Topological properties of graphs

Sushmita Roy
sroy@biostat.wisc.edu

Computational Network Biology
Biostatistics & Medical Informatics 826
https://compnetbiocourse.discovery.wisc.edu

Oct 25th 2018

RECAP of problems in network biology

Biological problem

- ❑ Mapping regulatory network structure
- ❑ Dynamics and context specificity of networks
- ❑ Understanding design principles of biological networks
- ❑ Interpretation of sequence variants
- ❑ Identification of important genes
- ❑ Integrating different types of molecular genomic data

Computational approaches

- ❑ Probabilistic graphical models
- ❑ Graph structure learning
- ❑ Multiple network learning
- ❑ Topological properties of graphs
- ❑ Graph clustering
- ❑ Graph alignment
- ❑ Diffusion on graphs

Topics in this section

- Topological properties of networks
- Modules in biological networks
- Algorithms for graph clustering

Goals for today

- Commonly measured network properties
  - Degree distribution
  - Average shortest path length
  - Network motifs
- Studying modularity of biological networks
  - Modularity
  - Clustering coefficient
  - Algorithms to find modules on graphs

Why should we care about network measures?

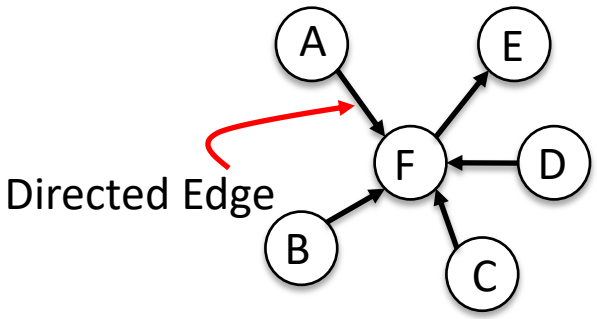
- From Barabasi and Oltvai 2004:

"Probably the most important discovery of network theory was the realization that despite the remarkable diversity of networks in nature, their architecture is governed by a few simple principles that are common to most networks of major scientific and technological interest"

Barabasi & Oltvai 2004, Nature Genetics Review

Node degree

- Undirected network
  - Degree, k: Number of neighbors of a node
- Directed network
  - In degree, k_{in}: Number of incoming edges
  - Out degree, k_{out}: Number of outgoing edges



In degree of F is 4
Out degree of E is 0

Average degree

- Consider an undirected network with N nodes
- Let k_i denote the degree of node i
- Average degree is

$$< k >= \frac{\sum_{i=1}^{N} k_i}{N}$$

Degree distribution

- $P(k)$ the probability that a node has k edges
- Different networks can have different degree distributions
- A fundamental property that can be used to characterize a network

Different degree distributions

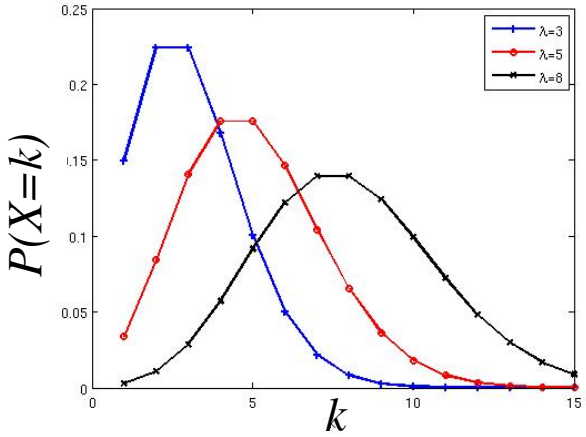
- Poisson distribution
  - The mean is a good representation of k_i of all nodes
  - Networks that have a Poisson degree distribution are called Erdos Renyi or random networks

- Power law distribution
  - Also called *scale free*
  - There is no "typical" node that captures the degree of nodes.

Poisson distribution

- A discrete distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



- The Poisson is parameterized by λ which can be easily estimated by maximum likelihood

$$\lambda_{MLE} = \frac{\sum_{i=1}^{n} k_i}{n}$$

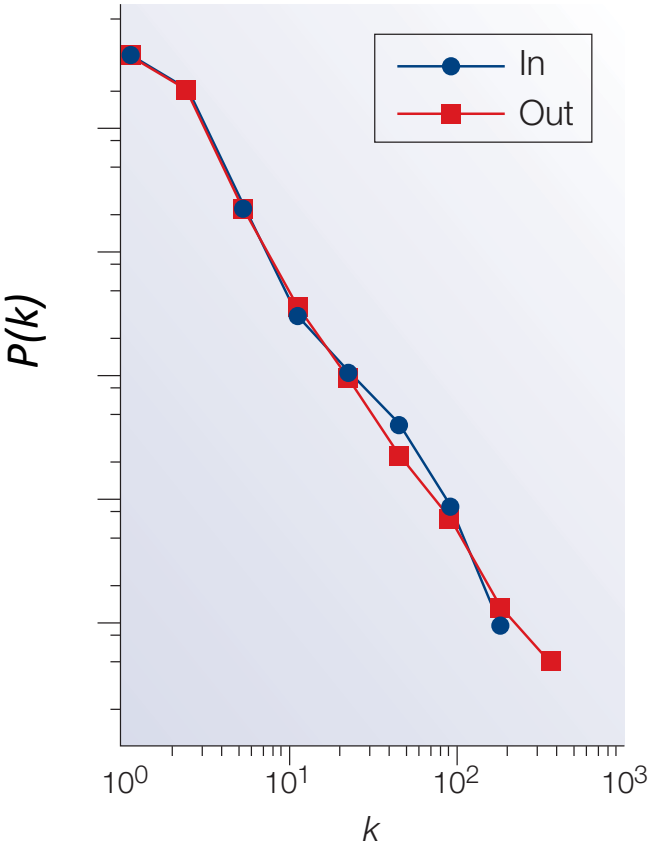
Power law distribution

- Used to capture the degree distribution of most real networks

$$P(X = k) \propto k^{-\gamma}$$

- Typical value of γ is between 2 and 3.

- MLE exists but is more complicated
  - See *Power-Law Distributions in Empirical Data.* Clauset, Shalizi and Newman, 2009 for details



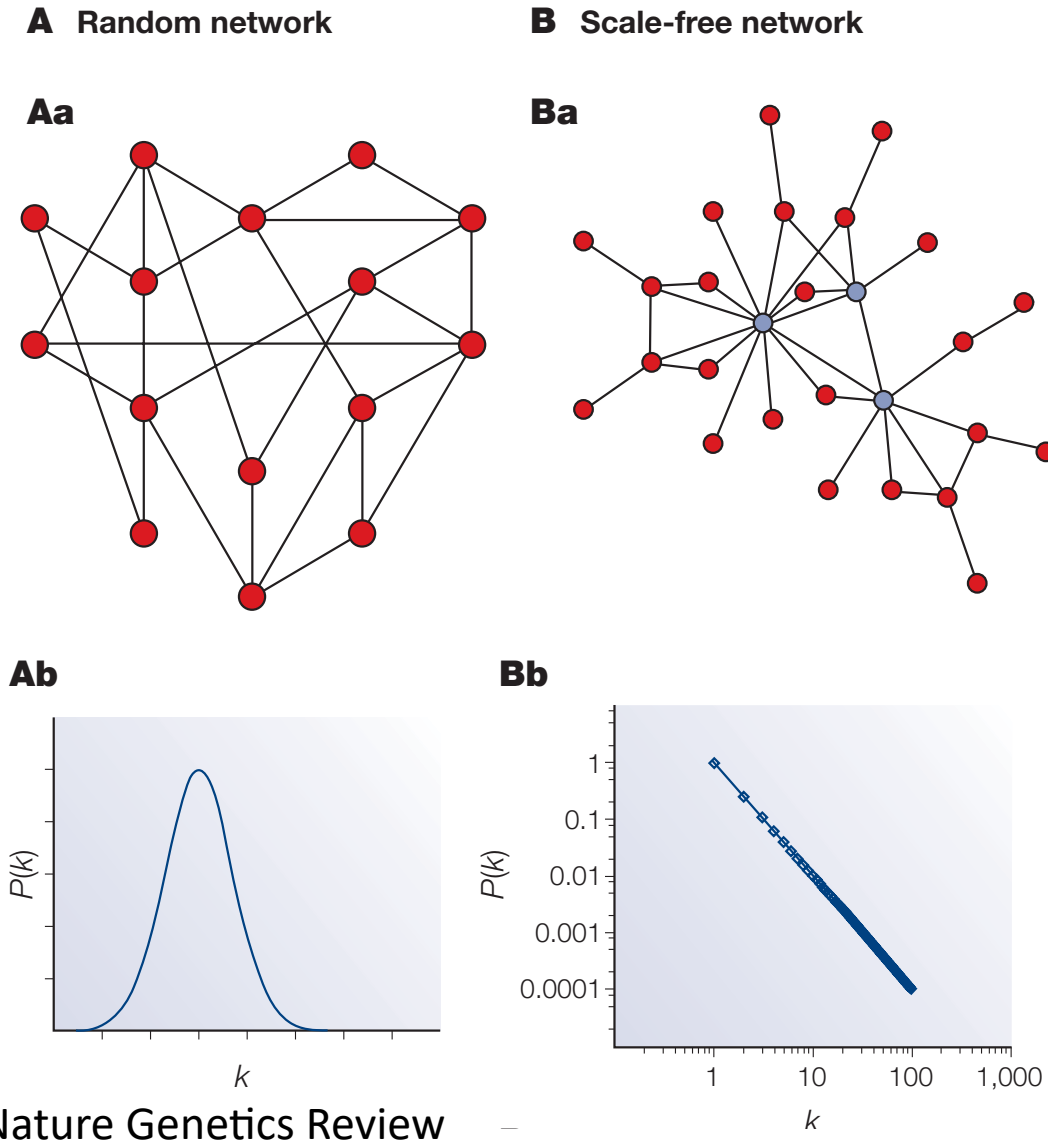
Erdos Renyi random graphs

- Dates back to 1960 due to two mathematicians Paul Erdos and Alfred Renyi.
- Provides a probabilistic model to generate a graph
- Starts with N nodes and connects two nodes with probability p
- Node degrees follow a Poisson distribution
- Tail falls off exponentially, suggesting that nodes with degrees different from the mean are very rare

Scale free networks

- Degree distribution is captured by a power law distribution
- There is no “typical” node that describes the degree of all other nodes
- Such networks are ubiquitous in nature

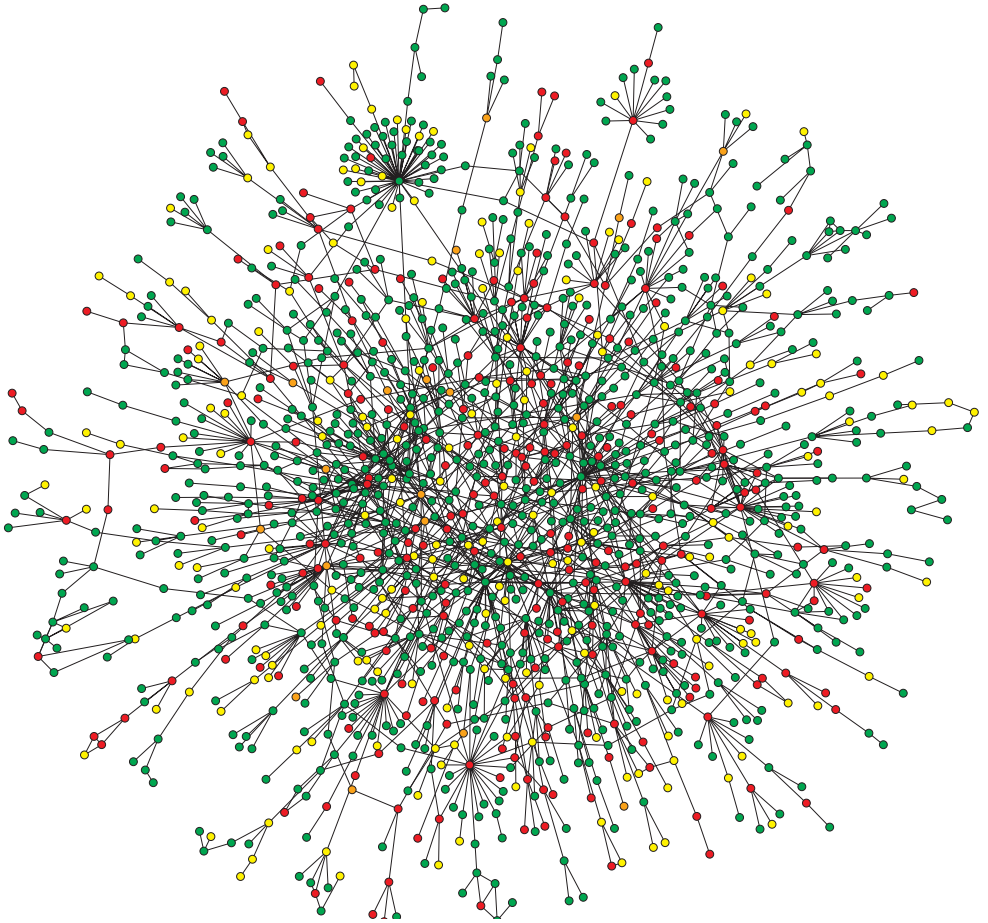
Poisson versus Scale free



Barabasi & Oltvai 2004, Nature Genetics Review

Yeast protein interaction network is believed to be scale free

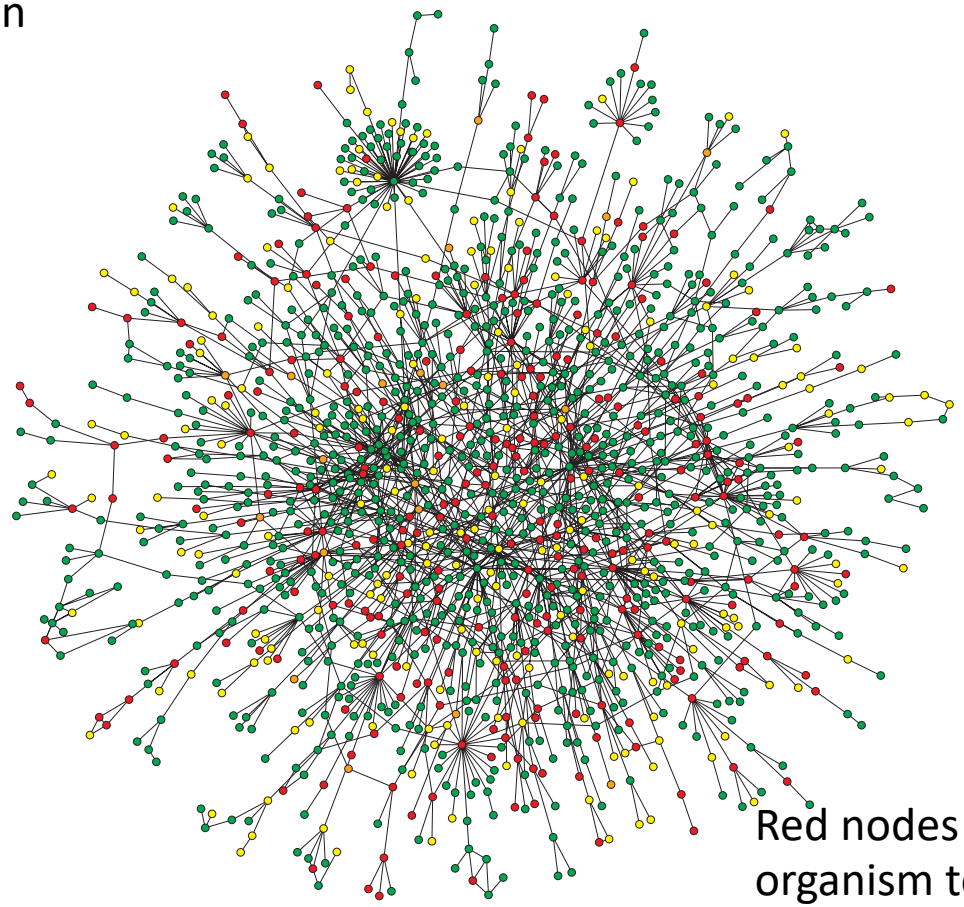
- "Whereas most proteins participate in only a few interactions, a few participate in dozens"
- Such high degree nodes are called *hubs*



Barabasi & Oltvai 2004, Nature Genetics Review

Degree of a node is correlated to functional importance of a node

Yeast protein-protein interaction network



Red nodes on deletion cause the organism to die

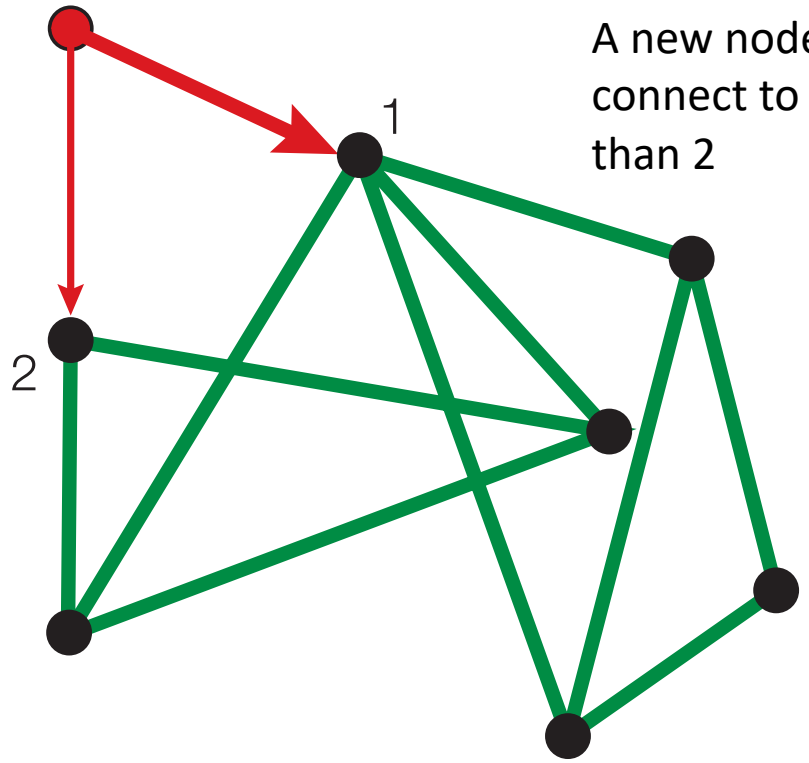
Red nodes also among the nodes with the highest degree

Origin of scale free networks

- Scale free networks are ubiquitous is nature
- How do such networks form?
- Such networks are the result of two processes
  - a growth process where new nodes join the network over an extended period of time
    - Think about how the internet has grown
  - preferential attachment: new nodes tend to connect to nodes with many neighbors
    - Rich get richer.

Barabasi & Oltvai 2004, Nature Genetics Review

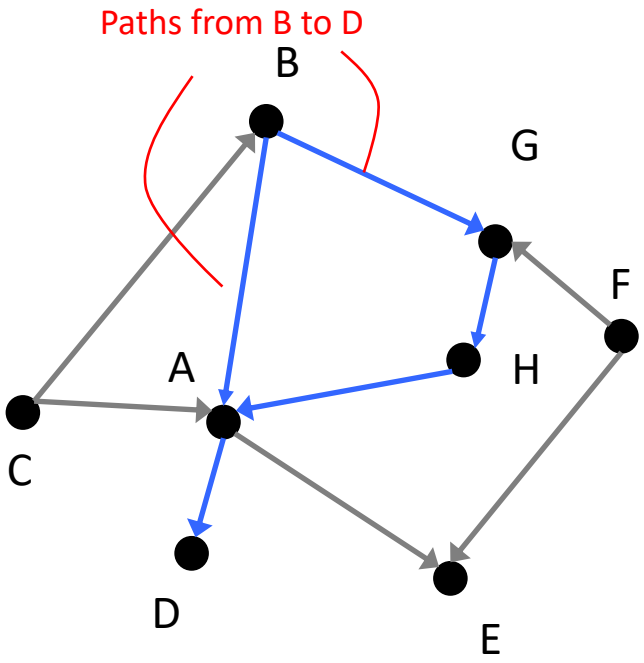
Growth and preferential attachment in scale free networks



A new node (red) is more likely to connect to node 1 than 2

Barabasi & Oltvai 2004, Nature Genetics Review

Paths on a graph



Path: a set of connected edges from one node to another

There are two paths from B-D: B->A->D and B->G->H->A->D

Path lengths

- The shortest path length between two nodes A and B:
  - The smallest number of edges that need to be traversed to get from A to B
- Mean path length is the average of all shortest path lengths
- Diameter of a graph is the longest of all shortest paths in the network

Scale-free networks tend to be ultra-small

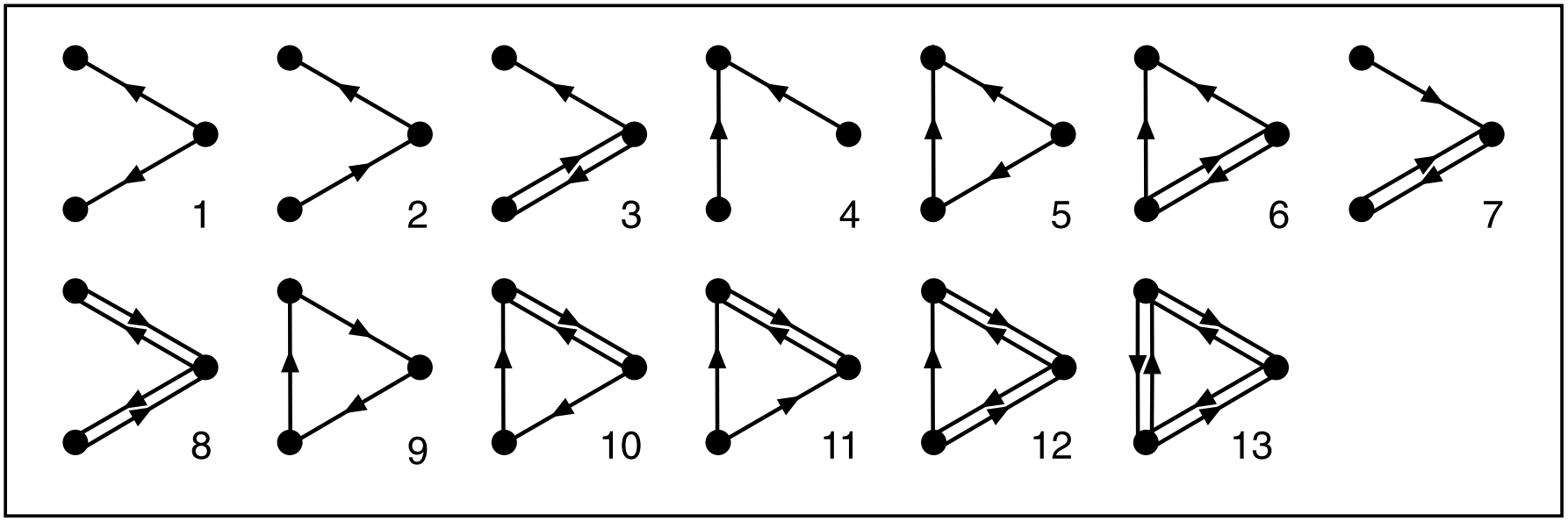
- Two nodes on the network are connected by a small number of edges
- Average path length is proportional to $log(log(N))$, where N is the number of nodes in the network
- In a random network (Erdos Renyi network) the average path length is proportional to $log(N)$

Network motifs

- Network motifs are defined as small recurring subgraphs that occur much more than a randomized network
- A subgraph is called a network motif of a network if its occurrence in randomized networks is significantly less than the original network.
- Network motifs are often called “building blocks” of complex networks
- Bio-molecular networks tend to exhibit certain types of network motifs more frequently than random
- Some motifs are associated with specific network dynamics

Milo Science 2002

Network motifs of size 3 in a directed network

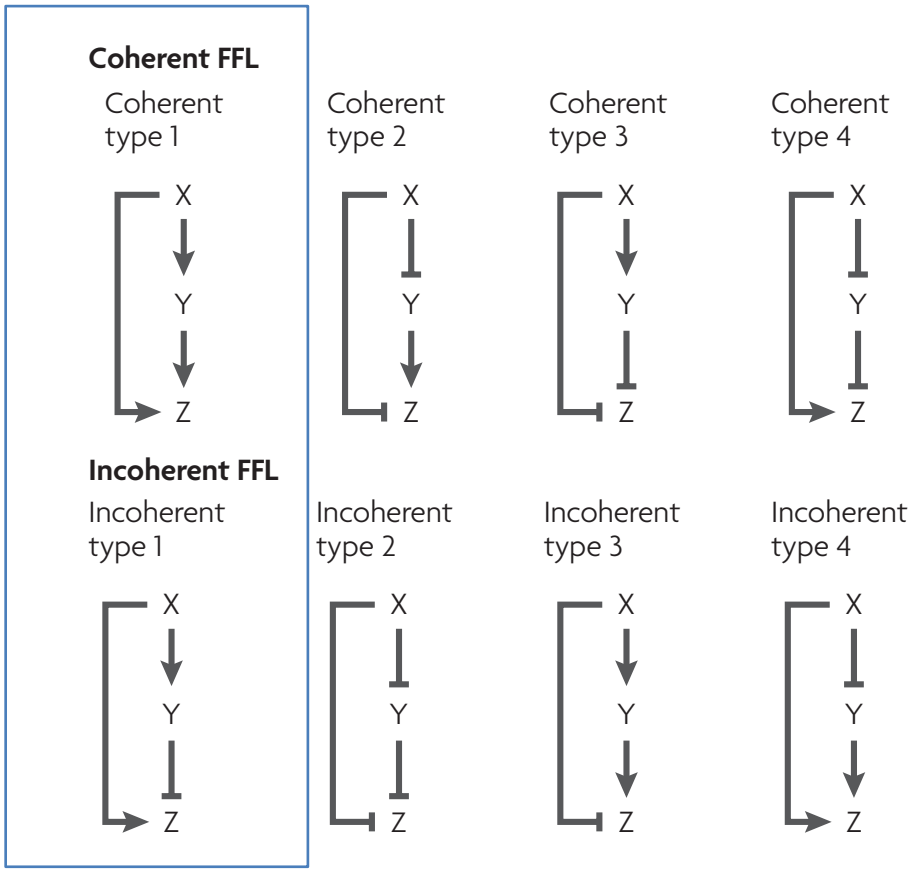


Additional possibilities if we consider the sign of an edge

Milo Science 2002

Different types of feed-forward motifs

These appear much more in transcriptional networks than other FF motifs

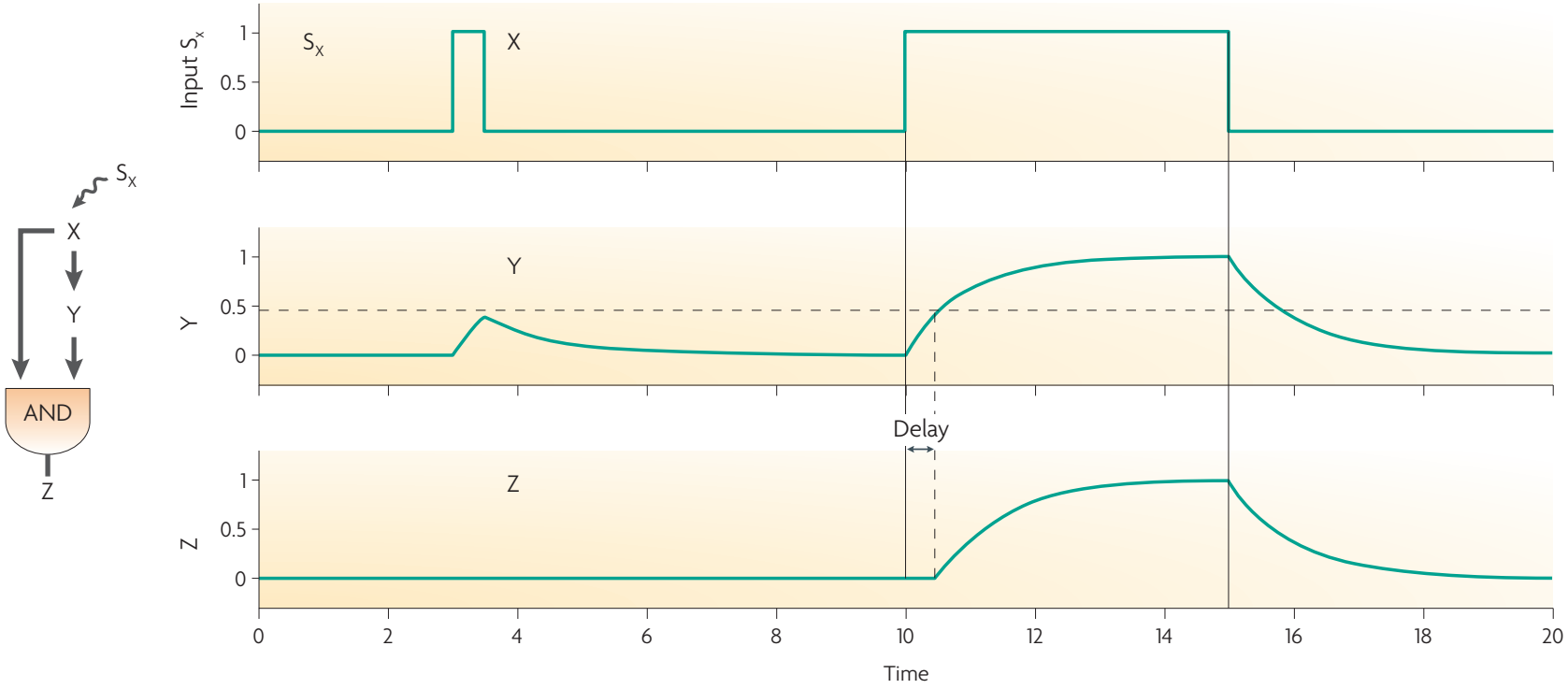


Coherent: sign of the direct path is the same as the indirect path
Incoherent: sign of the direct path and indirect path are not the same

Alon 2007, Nature Review Genetics

Network motifs are associated with dynamics

A feed-forward motif with an AND gate can encode a sign sensitive delay



Alon 2007, Nature Review Genetics Cited >2000 times!

How to find network motifs?

- Given an input network G we need to address two problems
  - *Subgraph enumeration*: Find which subgraphs occur in *G* and how many times
  - *Significance of the number of occurrences*: Compare to the number of occurrences of subgraphs in randomized networks
- Software to find network motifs
  - FANMOD: a tool for fast network motif detection http://bioinformatics.oxfordjournals.org/content/22/9/1152.full

Wernicke 2005: http://theinf1.informatik.uni-jena.de/publications/network-motifs-wabi05.pdf

Algorithm for enumerating subgraphs: Edge sampling

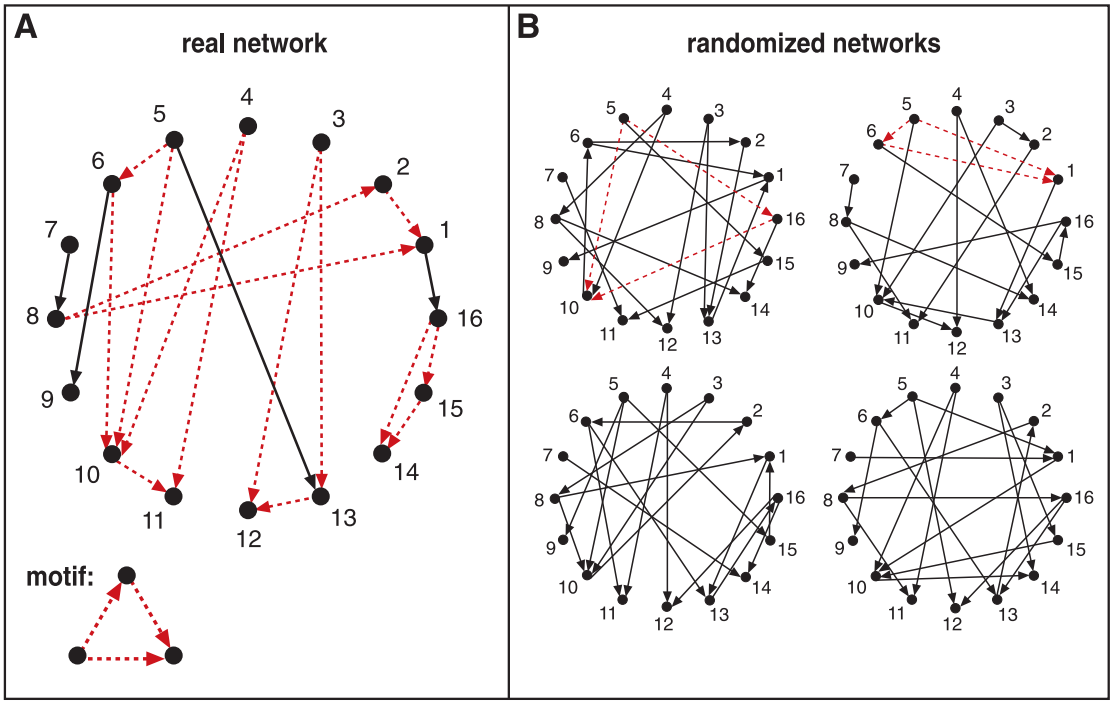
Input: A graph $G = (V, E)$ and an integer $2 \leq k \leq |V|$.
Output: Vertices of a randomly chosen size-k subgraph in G.

01 $\{u, v\} \leftarrow$ random edge from E
02 $V' \leftarrow \{u, v\}$
03 **while** $|V'| \neq k$ **do**
04 $\quad \{u, v\} \leftarrow$ random edge from $V' \times N(V')$
05 $\quad V' \leftarrow V' \cup \{u\} \cup \{v\}$
06 **return** V'

N(V'): set of neighbors of all vertices in *V'*

Kashtan et al., Bioinformatics 2004

Assessing the significance of a network motif

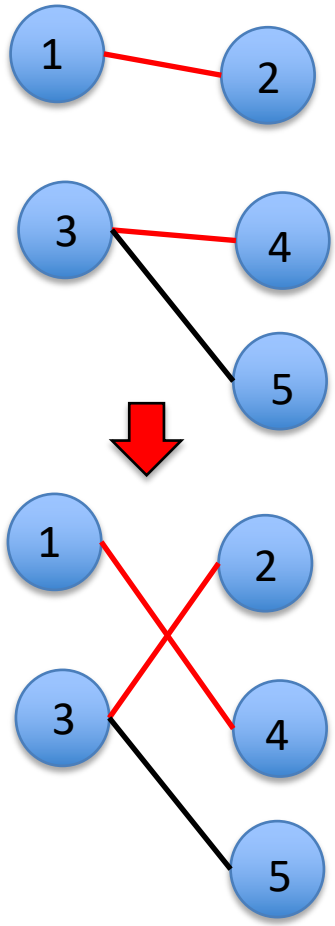


The motif (red dashed edges) occurs much more frequently in the real network than in any randomized network

Generating a randomized network

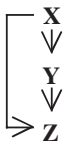
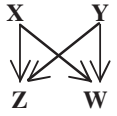
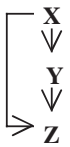
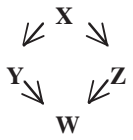
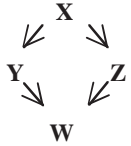
- While an Erdos Renyi network is random, it does not have the same degree distribution as a given network
- How to generate a randomized network with the same degree distribution?

Strategy to generate randomized networks



Select two edges connecting four vertices and swap the end points. Repeat.

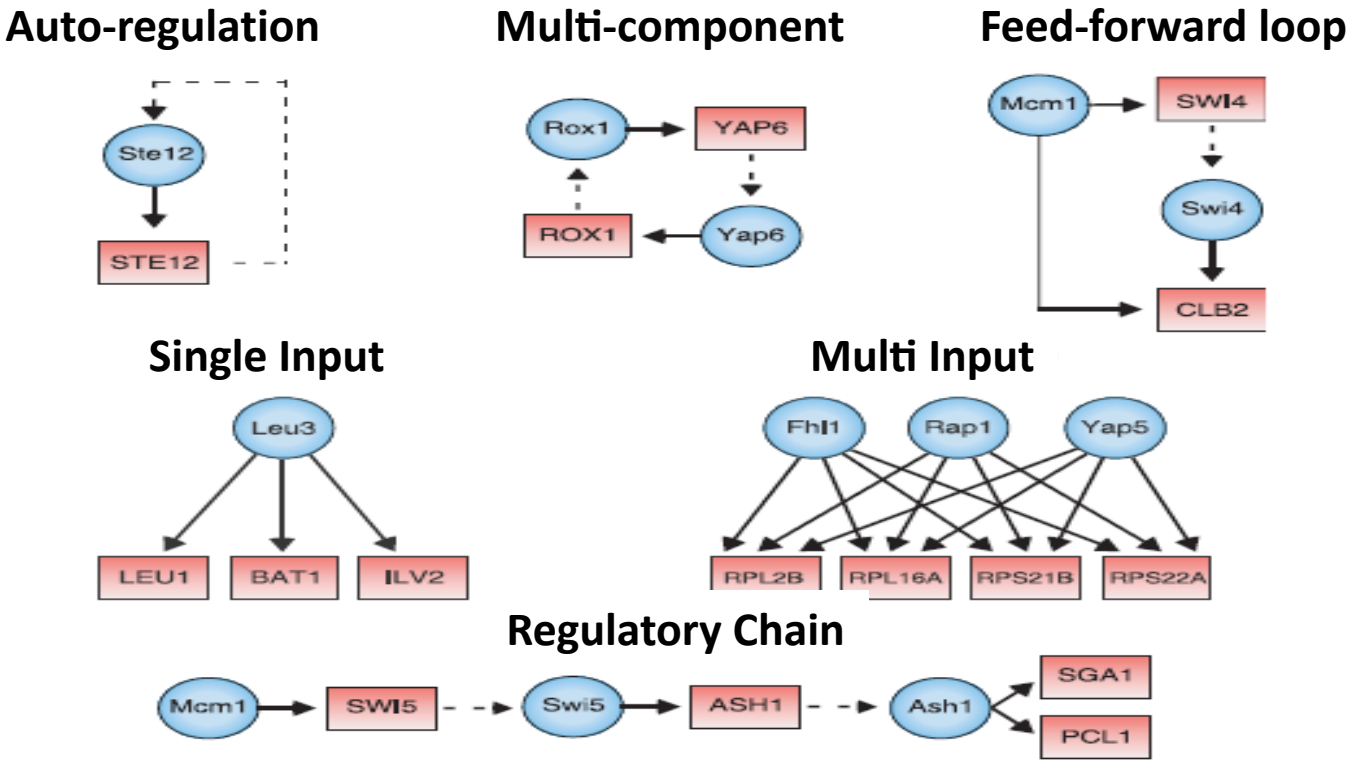
Network motifs found in many complex networks

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

The occurrence of the feedforward loop in both networks suggests a fundamental similarity in the design on these networks

Milo et al., Science 2002

Structural common motifs seen in the yeast regulatory network



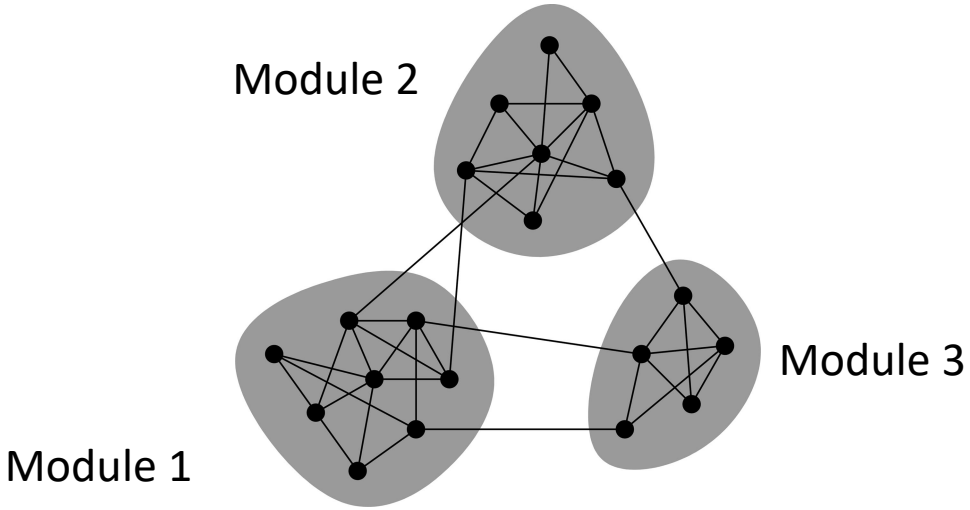
Feed-forward loops involved in speeding up in response of target gene

Lee *et.al. 2002,* Mangan & Alon, 2003

Modularity in networks

- Modularity "*refers to a group of physically or functionally linked nodes that work together to achieve a distinct function*" -- Barabasi & Oltvai
- Modularity is an important principle of biological systems
- Genes tend to interact with a select set of other genes exhibiting a clustering of interactions
- Module detection can help to
  - Understand the organizational properties of the network
  - Can be used to predict function of genes based on their grouping behavior

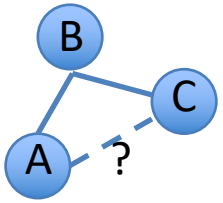
A modular network



M. E. J. Newman, Modularity and community structure in networks, PNAS 2006

Clustering coefficient

- Measure of transitivity in the network that asks
  - If A is connected to B, and B is connected to C, how often is A connected to C?
- Clustering coefficient C_i for each node i is



$$C_i = \frac{2n_i}{k_i(k_i - 1)}$$

- k_i Degree of node i
- n_i is the number of edges among neighbors of i
- Average clustering coefficient gives a measure of "modularity" of the network

Summary of topological properties of networks

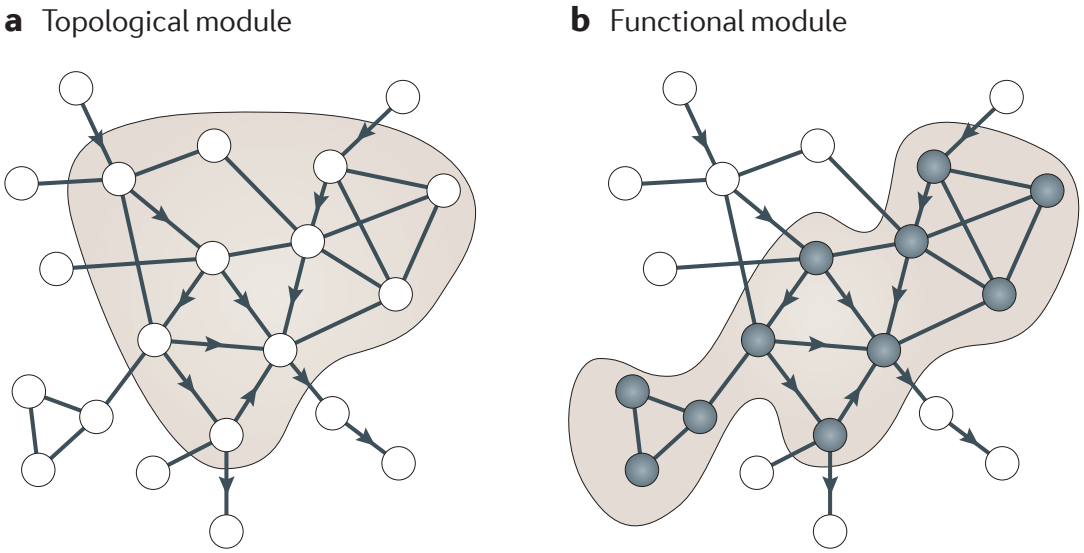
- Given a network, its topology can be characterized using different measures
  - Degree distribution
  - Average path length
  - Clustering coefficient (also used to measure modularity of a network)
- Degree distribution can be
  - Poisson
  - Power law
    - Such networks are called scale free
- Network modularity
  - Clustering coefficient
- Network motifs
  - Building blocks of complex networks
  - Over represented subgraphs of specific types

Goals for today

- Commonly measured network properties
  - Degree distribution
  - Average shortest path length
  - Network motifs
- Studying modularity of biological networks
  - Modularity
  - Clustering coefficient
  - Algorithms to find modules on graphs

Different types of network modules

- Topological modules
  - Defined solely based on the graph connectivity of nodes
- Functional modules
  - Based on graph connectivity & other node attributes
  - Can be further grouped into
    - Active modules
    - Integrative modules
    - Disease modules



Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo, Nature Review Genetics 2011, Mitra et al., Nature Review Genetics 2014

Studying modularity in biological systems

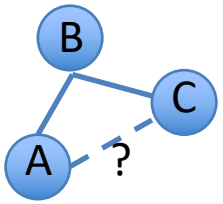
- Given a network is it modular?
- Given a network what are the modules in the network?

Given a graph, is it modular?

- Given a graph, we would like to know if it is modular
- This requires us to quantify modularity
- Different measures of modularity exist
  - Clustering coefficient
  - Q Modularity: measures the relative density of edges between and within the groups

Clustering coefficient

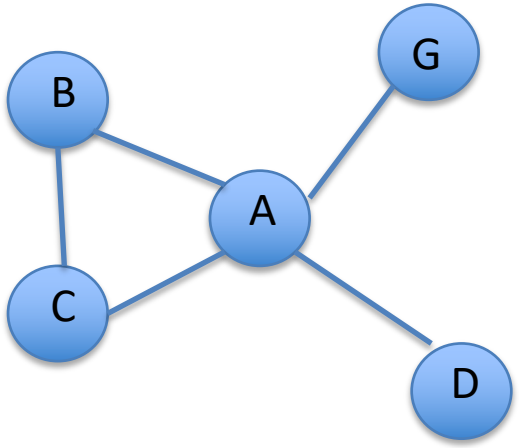
- Measure of transitivity in the network that asks
  - If A is connected to B, and B is connected to C, how often is A connected to C?
- Clustering coefficient C_i for each node i is



$$C_i = \frac{2n_i}{k_i(k_i - 1)}$$

- k_i Degree of node i
- n_i is the number of edges among neighbors of i
- Average clustering coefficient gives a measure of "modularity" of the network

Clustering coefficient example



$$C_A = \frac{2 * 1}{4 * 3}$$

$$C_G = 0$$

Q measure for modularity

- Suppose the nodes in the graph belong to K groups (communities)
- Modularity (Q) can be assessed as follows:
  - difference between within group (community) connections and expected connections within a group
- This measure assess how good a particular grouping is

$$Q = \sum_{i=1}^{K} (e_{ii} - a_i^2)$$

Fraction of edges within group i

Fraction of edges without regard to community structure

K: number of groups

e_{ij}: Fraction of total edges that link nodes in group i to group j

$$a_i = \sum_j e_{ij}$$

Detecting community structure in networks. M.E. J Newman

Studying modularity in biological systems

- Given a network is it modular?
- Given a network what are the modules in the network?

Given a graph, what are the modules

- Given a graph find the modules
  - Modules are represented by densely connected subgraphs
- The graph can be partitioned into modules using "Graph clustering" or "Graph partitioning"
- Clusters, modules are also called "communities"

Problem definition of (topological) module detection

- Given:
  - A graph
  - A measure of cluster quality
- Do
  - Partition the vertices into groups such that they form densely connected subgraphs (e.g. as measured by the cluster quality)

Common graph clustering algorithms

- Hierarchical or flat clustering using a notion of similarity between nodes
- Girvan-Newman algorithm
- Hierarchical Agglomerative clustering
- Spectral clustering
- Markov clustering algorithm
- Affinity propagation

Clustering

- Types of clustering
  - Flat clustering
    - K-means
    - Gaussian mixture models
  - Hierarchical clustering
- Clustering algorithms differ in the distance measure used to group objects together

Task definition: clustering objects

- Given: attributes for a set of objects we wish to cluster
- Do: organize objects into groups such that
  - Objects in the same cluster are highly similar to each other
  - Objects from different clusters have low similarity to each other

Flat clustering

- Cluster objects into K clusters
- K : number of clusters is a user defined argument
- Two example algorithms
  - *K*-means
  - Gaussian mixture model-based clustering

Hierarchical clustering

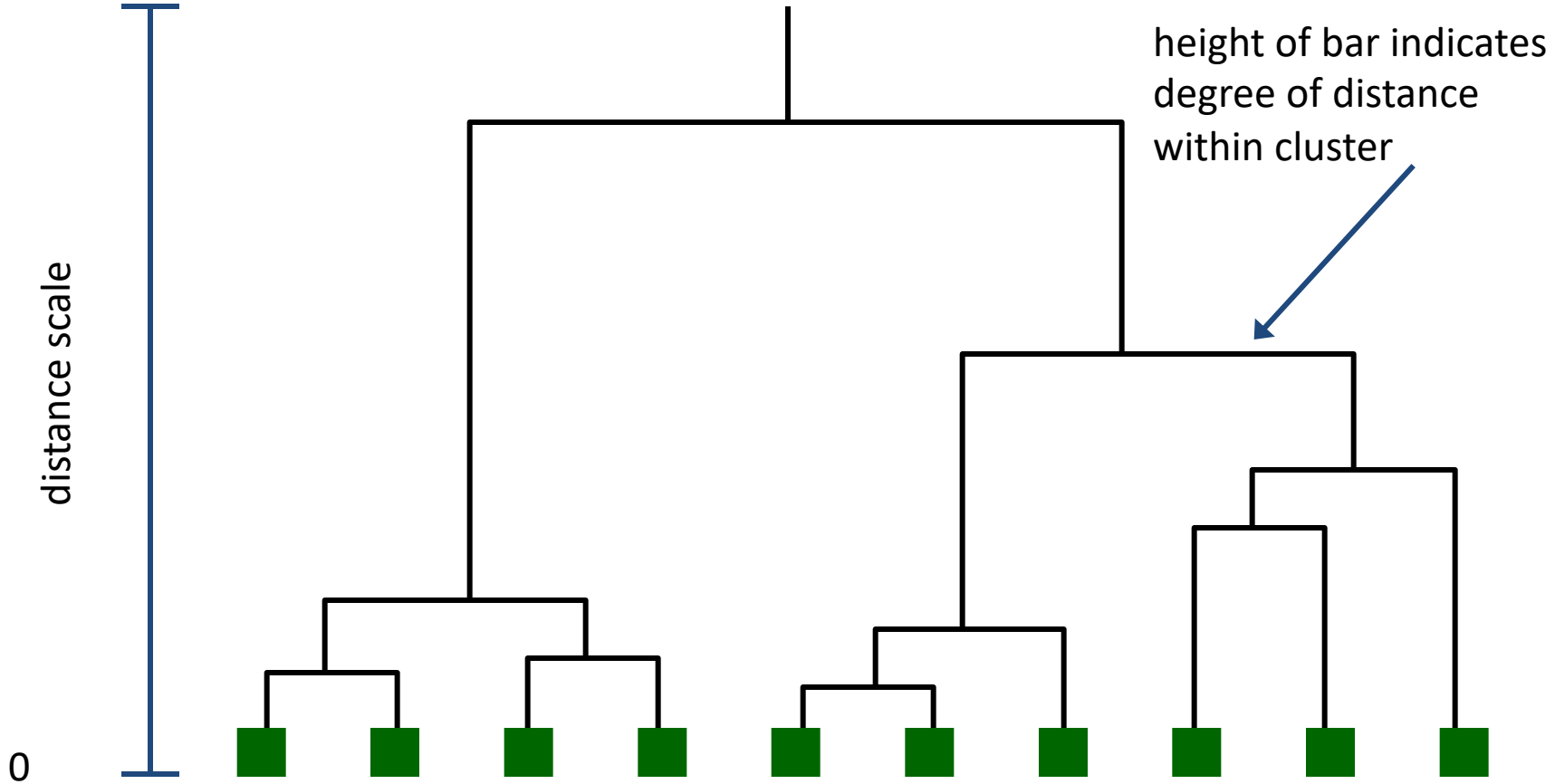
- Hierarchical clustering is a widely used clustering technique
- Instead of the number of clusters, it requires us to specify how much dissimilarity we will tolerate between groups
- The hierarchical clustering is represented by a tree structure called a dendrogram

Slides adapted from Prof. Mark Craven; BMI 576

Types of hierarchical clustering strategies

- Agglomerative (bottom-up)
  - Start from the individual objects
  - Group objects or clusters of objects
- Divisive (or top-down)
  - Start from all objects in a single cluster
  - Break down each cluster into smaller clusters

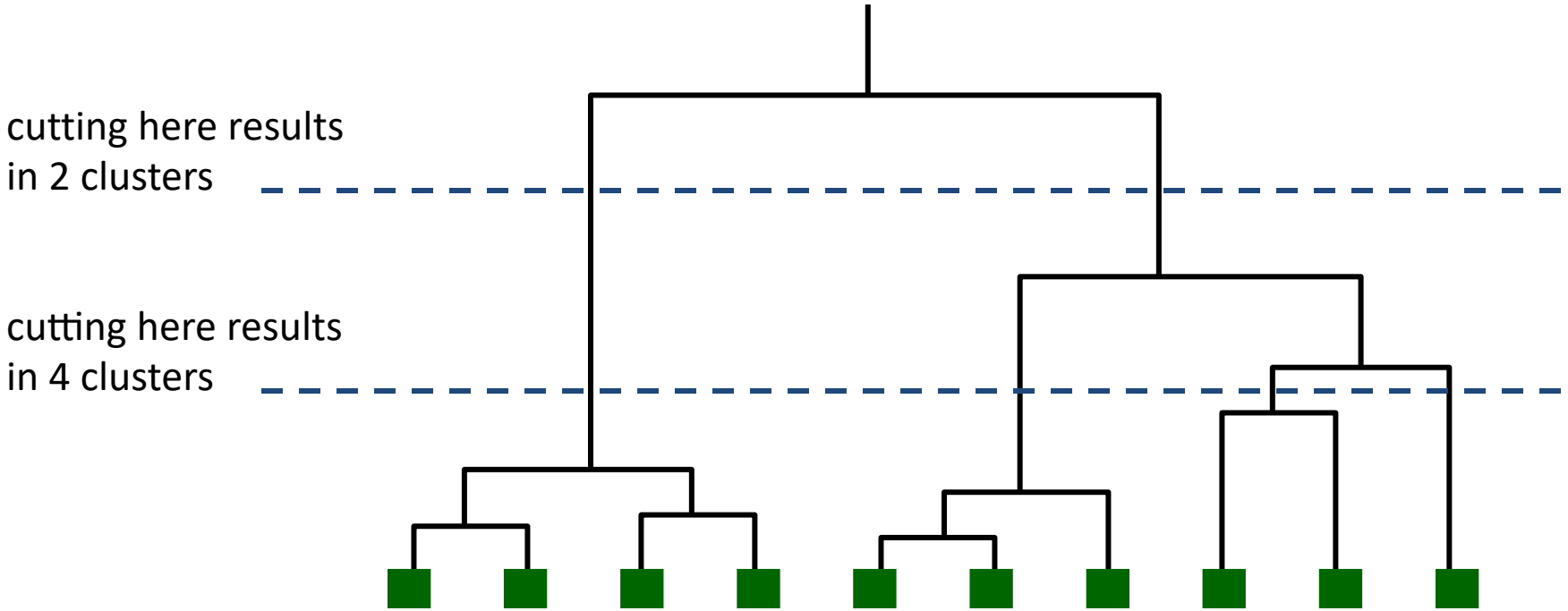
Hierarchical clustering



Slides from Prof. Mark Craven

Flat clustering from a hierarchical clustering

- We can always generate a flat clustering from a hierarchical clustering by “cutting” the tree at some distance threshold



Slides from Prof. Mark Craven

Hierarchical clustering to find modules on graphs

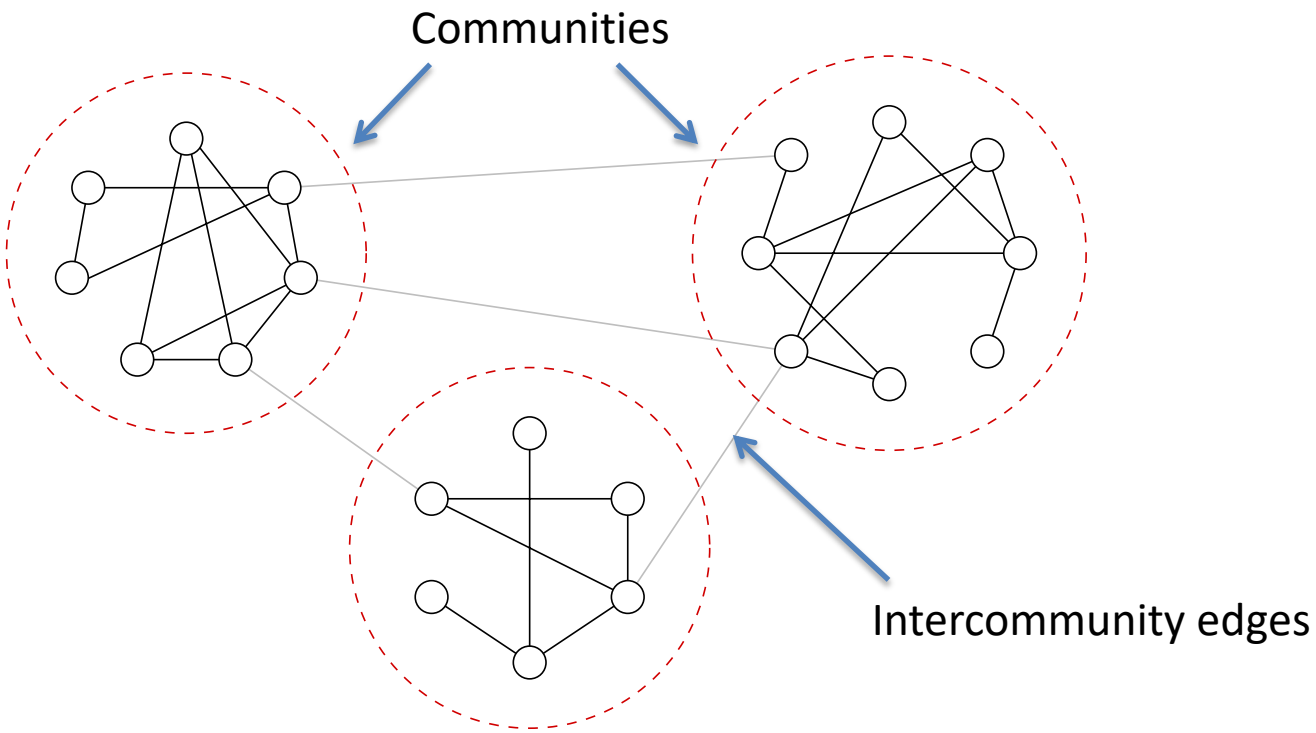
- What is a good measure of similarity to cluster nodes on a graph?
- One approach is to use local topological overlap
  - Find the similarity between the local neighborhoods of two nodes *i* and *j*

$$t_{ij} = \frac{|N_1(i) \cap N_1(j)| + a_{ij}}{\min(|N_1(i)|, |N_1(j)|) + 1 - a_{ij}}$$

- $N_1(i)$: the set of immediate neighbors of i
- a_{ij}: corresponding entry in the adjacency matrix of the graph

Yip and Horvath, BMC Bioinformatics 2007; WGCNA: an R package for weighted correlation network analysis. Peter Langfelder and Steve Horvath

Community structure in a graph



- A graph that has a grouping (community) structure is going to have few intercommunity edges.
- Community structure can be revealed by removing such intercommunity edges

Detecting community structure in networks, M. E. J. Newman

Girvan-Newman algorithm

- General idea: “If two communities are joined by only a few inter-community edges, then all paths through the network from vertices in one community to vertices in the other must pass along one of those few edges.”
- Community structure can be revealed by removing edges that with high betweenness
- Algorithm is based on a divisive clustering idea

M. E. J. Newman and M. Girvan. Finding and evaluating community structure

Betweenness of an edge

- Betweenness of an edge e is defined as the number of shortest paths that include e
- Edges that lie between communities tend to have high betweenness

$$B(e) = \frac{\text{Shortest path including } e}{\text{Number of total shortest paths}}$$

Girvan-Newman algorithm

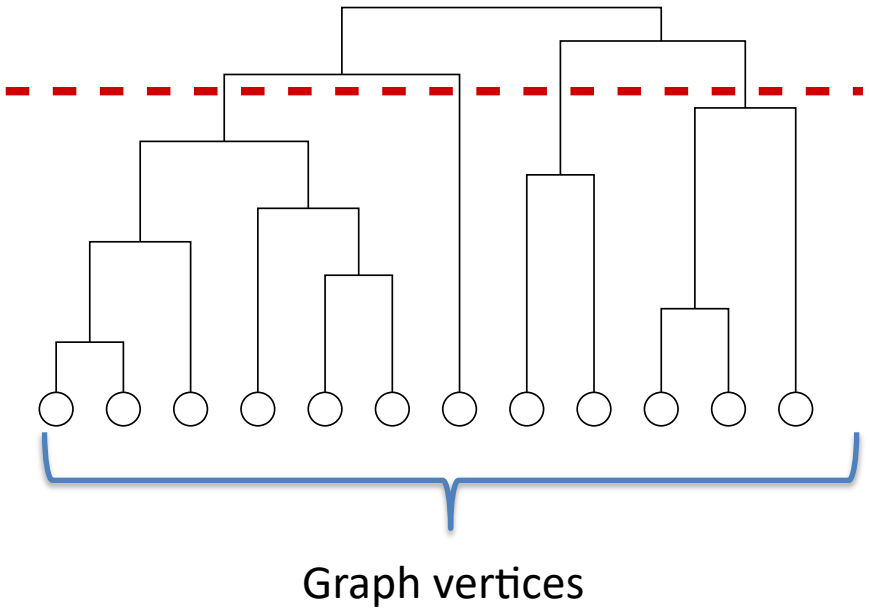
- Initialize
  - Compute betweenness for all edges
- Repeat until convergence criteria
  1. Remove the edge with the highest betweenness
  2. Recompute betweenness of remaining edges

- Convergence criteria can be
  - No more edges
  - Desired modularity

M. E. J. Newman and M. Girvan. Finding and evaluating community structure

Girvan-Newman algorithm as a hierarchical clustering algorithm

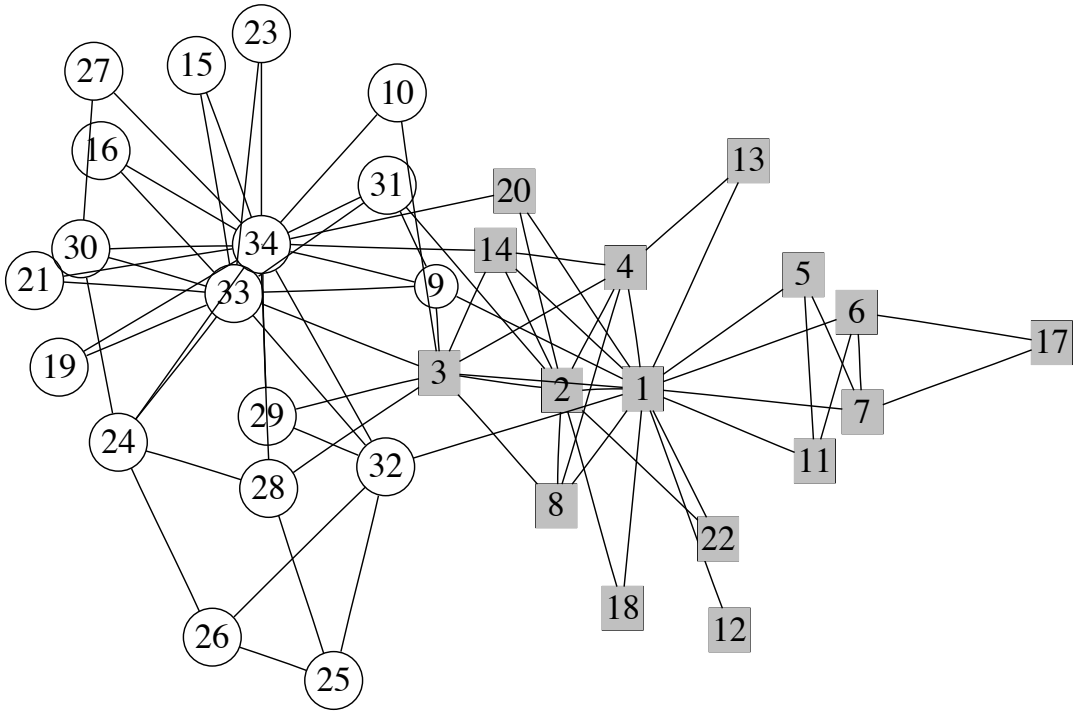
- One can view this algorithm as a top-down (divisive) hierarchical clustering algorithm
- The root of the dendrogram groups all nodes into one community
- Each branch of the tree represents the order of splitting the network as edges are removed

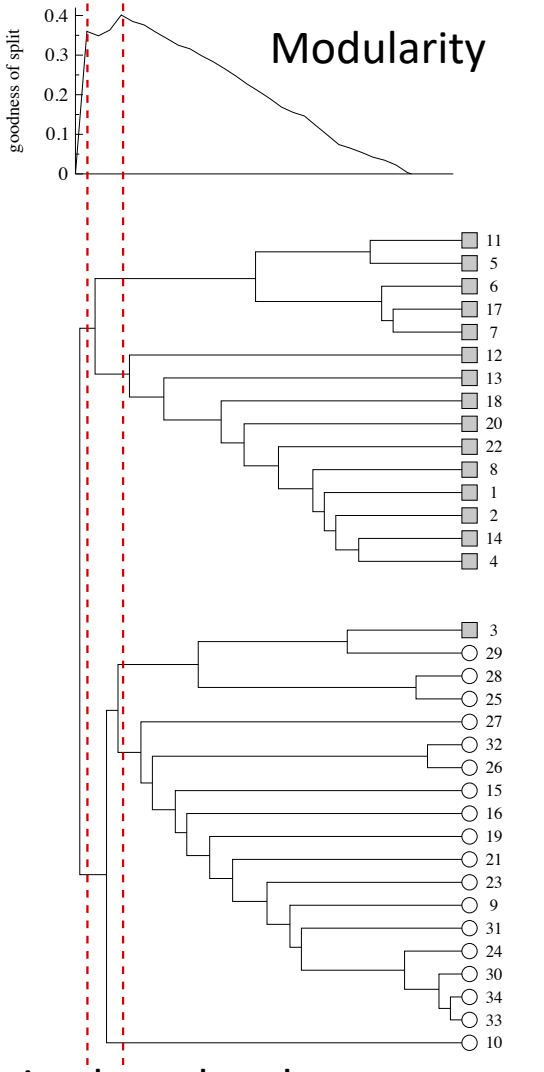


Applying the Girvan-Newman algorithm to Zachary's karate club network

- Dataset collected by Wayne Zachary over 2 years who observed social interactions among members of a karate club
- Zachary's karate club network is a well-known example of a social network with community structure
- Network represents the friendships among members of a karate club
- Due to a dispute the club split into two factions
- Can a graph clustering/module detection algorithm predict the factions?

Zachary's karate club study





Node grouping based on betweenness

Each node is an individual and edges represent social interactions among individuals.
The shape and colors represent different groups.

Take away points

- Biological networks are modular
- Modules can be topological or functional
- Modularity can be measured using
  - Clustering coefficient
  - Q measure
- We have seen one example of topological clustering algorithms
  - Girvan-Newman algorithm
    - based on edge-betweenness
    - Can be viewed as top-down/divisive clustering algorithm

References

- A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101-113, Feb. 2004. [Online]. Available: http://dx.doi.org/10.1038/nrg1272
- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824-827, Oct. 2002. [Online]. Available: http://dx.doi.org/10.1126/science.298.5594.824
- U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 450-461, Jun. 2007. [Online]. Available: http://dx.doi.org/10.1038/nrg2102
- S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152-1153, May 2006. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btl038
- M. E. J. Newman and M. Girvan. “Finding and evaluating community structure in networks”, 2003, 10.1103/PhysRevE.69.026113
- M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577-8582, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1073/pnas.0601602103
- A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661-703, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1137/070710111