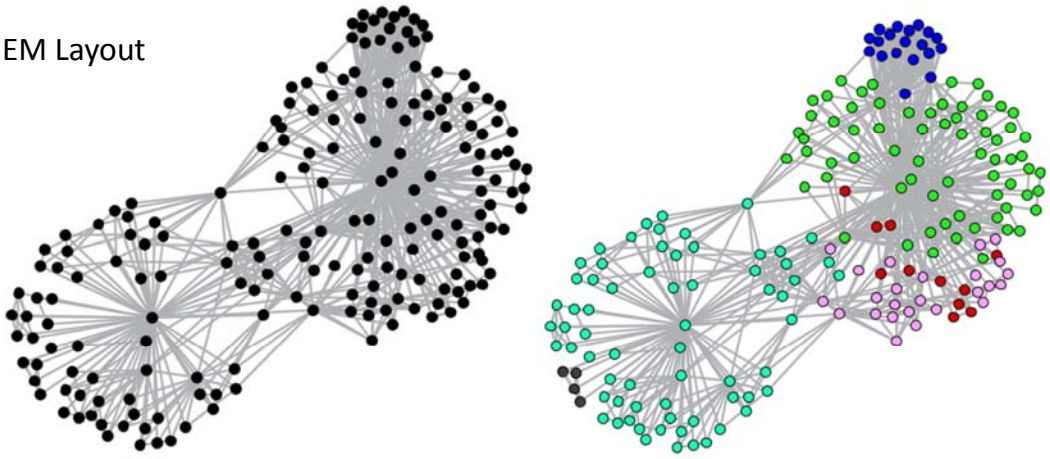
Agglomerative clustering using Blondel Community Detection

Algorithm reads a network and calculates additional attributes for each node (up to three community levels). Links are not modified.

GEM Layout



Blondel, Vincent D. Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics*. P10008. doi:10.1088/1742-5468/2008/10/P10008

Information Visualization MOOC

Unit 6 – "With Whom": Networks

Clustering

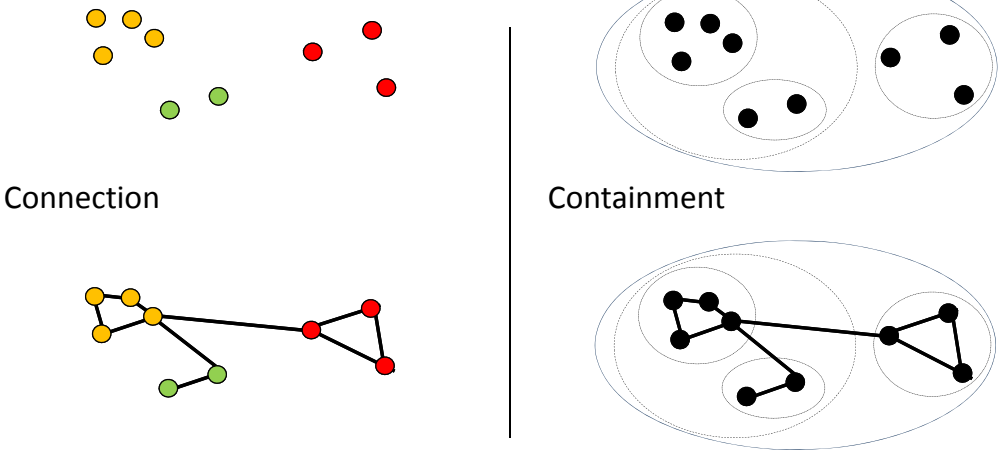
Reference:

Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486:75-174. http://arxiv.org/pdf/0906.0612.pdf

Goal

Clustering of graphs (also called community detection) aims to identify modules using the information encoded in the graph topology—i.e., geometric position and spatial relations.

Results might be visualized using graphic variable types (see Unit 1) or additional cluster boundaries.



Connection

Containment

Different Clustering Approaches

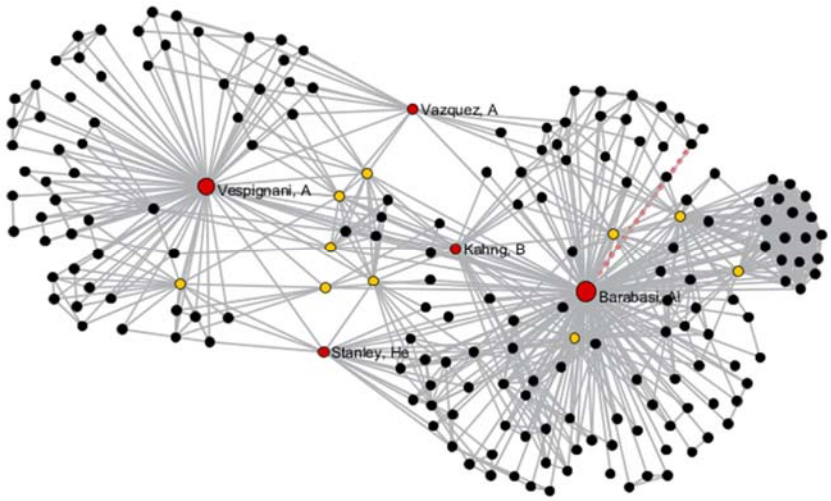
- Divisive algorithms: detect inter-community nodes or links and remove them from the network—e.g., using **betweenness centrality** thresholding.
- Agglomerative algorithms: merge similar nodes/communities recursively—e.g., **Blondel community detection**.
- Optimization methods: maximization of an objective function. See Clauset A, Newman M E J and Moore C, 2004 *Phys. Rev. E* 70 066111.

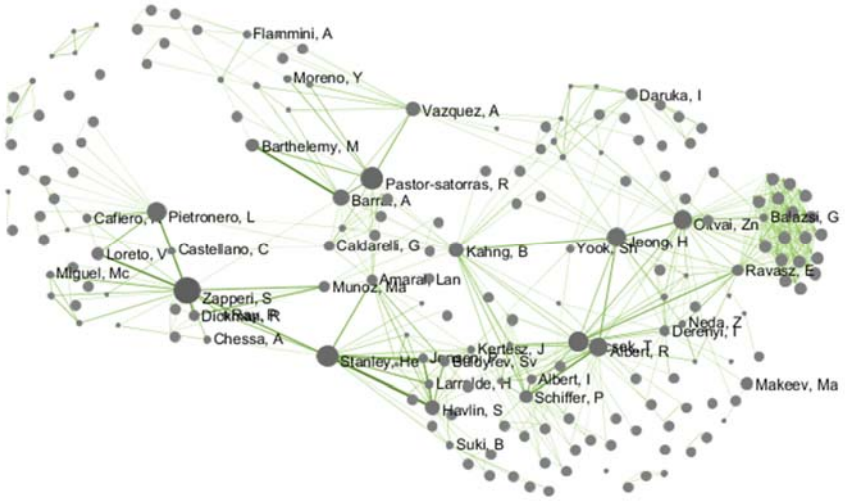
See examples on subsequent slides.

Calculate node betweenness centrality (BC) and delete high BC nodes

Collaboration network of Vespignani and Barabási

Nodes are sized and color coded by *betweenness centrality*



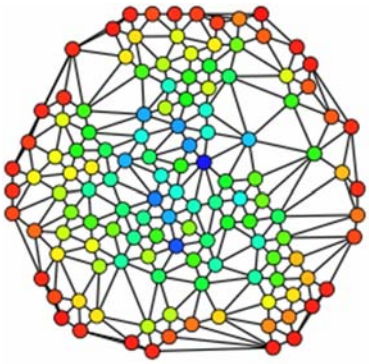


Betweenness Centrality (BC)

measures a node's centrality, load, or importance in a network. It equals the number of shortest paths from all nodes to all others that pass through that node. That is, nodes with higher 'betweenness' occur on more paths between other nodes.

BC of a node *n* in a network graph *G:=(N, L)* with *N* nodes is computed as follows:

- For each pair of nodes (*s*, *t*), compute the shortest paths between them.
- For each pair of nodes (*s*, *t*), determine the fraction of shortest paths that pass through the node in question (here, node *n*).
- Sum this fraction over all pairs of nodes (*s*, *t*).



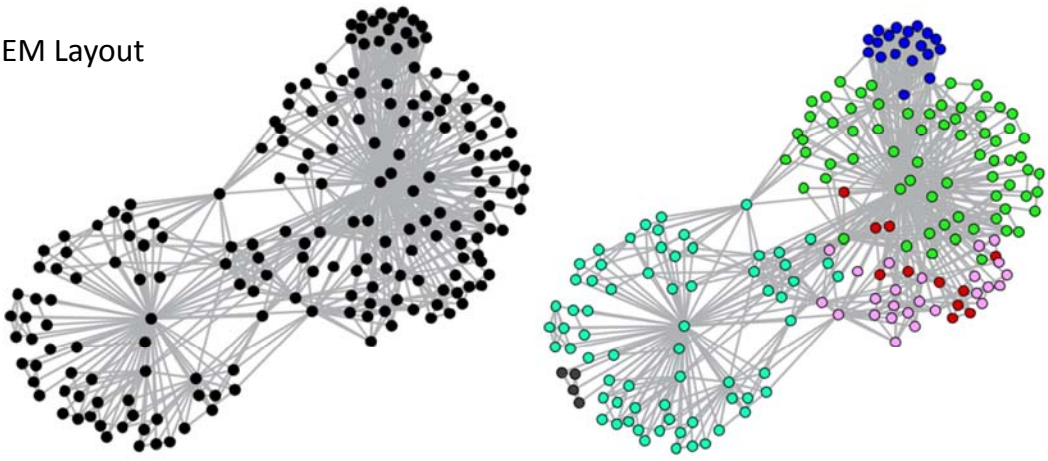
In the network to the right, node BC is represented by hue (from red=0 to blue=max).

See also http://en.wikipedia.org/wiki/Centrality

Agglomerative clustering using Blondel Community Detection

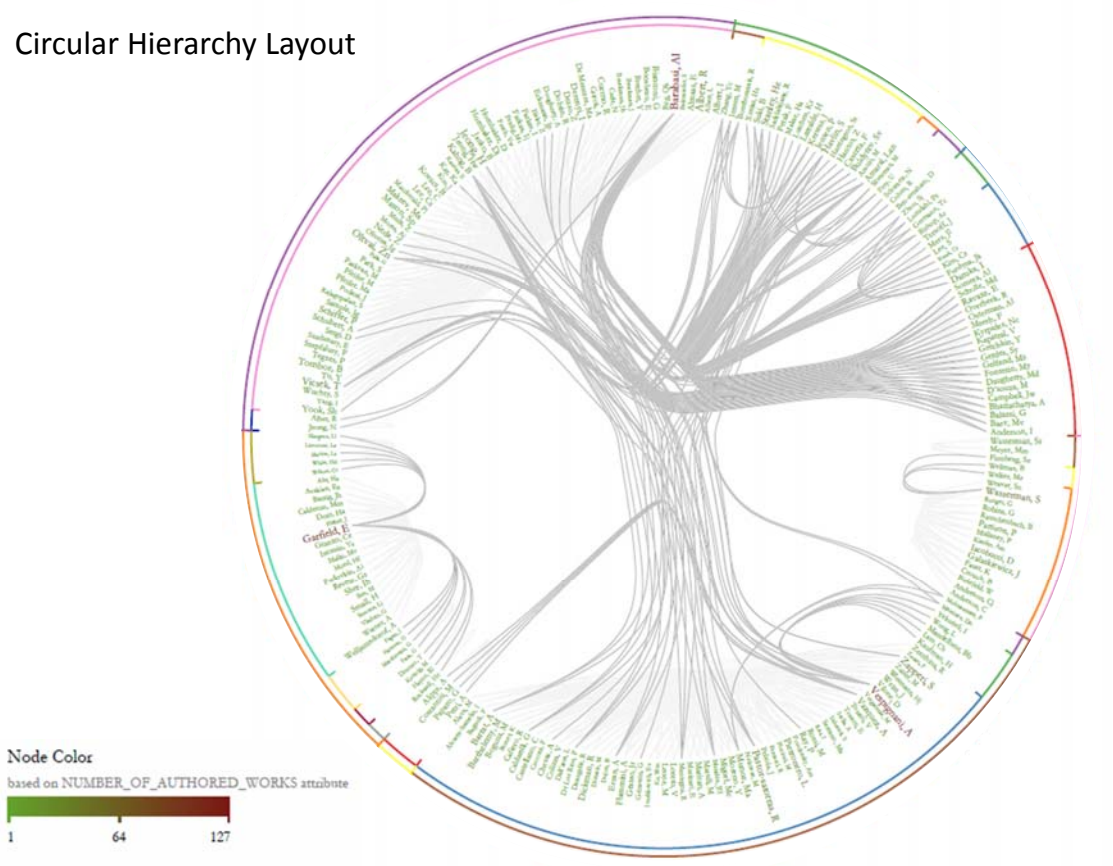
Algorithm reads a network and calculates additional attributes for each node (up to three community levels). Links are not modified.

GEM Layout



Blondel, Vincent D. Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics*. P10008. doi:10.1088/1742-5468/2008/10/P10008

Circular Hierarchy Layout



Blondel Community Detection

- Aims to partition a network into communities of densely connected nodes, with the nodes belonging to different communities being only sparsely connected.
- The quality of communities within a cluster partition is measured by the modularity of the partition. Modularity is a scalar value between -1 and 1 that measures the density of links within communities as compared to links between communities.
- Modularity can be used as an objective function to arrive at the best communities but also to compare different methods.

Blondel, Vincent D. Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics*. P10008. doi:10.1088/1742-5468/2008/10/P10008

Blondel Community Detection

Initialization

Algorithm starts with a weighted network of N nodes.

Each node is assigned to a different community—i.e., #nodes = #communities.

Phase I

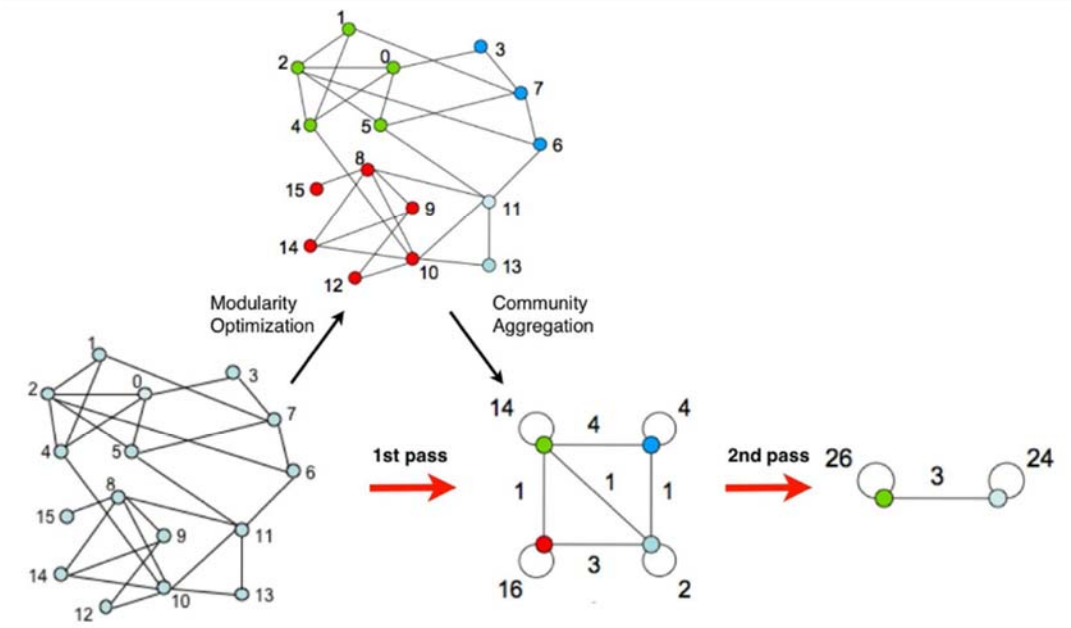
For each node *i* we consider the neighbors *j* of *i* and calculate the gain of modularity that would take place by removing *i* from its community and by placing it in the community of *j*. The node *i* is then placed in the community for which this gain is maximum (in case of a tie we use a breaking rule), but only if this gain is positive. If no positive gain is possible, *i* stays in its original community. This process is applied repeatedly and sequentially for all nodes until a local maxima of the modularity is attained—i.e., when no individual move can improve the modularity.

Phase II

Found communities are aggregated in order to build a new network of communities. Here, weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities. Links between nodes of the same community lead to self-loops for this community in the new network.

The passes are repeated iteratively until no increase of modularity is possible.

Blondel Community Detection



Blondel, Vincent D. Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics*. P10008. doi:10.1088/1742-5468/2008/10/P10008