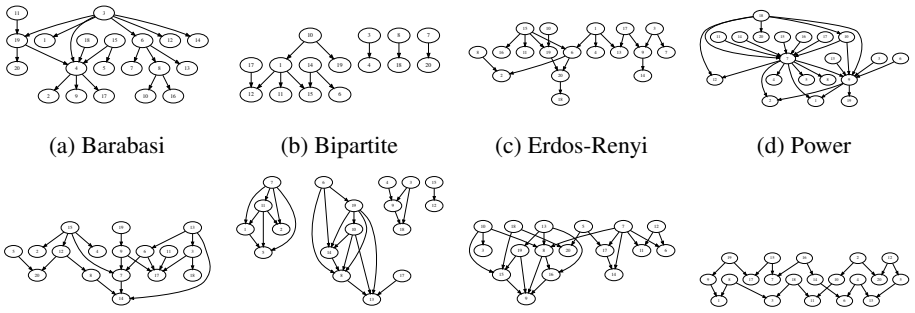
Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

1 Example DAGs

In order to evaluate the different causal discovery methods we generated random Directed Acyclic Graphs (DAGs) using eight different generation methods of the randDAG function of the R pcalg library[1]. These generation methods were *regular* - a graph where every node has exactly d incident edges, *er* - an Erdos-Renyi graph where every edge is present independently, *watts* - an interpolation between regular graph and Erdoes-Renyi graph, *power* - a graph with power-law degree distribution, *bipartite* - a bipartite graph, *barabasi* - a graph with power-law degree distribution and preferential attachment, *geometric* - a geometric random graph, and *interEr* - a graph with two islands of Erdoes-Renyi graphs connected by a small number of edges. An example of a graph produced by these different methods can be seen in Figure 1.



(a) Barabasi (b) Bipartite (c) Erdos-Renyi (d) Power

(e) Watts (f) Geometric (g) Inter-Er (h) Regular

Figure 1: Examples of causal graphs with $N = 20$ nodes and $E = 2$ expected edges per node, generated using each method of the PCALG random DAG function. We generated 1140 graphs total with $N = 20, N = 40, N = 60$ nodes and $E = 1$, $E = 2$, $E = 3$, $E = 4$, and $E = 5$ expected edges per node.

2 Stability Algorithm Selection Heuristic

We focus on detailed evaluation of the agreement algorithm selection heuristic in the main paper, but we also explored an alternative heuristic of algorithm stability, or robustness to sampling of the data, as another potential indicator of algorithm accuracy. We measure the stability of a given algorithm by applying the algorithm to multiple samples of the data. For each edge or node that appears in any graph generated from these samples, we calculated the percentage of the graphs in which the given edge or node appears. An edge which appears in all graphs is highly stable, while one that appears only in one is not. This definition of stability focuses on the consistency of edge presence rather

[1] https://www.rdocumentation.org/packages/pcalg/versions/2.6-8/topics/randDAG

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

than the consistency of edge absence. To calculate the overall stability of the graph, we average the stability of each individual edge or node.

In Figure 2, we show that algorithm stability is positively correlated with the accuracy of the algorithm. In fact, the correlation is observed to be positive for all graph generation methods and parameters. However, the level of correlation is significantly less on average than is observed for the agreement scores. Therefore, we expect that the agreement score will provide a better indicator for accuracy and incorporate this score in both the algorithm selection and ensemble methods that we develop.

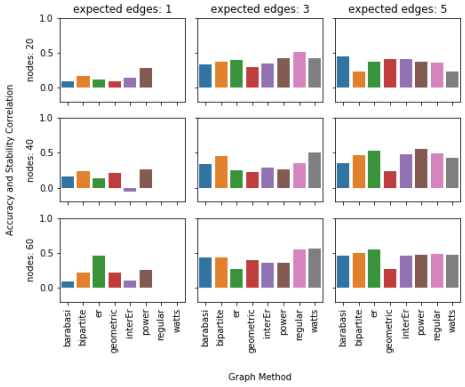


Figure 2: The Spearman rank correlation of algorithm accuracy and algorithm stability for PCALG data generated from graphs with different structural properties.

3 Parameter Optimization

The best performing ensemble method used the subseleted four algorithms, no upweighting of agreeable algorithms, and an edge weight threshold of 0.65. In Figure 3, we show the a comparison of the ensemble performance across the algorithm selection and upweighting ratio parameters for datasets generated from graphs with different structural properties. We find that for the majority of graph types, the four-algorithm ensemble outperforms the full ensemble for the full range of ratio settings. This indicates that the performance of the ensemble approach is sensitive to the selection of specific algorithms to include, with lower performing algorithms reducing the overall performance of the ensemble.We also find that the two different algorithm selections have different patterns with respect to the upweighting ratio. We find that increasing the weight of the most agreeable algorithm increases the performance of the full ensemble. In contrast, for the four-algorithm ensemble, we find that upweighting the most agreeable algorithms reduces the performance. This difference is likely due to two factors. Firstly, with fewer algorithms included the overall agreement score may become less meaningful because there are fewer inter-algorithm comparison points. Secondly, because the agreement scores are correlated with accuracy, the agreement score upweighting in the full ensemble is effectively downweighting the less accurate algorithms. These algorithms are excluded entirely from the selective ensemble, so the downweighting becomes less necessary.

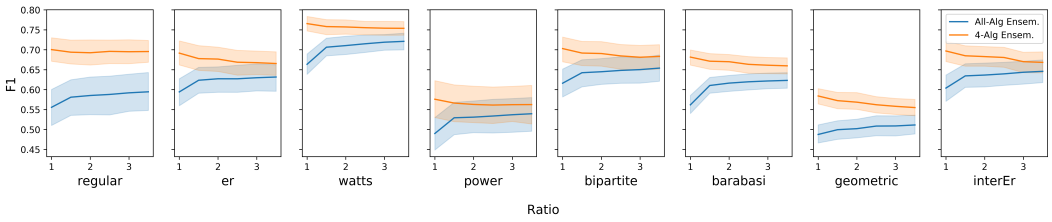


Figure 3: Causal discovery performance for the all-algorithm and the four-algorithm ensembles as a function of how much the most agreeable algorithm is upweighted (the ratio of the most agreeable algorithm weight to the least agreeable weight) evaluated on PCALG data.