Bioinformatics: Biological Networks

Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

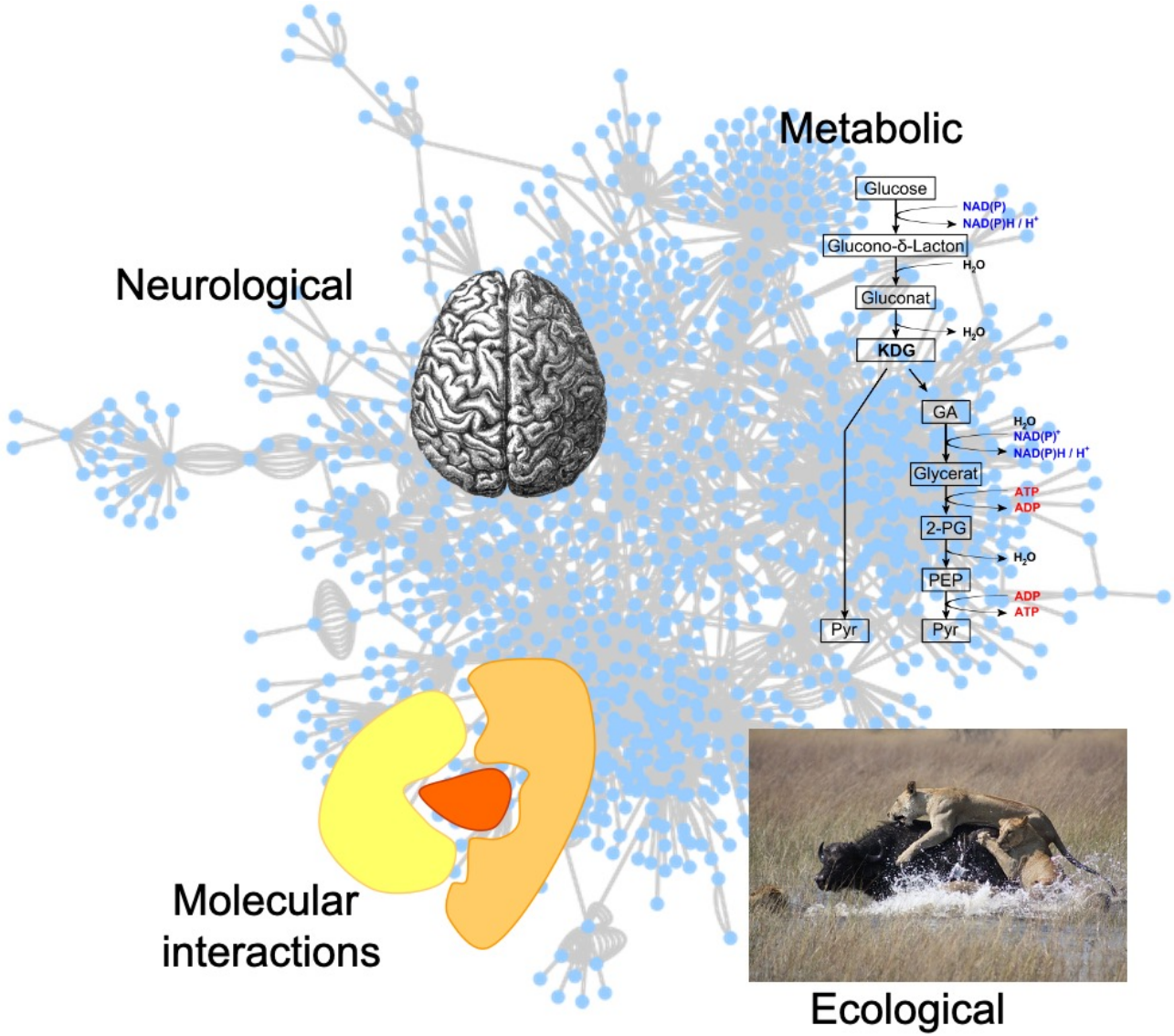
References

Network Biology

- Biological systems are often represented as networks which are complex sets of interactions or relations between different entities.
- Every biological entity has interactions with other biological entities, from the molecular to the ecosystem level, providing us with the opportunity to model biology using many different types of networks.
- The data explosion that originated in the -omics era of biological research necessitated the development of more systemic approaches to data analysis and a move away from the single gene/protein perspective.
- Network biology allows the representation and analysis of biological systems using tools derived from graph theory.
- Biological network analysis historically originated from the tools and concepts of social network analysis and the application of graph theory to the social sciences.

BRAC UNIVERSITY
Inspiring Excellence

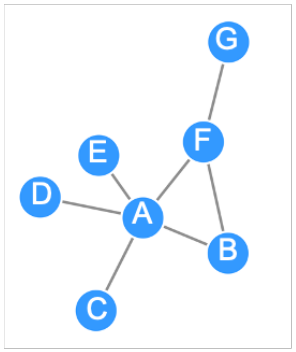
Network Biology



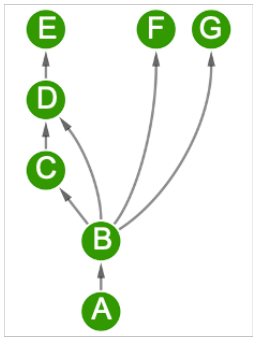
Graph types and edge properties

- The nodes represent different entities (e.g. proteins or genes in biological networks), and edges convey information about the links between the nodes.

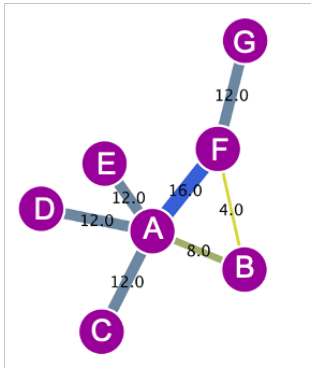
Undirected



Directed



Weighted

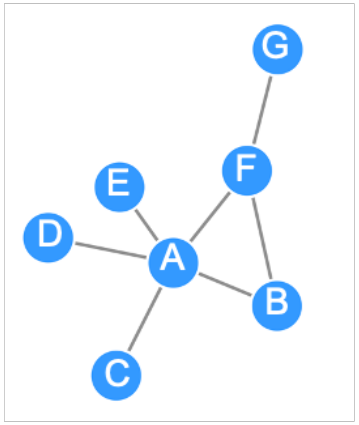


- Undirected edges - Protein Protein Interaction Networks
- Directed edges - metabolic or gene regulation networks. There is a clear flow of signal implied and the network can be organised hierarchically
- Weighted edges - reliability of an interaction, the quantitative expression change that a gene induces over another or even how closely related two genes are in terms of sequence similarity.

BRAC UNIVERSITY
Inspiring Excellence

Adjacency Matrix

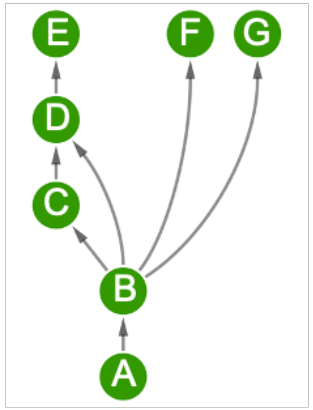
Undirected



	A	B	C	D	E	F	G	Degree
A	0	1	1	1	1	1	0	5
B	1	0	0	0	0	1	0	2
C	1	0	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	1
F	1	1	0	0	0	0	1	3
G	0	0	0	0	0	1	0	1

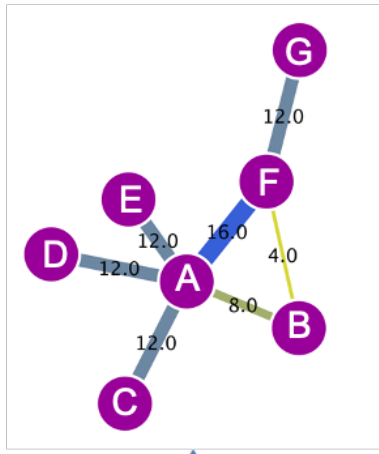
Adjacency matrices

Directed



	A	B	C	D	E	F	G	Out-degree
A	0	1	0	0	0	0	0	1
B	0	0	1	1	0	1	1	4
C	0	0	0	1	0	0	0	1
D	0	0	0	0	1	0	0	1
E	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0

Weighted

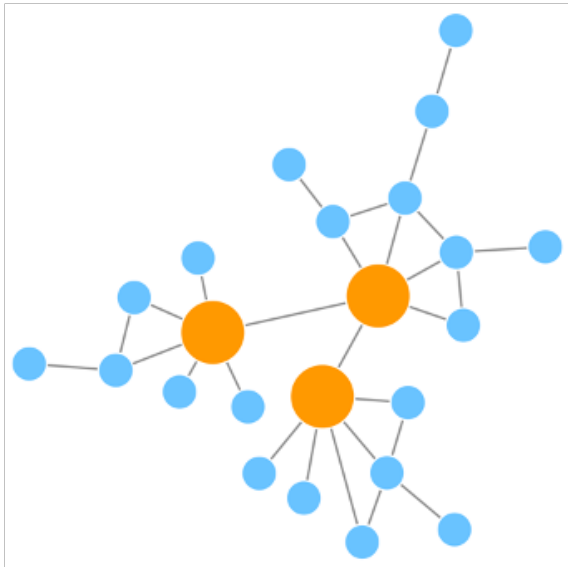


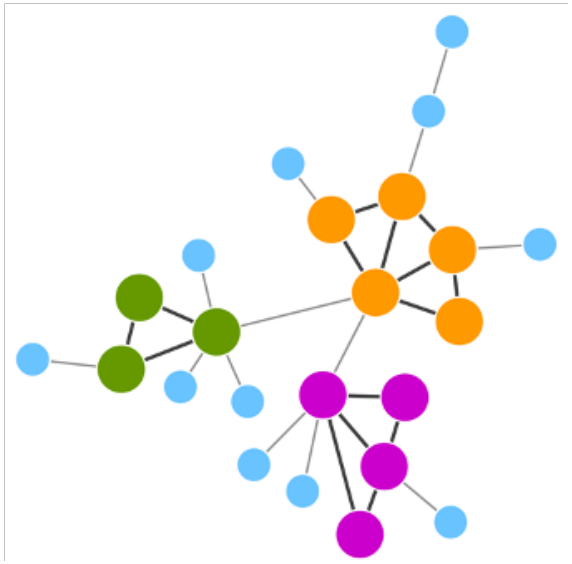
	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

BRAC UNIVERSITY
Inspiring Excellence

Network Topology

- The degree of a network – The degree is the number of edges that connect to a node.
- Shortest paths – Shortest paths, or the shortest distance between any two nodes, is used to model how information flows.
- Scale-free networks – In scale-free networks most of the nodes are connected to a low number of neighbours and there are a small number of high-degree nodes (hubs) that provide high connectivity to the network.
- Transitivity – Transitivity relates to the presence of tightly interconnected nodes in the network called clusters or communities.

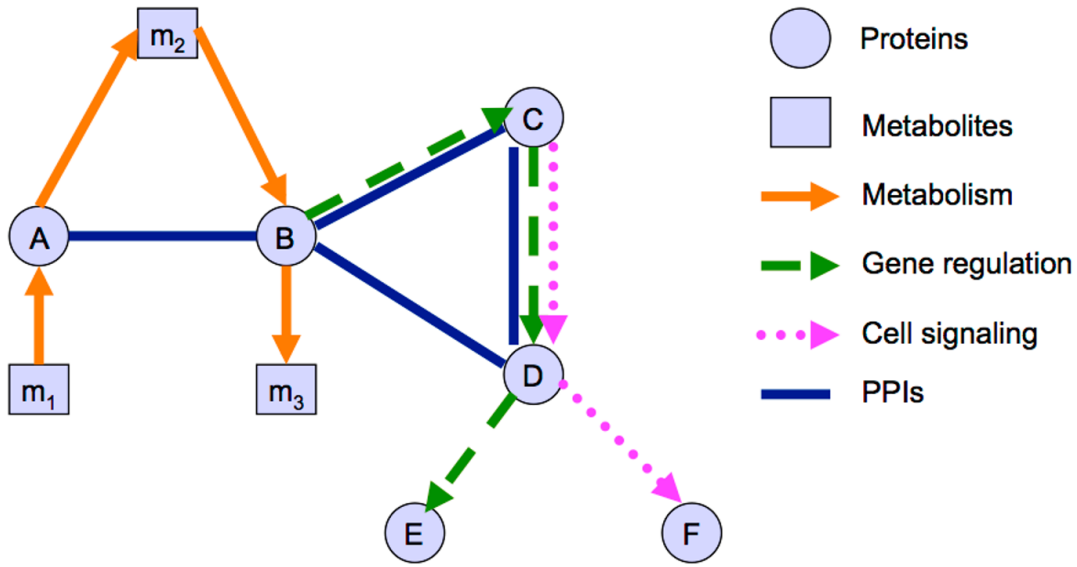




BRAC UNIVERSITY
Inspiring Excellence

Types of Biological Networks

- Protein-protein interaction networks
- Metabolic networks
- Genetic interaction networks
- Gene / transcriptional regulatory networks
- Cell signalling networks



- Sources: Manual curation of scientific literature, High-throughput datasets, Computational predictions, Literature text-mining.

BRAC UNIVERSITY
Inspiring Excellence

Protein Protein Interaction Networks

Protein-protein interaction networks (PPIN) are mathematical representations of the physical contacts between proteins in the cell. These contacts:

- are specific
- occur between defined binding regions in the proteins
- have a particular biological meaning (i.e., they serve a specific function)

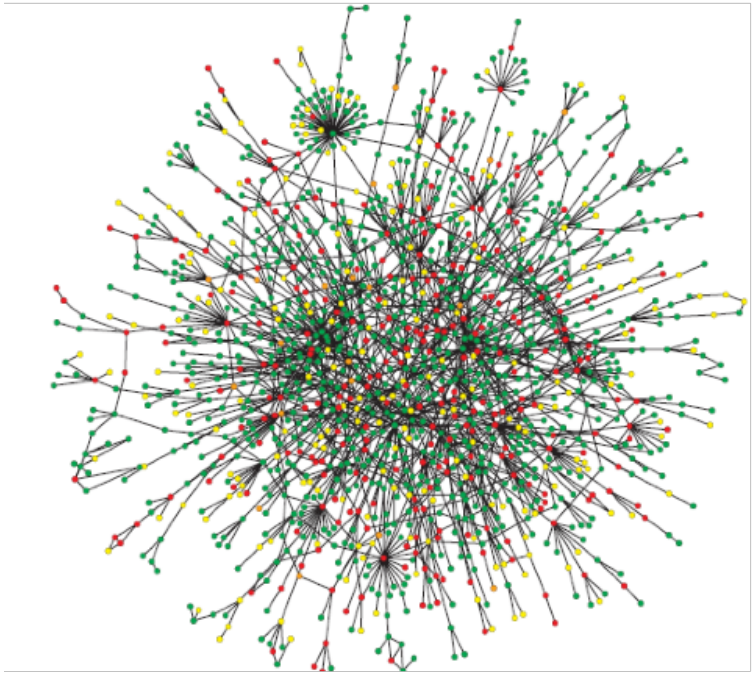
Knowledge of PPIs can be used to:

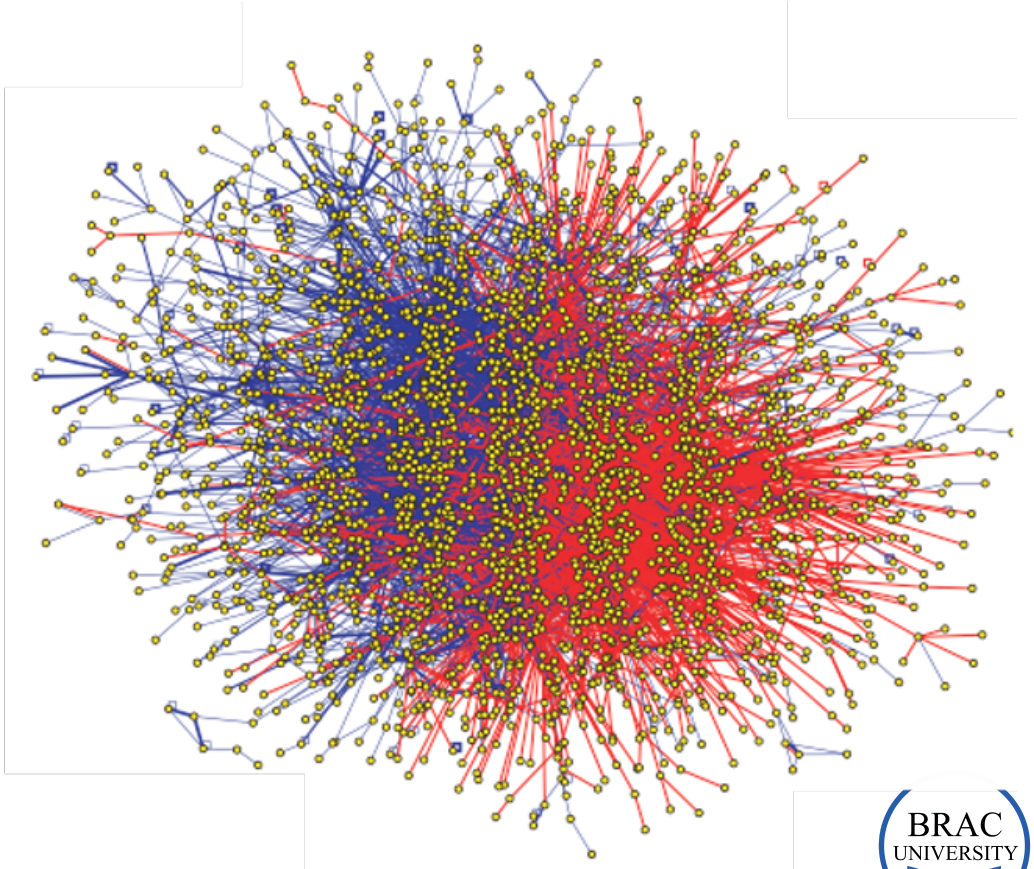
- assign putative roles to uncharacterised proteins
- add fine-grained detail about the steps within a signalling pathway
- characterise the relationships between proteins that form multi-molecular complexes such as the proteasome

BRAC UNIVERSITY
Inspiring Excellence

The Interactome

- The interactome is the totality of PPIs that happen in a cell, an organism or a specific biological context.

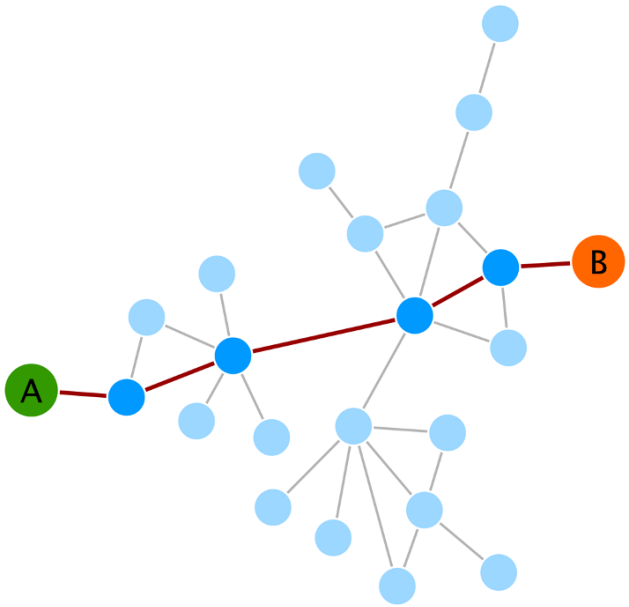




BRAC UNIVERSITY
Inspiring Excellence

The small world effect

- Protein-protein interaction networks show a small world effect meaning that there is great connectivity between proteins.
- Network's diameter (the maximum number of steps separating any two nodes) is small, no matter how big the network is.
- This level of connectivity has important biological consequences, since it allows for an efficient and quick flow of signals within the network.
- if the network is so tightly connected, why don't perturbations in a single gene or protein have dramatic consequences for the network?

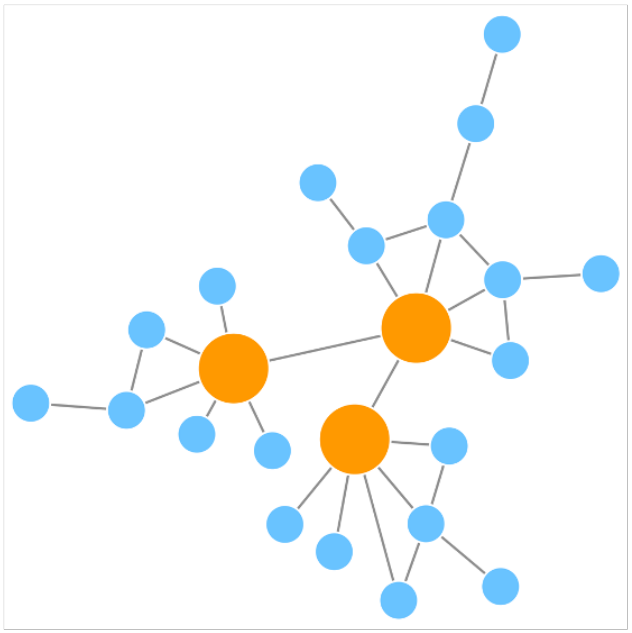


BRAC UNIVERSITY
Inspiring Excellence

Scale-free networks

- small number of nodes with high degree (the hubs) and a large number of nodes with a low degree

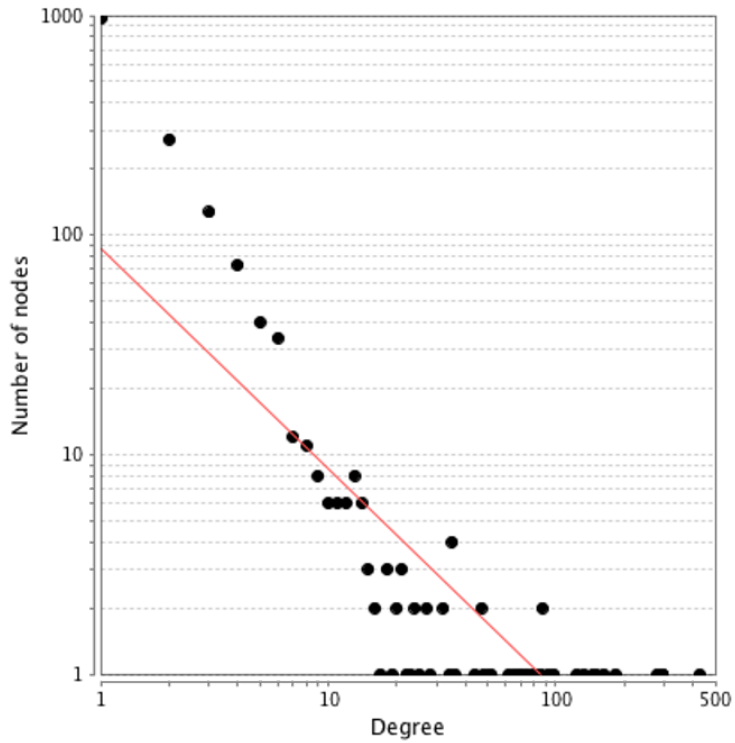
A



Scale-free network

B

Scale-free network degree distribution



BRAC UNIVERSITY
Inspiring Excellence

Scale-free networks

- Stability
  - If failures occur at random, and the vast majority of proteins are those with a small degree of connectivity, the likelihood that a hub would be affected is small
  - If a hub-failure occurs, the network will generally not lose its connectedness, due to the remaining hubs
- Invariant to changes of scale
  - No matter how many nodes or edges the network has, its properties remain stable
  - The presence of hubs is what allows for the small-world effect to be present regardless of the size of the network
- Vulnerable to targeted attack
  - If we lose a few major hubs from the network, the network is turned into a set of rather isolated graphs
  - Hubs are enriched with essential/lethal genes. For example, many cancer-linked proteins are hub proteins (e.g. the tumour suppressor protein p53)

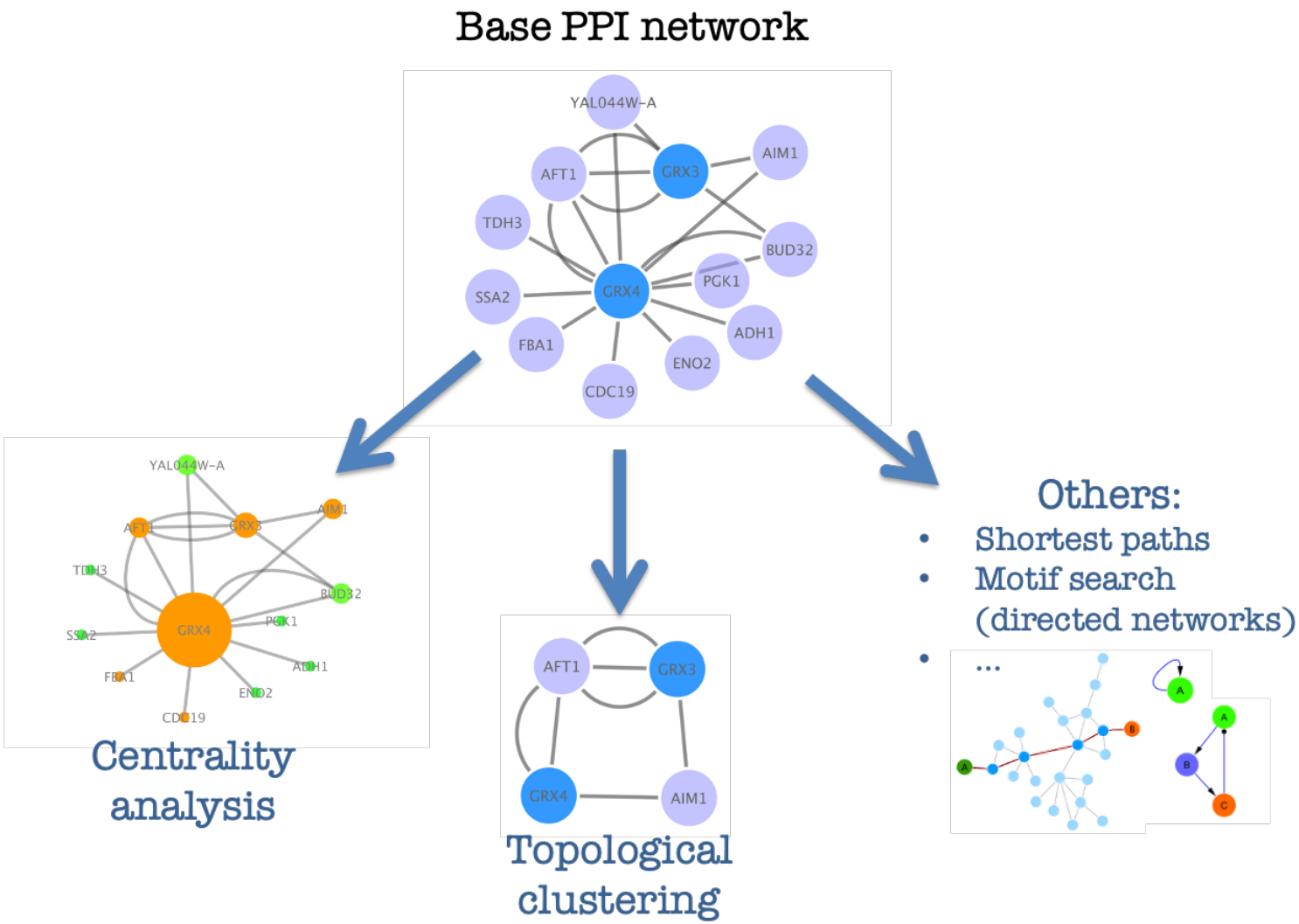
BRAC UNIVERSITY
Inspiring Excellence

Transitivity

- The transitivity or clustering coefficient of a network is a measure of the tendency of the nodes to cluster together.
  - High transitivity means that the network contains communities or groups of nodes that are densely connected internally.
  - In biological networks, finding these communities is very important, because they can reflect functional modules and protein complexes
- 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 placed in a different context.
  - Modules help reduce the complexity of biological networks by giving us a set of reducible, functional units that can be studied as an integrated entity.

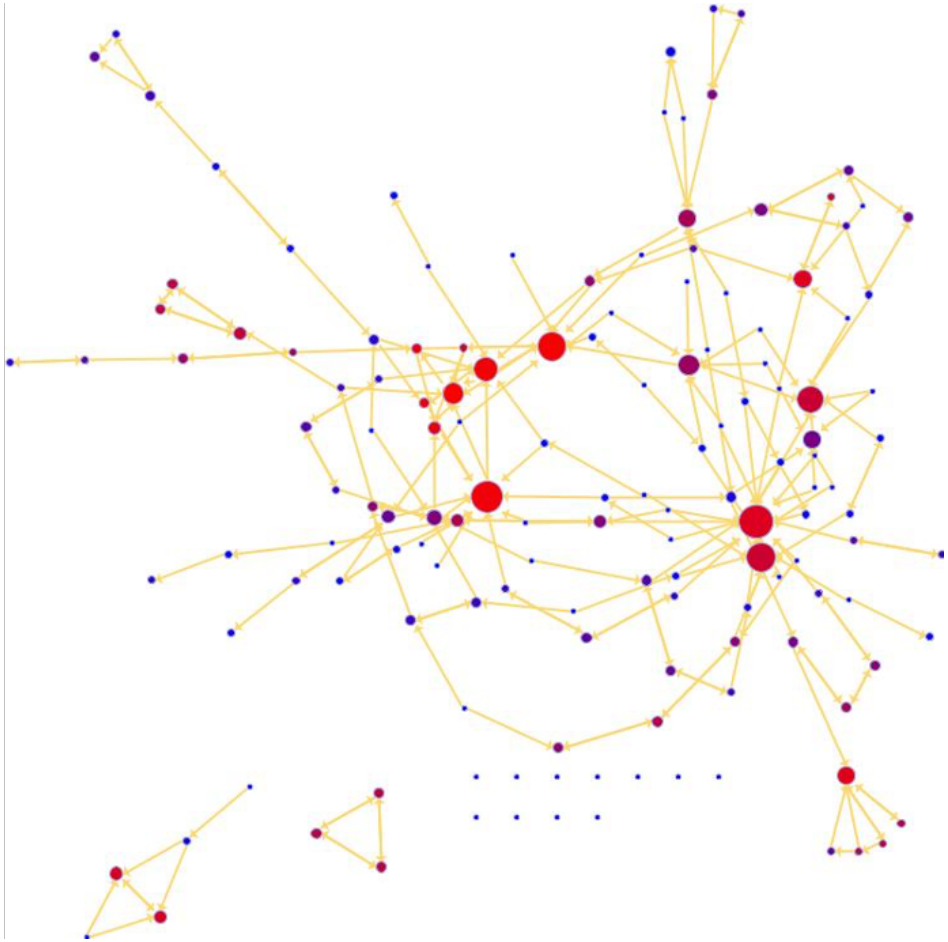
BRAC UNIVERSITY
Inspiring Excellence

Topological Analysis



Centrality Analysis

- Which protein is the most important and why?



- degree, betweenness, closeness, random walks, etc

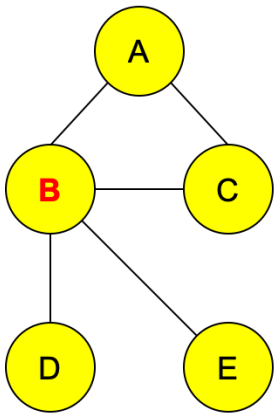
BRAC UNIVERSITY
Inspiring Excellence

Closeness centrality

- Closeness centrality measures how short the shortest paths are from node *i* to all nodes. It is usually expressed as the normalised inverse of the sum of the topological distances in the graph

$$CC(i) = \frac{N-1}{\sum_j d(i,j)}$$

where
$i \neq j$,
d_{ij} is the length of the shortest path between nodes i and j in the network,
N is the number of nodes.



	A	B	C	D	E	farness $\sum_{j=1}^{n} d(i,j)$	$CC(i) = \frac{N-1}{\sum_j d(i,j)}$
A	0	1	1	2	2	6	(5-1)/6= 0.67
B	1	0	1	1	1	4	1.00
C	1	1	0	2	2	6	0.67
D	2	1	2	0	2	7	0.57
E	2	1	2	2	0	7	0.57

N = 5 (# of nodes)

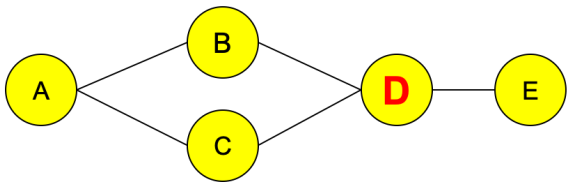
BRAC UNIVERSITY
Inspiring Excellence

Betweenness centrality

- Betweenness centrality is based on communication flow.
- Nodes with a high betweenness centrality are interesting because they lie on communication paths and can control information flow. These nodes can represent important proteins in signalling pathways and can form targets for drug discovery.
- defined as the number of shortest paths in the graph that pass through the node divided by the total number of shortest paths.

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

Where g_{jk} = the number of geodesics (shortest paths) connecting jk, and $g_{jk}(n_i)$ = the number that node i is on.



Clustering Analysis

- **Community / Cluster**: A group of nodes that are more connected within themselves than with the rest of the network. When talking about PPINs, communities fall into two categories: functional modules and protein complexes.
- **Module:** In biology, modules are exchangeable functional units in which the nodes (proteins) do not have to be interacting in the same time or space. The most important characteristic of a module is that its intrinsic functional properties do not change when it is placed in a different context.
- **Complex:** A complex is a group of proteins that interact with each other at the same time and in the same space, forming relatively stable multi-protein machinery.
- **Clique:** A subset of nodes in which every node is connected with every other member of the clique.
- **Motif:** Motifs are statistically over-represented sub-graphs in a network. They correspond with a pattern of connections that generates a characteristic dynamical response (e.g. a negative feedback loop).

BRAC UNIVERSITY
Inspiring Excellence

A greedy algorithm - Newman-Girvan

- Identifies communities by using the edge betweenness centrality measure. Edges that connect different communities have higher centrality values, since a larger proportion of shortest paths will pass through them
- To define communities it uses the edge betweenness centrality scores to rank the edges of the network, then removes the most central edges and then re-calculates the betweenness scores until no edges are left. Edges affected by the removal are deemed to be part of the same community
- Can be considered a 'naïve' approach that will define communities even when they are only marginally more connected than the rest of the network

BRAC UNIVERSITY
Inspiring Excellence