A Selecting the Graphs H^+ and H^- in Algorithm 5

Algorithm 6: Find graphs H^+ and H^-

- *Inputs:* $\mathfrak{X}$, $\widehat{E}$ and $\widehat{F}$
- **for** $\ell = 1, \dots, L$
  - Estimate a graph $\widehat{G}^{(\ell)}$ using random subsamples such that
    - $(i,j) \in E(\widehat{G}^{(\ell)}) \; \forall \; (i,j) \in \widehat{E}$
    - $(i,j) \notin E(\widehat{G}^{(\ell)}) \; \forall \; (i,j) \in \widehat{F}$
- $Q_{ij} \leftarrow$ Fraction of times the edge (i,j) appears in the graphs $\widehat{G}^{(1)}, \dots, \widehat{G}^{(L)}$
- Find H^+ s.t. $(i,j) \in E(H^+)$ for $Q_{ij} \geq \alpha^+$
- Find H^- s.t. $(i,j) \in E(H^-)$ for $Q_{ij} \geq \alpha^-$
- **return** H^+ and H^-

In this section, we discuss the algorithm to estimate the graphs H^+ and H^- in Algorithm 2. The main idea, outlined in Algorithm 6 above, is to use stability selection [21] to estimate edges in the unknown graph G^*. We first estimate multiple different graphs using L (30 in our simulations) randomly subsampled measurements (Line 1). The graphs are estimated in such a way that all edges in $\widehat{E}$ are in the estimated graph and all edges in $\widehat{F}$ are not in the estimated graph. This is done so that the graphs estimated are consistent with prior estimates of G^*. Next, for each edge $(i,j) \in V \times V$, we compute the fraction of times it appears in one of the estimated graphs. We store all these values in the matrix Q_{ij}. We choose H^- so that it contains all edges for which $Q_{ij} \geq \alpha^-$. We choose H^+ so that it contains all edges for which $Q_{ij} \geq \alpha^+$. Both α^- and α^+ influence the performance of the active learning algorithm. We conservatively choose them so that $\alpha^- = 1.0$ and $\alpha^+ = 0.1$.

B Proof of Theorem 4.1

In this section, we analyze Algorithm 3. The proof methodology is motivated from [24]. Throughout this section, we assume that $\widehat{G} = \mathsf{CIT}(\mathfrak{X}^n, \eta, \tau_n)$, where CIT is outlined in Algorithm 1. We are interested in finding conditions under which $\widehat{G} = G^*$ with high probability. To this end, define the set B_η as follows

$$B_\eta = \{(i,j,S) : i,j \in V, i \neq j, S \subseteq V \backslash \{i,j\}, |S| \leq \eta\} \, . \tag{8}$$

The following concentration inequality follows from [24] and [4]

Lemma B.1. *Under Assumptions (A1) and (A3), there exists constants c_1 and c_2 such that for $0 < \epsilon < 1$,*

$$\sup_{(i,j,S) \in B_\eta} \mathbb{P}\left(||\rho_{ij|S}| - |\widehat{\rho}_{ij|S}|| > \epsilon \right) \leq c_1 \exp\left(-c_2 (n - \eta) \epsilon^2 \right) \, , \tag{9}$$

where C_1 is a constant, and n is the number of vector valued measurements made of X_i, X_j, and X_S.

Proof. Applies Lemma 2 from [24] to the result in Lemma 18 in [4]. □

Let $p_e = \mathbb{P}(\widehat{G} \neq G)$, where the probability measure $\mathbb{P}$ is with respect to P_X. Recall that we threshold the empirical conditional partial correlation $\widehat{\rho}_{ij|S}$ to test for conditional independence, i.e., $\widehat{\rho}_{ij|S} \leq \tau_n \implies X_i \perp\!\!\!\perp X_j | X_S$. An error may occur if there exists two distinct vertices i and j such that either $\rho_{ij|S} = 0$ and $|\widehat{\rho}_{ij|S}| > \lambda_n$ or $|\rho_{ij|S}| > 0$ and $|\widehat{\rho}_{ij|S}| \leq \tau_n$. Thus, we have

$$p_e \leq \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2) \, , \tag{10}$$

$$\mathbb{P}(\mathcal{E}_1) = \mathbb{P}\left(\bigcup_{(i,j) \notin G} \{\exists \; S \text{ s.t. } |\widehat{\rho}_{ij|S}| > \lambda_n\} \right) \tag{11}$$

$$\mathbb{P}(\mathcal{E}_2) = \mathbb{P}\left(\bigcup_{(i,j) \in G} \{\exists \; S \text{ s.t. } |\widehat{\rho}_{ij|S}| \leq \lambda_n\} \right) \, . \tag{12}$$

We will find conditions under which $\mathbb{P}(\mathcal{E}_1) \to 0$ and $\mathbb{P}(\mathcal{E}_2) \to 0$ which will imply that $P_e \to 0$. The term $\mathbb{P}(\mathcal{E}_1)$, probability of including an edge in $\widehat{G}$ that does not belong to the true graph, can be upper bounded as follows:

$$\mathbb{P}(\mathcal{E}_1) \leq \mathbb{P}\left(\bigcup_{(i,j)\notin G} \{\exists\ S \text{ s.t. } |\widehat{\rho}_{ij|S}| > \tau_n\}\right) \leq \mathbb{P}\left(\bigcup_{(i,j)\notin G, S\subset V\setminus\{i,j\}} \{|\widehat{\rho}_{ij|S}| > \tau_n\}\right) \tag{13}$$

$$\leq p^{\eta+2} \sup_{(i,j,S)\in B_\eta} \mathbb{P}\left(|\widehat{\rho}_{ij|S}| > \tau_n\right) \tag{14}$$

$$\leq c_1 p^{\eta+2} \exp\left(-c_2(n-\eta)\tau_n^2\right) = c_1 \exp\left((\eta+2)\log(p) - c_2(n-\eta)\tau_n^2\right) \tag{15}$$

The terms $p^{\eta+2}$ comes from the fact that there are at most p^2 number of edges and the algorithm searches over at most p^η number of separators for each edge. Choosing τ_n such that $\tau_n > c_1(\eta+2)\log p/(n-\eta)$ ensures that $\mathbb{P}(\mathcal{E}_1) \to 0$ as $p \to \infty$.

Suppose we select $\tau_n < c_3\rho_{\min}$ for a constant $c_3 < 1$. The term $\mathbb{P}(\mathcal{E}_2)$, probability of not including an edge in $\widehat{G}$ that does belong to the true graph, can be upper bounded as follows:

$$\mathbb{P}(\mathcal{E}_2) \leq \mathbb{P}\left(\bigcup_{(i,j)\in G} \{\exists\ S \text{ s.t. } |\widehat{\rho}_{ij|S}| \leq \tau_n\}\right) \tag{16}$$

$$\leq \mathbb{P}\left(\bigcup_{(i,j)\in G, S\subset V\setminus\{i,j\}} |\rho_{ij|S}| - |\widehat{\rho}_{ij|S}| > |\rho_{ij|S}| - \tau_n\right) \tag{17}$$

$$\leq p^{\eta+2} \sup_{(i,j,S)\in B_\eta} \mathbb{P}\left(|\rho_{ij|S}| - |\widehat{\rho}_{ij|S}| > |\rho_{ij|S}| - \tau_n\right) \tag{18}$$

$$\leq p^{\eta+2} \sup_{(i,j,S)\in B_\eta} \mathbb{P}\left(||\rho_{ij|S}| - |\widehat{\rho}_{ij|S}|| > \rho_{\min} - \tau_n\right) \tag{19}$$

$$\leq c_1 p^{\eta+2} \exp\left(-c_2(n-\eta)(\rho_{\min}-\tau_n)^2\right) = c_1 \exp\left((\eta+2)\log(p) - c_4(n-\eta)\rho_{\min}^2\right). \tag{20}$$

To obtain (20), we use the choice of τ_n so that $(\rho_{\min} - \tau_n) > (1-c_3)\rho_{\min}$. For an appropriate constant $c_5 > 0$, choosing $n > \eta + c_5\rho_{\min}^{-2}(\eta+2)\log(p)$ ensures $\mathbb{P}(\mathcal{E}_2) \to 0$ as $n, p \to \infty$. We note that the choice of c_5 only depends on M. This concludes the proof. □

C Proof of Theorem 6.1

In this section, we analyze Algorithm 4. Recall the assumption that there exists sets of vertices V_1, V_2, and T such that there are no edges between $V_1\setminus T$ and $V_2\setminus T$ in G^*. Note that we only assume the *existence* of these clusters and the corresponding graph decomposition. Now, let $\widehat{G}$ be the graph estimated after Step 2 of Algorithm 4, i.e., after drawing n_0 measurements. Define the event $\mathcal{D}$ as

$$\mathcal{D} = \{\widehat{G}[V_2] = G^*[V_2] \text{ and } \forall\ i \in V_1\setminus T \text{ and } \forall\ j \in V_2\setminus T,\ (i,j) \notin \widehat{G}\}.$$

In words, $\mathcal{D}$ defines the event that after n_0 measurements, the CIT algorithm is able to accurately identify all the edges and the non-edges over V_2 and all the non-edges that connect V_1 and V_2. Given that $\mathcal{D}$ is true, it is easy to see that for any two-cluster decomposition of $\widehat{G}$ over clusters $\widehat{V}_1$ and $\widehat{V}_2$ such that $\widehat{G}[\widehat{V}_2] = G^*[\widehat{V}_2]$, we have that $\widehat{V}_1 \subseteq V_1$ and $V_2 \subseteq \widehat{V}_2$.

Let $\widehat{G}_1$ be the graph estimated in Step 5 of Algorithm 4 and let $\widehat{G}_F = \widehat{G}[\widehat{V}_2] \cup \widehat{G}_1$ be the output of Algorithm 4. Conditioning on $\mathcal{D}$, and using the assumption that $\widehat{V}_1 = V_1$, we have

$$\mathbb{P}(\widehat{G}_F \neq G^*) = \mathbb{P}(\widehat{G}_F \neq G^*|\mathcal{D})\mathbb{P}(\mathcal{D}) + \mathbb{P}(\widehat{G}_F \neq G^*|\mathcal{D}^c)\mathbb{P}(\mathcal{D}^c) \tag{21}$$

$$\leq \mathbb{P}(\widehat{G}_F \neq G^*|\mathcal{D}) + \mathbb{P}(\mathcal{D}^c). \tag{22}$$

We now make use of Theorem 4.1. Given the scaling of n_0, it is clear that $\mathbb{P}(\mathcal{D}^c) \to 0$ as $p \to \infty$. Furthermore, given that $\mathcal{D}$ holds, we only need to estimate the edges over $\widehat{V}_1$ in Step 5 of Algorithm 4. Since $\widehat{V}_1 \subseteq V_1$ when given $\mathcal{D}$, it follows that $\mathbb{P}(\widehat{G}_F \neq G^*|\mathcal{D}) \to 0$ as $p \to \infty$ given the scaling of $n_0 + n_1$. This concludes the proof. □

D Proof of Theorem 6.3

Once V_1 and V_2 have been identified, by the global Markov property of graphical models, the graph learning can be decomposed into two independent problems of learning the edges in $G[V_1]$ and learning the edges in $G[V_2]$. Thus, $p_e(\psi)$ can be lower bounded by

$$\max_{\Theta(G)} \left\{ \max \left[\mathbb{P}(\psi(\mathfrak{X}^n_{V_1}) \neq G[V_1]), \mathbb{P}(\psi(\mathfrak{X}^n_{V_2}) \neq G[V_2]) \right] \right\} ,$$

where $\Theta(G) \in \mathcal{G}_{p,p_1,p_2,\eta,d}(\theta_1, \theta_2)$. By definition, we know that $G[V_1]$ is sampled uniformly from $\mathcal{G}_{p_1,\eta,d}$. By identifying that $\mathcal{G}_{p_1,0,d} \subseteq \mathcal{G}_{p_1,\eta,d}$, we can now make use of the results in [29] for degree bounded graphs. In particular, we have from [29] that if

$$n \leq \max \left\{ \frac{\log \binom{p_1 - d}{2} - 1}{4\theta_1^2}, \frac{\log \binom{p_2 - d}{2} - 1}{4\theta_2^2} \right\} ,$$

then $p_e(\psi) \to 1$ as $n \to \infty$. This leads to the necessary condition in the statement of the theorem and concludes the proof. □

E Numerical Results on Scale-Free Graphs

Tables 2 shows results for scale-free graphs. It is typical for scale-free graphs to contain a small number of vertices that act as hubs and are connected to many other vertices in the graph. The inverse covariance is constructed as in the Hub graph case. For this graphical model, the weak edges correspond to all edges that connect to vertices with high degree. We again see that active learning results in superior performance than passive learning.

Table 2: Scale-free graph with $p = 400$ vertices

		Oracle Results			Model Selection Results		
n	Alg	TPR	FDR	ED	TPR	FDR	ED
200	Nonactive	0.405 (0.001)	0.059 (0.002)	247 (0.410)	0.382 (0.000)	0.033 (0.000)	251 (0.121)
	Active	0.422 (0.001)	0.040 (0.002)	237 (0.402)	0.391 (0.000)	0.017 (0.000)	245 (0.100)
400	Nonactive	0.522 (0.002)	0.043 (0.002)	200 (0.341)	0.500 (0.000)	0.023 (0.000)	204 (0.107)
	Active	0.545 (0.001)	0.036 (0.002)	189 (0.340)	0.509 (0.000)	0.001 (0.000)	197 (0.121)
600	Nonactive	0.605 (0.001)	0.0346 (0.001)	166 (0.361)	0.582 (0.000)	0.021 (0.000)	171 (0.123)
	Active	0.634 (0.001)	0.0321 (0.001)	154 (0.378)	0.592 (0.000)	0.008 (0.000)	164 (0.131)