Protein Interactions and Biological Networks

Xiang Fu

October 29, 2024

1. Measuring Protein-Protein Interactions (PPIs)

Affinity purification+ mass spectrometry (AP-MS)

- Principle: Isolate protein complexes under native conditions
- Key steps:
  - Tag protein of interest ("bait")
  - Express in cells
  - Lyse cells and purify complexes
  - Identify interacting partners via MS
- Advantages:
  - Detects interactions in native cellular context
  - Can identify entire complexes simultaneously
  - Quantitative data possible
- Limitations:
  - May lose weak/transient interactions
  - Background contaminants can be problematic
  - Requires specialized equipment

—

Yeast-two hybrid (Y2H)

- Principle: Split transcription factor approach in yeast
- Components:
  - Binding Domain (BD): DNA-binding portion
  - Activation Domain (AD): Transcription activation portion
  - Reporter gene: Activated only when BD and AD interact
- Method:
  - Fuse bait protein to BD
  - Fuse prey protein to AD
  - Co-express in yeast
  - Screen for reporter gene activation
- Advantages:
  - Fast and relatively inexpensive
  - Highly scalable for screening
  - Works well for binary interactions
  - Can detect weak/transient interactions
- Limitations:
  - High false positive/negative rates
  - Artificial nuclear environment
  - Not suitable for membrane proteins
  - May miss context-dependent interactions

2. The Interactome

1. Definition and Scope
    - Network of all protein-protein interactions in an organism
    - Key examples:
        - Yeast interactome (6,000 proteins)
        - Human interactome (20,000 proteins)
        - Bacterial interactomes
    - Dynamic and context-dependent

2. Current State of Knowledge
    - Incomplete mapping
        - Estimated 130,000-650,000 interactions in humans
        - Only 40% discovered so far
    - Data quality issues
        - High false positive rates
        - Missing weak/transient interactions
    - The “Hairball Problem”
        - Visual representation becomes a tangled mess
        - Difficult to extract meaningful patterns
        - Challenge of information overload

3. Computational Analysis Approaches
    - Network Theory Applications
        - Graph theory metrics
        - Topology analysis
        - Centrality measures
    - Key Questions:
        - Which proteins are network hubs?
        - How are functional modules organized?
        - Can we predict protein function?
        - What are the critical nodes?

4. Analytical Goals
    - Identify functional protein clusters
    - Find central/important proteins
    - Predict protein functions
    - Understand disease mechanisms
    - Drug target identification
    - Pathway reconstruction

5. Emerging Methods
    - Machine learning approaches
    - Integration with other data types
        - Expression data
        - Structural information
        - Literature mining
    - Dynamic network analysis

- Temporal changes
- Condition-specific networks

3. Properties of PPI Networks

Very similar to social networks!

—

1. Small World

The **diameter** (the maximuym number of steps separating any two nodes) of the network is small, <6.

“six-degrees of separation”

How is diamter calcualted?

Calculate all shortest paths (e.g., with Dijksta’s algorithm) and take the max.

This level of connectivity has important biological consequences, since it allows for an **efficient and quick flow of signals** within the network.

However, it also poses an interesting question: if the network is so tightly connected, why don’t perturbations in a protein have dramatic consequences for the network? Biological networks have been shown to be **extremely robust**.

—

2. Scale Free

The **degree** of a node is number of edges it has. A network is scale free if its degree distribution follows a power law:

Most nodes have low degree.

A few “hubs” with very high degree.

Scale-free networks form naturally by preferential attachment.

“the rich get richer”.

Because they are scale free, protein networks are:

Stable

- If failures occur at random, it is far more likely that a non-hub protein would be affected than a hub
- If a hub failure occurs, the network will generally not lose its connectedness, due to the remaining hubs (hubs are very likely to be connect to other hubs)

—

2. Invaraint to changes of scale

- If we select any connected subgraph, it will also be scale-free. This means the subgraphs has the same properties

—

Vulnerable to targedted attack

- If failure is *not* random, knocking out hubs will fragment the graph

- Hubs are enriched with essential proteins. (For example, many cancer-linked proteins like the tumour suppressor p53 are hubs)

—

3. High clustering coefficient

The **clustering coefficient** of a network is a measure of the tendency of the nodes to cluster together.

A high clustering coefficient means that the network contains **communities** or groups of nodes that are densely connect internally.

“The friend of my friends are my friends”

How is the clustering coefficient calculated?

Count the triangles:

$$C = \frac{\text{number of closed triplets}}{\text{number of all triplets (open and closed)}} \tag{1}$$

—

For PPIs, these communities can reflect “modules”.

A **module** is an exchangeable functional unit. They are self-contained components of a system with well-defined interfaces with other components. The defining feature of a module is that its intrinsic functional properties do not change when it is plsced in a different context.

Modules also help define **intermodular interactions and proteins** These are the edges/ nodes that link different modules within a network. They can act as switches or high-level modulators that, for example, mediate cross-talk between different complexes or pathways.

Modules help reduce the complexity of biological networks by giving us a set of reducible, functional units that can be studied as an intergrated entity.

4. Topological Analysis of PPI Networks

Topological analysis: the study of the ferature sof a network.

Goal 1: Which proteins are the most important and why?

Centrality is a measure of how “central” (importamt) is each node in the graph.

—

Idea 1: Degree Centrality

Nodes with a high degree (hubs) are key in maintaining some characteristics of scale-free networks such as their robustness and the small-world effect.

However, degree is a **local** measure since it does not take into account the rest of the network.

—

Idea 2: Global Centrality

Global centrality measures take into account the whole network.

- Closeness centrality
- Betweeness centrality

—

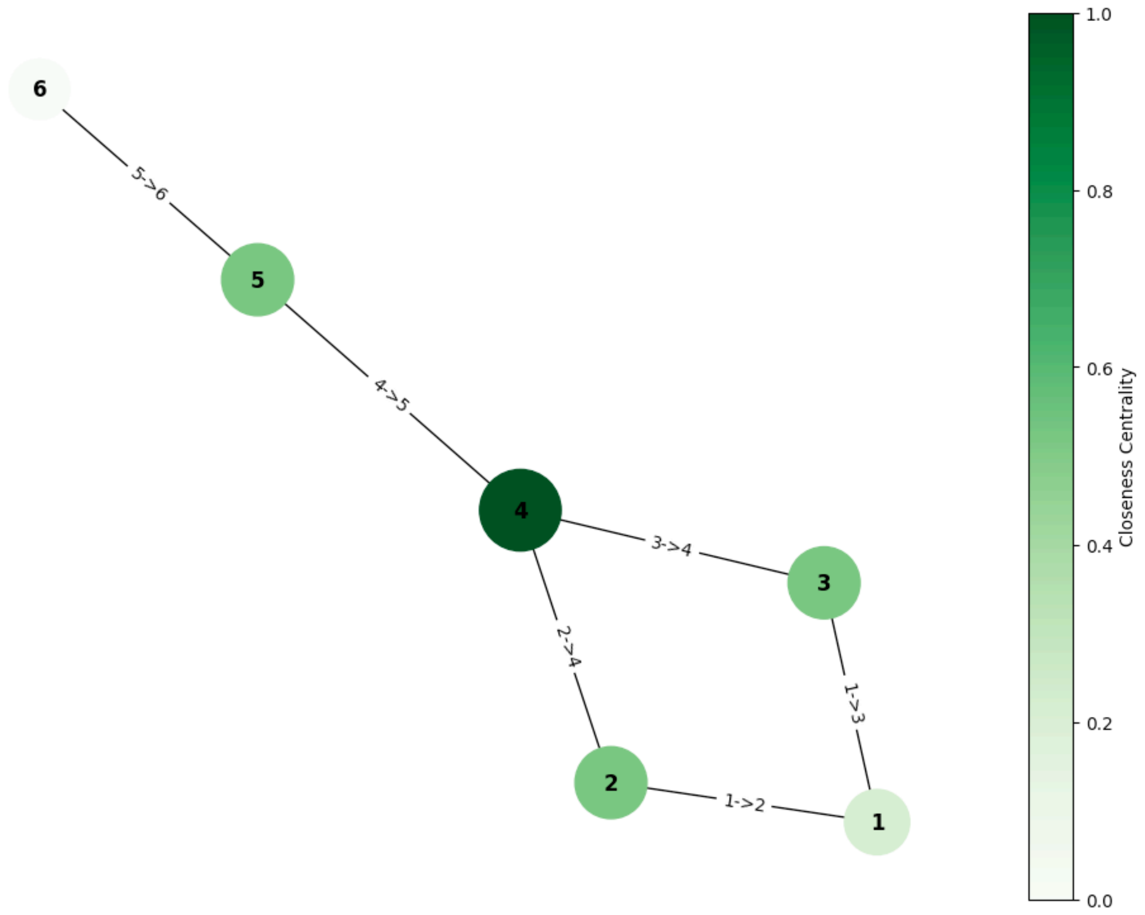
Global Centrality Measures - Closeness Centrality

Closeness centrality is based on the idea that a central node is close to *all* other nodes.

How is calculated?

Measure the length of shortest paths from the given node to all other nodes (again using Dijkstra’s). Closeness centrality is the mimum possible distance to all other nodes ($N - 1$) over the sum of actual shortest distances:

$$C(v) = \frac{N - 1}{\sum_u d(u, v)} \tag{2}$$



- Node 1: Closeness Centrality = 0.4545
- Node 2: Closeness Centrality = 0.5556
- Node 3: Closeness Centrality = 0.5556
- Node 4: Closeness Centrality = 0.7143
- Node 5: Closeness Centrality = 0.5556
- Node 6: Closeness Centrality = 0.3846

—

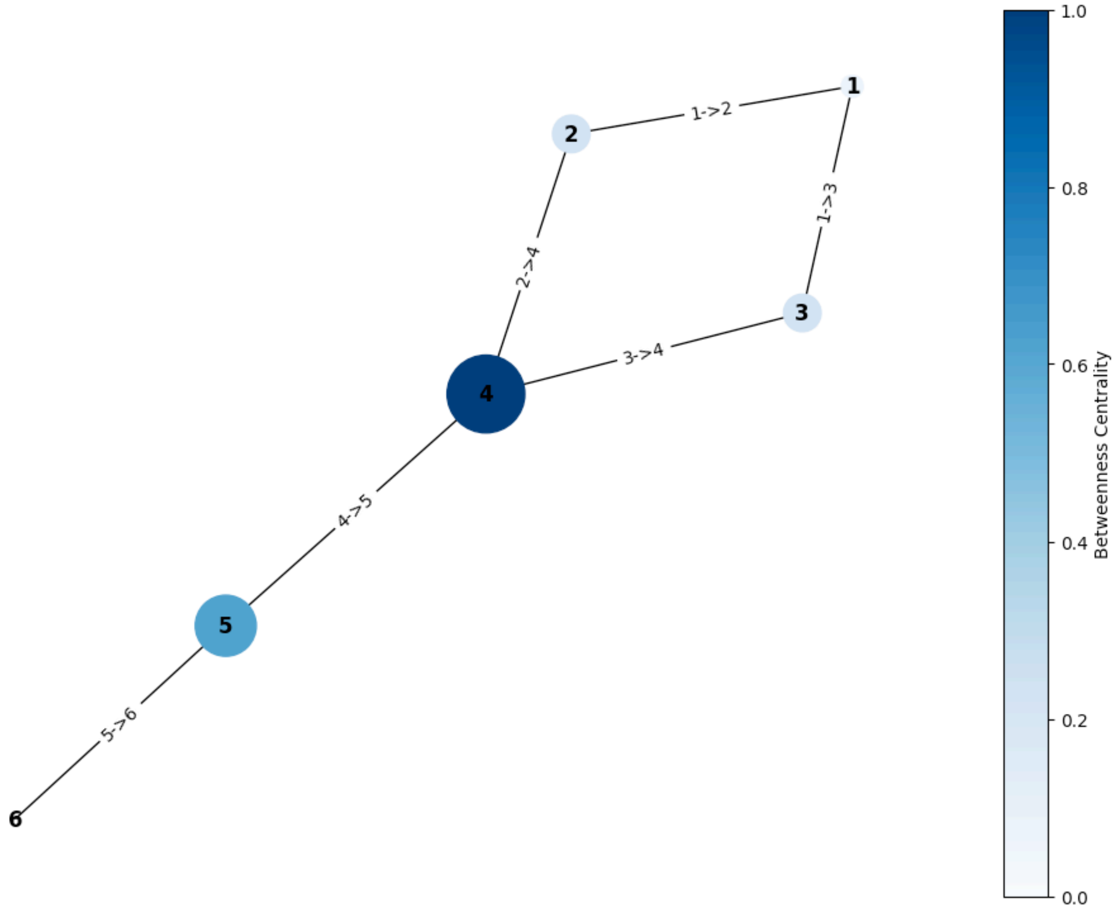
Global Centrality Measures - Betweenness Centrality

Betweenness centrality is based on the idea that soem nodes are important because they lie on communnication *paths*. They might represent important proteins in signaling pathways and can form targets for drug discovery.

How is calculated?

Find the number of shortest paths in the graph that pass through the node divided by the total number of shortest paths.

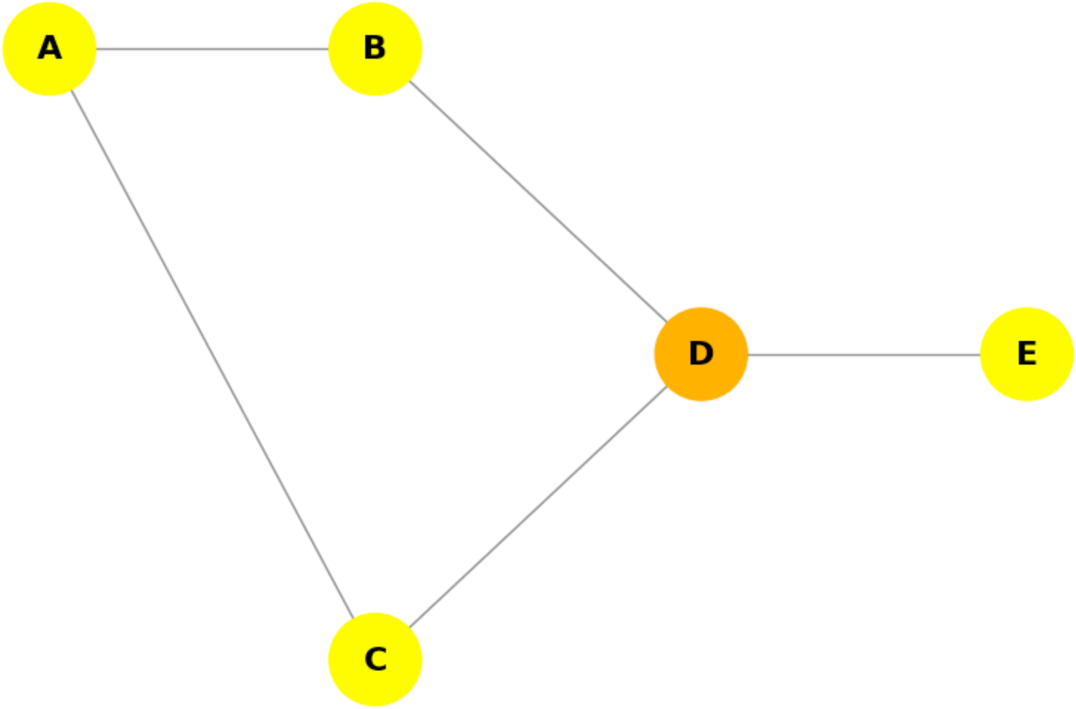
$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$



- Node 1: Betweenness Centrality = 0.0500
- Node 2: Betweenness Centrality = 0.1500
- Node 3: Betweenness Centrality = 0.1500
- Node 4: Betweenness Centrality = 0.6500
- Node 5: Betweenness Centrality = 0.4000
- Node 6: Betweenness Centrality = 0.0000

—

Another example:



All shortest paths that don't start or end on D:

- **A to B**: AB
- **A to C**: AC
- **A to E**: ABDE, ACDE
- **B to A**: BA
- **B to C**: BAC, BDC
- **B to E**: BDE
- **C to A**: CA
- **C to B**: CAB, CBD
- **C to E**: CDE
- **E to A**: EDBA, EDCA
- **E to B**: EDB
- **E to C**: EDC

Goal 2: Which groups of proteins form modules?

As a general approach, we can try to remove as few edges as possible to break apart the graph.

Approach 1: find the mimimum cuts (cuts that cross the fewest edges) in the graph and break it apart there

Approach 2: remove edges with the highest "edge betweenness" the equivalent of **betweeness centrality** for edges

Note: there are *many* other approaches to finding clusters in a network.

—

Finding Modules - Using the Mimimum Cut

Randomized min-cut algorithm

Pick an edge uniformly at random and merge the two vertices at its end-points

- If as a result there are several edges beteween some pairs of (newly-formed) vertices retain them all
- Edges between vertices that are merged are removed (no self-loops)

—

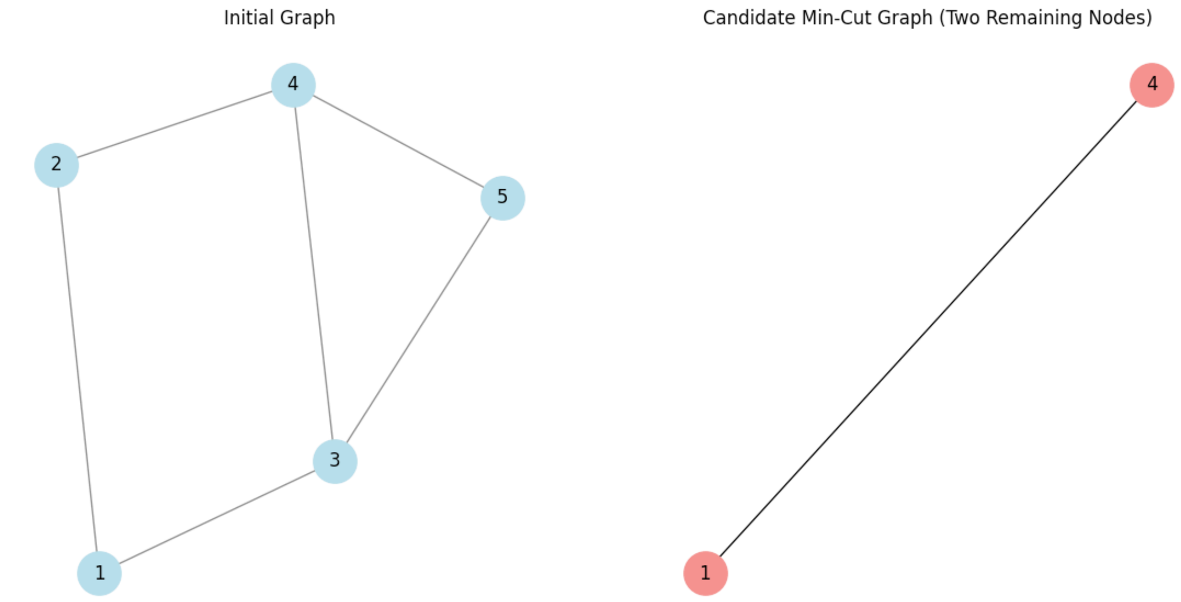
Repeat until only two vertices remain

The set of edges between these two vertices is a cut in the network and is output as a *candidate* min-cut.

If you run the algorithm $\frac{n^2}{2}$ times, the probability that a min-cut is not found becomes vanishingly small.

But this only splits the network in two, how do we get more modules?

Just run the algorithm recursively on the two networks! **Note:** this does *not* guarantee the optimal sets of cuts overall, but it is fast.


Initial Graph
4
2
5
3
1
Candidate Min-Cut Graph (Two Remaining Nodes)
4
1