

Analysis of Complex Systems

Lecture 4: Local network organisation:
Clusters, motifs, Jaccard index, betweenness
centrality

Marcus Kaiser
m.kaiser@ncl.ac.uk

Objectives

- Clusters
- Motifs
- Jaccard index (matching index)
- Edge and node betweenness

Clusters

Clusters

Clusters: nodes within a cluster tend to connect to nodes in the same cluster but are less likely to connect to nodes in other clusters

Quantitative measure: modularity Q
(Newman & Girvan, Physical Review E, 2004)

important terms:

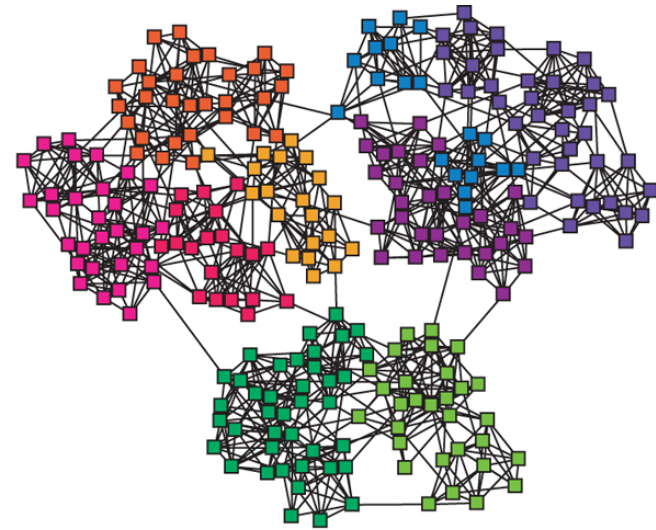
hierarchical (cluster, sub-cluster, ...)

overlapping or *non-overlapping*

(one node can only be member of one cluster)

predefined number of clusters

(e.g. k-means algorithm)



Potential time problem for large networks, $O(k^N)$

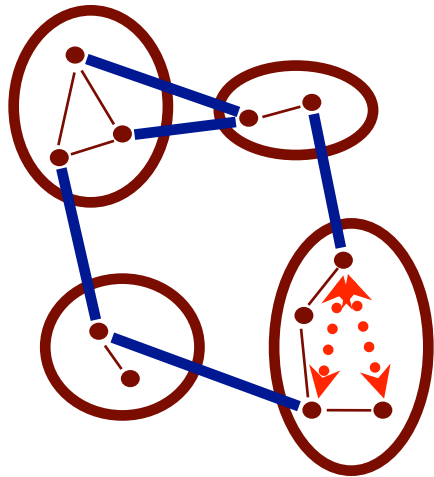
Hundreds of algorithms for cluster detection!



Cluster detection – algorithm 1

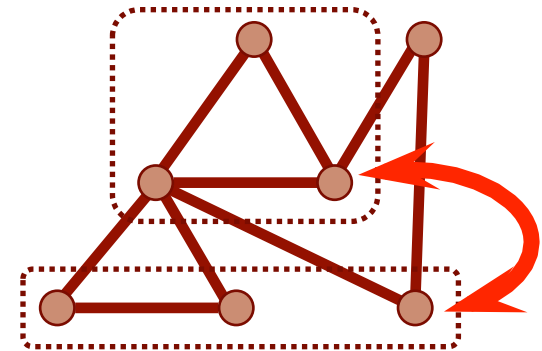
Non-hierarchical, overlapping

Genetic algorithm



- Have as few as possible connections **between** them
- Have as few as possible absent connections **within** them

Procedure



- Random starting configurations
- Evolution:
 - Mutation : Area relocation
 - Evaluation : Cost function
 - Selection : Threshold
- Validation

Hilgetag et al. (2000) Phil. Trans. Roy. Soc. Lond. B.

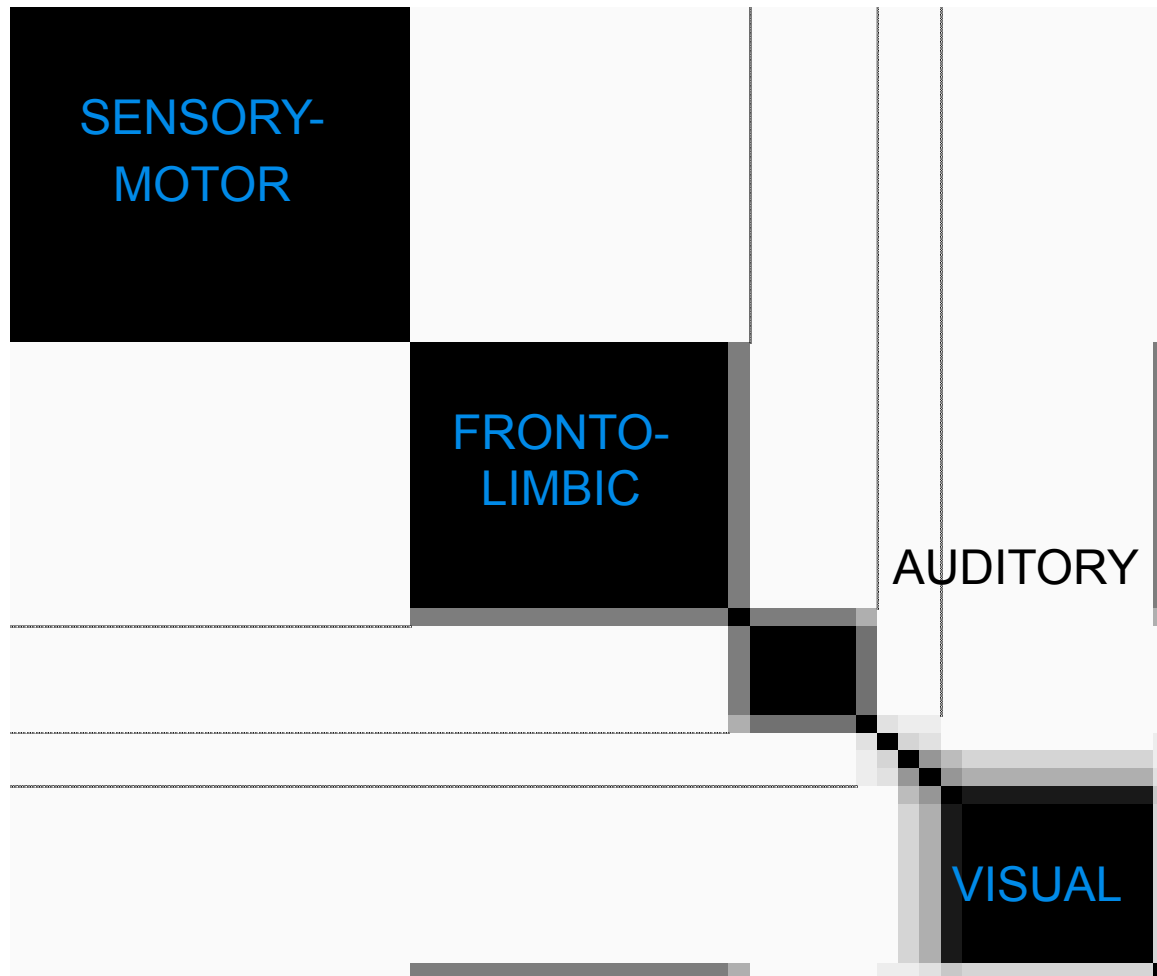
Cluster detection – algorithm 1

Example:

Cat cortical network

Black:
same cluster

Graylevel:
Ambiguous cases

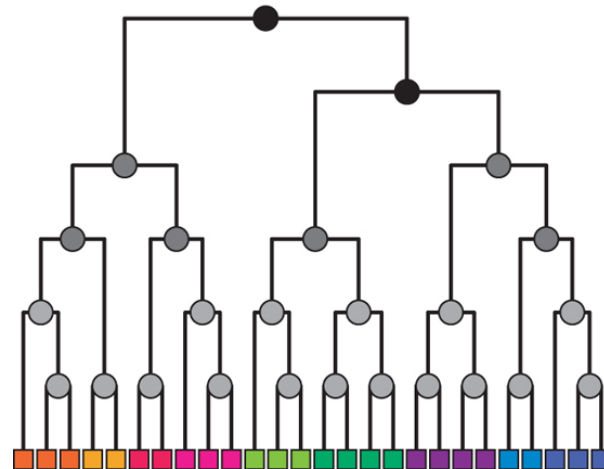
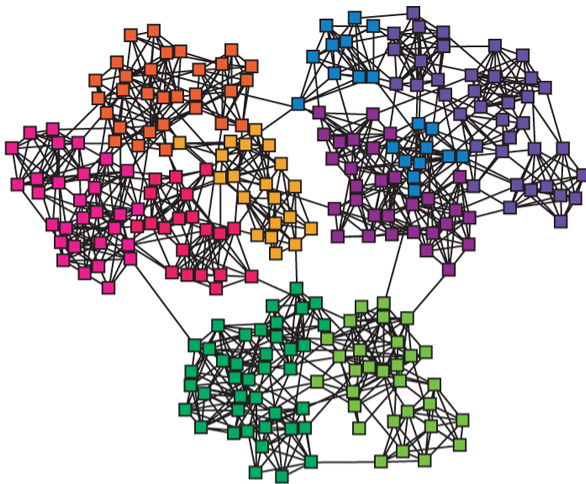


Cluster detection – algorithm 2

hierarchical, non-overlapping

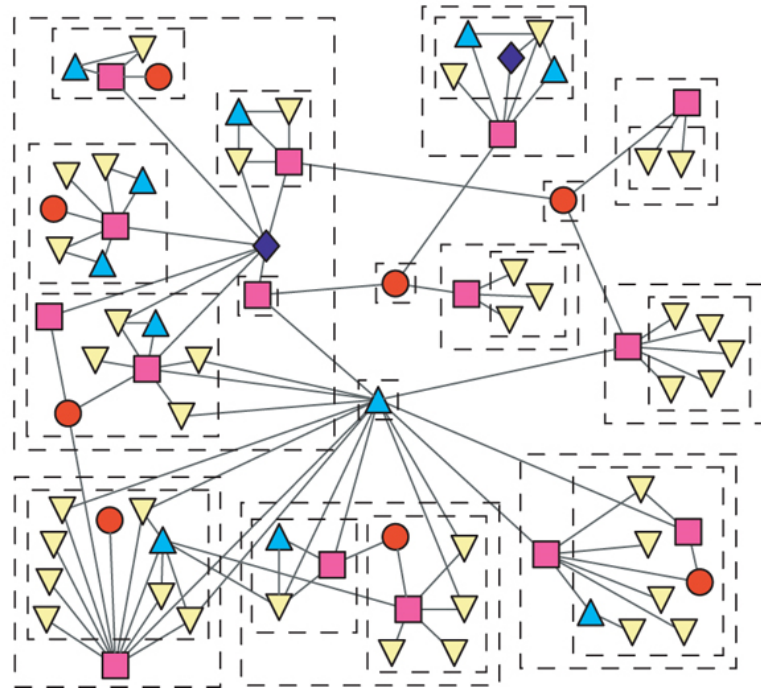
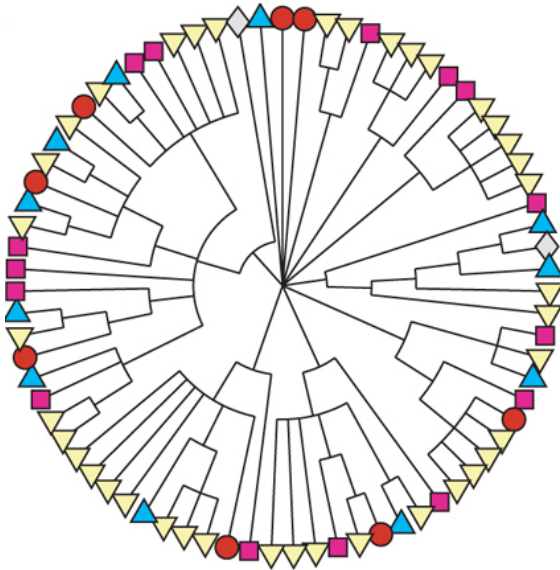
Monte-Carlo approach

Each internal node r of the *dendrogram* is associated with a probability p_r that a pair of vertices in the left and right subtrees of that node are connected. (The shades of the internal nodes in the figure represent the probabilities.)



Cluster detection – algorithm 2

Example result for a food web



Clauset et al. (2008). Nature 453, 98-101

Motifs

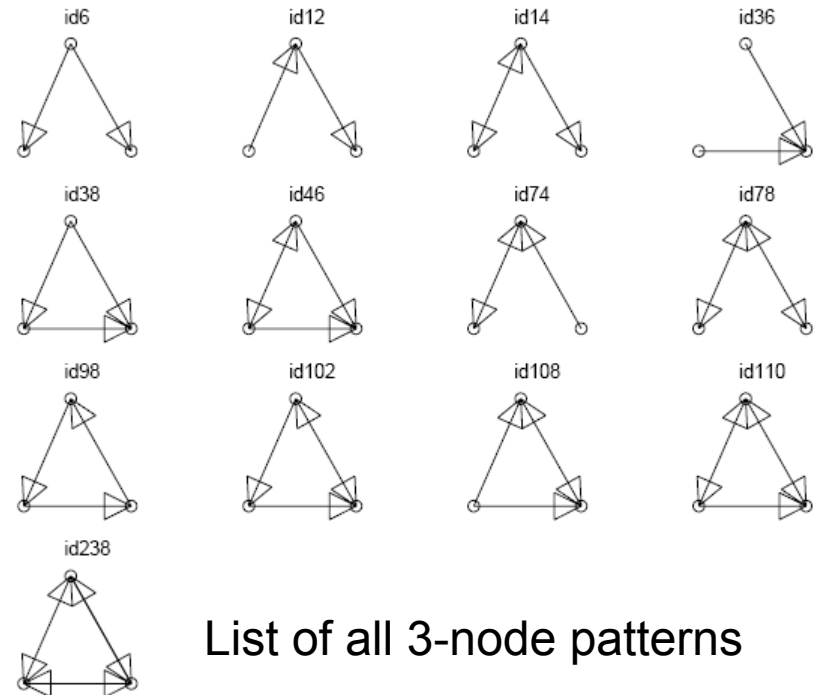
Motifs

Idea: determine building blocks of networks.

Hope: structural building blocks correspond to functional units.

Pattern: possible connection configuration for a k-node subgraph
(see list of all 3-node configurations)

Motif: pattern that occurs significantly more often than for rewired benchmark networks
(same number of nodes and edges and same degree distribution)



List of all 3-node patterns

* Milo et al. (2002) Science;
<http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html>

Motif detection – algorithm

Network name: network_exmp.txt
 Network type: Directed
 Num of Nodes: 16 Num of Edges: 19
 Num of Nodes with edges: 16
 Maximal out degree (out-hub) : 3
 Maximal in degree (in-hub) : 3
 Roots num: 4 Leaves num: 4
 Single Edges num: 19 Mutual Edges num: 0

Motif size searched 3
 Total number of 3-node subgraphs : 21
 Number of random networks generated : 100
 Random networks generation method: Switches
 Num of Switches range: 100.0-200.0,
 Success switches Ratio:0.652+0.01

The following motifs were found:

Criteria taken : Nreal Zscore > 2.00
 Pval ignored (due to small number of random networks)

Mfactor > 1.10
 Uniqueness >= 4

Appearances
in the real
network

Random
networks:
mean+- SD

Uniqueness

Concentration
X10⁻³

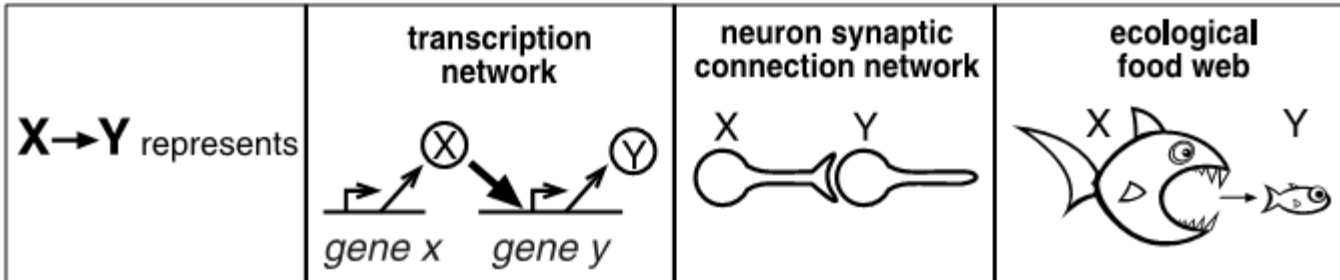
Full list includes 1 motifs

MOTIF ID	NREAL	NRAND STATS	NREAL ZSCORE	NREAL PVAL	UNIQ VAL	CREAL [MILI]
38	5	0.6+0.6	6.93	0.000	4	238.10

0 1 1
 0 0 1
 0 0 0

Motif
Adjacency
Matrix

Motif detection – results



Network	Nodes	Edges	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score
Gene regulation (transcription)			Feed-forward loop			Bi-fan					
<i>E. coli</i>	424	519	40	7 ± 3	10	203	47 ± 12	13			
<i>S. cerevisiae</i> *	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons			Feed-forward loop			Bi-fan			Bi-parallel		
<i>C. elegans</i> †	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20

Motif detection – problems

Advantages:

- Identify special network patterns which *might* represent functional modules

Disadvantages:

- Slow for large networks and
unfeasible for large (e.g. 5-node) motifs
(#patterns: 3-node – 13; 4-node – 199; 5-node: 9364; 6-node - 1,530,843)
- Rewired benchmark networks do not retain *clusters*;
most patterns become insignificant for clustered benchmark networks*

*Kaiser, unpublished

Jaccard index (matching index)

Jaccard index

- Jaccard index = similarity of incoming and outgoing connections of two nodes

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$|M|$: number of elements in the set M

$A \cap B$: common elements in sets A and B (intersection)

$A \cup B$: all elements in sets A and B (union)

A=0 1 0 0 0 0 1 0 1

B=0 1 0 1 0 0 1 0 0

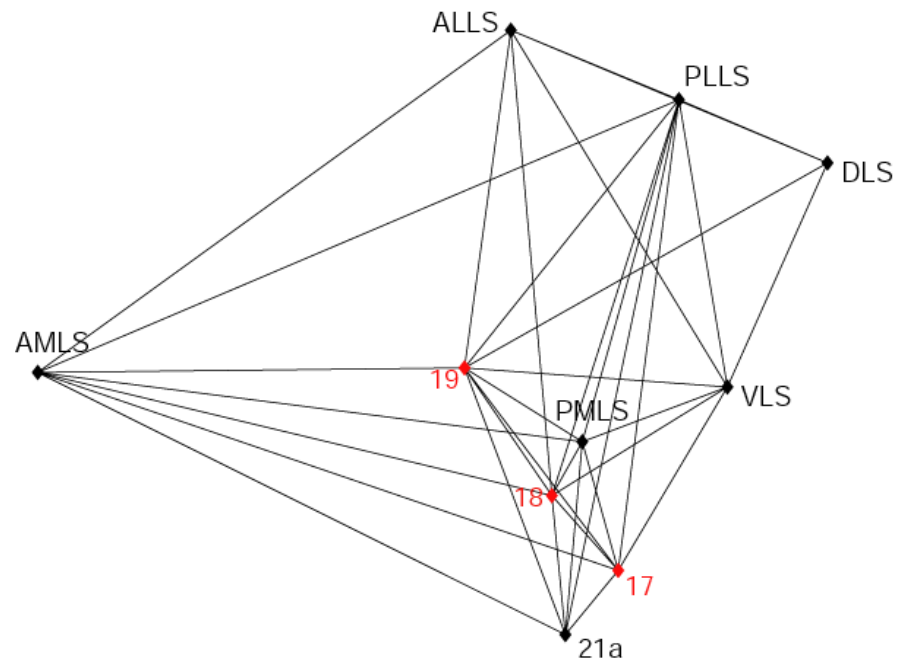
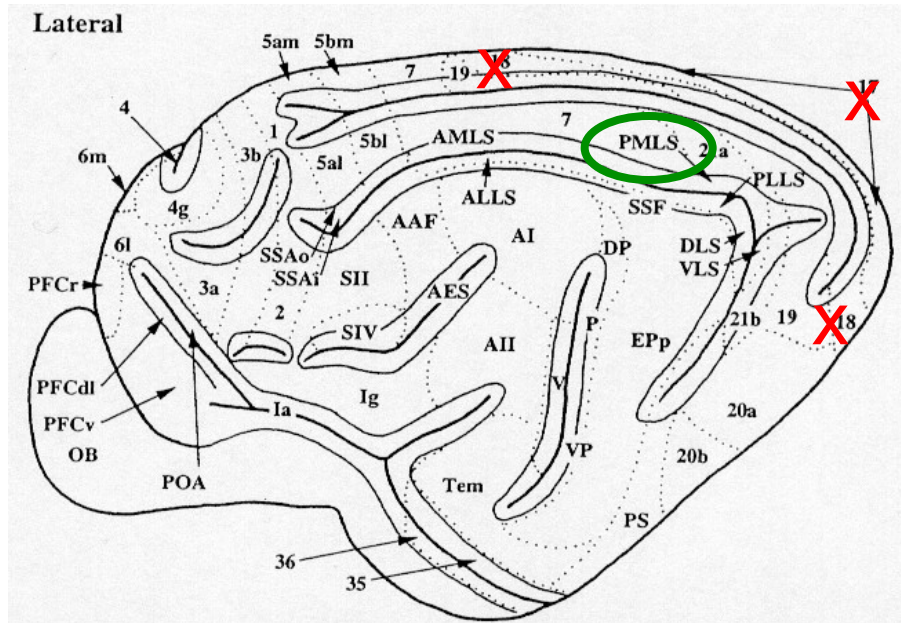
$$|A \cap B| = 2$$

$$|A \cup B| = 4$$

$$J(A, B) = 2 / 4 = 0.5$$

Similarity and compensation

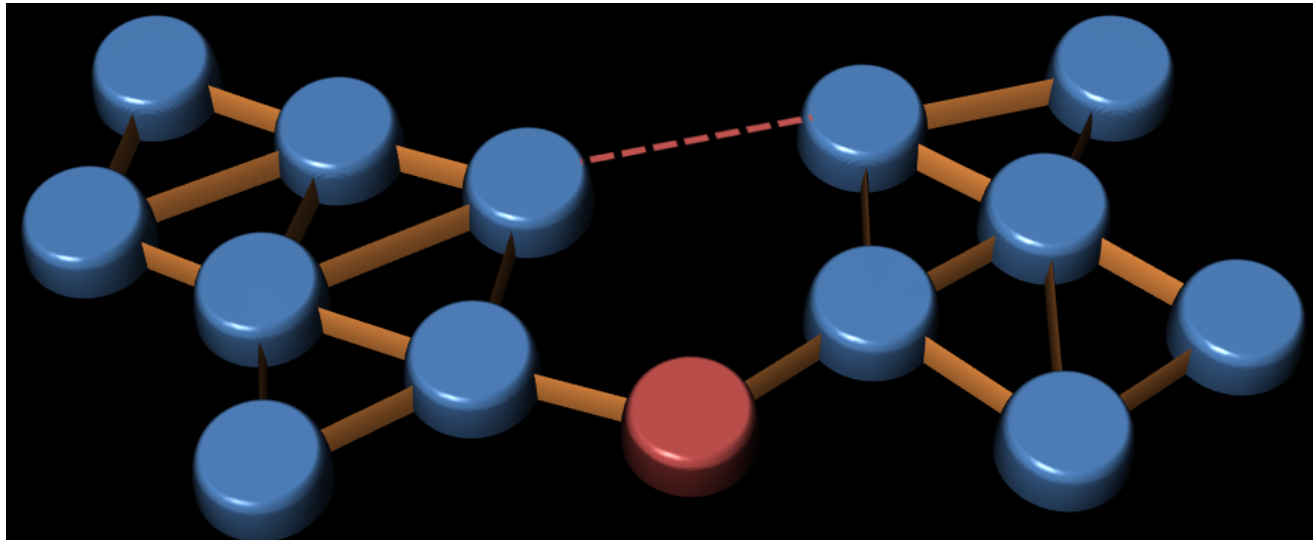
- Use of non-metric multi-dimensional scaling (NMDS)
- Idea: similar connections -> similar function



Centrality measures

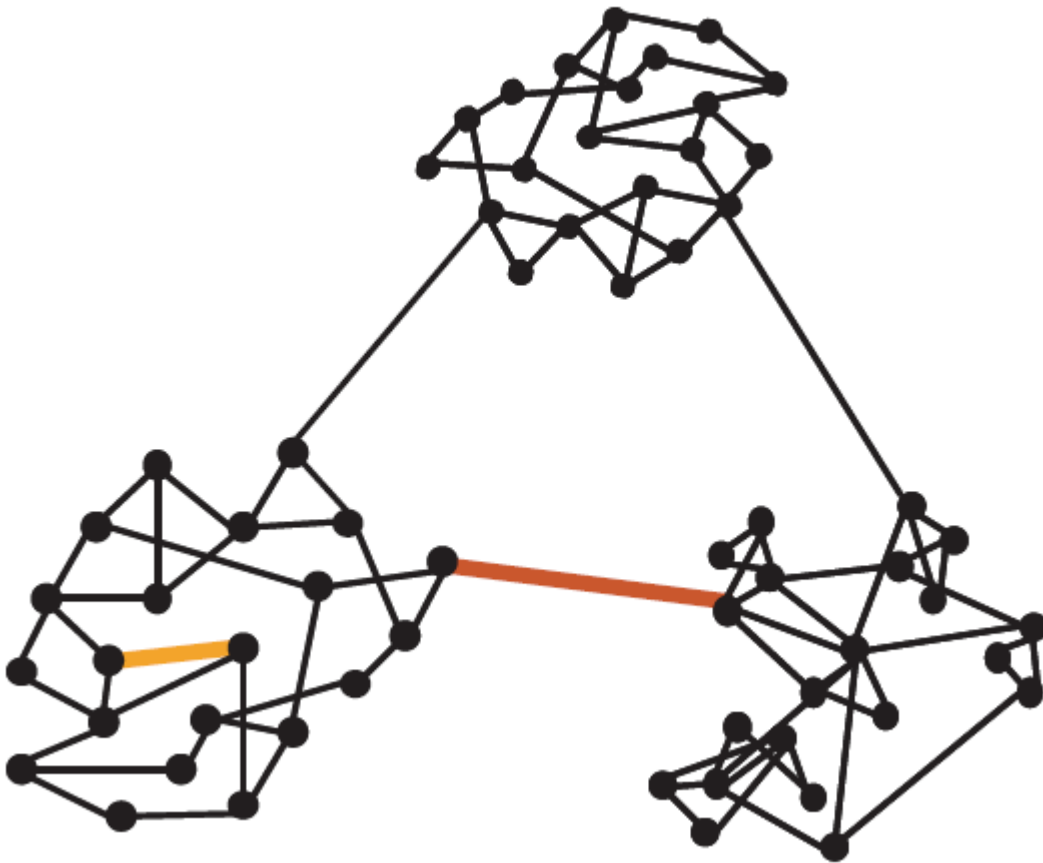
Node betweenness

- Node betweenness:
number of shortest paths that go through one node



Edge betweenness

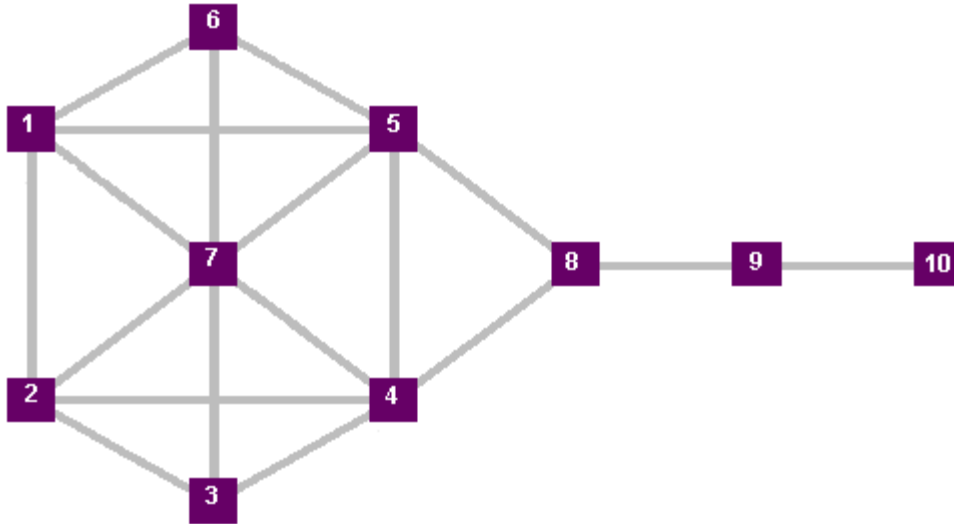
- Edge betweenness:
number of shortest paths that go through one edge



**High edge
betweenness**

**Low edge
betweenness**

Centrality measure example



Node 8 has the highest node betweenness
Edge 8-9 has the highest edge betweenness

Summary

- What are clusters? Which kinds of clustering algorithms exist?
- What are motifs? Which features of the original network are retained for the benchmark networks and which are lost?
- What is the Jaccard index? What could similar connectivity indicate?
- What are betweenness measures?

Q&A – 1

1. A cluster algorithm is hierarchical, non-overlapping, and has a predetermined number of clusters. Which of these features is the most devastating one for biological network analysis?
2. The Jaccard index for two proteins is very high. Would you expect the nodes to be in the same or in different protein clusters?
3. You have identified the edge with the highest edge betweenness. How would network properties change if that edge were removed? Which measure would be most affected?