\pi$ is induced by (RU). Lastly, $\mathcal{G}_\pi = \mathcal{H}$ follows from **Definition 3.4** if $\mathcal{H} \in \mathtt{SGS}(\mathcal{P})$. □

Lemma B.3 *[Solus et al., 2021] Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any $\pi, \tau \in \Pi(\mathbf{v})$ where τ is (j,k)-different from π for some $j, k \in \mathbf{v}$. Then $\mathcal{G}_\pi = \mathcal{G}_\tau$ if and only if $X_j \perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\mathtt{Pre}(j,\pi)}$.*

Proof. Suppose that $X_j \not\perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\mathtt{Pre}(j,\pi)}$. By (RU), we have $(j \to k) \in \mathtt{E}(\mathcal{G}_\pi)$. Note that $(j \to k) \notin \mathtt{E}(\mathcal{G}_\tau)$ since $\tau[k] < \tau[j]$ and τ is a causal order of $\mathcal{G}_\tau$ by construction. Hence, $\mathcal{G}_\pi \neq \mathcal{G}_\tau$.

On the other hand, suppose that $X_j \perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\mathtt{Pre}(j,\pi)}$. Since τ is (j,k)-different from π, we have $\pi = \langle \boldsymbol{\delta}_{<j}, j, k, \boldsymbol{\delta}_{>k} \rangle$ and $\tau = \langle \boldsymbol{\delta}_{<j}, k, j, \boldsymbol{\delta}_{>k} \rangle$ according to **Definition B.1** (a). By (RU), we know that $(k \to j) \notin \mathtt{E}(\mathcal{G}_\tau)$. Hence, π is a causal order of $\mathcal{G}_\tau$. By **Theorem 3.5**, $\mathcal{G}_\tau \in \mathtt{SGS}(\mathcal{P})$. Therefore, it follows from **Lemma B.2** that $\mathcal{G}_\tau = \mathcal{G}_\pi$. □

Lemma B.4 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any $\pi \in \Pi(\mathbf{v})$. Suppose that $\mathcal{G}_\pi$ contains a covered edge $j \to k$ where $\pi = \langle \boldsymbol{\delta}_{<j}, j, \boldsymbol{\delta}_{j\sim k}, k, \boldsymbol{\delta}_{>k} \rangle$. If $\tau = \langle \boldsymbol{\delta}_{<j}, j, k, \boldsymbol{\delta}_{j\sim k}, \boldsymbol{\delta}_{>k} \rangle$, then $\mathcal{G}_\pi = \mathcal{G}_\tau$.*

Proof. Since $j \to k$ is a covered edge in $\mathcal{G}_\pi$, it follows that $(i \to k) \notin \mathtt{E}(\mathcal{G}_\pi)$ for each $i \in \boldsymbol{\delta}_{j\sim k}$, and thus $X_i \perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\mathtt{Pre}(i,\pi)}$ by (RU). Hence, $\mathcal{G}_\pi = \mathcal{G}_\tau$ can be obtained after $|\boldsymbol{\delta}_{j\sim k}|$ applications of **Lemma B.3**. □

[1] The two DAG-inducing methods were not differentiated in [Raskutti and Uhler, 2018]. Thus, we provide a proof of **Theorem 3.5**.

Theorem B.5 *[Zhang, 2013] Given a set of variables* $\mathbf{V}$*, for any* $\mathcal{G}, \mathcal{H} \in \texttt{DAG}(\mathbf{V})$*, if* $\texttt{E}(\mathcal{G}) \subseteq \texttt{E}(\mathcal{H})$*, then* $\texttt{I}(\mathcal{H}) \subseteq \texttt{I}(\mathcal{G})$*.*

Lemma B.6 *[Chickering, 1995] Consider any DAG* $\mathcal{G}$*. Let* $\mathcal{H}$ *be the result of reversing* $(i \to j) \in \texttt{E}(\mathcal{G})$*. Then* $\mathcal{H} \in \texttt{MEC}(\mathcal{G})$ *if and only if* $i \to j$ *is a covered edge.*

Theorem B.7 *[Chickering, 1995] Consider any pair of DAGs* $\mathcal{G}$ *and* $\mathcal{H}$ *over the same set of variables s.t.* $\mathcal{H} \in \texttt{MEC}(\mathcal{G})$*, and for which there are* k *edges in* $\mathcal{G}$ *that have opposite orientation in* $\mathcal{H}$*. Then there exists a sequence of* k *distinct covered edge reversals in* $\mathcal{G}$ *s.t.* $\mathcal{G}$ *becomes* $\mathcal{H}$ *after all reversals.*

Lemma B.8 *Given a graphoid* $\mathcal{P}$ *over* $\mathbf{V}$*, consider any* $\pi \in \Pi(\mathbf{v})$*. Suppose that* $(j \to k) \in \texttt{E}(\mathcal{G}_\pi)$ *is a covered edge, and let* $\mathcal{H}$ *be the DAG resulted from reversing* $(j \to k)$ *in* $\mathcal{G}_\pi$*. If* $\tau = \mathit{tuck}(\pi, j, k)$*, then*

(a) τ *is a causal order of* $\mathcal{H}$*;*

(b) $\texttt{E}(\mathcal{G}_\tau) \subseteq \texttt{E}(\mathcal{H})$*;*

(c) $|\texttt{E}(\mathcal{G}_\tau)| \leq |\texttt{E}(\mathcal{G}_\pi)|$*;*

(d) $\texttt{I}(\mathcal{G}_\pi) \subseteq \texttt{I}(\mathcal{G}_\tau)$*.*

Proof. A similar lemma has been shown in [Solus et al., 2021]. First, we write $\pi = \langle \boldsymbol{\delta}_{<j}, j, \boldsymbol{\delta}_{j\sim k}, k, \boldsymbol{\delta}_{>k} \rangle$ as usual. Consider $\pi' = \langle \boldsymbol{\delta}_{<j}, j, k, \boldsymbol{\delta}_{j\sim k}, \boldsymbol{\delta}_{>k} \rangle$. By **Lemma B.4**, we have $\mathcal{G}_\pi = \mathcal{G}_{\pi'}$. Note that $\tau = \mathit{tuck}(\pi, j, k) = \langle \boldsymbol{\delta}_{<j}, k, j, \boldsymbol{\delta}_{j\sim k}, \boldsymbol{\delta}_{>k} \rangle$ because $\mathit{tuck}(\pi, j, k)$ is a covered tuck. Thus, τ is (j, k)-different from π'. Also, since π' is a causal order of $\mathcal{G}_\pi$, it follows that τ is a causal order of $\mathcal{H}$ and thus (a) is proven.

Next, observe that $\texttt{I}(\mathcal{G}_\pi) = \texttt{I}(\mathcal{H})$ from **Lemma B.6**. From $\mathcal{G}_\pi \in \texttt{CMC}(\mathcal{P})$ by **Theorem 3.5**, we know that $\mathcal{H} \in \texttt{CMC}(\mathcal{P})$. Thus, (b) immediately follows from (a) and **Lemma B.2**. Also, (c) is entailed by $|\texttt{E}(\mathcal{G}_\tau)| \leq |\texttt{E}(\mathcal{H})| = |\texttt{E}(\mathcal{G}_\tau)|$. Finally, by **Theorem B.5**, we have $\texttt{I}(\mathcal{G}_\pi) = \texttt{I}(\mathcal{H}) \subseteq \texttt{I}(\mathcal{G}_\tau)$ as desired in (d). □

Before we compare TSP and unbounded GRaSP$_0$, we want to make an assumption related to how the set of covered edges in any particular DAG is ordered. To see the importance of such an assumption, observe that different orderings of $\texttt{E}^0(\mathcal{G}_\pi)$ (i.e., the set of covered edges in an induced DAG $\mathcal{G}_\pi$) can alter the output of TSP and also GRaSP$_0$. For example, suppose that $(j \to k), (j' \to k') \in \texttt{E}^0(\mathcal{G}_\pi)$. Say the DFS of GRaSP$_0$ starts with performing $\mathit{tuck}(\pi, j, k)$ and leads to some permutation τ. However, choosing to perform $\mathit{tuck}(\pi, j', k')$ instead at the beginning of the DFS procedure can lead to some τ' where $\mathcal{G}_\tau \neq \mathcal{G}_{\tau'}$. Hence, we enforce the assumption that the ordering of $\texttt{E}^0(\mathcal{G})$ for any DAG $\mathcal{G}$ is fixed arbitrarily. For instance, $(j \to k)$ precedes $(j' \to k')$ in $\texttt{E}^0(\mathcal{G})$ if $j < j'$, or $j = j'$ and $k < k'$. Consequently, the issue of order-dependence can be avoided even when comparing a Chickering sequence found by TSP and a ct-sequence found by unbounded GRaSP$_0$. In the following, this assumption will be made implicitly.

Now we revisit how TSP works. Given a graphoid $\mathcal{P}$ over $\mathbf{V}$ and an initial permutation $\pi \in \Pi(\mathbf{v})$, TSP begins with setting $\mathcal{G}$ as the induced $\mathcal{G}_\pi$. Starting with the root $\mathcal{G}$, TSP performs DFS to identify a SGS-minimal DAG $\mathcal{H}$ connected by a Chickering sequence from $\mathcal{G}$ such that $|\texttt{E}(\mathcal{G})| > |\texttt{E}(\mathcal{H})|$. TSP returns $\mathcal{G}$ if no such $\mathcal{H}$ is found. Otherwise, it updates $\mathcal{G}$ as $\mathcal{H}$ and repeat the procedure.

The DFS procedure of TSP aims to traverse from one SGS-minimal DAG to another SGS-minimal DAG by the construction of a Chickering sequence. Though we know that a Chickering sequence is obtained by the reversals of covered edges and deletions of directed edges, Solus et al. [2021] did not specify any ordering of these operations. Below we provide a more precise definition of the Chickering sequences considered by TSP.

Definition B.9 *Given a graphoid* $\mathcal{P}$ *over* $\mathbf{V}$*, a TSP-Chickering sequence* $\mathfrak{C} = \langle \mathcal{G}^1, ..., \mathcal{G}^m \rangle$ *is a Chickering sequence satisfying the following condition:*

(a) $\mathcal{G}^1, \mathcal{G}^m \in \texttt{SGS}(\mathcal{P})$*;*

(b) $\mathcal{G}^i$ *and* $\mathcal{G}^{i'}$ *are pairwise distinct for* $1 \leq i < i' \leq m$*;*

(c) if $|\texttt{E}(\mathcal{G}^1)| = |\texttt{E}(\mathcal{G}^m)|$*, then* $\mathcal{G}^1, ..., \mathcal{G}^m \in \texttt{SGS}(\mathcal{P})$ *where they differ by the reversals of some covered edges;*

(d) otherwise, there exists a turning index $1 < l < m$ *such that (i)* $\mathcal{G}^1, ..., \mathcal{G}^{l-1} \in \texttt{SGS}(\mathcal{P})$*, (ii)* $\mathcal{G}^1, ..., \mathcal{G}^l$ *differ by the reversals of some covered edges, and (iii)* $\mathcal{G}^{i+1}$ *is obtained from deleting a directed edge in* $\mathcal{G}^i \notin \texttt{SGS}(\mathcal{P})$ *for each* $l \leq i < m$*.*

Readers are suggested to find the original pseudocode of TSP in Solus et al. [2021] to verify that our **Definition B.9** is a fair description of the Chickering sequences considered by TSP. Conditions (a) and (b) are straightforward. (c) refers to the case where TSP cannot find a sparser SGS-minimal DAG. So if any $\mathcal{G}^i$ in $\mathfrak{C}$ were non-SGS-minimal, then TSP would have obtained a proper subgraph of $\mathcal{G}^i$ which is SGS-minimal by a series of edge-deletion. (d) refers to the case where TSP manages to find a sparser SGS-minimal DAG. Notice that $\mathcal{G}^2$ must be obtained by a covered edge reversal from $\mathcal{G}^1$ since $\mathcal{G}^1 \in \texttt{SGS}(\mathcal{P})$. If $\mathcal{G}^2 \notin \texttt{SGS}(\mathcal{P})$, then TSP can obtain the desired SGS-minimal DAG by a series of edge-deletion from $\mathcal{G}^2$. But if $\mathcal{G}^2 \in \texttt{SGS}(\mathcal{P})$, the procedure above repeats until finding the turning index l such that $\mathcal{G}^l \notin \texttt{SGS}(\mathcal{P})$ and then the sparser $\mathcal{G}^m \in \texttt{SGS}(\mathcal{P})$ can be obtained by a series of edge-deletion from $\mathcal{G}^l$.

Now we compare TSP and unbounded GRaSP$_0$ by considering their respective sequences traversed in the DFS procedure.

Lemma B.10 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any $\pi \in \Pi(\mathbf{v})$ and $\tau = tuck(\pi, j, k)$ where $(j \to k) \in \texttt{E}^0(\mathcal{G}_\pi)$. Given that $\mathfrak{T} = \langle \pi, \tau \rangle$ is a ct-sequence,*

(a) if $|\texttt{E}(\mathcal{G}_\tau)| = |\texttt{E}(\mathcal{G}_\pi)|$, then $\mathfrak{C} = \langle \mathcal{G}_\pi, \mathcal{G}_\tau \rangle$ is a TSP-Chickering sequence where $\mathcal{G}_\tau$ is obtained from reversing $(j \to k) \in \texttt{E}^0(\mathcal{G}_\pi)$;

(b) otherwise, there exists a TSP-Chickering sequence $\mathfrak{C} = \langle \mathcal{G}_\pi = \mathcal{G}^1, ..., \mathcal{G}^m = \mathcal{G}_\tau \rangle$ s.t. $\mathcal{G}^2$ is obtained from reversing $(j \to k) \in \texttt{E}^0(\mathcal{G}_\pi)$, and $\mathcal{G}^{i+1}$ is obtained from deleting a directed edge in $\mathcal{G}^i$ for each $2 \leq i < m$.

Proof. First, consider the DAG $\mathcal{H}$ obtained from reversing $(j \to k) \in \texttt{E}^0(\mathcal{G}_\pi)$. We start with the case in (a) where $|\texttt{E}(\mathcal{G}_\tau)| = |\texttt{E}(\mathcal{G}_\pi)| = |\texttt{E}(\mathcal{H})|$. We want to show that $\mathcal{G}_\tau = \mathcal{H}$. By **Lemma B.8** (b), we have $\texttt{E}(\mathcal{G}_\tau) \subseteq \texttt{E}(\mathcal{H})$. If $\texttt{E}(\mathcal{G}_\tau) \subset \texttt{E}(\mathcal{H})$ holds, then $|\texttt{E}(\mathcal{G}_\pi)| = |\texttt{E}(\mathcal{H})|$ will be violated. Hence, we have $\mathcal{G}_\tau = \mathcal{H}$ and thus $\mathfrak{C} = \langle \mathcal{G}_\pi, \mathcal{H} = \mathcal{G}_\tau \rangle$ is our desired TSP-Chickering sequence.

For (b), it follows from **Lemma B.8** (c) that $|\texttt{E}(\mathcal{G}_\tau)| < |\texttt{E}(\mathcal{H})| = |\texttt{E}(\mathcal{G}_\pi)|$. Let $\mathcal{G}_\pi$ and $\mathcal{H}$ be $\mathcal{G}^1$ and $\mathcal{G}^2$ respectively. By **Lemma B.8** (b) again, we have $\texttt{E}(\mathcal{G}_\tau) \subset \texttt{E}(\mathcal{G}^2)$ such that we can remove a directed edge from $\mathcal{G}^2$ once at a time until obtaining $\mathcal{G}_\tau$. Therefore, we have the desired TSP-Chickering sequence in (b). □

Lemma B.11 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any TSP-Chickering sequence $\mathfrak{C} = \langle \mathcal{G}^1, ..., \mathcal{G}^m \rangle$. Let π^1 be a causal order of $\mathcal{G}^1$. Then*

(a) if $|\texttt{E}(\mathcal{G}^1)| = |\texttt{E}(\mathcal{G}^m)|$, then $\mathcal{G}^{i+1} = \mathcal{G}_{\pi^{i+1}} = \mathcal{G}_{tuck(\pi,j,k)}$ where $j \to k$ is the covered edge reversed to obtain $\mathcal{G}^{i+1}$ from $\mathcal{G}^i$ for each $1 \leq i < m$ s.t. $\mathfrak{T} = \langle \pi^1, ..., \pi^m \rangle$ is a ct-sequence;

(b) otherwise, then $\mathcal{G}^{i+1} = \mathcal{G}_{\pi^{i+1}} = \mathcal{G}_{tuck(\pi,j,k)}$ where $j \to k$ is the covered edge reversed to obtain $\mathcal{G}^{i+1}$ from $\mathcal{G}^i$ for each $1 \leq i < l$ where l is the turning index of $\mathfrak{C}$ and $\mathcal{G}_{\pi^l} = \mathcal{G}^m$ s.t. $\mathfrak{T} = \langle \pi^1, ..., \pi^l \rangle$ is a ct-sequence.

Proof. (a) can be easily shown by **Lemma B.8**(a) and **Lemma B.2**. For (b), the proof of $\mathcal{G}^i = \mathcal{G}_{\pi^i}$ for each $1 \leq i < l$ is similar to that in (a). So we consider l where $\mathcal{G}^l \notin \texttt{SGS}(\mathcal{P})$ according to **Definition B.9**(d). However, it follows from **Lemma B.8** (a) that π^l is a causal order of $\mathcal{G}^l$. Since $\texttt{E}(\mathcal{G}^m) \subset \texttt{E}(\mathcal{G}^l)$, we know that π^l is also a causal order of $\mathcal{G}^m$. Lastly, given that $\mathcal{G}^m \in \texttt{SGS}(\mathcal{P})$, it follows from **Lemma B.2** that $\mathcal{G}_{\pi^l} = \mathcal{G}^m$. □

Lemma 4.4 *Given a graphoid $\mathcal{P}$, for any $\pi \in \Pi(\mathbf{v})$ and any Chickering sequence from $\mathcal{G}_\pi$ to some $\mathcal{H} \in \texttt{SGS}(\mathcal{P})$ considered by TSP, there exists a ct-sequence $\langle \pi, ..., \tau \rangle$ s.t. $\mathcal{G}_\tau = \mathcal{H}$.*

Proof. Given that a Chickering sequence considered by TSP is simply a TSP-Chickering sequence defined in **Definition B.9**, the lemma follows immediately from **Lemma B.11**. □

Theorem 4.7 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$ and any initial permutation $\pi \in \Pi(\mathbf{v})$, the DAG induced by the output of unbounded GRaSP$_0$ is equivalent to the DAG returned by TSP.*

Proof. Immediate from **Lemma B.10** and **Lemma B.11**. □

Now we turn to the discussion on the correctness of GRaSP$_0$ under faithfulness.

Lemma B.12 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$ and any $\pi \in \Pi(\mathbf{v})$, if $\mathcal{G}_\pi \notin \texttt{Pm}(\mathcal{P})$, then there exists a ct-sequence $\mathfrak{T} = \langle \pi, ..., \tau \rangle$ s.t. $\texttt{I}(\mathcal{G}_\pi) \subset \texttt{I}(\mathcal{G}_\tau)$.*

Proof. Suppose that $\mathcal{G}_\pi \notin \texttt{Pm}(\mathcal{P})$. By **Definition 3.7**, it follows that there exists $\mathcal{H} \in \texttt{CMC}(\mathcal{P})$ s.t. $\texttt{I}(\mathcal{G}_\pi) \subset \texttt{I}(\mathcal{H}) \subseteq \texttt{I}(\mathcal{P})$. By **Theorem 3.6**, we know that there exists a Chickering sequence $\mathfrak{C}_0 = \langle \mathcal{G}_\pi = \mathcal{G}^1, ..., \mathcal{G}^l = \mathcal{H} \rangle$. Without loss of generality, suppose that $\mathfrak{C}_0$ is the shortest Chickering sequence where each $\mathcal{G}^{i+1}$ differs from $\mathcal{G}^i \in \texttt{SGS}(\mathcal{P})$ by the reversal of a covered edge in $\texttt{E}^0(\mathcal{G}^i)$ for each $1 \leq i < l-1$, and $\mathcal{G}^l$ is obtained from deleting a directed edge in $\mathcal{G}^{l-1}$. Notice that $|\texttt{E}(\mathcal{G}^l)| < |\texttt{E}(\mathcal{G}^1)|$ due to the edge deletion. If $\mathcal{G}^l \in \texttt{SGS}(\mathcal{P})$, then $\mathfrak{C}_0$ is a TSP-Chickering sequence. Otherwise, we can easily construct a TSP-Chickering sequence $\mathfrak{C} = \langle \mathcal{G}^1, ..., \mathcal{G}^m \rangle$ with $l-1$ as the turning index and $\mathcal{G}^m \in \texttt{SGS}(\mathcal{P})$ obtained by repeated edge-deletion from $\mathcal{G}^l$ such that $\texttt{I}(\mathcal{G}^1) \subset \texttt{I}(\mathcal{G}^l) \subset \texttt{I}(\mathcal{G}^m)$. By **Lemma B.11** (b), we have the desired ct-sequence. □

Theorem 4.5 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$ and any $\pi \in \Pi(\mathbf{v})$, if $\mathcal{G}_\pi \notin \texttt{Pm}(\mathcal{P})$, then there exists a ct-sequence $\mathfrak{T} = \langle \pi, ..., \tau \rangle$ s.t. $\mathcal{G}_\tau \in \texttt{Pm}(\mathcal{P})$.*

Proof. Immediate from **Lemma B.12**. □

Theorem B.13 *Unbounded GRaSP$_0$ is correct and pointwise consistent under faithfulness.*

Proof. We review the argument for the correctness of unbounded GRaSP$_0$ under faithfulness given in the main paper. Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any initial permutation $\pi \in \Pi(\mathbf{v})$. Given that unbounded GRaSP$_0$ greedily search for a ct-sequence from π, it is guaranteed by **Theorem 4.5** that τ returned by unbounded GRaSP$_0$ in **Algorithm 2** induces a P-minimal DAG. Under faithfulness, we have $\mathcal{G}_\tau \in \texttt{MEC}(\mathcal{G}^*)$ due to **Theorem 3.8** and hence unbounded GRaSP$_0$ is correct.

Alternatively, the correctness and pointwise consistency of unbounded GRaSP$_0$ can also be proven directly from **Theorem 4.7** and the corresponding results of TSP in [Solus et al., 2021]. □

Corollary 4.6 *Unbounded GRaSP$_0$, GRaSP$_1$, and GRaSP$_2$ are correct and pointwise consistent under faithfulness.*

In the following, we want to prove that faithfulness is not only sufficient, but also *necessary* for the correctness of TSP and unbounded GRaSP$_0$. We first want to prove an interesting and novel equivalence between two causal razors: faithfulness and u-P-minimality.

Lemma B.14 *Given a joint probability distribution $\mathcal{P}$ over $\mathbf{V}$, for any $\langle X_i, X_j \,|\, \mathbf{X_k} \rangle \in \texttt{I}(\mathcal{P})$, there exists $\mathcal{G} \in \texttt{DAG}(\mathbf{V})$ s.t. $\texttt{I}(\mathcal{G}) = \{\langle X_i, X_j \,|\, \mathbf{X_k} \rangle\}$.*

Proof. Consider $\mathbf{V} = \{X_1, ..., X_m\}$. An empty DAG suffices when $m = 2$. So assume that $m \geq 3$. Without loss of generality, consider $\langle X_1, X_{k+2} \,|\, \mathbf{X_k} \rangle \in \texttt{I}(\mathcal{P})$ where $\mathbf{k} = \langle 2, ..., k+1 \rangle$, and the remaining vertices are $\langle k+3, ..., m \rangle$. We propose a procedure which guarantees the existence of the desired DAG $\mathcal{G}$.

---

1 $\mathcal{G} \leftarrow$ a complete undirected graph over $\mathbf{v}$
2 remove the adjacency $1 - k+2$ in $\mathcal{G}$
3 **foreach** (j, k) *that are adjacent in* $\mathcal{G}$ **do**
4 **if** $j < k$ **then**
5 orient $j \to k$ in $\mathcal{G}$
6 return $\mathcal{G}$

Line 3 to 5 guarantee that $\mathcal{G}$ is a DAG since all edges are directed and pointing from lower indices to higher indices such that no directed cycle can occur. Finally, $1 \perp_{\mathcal{G}} k+2 \,|\, \mathbf{k}$ holds because all paths from 1 to $k+2$ either contain a non-collider $i \in \mathbf{k}$ or contain a collider $i \notin \mathbf{k}$. Therefore, $\texttt{I}(\mathcal{G}) = \{\langle X_1, X_{k+2} \,|\, \mathbf{X_k} \rangle\}$ because no other d-separation relations hold in $\mathcal{G}$. □

Theorem B.15 *For any joint probability distribution $\mathcal{P}$, $\texttt{CFC}(\mathcal{P}) = \texttt{uPm}(\mathcal{P})$.*

Proof. [$\subseteq$] Suppose that $\mathcal{G} \in \texttt{CFC}(\mathcal{P})$. It follows that $\mathcal{G} \in \texttt{Pm}(\mathcal{P})$ by **Definition 3.7**. For any $\mathcal{G}' \in \texttt{CMC}(\mathcal{P})$, if $\texttt{I}(\mathcal{G}') \subset \texttt{I}(\mathcal{G})$, then $\mathcal{G}' \notin \texttt{Pm}(\mathcal{P})$. Hence, if $\mathcal{G}' \in \texttt{Pm}(\mathcal{P})$, then $\texttt{I}(\mathcal{G}') = \texttt{I}(\mathcal{G})$. Hence, $\mathcal{G} \in \texttt{uPm}(\mathcal{P})$.

[$\supseteq$] Suppose that $\mathcal{G} \notin \texttt{CFC}(\mathcal{P})$. Since $\texttt{uPm}(\mathcal{P}) \subseteq \texttt{Pm}(\mathcal{P})$ by **Definition 3.7**, if $\mathcal{G} \notin \texttt{Pm}(\mathcal{P})$, we have $\mathcal{G} \notin \texttt{uPm}(\mathcal{P})$ immediately. So consider the case where $\mathcal{G} \in \texttt{Pm}(\mathcal{P})$. It follows from $\mathcal{G} \notin \texttt{CFC}(\mathcal{P})$ that there exists a CI relation $\psi \in \texttt{I}(\mathcal{P}) \setminus \texttt{I}(\mathcal{G})$. By **Lemma B.14**, we can construct a DAG $\mathcal{G}^0$ such that $\texttt{I}(\mathcal{G}^0) = \{\psi\}$. Consequently, there exists $\mathcal{G}^1 \in \texttt{Pm}(\mathcal{P})$ such that $\texttt{I}(\mathcal{G}^0) \subseteq \texttt{I}(\mathcal{G}^1) \subseteq \texttt{I}(\mathcal{P})$. Since $\psi \in \texttt{I}(\mathcal{G}^1)$, we know that $\mathcal{G}^1 \notin \texttt{MEC}(\mathcal{G})$. Given that both $\mathcal{G}, \mathcal{G}^1 \in \texttt{Pm}(\mathcal{P})$, we have $\mathcal{G} \notin \texttt{uPm}(\mathcal{P})$. □

Theorem 4.8 *Given a graphoid $\mathcal{P}$, faithfulness is necessary for the correctness of TSP.*

Proof. Suppose that $(\mathcal{G}^*, \mathcal{P})$ is unfaithful. We consider the two kinds of unfaithfulness in [Zhang and Spirtes, 2008]: *detectable* (i.e., $\texttt{CFC}(\mathcal{P}) = \varnothing$) versus *undetectable* (i.e., $\mathcal{G}' \in \texttt{CFC}(\mathcal{P})$ where $\mathcal{G}' \notin \texttt{MEC}(\mathcal{G}^*)$). For the latter, TSP can identify $\mathcal{G}_\tau \in \texttt{Pm}(\mathcal{P}) = \texttt{CFC}(\mathcal{P}) = \texttt{MEC}(\mathcal{G}')$. However, TSP is incorrect because $\mathcal{G}_\tau \notin \texttt{MEC}(\mathcal{G}^*)$.

On the other hand, consider the case that $\texttt{CFC}(\mathcal{P}) = \varnothing$. By **Theorem B.15**, there exists $\mathcal{G} \in \texttt{Pm}(\mathcal{P})$ such that $\mathcal{G} \notin \texttt{MEC}(\mathcal{G}^*)$ even if $\mathcal{G}^* \in \texttt{Pm}(\mathcal{P})$. Recall that Chickering algorithm can only allow us to traverse to a DAG $\mathcal{H}$ from $\mathcal{G}$ satisfying $\texttt{I}(\mathcal{G}) \subseteq \texttt{I}(\mathcal{H})$. It entails that Chickering algorithm can only obtain DAGs that are in $\texttt{MEC}(\mathcal{G})$ since $\mathcal{G} \in \texttt{Pm}(\mathcal{P})$ and hence never be able to reach $\mathcal{G}^*$ where $\texttt{I}(\mathcal{G}_\pi) \nsubseteq \texttt{I}(\mathcal{G}^*)$. Therefore, by setting π as the initial permutation to TSP where $\mathcal{G}_\pi = \mathcal{G}$, TSP will return $\mathcal{G}_\pi$ incorrectly. □

Notice that **Theorem 4.8** is contrary to what Solus et al. [2021] suggested. They proposed an example arguing that TSP can be correct even under (detectable) unfaithfulness.[2] However, the distribution used in the example is not a semigraphoid. This renders their example illegitimate because every joint probability distribution is a semigraphoid.

C ESP AND GRASP-1

As shown in **Theorem 4.8** in the last section, TSP cannot be correct under unfaithfulness by choosing an arbitrary initial permutation. Consequently, one important question is how to relax the search space of TSP to identify a sparser permutation under unfaithfulness. Solus et al. [2021] proposed the *Edge SP* (ESP) algorithm based on an assumption strictly weaker than that assumed by TSP. However, unlike TSP, they did not provide an operational version of ESP in their work. In this section, we are going to show a theorem similar to **Theorem 4.7** but with respect to ESP and unbounded GRaSP_1. In other words, unbounded GRaSP_1 is an operational version of ESP. In the following, we first examine some technical notations used in [Mohammadi et al., 2018] and [Solus et al., 2021]. Readers are strongly suggested to visit [Solus et al., 2021] for the full discussion of ESP and relevant notations.

Given a set of measured variables $\mathbf{V}$, a *permutohedron* on $\mathbf{v}$, denoted $\mathsf{A}_\mathbf{v}$, is the convex hull in $\mathbb{R}^{|\mathbf{v}|}$ of all permutations in $\Pi(\mathbf{v})$. In simpler terms, $\mathsf{A}_\mathbf{v}$ is the *state space* with each *state* being a permutation $\pi \in \Pi(\mathbf{v})$. The neighborhood of states in $\mathsf{A}_\mathbf{v}$ is defined by adjacent transpositions (ATs) as in **Definition B.1** (b).

Notice that different states in $\mathsf{A}_\mathbf{v}$ can induce the same DAG given a graphoid $\mathcal{P}$. Thus, a natural way to narrow down the search space is to identify permutations inducing the same DAG. **Lemma B.3** provides such a characterization. Construct $\mathsf{A}_\mathbf{v}(\mathcal{P})$ by *contracting* neighborhood in $\mathsf{A}_\mathbf{v}$ to ATs that correspond to the CI relations in $\texttt{I}(\mathcal{P})$ specified in **Lemma B.3**. To be more specified, the contracted permutohedron $\mathsf{A}_\mathbf{v}(\mathcal{P})$, also known as the *DAG associahedron*, is the state space with each state being an induced DAG.[3] Two states $\mathcal{G}_1, \mathcal{G}_2$ in $\mathsf{A}_\mathbf{v}(\mathcal{P})$ are neighbors if and only if there exist $\pi^1, \pi^2 \in \Pi(\mathbf{v})$ s.t. $\mathcal{G}_{\pi^1} = \mathcal{G}^1$, $\mathcal{G}_{\pi^2} = \mathcal{G}^2$, and π^1 and π^2 are neighbors in the permutohedron $\mathsf{A}_\mathbf{v}$. As shown by Mohammadi et al. [2018], the DAG associahedron is a convex polytope where each vertex of $\mathsf{A}_\mathbf{v}(\mathcal{P})$ corresponds to a different DAG.

To draw a clearer picture, consider any $\pi, \tau \in \Pi(\mathbf{v})$ where τ is (j,k)-different from π for some $j, k \in \mathbf{v}$. They are neighbors in $\mathsf{A}_\mathbf{v}$ but they do not necessarily induce the same DAG. If $X_j \perp\!\!\!\perp_\mathcal{P} X_k \mid \mathbf{X}_{\texttt{Pre}(j,\pi)}$ holds, they induce the same DAG and

[2] See Figure 2 in the supplementary materials of [Solus et al., 2021].

[3] One can equivalently express each state in the DAG associahedron as the set of permutations which induce the same DAG. This is the original representation in [Mohammadi et al., 2018]. However, we prefer the representation given in [Solus et al., 2021] in the sense that one can easily compare DAGs that are in neighborhood.

thus correspond to the same state $\mathcal{G}_\pi$ in the DAG associahedron $\mathsf{A}_\mathbf{v}(\mathcal{P})$. But if the CI relation does not hold, then $\mathcal{G}_\pi$ and $\mathcal{G}_\tau$ are neighbors in $\mathsf{A}_\mathbf{v}(\mathcal{P})$. See Figure 1 for an example from [Solus et al., 2021].

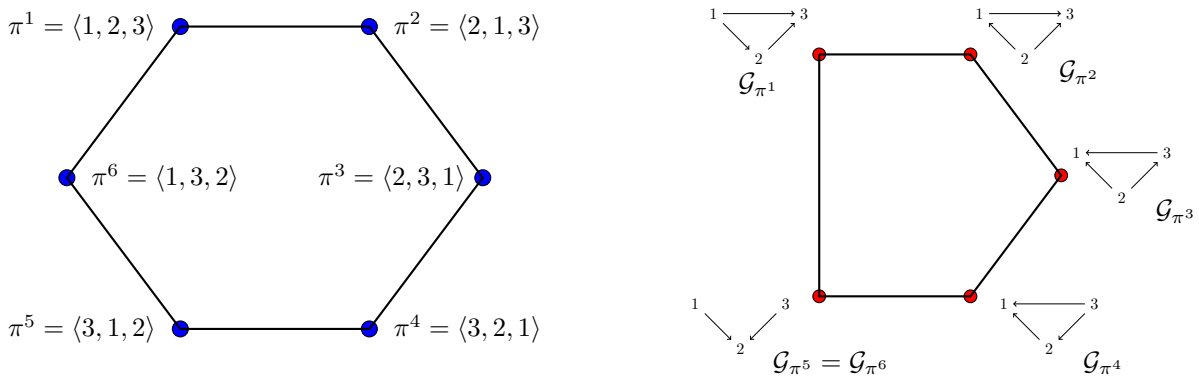


Figure 1: Given $\mathbf{V} = \{X_1, X_2, X_3\}$, consider $\mathtt{I}(\mathcal{P}) = \{\langle X_1, X_3 \,|\, \varnothing \rangle\}$. The diagram on the left is the permutohedron $\mathsf{A}_\mathbf{v}$ where each state is a permutation in $\Pi(\mathbf{v})$. The one on the right is the DAG associahedron $\mathsf{A}_\mathbf{v}(\mathcal{P})$ where each state is a different DAG in $\mathtt{SGS}(\mathcal{P})$. In particular, the two states π^5 and π^6 in $\mathsf{A}_\mathbf{v}$ are collapsed into a single state in $\mathsf{A}_\mathbf{v}(\mathcal{P})$ because they induce the same DAG.

Observe that each state in the DAG associahedron $\mathsf{A}_\mathbf{v}(\mathcal{P})$ corresponds to a SGS-minimal DAG according to **Theorem 3.5**. ESP performs a greedy DFS in $\mathsf{A}_\mathbf{v}(\mathcal{P})$. Given an initial permutation $\pi \in \Pi(\mathbf{v})$, set $\mathcal{G}$ as the induced $\mathcal{G}_\pi$ and traverse through $\mathsf{A}_\mathbf{v}(\mathcal{P})$ by a *weakly decreasing walk* to obtain $\mathcal{H}$ where $|\mathtt{E}(\mathcal{H})| < |\mathtt{E}(\mathcal{G}_\pi)|$.[4] If no such $\mathcal{H}$ exists, ESP returns $\mathcal{G} = \mathcal{G}_\pi$; else $\mathcal{G}$ is reset as $\mathcal{H}$ and repeat.

As noted by Solus et al. [2021], the construction of $\mathsf{A}_\mathbf{v}(\mathcal{P})$ is inefficient since one is only required to know the neighboring states instead of the entire $\mathsf{A}_\mathbf{v}(\mathcal{P})$ to perform the traversal. Below we show that unbounded GRaSP$_1$ can efficiently learn the neighbors of each state in $\mathsf{A}_\mathbf{v}(\mathcal{P})$ by our permutation-based operation *tuck* performed on singular edges. Before examining this claim, we introduce some useful definitions.

Given the permutohedron $\mathsf{A}_\mathbf{v}$, a *walk* $\mathfrak{W} = \langle \pi^1, ..., \pi^m \rangle$ is a sequence of neighboring states in $\mathsf{A}_\mathbf{v}$ such that $\pi^i, \pi^{i+1} \in \Pi(\mathbf{v})$ are in AT for each $1 \leq i < m$.

Definition C.1 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, for any walk $\mathfrak{W} = \langle \pi^1, ..., \pi^m \rangle$ in $\mathsf{A}_\mathbf{v}$,*

(a) $\mathfrak{W}$ is said to be DAG-preserving if $\mathcal{G}_{\pi^1} = ... = \mathcal{G}_{\pi^m}$;

(b) $\mathfrak{W}$ is said to be DAG-changing if $\langle \pi^1, ..., \pi^{m-1} \rangle$ is DAG-preserving and $\mathcal{G}_{\pi^{m-1}} \neq \mathcal{G}_{\pi^m}$.

In addition, for each DAG-changing walk $\mathfrak{W} = \langle \pi^1, ..., \pi^m \rangle$, we say that $\mathfrak{W}$ is relative to (j, k) if π^m is (j, k)-different from π^{m-1} for some $j, k \in \mathbf{v}$. Thus, each DAG-changing walk is relative to a pair of vertices corresponding to the last AT performed in the walk.

Definition C.2 *Given a set of variables $\mathbf{V}$, consider any $j, k \in \mathbf{v}$. Two DAGs $\mathcal{G}, \mathcal{H} \in \mathtt{DAG}(\mathbf{V})$ are said to be (j, k)-reverse if $(j \to k) \in \mathtt{E}(\mathcal{G})$ and $(k \to j) \in \mathtt{E}(\mathcal{H})$, and there does not exist any other $j', k' \in \mathbf{v}$ s.t. $(j' \to k') \in \mathtt{E}(\mathcal{G})$ and $(k' \to j') \in \mathtt{E}(\mathcal{H})$.*

Lemma C.3 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any $j, k \in \mathbf{v}$. Then there exists a DAG-changing walk $\mathfrak{W} = \langle \pi^1, ..., \pi^m \rangle$ relative to (j, k) in $\mathsf{A}_\mathbf{v}$ if and only if $\mathcal{G}_{\pi^1}$ and $\mathcal{G}_{\pi^m}$ are neighbors in $\mathsf{A}_\mathbf{v}(\mathcal{P})$ that are (j, k)-reverse.*

Proof. For the forward direction, given that π^{m-1} and π^m are (j, k)-different but induce different DAGs, it follows from the definition of $\mathsf{A}_\mathbf{v}(\mathcal{P})$ that $\mathcal{G}_{\pi^1} = \mathcal{G}_{\pi^{m-1}}$ and $\mathcal{G}_{\pi^m}$ are neighbors in $\mathsf{A}_\mathbf{v}(\mathcal{P})$. Also, we know that $(j \to k) \in \mathtt{E}(\mathcal{G}_{\pi^{m-1}})$ and

[4] In [Solus et al., 2021], their pseudocode does not indicate that such a walk needs to be weakly decreasing but such a requirement is imposed in the description of the algorithm.

$(k \to j) \in \mathtt{E}(\mathcal{G}_{\pi^m})$ by **Lemma B.3** and (RU). The fact that $\mathcal{G}_{\pi^1} = \mathcal{G}_{\pi^{m-1}}$ and $\mathcal{G}_{\pi^m}$ are (j,k)-reverse follows immediately from (RU) and the assumption that π^{m-1} is (j,k)-different from π^m.

For the backward direction, suppose that $\mathcal{G}_\pi$ and $\mathcal{G}_\tau$ are neighbors in $\mathsf{A}_{\mathbf{v}}(\mathcal{P})$ that are (j,k)-reverse. It entails from (RU) that there exist $\pi', \tau' \in \Pi(\mathbf{v})$ such that π' and τ' are (j,k)-different where $\mathcal{G}_\tau = \mathcal{G}_{\tau'}$ and $\mathcal{G}_\pi = \mathcal{G}_{\pi'}$. Hence, $\langle \pi', \tau' \rangle$ is our desired DAG-changing walk relative to (j,k) in $\mathsf{A}_{\mathbf{v}}$. □

Lemma C.4 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any pair $\pi^1, \tau^1 \in \Pi(\mathbf{v})$ such that $\pi^1 = \langle \boldsymbol{\delta}_1, j, k, \boldsymbol{\delta}_2 \rangle$ for some sub-sequences $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$ of π^1, and $\tau^1 = \langle \boldsymbol{\zeta}_1, j, k, \boldsymbol{\zeta}_2 \rangle$ for some sub-sequences $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$ of τ^1. Further consider $\pi^2 = \langle \boldsymbol{\delta}_1, k, j, \boldsymbol{\delta}_2 \rangle$ and $\tau^2 = \langle \boldsymbol{\zeta}_1, k, j, \boldsymbol{\zeta}_2 \rangle$. If $\mathcal{G}_{\pi^1} = \mathcal{G}_{\tau^1}$, then $\mathcal{G}_{\pi^2} = \mathcal{G}_{\tau^2}$.*

Proof. Notice that $\mathcal{G}_{\pi^2} \in \mathtt{SGS}(\mathcal{P})$ by **Theorem 3.5**. If we can show that τ^2 is a causal order of $\mathcal{G}_{\pi^2}$, it follows from **Lemma B.2** that $\mathcal{G}_{\pi^2} = \mathcal{G}_{\tau^2}$. To do so, it suffices to show the following. For any $i \in \mathbf{v} \setminus \{j,k\}$,

(i) if $(i \to j) \in \mathtt{E}(\mathcal{G}_{\pi^2})$, then $i \in \boldsymbol{\zeta}_1$;
(ii) if $(i \to k) \in \mathtt{E}(\mathcal{G}_{\pi^2})$, then $i \in \boldsymbol{\zeta}_1$;
(iii) if $(j \to i) \in \mathtt{E}(\mathcal{G}_{\pi^2})$, then $i \in \boldsymbol{\zeta}_2$;
(iv) if $(k \to i) \in \mathtt{E}(\mathcal{G}_{\pi^2})$, then $i \in \boldsymbol{\zeta}_2$.

For (i), suppose that $(i \to j) \in \mathtt{E}(\mathcal{G}_{\pi^2})$. If $(i \to j) \in \mathtt{E}(\mathcal{G}_{\pi^1})$ as well, then $(i \to j) \in \mathtt{E}(\mathcal{G}_{\tau^1})$ since $\mathcal{G}_{\pi^1} = \mathcal{G}_{\tau^1}$. This entails that $i \in \boldsymbol{\zeta}_1$. On the other hand, consider the case that $(i \to j) \notin \mathtt{E}(\mathcal{G}_{\pi^1})$. Then

$$X_i \not\perp\!\!\!\perp_{\mathcal{P}} X_j \,|\, \mathbf{X}_{\boldsymbol{\delta}_1 \setminus \{i\}} \cup \{X_k\} \qquad \because (i \to j) \in \mathtt{E}(\mathcal{G}_{\pi^2}) \tag{4}$$
$$X_i \not\perp\!\!\!\perp_{\mathcal{P}} \{X_j, X_k\} \,|\, \mathbf{X}_{\boldsymbol{\delta}_1 \setminus \{i\}} \qquad \because (4), \textit{weak union} \tag{5}$$
$$X_i \perp\!\!\!\perp_{\mathcal{P}} X_j \,|\, \mathbf{X}_{\boldsymbol{\delta}_1 \setminus \{i\}} \qquad \because (i \to j) \notin \mathtt{E}(\mathcal{G}_{\pi^1}) \tag{6}$$
$$X_i \not\perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\boldsymbol{\delta}_1 \setminus \{i\}} \cup \{X_j\} \qquad \because (5), (6), \textit{contraction} \tag{7}$$

By (RU), (8) entails that $(i \to k) \in \mathtt{E}(\mathcal{G}_{\pi^1}) = \mathtt{E}(\mathcal{G}_{\tau^1})$. Since τ^1 is a causal order of $\mathcal{G}_{\tau^1}$, we have $i \in \boldsymbol{\zeta}_1$.

For (ii), suppose that $(i \to k) \in \mathtt{E}(\mathcal{G}_{\pi^2})$. Similar to (i), the case for $(i \to k) \in \mathtt{E}(\mathcal{G}_{\pi^1})$ is simple. So consider the case where $(i \to k) \notin \mathtt{E}(\mathcal{G}_{\pi^1})$.

$$X_i \not\perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\boldsymbol{\delta}_1 \setminus \{i\}} \qquad \because (i \to k) \in \mathtt{E}(\mathcal{G}_{\pi^2}) \tag{8}$$
$$X_i \not\perp\!\!\!\perp_{\mathcal{P}} \{X_j, X_k\} \,|\, \mathbf{X}_{\boldsymbol{\delta}_1 \setminus \{i\}} \qquad \because (8), \textit{decomposition} \tag{9}$$
$$X_i \perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\boldsymbol{\delta}_1 \setminus \{i\}} \cup \{X_j\} \qquad \because (i \to k) \notin \mathtt{E}(\mathcal{G}_{\pi^1}) \tag{10}$$
$$X_i \not\perp\!\!\!\perp_{\mathcal{P}} X_j \,|\, \mathbf{X}_{\boldsymbol{\delta}_1 \setminus \{i\}} \qquad \because (9), (10), \textit{contraction} \tag{11}$$

By (RU), (13) entails that $(i \to j) \in \mathtt{E}(\mathcal{G}_{\pi^1}) = \mathtt{E}(\mathcal{G}_{\tau^1})$ and hence $i \in \boldsymbol{\zeta}_1$.

For (iii), suppose that $(j \to i) \in \mathtt{E}(\mathcal{G}_{\pi^2})$. Then we have $(j \to i) \in \mathtt{E}(\mathcal{G}_{\pi^1})$ by (RU) because $\mathtt{Pre}(i, \pi^1) = \mathtt{Pre}(i, \pi^2)$. Hence $(j \to i) \in \mathtt{E}(\mathcal{G}_{\tau^1})$ since $\mathcal{G}_{\pi^1} = \mathcal{G}_{\tau^1}$. Given that τ^1 is a causal order of $\mathcal{G}_{\tau^1}$, we have $i \in \boldsymbol{\zeta}_2$. (iv) is analogous to (iii). □

Lemma C.5 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any two DAG-changing walks $\mathfrak{W} = \langle \pi^1, ..., \pi^m \rangle$ and $\mathfrak{W}' = \langle \tau^1, ..., \tau^n \rangle$ in $\mathsf{A}_{\mathbf{v}}$ where $\pi^1 = \tau^1$. If $\mathfrak{W}$ and $\mathfrak{W}'$ are both relative to the same (j,k) for some $j, k \in \mathbf{v}$, then $\mathcal{G}_{\pi^m} = \mathcal{G}_{\tau^n}$.*

Proof. Immediate from **Definition C.1** and **Lemma C.4**. □

Lemma C.6 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any DAG-changing walk $\mathfrak{W} = \langle \pi^1, ..., \pi^m \rangle$ in $\mathsf{A}_{\mathbf{v}}$ which is relative to (j,k) for some $j, k \in \mathbf{v}$. Then $j \to k$ is a singular edge in $\mathcal{G}_{\pi^1}$.*

Proof. Let $\mathfrak{W}_0$ denotes the DAG-preserving walk $\langle \pi^1, ..., \pi^{m-1} \rangle$. Given that π^m is (j,k)-different from π^{m-1}, it follows from **Lemma B.3** and (RU) that $(j \to k) \in \mathtt{E}(\mathcal{G}_{\pi^{m-1}})$. Since $\mathfrak{W}_0$ is a DAG-preserving walk in $\mathsf{A}_{\mathbf{v}}$, we have $(j \to k) \in \mathtt{E}(\mathcal{G}_{\pi^1}) = \mathtt{E}(\mathcal{G}_{\pi^{m-1}})$.

Next, suppose by reductio that $j \to k$ is not a singular edge in $\mathcal{G}_{\pi^1}$. Then there is a directed path from j to k other than $j \to k$ in $\mathcal{G}_{\pi^1}$. So there exists $l \in \mathbf{v}$ such that $l \in \mathtt{De}(j, \mathcal{G}_{\pi^1}) \cap \mathtt{An}(k, \mathcal{G}_{\pi^1})$. In order to ensure that j and k are π^{m-1}-adjacent, either $\pi^{m-1}[l] < \pi^{m-1}[j]$ or $\pi^{m-1}[l] > \pi^{m-1}[k]$ holds. However, either case will violate that π^{m-1} is a causal order of $\mathcal{G}_{\pi^{m-1}} = \mathcal{G}_{\pi^1}$. □

Lemma C.7 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider $\pi \in \Pi(\mathbf{v})$ where $(j \to k) \in \mathtt{E}(\mathcal{G}_\pi)$ is a singular edge for some $j, k \in \mathbf{v}$. Then there exists a DAG-changing walk $\mathfrak{W} = \langle \pi, ..., \tau \rangle$ in $\mathsf{A}_\mathbf{v}$ relative to (j, k) where $\tau = tuck(\pi, j, k)$.*

Proof. First, we rewrite $\pi = \langle \boldsymbol{\delta}_{<j}, j, \boldsymbol{\delta}_{j\sim k}, k, \boldsymbol{\delta}_{>k} \rangle$ as usual. Then we partition $\boldsymbol{\delta}_{j\sim k}$ as follows: $\boldsymbol{\zeta}_a = \langle i \in \boldsymbol{\delta}_{j\sim k} : i \in \mathtt{An}(k, \mathcal{G}_\pi) \rangle$, and $\boldsymbol{\zeta}_b = \langle i \in \boldsymbol{\delta}_{j\sim k} : i \notin \mathtt{An}(k, \mathcal{G}_\pi) \rangle$. Given that $(j \to k)$ is a singular edge, we know that $\mathtt{De}(j, \mathcal{G}_{\pi^1}) \cap \mathtt{An}(k, \mathcal{G}_{\pi^1}) = \varnothing$. In other words, we know that (i) each vertex in $\boldsymbol{\zeta}_a$ has no ancestor in $\boldsymbol{\delta}_{j\sim k} \setminus \boldsymbol{\zeta}_a$ in $\mathcal{G}_\pi$ and (ii) each vertex in $\boldsymbol{\zeta}_b$ has no descendant in $\boldsymbol{\delta}_{j\sim k} \setminus \boldsymbol{\zeta}_b$ in $\mathcal{G}_\pi$.

Now consider the permutation $\tau' = \langle \boldsymbol{\delta}_{<j}, \boldsymbol{\zeta}_a, j, k, \boldsymbol{\zeta}_b, \boldsymbol{\delta}_{>k} \rangle$ in particular. We want to show that there exists a DAG-preserving walk from π to τ'. Such a walk is easy to construct. First, perform repeated ATs by moving each $i \in \boldsymbol{\zeta}_a$ prior to j from left to right, and then repeated ATs by moving each $i \in \boldsymbol{\zeta}_b$ behind k from right to left. The two sets of ATs are licensed by (i) and (ii) respectively. Hence, we have $\mathcal{G}_{\tau'} = \mathcal{G}_\pi$. Finally, consider $\tau = tuck(\pi, j, k) = \langle \boldsymbol{\delta}_{<j}, \boldsymbol{\zeta}_a, k, j, \boldsymbol{\zeta}_b, \boldsymbol{\delta}_{>k} \rangle$ which is (j, k)-different from τ'. By (RU) and **Lemma B.3**, we know that $\mathcal{G}_\tau \neq \mathcal{G}_{\tau'}$ and thus $\langle \pi, ..., \tau', \tau \rangle$ is a DAG-changing walk in $\mathsf{A}_\mathbf{v}$ relative to (j, k). □

Theorem C.8 *Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, consider any DAG-changing walk $\mathfrak{W} = \langle \pi^1, ..., \pi^m \rangle$ in $\mathsf{A}_\mathbf{v}$ which is relative to (j, k) for some $j, k \in \mathbf{v}$. Then $\mathcal{G}_{\pi^m} = \mathcal{G}_\tau$ where $\tau = tuck(\pi^1, j, k)$.*

Proof. We obtain a DAG-changing walk $\mathfrak{W}' = \langle \pi^1, ..., \tau \rangle$ in $\mathsf{A}_\mathbf{v}$ relative to (j, k) by **Lemma C.7**. Since both $\mathfrak{W}$ and $\mathfrak{W}'$ are relative to the same (j, k), it follows from **Lemma C.5** that $\mathcal{G}_{\pi^m} = \mathcal{G}_\tau$. □

Similar to the discussion in Appendix **B**, we want to fix the ordering of the set of singular edges in any DAG. This ensures that ESP and unbounded GRasP$_1$ will not yield different DAGs simply due to the issue of order-dependence. Below we prove that ESP and unbounded GRaSP$_1$ are equivalent algorithms.

Theorem C.9 *Given a graphoid $\mathcal{P}$ and any initial permutation $\pi \in \Pi(\mathbf{v})$, the DAG induced by the output of unbounded GRaSP$_1$ is equivalent to the DAG returned by ESP.*

Proof. Consider any $j, k \in \mathbf{v}$. By **Lemma C.6** and **Lemma C.7**, every DAG-changing walk $\mathfrak{W} = \langle \pi^1, ..., \pi^m \rangle$ in $\mathsf{A}_\mathbf{v}$ relative to (j, k) corresponds to a *tuck* operation of the singular edge $j \to k$ in $\mathtt{E}(\mathcal{G}_{\pi^1})$. Hence, by **Lemma C.3**, we know that $tuck(\pi^1, j, k)$ corresponds to the neighboring relation between $\mathcal{G}_{\pi^1}$ and $\mathcal{G}_{\pi^m}$ in $\mathsf{A}_\mathbf{v}(\mathcal{P})$ that are (j, k)-reverse. Therefore, every step taken by ESP to move to a neighboring state in $\mathsf{A}_\mathbf{v}(\mathcal{P})$ (relative to a unique pair of vertices) is equivalent to the tuck operation taken by GRaSP$_1$ over the same pair of vertices. □

D CAUSAL RAZORS AND GRASP

In this section, we first provide a logical analysis of the causal razors discussed in the main text.[5] Then we construct new causal razors with respect to each tier of GRaSP, and show how a higher tier of GRaSP requires a strictly weaker causal razor.

Theorem D.1 *The following statements are true:*

(a) For any joint probability distribution $\mathcal{P}$, $\mathtt{uPm}(\mathcal{P}) = \mathtt{CFC}(\mathcal{P}) \subseteq \mathtt{uFr}(\mathcal{P}) \subseteq \mathtt{Fr}(\mathcal{P}) \subseteq \mathtt{Pm}(\mathcal{P}) \subseteq \mathtt{SGS}(\mathcal{P})$.

(b) For any joint probability distribution $\mathcal{P}$, if faithfulness is satisfied, $\mathtt{CFC}(\mathcal{P}) = \mathtt{uFr}(\mathcal{P}) = \mathtt{Fr}(\mathcal{P}) = \mathtt{Pm}(\mathcal{P})$.

(c) There exists a joint probability distribution s.t. $\mathtt{CFC}(\mathcal{P}) \subset \mathtt{uFr}(\mathcal{P})$.

(d) There exists a joint probability distribution s.t. $\mathtt{uFr}(\mathcal{P}) \subset \mathtt{Fr}(\mathcal{P})$.

(e) There exists a joint probability distribution s.t. $\mathtt{Fr}(\mathcal{P}) \subset \mathtt{Pm}(\mathcal{P})$.

(f) There exists a joint probability distribution s.t. $\mathtt{Pm}(\mathcal{P}) \subset \mathtt{SGS}(\mathcal{P})$.

[5] There are other causal razors discussed in the literature, including, but not limited to, *adjacency-faithfulness* and *orientation-faithfulness* in [Ramsey et al., 2006], and *triangle-faithfulness* in Zhang [2013]. But they do not have a strong connection with our discussion of GRaSP and so will not analyzed in this work.

Proof. For (a), $\texttt{uPm}(\mathcal{P}) = \texttt{CFC}(\mathcal{P})$ is our result in **Theorem B.15**. $\texttt{CFC}(\mathcal{P}) \subseteq \texttt{uFr}(\mathcal{P})$ is proven in [Raskutti and Uhler, 2018], $\texttt{uFr}(\mathcal{P}) \subseteq \texttt{Fr}(\mathcal{P})$ is true by **Definition 3.3**, $\texttt{Fr}(\mathcal{P}) \subseteq \texttt{Pm}(\mathcal{P})$ in [Forster et al., 2020], and $\texttt{Pm}(\mathcal{P}) \subseteq \texttt{SGS}(\mathcal{P})$ in [Zhang, 2013]. (b) is a direct consequence of (a) and **Theorem 3.8**.

For (c), see [Raskutti and Uhler, 2018, Theorem 2.4]. For (d), see [Forster et al., 2020, Figure 6]. For (e), see [Raskutti and Uhler, 2018, Theorem 2.5]. For (f), see [Zhang, 2013, Figure 2]. Additionally, the example in **Theorem D.6** and its corresponding Figure 2 verifies (c) and (e): $\mathcal{G}^* \in \texttt{uFr}(\mathcal{P}) \setminus \texttt{CFC}(\mathcal{P})$ and $\mathcal{G}_\pi \in \texttt{Pm}(\mathcal{P}) \setminus \texttt{Fr}(\mathcal{P})$. On the other hand, each of $\mathcal{G}_{\pi^1}, \mathcal{G}_{\pi^2}, \mathcal{G}_{\pi^3}$, and $\mathcal{G}_{\pi^4}$ in the DAG-associahedron in Figure 1 is in $\texttt{SGS}(\mathcal{P}) \setminus \texttt{Pm}(\mathcal{P})$ verifying (f). □

Definition D.2 *(TSP-razor and ESP-razor) Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, let $tsp(\mathcal{P}, \pi)$ be the DAG returned by TSP on $\mathcal{P}$ by setting π as the initial permutation. Define $\texttt{TSP}(\mathcal{P}) = \{\mathcal{G} \in \texttt{DAG}(\mathbf{V}) : \pi \in \Pi(\mathbf{v})$ and $\mathcal{G} = tsp(\mathcal{P}, \pi)\}$ as the set of DAGs returned by TSP on $\mathcal{P}$ over each initial permutation in $\Pi(\mathbf{v})$. Further define*

$$\texttt{TSPr}(\mathcal{P}) = \{\mathcal{G} \in \texttt{TSP}(\mathcal{P}) : \neg\exists \mathcal{G}' \in \texttt{TSP}(\mathcal{P}) \text{ s.t. } \mathcal{G}' \notin \texttt{MEC}(\mathcal{G})\}.$$

$(\mathcal{G}^*, \mathcal{P})$ satisfies the TSP-razor if $\mathcal{G}^* \in \texttt{TSPr}(\mathcal{P})$. Similarly for ESP, esp, ESP, ESPr, and ESP-razor.

One can observe that $\texttt{TSPr}(\mathcal{P}) = \texttt{TSP}(\mathcal{P})$ if every DAG in $\texttt{TSP}(\mathcal{P})$ belongs to the same MEC, and $\texttt{TSPr}(\mathcal{P}) = \varnothing$ otherwise. The same is also true for $\texttt{ESPr}(\mathcal{P})$ and $\texttt{ESP}(\mathcal{P})$. These definitions will be proven useful when we compare them with the classes of DAGs discussed in **Theorem D.1**. Below we provide a similar definition for each tier of GRaSP.

Definition D.3 *(GRaSP$_t$-razor) Given a graphoid $\mathcal{P}$ over $\mathbf{V}$, for $t \in \{0, 1, 2\}$, define $\texttt{GRaSP}_t(\mathcal{P}) = \{\mathcal{G}_\tau \in \texttt{DAG}(\mathbf{V}) : \pi \in \Pi(\mathbf{v})$ and $\tau = grasp(\mathcal{P}, \pi, |\mathbf{v}|, t)\}$ as the set of DAGs returned by unbounded GRaSP$_t$ on $\mathcal{P}$ over each initial permutation in $\Pi(\mathbf{v})$. Further define*

$$\texttt{GRaSP}_t\texttt{r}(\mathcal{P}) = \{\mathcal{G} \in \texttt{GRaSP}_t(\mathcal{P}) : \neg\exists \mathcal{G}' \in \texttt{GRaSP}_t(\mathcal{P}) \text{ s.t. } \mathcal{G}' \notin \texttt{MEC}(\mathcal{G})\}.$$

$(\mathcal{G}^*, \mathcal{P})$ satisfies the GRaSP$_t$-razor if $\mathcal{G}^* \in \texttt{GRaSP}_t\texttt{r}(\mathcal{P})$.

Theorem D.4 *Given a graphoid $\mathcal{P}$, the following statement is true:*

$$\texttt{CFC}(\mathcal{P}) = \texttt{TSPr}(\mathcal{P}) = \texttt{GRaSP}_0\texttt{r}(\mathcal{P}) \subseteq \texttt{ESPr}(\mathcal{P}) = \texttt{GRaSP}_1\texttt{r}(\mathcal{P}) \subseteq \texttt{GRaSP}_2\texttt{r}(\mathcal{P}) \subseteq \texttt{uFr}(\mathcal{P}).$$

Proof. $\texttt{CFC}(\mathcal{P}) = \texttt{TSPr}(\mathcal{P}) = \texttt{GRaSP}_0\texttt{r}(\mathcal{P})$ is directly entailed by **Theorem 4.7** and **Theorem 4.8**. Solus et al. [2021] showed that $\texttt{TSPr}(\mathcal{P}) \subseteq \texttt{ESPr}(\mathcal{P})$. $\texttt{ESPr}(\mathcal{P}) = \texttt{GRaSP}_1\texttt{r}(\mathcal{P})$ is entailed by **Theorem C.9**.

Next, to show that $\texttt{GRaSP}_1\texttt{r}(\mathcal{P}) \subseteq \texttt{GRaSP}_2\texttt{r}(\mathcal{P})$, notice that $\texttt{GRaSP}_1\texttt{r}(\mathcal{P}) = \texttt{GRaSP}_1(\mathcal{P})$ when all DAGs in $\texttt{GRaSP}_1(\mathcal{P})$ belong to the same MEC, and $\texttt{GRaSP}_1\texttt{r}(\mathcal{P}) = \varnothing$ otherwise. The latter case validates $\texttt{GRaSP}_1\texttt{r}(\mathcal{P}) \subseteq \texttt{GRaSP}_2\texttt{r}(\mathcal{P})$ trivially. Now consider the former case where all DAGs in the non-empty $\texttt{GRaSP}_1(\mathcal{P})$ belong to the same MEC and so they have the same number of edges. Now consider any $\pi \in \Pi(\mathbf{v})$ satisfying $\mathcal{G}_\pi \in \texttt{Fr}(\mathcal{P})$ (where $\texttt{Fr}(\mathcal{P})$ is necessarily non-empty). We know that $\mathcal{G}_\pi \in \texttt{GRaSP}_1(\mathcal{P})$. This is because every initial permutation in $\Pi(\mathbf{v})$ is considered and unbounded GRaSP$_1$ will never return a denser permutation than its initial permutation. Hence, every DAG in $\texttt{GRaSP}_1(\mathcal{P})$ is the sparsest Markovian DAG. (The same also holds when $\texttt{GRaSP}_2\texttt{r}(\mathcal{P}) \neq \varnothing$.) Then the construction of **Algorithm 2** entails that GRaSP$_2$ will return the same permutation as GRaSP$_1$. Hence, $\texttt{GRaSP}_1\texttt{r}(\mathcal{P}) = \texttt{GRaSP}_2\texttt{r}(\mathcal{P})$ when all DAGs in $\texttt{GRaSP}_1(\mathcal{P})$ belongs to the same MEC.

Lastly, to show that $\texttt{GRaSP}_2\texttt{r}(\mathcal{P}) \subseteq \texttt{uFr}(\mathcal{P})$, we use a proof similar to the above. First, the case where $\texttt{GRaSP}_2\texttt{r}(\mathcal{P}) = \varnothing$ is trivial. Consider the case where $\texttt{GRaSP}_2\texttt{r}(\mathcal{P}) = \texttt{GRaSP}_2(\mathcal{P})$ s.t. all DAGs in $\texttt{GRaSP}_2(\mathcal{P})$ are in the same MEC. Using a similar inference used in the last paragraph, we know that every DAG in $\texttt{GRaSP}_2(\mathcal{P})$ is the sparsest Markovian DAG. Therefore, $\texttt{GRaSP}_2\texttt{r}(\mathcal{P}) = \texttt{uFr}(\mathcal{P})$ when all DAGs in $\texttt{GRaSP}_2(\mathcal{P})$ belongs to the same MEC. □

Theorem D.5 *There exists a graphoid $\mathcal{P}$ s.t. $\texttt{GRaSP}_0\texttt{r}(\mathcal{P}) \subset \texttt{GRaSP}_1\texttt{r}(\mathcal{P})$.*

Proof. Given the equivalence between TSP and unbounded GRaSP$_0$ shown in **Theorem 4.7**, and that between ESP and unbounded GRaSP$_1$ in **Theorem C.9**, we can borrow the example from [Solus et al., 2021] on how ESP requires a strictly weaker causal razor than TSP. We refer the readers to Figure 3 in the supplementary materials of [Solus et al., 2021]. □

In the remainder of this section, we discuss two examples: how unbounded GRaSP$_2$ requires a strictly weaker causal razor than unbounded GRaSP$_1$, and how unbounded GRaSP$_2$ requires a strictly stronger causal razor than u-frugality. The joint distribution of each example below is a compositional graphoid. For the sake of simplicity, we only include CI relations that hold between two *singleton* sets of variables such that all other CI relations entailed by each of the graphoid axioms discussed in Appendix **A.2** are understood.

Theorem D.6 *There exists a graphoid $\mathcal{P}$ s.t.* $\texttt{GRaSP}_1\texttt{r}(\mathcal{P}) \subset \texttt{GRaSP}_2\texttt{r}(\mathcal{P})$.

Proof. Given $\mathbf{V} = \{X_1, ..., X_4\}$, consider the unfaithful model $(\mathcal{G}^*, \mathcal{P})$ where the true DAG $\mathcal{G}^*$ is shown on the left in Figure 2, and $\texttt{I}(\mathcal{P}) = \Phi \cup \Psi$ where Φ is the set of faithful CI relations and Ψ is the set of unfaithful CI relations as listed below:

$$\Phi = \{\phi_1 = \langle X_1, X_3 \,|\, \{X_2\}\rangle, \phi_2 = \langle X_2, X_4 \,|\, \{X_1, X_3\}\rangle\};$$
$$\Psi = \{\psi_1 = \langle X_2, X_4 \,|\, \varnothing\rangle\}.$$

For every $\mathcal{G} \in \texttt{CMC}(\mathcal{P})$ where $\psi_1 \in \texttt{I}(\mathcal{G})$, we have $5 = |\texttt{E}(\mathcal{G})| > |\texttt{E}(\mathcal{G}^*)| = 4$. Also, all 4-edge Markovian DAGs are in the same MEC. Hence, u-frugality is satisfied. Consider feeding the initial permutation $\pi = \langle 2, 4, 1, 3\rangle$ to unbounded GRaSP$_1$. It will return the same π after the DFS procedure and the induced $\mathcal{G}_\pi$, as shown on the right in Figure 2, contains 5 edges. Therefore, unbounded GRaSP$_1$ fails to return the sparsest permutation under some initial permutation and $\texttt{GRaSP}_1\texttt{r}(\mathcal{P}) = \varnothing$.

On the contrary, $|\mathbf{v}|! = 24$ initial permutations have been tested on unbounded GRaSP$_2$ and it returns $\hat{\tau}$ where $\mathcal{G}_{\hat{\tau}} \in \texttt{MEC}(\mathcal{G}^*)$ for each initial permutation. Hence, $\texttt{GRaSP}_2\texttt{r}(\mathcal{P}) \neq \varnothing$. □

$\mathcal{G}^*$ $\quad\quad$ $\mathcal{G}_\pi$

Figure 2: An unfaithful model satisfying u-frugality. The true DAG $\mathcal{G}^*$ is shown on the left where the two shaded vertices indicate the unfaithful marginal independence $X_2 \perp\!\!\!\perp_{\mathcal{P}} X_4 \,|\, \varnothing$. Unbounded GRaSP$_1$ returns its initial permutation $\pi = \langle 2, 4, 1, 3\rangle$. The induced DAG $\mathcal{G}_\pi$ is shown on the right with 5 edges. However, unbounded GRaSP$_2$ manages to return one of the sparsest permutations under every initial permutation.

Theorem D.7 *There exists a graphoid $\mathcal{P}$ s.t.* $\texttt{GRaSP}_2\texttt{r}(\mathcal{P}) \subset \texttt{uFr}(\mathcal{P})$.

Proof. The example below is one of the uDAGs studied in Table 1 in Section **5.1** where GRaSP$_2$ fails to return one of the sparsest permutations under u-frugality. Given $\mathbf{V} = \{X_1, ..., X_5\}$, consider the unfaithful model $(\mathcal{G}^*, \mathcal{P})$ where the true DAG $\mathcal{G}^*$ is shown on the left in Figure 3, and $\texttt{I}(\mathcal{P}) = \Phi \cup \Psi$ where Φ is the set of faithful CI relations and Ψ is the set of unfaithful CI relations as listed below:

$$\begin{aligned}\Phi = \{&\phi_1 = \langle X_1, X_2 \,|\, \varnothing\rangle, \phi_2 = \langle X_1, X_2 \,|\, \{X_3\}\rangle,\\ &\phi_3 = \langle X_2, X_3 \,|\, \varnothing\rangle, \phi_4 = \langle X_2, X_3 \,|\, \{X_1\}\rangle,\\ &\phi_5 = \langle X_2, X_5 \,|\, \{X_1, X_3, X_4\}\rangle\};\\ \Psi = \{&\psi_1 = \langle X_1, X_5 \,|\, \varnothing\rangle\}.\end{aligned}$$

For every $\mathcal{G} \in \texttt{CMC}(\mathcal{P})$ where $\psi_1 \in \texttt{I}(\mathcal{G})$, we have $|\texttt{E}(\mathcal{G})| > |\texttt{E}(\mathcal{G}^*)| = 7$. Also, all 7-edge Markovian DAGs are in the same MEC and there exists no sparser Markovian DAG. Hence, u-frugality is satisfied s.t. $\texttt{uFr}(\mathcal{P}) \neq \varnothing$.

Next, consider feeding the initial permutation $\pi = \langle 5, 1, 3, 4, 2\rangle$ to unbounded GRaSP$_2$. It will return the same π after the DFS procedure and the induced $\mathcal{G}_\pi$, as shown on the right in Figure 3, contains 8 edges. Therefore, unbounded GRaSP$_2$ fails to return one of the sparsest permutations under some initial permutation and $\texttt{GRaSP}_2\texttt{r}(\mathcal{P}) = \varnothing$. □

Figure 3: An unfaithful model satisfying u-frugality. The true DAG $\mathcal{G}^*$ is shown on the left where the two shaded vertices indicate the unfaithful marginal independence $X_1 \perp\!\!\!\perp_{\mathcal{P}} X_5 \,|\, \varnothing$. Unbounded GRaSP$_2$ returns its initial permutation $\pi = \langle 5, 1, 3, 4, 2 \rangle$. The induced DAG $\mathcal{G}_\pi$ is shown on the right with 8 edges. Hence, GRaSP$_2$ is not correct under u-frugality alone.

Corollary 4.9 *Given a graphoid $\mathcal{P}$, unbounded GRaSP$_2$ is correct under a strictly weaker causal razor than unbounded GRaSP$_1$, which is correct under a strictly weaker causal razor than unbounded GRaSP$_0$.*

E GROW-SHRINK ALGORITHM AND ITS PROPERTIES

Definition E.1 *Given an observational dataset $\mathcal{D}$ with n i.i.d. observations from a joint probability distribution $\mathcal{P}$ over $\mathbf{V}$ that belongs to a curved exponential family[6], for every $X \in \mathbf{V}$ and every $\mathbf{M} \subseteq \mathbf{V} \setminus X$,*

$$\texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}) = \ell_{X|\mathbf{M}}(\hat{\theta}_{\texttt{mle}} \mid \mathcal{D}) + c\,\frac{|\hat{\theta}_{\texttt{mle}}|}{2}\log(n)$$

where $\ell_{X|\mathbf{M}}$ is the conditional log likelihood function, $|\hat{\theta}_{\texttt{mle}}|$ is the absolute value of the maximum likelihood estimate, and c is a multiplier for the parameter penalty.

BIC score is a *decomposable* scoring function in the sense that the BIC score of any DAG $\mathcal{G}$ (over the same set of variables $\mathbf{V}$ as the observational dataset $\mathcal{D}$), denoted as $\texttt{BIC}_{\mathcal{D}}(\mathcal{G})$, satisfies the following:

$$\texttt{BIC}_{\mathcal{D}}(\mathcal{G}) = \sum_{i \in \mathbf{v}} \texttt{BIC}_{\mathcal{D}}(X_i, \mathbf{X}_{\texttt{Pa}(i,\mathcal{G})}).$$

In addition, since we will be using BIC throughout this appendix, we assume that every joint probability distribution $\mathcal{P}$ belongs to a curved exponential family in this section.

Algorithm 1: GROW: *grow*$(\mathcal{D}, X, \mathbf{Z})$

Input: (a) $\mathcal{D}$: an observational dataset over $\mathbf{V}$; (b) $X \in \mathbf{V}$; (c) $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X\}$
Output: $\mathbf{M}_{gr} \subseteq \mathbf{Z}$

```
s ← BIC_D(X, ∅)
s' ← s
M_gr ← ∅
do
    s ← s'
    s' ← max_{Y ∈ Z \ M_gr} BIC_D(X, M_gr ∪ {Y})
    Y' ← argmax_{Y ∈ Z \ M_gr} BIC_D(X, M_gr ∪ {Y})
    if s' > s then
        M_gr ← M_gr ∪ {Y'}
while s' > s
return M_gr
```

[6] See [Kass and Vos, 2011] for an in-depth analysis of curved exponential families.

Algorithm 2: SHRINK: *shrink*$(\mathcal{D}, X, \mathbf{Z})$

Input: (a) $\mathcal{D}$: an observational dataset over $\mathbf{V}$; (b) $X \in \mathbf{V}$; (c) $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X\}$
Output: (i) $\mathbf{M}_{sh} \subseteq \mathbf{Z}$; (ii) $s = \texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}_{sh})$

$s \leftarrow \texttt{BIC}_{\mathcal{D}}(X, \mathbf{Z})$
$s' \leftarrow s$
$\mathbf{M}_{sh} \leftarrow \mathbf{Z}$
do
  $s \leftarrow s'$
  $s' \leftarrow \max_{Y \in \mathbf{M}_{sh}} \texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}_{sh} \setminus \{Y\})$
  $Y' \leftarrow \text{argmax}_{Y \in \mathbf{M}_{sh}} \texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}_{gs} \setminus \{Y\})$
  if $s' > s$ **then**
    $\mathbf{M}_{sh} \leftarrow \mathbf{M}_{sh} \setminus \{Y'\}$
while $s' > s$
return $\mathbf{M}_{sh}, s$

Theorem E.2 *[Chickering, 2002] Given an observational dataset $\mathcal{D}$ with n i.i.d. observations from a joint probability distribution $\mathcal{P}$ over $\mathbf{V}$, consider $\mathcal{G}, \mathcal{G}' \in \texttt{DAG}(\mathbf{V})$ where $\mathcal{G}'$ is resulted from adding the edge $j \to k$ in $\mathcal{G}$. In the large sample limit of n,*

(a) if $X_j \not\perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\texttt{Pa}(k,\mathcal{G})}$, then $\texttt{BIC}_{\mathcal{D}}(\mathcal{G}') > \texttt{BIC}_{\mathcal{D}}(\mathcal{G})$;

(b) if $X_j \perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\texttt{Pa}(k,\mathcal{G})}$, then $\texttt{BIC}_{\mathcal{D}}(\mathcal{G}') < \texttt{BIC}_{\mathcal{D}}(\mathcal{G})$.

The theorem above is known as the *local consistency* of BIC score over DAGs. We can easily derive a lemma which concerns the BIC score of a variable (relative to a set of variables).

Lemma E.3 *Given an observational dataset $\mathcal{D}$ with n i.i.d. observations from a joint probability distribution $\mathcal{P}$ over $\mathbf{V}$, consider any distinct $j, k \in \mathbf{v}$ and $\mathbf{i} \subseteq \mathbf{v} \setminus \{j, k\}$. In the large sample limit of n,*

(a) if $X_j \not\perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\mathbf{i}}$, then $\texttt{BIC}_{\mathcal{D}}(X_k, \mathbf{X}_{\mathbf{i}} \cup \{X_j\}) > \texttt{BIC}_{\mathcal{D}}(X_k, \mathbf{X}_{\mathbf{i}})$;

(b) if $X_j \perp\!\!\!\perp_{\mathcal{P}} X_k \,|\, \mathbf{X}_{\mathbf{i}}$, then $\texttt{BIC}_{\mathcal{D}}(X_k, \mathbf{X}_{\mathbf{i}} \cup \{X_j\}) < \texttt{BIC}_{\mathcal{D}}(X_k, \mathbf{X}_{\mathbf{i}})$.

Proof. Construct a DAG $\mathcal{G} \in \texttt{DAG}(\mathbf{V})$ by drawing all and only directed edges from each vertex in $\mathbf{i}$ to k, and another DAG $\mathcal{G}' \in \texttt{DAG}(\mathbf{V})$ by adding $j \to k$ in $\mathcal{G}$. Then the lemma immediately follows from **Theorem E.2** and the decomposable feature of BIC scores. □

Lemma E.4 *Consider an observational dataset $\mathcal{D}$ with n i.i.d. observations from a compositional graphoid $\mathcal{P}$ over $\mathbf{V}$. In the large sample limit of n, for any $X \in \mathbf{V}$ and any $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X\}$, $\texttt{MB}(X, \mathbf{Z}) \subseteq \mathbf{M}_{gr}$ where $\mathbf{M}_{gr} = \textit{grow}(\mathcal{D}, X, \mathbf{Z}) \subseteq \mathbf{Z}$.*

Proof. First, **Algorithm 1** requires that $\mathbf{M}_{gr} \subseteq \mathbf{Z}$, and $\texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}_{gr} \cup \{Y\}) < \texttt{BIC}_{D}(X, \mathbf{M}_{gr})$ for every $Y \in \mathbf{Z} \setminus \mathbf{M}_{gr}$. By **Lemma E.3**, we have $X \perp\!\!\!\perp_{\mathcal{P}} Y \,|\, \mathbf{M}_{gr}$ for each $Y \in \mathbf{Z} \setminus \mathbf{M}_{gr}$. By composition, we have $X \perp\!\!\!\perp_{\mathcal{P}} (\mathbf{Z} \setminus \mathbf{M}_{gr}) \,|\, \mathbf{M}_{gr}$. Therefore, by **Definition A.1** and **Lemma A.2**, we have $\texttt{MB}(X, \mathbf{Z}) \subseteq \mathbf{M}_{gr}$. □

Lemma E.5 *Consider an observational dataset $\mathcal{D}$ with n i.i.d. observations from a graphoid $\mathcal{P}$ over $\mathbf{V}$. In the large sample limit of n, for any $X \in \mathbf{V}$ and any $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X\}$, $\texttt{MB}(X, \mathbf{Z}) = \mathbf{M}_{sh}$ where $\mathbf{M}_{sh} = \textit{shrink}(\mathcal{D}, X, \mathbf{Z}) \subseteq \mathbf{Z}$.*

Proof. We show the lemma by $\mathbf{M}_{sh} \subseteq \texttt{MB}(X, \mathbf{Z})$ and $\mathbf{M}_{sh} \supseteq \texttt{MB}(X, \mathbf{Z})$.

[$\subseteq$] By reductio, suppose that there exists $Y \in \mathbf{M}_{sh} \subseteq \mathbf{Z}$ but $Y \notin \texttt{MB}(X, \mathbf{Z})$. Let $\mathbf{S}$ be $\mathbf{M}_{sh} \setminus \{Y\}$. **Algorithm 2** requires that $\texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}_{sh} \setminus \{Y\}) < \texttt{BIC}_{D}(X, \mathbf{M}_{sh})$. In other words, we have $\texttt{BIC}_{D}(X, \mathbf{S}) < \texttt{BIC}_{\mathcal{D}}(X, \mathbf{S} \cup \{Y\})$. By **Lemma E.3**, we have $X \not\perp\!\!\!\perp_{\mathcal{P}} Y \,|\, \mathbf{S}$.

Let $\mathbf{W} = \mathbf{S} \setminus \texttt{MB}(X, \mathbf{Z})$. From $Y \notin \texttt{MB}(X, \mathbf{Z})$ and $Y \notin \mathbf{S}$, we have $\{Y\} \cup \mathbf{W} \subseteq \mathbf{Z} \setminus \texttt{MB}(X, \mathbf{Z})$. Recall **Definition A.1** that $X \perp\!\!\!\perp_{\mathcal{P}} \mathbf{Z} \setminus \texttt{MB}(X, \mathbf{Z}) \,|\, \texttt{MB}(X, \mathbf{Z})$. Thus,

$$X \perp\!\!\!\perp_{\mathcal{P}} \{Y\} \cup \mathbf{W} \,|\, \texttt{MB}(X, \mathbf{Z}) \qquad \because X \perp\!\!\!\perp_{\mathcal{P}} \mathbf{Z} \setminus \texttt{MB}(X, \mathbf{Z}) \,|\, \texttt{MB}(X, \mathbf{Z}),\ \textit{decomposition} \tag{12}$$

$$X \perp\!\!\!\perp_{\mathcal{P}} Y \,|\, \texttt{MB}(X, \mathbf{Z}) \cup \mathbf{W} \qquad \because (12),\ \textit{weak union} \tag{13}$$

$$X \perp\!\!\!\perp_{\mathcal{P}} Y \,|\, \mathbf{S} \qquad \because (13), \mathbf{W} = \mathbf{S} \setminus \texttt{MB}(X, \mathbf{Z}) \tag{14}$$

Contradiction arises with $X \not\perp\!\!\!\perp_{\mathcal{P}} Y \,|\, \mathbf{S}$.

[$\supseteq$] Observe that **Algorithm 2** removes one variable in $\mathbf{Z}$ one at a time repeatedly to form $\mathbf{M}_{sh}$. Thus, the shrink-procedure corresponds to a sequence of sets of variables $\langle \mathbf{M}^0, ..., \mathbf{M}^k \rangle$ and a sequence of variables $\mathbf{W} = \langle W_1, ..., W_k \rangle = \mathbf{Z} \setminus \mathbf{M}_{sh}$ such that $\mathbf{M}^0 = \mathbf{Z}$, $\mathbf{M}^k = \mathbf{M}_{sh}$, and $\mathbf{M}^i = \mathbf{M}^{i-1} \setminus \{W_i\}$ (where $W_i \in \mathbf{M}^{i-1}$) for each $1 < i \leq k$.

Notice that $\mathbf{M}^{i-1} = \mathbf{M}^i \cup \{W_i\}$. **Algorithm 2** requires that $\texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}^i) > \texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}^{i-1}) = \texttt{BIC}_{D}(X, \mathbf{M}^i \cup \{W_i\})$. We then have

$$X \perp\!\!\!\perp_{\mathcal{P}} W_1 \,|\, \mathbf{M}^1 \qquad \because \texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}^1) > \texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}^0), \textbf{Lemma E.3} \tag{15}$$

$$X \perp\!\!\!\perp_{\mathcal{P}} W_2 \,|\, \mathbf{M}^2 \qquad \because \texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}^2) > \texttt{BIC}_{\mathcal{D}}(X, \mathbf{M}^1), \textbf{Lemma E.3} \tag{16}$$

$$X \perp\!\!\!\perp_{\mathcal{P}} W_1 \,|\, \mathbf{M}^2 \cup \{W_2\} \qquad \because (15), \mathbf{M}^1 = \mathbf{M}^2 \cup \{W_2\} \tag{17}$$

$$X \perp\!\!\!\perp_{\mathcal{P}} \{W_1, W_2\} \,|\, \mathbf{M}^2 \qquad \because (16), (17), \textit{contraction} \tag{18}$$

$$\vdots$$

$$X \perp\!\!\!\perp_{\mathcal{P}} \{W_1, ..., W_k\} \,|\, \mathbf{M}^k \qquad \because ..., \textit{contraction} \tag{19}$$

$$X \perp\!\!\!\perp_{\mathcal{P}} \mathbf{W} \,|\, \mathbf{M}_{sh} \qquad \because (19), \mathbf{W} = \langle W_1, ..., W_k \rangle \text{ and } \mathbf{M}^k = \mathbf{M}_{sh} \tag{20}$$

$$X \perp\!\!\!\perp_{\mathcal{P}} \mathbf{Z} \setminus \mathbf{M}_{sh} \,|\, \mathbf{M}_{sh} \qquad \because (20), \mathbf{W} = \mathbf{Z} \setminus \mathbf{M}_{sh} \tag{21}$$

Hence, it follows from **Definition A.1** that $\mathbf{M}_{sh} \supseteq \texttt{MB}(X, \mathbf{Z})$. □

Theorem E.6 *Consider an observational dataset $\mathcal{D}$ with n i.i.d. observations from a compositional graphoid $\mathcal{P}$ over $\mathbf{V}$. In the large sample limit of n, for any $X \in \mathbf{V}$ and any $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X\}$, $\texttt{MB}(X, \mathbf{Z}) = \mathbf{M}_{gs}$ where $\mathbf{M}_{gs} =$ shrink($\mathcal{D}, X,$ grow($\mathcal{D}, X, \mathbf{Z}$)).*

Proof. Immediate from **Lemma E.4** and **Lemma E.5**. □

Theorem E.7 *Consider an observational dataset $\mathcal{D}$ with n i.i.d. observations from a (compositional) graphoid $\mathcal{P}$ over $\mathbf{V} = \{X_1, ..., X_m\}$, and any $\pi \in \Pi(\mathbf{v})$. Let s_i and $\mathbf{M}_i$ be the score and the set of variables returned by shrink($\mathcal{D}, X_i, \mathbf{X}_{\texttt{Pre}(i,\pi)}$) (or shrink($\mathcal{D}, X_i,$ grow($\mathcal{D}, X_i, \mathbf{X}_{\texttt{Pre}(i,\pi)}$)) if $\mathcal{P}$ is a compositional graphoid) respectively. Denote s_π as $\sum_{i \in \mathbf{v}} s_i$. In the large sample limit of n, $\texttt{BIC}_{\mathcal{D}}(\mathcal{G}_\pi) = s_\pi$ where $\mathcal{G}_\pi$ is induced from π by (VP).*

Proof. Immediate from the decomposable feature of BIC scores, **Lemma E.5** and **Theorem E.6**. □

Lastly, though a compositional graphoid is a sufficient condition for the correct identification of the unique Markov boundary using the grow-shrink algorithm, we are aware of an assumption weaker than compositional graphoid to validate such an identification. Nevertheless, this discussion will be beyond the scope of this paper and we will leave the formal proof to future work.

F ADDITIONAL EXAMPLES

F.1 LU ET AL. COMPARISON

Reported below are average statistics obtained by running GRaSP$_2$ on the published datasets used to generate Figure 6 in [Lu et al., 2021][7]. We cannot compare these results to Lu et al. precisely, since their statistics are given in figures and not

[7] https://github.com/ninalu/urlearning-cpp/tree/master/triplet_data

exactly in tables, though judging from their figures it appears that GRaSP$_2$ is dominating for adjacency precision and recall, arrowhead recall, and most results for arrowhead precision. Timing results are not reported by Lu et al.; we include these to show that GRASP$_2$ returns quickly for all of these examples, where we know (personal communication) that some of the results for Triple A* take much longer. Adjacencies in these graphs are sampled with uniform probability, "Edge-prob".

Edge-prob	0.03	0.04	0.05	0.06	0.07	0.08
Precision	0.964	0.976	0.979	0.980	0.982	0.976
Recall	0.985	0.982	0.986	0.986	0.985	0.985
F1	0.974	0.979	0.983	0.983	0.983	0.980

Table 1: GRaSP$_2$ Adjacency Statistics

Edge-prob	0.03	0.04	0.05	0.06	0.07	0.08
Precision	0.907	0.914	0.933	0.949	0.946	0.945
Recall	0.897	0.916	0.933	0.952	0.952	0.955
F1	0.898	0.913	0.932	0.950	0.948	0.950

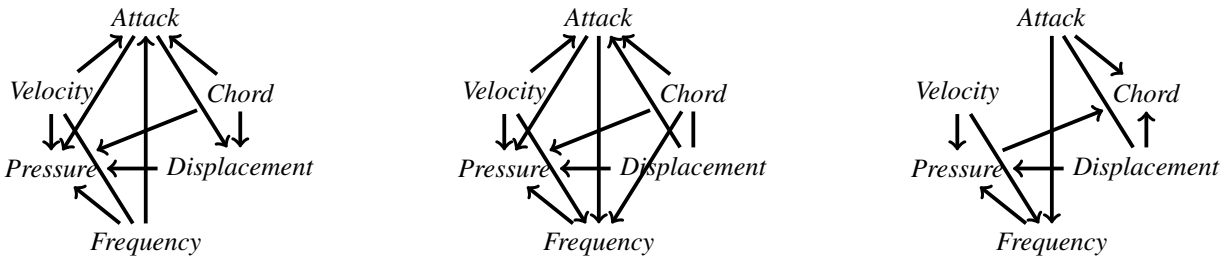
Table 2: GRaSP$_2$ Arrowhead Statistics

Edge-prob	0.03	0.04	0.05	0.06	0.07	0.08
Seconds	0.405	0.755	1.403	2.703	4.795	7.161

Table 3: GRaSP$_2$ Timing Statistics

F.2 AIRFOIL EXAMPLE

Figure 4 gives the results of running GRaSP$_2$, PC, and fGES on the Airfoil empirical example described in Section **6**. GRaSP$_2$ gets the same uniquely frugal result as SP. To improve readability, we use the names of the variables (instead of numerals) to label the vertices.



(a) GRaSP$_2$ result (b) fGES result (c) PC result

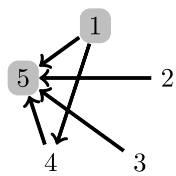
Figure 4: Results of algorithms on NASA airfoil experiment.

Note that both the GRaSP$_2$ and FGES results use the linear, Gaussian BIC score with a penalty multiplier of 2. For the GRaSP$_2$ result in (a), *Attack* is not exogenous, which is counter-intuitive, since it is experimentally controlled. Allowing for latent variables could resolve this issues. However, we leave the development of such an algorithm to future work. On the other hand, the FGES result in (b) is notably not the same as the SP result and so is not frugal. Also, the orientation between *Attack* and *Displacement* is reversed.

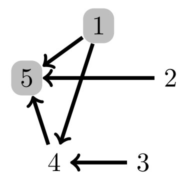
The PC result in (c), which uses the zero partial correlation test with a significance level of 0.01, in fact has fewer edges than the frugal result and makes *Chord*, another experimental variable, endogenous. Causally, PC is giving incorrect and incomplete information.

G UNIT TESTS

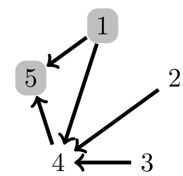
We consider path cancellations in DAGs between pairs of vertices, one of which is exogenous, connected by two or more unique treks. Furthermore, the path cancellations we consider elicit a marginal independence between the two vertices in question. Below, we enumerate all possible path cancellations of this type (up to vertex relabeling). Each graph illustrates a case where an unfaithful marginal independence is elicited between the two gray vertices due to path cancellation. A complete list of all unfaithful CI relations (symmetry assumed) where the independent sets are singletons is also provided for each graph.



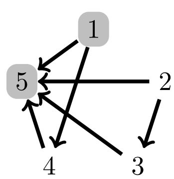
$X_1 \perp\!\!\!\perp X_5 \mid X_2, X_3$
$X_1 \perp\!\!\!\perp X_2 \mid X_3, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_3$
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



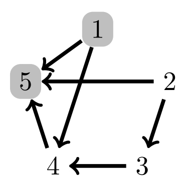
$X_1 \perp\!\!\!\perp X_5 \mid X_2, X_3$
$X_1 \perp\!\!\!\perp X_2 \mid X_3, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_3$
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



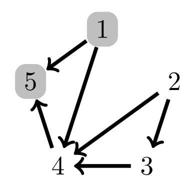
$X_1 \perp\!\!\!\perp X_5 \mid X_2, X_3$
$X_1 \perp\!\!\!\perp X_2 \mid X_3, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_3$
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



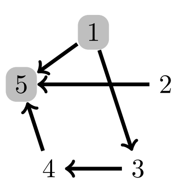
$X_1 \perp\!\!\!\perp X_5 \mid X_2, X_3$
$X_1 \perp\!\!\!\perp X_2 \mid X_3, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_3$
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



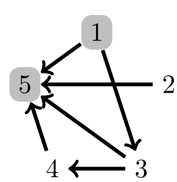
$X_1 \perp\!\!\!\perp X_5 \mid X_2, X_3$
$X_1 \perp\!\!\!\perp X_2 \mid X_3, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_3$
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



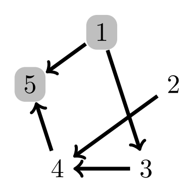
$X_1 \perp\!\!\!\perp X_5 \mid X_2, X_3$
$X_1 \perp\!\!\!\perp X_2 \mid X_3, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_5$
$X_1 \perp\!\!\!\perp X_3 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_3$
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



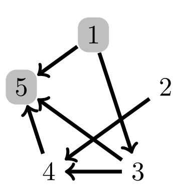
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



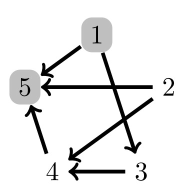
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



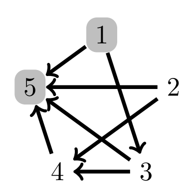
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



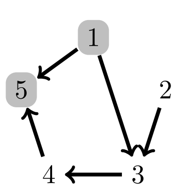
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



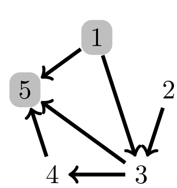
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



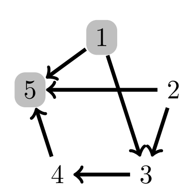
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



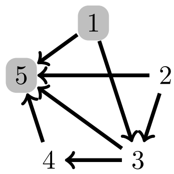
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



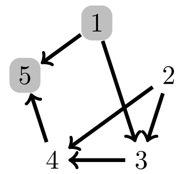
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



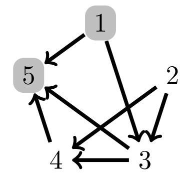
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



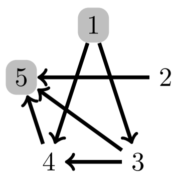
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



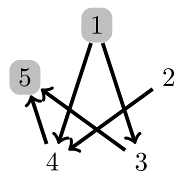
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



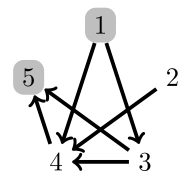
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



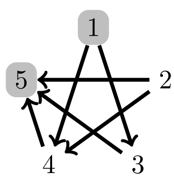
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



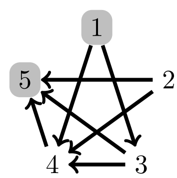
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



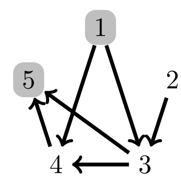
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



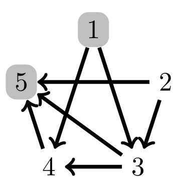
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



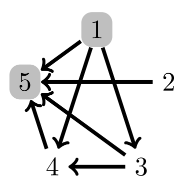
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



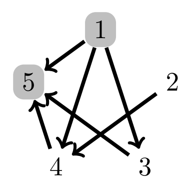
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



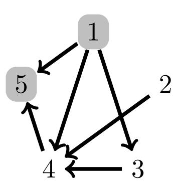
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



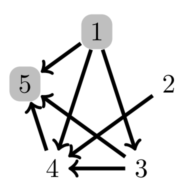
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



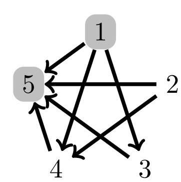
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



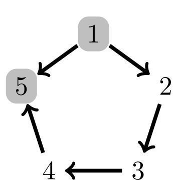
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



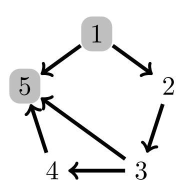
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



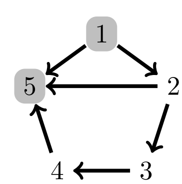
$X_1 \perp\!\!\!\perp X_5$
$X_1 \perp\!\!\!\perp X_2 \mid X_5$
$X_1 \perp\!\!\!\perp X_5 \mid X_2$



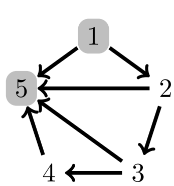
$X_1 \perp\!\!\!\perp X_5$



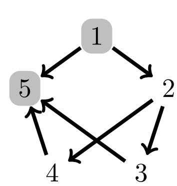
$X_1 \perp\!\!\!\perp X_5$



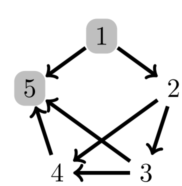
$X_1 \perp\!\!\!\perp X_5$



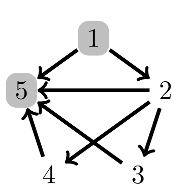
$X_1 \perp\!\!\!\perp X_5$



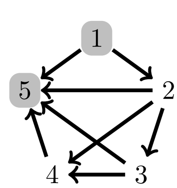
$X_1 \perp\!\!\!\perp X_5$



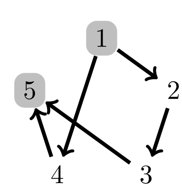
$X_1 \perp\!\!\!\perp X_5$



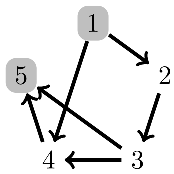
$X_1 \perp\!\!\!\perp X_5$



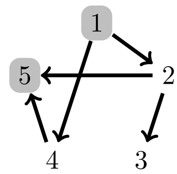
$X_1 \perp\!\!\!\perp X_5$



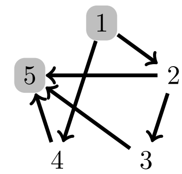
$X_1 \perp\!\!\!\perp X_5$



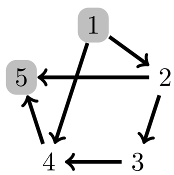
$X_1 \perp\!\!\!\perp X_5$



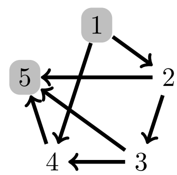
$X_1 \perp\!\!\!\perp X_5$



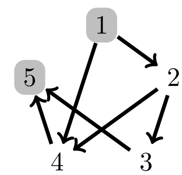
$X_1 \perp\!\!\!\perp X_5$



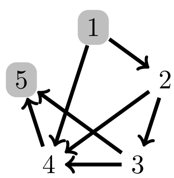
$X_1 \perp\!\!\!\perp X_5$



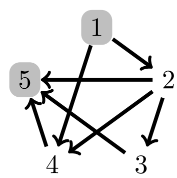
$X_1 \perp\!\!\!\perp X_5$



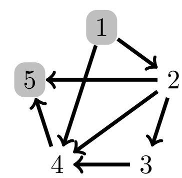
$X_1 \perp\!\!\!\perp X_5$



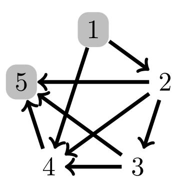
$X_1 \perp\!\!\!\perp X_5$



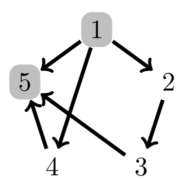
$X_1 \perp\!\!\!\perp X_5$



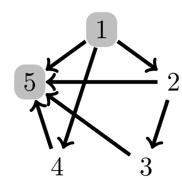
$X_1 \perp\!\!\!\perp X_5$



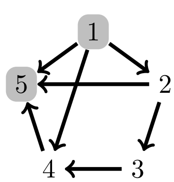
$X_1 \perp\!\!\!\perp X_5$



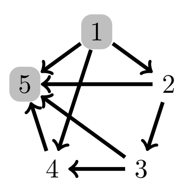
$X_1 \perp\!\!\!\perp X_5$



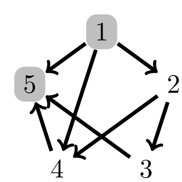
$X_1 \perp\!\!\!\perp X_5$



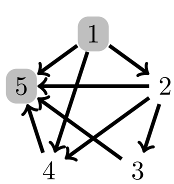
$X_1 \perp\!\!\!\perp X_5$



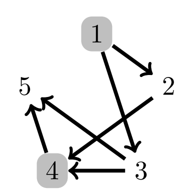
$X_1 \perp\!\!\!\perp X_5$



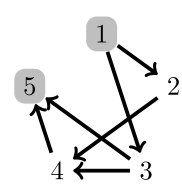
$X_1 \perp\!\!\!\perp X_5$



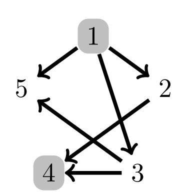
$X_1 \perp\!\!\!\perp X_5$



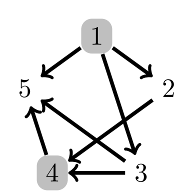
$X_1 \perp\!\!\!\perp X_4$



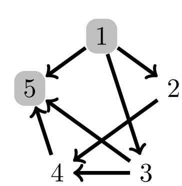
$X_1 \perp\!\!\!\perp X_5$



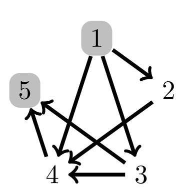
$X_1 \perp\!\!\!\perp X_4$



$X_1 \perp\!\!\!\perp X_4$



$X_1 \perp\!\!\!\perp X_5$



$X_1 \perp\!\!\!\perp X_5$

References

David Maxwell Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, page 87–98, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3: 507–554, 2002.

Malcolm Forster, Garvesh Raskutti, Reuben Stern, and Naftali Weinberger. The frugal inference of causal relations. *The British Journal for the Philosophy of Science*, 2020.

Robert E Kass and Paul W Vos. *Geometrical Foundations of Asymptotic Inference*. John Wiley & Sons, Hoboken, NJ, 2011.

Ni Y Lu, Kun Zhang, and Changhe Yuan. Improving causal discovery by optimal Bayesian network learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8741–8748, 2021.

Fatemeh Mohammadi, Caroline Uhler, Charles Wang, and Josephine Yu. Generalized permutohedra from probabilistic graphical models. *SIAM Journal on Discrete Mathematics*, 32:64–93, 2018.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

Joseph Ramsey, Jiji Zhang, and Peter Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the conference on Uncertainty in artificial intelligence*, pages 401–408, 2006.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7, 2018.

Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108:795–814, 2021.

Wolfgang Spohn. On the properties of conditional independence. In Paul Humphreys, editor, *Patrick Suppes: Scientific Philosopher: Volume 1. Probability and Probabilistic Causality*, pages 173–196. Springer Netherlands, 1994.

Milan Studený. *Probabilistic Conditional Independence Structures*. Information Science and Statistics. Springer, 2005.

Thomas Verma and Judea Pearl. Causal networks: semantics and expressiveness. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 69–78, 1988.

Jiji Zhang. A comparison of three Occam's razors for Markovian causal models. *The British journal for the philosophy of science*, 64:423–448, 2013.

Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.