Supporting Information for

Model Selection over partially ordered sets

Armeen Taeb, Peter Bühlmann and Venkat Chandrasekaran

Armeen Taeb
E-mail:ataeb@uw.edu

This PDF file includes:

Supporting text
Fig. S1
SI References

Supporting Information Text

Subhead. This document includes proofs of claims in the main text.

I. Meet Semi-lattice and Join Semi-lattice Properties and Posets in Examples 1-9

The Boolean poset (Example 1), partition poset (Examples 2-3), integer poset (Example 5), permutation poset (Example 7), and subspace poset (Example 8) are all known in the literature to be lattices (and consequently meet-semi and join semi-lattices); see (1).

We next show that for Examples 6 and 9 associated with partial ranking and blind-source separation, the corresponding posets are also meet semi-lattices. Consider the partial ranking setting in Example 6. Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be two relations that are irreflexive, asymmetric, and transitive. Recalling that the partial ordering is based on inclusion, it is clear that the relations $\mathcal{R} = \{(a,b) : (a,b) \in \mathcal{R}_1, (a,b) \in \mathcal{R}_2\}$ is the unique largest rank element in the partial ranking poset such that $\mathcal{R} \preceq \mathcal{R}_1$ and $\mathcal{R} \preceq \mathcal{R}_2$. Furthermore, for any $\tilde{\mathcal{R}}$ with $\tilde{\mathcal{R}} \preceq \mathcal{R}_1$ and $\tilde{\mathcal{R}} \preceq \mathcal{R}_2$, we clearly have that $\tilde{\mathcal{R}} \preceq \mathcal{R}$. Consider the blind-source separation setting in Example 9. Let x_1 and x_2 be two sets of linearly independent subsets of unit norm vectors. Recalling that the partial ordering in the associated poset is based on inclusion, it is clear that the set $y = x_1 \cap x_2$ is the unique largest rank element in the partial ranking poset such that $y \preceq x_1$ and $y \preceq x_2$. Furthermore, for every z with $z \preceq x_1$ and $z \preceq x_2$, we have that $z \preceq y$.

We show that the poset corresponding to causal structure learning setting (Example 4) is not meet semi-lattice or join semi-lattice. As a counterexample, consider the CPDAGs $\mathcal{C}_i$ for $i = 1, 2, 3, 4$ shown in Figure S1. Notice that $\mathcal{C}_3 \preceq \mathcal{C}_1$, $\mathcal{C}_3 \preceq \mathcal{C}_2$, $\mathcal{C}_4 \preceq \mathcal{C}_1$, and $\mathcal{C}_4 \preceq \mathcal{C}_2$. Notice also that $\mathcal{C}_3$ and $\mathcal{C}_4$ are both CPDAGs with the largest rank that are smaller (in a partial order sense) than $\mathcal{C}_1$ and $\mathcal{C}_2$. We thus can conclude that the poset is not meet semi-lattice. Similarly, $\mathcal{C}_1$ and $\mathcal{C}_2$ are both CPDAGs with the smallest rank that are larger (in a partial order sense) than $\mathcal{C}_3$ and $\mathcal{C}_4$. We thus can conclude that the poset is not join semi-lattice.

We next show that the poset for Example 6 is not join semi-lattice with a simple counterexample. Consider as an example elements $x_1 = \{(1,2)\}$ and $x_2 = \{(2,1)\}$. Note that there does not exist an element z such that $x_1 \preceq z$ and $x_2 \preceq z$. Thus, the poset is not join semi-lattice.

Finally, we show that the poset corresponding to blind-source separation (Example 9) is not join semi-lattice. Consider a collection of $p+1$ rank-1 elements in this poset, each element consisting of a single p dimensional vector. Then, evidently, there cannot exist an element z consisting of a set of vectors that contains all of the vectors in the rank-1 elements, while satisfying the linear independence condition.

II. Proof that Eq. (1) is a Similarity Valuation Function

Recall that

$$\rho_{\text{meet}}(x, y) = \max_{z \preceq x, z \preceq y} \text{rank}(z). \qquad [14]$$

By definition, $\rho_{\text{meet}}(\cdot,\cdot)$ is a symmetric function. We will now show that it satisfies the three properties in Definition 1 for any pair of elements $x, y \in \mathcal{L}$. For the first property, we can conclude $\rho_{\text{meet}}(x, y) \geq 0$ since by definition, the rank function returns a non-negative integer for all the elements in the poset. Again, because of the property of the rank function in a graded poset, a feasible z (satisfying the constraints $z \preceq x$, $z \preceq y$) will necessarily have $\text{rank}(z) \leq \min\{\text{rank}(x), \text{rank}(y)\}$. For the second property, consider any $w \in \mathcal{L}$ with $x \preceq w$. Note that:

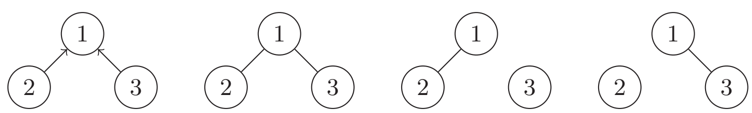
$$\rho_{\text{meet}}(w, y) = \max_{z \preceq w, z \preceq y} \text{rank}(z). \qquad [15]$$

Then, any feasible z in Eq. (14) is also feasible in Eq. (15) by the transitive property of posets. Therefore, $\rho_{\text{meet}}(x, y) \leq \rho_{\text{meet}}(w, y)$. For the third property, first note that if $x \preceq y$, then $z = x$ is feasible in Eq. (14) and thus $\rho_{\text{meet}}(x, y) \geq \text{rank}(x)$. Since also $\rho_{\text{meet}}(x, y) \leq \text{rank}(x)$ by the second property of similarity valuations, we have that $\rho_{\text{meet}}(x, y) = \text{rank}(x)$. Now suppose that $\rho_{\text{meet}}(x, y) = \text{rank}(x)$. By Eq. (14), we conclude that there exists a feasible z ($z \preceq x$, $z \preceq y$) such that $\text{rank}(z) = \text{rank}(x)$. By the property of the rank function, we have that if $\text{rank}(z) = \text{rank}(x)$ and $z \preceq x$, then $z = x$. Since we have additionally that $z \preceq y$, we conclude that $x \preceq y$.

III. Proof of Lemmas 8-9

Proof of Lemma 8. Recall the telescoping sum decomposition Eq. (5) that $\text{FD}(x_k, x^\star) = \sum_{i=1}^{k} 1 - [f(x_{i-1}, x_i; x^\star)]$. From the first property of similarity valuation that it yields non-negative values, second property of similarity valuation that $\rho(x, y) \leq \rho(z, y)$ for $x \preceq z$, and that the ρ is an integer-valued similarity valuation, we have that $\text{FD}(x, x^\star) \leq \sum_{i=1}^{k} \mathbb{I}[(x_{i-1}, x_i) \in \mathcal{T}_{\text{null}}]$. □

Proof of Lemma 9. For any covering pairs (x, y) and (u, v) with $v \preceq x$, we cannot have that $f(x, y; z) = f(u, v; z)$ for all $z \in \mathcal{L}$. Suppose as a point of contradiction that for every $z \in \mathcal{L}$, $f(x, y; z) = f(u, v; z)$. Let $z = v$. Then, by the third property of a similarity valuation (see Definition 1), $\rho(u, z) = \text{rank}(u)$ and $\rho(v, z) = \text{rank}(v)$; thus, for this choice of z, $f(u, v; z) = 1$. On the other hand, again by the third property of a similarity valuation and for the choice of $z = v$, since $u \preceq v \preceq x \preceq y$, $\rho(x, z) = \rho(y, z) = \text{rank}(v)$ and thus $f(x, y; z) = 0$. □



(a) $\mathcal{C}_1$ **(b)** $\mathcal{C}_2$ **(c)** $\mathcal{C}_3$ **(d)** $\mathcal{C}_4$

Fig. S1. Four CPDAGs. Here, CPDAGs $\mathcal{C}_3$ and $\mathcal{C}_4$ are both largest complexity models that are smaller (in partial order sense) than $\mathcal{C}_1$ and $\mathcal{C}_2$. Similarly, CPDAGs $\mathcal{C}_1$ and $\mathcal{C}_2$ are the smallest complexity models that are larger (in a partial order sense) than $\mathcal{C}_3$ and $\mathcal{C}_4$.

IV. Analysis in the Continuous Examples 8 and 9

For notational ease, we let $\hat{x}_{\text{base}}^{(\ell)} = \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})$. Notice that for any $l = 1, 2, \dots, B$:

$$\begin{aligned}\text{FD}(\hat{x}_{\text{stable}}, x^\star) &= \text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, x^\star)\\ &= \left[\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)})\right] + \left[\text{rank}(\hat{x}_{\text{base}}^{(\ell)}) - \rho(\hat{x}_{\text{base}}^{(\ell)}, x^\star)\right] + \kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}),\end{aligned}$$

where

$$\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}) := \rho(\hat{x}_{\text{base}}^{(\ell)}, x^\star) - \text{rank}(\hat{x}_{\text{base}}^{(\ell)}) + \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)}) - \rho(\hat{x}_{\text{stable}}, x^\star).$$

Since the choice of l was arbitrary, we note that:

$$\begin{aligned}\text{FD}(\hat{x}_{\text{stable}}, x^\star) &= \frac{2}{B}\sum_{\ell=1}^{B/2} \min_{t\in\{0,1\}}\left\{\left[\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(2\ell-t)})\right] + \left[\text{rank}(\hat{x}_{\text{base}}^{(2\ell-t)}) - \rho(\hat{x}_{\text{base}}^{(2\ell-t)}, x^\star)\right] + \kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(2\ell-t)})\right\}\\ &\leq \frac{2}{B}\sum_{\ell=1}^{B/2} \min_{t\in\{0,1\}}\left\{\left[\text{rank}(\hat{x}_{\text{base}}^{(2\ell-t)}) - \rho(\hat{x}_{\text{base}}^{(2\ell-t)}, x^\star)\right]\right\} + \frac{2}{B}\sum_{\ell=1}^{B}\left[\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)})\right]\\ &\quad + \frac{2}{B}\sum_{\ell=1}^{B}\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)})\\ &\leq \frac{2}{B}\sum_{\ell=1}^{B/2}\prod_{t\in\{0,1\}}\sqrt{\text{rank}(\hat{x}_{\text{base}}^{(2\ell-t)}) - \rho(\hat{x}_{\text{base}}^{(2\ell-t)}, x^\star)} + 2\alpha\text{rank}(\hat{x}_{\text{stable}}) + \frac{2}{B}\sum_{\ell=1}^{B}\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}).\end{aligned}$$

Here, the second inequality follows from $\min\{a+b, c+d\} \leq \min\{a, c\} + b + d$ for $a, b, c, d \geq 0$. The third inequality follows from $\min\{a, b\} \leq \sqrt{ab}$ for $a, b \geq 0$ and

$$\frac{1}{B}\sum_{\ell=1}^{B}\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)}) = \sum_{k=1}^{\text{rank}(\hat{x}_{\text{stable}})}\frac{1}{B}\sum_{\ell=1}^{B} 1 - [\rho(x_k, \hat{x}_{\text{base}}^{(\ell)}) - \rho(x_{k-1}, \hat{x}_{\text{base}}^{(\ell)})] \leq \alpha\text{rank}(\hat{x}_{\text{stable}}), \tag{16}$$

where $(x_0, x_1, \dots, x_{\hat{k}})$ is a sequence specifying a path from the least element x_0 to $x_{\hat{k}} = \hat{x}_{\text{stable}}$ with $\text{rank}(\hat{x}_{\text{stable}}) = \hat{k}$. Thus, $\frac{1}{B}\sum_{\ell=1}^{B}\rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)}) \geq (1-\alpha)\text{rank}(\hat{x}_{\text{stable}})$. As $\rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)}) \leq \text{rank}(\hat{x}_{\text{base}}^{(\ell)})$, we can then conclude that $\mathbb{E}[\text{rank}(\hat{x}_{\text{stable}})] \leq \frac{\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})]}{1-\alpha}$. Taking expectations and using the fact that the data across complementary bags is IID, we obtain:

$$\text{FD}(\hat{x}_{\text{stable}}, x^\star) \leq \mathbb{E}[\sqrt{\text{FD}(\hat{x}_{\text{sub}}, x^\star)}]^2 + \frac{2\alpha}{1-\alpha}\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})] + \frac{2}{B}\sum_{\ell=1}^{B}\mathbb{E}[\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)})].$$

It remains to bound $\frac{2}{B}\sum_{\ell=1}^{B}\mathbb{E}[\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)})]$ for subspace selection and blind-source separation.

Subspace-selection: We will use the similarity valuation $\rho := \rho_{\text{subspace}}$ in Eq. (3). Note that:

$$\begin{aligned}\text{rank}(x) - \rho(x, y) = \text{trace}\left(\mathcal{P}_x\mathcal{P}_{y^\perp}\right) &= \text{trace}\left(\mathcal{P}_x\mathcal{P}_z\mathcal{P}_{y^\perp}\mathcal{P}_z\right) + \text{trace}\left(\mathcal{P}_x\mathcal{P}_{z^\perp}\mathcal{P}_{y^\perp}\mathcal{P}_{z^\perp}\right)\\ &\quad + \text{trace}\left(\mathcal{P}_x\mathcal{P}_{z^\perp}\mathcal{P}_{y^\perp}\mathcal{P}_z\right) + \text{trace}\left(\mathcal{P}_x\mathcal{P}_z\mathcal{P}_{y^\perp}\mathcal{P}_{z^\perp}\right)\\ &\leq \text{trace}\left(\mathcal{P}_{y^\perp}\mathcal{P}_z\right) + \text{trace}\left(\mathcal{P}_x\mathcal{P}_{z^\perp}\right) + \text{trace}\left([\mathcal{P}_x, \mathcal{P}_{z^\perp}]\left[\mathcal{P}_z, \mathcal{P}_{y^\perp}\right]\right)\\ &= \text{rank}(z) - \rho(y, z) + \text{rank}(x) - \rho(x, z) + \text{trace}\left([\mathcal{P}_x, \mathcal{P}_{z^\perp}]\left[\mathcal{P}_z, \mathcal{P}_{y^\perp}\right]\right).\end{aligned} \tag{17}$$

Here, for matrices $A, B \in \mathbb{R}^{p\times p}$, $[A, B] = AB - BA$ represents the commutator. Furthermore, note that:

$$\begin{aligned}\text{trace}\left([\mathcal{P}_x, \mathcal{P}_{z^\perp}]\left[\mathcal{P}_z, \mathcal{P}_{y^\perp}\right]\right) &\leq \|[\mathcal{P}_x, \mathcal{P}_{z^\perp}]\|_\star \|\left[\mathcal{P}_z, \mathcal{P}_{y^\perp}\right]\|_2\\ &\leq 2\sqrt{\text{rank}(x)}\sqrt{\text{rank}(x) - \rho(x, z)}\|[\mathcal{P}_z, \mathcal{P}_y]\|_2.\end{aligned} \tag{18}$$

Combining the bounds Eq. (17) and Eq. (18), we find that:

$$\begin{aligned}\text{rank}(x) - \rho(x, y) &\leq \text{rank}(z) - \rho(y, z) + \text{rank}(x) - \rho(x, z) + 2\sqrt{\text{rank}(x)}\sqrt{\text{rank}(x) - \rho(x, z)}\|[\mathcal{P}_z, \mathcal{P}_y]\|_2\\ &\leq \text{rank}(z) - \rho(y, z) + \text{rank}(x) - \rho(x, z) + \sqrt{\text{rank}(x)}\sqrt{\text{rank}(x) - \rho(x, z)}.\end{aligned}$$

Here, the second inequality follows from the fact that for projection matrices A and B, $\|[A,B]\|_2 \le 1/2$. From this inequality, we conclude that in the subspace selection setting,

$$\begin{aligned}
\frac{1}{B}\sum_{\ell=1}^{B}\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}) &\le \sqrt{\text{rank}(\hat{x}_{\text{stable}})}\frac{1}{B}\sum_{l=1}^{B}\sqrt{\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)})} \\
&\le \sqrt{\text{rank}(\hat{x}_{\text{stable}})}\sqrt{\frac{1}{B}\sum_{l=1}^{B}\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)})} \\
&\le \sqrt{\alpha}\text{rank}(\hat{x}_{\text{stable}}).
\end{aligned}$$

Here, the second equality follows from Cauchy-Schwartz and the last inequality follows from the bound Eq. (16). Recalling that $\mathbb{E}[\text{rank}(\hat{x}_{\text{stable}})] \le \frac{\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})]}{1-\alpha}$, we obtain the final bound:

$$\text{FD}(\hat{x}_{\text{stable}}, x^\star) \le \mathbb{E}[\sqrt{\text{FD}(\hat{x}_{\text{sub}}, x^\star)}]^2 + \frac{2\alpha + \sqrt{\alpha}}{1-\alpha}\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})].$$

Blind-source separation We will use the similarity valuation $\rho := \rho_{\text{source-separation}}$ in Eq. (4). For simplicity of notation, associated with any element $z \in \mathcal{L}$, we consider a block-diagonal $p^2 \times p^2$ projection matrix where each $p \times p$ block is a projection matrix of the subspace spanned by a vector in z. We denote this projection matrix $\mathcal{P}_z$. Then, $\rho(x,y) = \max_{\Pi \in \mathbb{S}_{\text{block}}^{p^2}} \text{trace}\left(\mathcal{P}_x \Pi \mathcal{P}_y \Pi^T\right)$ where $\mathbb{S}_{\text{block}}^{p^2}$ is the space of $p^2 \times p^2$ permutation matrices that are block-diagonal where each block is of size $p \times p$.

Note that:

$$\begin{aligned}
\text{rank}(x) - \rho(x,y) &= \min_{\Pi \in \mathbb{S}_{\text{block}}^{p^2}} \text{trace}\left(\mathcal{P}_x \Pi \mathcal{P}_{y^\perp} \Pi^T\right) \\
&\le \min_{\tilde{\Pi} \in \mathbb{S}_{\text{block}}^{p^2}} \min_{\Pi \in \mathbb{S}_{\text{block}}^{p^2}} \text{trace}\left(\Pi \mathcal{P}_{y^\perp} \Pi^T \tilde{\Pi} \mathcal{P}_z \tilde{\Pi}^T\right) + \text{trace}\left(\mathcal{P}_x \tilde{\Pi} \mathcal{P}_{z^\perp} \tilde{\Pi}^T\right) \\
&\qquad + 2\sqrt{\text{rank}(x)}\sqrt{\text{trace}\left(\mathcal{P}_x \tilde{\Pi} \mathcal{P}_{z^\perp} \tilde{\Pi}^T\right)}\|[\tilde{\Pi}\mathcal{P}_z\tilde{\Pi}^T, \Pi\mathcal{P}_y\Pi^T]\|_2 \\
&\le \min_{\tilde{\Pi} \in \mathbb{S}_{\text{block}}^{p^2}} \text{trace}\left(\mathcal{P}_x \tilde{\Pi} \mathcal{P}_{z^\perp} \tilde{\Pi}^T\right) + 2\sqrt{\text{rank}(x)}\sqrt{\text{trace}\left(\mathcal{P}_x \tilde{\Pi} \mathcal{P}_{z^\perp} \tilde{\Pi}^T\right)} \max_{\bar{\Pi}, \bar{\bar{\Pi}} \in \mathbb{S}_{\text{block}}^{p^2}} \|[\bar{\bar{\Pi}}\mathcal{P}_z\bar{\bar{\Pi}}^T, \bar{\Pi}\mathcal{P}_y\bar{\Pi}^T]\|_2 \\
&\qquad + \max_{\tilde{\Pi} \in \mathbb{S}_{\text{block}}^{p}} \min_{\Pi \in \mathbb{S}_{\text{block}}^{p}} \text{trace}\left(\Pi(\text{Id} - \mathcal{P}_y)\Pi^T \tilde{\Pi}\mathcal{P}_z\tilde{\Pi}^T\right) \\
&= [\text{rank}(x) - \rho(x,z)] + [\text{rank}(z) - \rho(z,y)] \\
&\qquad + 2\sqrt{\text{rank}(x)}\sqrt{\text{rank}(x) - \rho(x,z)} \max_{\bar{\Pi}, \bar{\bar{\Pi}} \in \mathbb{S}_{\text{block}}^{p^2}} \|[\bar{\bar{\Pi}}\mathcal{P}_z\bar{\bar{\Pi}}^T, \bar{\Pi}\mathcal{P}_y\bar{\Pi}^T]\|_2.
\end{aligned}$$

Here, the first inequality follows from a similar analysis as arriving to Eq. (17) in subspace selection. The second inequality follows from the fact that $\min_{a,b} f(a) + g(b) \le \min_a f(a) + \max_b f(b)$. Note that projection matrices A, B, $[A,B] \le \frac{1}{2}$. Then, following the same exact reasoning as the subspace case, we have that in the blind-source separation setting $\frac{1}{B}\sum_{\ell=1}^{B}\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}) \le \sqrt{\alpha}\text{rank}(\hat{x}_{\text{sub}})$. The result follows subsequently.

V. Specializing Bound Eq. (8) for Different Problem Settings

V.I. Partial Ranking. Let $S = \{a_1, a_2, \ldots, a_p\}$ be the set of p elements. We use the similarity valuation $\rho := \rho_{\text{meet}}$ in Eq. (1) of the main paper.

V.I.1. Characterizing $\mathcal{S}$ for Partial Ranking. We construct a set $\mathcal{S}$ satisfying the properties in Definition 3 of the main paper. Specifically, we let:

$$\mathcal{S} = \{(a_i, a_j) : i \neq j\},$$

with $|\mathcal{S}_1| = p(p-1)$ and $\mathcal{S}_k = \emptyset$ for every $k \ge 2$.

We will show that set $\mathcal{S}$ as constructed above satisfies Definition 3. First, consider any covering pair $(u', v') \notin \mathcal{S}$. Here, u' and v' are relations and $v' = u' \cup (a_i, a_j)$ for some $i \neq j$. Then, for any $z \in \mathcal{L}$, it is easy to see that

$$\rho(v', z) - \rho(u', z) = \mathbb{I}[(a_i, a_j) \in z] = \rho(v, z) - \rho(u, z),$$

where $v = \{(a_i, a_j)\}$ and $u = \emptyset$. Clearly, $\text{rank}(v) \le \text{rank}(v')$.

To show the second property, consider covering pairs $(\{(a_i, a_j)\}, \emptyset) \in \mathcal{S}$ and $(\{(a_k, a_l)\}, \emptyset) \in \mathcal{S}$. By construction of the set $\mathcal{S}$, $(a_i, a_j) \neq (a_k, a_l)$. Let $z = \{(a_i, a_j)\}$. Then, it is straightforward to see that $\rho(\{(a_i, a_j)\}, z) - \rho(\emptyset, z) = 1$ but $\rho(\{(a_k, a_l)\}, z) - \rho(\emptyset, z) = 0$.

V.I.2. Characterizing $c_{\mathcal{L}}(x, y)$ for Covering Pair (x, y). Since for any z, $\rho(y, z) - \rho(x, z) = \mathbb{I}((a_i, a_j) \in z)$ for some (a_i, a_j). Thus, $c_{\mathcal{L}}(x, y) = 1$.

V.I.3. Refined False Discovery Bound for Partial Ranking. Let $\hat{x}_{\text{stable}}$ be output of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$. Then:

$$\mathbb{E}[\text{FD}(\hat{x}_{\text{stable}}, x^\star)] \leq \frac{q_1^2}{(1-2\alpha)p(p-1)},$$

where

$$q_1 = \sum_{i \neq j} \mathbb{I}[(a_i, a_j) \in \hat{x}_{\text{sub}}].$$

Here, $\hat{x}_{\text{sub}}$ is the estimated partial ranking from supplying $n/2$ samples to the base estimator. We can use the following data-driven approximation for q_1:$q_1 \approx \frac{1}{B}\sum_{\ell=1}^{B}\sum_{i \neq j} \mathbb{I}[(a_i, a_j) \in \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})]$ with $\hat{x}_{\text{base}}(\mathcal{D}^{(\ell)}), l = 1, 2, \ldots, B$ representing the estimates from subsampling.

V.II. Total Ranking. Let $S = \{a_1, a_2, \ldots, a_p\}$ be the set of p elements. Let $\pi_{\text{null}}(a_i) = i$ for every $i = 1, 2, \ldots, p$. We use the similarity valuation $\rho := \rho_{\text{total-ranking}}$ in Eq. (2) of the main paper. As each element in the poset corresponds to a function $\pi : S \to S$, we will use this functional notation throughout.

V.II.1. Characterizing $\mathcal{S}$ for Total Ranking. We construct a set $\mathcal{S}$ satisfying the properties in Definition 3 of the main paper. Initialize $\mathcal{S} = \emptyset$. Then, for every relation (a_i, a_j) with $i < j$, we augment $\mathcal{S}$ as follows:

$$\mathcal{S} = \mathcal{S} \cup (\pi_1, \pi_2),$$

where π_1, π_2 are covering pairs. Here, π_2 is any rank $j - i$ element in the poset with the relation (a_i, a_j) in its corresponding inversion set. Furthermore, we let π_1 be a rank $j - i - 1$ element that is covered by π_2 and does not contain (a_i, a_j) in its inversion set. Recalling that $\mathcal{S}_k = \{(\pi_1, \pi_2) \in \mathcal{S}, \text{rank}(\pi_2) = k\}$, we have that for every $k = 1, 2, \ldots, p-1$

$$|\mathcal{S}_k| = p - k.$$

We will show that set $\mathcal{S}$ as constructed above satisfies Definition 3. First, consider any covering pair $(\tilde{\pi}_1, \tilde{\pi}_2) \notin \mathcal{S}$. Then by definition, the corresponding inversion sets are nested, i.e. $\text{inv}(\tilde{\pi}_2; \pi_{\text{null}}) \supseteq \text{inv}(\tilde{\pi}_1; \pi_{\text{null}})$ with the difference being a single relation. We will denote this relation by (a_i, a_j) with $j > i$. Consider the covering pair $(\pi_1, \pi_2) \in \mathcal{S}$ where (a_i, a_j) is in the inversion set of π_2 but not in the inversion set of π_1. Then, for any π, we have that

$$\rho(\pi_2, \pi) - \rho(\pi_1, \pi) = \mathbb{I}((a_i, a_j) \in \text{inv}(\pi; \pi_{\text{null}})) = \rho(\tilde{\pi}_2, \pi) - \rho(\tilde{\pi}_1, \pi).$$

Furthermore, it is straightforward to check that $\text{rank}(\tilde{\pi}_2) \geq j - i = \text{rank}(\pi_2)$. We have thus shown that $\mathcal{S}$ satisfies the first property in Definition 3.

To show the second property, consider covering pairs $(\pi_1, \pi_2) \in \mathcal{S}$ where the difference between the two inversion sets is the relation (a_i, a_j). Let $(\pi_3, \pi_4) \in \mathcal{S}$ where the difference between the two inversion sets is the relation (a_k, a_l). By construction of the set $\mathcal{S}$, $(a_i, a_j) \neq (a_k, a_l)$. Let π be a permutation with (a_i, a_j) in its inversion set. Then, as desired,

$$\rho(\pi_2, \pi) - \rho(\pi_1, \pi) = \mathbb{I}((a_i, a_j) \in \text{inv}(\pi; \pi_{\text{null}})) \neq \rho(\pi_4, \pi) - \rho(\pi_3, \pi).$$

V.II.2. Characterizing $c_{\mathcal{L}}(\pi_1, \pi_2)$ for Covering Pair (π_1, π_2). Since for any π, $\rho(\pi_2, \pi) - \rho(\pi_1, \pi) = \mathbb{I}((a_i, a_j) \in \text{inv}(\pi; \pi_{\text{null}}))$ for some pair of elements (a_i, a_j), then $c_{\mathcal{L}}(\pi_1, \pi_2) = 1$.

V.II.3. Refined False Discovery Bound for Total Ranking. Let $\hat{\pi}_{\text{stable}}$ be output of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$. Then:

$$\mathbb{E}[\text{FD}(\hat{\pi}_{\text{stable}}, \pi^\star)] \leq \sum_{k=1}^{p-1} \frac{q_k^2}{(1-2\alpha)(p-k)},$$

where

$$q_k = \sum_{(\pi_1, \pi_2) \in \mathcal{S}_k} \mathbb{E}[\rho(\pi_2, \hat{\pi}_{\text{sub}}) - \rho(\pi_1, \hat{\pi}_{\text{sub}})] = \sum_{(i,j), j-i=k} [\mathbb{I}[(a_i, a_j) \in \text{inv}(\hat{\pi}_{\text{sub}}; \pi_{\text{null}})]].$$

Here, $\hat{\pi}_{\text{sub}}$ represents ranking from supplying $n/2$ samples to the base estimator. We can use the following data-driven approximation for q_k:$q_k \approx \frac{1}{B}\sum_{(i,j), j-i=k}\sum_{\ell=1}^{B}\left[\mathbb{I}[(a_i, a_j) \in \text{inv}(\hat{\pi}_{\text{base}}(\mathcal{D}^{(\ell)}); \pi_{\text{null}})]\right]$, where $\hat{\pi}_{\text{base}}(\mathcal{D}^{(\ell)})$ represents the total ranking obtained by supplying the base estimator on dataset $\mathcal{D}^{(\ell)}$.

 Armeen Taeb, Peter Bühlmann and Venkat Chandrasekaran

V.III. Clustering. We have a collection of p items $\{a_1, a_2, \ldots, a_p\}$ that we wish to cluster. We let $x_0 = \{\{a_1\}, \{a_2\}, \ldots, \{a_p\}\}$ be the least element. As described in the main paper, will use the similarity valuation $\rho := \rho_{\text{meet}}$ defined in Eq. (1) of the main paper. Since the clustering poset is meet semi-lattice, ρ computes the rank of the meet of two elements; in this setting, the meet $x \wedge z$ of $x = \{G_1, \ldots, G_q\}$ and $z = \{\tilde{G}_1, \ldots, \tilde{G}_s\}$ is

$$x \wedge z = \{G_i \cap \tilde{G}_j : G_i \cap \tilde{G}_j \neq \emptyset\}.$$

Subsequently, $\rho(x, z) = \text{rank}(x \wedge y)$ is $p - \#$ groups in $x \wedge z$, which can be equivalently expressed as:

$$\rho(x, z) = \sum_{i,j : |G_i \cap \tilde{G}_j| \neq \emptyset} |G_i \cap \tilde{G}_j| - 1.$$

For sets $G_1, G_2 \subseteq \{1, 2, \ldots, p\}$ with $G_1 \cap G_2 = \emptyset$, we define:

$$\mathcal{R}_{G_1,G_2} := \{\{a_1\}, \{a_2\}, \ldots, \{a_p\}\} \setminus \{\{a_i\} : a_i \in G_1 \cup G_2\}.$$

V.III.1. Characterizing $\mathcal{S}$ for Clustering. We construct a set $\mathcal{S}$ satisfying the properties in Definition 3. Initialize $\mathcal{S} = \emptyset$. Then, for every $k = 1, 2, \ldots, p-1$ and pairs of groups of variables $G_1 \subseteq \{a_1, \ldots, a_p\}$ and $G_2 \subseteq \{a_1, \ldots, a_p\}$ with $|G_1| + |G_2| = k + 1$ and $G_1 \cap G_2 = \emptyset$, we generate covering pairs (x, y) with $y = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$ and $x = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$, and let

$$S = S \cup (x, y).$$

Recalling that $\mathcal{S}_k = \{(x, y) \in \mathcal{S}, \text{rank}(y) = k\}$, it is straightforward to check that for every $k = 1, 2, \ldots, p-1$

$$|\mathcal{S}_k| = \binom{p}{k+1} \sum_{\ell=1}^{k} \binom{k+1}{l}.$$

Here, the terms $\binom{p}{k+1}$ counts the number of possible items in $G_1 \cup G_2$ and the term $\sum_{\ell=1}^{k+1} \binom{k+1}{l}$ counts the number of possible configurations of the group G_2. We will show that the constructed set $\mathcal{S}$ satisfies Definition 3 of the main paper. Our analysis is based on the following lemma.

Lemma 10. *Consider the covering pairs (x, y) with $x = \{G_1, G_2, \ldots, G_q\}$ and $y = \{G_1 \cup G_2, G_3, \ldots, G_q\}$ where $G_i \subseteq \{1, 2, \ldots, px\}$ and $G_i \cap G_j = \emptyset$ for every $i \neq j$. Let $(\tilde{x}, \tilde{y})$ be covering pairs with $\tilde{y} = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$ and $\tilde{x} = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$. Then, for every $z \in \mathcal{L}$, $\rho(y, z) - \rho(x, z) = \rho(\tilde{y}, z) - \rho(\tilde{x}, z)$.*

Proof of Lemma 10. Let $z = \{\tilde{G}_1, \ldots, \tilde{G}_s\}$ with $\tilde{G}_i \subseteq \{a_1, a_2, \ldots, a_p\}$ and $\tilde{G}_i \cap \tilde{G}_j = \emptyset$ for every $i \neq j$. Then:

$$\rho(y, z) = \sum_{j : (G_1 \cup G_2) \cap \tilde{G}_j \neq \emptyset} |(G_1 \cup G_2) \cap \tilde{G}_j| - 1 + \sum_{i \geq 3, j : G_i \cap \tilde{G}_j \neq \emptyset} |G_i \cap \tilde{G}_j| - 1,$$

and

$$\rho(x, z) = \sum_{j : G_1 \cap \tilde{G}_j \neq \emptyset} |G_1 \cap \tilde{G}_j| - 1 + \sum_{j : G_2 \cap \tilde{G}_j \neq \emptyset} |G_2 \cap \tilde{G}_j| - 1 + \sum_{i \geq 3, j : G_i \cap \tilde{G}_j \neq \emptyset} |G_i \cap \tilde{G}_j| - 1.$$

Since $\mathcal{R}_{G_1,G_2}$ consists of groups of size one, we have that:

$$\rho(\tilde{y}, z) = \sum_{j : (G_1 \cup G_2) \cap \tilde{G}_j \neq \emptyset} |(G_1 \cup G_2) \cap \tilde{G}_j| - 1,$$

and

$$\rho(\tilde{x}, z) = \sum_{j : G_1 \cap \tilde{G}_j \neq \emptyset} |G_1 \cap \tilde{G}_j| - 1 + \sum_{j : G_2 \cap \tilde{G}_j \neq \emptyset} |G_2 \cap \tilde{G}_j| - 1.$$

We thus can see that $\rho(y, z) - \rho(x, z) = \rho(\tilde{y}, z) - \rho(\tilde{x}, z)$. □

<u>Showing $\mathcal{S}$ satisfies Definition 3</u> With Lemma 10 at hand, we show that out constructed $\mathcal{S}$ satisfies Definition 3 of the main paper. We start with the first property. Consider any $(u', v') \subseteq \mathcal{L}$. Without loss of generality, we take $v' = \{G_1 \cup G_2, G_3, \ldots, G_q\}$ and $u' = \{G_1, G_2, \ldots, G_q\}$. We let $v = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$ and $u = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$. Then, according to Lemma 10, we have that $\rho(v', z) - \rho(u', z) = \rho(v, z) - \rho(u, z)$. Furthermore, since $\text{rank}(x) = p - \#$ groups in x, we have that $\text{rank}(v) \leq \text{rank}(v')$. Thus, the first property of $\mathcal{S}$ is satisfied. We demonstrate the second property. Consider any $(u, v) \in \mathcal{S}$ and $(u', v') \in \mathcal{S}$ that are different. Let $u = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$ and $v = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$. Additionally, let $u' = \{G'_1, G'_2, \mathcal{R}_{G'_1,G'_2}\}$ and $v' = \{G'_1 \cup G'_2, \mathcal{R}_{G'_1,G'_2}\}$. Since the covering pairs (u, v) and (u', v') are different, there must exist two items a_i, a_j such that either (a_i, a_j) are grouped together in v but are not together in u or (a_i, a_j) are grouped together in v' but are not together in u'. Let $z = \{\{a_i, a_j\}, \mathcal{R}_{\{a_i\},\{a_j\}}\}$. Since $\rho(v, z) - \rho(u, z) = \mathbb{I}[(a_i, a_j)$ grouped together in v but not in $u]$ and $\rho(v', z) - \rho(u', z) = \mathbb{I}[(a_i, a_j)$ grouped together in v' but not in $u']$, we have that $\rho(v, z) - \rho(u, z) \neq \rho(v', z) - \rho(u', z)$.

V.III.2. Characterizing $c_{\mathcal{L}}(u,v)$ for Covering Pair (u,v).

Lemma 11. *Let $v = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$ and $u = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$ be a covering pair $(u,v) \in \mathcal{S}$. Then, $c_{\mathcal{L}}(u,v) = \min\{|G_1|,|G_2|\}$.*

Proof of Lemma 11. Let $z = \{\tilde{G}_1, \ldots, \tilde{G}_q\}$. Then, from proof of Lemma 10, we have that:

$$\rho(v,z) - \rho(u,z) = \left[\sum_{j:(G_1\cup G_2)\cap\tilde{G}_j\neq\emptyset} |(G_1 \cup G_2)\cap \tilde{G}_j| - 1\right] - \left[\sum_{j:G_1\cap\tilde{G}_j\neq\emptyset} |G_1 \cap \tilde{G}_j| - 1\right] - \left[\sum_{j:G_2\cap\tilde{G}_j\neq\emptyset} |G_2 \cap \tilde{G}_j| - 1\right].$$

Let $I_1 := \{j : \tilde{G}_j \cap G_1 \neq \emptyset\}$ and $I_2 := \{j : \tilde{G}_j \cap G_2 \neq \emptyset\}$. Then,

$$\rho(v,z) - \rho(u,z) = \left[\sum_{j\in I_1\cup I_2} |(G_1 \cup G_2)\cap \tilde{G}_j| - 1\right] - \left[\sum_{j\in I_1} |G_1 \cap \tilde{G}_j| - 1\right] - \left[\sum_{j\in I_2} |G_2 \cap \tilde{G}_j| - 1\right].$$

Simple manipulations yield:

$$\rho(v,z) - \rho(u,z) = \left[\sum_{j\in I_1\cap I_2} |(G_1 \cup G_2)\cap \tilde{G}_j| - 1\right] - \left[\sum_{j\in I_1\cap I_2} |G_1 \cap \tilde{G}_j| - 1\right] - \left[\sum_{j\in I_1\cap I_2} |G_2 \cap \tilde{G}_j| - 1\right].$$

Clearly, if $I_1 \cap I_2 = \emptyset$, then $\rho(v,z) - \rho(u,z) = 0$. Suppose $I_1 \cap I_2 \neq \emptyset$. Then,

$$\rho(v,z) - \rho(u,z) = |I_1 \cap I_2| + \left[\sum_{j\in I_1\cap I_2} |(G_1 \cup G_2)\cap \tilde{G}_j| - |G_1 \cap \tilde{G}_j| - |G_2 \cap \tilde{G}_j|\right] = |I_1 \cap I_2|.$$

Notice that $|I_1 \cap I_2| \leq \min\{|G_1|,|G_2|\}$. Then, the upper bound can be achieved by for example setting $z = \{N, \{\{a_1\},\{a_2\},\ldots,\{a_p\}\backslash N\}$ with $N = \{(a_i,a_j) : a_i \in G_1, a_j \in G_2\}$. $\square$

V.III.3. Refined False Discovery Bound for Clustering. Let $\hat{x}_{\text{stable}}$ be output of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$. Then:

$$\mathbb{E}[\text{FD}(\hat{x}_{\text{stable}}, x^\star)] \leq \sum_{k=1}^{p-1} \frac{q_k^2}{(1-2\alpha)\binom{p}{k+1}\sum_{\ell=1}^{k}\binom{k+1}{l}},$$

where,

$$\begin{aligned} q_k &= \sum_{(u,v)\in\mathcal{S}_k} \frac{\mathbb{E}[\rho(v,\hat{x}_{\text{sub}}) - \rho(u,\hat{x}_{\text{sub}})]}{c(u,v)} \\ &= \sum_{\substack{G_1\subseteq\{a_1,\ldots,a_p\}, G_2\subseteq\{a_1,\ldots,a_p\}\\ G_1\cap G_2=\emptyset; |G_1|+|G_2|=k+1}} \frac{\mathbb{E}[\#\text{ groups } \hat{G}_j \text{ in } \hat{x}_{\text{sub}} \text{ satisfying } \hat{G}_j\cap G_1\neq\emptyset \text{ and } \hat{G}_j \cap G_2 \neq \emptyset]}{\min\{|G_1|,|G_2|\}}. \end{aligned} \tag{127}$$

Here, $\hat{x}_{\text{sub}}$ represents clustering from supplying $n/2$ samples to the base estimator. We will use the following data-driven approximation to estimate q_k

$$q_k \approx \frac{1}{B} \sum_{\substack{G_1\subseteq\{a_1,\ldots,a_p\}, G_2\subseteq\{a_1,\ldots,a_p\}\\ G_1\cap G_2=\emptyset; |G_1|+|G_2|=k+1}} \sum_{\ell=1}^{B} \frac{\#\text{ groups } \hat{G}_j \text{ in } \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)}) \text{ satisfying } \hat{G}_j\cap G_1\neq\emptyset \text{ and } \hat{G}_j \cap G_2 \neq \emptyset]}{\min\{|G_1|,|G_2|\}},$$

with $\hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})$ represents the partition obtained from supplying $\mathcal{D}^{(\ell)}$ to the base estimator.

V.IV. Causal Structure Learning. Throughout, we consider covering pairs $(\mathcal{C}_u, \mathcal{C}_v)$ where each connected component in the skeletons of $\mathcal{C}_u, \mathcal{C}_v$ have a diameter at most two. We denote this set by $\mathcal{T}$. Note that for any covering pair $(\mathcal{C}_u,\mathcal{C}_v) \in \mathcal{T}$, $\mathcal{C}_v$ is a polytree. Throughout, we will use the similarity valuation $\rho := \rho_{\text{meet}}$. Our analysis in this section will build on the following result.

Lemma 12. *Let $\mathcal{C}_u$ and $\mathcal{C}_v$ be two CPDAGs that are polytrees with $\mathcal{C}_u \preceq \mathcal{C}_v$. Then, the following statements hold:*

(a) for any pairs of nodes $\mathcal{E}$, the set of DAGs that result from removing edges among pairs $\mathcal{E}$ in any DAG $\mathcal{G}_v$ form a Markov equivalence class.

(b) for every DAG $\mathcal{G}_v \in \mathcal{C}_v$, there exists a DAG $\mathcal{G}_u \in \mathcal{C}_u$ such that $\mathcal{G}_u$ is a directed subgraph of $\mathcal{G}_v$.

Proof of Lemma 12. We first prove part (a). By the polytree assumption, it follows that for any DAG $\mathcal{G}_v$ in the CPDAG $\mathcal{C}_v$, removing the edges among pairs in $\mathcal{E}$ does not create any v-structures, and removes the same (potentially empty) v-structures. That means that the collection of DAGs obtained by taking any DAG in $\mathcal{C}_v$ and removing the edges between the pairs of nodes $\mathcal{E}$ will have the same skeleton and same v-structures, and are thus in the same Markov equivalence class.

We next prove part (b). Let (i, j) be the pair of nodes that are connected in $\mathcal{C}_v$ but not in $\mathcal{C}_u$. Recall that $\mathcal{C}_u \preceq \mathcal{C}_v$ implies there exists a DAG $\mathcal{G}_u \in \mathcal{C}_u$ and a DAG $\mathcal{G}_v \in \mathcal{C}_v$ where $\mathcal{G}_u$ is a subgraph of $\mathcal{G}_v$, where $\mathcal{G}_u$ does not have the edge among pairs (i, j). Appealing to the result in part (a), we have that removing the edge (i, j) from any other DAG in $\mathcal{C}_v$ results in a DAG in the same equivalence class, which is $\mathcal{C}_u$. □

V.IV.1. Characterizing $\mathcal{S}$ for Causal Structure Learning. We construct the set $\mathcal{S}$ as follows. Initialize $\mathcal{S} = \emptyset$. For every reference node, and $k = 1, \dots, p-1$, let $\mathcal{C}_y$ be a CPDAG generated with k edges, where every edge is between the reference node and another node; no other edges can be added without violating the condition that the largest undirected path has size less than or equal to two. A consequence of Lemma 12 is that there are k CPDAGs $\mathcal{C}_{x_1}, \dots, \mathcal{C}_{x_k}$ that form a covering pair with $\mathcal{C}_y$. We then let

$$\mathcal{S} = \mathcal{S} \cup (\mathcal{C}_{x_i}, \mathcal{C}_y),$$

for every $i = 1, 2, \dots, k$. Recall that $\mathcal{S}_k := \{(\mathcal{C}_x, \mathcal{C}_y) \in \mathcal{S}, \mathrm{rank}(\mathcal{C}_y) = k\}$. Then,

$$|\mathcal{S}_k| = p\binom{p-1}{k} \sum_{i \in \{0,2\dots,k\}} \binom{k}{i}.$$

The result above follows from noting that for every reference node and k other nodes, there are $\sum_{i \in \{0,2\dots,k\}} \binom{k}{i}$ possible CPDAGs that are polytrees can formed by connecting the k nodes to the reference node; the factor $p\binom{p-1}{k}$ comes from p total possible reference nodes and $\binom{p-1}{k}$ possible set of k nodes to connect to the reference node.

We will show that the constructed set $\mathcal{S}$ satisfies Definition 3 of the main paper. Our analysis is based on the following lemma.

Lemma 13. *Let $\mathcal{C}_{\tilde{y}}$ be a CPDAG that contains m disconnected subgraphs (both directed and undirected). Let $\mathcal{C}_{\tilde{y}_i}$ be each disconnected subgraph for $i = 1, 2, \dots, m$. Then, for any CPDAG $\mathcal{C}_z$,*

$$\rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z) = \sum_{i=1}^{m} \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z).$$

Proof. We will first show that $\rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z) \leq \sum_{i=1}^m \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z)$. Let $\mathcal{C}_{\tilde{x}} \in \mathrm{argmax}_{\mathcal{C}_x \preceq \mathcal{C}_{\tilde{y}}, \mathcal{C}_x \preceq \mathcal{C}_z} \mathrm{rank}(\mathcal{C}_x)$. By definition, $\mathcal{C}_x \preceq \mathcal{C}_{\tilde{y}}$ if there is a DAG $\mathcal{G}_x$ in $\mathcal{C}_x$ and a DAG $\mathcal{G}_{\tilde{y}}$ in $\mathcal{C}_{\tilde{y}}$ such that $\mathcal{G}_x$ is a subgraph of $\mathcal{G}_{\tilde{y}}$. Since $\mathcal{G}_{\tilde{y}}$ has disconnected components, so must $\mathcal{G}_x$. We let $\mathcal{C}_{\tilde{x}_i}$ be the subgraphs of $\mathcal{C}_{\tilde{x}}$ where every subgraph $\mathcal{C}_{\tilde{x}_i}$ only contains edges among nodes that are connected (to other nodes) in the graph $\mathcal{C}_{\tilde{y}_i}$. By construction, $\mathcal{C}_{\tilde{x}_i} \preceq \mathcal{C}_{\tilde{y}_i}$, $\mathrm{rank}(\mathcal{C}_{\tilde{x}}) = \sum_{i=1}^m \mathrm{rank}(\mathcal{C}_{\tilde{x}_i})$, and $\mathcal{C}_{\tilde{x}_i} \preceq \mathcal{C}_z$. Thus, $\mathrm{rank}(\mathcal{C}_{\tilde{x}_i}) \leq \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z)$. Then, we can conclude that

$$\sum_{i=1}^{m} \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z) \geq \sum_{i=1}^{m} \mathrm{rank}(\mathcal{C}_{\tilde{x}_i}) = \mathrm{rank}(\mathcal{C}_{\tilde{x}}) = \rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z).$$

Now we will show that $\rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z) \geq \sum_{i=1}^m \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z)$. Let $\mathcal{C}_{\tilde{x}_i} \in \mathrm{argmax}_{\mathcal{C}_x \preceq \mathcal{C}_{\tilde{y}_i}, \mathcal{C}_x \preceq \mathcal{C}_z} \mathrm{rank}(\mathcal{C}_x)$. Now form a CPDAG $\mathcal{C}_{\bar{y}}$ by combining all the disjoint graphs $\mathcal{C}_{\tilde{x}_i}$ for every $i = 1, 2, \dots, m$ into one graph. Since these graphs are disjoint (i.e. nodes that are connected in each graph are distinct), we have that $\mathcal{C}_{\bar{y}} \preceq \mathcal{C}_{\tilde{y}}$ and $\mathcal{C}_{\bar{y}} \preceq \mathcal{C}_z$ and that $\mathrm{rank}(\mathcal{C}_{\bar{y}}) = \sum_{i=1}^m \mathrm{rank}(\mathcal{C}_{\tilde{x}_i})$. So we conclude that

$$\rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z) \geq \mathrm{rank}(\mathcal{C}_{\bar{y}}) = \sum_{i=1}^{m} \mathrm{rank}(\mathcal{C}_{\tilde{x}_i}) = \sum_{i=1}^{m} \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z).$$

□

Showing $\mathcal{S}$ satisfies Definition 3 For the first property, consider covering pairs $(\mathcal{C}_{u'}, \mathcal{C}_{v'}) \in T$. Let (i, j) be the pair of nodes that are connected in $\mathcal{C}_{v'}$ and are not connected in $\mathcal{C}_{u'}$. Since every undirected path in $\mathcal{C}_{v'}$ has size at most 2, then $\mathcal{C}_{v'}$ decouples into two disconnected CPDAGs $\mathcal{C}_v$ and $\mathcal{C}_1$, where $\mathcal{C}_v$ only involves nodes adjacent to (i, j). Similarly, $\mathcal{C}_{u'}$ decouples into two disconnected CPDAGs $\mathcal{C}_u$ and $\mathcal{C}_2$, where $\mathcal{C}_2 = \mathcal{C}_1$ and $\mathcal{C}_u$ is covered by $\mathcal{C}_v$. From Lemma 13, we have that for any CPDAG $\mathcal{C}_z$

$$\rho(\mathcal{C}_{v'}, \mathcal{C}_z) - \rho(\mathcal{C}_{u'}, \mathcal{C}_z) = \rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z).$$

Notice that $(\mathcal{C}_u, \mathcal{C}_v) \in \mathcal{S}$. Furthermore, since the number of edges (directed and undirected) in $\mathcal{C}_{v'}$ is larger than $\mathcal{C}_v$, we have that $\mathrm{rank}(\mathcal{C}_v) \leq \mathrm{rank}(\mathcal{C}_{v'})$.

We next show the second property in Definition 3. Let $(\mathcal{C}_u, \mathcal{C}_v) \in \mathcal{S}$ and $(\mathcal{C}_{u'}, \mathcal{C}_{v'}) \in \mathcal{S}$. Our objective is to show that $\rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z) = \rho(\mathcal{C}_{v'}, \mathcal{C}_z) - \rho(\mathcal{C}_{u'}, \mathcal{C}_z)$ for all $\mathcal{C}_z \Leftrightarrow \mathcal{C}_u = \mathcal{C}_{u'}$ and $\mathcal{C}_v = \mathcal{C}_{v'}$. The direction $\leftarrow$ trivially holds, and hence we focus on the direction $\rightarrow$. We consider multiple scenarios; throughout the extra edge that is present in $\mathcal{C}_v$ and not in $\mathcal{C}_u$ is between the pair of nodes (i, j), and the extra edge that is present in $\mathcal{C}_{v'}$ and not in $\mathcal{C}_{u'}$ is between the pair of nodes (k, l).

(1) Suppose that the nodes (k,l) are not connected in $\mathcal{C}_v$. Letting $\mathcal{C}_z$ be a CPDAG with only an edge between nodes (k,l), we find that $\rho(\mathcal{C}_v,\mathcal{C}_z)-\rho(\mathcal{C}_u,\mathcal{C}_z)=0$ and $\rho(\mathcal{C}_{v'},\mathcal{C}_z)-\rho(\mathcal{C}_{u'},\mathcal{C}_z)=1$. So this scenario cannot occur.

(2) Suppose there is an edge between pairs (s,t) in $\mathcal{C}_{u'}$ that is missing in $\mathcal{C}_v$ (and as a result in $\mathcal{C}_u$). Construct CPDAG $\mathcal{C}_z$ with two edges, one between the pair (i,j) and another between the pair (s,t) with the property that $\mathcal{C}_z \not\preceq \mathcal{C}_{v'}$; this construction is possible since $(\mathcal{C}_{u'},\mathcal{C}_{v'})\in\mathcal{S}$, meaning that if there is an edge between pair of nodes (i,j) in $\mathcal{C}_{v'}$, this edge is incident to the edge between the pair of nodes (s,t). Then, it is evident that $\rho(\mathcal{C}_v,\mathcal{C}_z)-\rho(\mathcal{C}_u,\mathcal{C}_z)=1$ but $\rho(\mathcal{C}_{v'},\mathcal{C}_z)-\rho(\mathcal{C}_{u'},\mathcal{C}_z)=0$. So this scenario cannot occur.

(3) Suppose there is an edge between pairs (s,t) in $\mathcal{C}_{u'}$ that is missing in $\mathcal{C}_u$ but is not missing in $\mathcal{C}_v$. Let $\mathcal{C}_z$ be a CPDAG only containing an edge between (s,t). Then it follows that $\rho(\mathcal{C}_v,\mathcal{C}_z)-\rho(\mathcal{C}_u,\mathcal{C}_z)=1$ but $\rho(\mathcal{C}_{v'},\mathcal{C}_z)-\rho(\mathcal{C}_{u'},\mathcal{C}_z)=0$. So this scenario cannot occur.

From the impossibilities of scenarios 1-2, and noting that a similar argument can be made by swapping $\mathcal{C}_{u'}$ with $\mathcal{C}_u$, and $\mathcal{C}_{v'}$ with $\mathcal{C}_v$, we conclude that $\mathcal{C}_v,\mathcal{C}_{v'}$ have edges between the same pairs of nodes. Combining this result with the impossibility of scenario 3, we conclude that $\mathcal{C}_u,\mathcal{C}_{u'}$ have edges between the same pairs of nodes. We then continue with the final scenario.

(4) Suppose that $\mathcal{C}_v$ and $\mathcal{C}_{v'}$ are not identical CPDAGs. Since both $\mathcal{C}_v$ and $\mathcal{C}_{v'}$ have maximum undirected path length less than or equal to two, they both must have the same reference node i (where the other nodes are connected to). Furthermore, since $\mathcal{C}_v$ and $\mathcal{C}_{v'}$ have the same skeleton and are different, they must have strictly more than one edge, and they must have different v-structures. As a first sub-case, suppose $\mathcal{C}_{v'}$ have a v-structure $s\rightarrow i\leftarrow t$ that is not present in $\mathcal{C}_v$, so that $s\leftarrow i$ or $s-i$ in $\mathcal{C}_v$. Then, let $\mathcal{C}_z$ be a CPDAG containing two edges between the pairs (i,j) and (i,s) with $\mathcal{C}_z\preceq\mathcal{C}_v$. By construction, $\rho(\mathcal{C}_v,\mathcal{C}_z)-\rho(\mathcal{C}_u,\mathcal{C}_z)=1$ but $\rho(\mathcal{C}_{v'},\mathcal{C}_z)-\rho(\mathcal{C}_{u'},\mathcal{C}_z)=0$. Swapping $\mathcal{C}_{u'}$ with $\mathcal{C}_u$, and $\mathcal{C}_{v'}$ with $\mathcal{C}_v$, and following similar arguments, we arrive again at a contradiction if $\mathcal{C}_v$ has a v-structure that is not present in $\mathcal{C}_{v'}$.

From the impossibility of scenario 4, we conclude that $\mathcal{C}_v$ and $\mathcal{C}_{v'}$ have the same skeleton and v-structure and consequently $\mathcal{C}_v=\mathcal{C}_{v'}$. We thus have that $\mathcal{C}_u\preceq\mathcal{C}_v$ and $\mathcal{C}_{u'}\preceq\mathcal{C}_v$. Furthermore, since $\mathcal{C}_{u'}$ and $\mathcal{C}_u$ have the same skeleton, both are missing an edge between pair of nodes (i,j) that is connected in $\mathcal{C}_v$. Appealing to part a of Lemma 12, we conclude that $\mathcal{C}_u=\mathcal{C}_{u'}$.

V.IV.2. Characterizing $c_{\mathcal{L}}(\mathcal{C}_u,\mathcal{C}_v)$ for Covering Pairs $(\mathcal{C}_u,\mathcal{C}_v)$. We have the following lemma.

Lemma 14. *Let $(\mathcal{C}_u,\mathcal{C}_v)$ be CPDAGs that are polytrees and form a covering pair. Then, $c_{\mathcal{L}}(\mathcal{C}_u,\mathcal{C}_v)=1$.*

Proof. Let the pair of nodes (i,j) be connected in $\mathcal{C}_v$ and not connected in $\mathcal{C}_u$. Consider any CPDAG $\mathcal{C}_z$. Let $\mathcal{C}_{\tilde{y}}\in\operatorname{argmax}_{\mathcal{C}_y\preceq\mathcal{C}_v,\mathcal{C}_y\preceq\mathcal{C}_z}\operatorname{rank}(\mathcal{C}_y)$. Since the CPDAG $\mathcal{C}_v$ is a polytree, so is the CPDAG $\mathcal{C}_{\tilde{y}}$. Let $\mathcal{G}_v$ be any DAG in $\mathcal{C}_v$. Then, by Lemma 12, there exists DAGs $\mathcal{G}_{\tilde{y}}^{(1)}\in\mathcal{C}_{\tilde{y}}$ and $\mathcal{G}_u\in\mathcal{C}_u$ such that $\mathcal{G}_{\tilde{y}}^{(1)}$ and $\mathcal{G}_u$ are both subgraphs of $\mathcal{G}_v$. Suppose we remove an edge that may be present between the pair of nodes (i,j) in $\mathcal{G}_{\tilde{y}}^{(1)}$ and denote the resulting subgraph by $\mathcal{G}_x^{(1)}$. By construction, $\mathcal{G}_x^{(1)}$ is also a subgraph of $\mathcal{G}_u$. Since $\mathcal{C}_{\tilde{y}}\preceq\mathcal{C}_z$, there exists a DAG $\mathcal{G}_{\tilde{y}}^{(2)}\in\mathcal{C}_{\tilde{y}}$ and a DAG $\mathcal{G}_z\in\mathcal{C}_z$ such that $\mathcal{G}_{\tilde{y}}^{(2)}$ is a subgraph of $\mathcal{G}_z$. Suppose again we remove an edge that may be present between the pair of nodes (i,j) in $\mathcal{G}_{\tilde{y}}^{(2)}$ and denote the resulting subgraph by $\mathcal{G}_x^{(2)}$. By Lemma 12, $\mathcal{G}_x^{(2)}$ and $\mathcal{G}_x^{(1)}$ are in the same equivalence class, which we denote by $\mathcal{C}_x$. By construction, $\mathcal{C}_x\preceq z$ and $\mathcal{C}_x\preceq\mathcal{C}_u$. Furthermore, $\operatorname{rank}(\mathcal{C}_x)\geq\operatorname{rank}(\mathcal{C}_{\tilde{y}})-1$. Thus, we have shown that for any arbitrary $\mathcal{C}_z$: $\rho(\mathcal{C}_v,\mathcal{C}_z)-\rho(\mathcal{C}_u,\mathcal{C}_z)\leq 1$. □

V.IV.3. Refined False Discovery Bound for Causal Structure Learning. Let $\hat{\mathcal{C}}_{\text{stable}}$ be output of Algorithm 1 with $\Psi=\Psi_{\text{stable}}$. Let $\mathcal{C}^\star$ be the population CPDAG. Then:

$$\mathbb{E}[\text{FD}(\hat{\mathcal{C}}_{\text{stable}},\mathcal{C}^\star)]\leq\sum_{k=1}^{p-1}\frac{q_k^2}{(1-2\alpha)p\binom{p-1}{k}\sum_{i\in\{0,2\ldots,k\}}\binom{k}{i}},$$

where,

$$q_k=\sum_{(\mathcal{C}_u,\mathcal{C}_v)\in\mathcal{S}_k}\mathbb{E}[\rho(\mathcal{C}_v,\hat{\mathcal{C}}_{\text{sub}})-\rho(\mathcal{C}_u,\hat{\mathcal{C}}_{\text{sub}})].$$

Here, $\hat{\mathcal{C}}_{\text{sub}}$ represents the CPDAG from supplying $n/2$ samples to the base estimator. We will use the following data-driven approximation to estimate q_k

$$q_k\approx\frac{1}{B}\sum_{\ell=1}^{B}\sum_{(\mathcal{C}_u,\mathcal{C}_v)\in\mathcal{S}_k}\mathbb{E}[\rho(\mathcal{C}_v,\hat{\mathcal{C}}_{\text{base}}(\mathcal{D}^{(\ell)})-\rho(\mathcal{C}_u,\hat{\mathcal{C}}_{\text{base}}(\mathcal{D}^{(\ell)}))],$$

with $\hat{\mathcal{C}}_{\text{base}}(\mathcal{D}^{(\ell)})$ represents the CPDAGs obtained from supplying dataset $\mathcal{D}^{(\ell)}$ to base estimator $\hat{\mathcal{C}}_{\text{base}}$.

VI. Assumptions 1 and 2 of the Main Paper for the Total Ranking Problem in Example 7

Let $S = \{a_1, a_2, \dots, a_p\}$ be the set of p elements. Let $\pi_{\text{null}}(a_i) = i$ for every $i = 1, 2, \dots, p$. We use the similarity valuation $\rho := \rho_{\text{total-ranking}}$ in Eq. (2) of the main paper. As each element in the poset corresponds to a function $\pi : S \to S$, we will use this functional notation throughout. For a covering pair (π_1, π_2), there exists a single pair of elements $(a_i, a_j) \in \text{inv}(\pi_2; \pi_{\text{null}}) \setminus \text{inv}(\pi_1; \pi_{\text{null}})$ with $j > i$. Then, from the definition of ρ, for any permutation π, we have that

$$\rho(\pi_2, \pi) - \rho(\pi_1, \pi) = \mathbb{I}[(a_i, a_j) \in \text{inv}(\pi; \pi_{\text{null}})] = \mathbb{I}[\pi(a_j) < \pi(a_i)].$$

Let $\hat{\pi}_{\text{sub}}$ be the estimated ranking from applying a base procedure on a subsample of the data. Consider a fixed integer k with $1 \le k \le p - 1$. Define the sets S_1 and S_2:

$$S_1 = \{(a_i, a_j) \in \text{inv}(\pi^\star; \pi_{\text{null}}) : j - i = k\},$$
$$S_2 = \{(a_i, a_j) \notin \text{inv}(\pi^\star; \pi_{\text{null}}) : j - i = k\}.$$

The set S_1 corresponds to non-null pairs (as described in the main paper) and the set S_2 corresponds to null pairs.

Then, appealing to the definition of $\mathcal{S}$ and the constant $c_{\mathcal{L}}(\cdot, \cdot)$ in the total ranking case (see Section V.II), Assumption 1 of the main paper reduces to the following inequality being satisfied

$$\frac{\sum_{(a_i, a_j) \in S_1} \mathbb{P}(\hat{\pi}_{\text{sub}}(a_j) < \hat{\pi}_{\text{sub}}(a_i))}{\sum_{(a_i, a_j) \in S_2} \mathbb{P}(\hat{\pi}_{\text{sub}}(a_j) < \hat{\pi}_{\text{sub}}(a_i))} \ge \frac{|S_1|}{|S_2|}. \qquad [19]$$

Consider an estimator $\hat{\pi}_{\text{sub}} = \hat{\pi}_{\text{random}}$ that randomly selects a total ranking in the space of permutations. Then, for every i and j, $\mathbb{P}(\hat{\pi}_{\text{sub}}(a_j) < \hat{\pi}_{\text{sub}}(a_i)) = \frac{1}{2}$. Thus, in this case, Assumption 1 in Eq. (19) is satisfied with equality.

It is also straightforward to check that Assumption 2 of the main paper is reduced to

$$\mathbb{P}(\hat{\pi}_{\text{sub}}(a_j) < \hat{\pi}_{\text{sub}}(a_i)) \text{ being the same for every } (a_j, a_i) \in S_2.$$

References

1. RP Stanley, *Enumerative Combinatorics*, Cambridge Studies in Advanced Mathematics. (Cambridge University Press) Vol. 1, 2nd edition, (2011).