The aim in many sciences is to understand the mechanisms that underlie the observed distribution of variables, starting from a set of initial hypotheses. Causal discovery aims to find meaningful causal relationships using large-scale observational data. Causal relationships are often represented as a graph, where nodes are random variables and directed edges are cause-effect relationships between random variables (Spirtes et al., 2000b). Causal graphs have high expressive power as they allow us to investigate complex relationships between many variables simultaneously—making them relevant for many problems in science, economics, and decision systems (Pearl, 1995).

Exploring the graph search space to find the causal graph is an NP-hard problem. Causal discovery algorithms have benefited from some performance enhancements and parallelization strategies (Ramsey, 2015; Laborda et al.,2023; Lee & Kim, 2019). Recent work explores a distributed divide-and-conquer version of causal discovery by partitioning variables into subsets, locally estimating graphs, and merging graphs to resolve a causal graph. Divide-and-conquer methods do not provide theoretical guarantees for consistency; meaning in the infinite data limit they do not necessarily find the Markov Equivalence Class of the true causal graph. Existing algorithms also rely on an extra learning step to merge graphs which can be computationally expensive. Finally, these algorithms ignore the violations to causal assumptions when learning on subsets of variables (Spirtes et al., 2000b; Eberhardt, 2017).

To address these limitations in literature, we define a novel ***causal graph partition*** that allows for divide-and-conquer causal discovery with theoretical guarantees under the Maximal Ancestral Graph (MAG) class. A causal partition is a graph partition of the hypothesis space, defined by a superstructure, into overlapping variable sets. A causal partition allows for merging locally estimated graphs without an additional learning step. We can efficiently create a causal partition from any disjoint partition. This means that a causal partition can be an extension to any graph partitioning algorithm.

We are interested in causal discovery for high-dimensional scientific problems; in particular, biological network inference. Biological networks are organized into hierarchical scale-free sub-modules (Albert, 2005; Wuchty et al., 2006; Ravasz, 2009). The causal partition allows us to leverage the inherent, interpretable communities in these networks for scaling. Our contributions are as follows: **(A)** We define a novel causal partition which leverages a superstructure and extends any disjoint partition. **(B)** We prove, under certain assumptions, that learning with a causal partition is consistent without an additional learning procedure. **(C)** We show the efficacy of our algorithm on synthetic biologically-tuned networks up to 10,000 nodes.

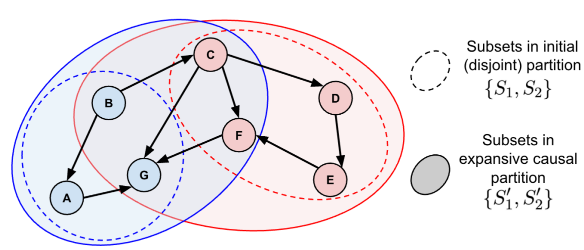

Figure: A causal partition, shown in solid lines, is an expansion of any disjoint partition, shown in dashed lines, by including the nodes in the ***outer boundary*** of the initial disjoint partition. This simple extension allows for an efficient creation of a causal partition from any initial partition. We show a causal partition enables ~13x (~25 hours to ~2 hours) speedup of causal discovery on 10,000 node graphs.