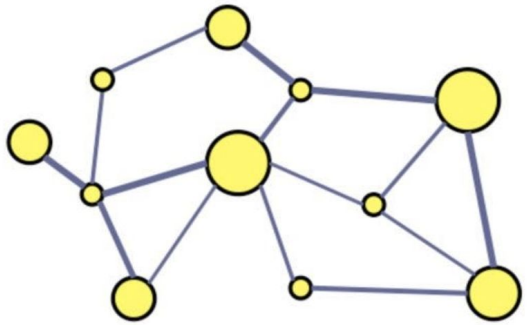
Prize-Collecting Steiner Tree (PCST): An Extension of MST for Genomics Problems

Davyd Sadovskyy

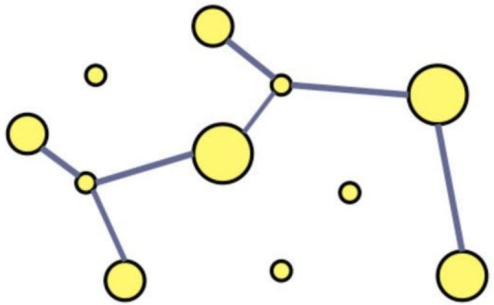
MST, ST, PCST

- Minimum Spanning Tree (MST)
  - Every node in graph must be visited
  - Goal is to minimize total edge cost
- Steiner Tree (ST)
  - Only a set of “terminal nodes” in graph must be visited
  - Goal is to minimize total edge cost
  - Additional nodes may be visited if doing so is cheaper
- Prize Collecting Steiner Tree (PCST)
  - Each node has an associated prize (value/reward)
  - Goal is to maximize total prize − total edge cost
  - No requirement on which nodes need to be visited

A) Input Graph, G = (V, E)



B) Prize-collecting Steiner Tree, G’ = (V’, E’)



Comparing IP Formulations

	MST	PCST
Objective Function	$\min \sum_{ij \in E} c_{ij} x_{ij}$	$\min \sum_{ij \in E} c_{ij} x_{ij} + \sum_{i \in V} p_i (1 - y_i)$

- Selecting a vertex incurs a prize which reduces objective value

	MST	PCST
Tree Size Constraint	$$\sum_{ij \in E} x_{ij} = n - 1$$ • A tree spanning all n vertices always has n-1 edges.	$$\sum_{ij \in E} x_{ij} \leq \sum_{i \in V} y_i - 1$$ • Only selected vertices must form a tree. • ≤ allows feasible intermediate solutions.
Subtour Elimination Constraint	$$\sum_{ij \in E : i,j \in S} x_{ij} \leq \lvert S \rvert - 1 \qquad \forall S \subseteq V, S \neq \emptyset$$ • Prevents cycles in any subset S. • Uses |S| because every vertex is required to be included	$$\sum_{ij \in E : i,j \in S} x_{ij} \leq \sum_{i \in S} y_i - 1 \qquad \forall S \subseteq V, S \neq \emptyset$$ • Prevents cycles among the **selected** vertices in S.
Variable Domain Constraint	$$x_{ij} \in \{0,1\} \qquad \forall ij \in E$$ • Each edge is either chosen or not. • MST does not need vertex variables since all vertices are included automatically	$$x_{ij} \in \{0,1\} \qquad \forall ij \in E \qquad y_i \in \{0,1\} \qquad \forall i \in V$$ • PCST introduces vertex selection variables yi.
Edge Vertex Consistency Constraint	None (all vertices are included by definition)	$$x_{ij} \leq y_i, \qquad x_{ij} \leq y_j \qquad \forall ij \in E$$ • An edge can only be selected if both of its endpoints are selected.
Rooting Constraint	None	$$y_r = 1$$ • Fixes a root to anchor the connected solution. • Prevents trivial solution x=0, y=0

PCST Algorithms

(Minkoff, 2000)

- PCST is NP-Hard. There are no MST-style polynomial algorithm (Prim/Kruskal/etc.)
- Unlike MST, PCST's LP relaxation isn't integral
- Best-known classical method is Goemans-Williamson primal-dual moat-growing algorithm
  - Gives a provably good approximate tree.
  - Too slow for large genomics networks.
- A modern scalable variant is Fast-PCST
  - same guarantee as Goemans-Williamson, but runs in nearly linear time. O(mlogn)

FAST PCST - Core Idea

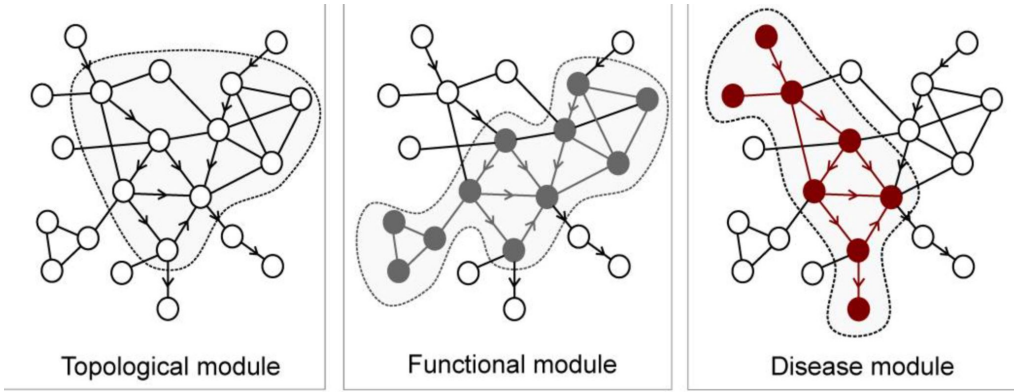
(Minkoff, 2000)

- Start with each node as its own cluster and grow “moats” (their dual variables) outward evenly.
- Whenever a moat hits an edge of the right cost, merge the two clusters. This simulates picking that edge.
- Stop when clusters have grown enough to justify their prize, then prune unused branches.

Network Medicine

(Barabási et al., 2011).

- Nodes = biological entities (genes, proteins).
- Edges = functional relationships (physical interaction, co-expression, regulation).
- A disease is not determined solely by the known function of the mutated gene, but also by the components with which the gene and its products interact. This represents a **disease module.**
- Sick Cell Disease
  - One mutated gene, but many symptoms (osteonecrosis, acute chest syndrome, stroke, and anemia)
  - The underlying disease module will likely include all disease modifying genes.



A disease is a result of the breakdown of a particular functional module

Impact

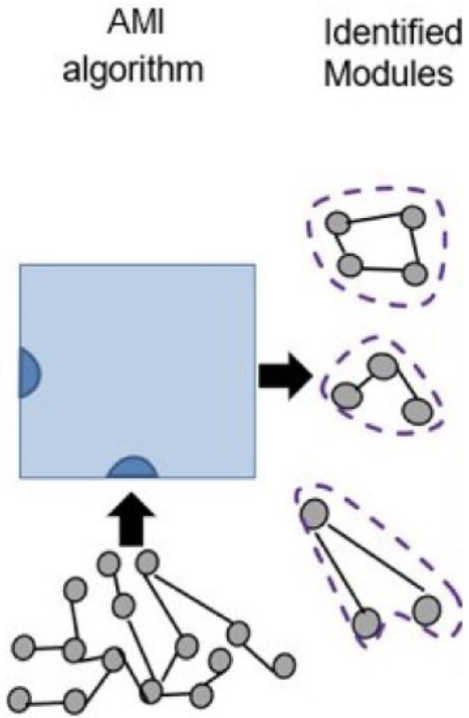
- Better Disease Treatment
  - Current approach: pick ONE gene/protein and make ONE drug that binds it.
  - Network approach: Identify the entire dysfunctional module driving disease and target several interacting components or the regulatory hub that controls them.

- Better Disease Prediction
  - For a patient, collect their gene expression data and assign these values to the nodes in a graph.
  - Check if the patient shows unusually strong perturbation in each module.

Active Module Identification

(Levi et al., 2021)

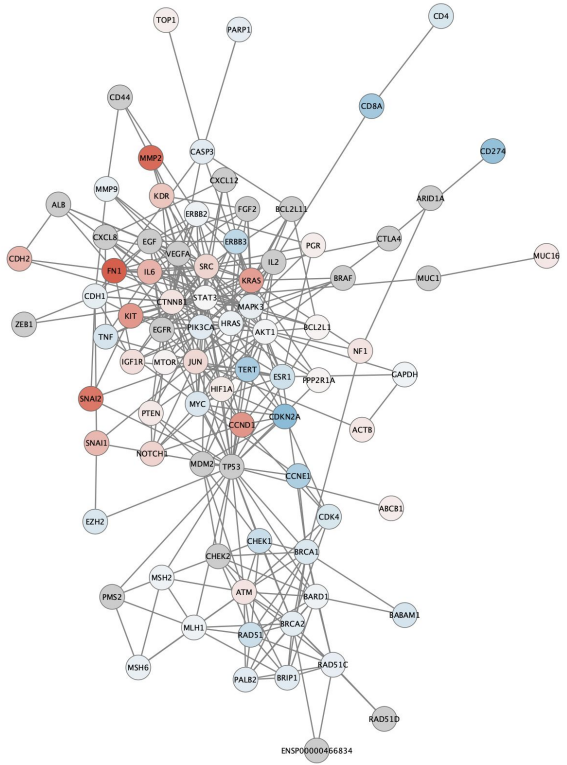
- Given a biological network (genes as nodes, interactions as edges) and an “activity score” for each gene, find a connected subgraph (module) whose total activity is high.
- Activity score is any numeric measure of how strongly a gene is implicated in the disease
  - Differential gene expression
  - GWAS based gene scores
  - Proteomics fold-changes
  - Methylation differences



AMI Algorithm: DOMINO (Discovery of active Modules In Networks using Omics)

(Levi et al., 2021)

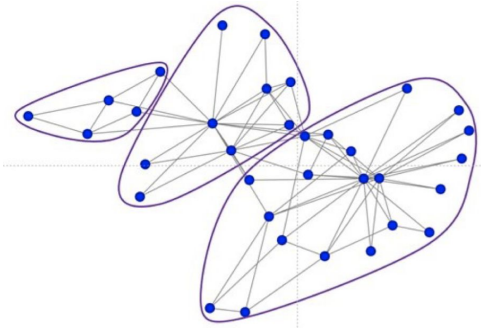
- Biological interaction graph
  - ~18,000 human genes (nodes)
  - Binary protein-protein interactions which represent gene-gene interactions (edges)
- Node Values (per gene activity score)
  - Values represents a particular gene's association with a given disease.
  - Derived from GWAS analysis.
- Active Module Identification
  - Find connected subgraphs (modules) with high total gene score.
  - DOMINO does this by incorporating a prize-collecting steiner tree (PCST)



Starting network is for one disease. Node Values are based on gene association to that disease.

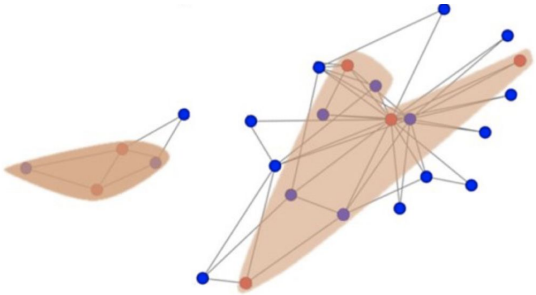
3 Main DOMINO Steps

(Levi et al., 2021)



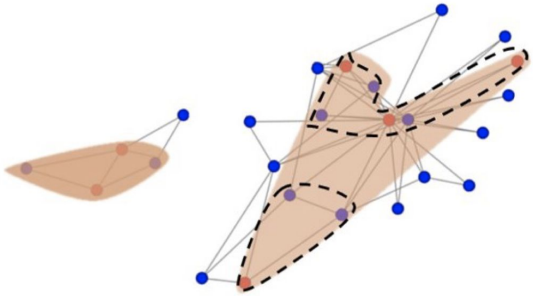
Step 1 Community Detection (Clustering)

- Clustering based on edge weights (gene-gene interactions)
- Reduces the network into manageable connected regions.
- The result represents coarse regions of the network that may contain biological signal.



Step 2 Optimization

- Applies Prize Collecting Steiner Tree (PCST) on step 1 clusters.
- Selects as many disease relevant genes as possible (collect prizes) while keeping the subgraph as small and connected as possible (minimize cost) and adds only the necessary intermediate genes that link them.
- Each orange region represents the minimal connected set of genes that explain the disease.

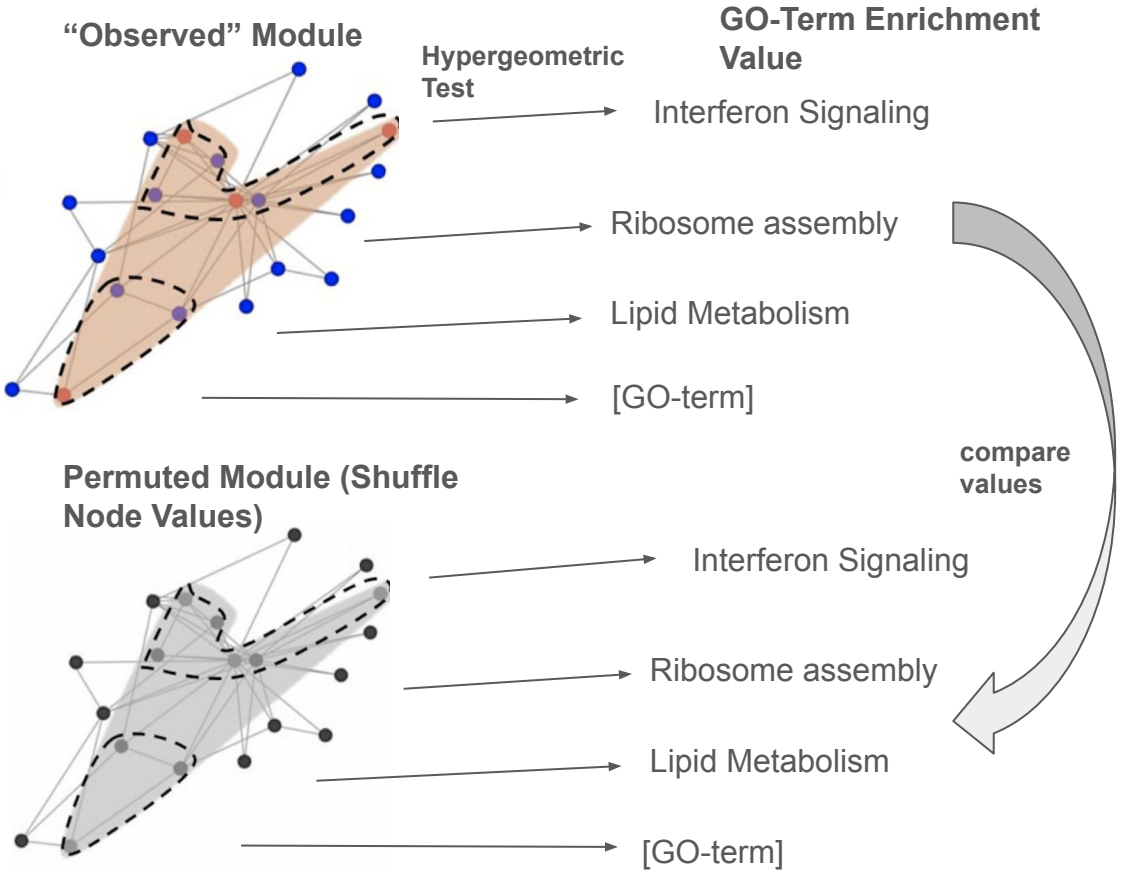


Step 3: Community Detection (Clustering)

- Clustering based on edges weights.
- The resulting clusters represent groups of genes that function together in a common process (immune activation vs metabolic regulation)

DOMINO Step 4 - Validation of Final Clusters

(Levi et al., 2021)



- <u>Are genes getting selected into modules simply because they are in a highly connected regions?</u>
- If a GO-Term is significant in observed module, but not in permuted module, then it's "real disease biology".
- If GO-Term is significant in both observed and permuted modules, there it's just a network artifact.

References

Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). *Network medicine: a network-based approach to human disease*. *Nature Reviews Genetics*, 12 (1), 56–68. doi:10.1038/nrg2918.

Minkoff, M. (2000). *The Prize Collecting Steiner Tree Problem* (Master's thesis). Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Levi, H., Elkon, R., & Shamir, R. (2021). *DOMINO: a network-based active module identification algorithm with reduced rate of false calls*. *Molecular Systems Biology*, 17, MSB20209593. doi:10.15252/msb.20209593.