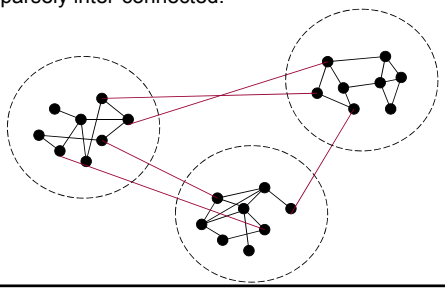
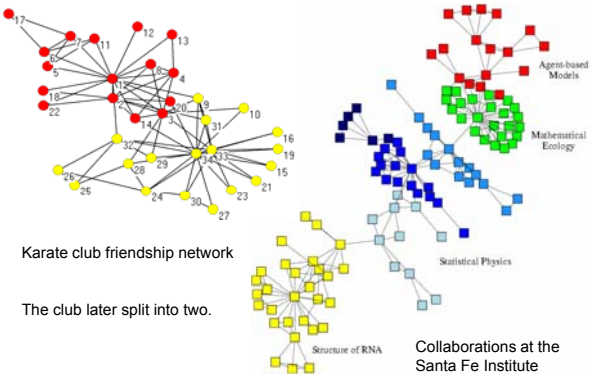
Community structure in networks

- Many real-world networks, especially social ones, exhibit community structure (also called modularity).
- Intuitively community structure can be defined as the existence of subgraphs that are densely connected but sparsely inter-connected.





Karate club friendship network

The club later split into two.

Collaborations at the Santa Fe Institute

Examples of communities

- Network : World Wide Web
  - Nodes : webpages
  - Edges : hyper-references
- Communities : Nodes on related topics

- Network : Friendship network
  - Nodes : people
  - Edges : friendship
- Communities : Group formation among people

- Network : Metabolic networks
  - Nodes : metabolites
  - Edges : participation in a chemical reaction
- Communities : Functional modules

Definitions of a community

- Cliques (completely connected subgraphs)
- Chain of cliques – adjacent cliques share every node except one

- k-clan – diameter (largest path length) is $\leq k$

- Definitions using the edges inside and outside a presumed community

  k_i^{in} – edges of node i that stay inside the community

  k_i^{out} – edges of node i that go outside of the community

  - Strong community: $k_i^{in} \geq k_i^{out}$ for every node i in the community
  - Weak community: $\sum_i k_i^{in} \geq \sum_i k_i^{out}$, where the sum is over nodes in the community

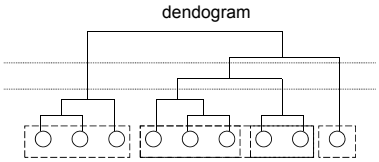
F. Radicchi et al., PNAS 101, 2658 (2004).

Community Detecting Algorithms

- Input: A network G(n,m)
- Output:
  - The number of communities
  - Classification of nodes into these communities
- Two families of methods:
  - Agglomerative (bottom up)
  - Divisive (top down)

Agglomerative method: hierarchical clustering

- Calculate a weight (connectivity measure) W_{ij} for every pair *i, j* of vertices
  - Example of weight: number of node-independent paths between *i* and *j*.
- Start with each node as a separate community
- Unite the highest-weight node pair(s)
- Calculate the weights between the newly formed community(ies) as averages over the nodes in the community
- Repeat



Divisive method: betweenness centrality algorithm

- Betweenness centrality of an edge is the number of shortest paths between pairs of vertices that run along it
- Algorithm:
  - Calculate the betweenness for all edges in the network
  - Remove the edge with highest betweenness
  - Recalculate the betweenness for all edges affected by the removal
  - Repeat

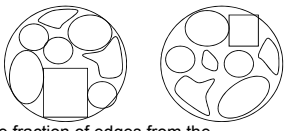
M.E.J. Newman, Phys. Rev. E 69, 066133, 2004.

This algorithm also leads to a dendogram

Q: When is it most meaningful to stop?

Strength of communities

To check if a particular division is meaningful, we can determine

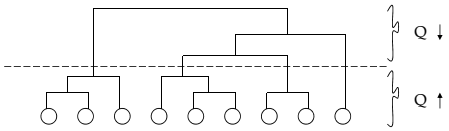


1. The fraction of edges within the community divided by the fraction of edges from the community to outside of the community
2. The observed number of edges within the community divided by the expected number of edges within a community if the edges are distributed randomly
3. Modularity measure Q, defined as the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random without regard for the community structure.

For either measure, the higher the result, the better the proposed community structure. M. Girvan and M.E.J. Newman, PNAS 99 (2002).

Modularity-based agglomeration

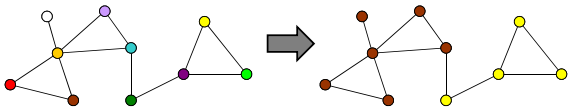
- If Q=0, implies the division gives no more within-community edges than would be expected by random chance. Q>0 indicates a significant community structure
- Begin with each node in its own community.
- Join two communities if that gives the maximum increase (or smallest decrease) in Q



Label propagation

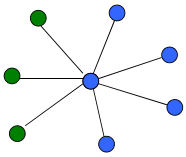
Motivation: within highly cohesive groups, individuals tend to have homogeneous beliefs.

- Each node is initialized with a separate label (is its own community)
- These labels flood across the network following a certain condition
- Stop when each node is in the community where most of its neighbors are.



Details of the algorithm

- Nodes updated in asynchronous rounds
- Label adoption condition: join the community to which the most adjacent nodes belong. Ties are broken randomly.



Converges after ~5 rounds of update.

No unique solution, but solutions are similar to each other.

Faster and as efficient as other algorithms.

U.N. Raghavan, R. Albert, S. Kumara
PRE 76, 036106 (2007).