Advanced Techniques for Mining Structured Data:

Graph Mining

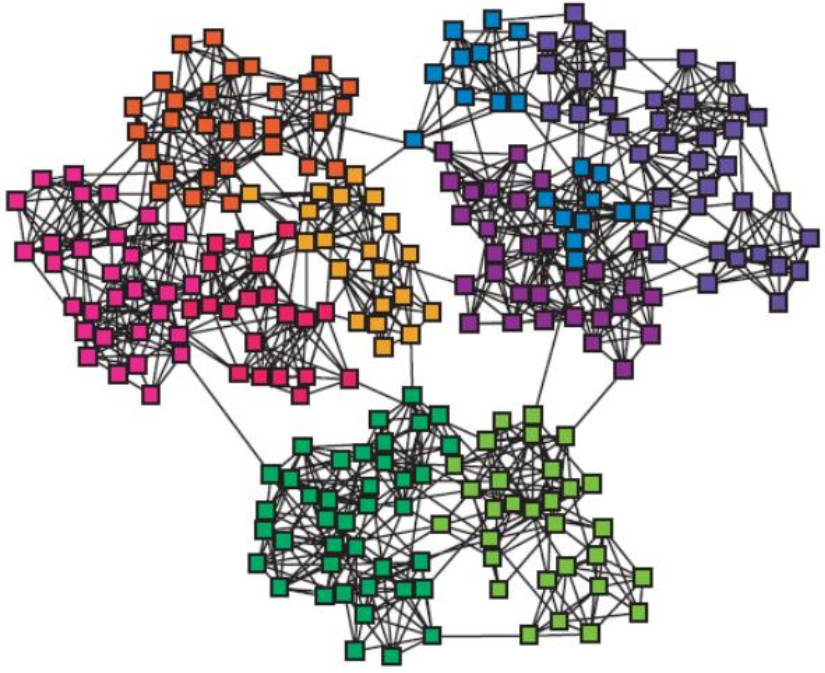
Node Clustering

Dr C.Loglisci

PhD Course in Computer Science and Mathematics XXXII cycle

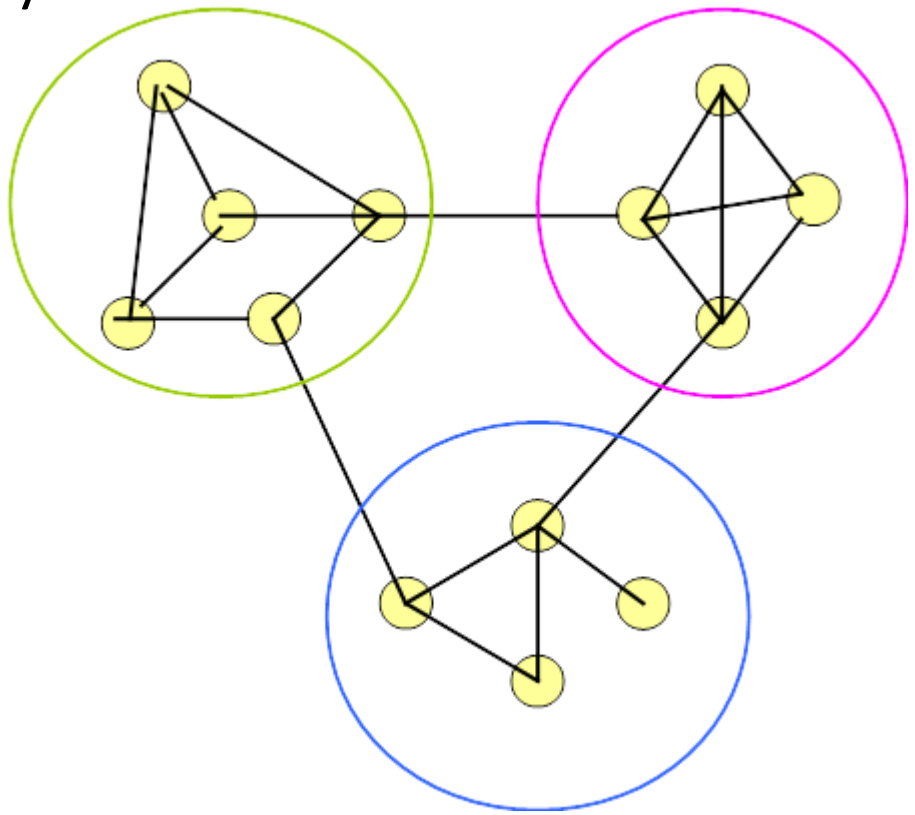
Networks and Communities

- We often think of networks being organized into modules, cluster, communities:…
- …..and we aim at finding densely linked communities/clusters
- Examples are
  - communities of biochemical network correspond to functional units of some kind.
  - communities of a web graph correspond to sets of web sites dealing with a related topics.



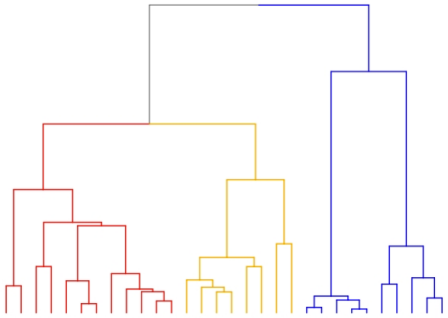
Networks and Communities

- These can be referred to groups of vertices within which connections are dense but between which they are sparser:
  - within-group(intra-group) edges, High density
  - between-group(inter-group) edges, Low density.

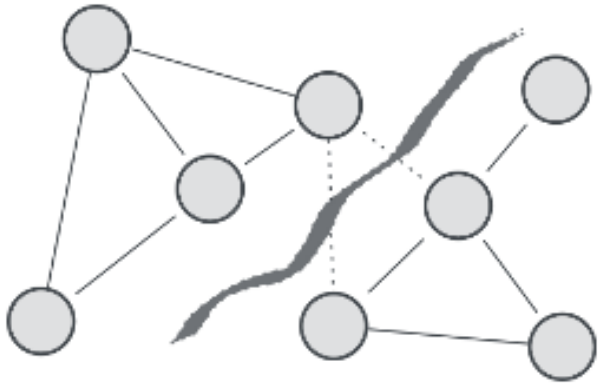


State-of-Art Approaches

- Hierarchical approaches
  - divisive (partitioning)
  - agglomerative



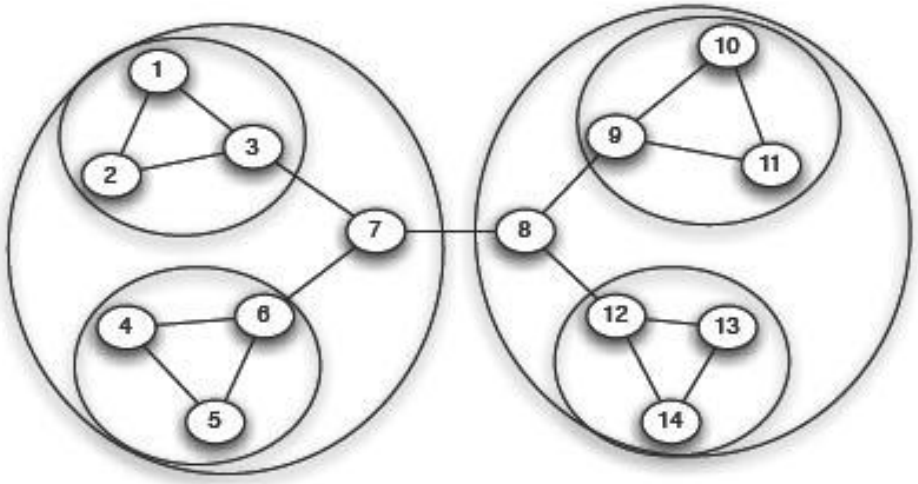
- Spectral clustering



We will work with undirected (unweighted) networks

State-of-Art Approaches

- Hierarchical approaches
  - divisive (partitioning)
  - agglomerative

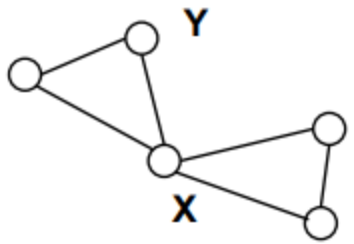


  - it doesn't require us to specify the size or number of groups
  - It doesn’t give indication to get the best partitioning of the network

Girvan-Newman algorithm

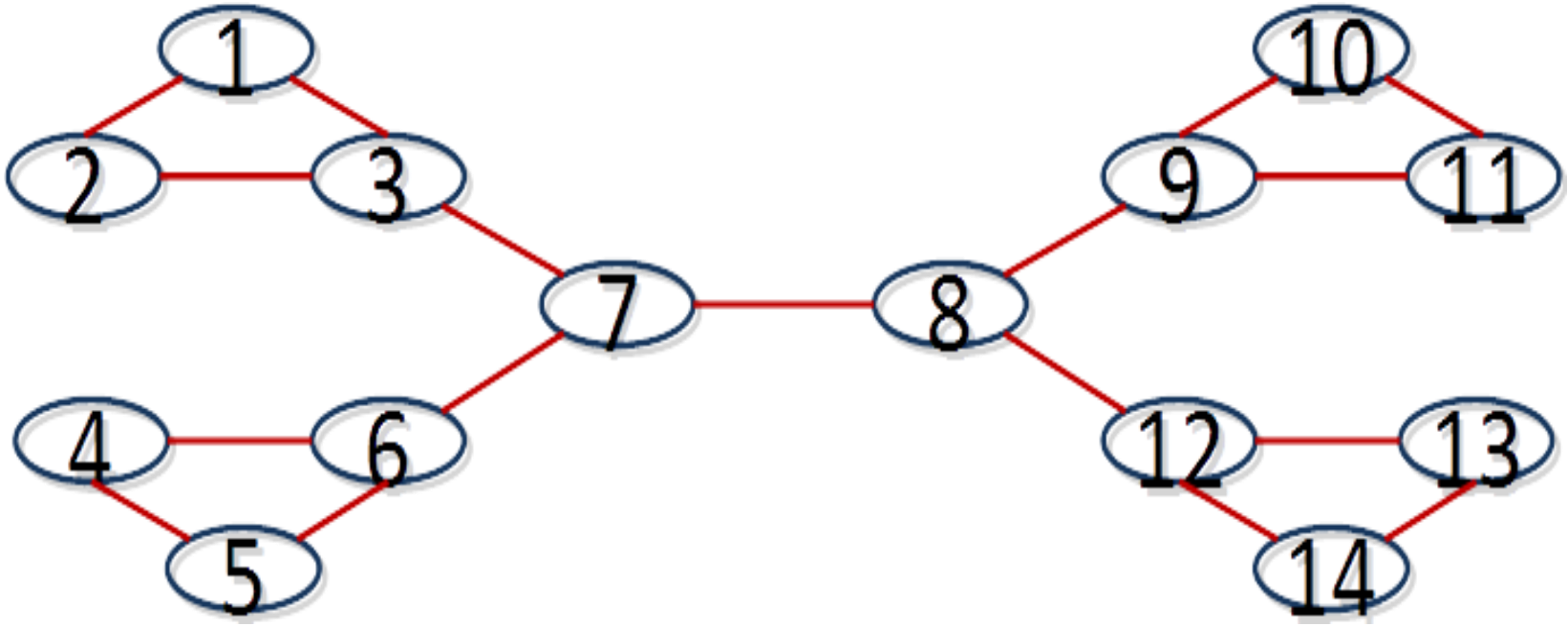
- Basic idea:
  - remove bridges, that is, partition by strongest ties, which may connect different communities
  - edges betweenness, number of shortest paths passing over the edge.
  - recall vertex betweenness:

$$c(v_i) = \sum_{j \neq i} \sum_{k \neq i, k > j} \frac{\eta_{jk}(v_i)}{\eta_{jk}}$$



Girvan-Newman algorithm

- Basic idea:
  - edges betweenness, number of shortest paths passing over the edge.
  - Total amount of “flow” an edge carries between all pairs of nodes

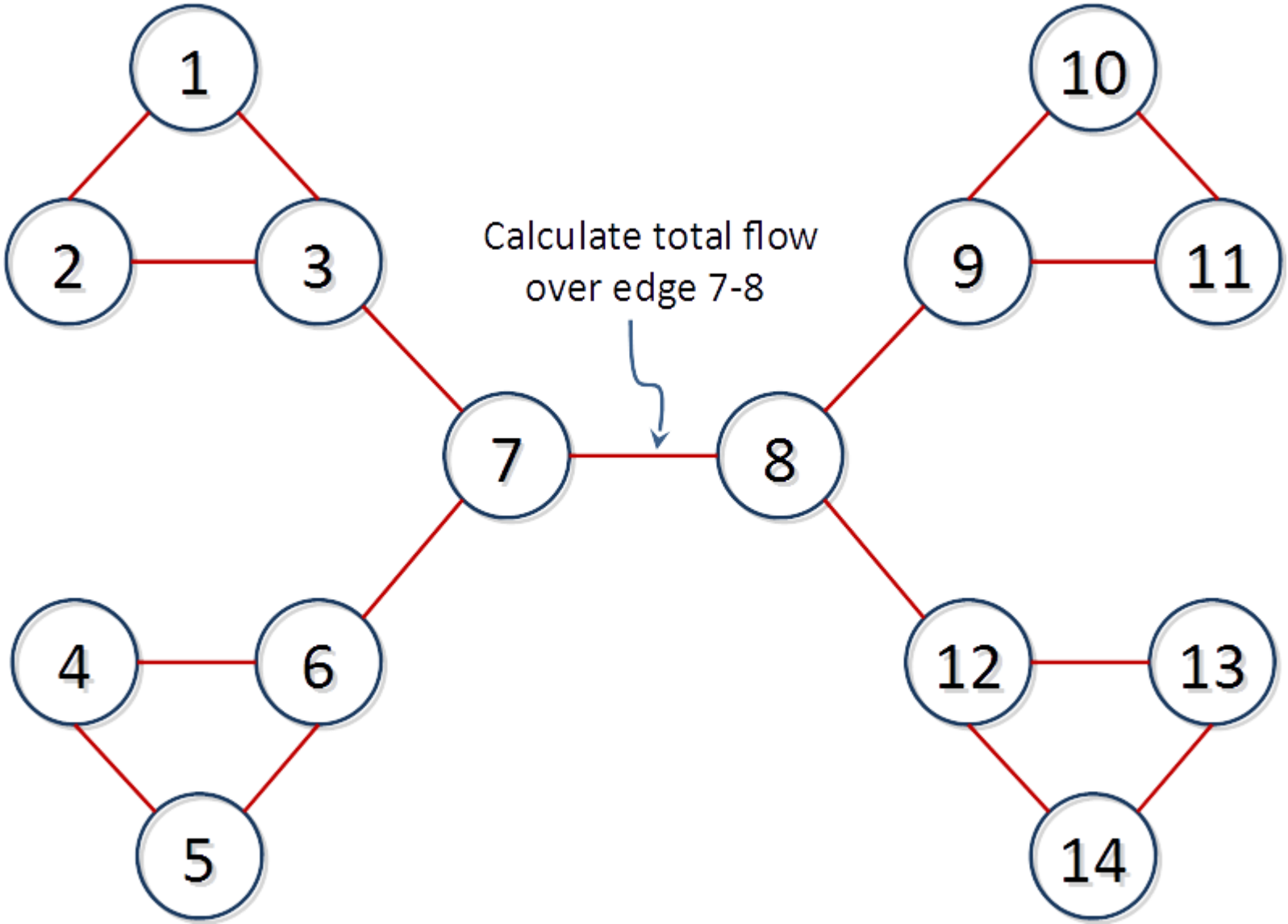


Girvan-Newman algorithm

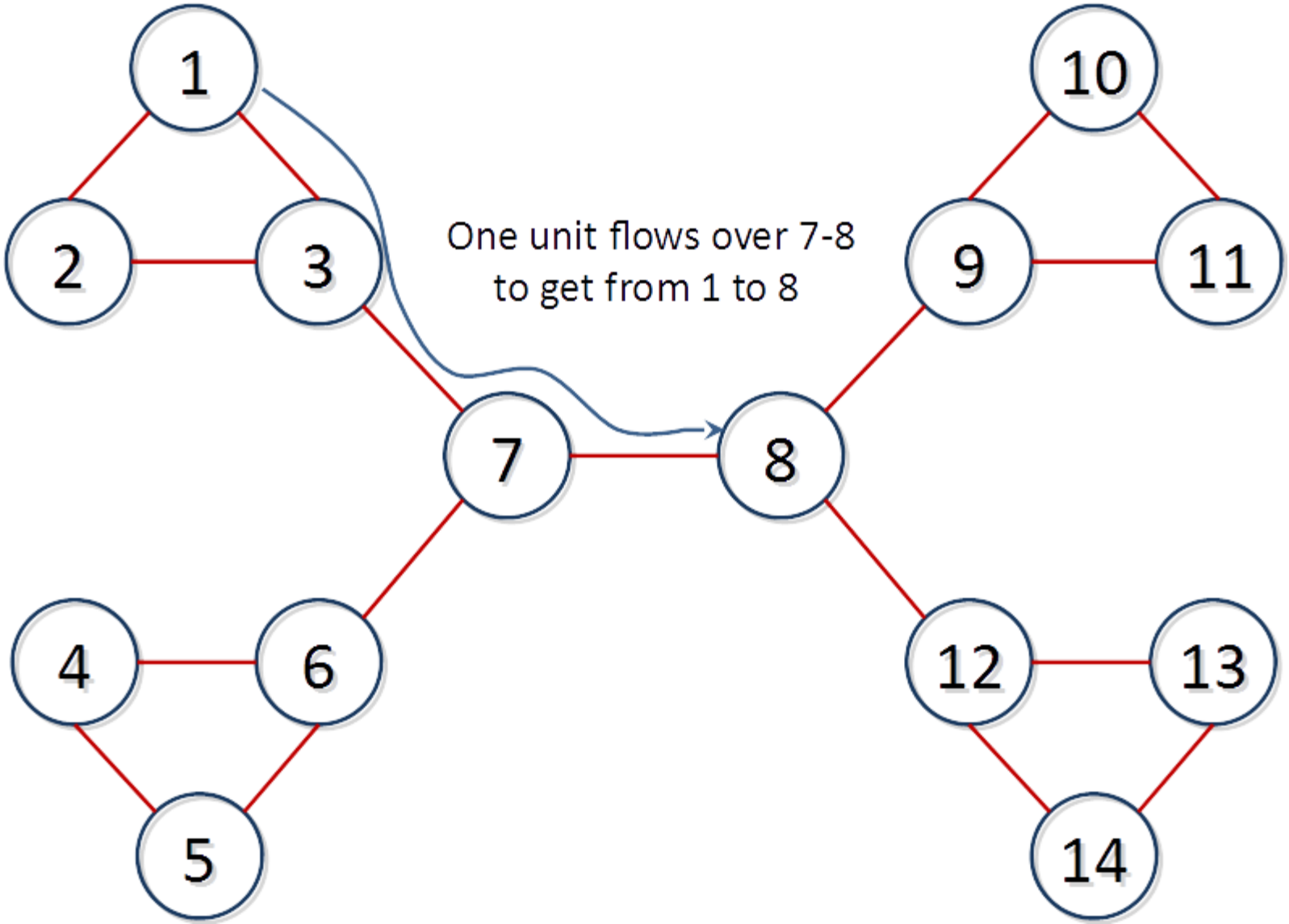
- Procedure
  1. Calculate the betweenness for all edges in the network.
  2. Remove the edge with the highest betweenness.
  3. Recalculate betweennesses for all edges affected by the removal.
  4. Repeat from step 2 until no edges remain.

  5. cross cut the dendogram of components.
  6. by removing these edges, we separate communities from one another as components.

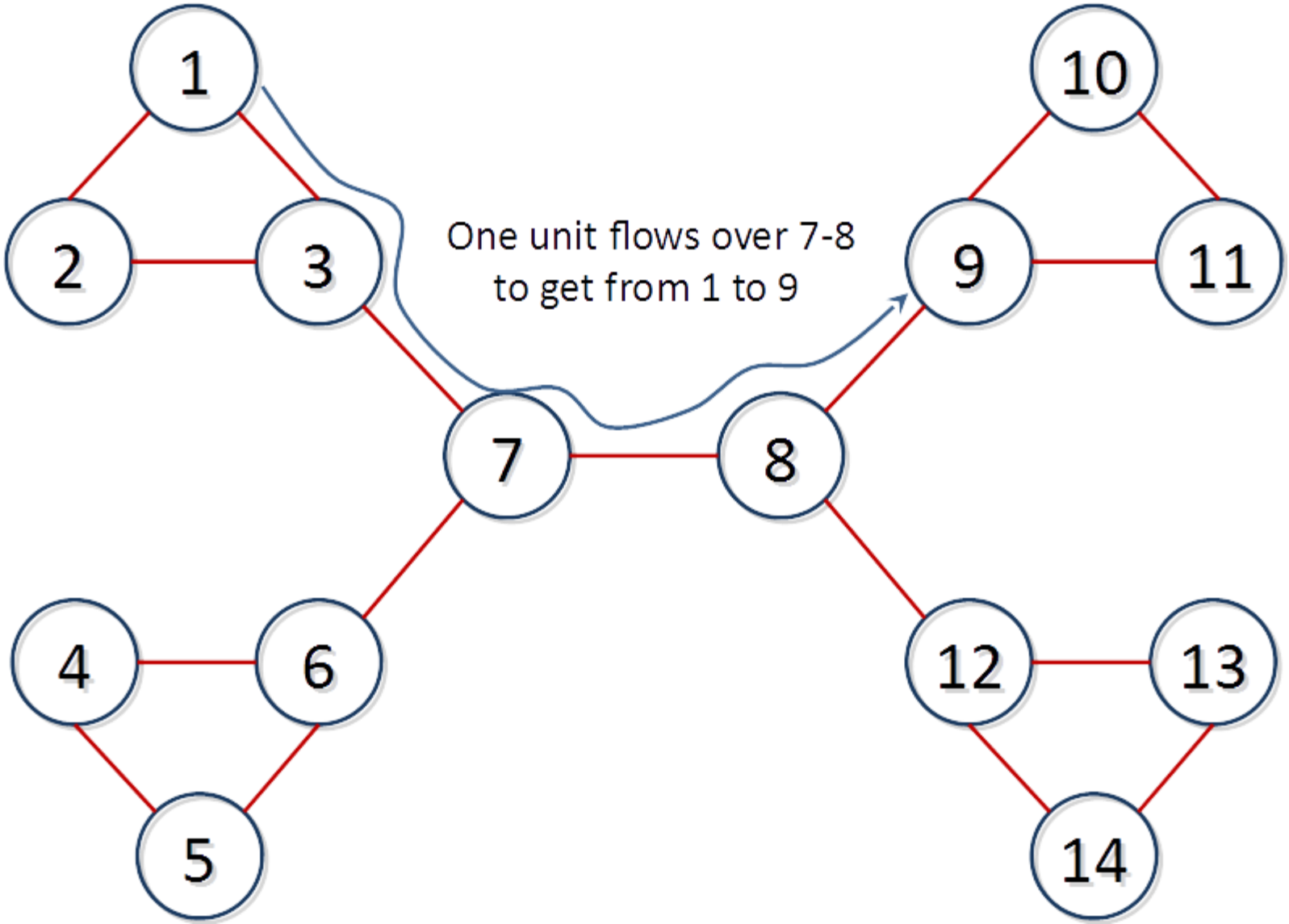
Girvan-Newman algorithm



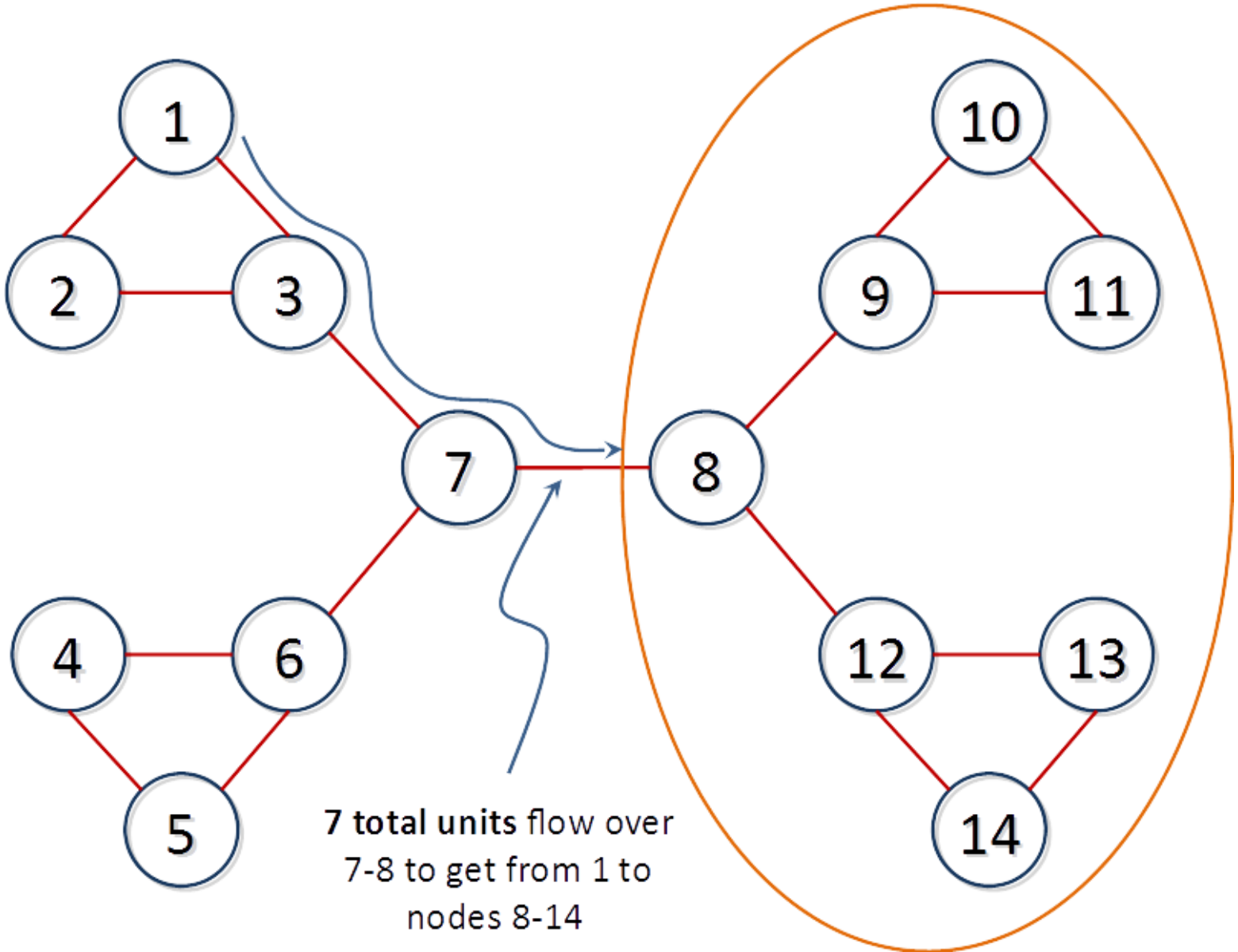
Girvan-Newman algorithm



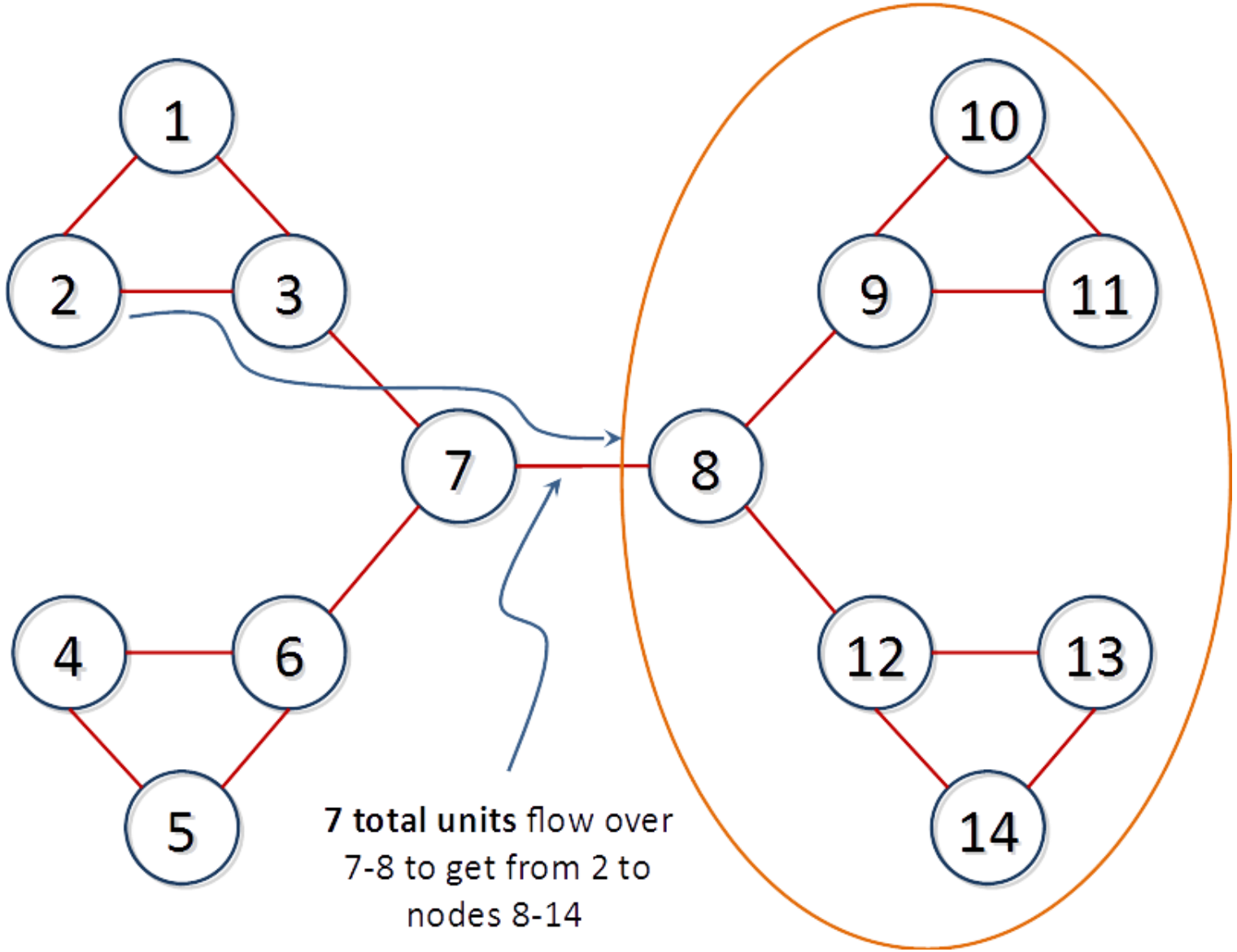
Girvan-Newman algorithm



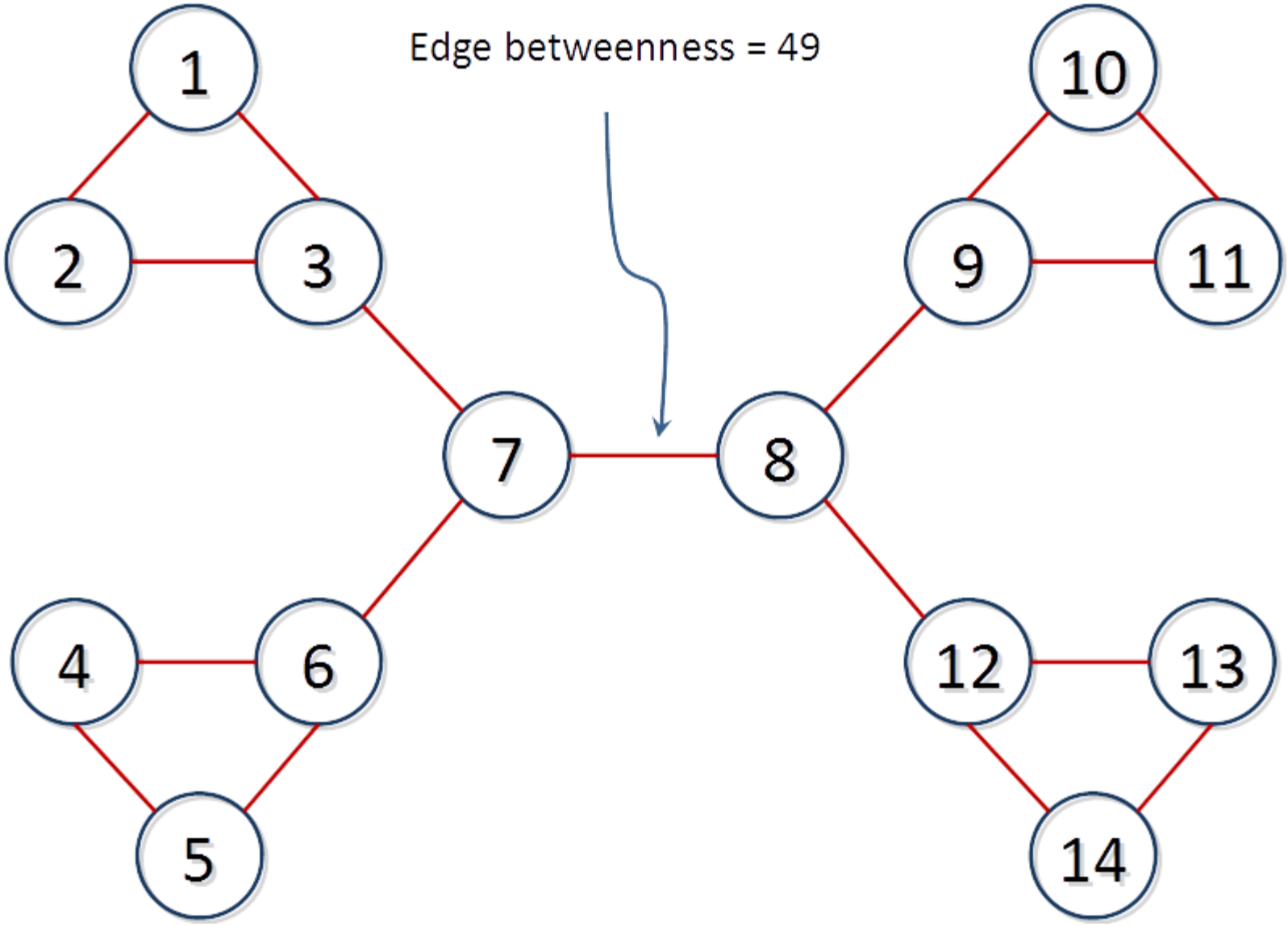
Girvan-Newman algorithm



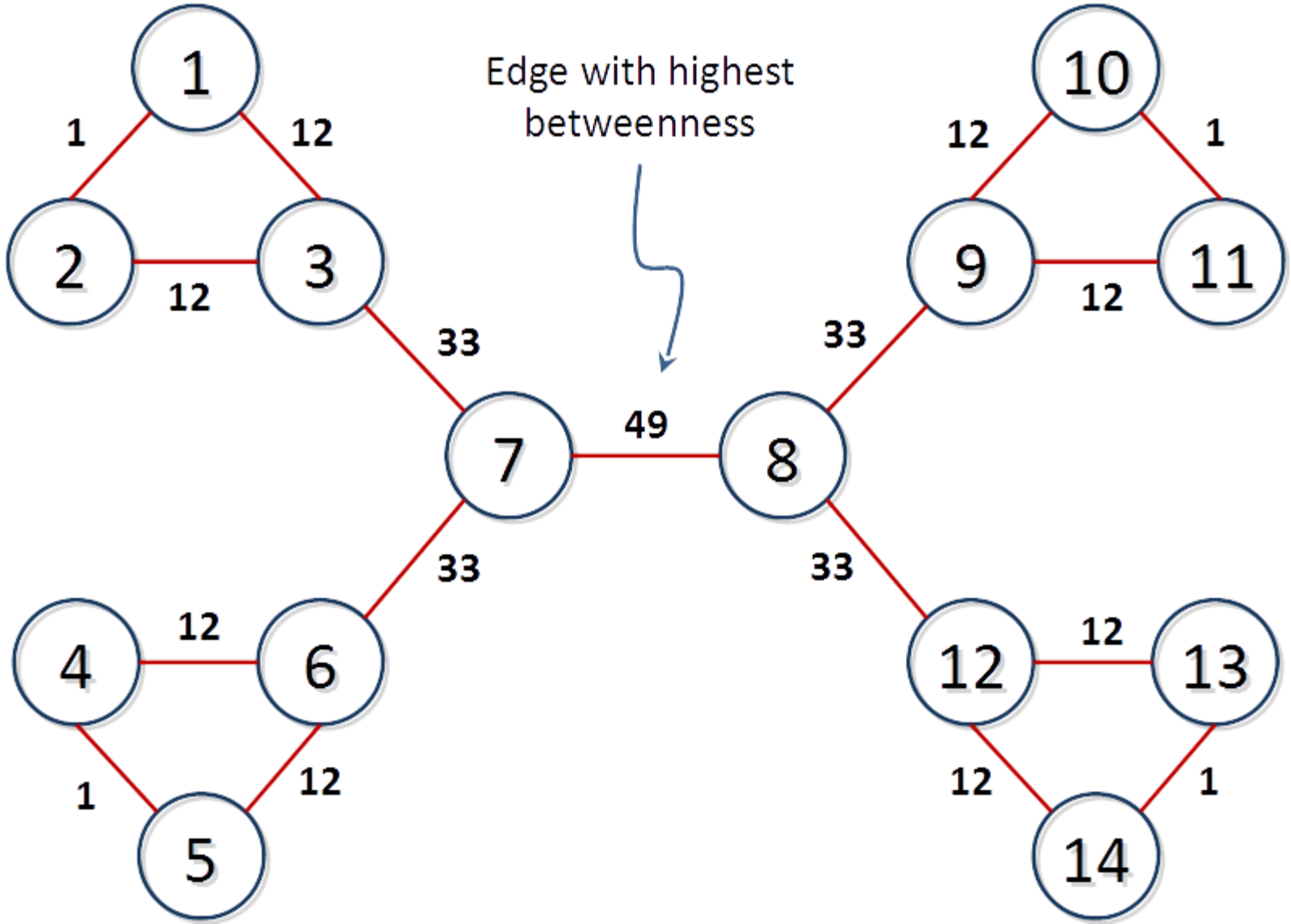
Girvan-Newman algorithm



Girvan-Newman algorithm

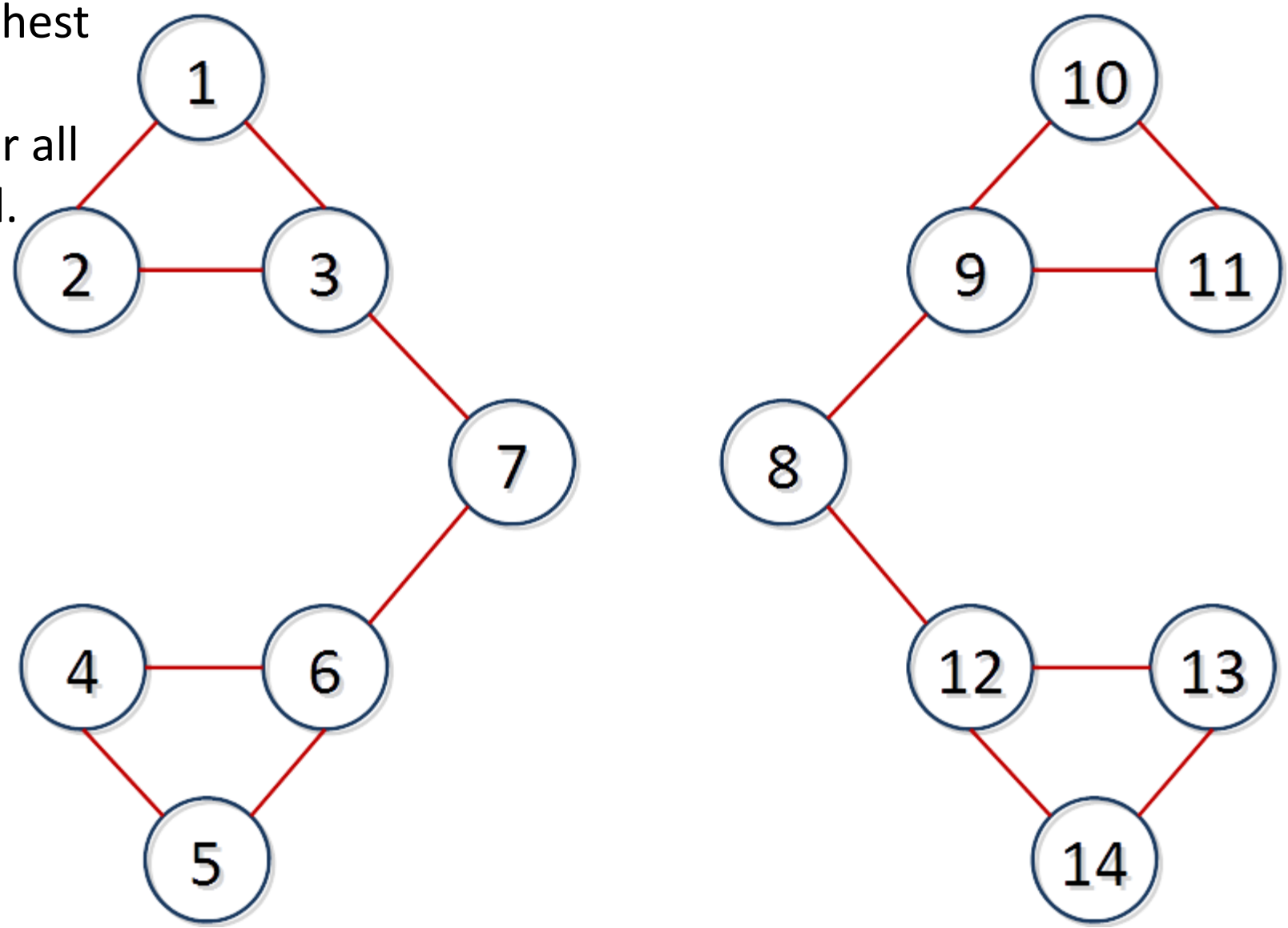


Girvan-Newman algorithm



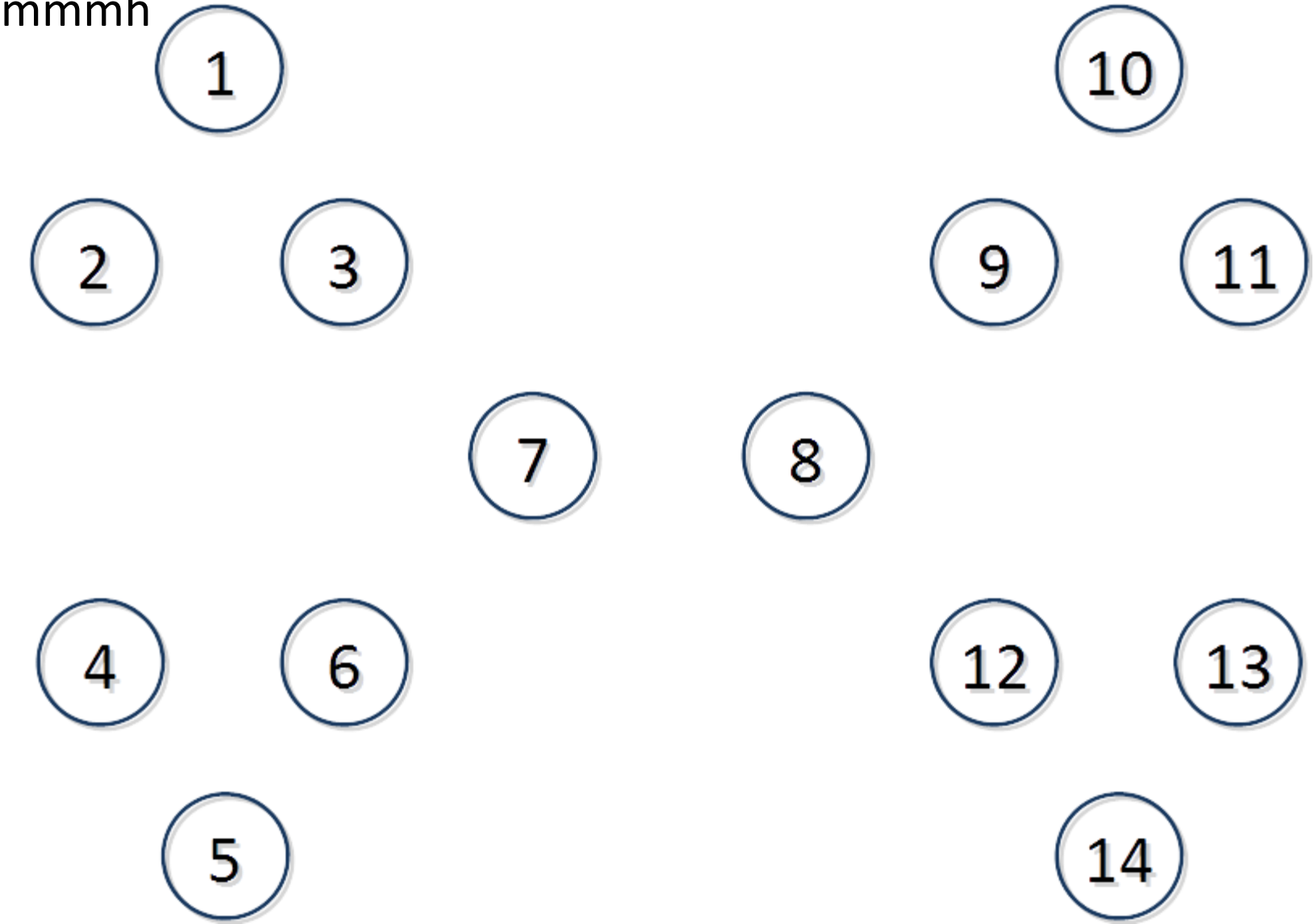
Girvan-Newman algorithm

- Remove the edge with the highest betweenness.
- Recalculate betweennesses for all edges affected by the removal.



Girvan-Newman algorithm

- until to have: …mmmh

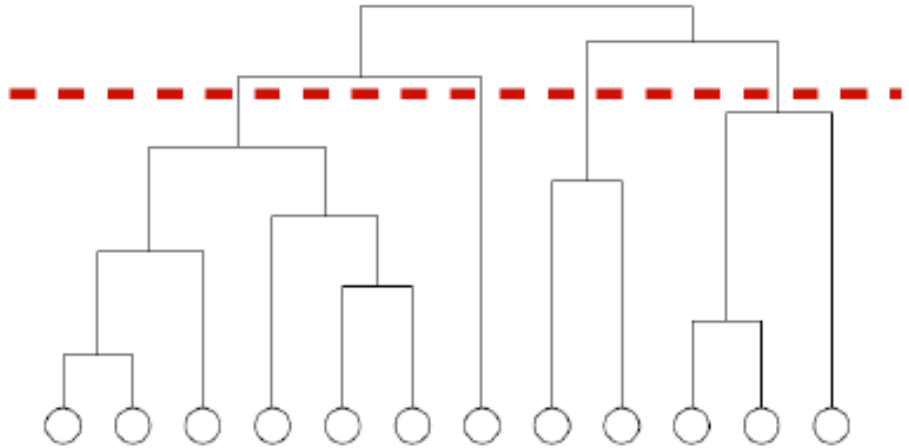


Girvan-Newman algorithm

- How to select the best partition of communities?
- Modularity: a measure of how well a network is partitioned into communities

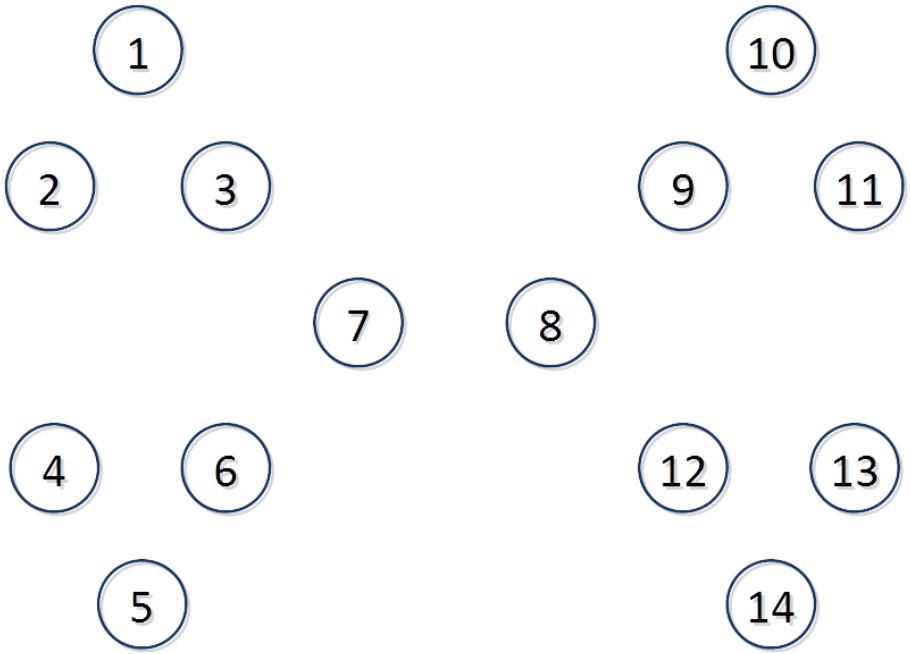
  $Q \sim \sum$(#edges within community s – # expected edges within community s),

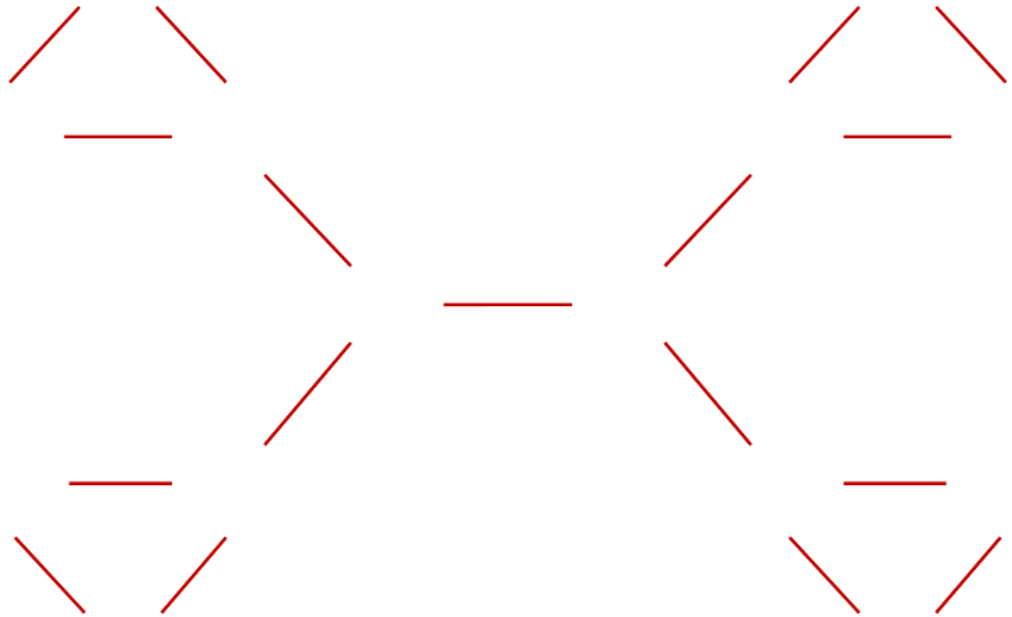
  $s \in S$, s community
- We need a reference graph to estimate the # expected edges



Girvan-Newman algorithm

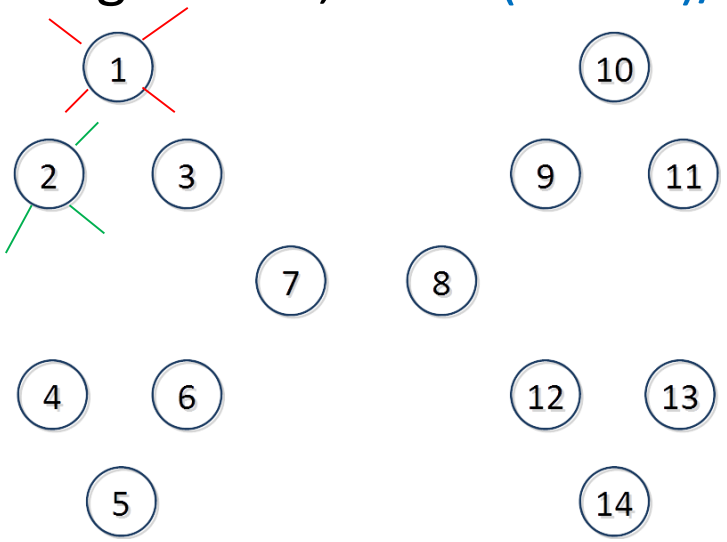
- Construct a network G' with same # nodes (n), same # edges (m), same degree distribution (see Graph Theory: Measures) but random connections





Girvan-Newman algorithm

- The expected number of edges between two u and v of degrees du, dv is: (du*dv)/2m

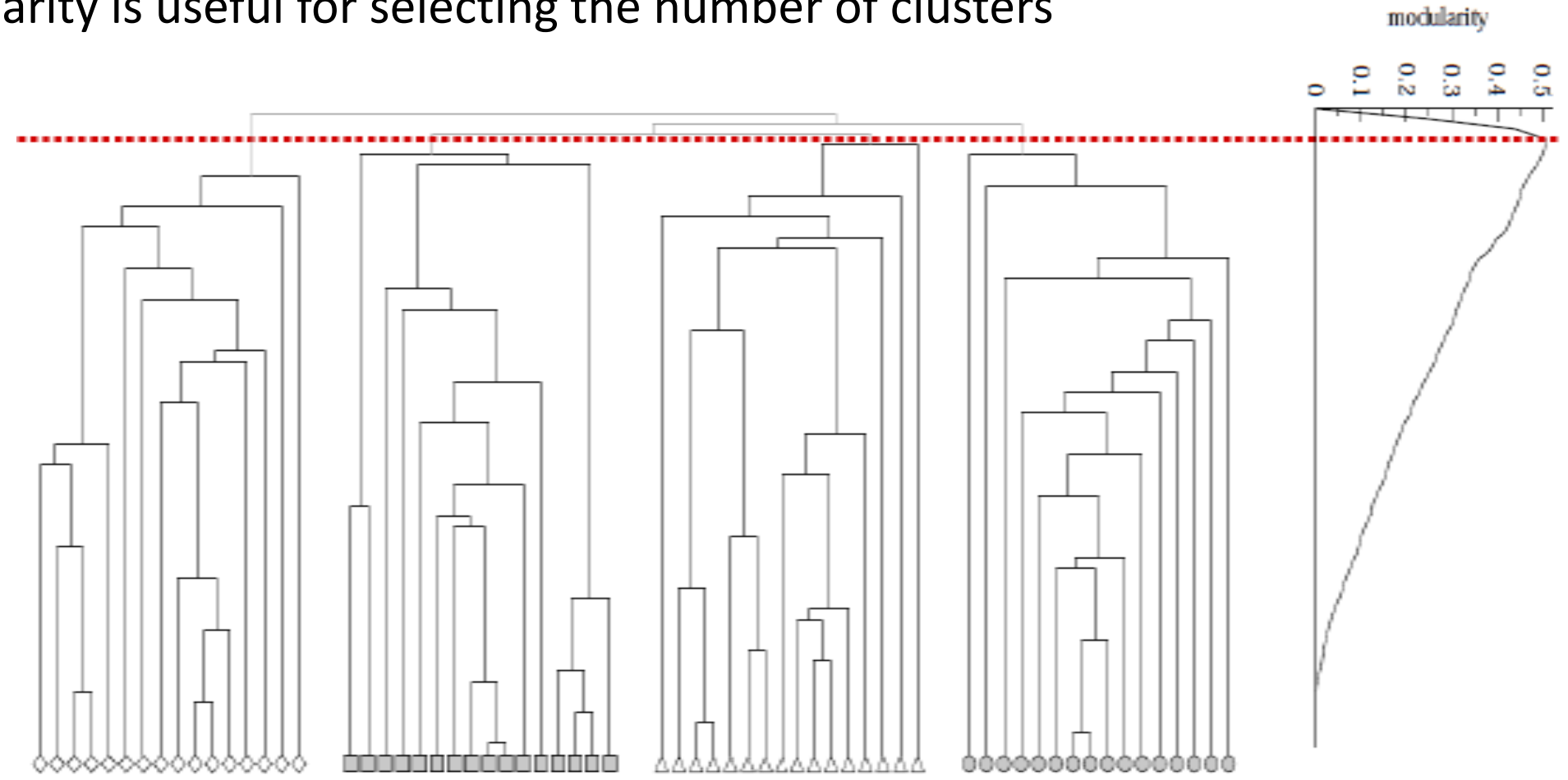


- Q ranges in [-1;1] and is positive if the number of edges within groups exceeds the expected number

- $\mathbf{0.3} < \boldsymbol{Q} < \mathbf{0.7}$ means significant community structure

Girvan-Newman algorithm

- Modularity is useful for selecting the number of clusters



Girvan-Newman algorithm

- Improvements:
  - Efficient computation of the betweenness (Breadth-first search)
  - Automatic determination of the modularity
  - Girvan-Newmam algorithm (and extended versions) available, for instance, in *R* software, package *igraph.*

References

- Andersen, R. and Lang, K.J. Communities from seed sets. WWW, 2006.
- Mcauley, J. and Leskovec, J. Discovering social circles in ego networks. *TKDD, 2014*.
- Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *TPAMI*, *22*(8),
- 888-905.