Algorithms for Biological Networks

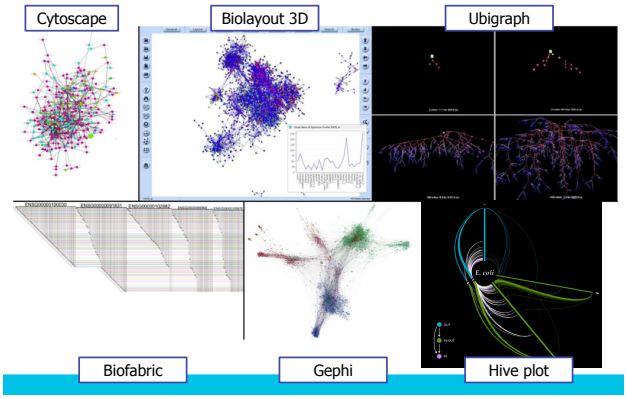
Day 1 Part C – Network Visualization and Topology

Jeroen de Ridder — **Delft University of Technology**
Dick de Ridder — WUR, Wageningen
Anton Feenstra — Vrije Universiteit, Amsterdam
Nicola Bonzanni — Vrije Universiteit, Amsterdam
Mohammed El-Kebir — CWI, Amsterdam
Aalt-Jan van Dijk — PRI/WUR, Wageningen

BioSB research school
Netherlands Bioinformatics and Systems Biology Research School

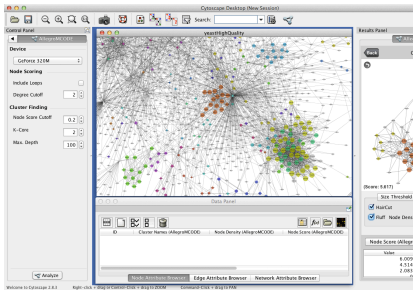
Network visualization

Graph visualization



Cytoscape

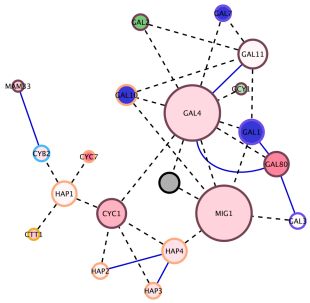
- Visualization
- Integration
- Analysis



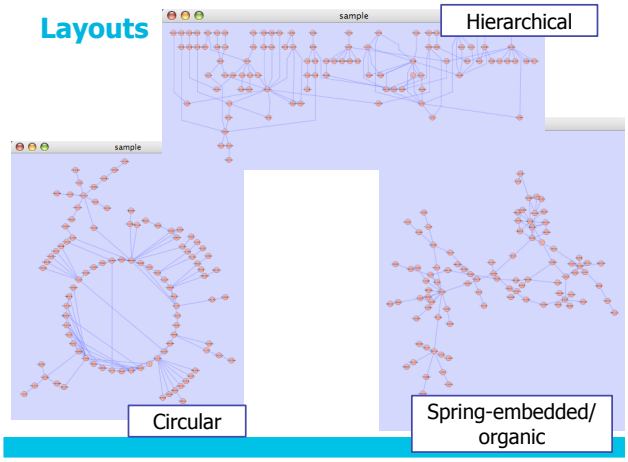
Mapping data on the nodes and edges

- Node-type
- Node-scores

- Link-type
- Link-score



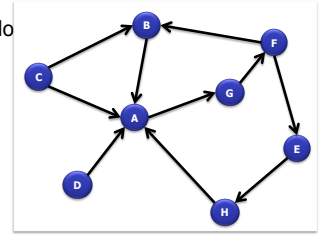
Layouts



Network properties

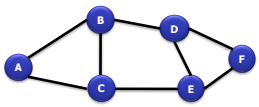
Directed graph

- What specific properties do biological network graphs possess?
  - Global vs Local
  - Shortest Paths
  - Centrality measures
  - Clustering coefficient



Undirected graphs

- Adjacency matrix is the matrix representation of a graph
- Symmetric for undirected graphs

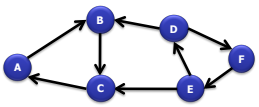


Adjacency matrix

$$A = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

Directed graphs

- Asymmetric for directed graphs

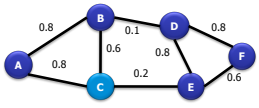


Adjacency matrix

$$A = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Weight graphs

- Weighted links are encoded in the adjacency matrix

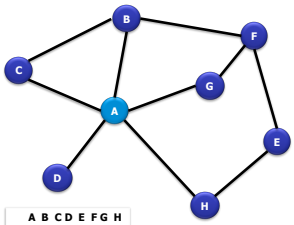


Weight matrix

$$W = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0 & 0.6 & 0.1 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0 & 0.8 & 0.8 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0.6 \\ 0 & 0 & 0 & 0.8 & 0.6 & 0 \end{pmatrix} \end{array}$$

Shortest paths

- Shortest path between two nodes: the minimum number of links to cross
- Local measure that characterizes the connection between two nodes



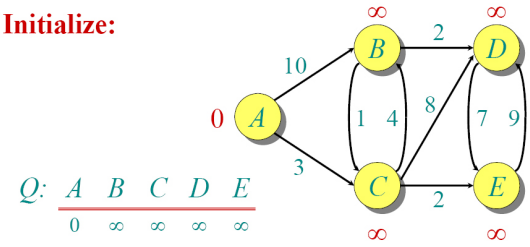
	A	B	C	D	E	F	G	H
A	0	1	1	1	2	2	1	1
B	0	0	1	2	2	1	2	2
C	0	0	0	2	3	2	2	2
D	0	0	0	0	3	3	2	2
E	0	0	0	0	0	1	2	1
F	0	0	0	0	0	0	1	2
G	0	0	0	0	0	0	1	2
H	0	0	0	0	0	0	0	0

Calculating shortest paths

- Dijkstra's Algorithm
  - a solution to the single-source shortest path problem in graph theory
  - takes into account weights on the edges
  - works on both directed and undirected graphs; However, all edges must have nonnegative weights
- Many other algorithms
  - Viterbi Algorithm - solves the shortest stochastic path problem with an additional probabilistic weight on each node
  - Floyd–Warshall algorithm – finds all pairwise shortest paths

Dijkstra Animated Example

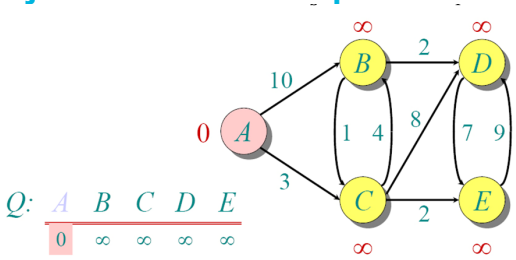
Initialize:



Q:	A	B	C	D	E
	0	∞	∞	∞	∞

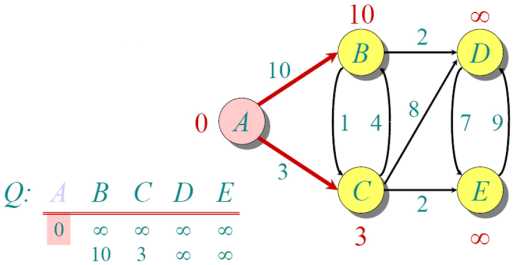
S: {}

Dijkstra Animated Example



Q:	A	B	C	D	E
	0	∞	∞	∞	∞

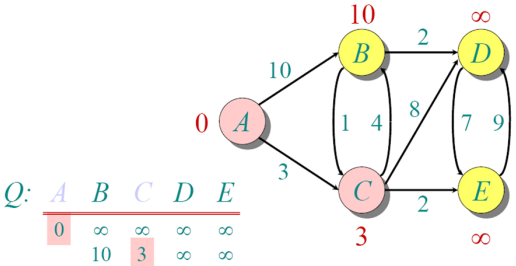
Dijkstra Animated Example



Q:	A	B	C	D	E
	0	∞	∞	∞	∞
		10	3	∞	∞

S: { A }

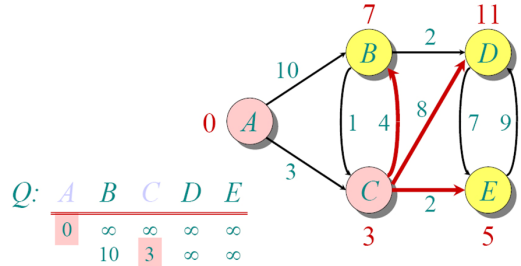
Dijkstra Animated Example



Q:	A	B	C	D	E
	0	∞	∞	∞	∞
		10	3	∞	∞

S: { A, C }

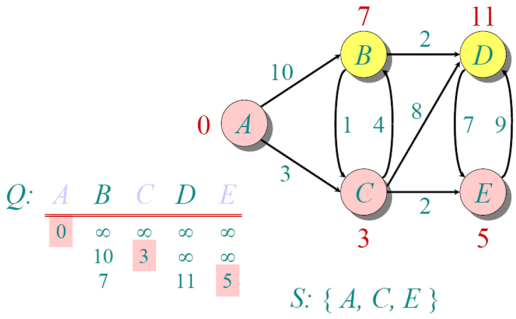
Dijkstra Animated Example



Q:	A	B	C	D	E
	0	∞	∞	∞	∞
		10	3	∞	∞
		7		11	5

S: { A, C }

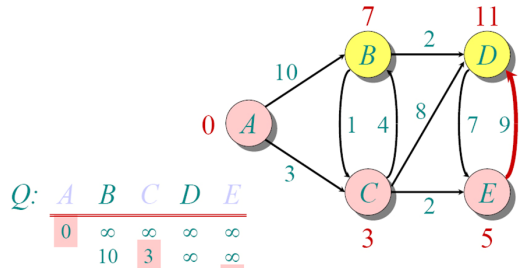
Dijkstra Animated Example



Q:	A	B	C	D	E
	0	∞	∞	∞	∞
		10	3	∞	∞
		7		11	5

S: { A, C, E }

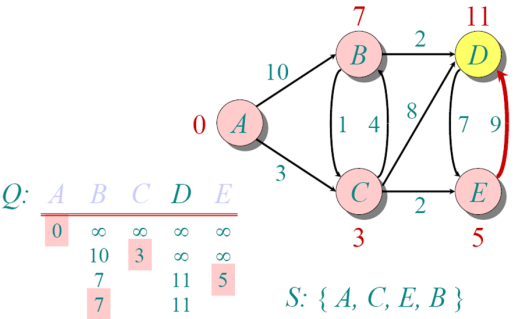
Dijkstra Animated Example



Q:	A	B	C	D	E
	0	∞	∞	∞	∞
		10	3	∞	∞
		7		11	5
		7		11	

S: { A, C, E }

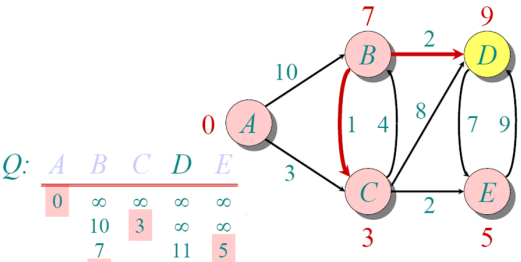
Dijkstra Animated Example



Q:	A	B	C	D	E
	0	∞	∞	∞	∞
		10	3	∞	∞
		7		11	5
		7		11	

S: { A, C, E, B }

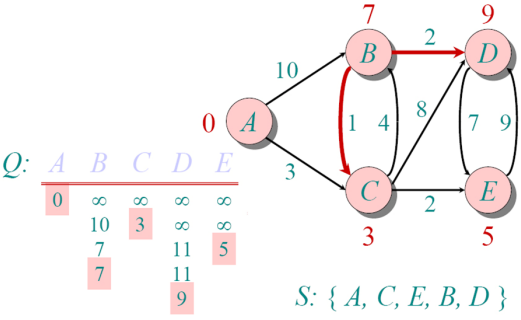
Dijkstra Animated Example



Q:	A	B	C	D	E
	0	∞	∞	∞	∞
		10	3	∞	∞
		7		11	5
		7		11	
				9	

S: { A, C, E, B }

Dijkstra Animated Example



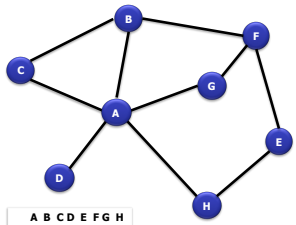
Q:	A	B	C	D	E
	0	∞	∞	∞	∞
		10	3	∞	∞
		7		11	5
		7		11	
				9	

S: { A, C, E, B, D }

Average path length

- Mean path length l_G: average over shortest paths between all pairs of nodes
- Global measure that characterizes the whole network

$$l_G = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j)$$



	A	B	C	D	E	F	G	H
A	0	1	1	1	2	2	1	1
B	1	0	1	2	2	1	2	2
C	1	1	0	2	3	2	2	2
D	1	2	2	0	3	3	2	2
E	2	2	3	3	0	1	2	1
F	2	1	2	3	1	0	1	2
G	1	2	2	2	2	1	0	2
H	1	2	2	2	1	2	2	0

$\Longrightarrow l_G = 1.75$

Closeness centrality

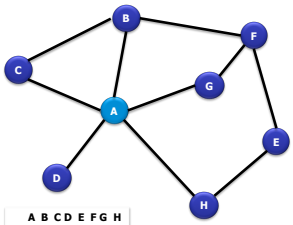
- Inverse of the farness (sum of all distances to a node)
- Closeness centrality:

$$C_C(i) = \frac{1}{\sum_j d(i,j)}$$

- To take into account disconnected components:

$$C_C(i) = \sum_j \frac{1}{d(i,j)}$$

- Local measure



	A	B	C	D	E	F	G	H
A	0	1	1	1	2	2	1	1
B	1	0	1	2	2	1	2	2
C	1	1	0	2	3	2	2	2
D	1	2	2	0	3	3	2	2
E	2	2	3	3	0	1	2	1
F	2	1	2	3	1	0	1	2
G	1	2	2	2	2	1	0	2
H	1	2	2	2	1	2	2	0

$\Rightarrow C_C(A) = 0.11$

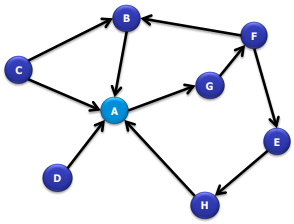
Degree centrality

How do you calculate degree centrality from the adjacency matrix?

- Degree centrality *:* number of links to other nodes:

$$C_D(i) = \deg(i)$$

- Indegree: number of incoming links
- Outdegree: number of outgoing links
- Local measure characterizing each node



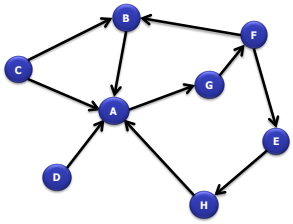
$C_D(A) = 5$
$C_{D,\text{in}}(A) = 4$
$C_{D,\text{out}}(A) = 1$

Average degree

- In undirected networks with N nodes and L links:

$$\bar{C}_D = \frac{2L}{N}$$

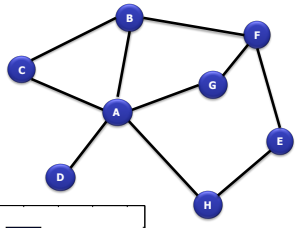
- Characterizes the structure of the whole network

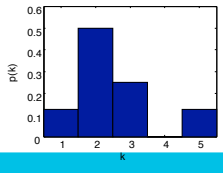


$\bar{C}_D = 2.5$

Degree distribution

- Degree distribution $p(k)$: probability that a selected node has exactly k links
- Count the number of nodes $N(k)$ for each k, and divide by total number of nodes N
- Global measure characterizing the whole network





Clustering coefficient

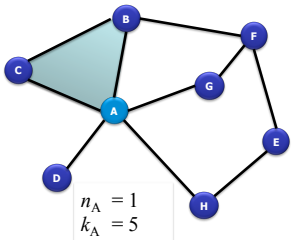
How can this measure be derived? Hint: what is the maximum number of links between neighbors

- Often, if A is linked to B and B to C, then A is also linked to C
- Quantify by a node's clustering coefficient:

$$C_A = \frac{2n_A}{k_A(k_A - 1)}$$

with n_A the number of links between the k_A neighbors of A

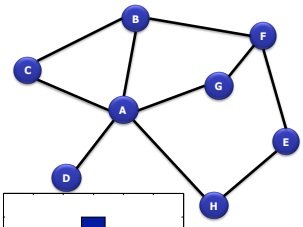
- Characterizes a single node

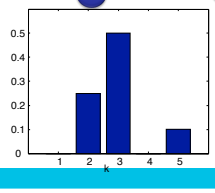


$n_A = 1$
$k_A = 5$
$C_A = 2/20$

Average clustering coefficient

- $\bar{C}$ is the average cluster coefficient: overall tendency to cluster
- $\bar{C}(k)$ is the average cluster coefficient for nodes with degree k
- Characterizes the structure of the whole network





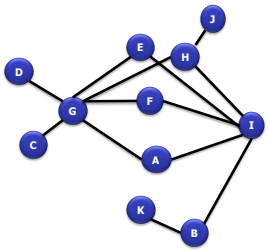
Jaccard index

What is the Jaccard index between G and I?

- Measures the fraction of shared nodes for a pair of nodes

$$J_{ij} = \frac{|k_i \cap k_j|}{|k_i \cup k_j|}$$

- k_i the neighbors of node i



$$J_{GI} = \frac{|\{E, F, A, H\}|}{|\{A, B, C, D, E, F, G, H, I\}|} = \frac{4}{9}$$

- Characterizes the connection of two nodes in the network

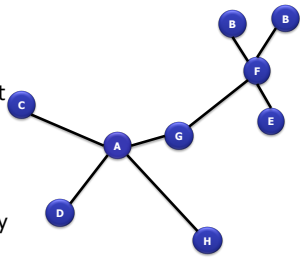
Betweenness centrality

Which node has the highest betweenness?

- Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes

$$C_B(i) = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

- $\sigma_{jk}(i)$ number of shortest paths from j to k through i
- σ_{jk} number of shortest paths from j to k



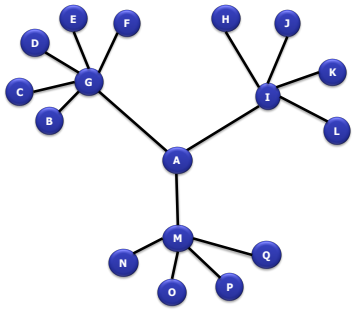
- Characterizes local topology

Illustration (1)

Which node is most central according to the degree?

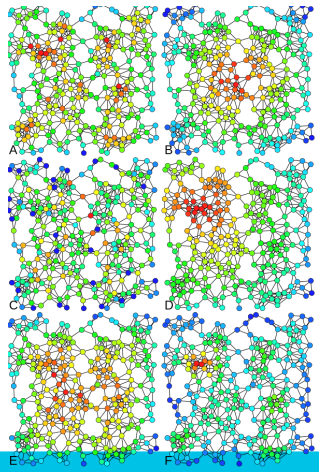
Which node is most central according to the closeness?

Which node is most central according to the betweenness?



Centralities

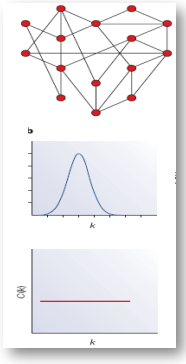
- A – Degree
- B – Closeness
- C – Betweenness
- D – Eigenvector
- E – Katz
- F – Alpha



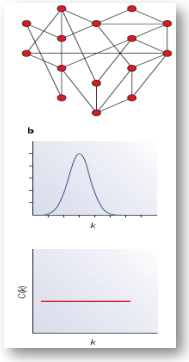
Network models

Random network (1)

- Basic random network:
  - Creation algorithm (Erdös-Rényi):
    - Connect each pair of nodes independently with probability p
    - Creates graph with approximately $pN(N\text{-}1) / 2$ randomly placed links

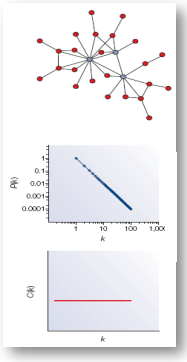


Random network (2)



- Basic random network:
  - Node degree follows a Poisson distribution
  - Each node has approximately the same number of links
  - Tail (nodes with high degree) decreases exponentially
  - Clustering coefficient is independent of a node's degree
  - Mean path length: $l_G \sim \log N$

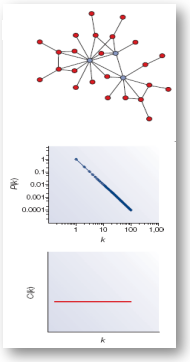
Scale-free network (1)



- Scale free networks:
  - Creation algorithm (Barabási-Albert):
    - Nodes are iteratively added with preferential attachment
    - A node with M links is added to network, connecting to already existing node A with probability $k_A / \sum k_I$

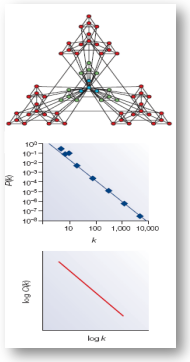
      where k denotes the degree

Scale-free network (2)



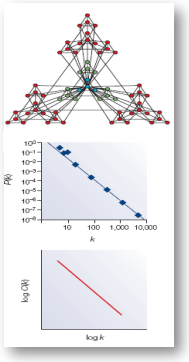
- Scale free networks:
  - Power-law degree distribution
    - ie: log-log, straight line
  - Probability that a node is highly connected is statistically more significant than in E-R networks
  - Network properties determined by relatively small number of highly connected nodes: *hubs*
  - Small-world property
- Ultra-small networks: $2 < \gamma < 3$; mean path length $l_G \sim \log \log N$

Hierarchical network (1)



- Hierarchical networks:
  - Creation algorithm:
    - Start with small cluster of four densely linked nodes (blue)
    - Generate three replicas
    - Connect three external nodes of replicated cluster to central node of old cluster
    - Generate three replicas of each new module
    - etc.

Hierarchical network (2)



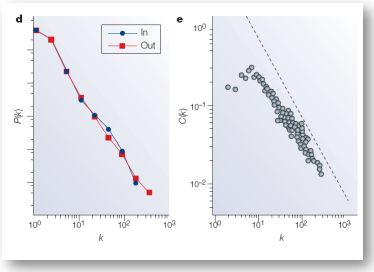
- Hierarchical networks:
  - Combine modularity, local clustering and scale-free topology
  - $\bar{C} \sim 0.6$
  - Most importantly: $C(k) = k^{-1}$
  - Sparsely connected nodes are part of highly clustered areas, maintained by a few hubs

Network topology and biology
Why look at topology?

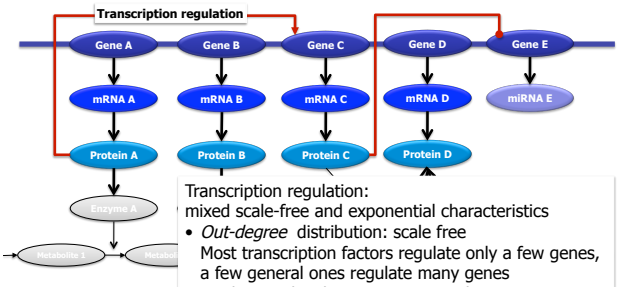
Global topology of E. coli

What is the global topology of this network?

- Example: *Escherichia coli* (*E. coli*) metabolic network



Scale-free networks in biology

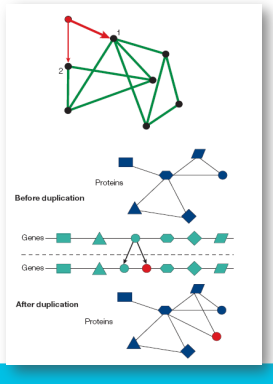


Transcription regulation:
mixed scale-free and exponential characteristics

- *Out-degree* distribution: scale free
  Most transcription factors regulate only a few genes, a few general ones regulate many genes
- *In-degree* distribution: exponential
  Most genes are regulated by 1-3 transcription factors

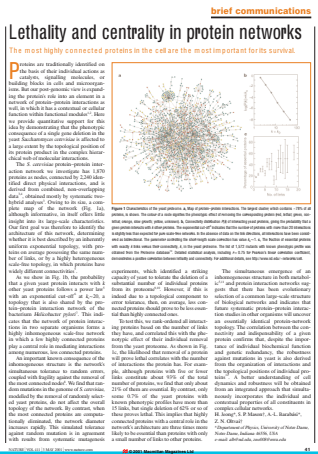
Scale-free networks in biology (2)

- Scale-free behavior may be explained by growth with preferential attachment (new nodes link to highly connected nodes, rich-get-richer)
- Protein interaction networks: possibly caused by gene duplication
  - During cell division, occasionally one or more genes are copied twice: growth of network
  - Copy encodes for same protein and interacts with similar proteins
  - Proteins with large number of interactions gain links more often:
    preferential attachment



Lethality

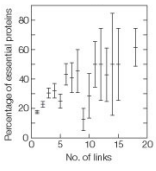
- In 2001 this paper appeared in Nature



Jeong et al. Nature 2001

Lethality and centrality in protein networks

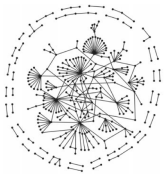
- Phenotypic consequence of a gene deletion is dependent on its local topology in the PPI network (degree)
- Scale-free network is tolerant to random errors but sensitive to removal of 'hubs'
- Conclusions:
  - Topology influences error tolerance
  - High degree proteins are more essential



Jeong et al. Nature 2001

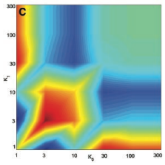
Specificity and stability

- Studies the probability that nodes of certain degrees interact
- High degree nodes tend to interact with low degree nodes
- Hub nodes have reduced likelihood to interact
  - Implies robustness – if one hub is damaged it will not affect other hubs

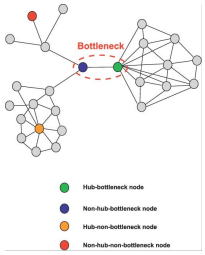


Maslov et al. Science. 2002

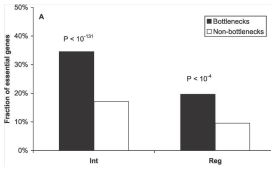
- Nodes with degrees between 4 and 9 tend to interact
  - indicative of protein complex formation



Bottlenecks



- Yu et al. 2007. used betweenness to identify bottleneck proteins
- Showed that bottleneck proteins tend to be more essential
- Showed that bottleneck proteins are significantly less well coexpressed with their neighbors than nonbottlenecks, implying that expression dynamics is wired into the network topology.



Yu et al. 2007. Plos Comp Biol

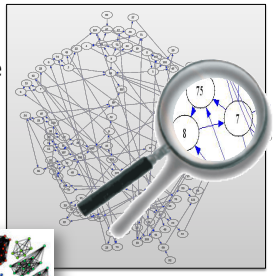
Function emerges from topology?

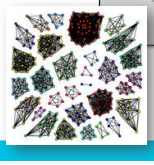
- Words of caution:
  - Coulomb et al. 2005. PNAS found that the relation between essentiality and centrality may be due to bias in the PPI networks
    - Proteins of interest are studied more and may therefore have higher degrees
  - He et al. 2006. PLoS Genetics explained the relation between essentiality and centrality by the simple fact that hubs have large numbers of PPIs, therefore high probabilities of engaging in essential PPIs

Network motifs (1)

- Many networks are too large to study as a whole
- Global properties may emerge from local structure in the network

- Network motifs: significantly recurring small subnetworks



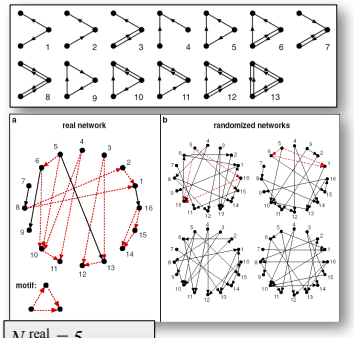


Network motifs (2)

- Enumerate all motifs of k nodes
- Generate a large number of random networks
- For each motif, calculate Z-score

$$Z = \frac{\left(N^{\text{real}} - \langle N^{\text{rand}} \rangle\right)}{\text{std}\left(N^{\text{rand}}\right)}$$

and p-value



$N^{\text{real}} = 5$
$N^{\text{rand}} = 0.5 \pm 0.6$
$Z = 7.5$

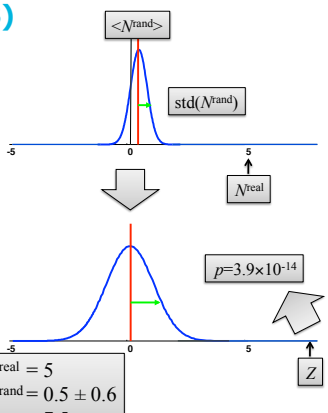
Milo et al. 2002 Science

Network motifs (3)

- Enumerate all motifs of k nodes
- Generate a large number of random networks
- For each motif, calculate Z-score

$$Z = \frac{\left(N^{\text{real}} - \langle N^{\text{rand}} \rangle\right)}{\text{std}\left(N^{\text{rand}}\right)}$$

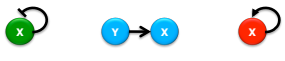
and p-value

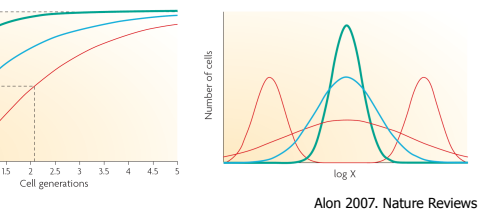


$N^{\text{real}} = 5$
$N^{\text{rand}} = 0.5 \pm 0.6$
$Z = 7.5$

Network motifs (4)

- Small differences in motifs can lead to different behavior

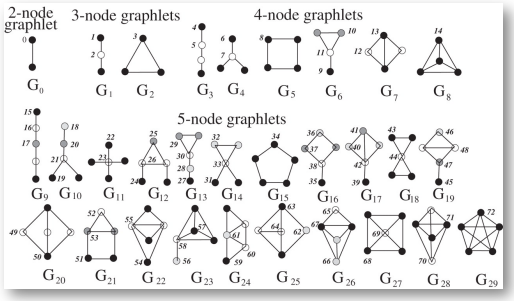




Alon 2007. Nature Reviews

Graphlets (1)

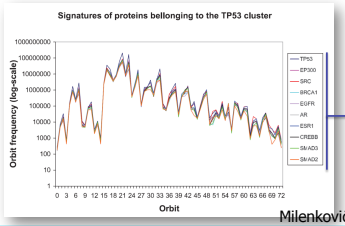
- Related to motifs: non-isomorphic subgraphs



Milenković et al. 2008. Cancer Informatics

Graphlets (2)

- Distribution of number of graphlets touching each node as signature of a network
  - Use to find clusters in PPI networks
  - Signatures predict function?



Cluster of known cancer genes

Milenković et al. 2008. Cancer Informatics

Wrap up

Global vs local

- Many local measures exist that
  - characterize the topology surrounding one node
  - characterize the connection between two nodes
- Some local measures can be extended to global measures
  - Degree distribution
  - Clustering coefficient distribution
  - Average path length
  - Counts on network motifs

Descriptive, suggestive, predictive

- Global topological characteristics provide interesting insights in biological networks, but yield no new hypotheses or support conclusions
  - Descriptive
- Local analysis of network structure, such as clustering or topological measures for each node, yields insight in function
  - Suggestive
- Using local topology to derive signatures can link function to topology
  - Predictive

Winterbach et al. 2013 BMC Systems Biology