582746 Modelling and Analysis in Bioinformatics

Lecture 1: Global Network Models

26.10.2015

Outline

Course introduction

Examples of biological networks

Global properties of networks
- Distance measures
- Degree measures
- Local clusters

Network Models
- Erdös-Renyi Model
- Watts-Strogatz Model
- Barabasi-Albert Model

More Network Properties

Statistical Testing of Network Properties

Outline

Course introduction

Examples of biological networks

Global properties of networks
- Distance measures
- Degree measures
- Local clusters

Network Models
- Erdös-Renyi Model
- Watts-Strogatz Model
- Barabasi-Albert Model

More Network Properties

Statistical Testing of Network Properties

Course topics

- Computational models for biological networks (Leena Salmela)
- Gene regulation (Antti Honkela)
- Probabilistic analysis of sequence level problems (Veli Mäkinen)

Practical arrangements

- Mondays: Lectures to introduce the topics
- Thursday mornings: Study group to deepen the knowledge on the subject
- Thursday afternoons: Exercise sessions
- 23.11.-27.11. Visiting lecturers (no exercise session)

How to pass the course?

- Attending study groups on Thursday mornings is mandatory
- Attending visiting lectures on Monday 23.11. and Thursday 26.11. is mandatory
- Submit the exercises and get at least 6 points for each three exercise sets (network models, gene regulation, probabilistic analysis of sequence-level problems)
- If you miss a study group or visiting lecture, contact the lecturers for an alternative assignment

Grading

- Grading is based on submitted exercises
- 60 points will be available
- 30 points $\Longrightarrow$ Passed, 50 points $\Longrightarrow$ 5
- No exam
- Not possible to pass with a separate exam

Outline

Course introduction

Examples of biological networks

Global properties of networks
- Distance measures
- Degree measures
- Local clusters

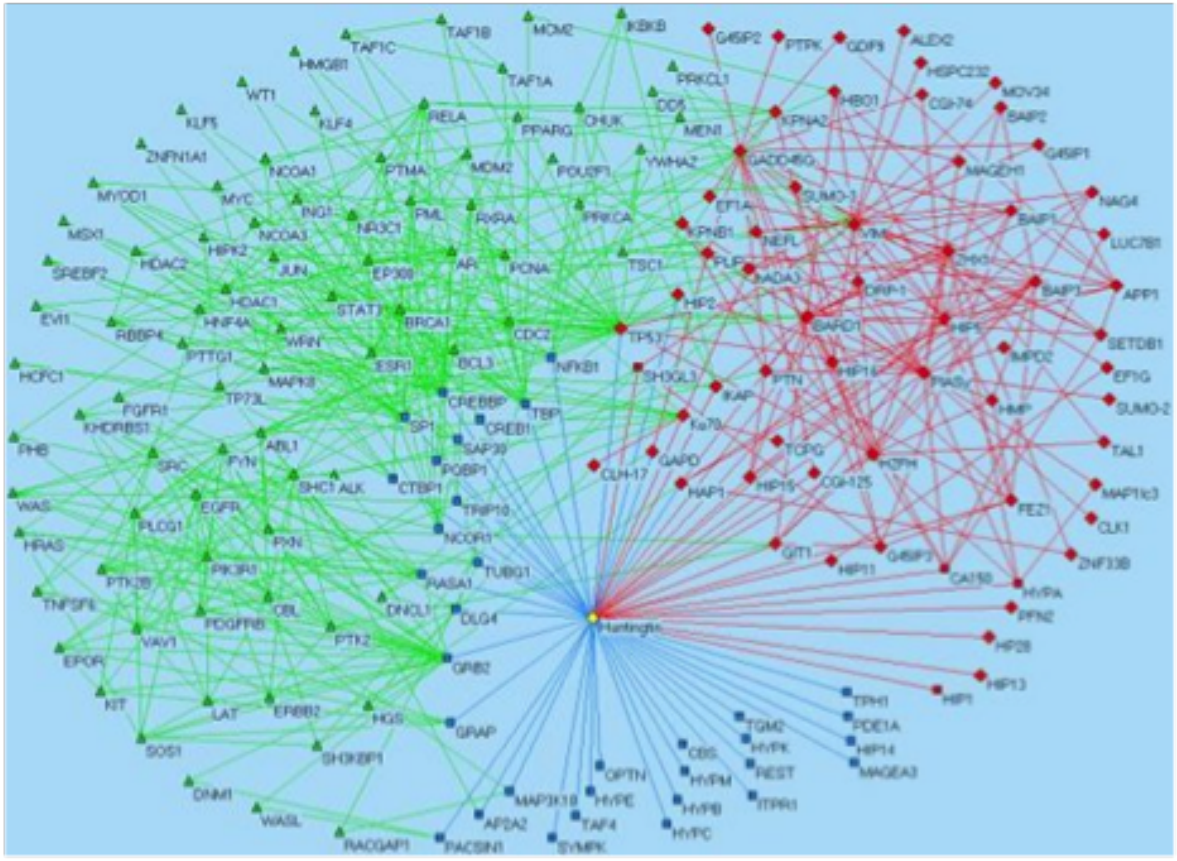
Network Models
- Erdös-Renyi Model
- Watts-Strogatz Model
- Barabasi-Albert Model

More Network Properties

Statistical Testing of Network Properties

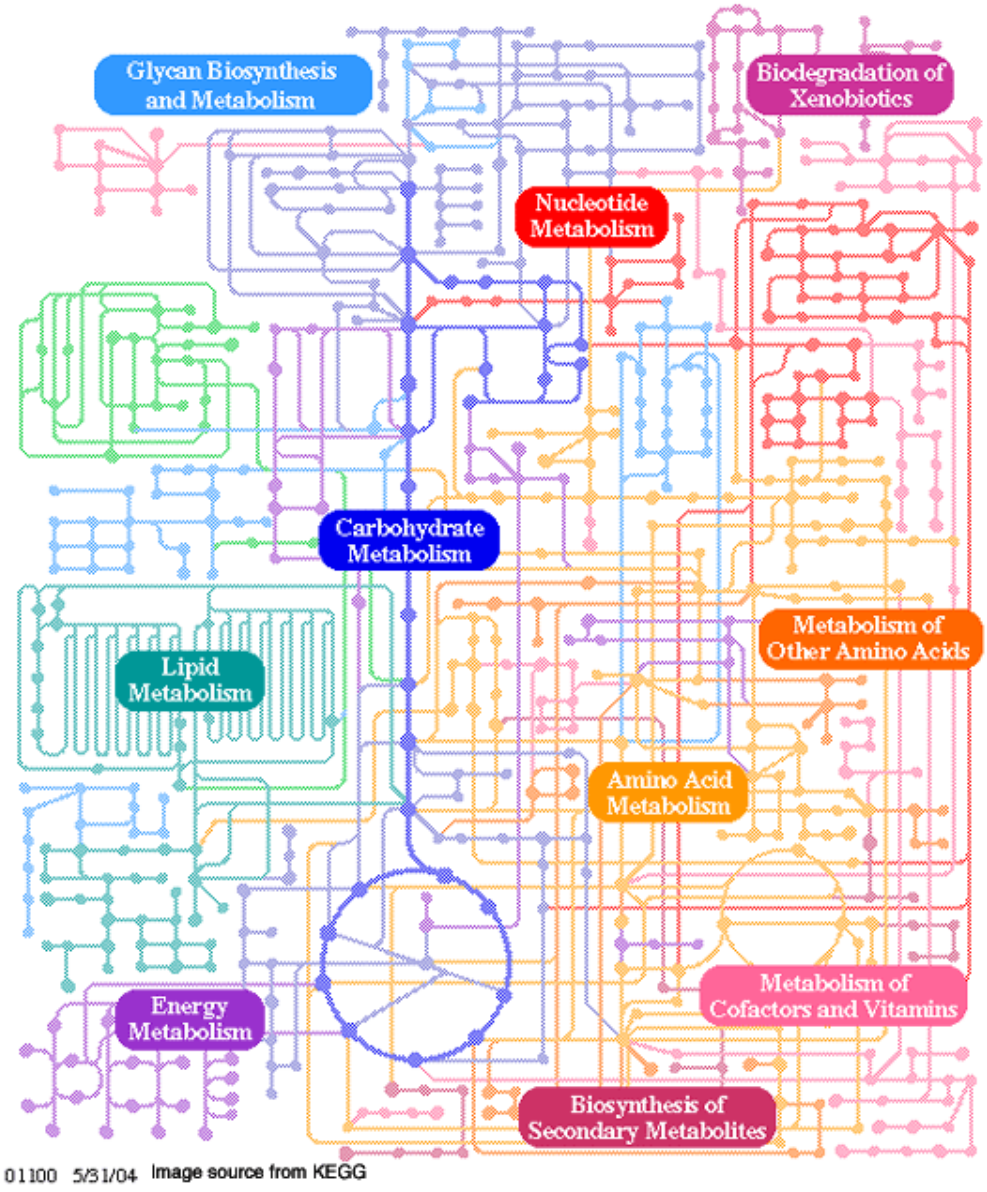
Protein-protein interaction network

- Vertices are proteins
- The proteins are connected if they interact with each other.



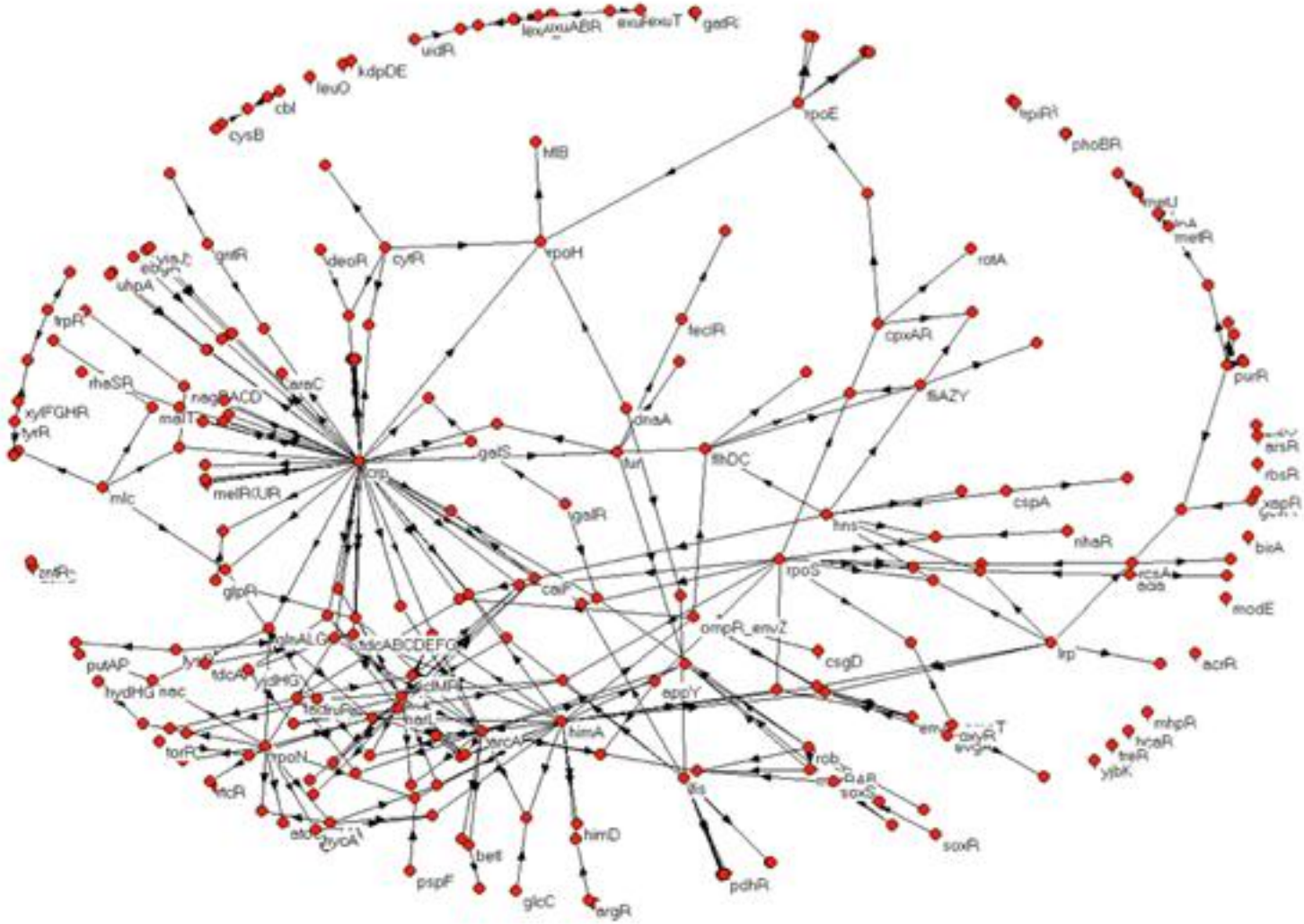
Metabolic network

- Vertices are metabolites, i.e. chemical compounds
- Edges describe how the cell can transform a metabolite into another



Gene regulatory network

- Vertices are genes
- Genes are linked if one regulates the other



Source: Shen-Orr SS, Milo R, Mangan S, Alon U. 2002

Outline

Course introduction

Examples of biological networks

Global properties of networks
- Distance measures
- Degree measures
- Local clusters

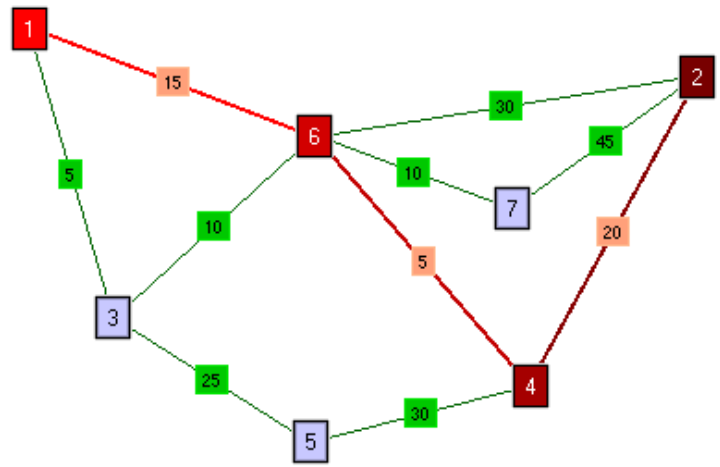
Network Models
- Erdös-Renyi Model
- Watts-Strogatz Model
- Barabasi-Albert Model

More Network Properties

Statistical Testing of Network Properties

(Shortest path) distance

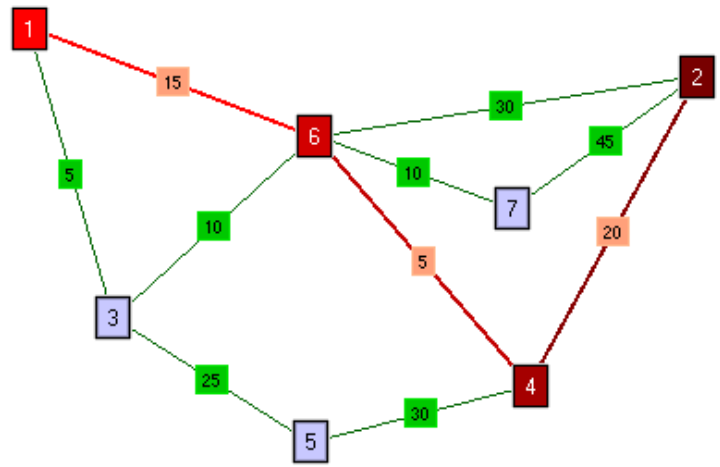
- Distance d_{ij} is the length of the shortest path between vertices n_i and n_j, i.e. the minimal number of edges one needs to traverse to get from n_i to n_j
- The shortest path may not be unique, but the *length* of the shortest path is unique
- In directed network, we may have $d_{ij} \neq d_{ji}$
- If there is no path between n_i and n_j, we have $d_{ij} = \infty$



Ignoring weights, above we have $d_{12} = 2$, $d_{13} = 1, d_{14} = 2, \ldots$

Diameter and average path length

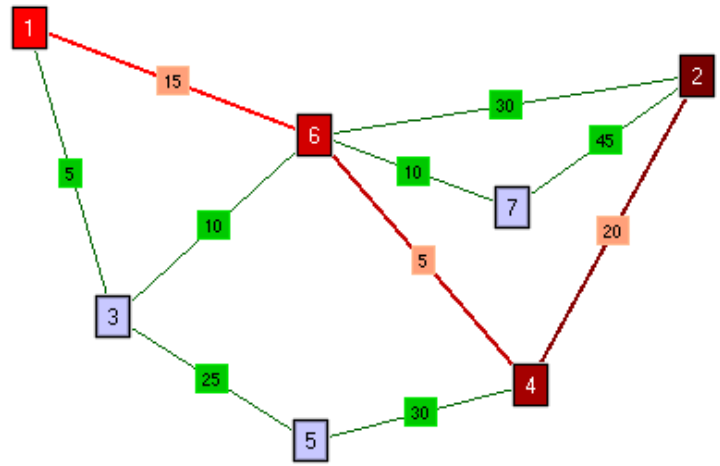
- The diameter $d_m = \max(d_{ij})$ is the maximal distance between any two nodes (= the longest shortest path)
- Average or characteristic path length
  $d = \langle d_{ij} \rangle = \frac{1}{N_V^2} \sum_{i=1}^{N_V} \sum_{j=1}^{N_V} d_{ij}$



Ignoring weights, above
$d_m = 3$, $d \approx 1.57$

Efficiency

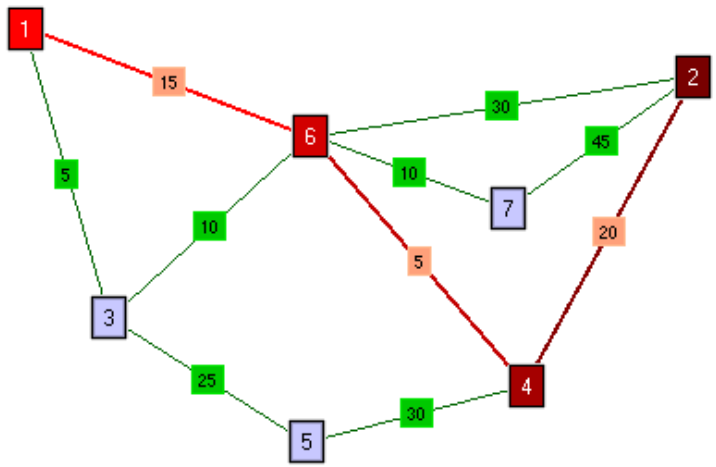
- Efficiency, or average inverse path length: $d_{eff} = \langle 1/d_{ij} \rangle$
- Useful when average path length is infinite (disconnected network)
- Fully connected network has efficiency $d_{eff} = 1$, graph with no edges has $d_{eff} = 0$



Ignoring weights, above $d_{eff} \approx 0.73$

Weighted graphs

- If the edges in the graph have associated weights w_{ij}, it is natural to define distances based on the weights:
- d_{ij} as the sum of weights in the minimum weight path between n_i and n_j
- Maximum and average path length as well as efficiency naturally generalize by changing the distance measure to the weighted version



With weights, $d_{12} = 40$ (red path), $d \approx 24.2$, $d_m = 50 = d_{25}, d_{eff} = 0.06$

Finding shortest paths

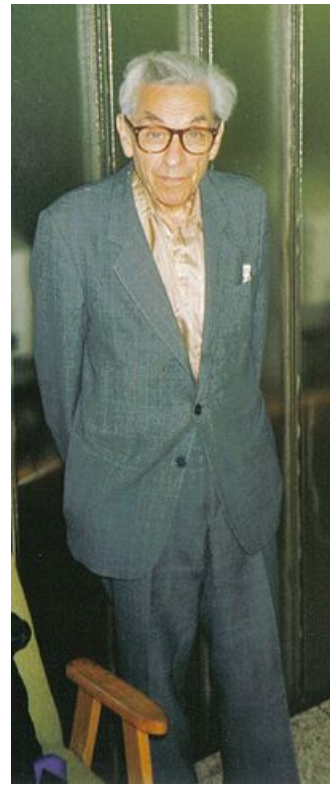
Finding shortest paths in graphs is part of classical algorithm theory, two efficient algorithms

- Dijkstra's algorithm: given a vertex find shortest paths to all other vertices, basic implementation runs in $O(N_V^2)$ time, can be implemented faster for sparse graphs
- Floyd-Warshall algorithm: find shortest paths for all pairs of vertices in the graph in $O(N_V^3)$ time; outputs a distance matrix $(d_{ij})_{i,j=1}^{N_V}$ in same time.
- Both work with weighted formulations

Shortest path distances in empirical networks

Path length analysis of many networks that occur in nature reveals the small-world property

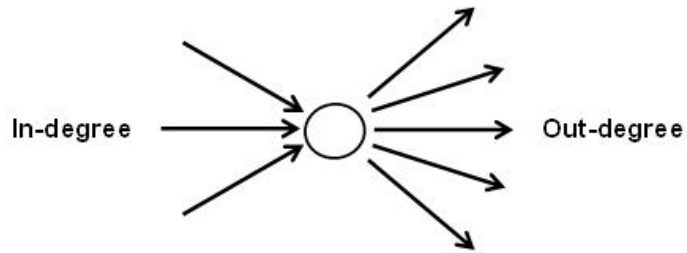
- Metabolite graphs: average path length $d \approx 3$ ($N_V \approx 10^3 - 10^4$)
- WWW: links chains between two web documents $d \approx 16$ ($N_V > 10^9$)
- Erdös number: shortest co-author chain to Paul Erdös, $d = 4.65$ ($N_V \approx 4 \times 10^5$)
- ...



Paul Erdös (1913-1996, a Hungarian mathematician, published over 1400 scientific papers over his lifetime with over 500 different co-authors

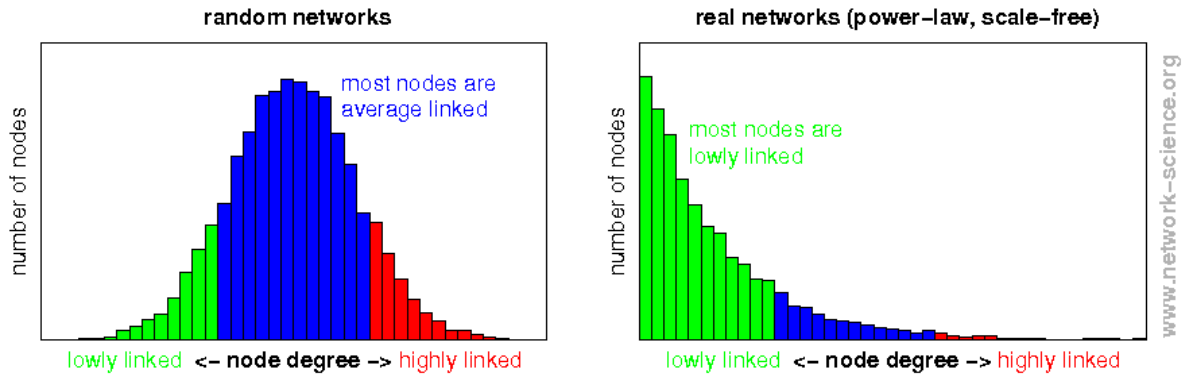
Node degree

- Degree k_i of vertex n_i is the number of edges adjacent to a vertex
- In a network without self-loops and without multiple edges between any pair of edges: degree = number of neighbours
- In directed networks: in-degree is the number of incoming edges and out-degree is the number outgoing edges



Degree distribution

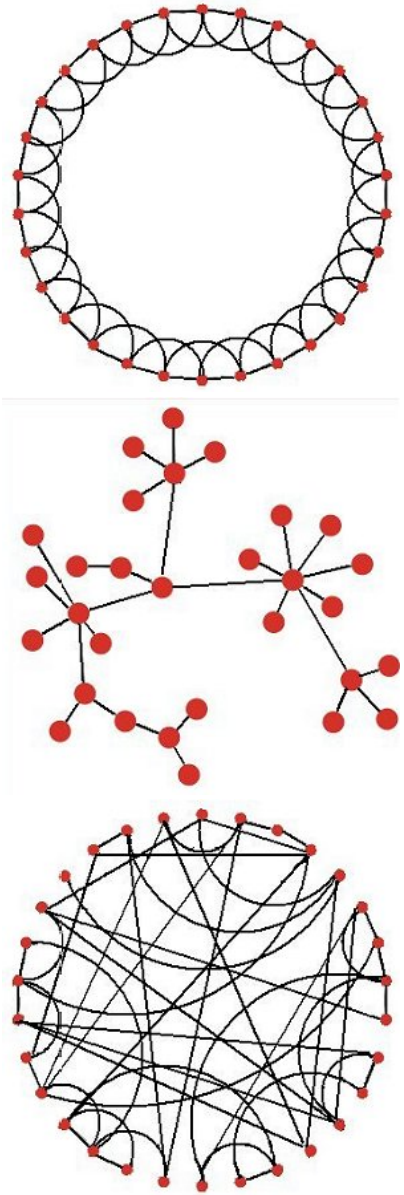
- Given a fixed set of vertices, $p(k)$ denotes the probability that a randomly chosen vertex has degree k.
- (Empirical) degree distribution is the list of probabilities (or relative frequencies) $p(k)$, $k = 0 \ldots N_V$.
- Analysis of the degree distribution is an important means to characterize networks



Degree distribution

Fitting the empirical degree distribution to a theoretical distribution given by a mathematical law is an important tool for network analysis

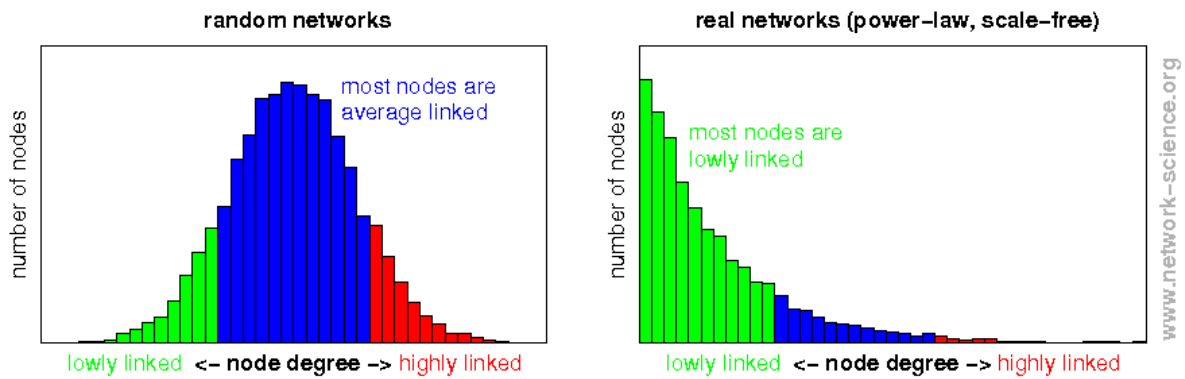
- Regular lattice: $p(k) \approx 1$, where k is a constant
- Scale free network: $p(k) \propto k^{-\gamma}$
- Random network: $p(k) \propto \binom{N_V - 1}{k} p^k (1-p)^{N_V - 1 - k}$



Degree distribution

The degree distributions of of scale-free network and random network look markedly different

- Scale free network: $p(k) \propto k^{-\gamma}$ (power law, heavy tail)
- Random network: $p(k) \propto \binom{N_V - 1}{k} p^k (1-p)^{N_V - 1 - k}$ (binomial, light tail)



Fitting degree distributions

- Typically the fitting of the empirical distribution is based on the histogram of observations for $p(k)$
- This is prone to errors in the region of high degree nodes due to low number of observations
- Binning can help: divide the range of k into intervals and put all observations in the interval into a common bin
- Cumulative degree distribution $p_c(k) = \sum_{l=k}^{\infty} p(l)$, the likelyhood that a given node has degree at least k, is more reliable and does not require binning

Degree correlations and assortative mixing

Degree correlation is a statistic that reveals additional information of the connection patterns of the nodes

- Assortative networks: high correlation between the degrees of adjacent nodes; highly connected nodes mostly connect to other highly connected nodes
- Disassortative networks: highly connected nodes mostly connect to low degree nodes
- Assortativity index $-1 \leq r \leq 1$: Pearson correlation coefficient of degrees of adjacent nodes, $r > 0$ assortative, $r < 0$ disassortative

Examples

Social networks are typically assortative, technological and biological networks tend to be disassortative

	Group	Network	Type	Size n	Assortativity r	Error σ_r
Social	a	Physics coauthorship	undirected	52 909	0.363	0.002
	a	Biology coauthorship	undirected	1 520 251	0.127	0.0004
	b	Mathematics coauthorship	undirected	253 339	0.120	0.002
	c	Film actor collaborations	undirected	449 913	0.208	0.0002
	d	Company directors	undirected	7 673	0.276	0.004
	e	Student relationships	undirected	573	−0.029	0.037
	f	Email address books	directed	16 881	0.092	0.004
Technological	g	Power grid	undirected	4 941	−0.003	0.013
	h	Internet	undirected	10 697	−0.189	0.002
	i	World Wide Web	directed	269 504	−0.067	0.0002
	j	Software dependencies	directed	3 162	−0.016	0.020
Biological	k	Protein interactions	undirected	2 115	−0.156	0.010
	l	Metabolic network	undirected	765	−0.240	0.007
	m	Neural network	directed	307	−0.226	0.016
	n	Marine food web	directed	134	−0.263	0.037
	o	Freshwater food web	directed	92	−0.326	0.031

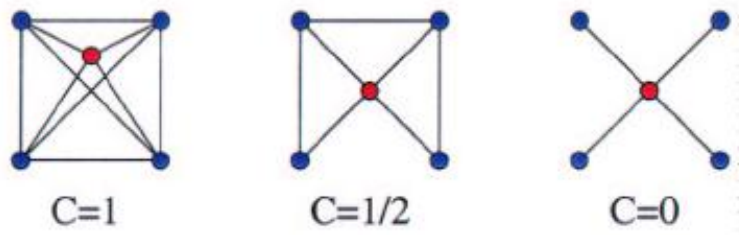
(M. Newman. Phys. Rev. E 67, 026126 (2003))

Clustering coefficient

- Clustering coefficient measures the probability that two vertices with a common neighbor are connected
- Let E_i denote the number of edges between the neighbors of v_i, and $E_{max} = k_i(k_i - 1)/2$ the theoretical maximum. Clustering coefficient for vertex $n_i$is now

$$C_i = \frac{E_i}{E_{max}} = \frac{2E_i}{k_i(k_i - 1)}$$

- Clustering coefficient for the whole graph is obtained by averaging over the vertices

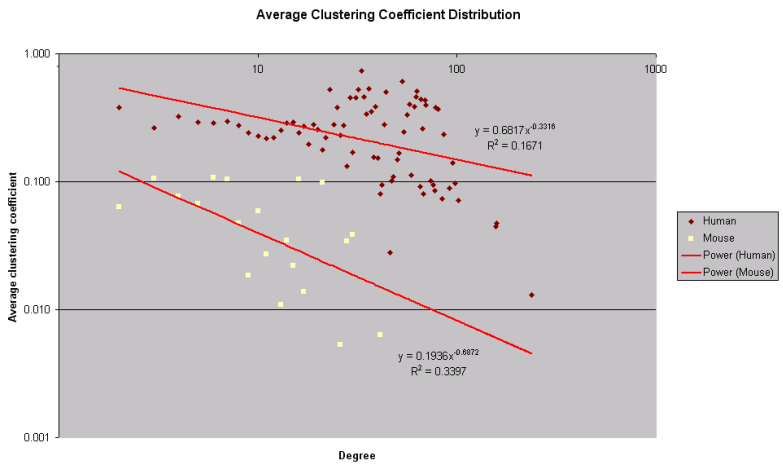


http://www.nd.edu/~swuchty/Download/Wuc.pdf, p. 24, Fig. 2

Clustering coefficient in natural networks

- Natural networks often have relatively high clustering coefficient indicating local clustering within the network
- Negative correlation between the degree and the clustering coefficient has also been observed;
  - Low degree nodes lie in local clusters, while the neighbors of high degree nodes are less often connected
  - Indicates modular network structure

Example: PPIs in Mouse and Human:

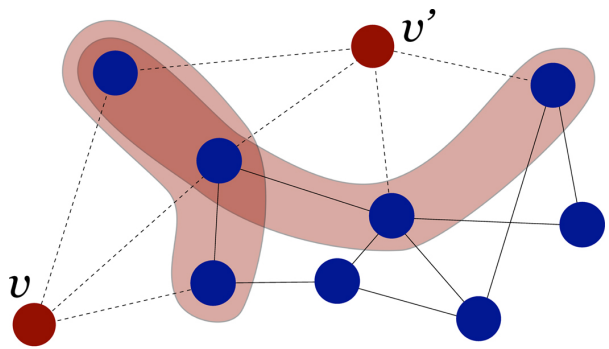


(http://bccs.bristol.ac.uk/toProgramme/_project/2008/Angela_Onslow_S08/)

Matching index

To be functionally related, two vertices do not need to be connected, examples:

- Two transcription factor proteins regulating the same gene
- Two metabolite molecules taking part in similar reactions



$$MI(v, v') = \frac{2}{5} = 0.4$$

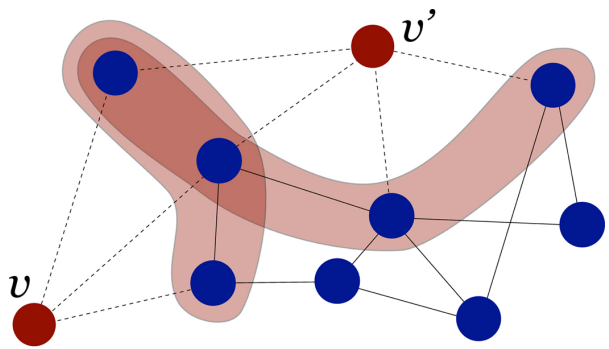
Zamora-Lopez. Frontiers in Neuroinformatics 4, 2010

Matching index

- Matching index measures the amount of neighbors the two nodes share:

$$MI_{ij} = \frac{Shared_{ij}}{k_i + k_j - Shared_{ij}}$$

- Similarity in terms of perceiving the neighborhood similarly



$$MI(v, v') = \frac{2}{5} = 0.4$$

Zamora-Lopez. Frontiers in Neuroinformatics 4, 2010

Outline

Course introduction

Examples of biological networks

Global properties of networks
- Distance measures
- Degree measures
- Local clusters

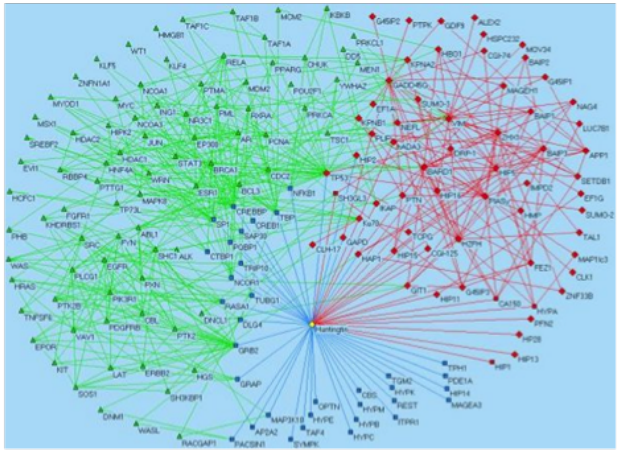
Network Models
- Erdös-Renyi Model
- Watts-Strogatz Model
- Barabasi-Albert Model

More Network Properties

Statistical Testing of Network Properties

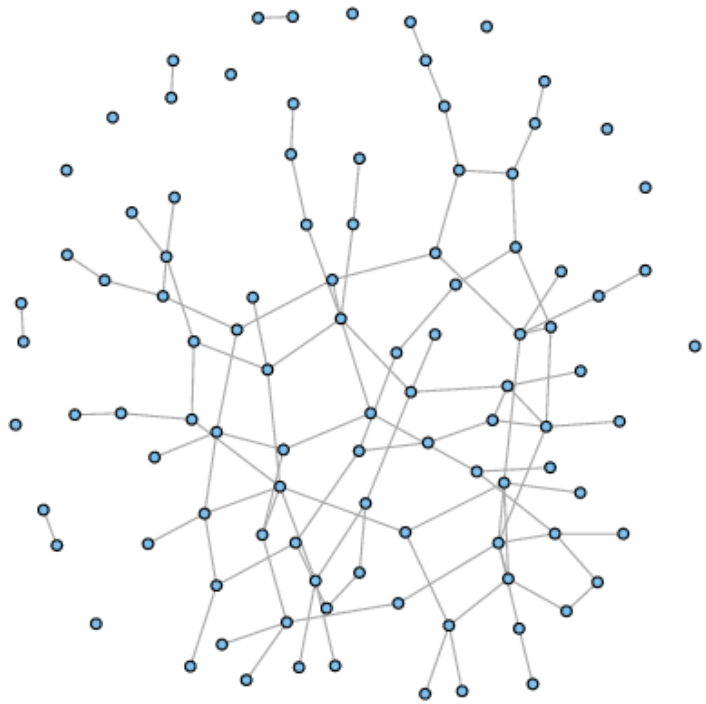
Models of complex networks

- Theoretical models of networks are needed as a basis for comparison to determine the significance of global properties or non-trivial substructures of natural networks.
- We will look at three specific models
  - Erdös-Renyi Model
  - Watts-Strogatz Model
  - Barabasi-Albert model



Erdös-Renyi Model

- ER network consists of N_V vertices
- Edge is drawn between a pair of nodes randomly with probability p
- Degree distribution of the ER model is binomial: $p(k) \propto \binom{N_V-1}{k} p^k (1-p)^{N_V-1-k}$
- Degree distribution can be approximated by Poisson distribution for large graphs



Erdös-Renyi Model

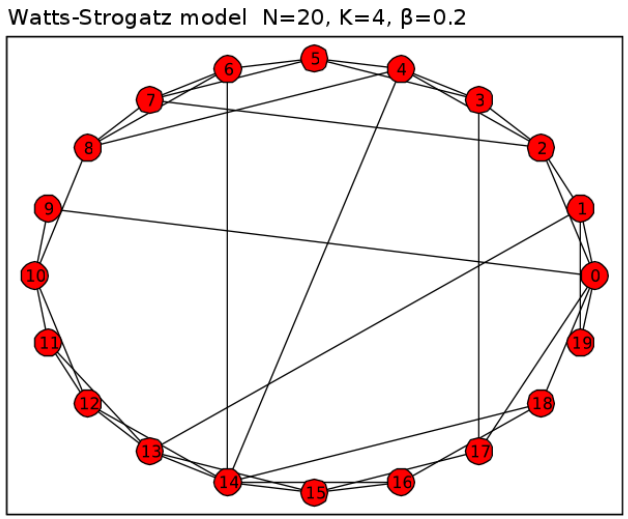
Significant body of theoretical research exists for the ER model, e.g.

- For $N_V p < 1$ the network almost surely has no large connected components
- For $N_V p \approx 1$ the network will almost surely have one large connected component
- For $N_V p > \log N_V$ the network will almost surely be connected
- ER network has the small-world property when $p > 1/N_V$ with average path length scaling as $l \sim \log N_V$
- No local clustering, expected clustering coefficient $C = p = \langle k \rangle / N_V$ for all nodes

Watts-Strogatz model

1. Arrange vertices in a ring structure
2. Connect each vertex to K closest neighbours
3. With probability p_{rew}, rewire each each edge by detaching from one end and attaching to a randomly chosen vertex.

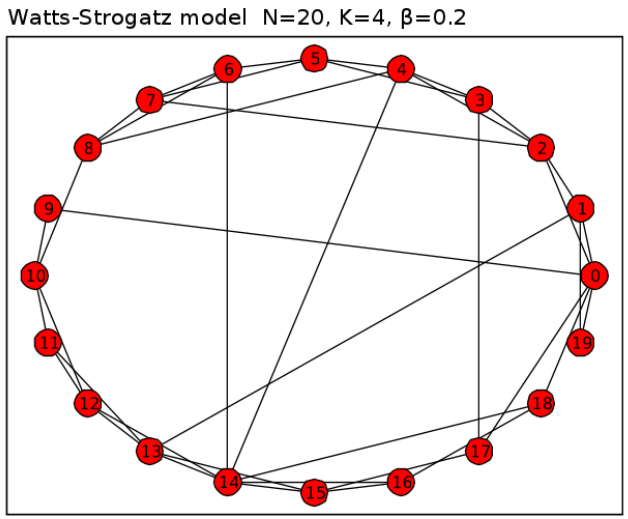
After steps (1-2) there is local clustering, step (3) lowers average path length by creating shortcuts



http://en.wikipedia.org/wiki/File:Watts_strogatz.svg

Watts-Strogatz model

- Even for low rewiring probability ($p_{rew} << 1$) the average path length goes down rapidly
- Small average path length and local clustering is retained for intermediate p_{rew}
- When $p_{rew} \to 1$, we get ER model, i.e. local clustering is destroyed
- Degree distribution is similar to ER graph: homogeneuos and peaked around $k = K$

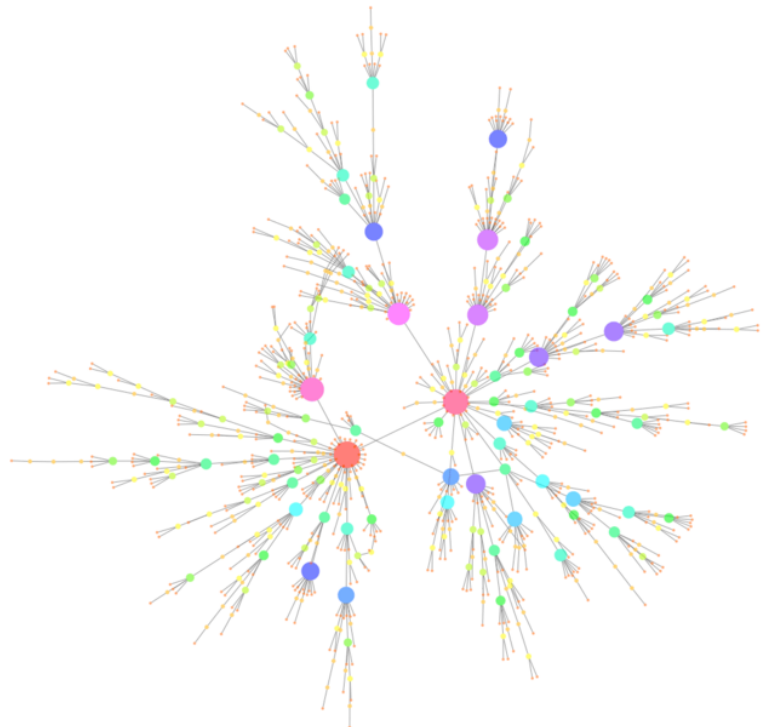


http://en.wikipedia.org/wiki/File:Watts_strogatz.svg

Barabasi-Albert model

- Start with an initial small connected network of N_0 vertices
- Iteratively add new vertices and connect the new vertex to $m \leq N_0$ vertices
- Draw the nodes that will be connected the the new vertex with probability proportional to their degree (preferential attachment):

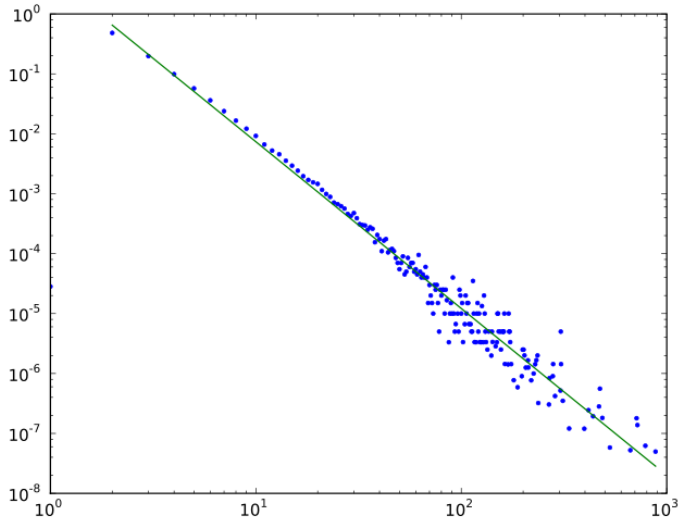
$$\rho(n_i) = k_i / \sum_j k_j$$



(BA graph from http://melihsozdinler.blogspot.com/)

Barabasi-Albert model

- Unlike ER or WS model, Barabasi-Albert model explain the inhomogeneuos degree distribution observed in natural graphs
- With enough iterations, the degree distribution of the BA model is scale-free, with $p(k) \sim k^{-3}$
- Average path length in BA networks has been found to be smaller than in ER and WS models



Outline

Course introduction

Examples of biological networks

Global properties of networks
- Distance measures
- Degree measures
- Local clusters

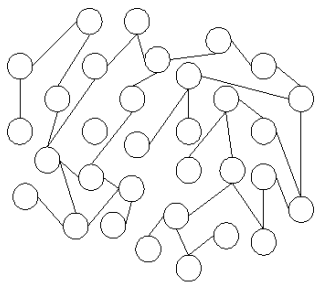
Network Models
- Erdös-Renyi Model
- Watts-Strogatz Model
- Barabasi-Albert Model

More Network Properties

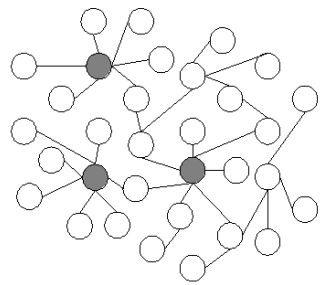
Statistical Testing of Network Properties

Robustness and Attack tolerance

- Robustness against pertubations (mutations, environment changes) is a preferable property for biological networks
- Networks analysis is interested in preservation of network topology under perturbations (usually: removals of vertices or edges)



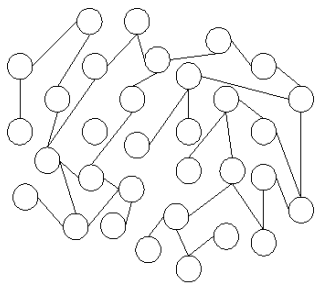
(a) Random network



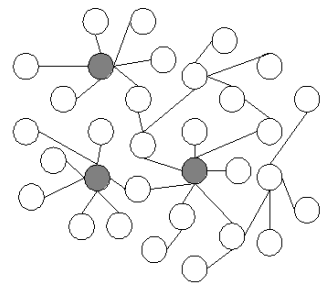
(b) Scale-free network

Robustness and Attack tolerance

- Both ER networks and scale-free networks (such as BA model) are robust towards random deletions of nodes and connections
  - A random mutation is likely to hit a low degree node in BA model
- Scale-free networks are not robust towards intentioanl attacks
  - Removal a set of highly connected nodes may collapse the global structure
  - "Robust, yet fragile"



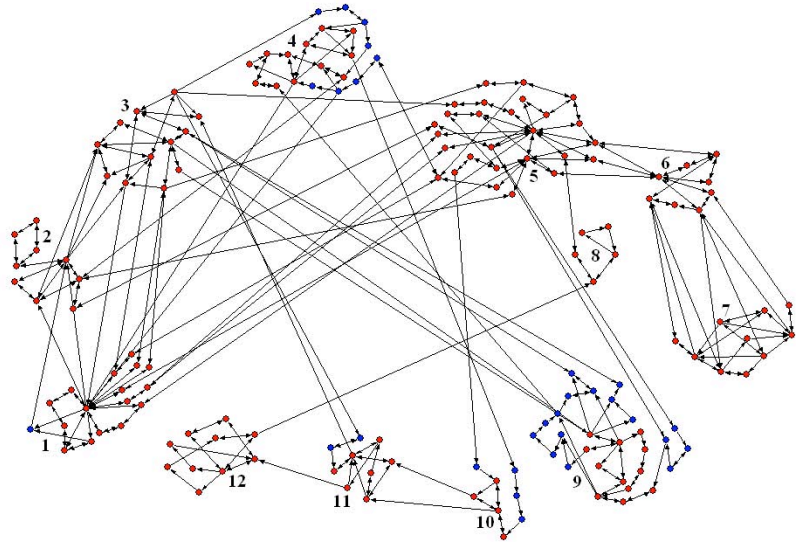
(a) Random network



(b) Scale-free network

Modularity and hierarchical organization

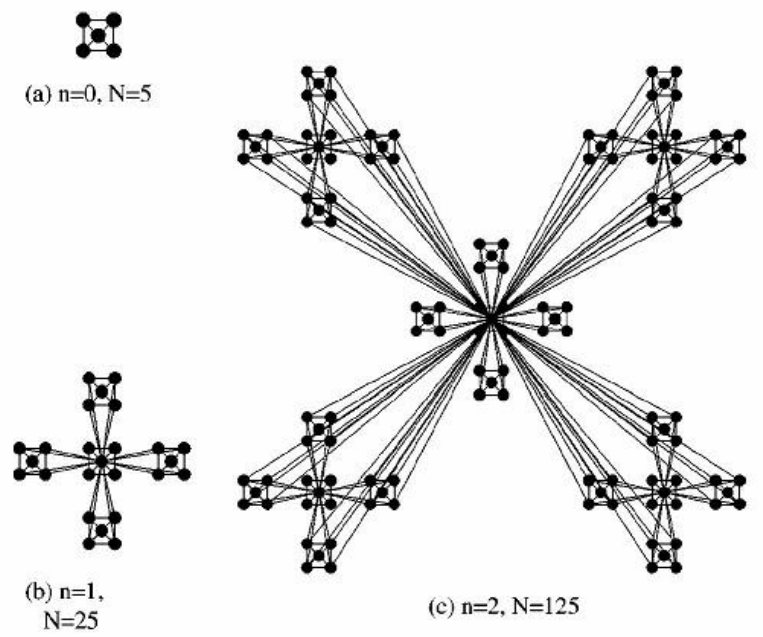
- Many natural networks are observed to posses modular structure with densely connected functional clusters of nodes that are sparsely connected to other nodes.
- Also, hierarchical organization of network structure can be observed
- The random network models discussed above, do not directly explain these phenomena



(Zhao et al. BMC Bioinformatics 2006, 7:386)

Modularity and hierarchical organization

- Barabasi and Albert model has been later extended to that direction
  - Based on replicating basic modules and wiring them to the central module of rest of the network
  - Recursive application leads to hierarchical organization
  - Deterministic rather than random procedure



Outline

Course introduction

Examples of biological networks

Global properties of networks
- Distance measures
- Degree measures
- Local clusters

Network Models
- Erdös-Renyi Model
- Watts-Strogatz Model
- Barabasi-Albert Model

More Network Properties

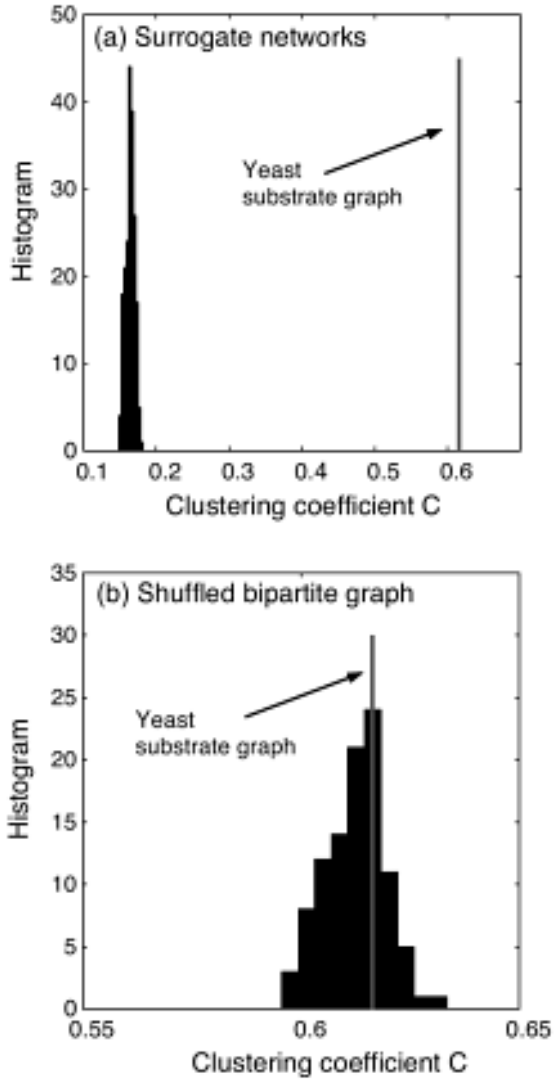
Statistical Testing of Network Properties

Statistical testing of network properties

- How to determine if an observed property of the network is significant or if it occured just by chance?
- Set up a null hypothesis
- Test if the observed property is consistent with the null hypothesis

Statistical testing of network properties: Example

- Suppose that we have observed a clustering coefficient C for a given network. Is the network highly clustered?
- Null hypothesis: The clustering coefficient is consistent with a network of the same size and degree distribution.
- Create an ensemble of random networks with same size and degree distribution and compute the clustering coefficient of each network
- Reject the null hypothesis if the probability of a network with clustering coefficient of at least C is low enough
- If the null hypothesis can be rejected, we can conclude that the network is highly clustered as compared to the null model.



What next?

- Thursday 10-12: Study group on analytical properties of ER networks
- Thursday 12-14: Exercise session

Moodle Enrolment

- `https://moodle.helsinki.fi/course/view.php?id=18471`
- Enrolment key: BIOMODELS