GRAPH PARTITIONING AND COMMUNITY DETECTION

Book chapter

E.D. Kolaczyk, Statistical Analysis of Network Data, Chapter 4.3

Papers:

Fortunato S., Community detection in graphs, Physics Reports, Volume 486, Issues
3-5, February 2010, Pages 75-174, ISSN 0370-1573,
https://doi.org/10.1016/j.physrep.2009.11.002.
(http://www.sciencedirect.com/science/article/pii/S0370157309002841)

Network Cohesion

- Many questions in network analysis (not only social networks) boil down to questions involving network cohesion, the extent to which subsets of nodes are cohesive wrt the relation defining edges in the network

- Do friends of a given actor in a social network tend to be friends of one another as well?
- What collections of proteins in a cell appear to work closely together?
- Does the structure of the pages in the World Wide Web tend to separate with respect to distinct types of content?

- There are many ways that we can define network cohesion

- Definitions differ, for example, in scale, ranging from local (e.g., triads) to global (e.g., giant components)

- and also in the extent to which they are specified explicitly (e.g., cliques) versus implicitly (e.g., 'clusters').

- More generally we will call these aggregations of nodes as **communities**

Network Cohesion

- Social networks of various kinds demonstrate a strong community effect. Actors in a network tend to form closely-knit groups.

- groups are also called communities, clusters, cohesive subgroups or modules

- Generally speaking, individuals interact more frequently with members within group than outside
- Detecting cohesive groups in a social network (i.e., community detection) remains a core problem in explorative network analysis.
- Many approaches. These approaches can be separated into four categories: node-centric, group-centric, network-centric, and hierarchy-centric

Node centric

- Many notions of coherent network structure are based on the principle that a coherent subset of nodes should be locally 'dense' (often maximal dense)

- The most obvious concept to employ in this regard is that of a clique – a complete subgraph *H* of *G*.

- Cliques are subsets of vertices that are fully cohesive, in the sense that all vertices within the subset are connected
- A case of common practical interest, particularly in social network analysis, is that of 3-cliques (i.e., triangles).
- In practice, large cliques are relatively rare, as they necessarily require that *G* itself be fairly dense.

- cliques tend to be an overly restrictive definition of network cohesion

- weakened versions of this idea tend to be more practical

Node centric

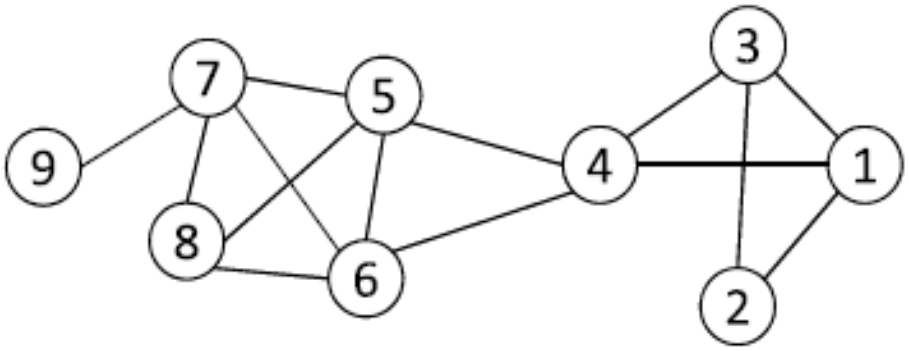


Figure 1.1: A social network of 9 actors and 14 connections. The diameter of the network is 5. The clustering coefficients of nodes 1-9 are: $C_1 = 2/3$, $C_2 = 1$, $C_3 = 2/3$, $C_4 = 1/3$, $C_5 = 2/3$, $C_6 = 2/3$, $C_7 = 1/2$, $C_8 = 1$, $C_9 = 0$.

Table 1.3: Adjacency Matrix

Node	1	2	3	4	5	6	7	8	9
1	-	1	1	1	0	0	0	0	0
2	1	-	1	0	0	0	0	0	0
3	1	1	-	1	0	0	0	0	0
4	1	0	1	-	1	1	0	0	0
5	0	0	0	1	-	1	1	1	0
6	0	0	0	1	1	-	1	1	0
7	0	0	0	0	1	1	-	1	1
8	0	0	0	0	1	1	1	-	0
9	0	0	0	0	0	0	1	0	-

Node centric

- An ideal cohesive subgroup is a clique. It is a maximum complete subgraph in which all nodes are adjacent to each other.
- Typically, cliques of larger sizes are of much more interest. However, the search for the maximum cliques in a graph is an NP-hard problem.
- One brute-force approach is to traverse all nodes in a network. For each node, check whether there is any clique of a specified size that contains the node.

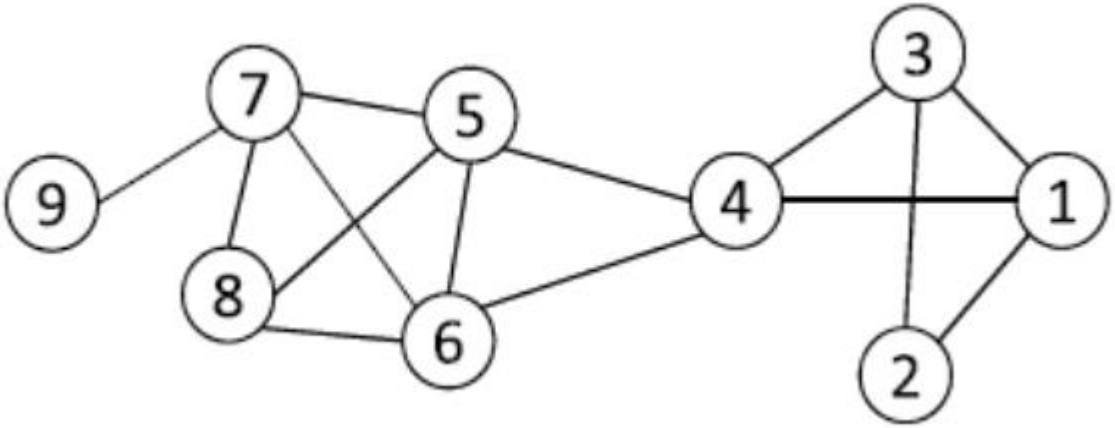
Node centric: brute force

Suppose we now look at node v_C. We can maintain a queue of cliques. It is initialized with a clique of one single node $\{v_C\}$.

Then we perform the following:

- Pop a clique from the queue, say, a clique B_k of size k. Let v_i denote the last added node into B_k.
- For each of v_i's neighbor v_j (to remove duplicates, we may look at only those nodes whose index is larger than v_i), form a new candidate set $B_{k+1} = B_k \cup \{v_j\}$.
- Validate whether B_{k+1} is a clique by checking whether v_j is adjacent to all nodes in B_k. Add to the queue if B_{k+1} is a clique.

Node centric: brute force



Take the network in Figure 1.1 as an example. Suppose we start from node $B_1 = \{4\}$. For each of its friends with a larger index, we obtain a clique of size 2. Thus, we have $\{4, 5\}$ and $\{4, 6\}$ added into the queue. Now suppose we pop $B_2 = \{4, 5\}$ from the queue. Its last added element is node 5. We can expand the set following node 5's connections and generate three candidate sets: $\{4, 5, 6\}$, $\{4, 5, 7\}$ and $\{4, 5, 8\}$. Among them, only $\{4, 5, 6\}$ is a clique as node 6 is connected both nodes 4 and 5. Thus, $\{4, 5, 6\}$ is appended to the queue for further expansion for larger cliques.

Node centric: Pruning

- The exhaustive search above works for small-scale networks
- impractical for large-scale networks
- If the goal is to find out a maximum clique, then a strategy is to effectively prune those nodes and edges that are unlikely to be contained in the maximum clique.
- For a clique of size k, each node in the clique should maintain at least degree k − 1.
- Hence, those nodes with degree less than k − 1 cannot be included in the maximum clique, thus can be pruned

Node centric: Pruning

We can recursively apply the pruning procedure below to a given network:

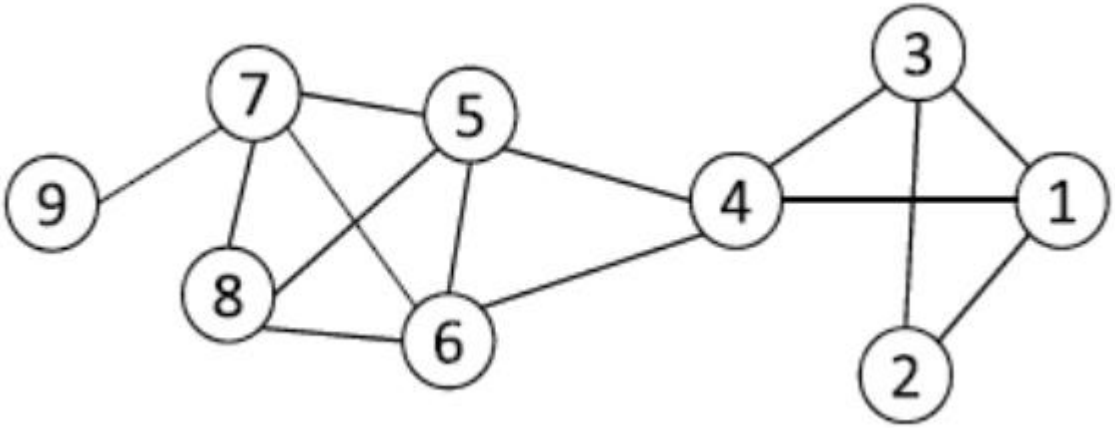
- A sub-network is sampled from the given network. A clique in the sub-network can be found in a greedy manner, e.g., expanding a clique by adding an adjacent node with the highest degree.

- The maximum clique found on the sub-network (say, it contains k nodes) serves as the lower bound for pruning. That is, the maximum clique in the original network should contain at least k members.

Hence, in order to find a clique of size larger than k, subgraph composed of the nodes with degree less than or equal to $k - 1$ and their connections can be removed from future consideration.

As real social networks follow a power law distribution for node degrees, i.e., the majority of nodes have a low degree, this pruning strategy can reduce the network size significantly.

Node centric: Pruning



Suppose we randomly sample a sub-network from the network in Figure 1.1. It consists of nodes 1 to 6. A maximal clique in the sub-network is of size 3 ({1, 2, 3} or {1, 3, 4}). If there exists a larger clique (i.e., size $>$ 3) in the original network, all the nodes of degree less than or equal to 2 can be removed from consideration. Hence, nodes 9 and 2 can be pruned. Then, the degree of nodes 1 and 3 is reduced to 2, thus they can also be removed. This further leaves node 4 with only 2 connections, which can be removed as well. After this pruning, we obtain a much smaller network of nodes {5, 6, 7, 8}. And in this pruned network, a clique of size 4 can be identified. It is exactly the maximum clique.

Node centric: Rechability

- This type of community considers the reachability among actors. In the extreme case, two nodes can be considered as belonging to one community if there exists a path between the two nodes.

Thus in principle each connected component is a community if we follow the reachability approach

- In real-world networks, a giant component tends to form while many others are singletons and minor communities.

- Conceptually, there should be a short path between any two nodes in a group.

Node centric: Rechability

- Some structures based on rechability approach are the following:

1. k-clique is a maximal subgraph in which the largest geodesic distance between any two nodes is no greater than k. The largest geodesic is computed on the original network. So that:

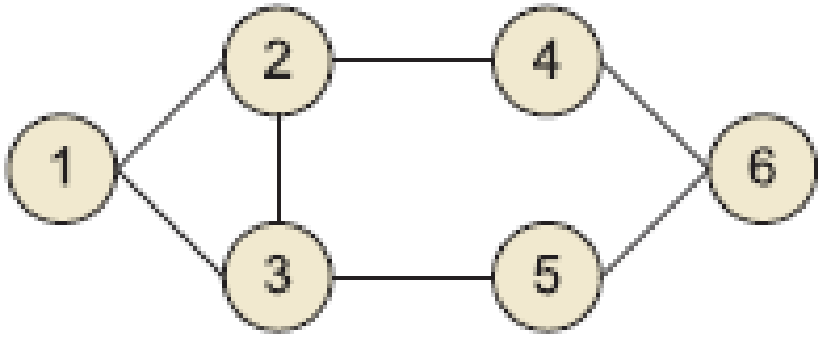
$$d(v_i, v_j) \leq k \quad \forall v_i, v_j \in V_s$$

where V_s is the set of nodes in the subgraph

2. k-club restricts the geodesic distance within the group to be no greater than k. It is a maximal substructure of diameter k.

The definition of k-club is more strict than that of k-clique. A k-club is often a subset of a k-clique.

Node centric: Rechability



cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5, }, {2, 3, 4, 5, 6}
2-clubs: {1, 2, 3, 4, }, {1, 2, 3, 5}, {2, 3, 4, 5, 6}

{1, 2, 3, 4, 5} form a 2-clique. But the geodesic distance between nodes 4 and 5 within the group is 3.

The 2-clique structure {1, 2, 3, 4, 5} contains two 2-clubs, {1, 2, 3, 4} and {1, 2, 3, 5}.

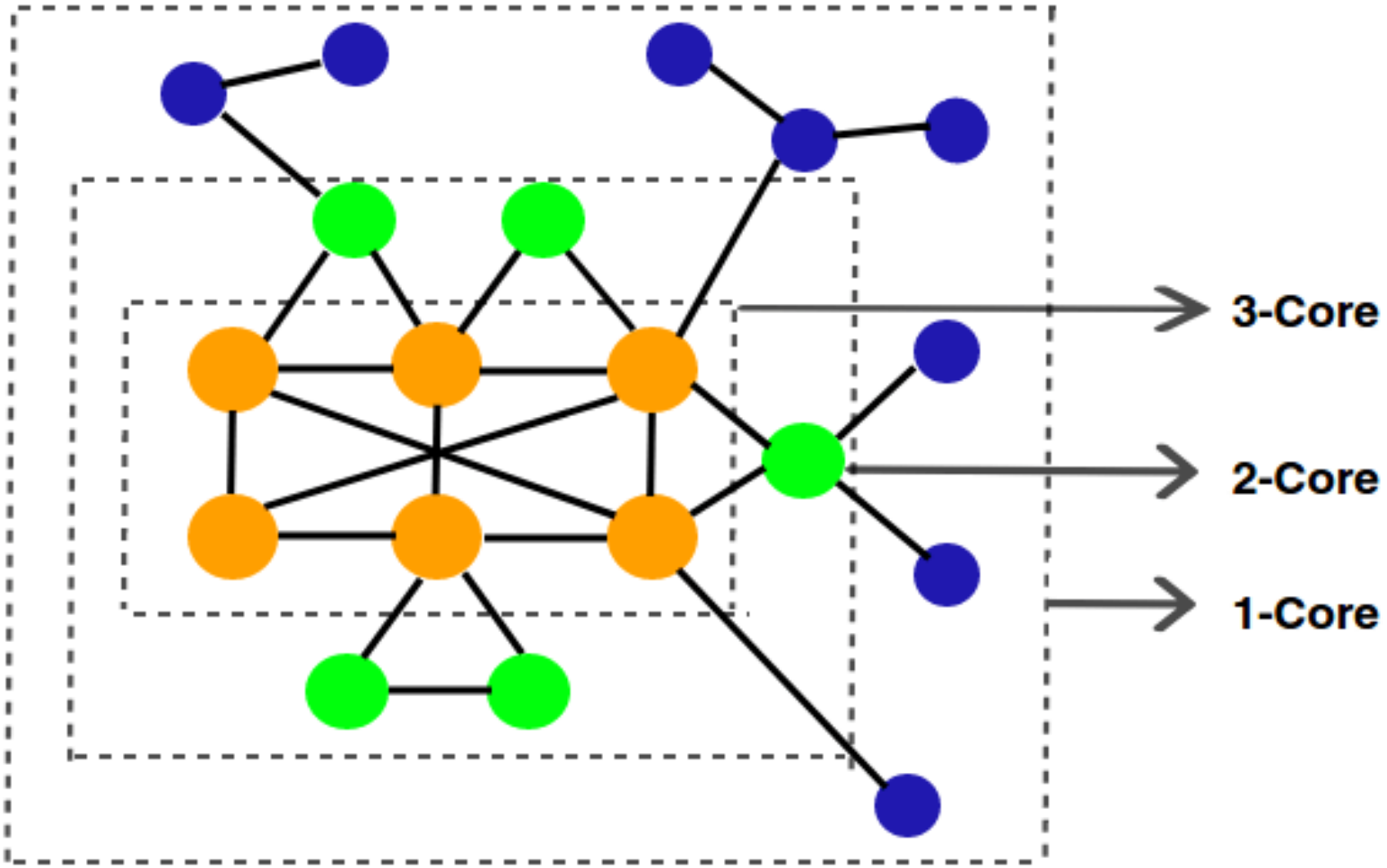
Node centric: k-cores

• Maximal subgraph C such that each vertex is adjacent to at least k other vertices in the subgraph (k integer greater or equal to 0)

- In C all vertex degrees are at least k, and no other subgraph obeying the same condition contains it (i.e., it is maximal in this property)
- k-cores are one natural way to look at group structure across a graph G.

k-core decomposition has been at the heart of a number of proposals for the representation of large Networks (see section **3.5.2.3** of the textbook)

Node centric: k-cores



Brief summary

Nodes are all adjacent to each other or we use a relaxation based on geodesic/degree

$$\delta^{int}(C) = 1$$

- Local definition
- Triangles are frequent; larger cliques are rare.
- Communities do not necessarily correspond to complete subgraphs, as many of their nodes do not link directly to each other.
- Among the others the notion of k-clubs and k-core is rather important

What about communities in social networks for example?

Disjoint communities (i.e., groups of friends who don't know each other) e.g. my American friends and my Australian friends

Overlapping communities (i.e., groups with some intersection) e.g. my friends and my girlfriend's friends

Intuition:

There are more edges inside a community than edges connected with the rest of the network

Types of communities

Two types of communities:

-Explicit Groups: formed by user subscriptions

- Implicit Groups: implicitly formed by social interactions

generally the concept of community in community detection is a relaxation of the communities found in the nodal-based approach

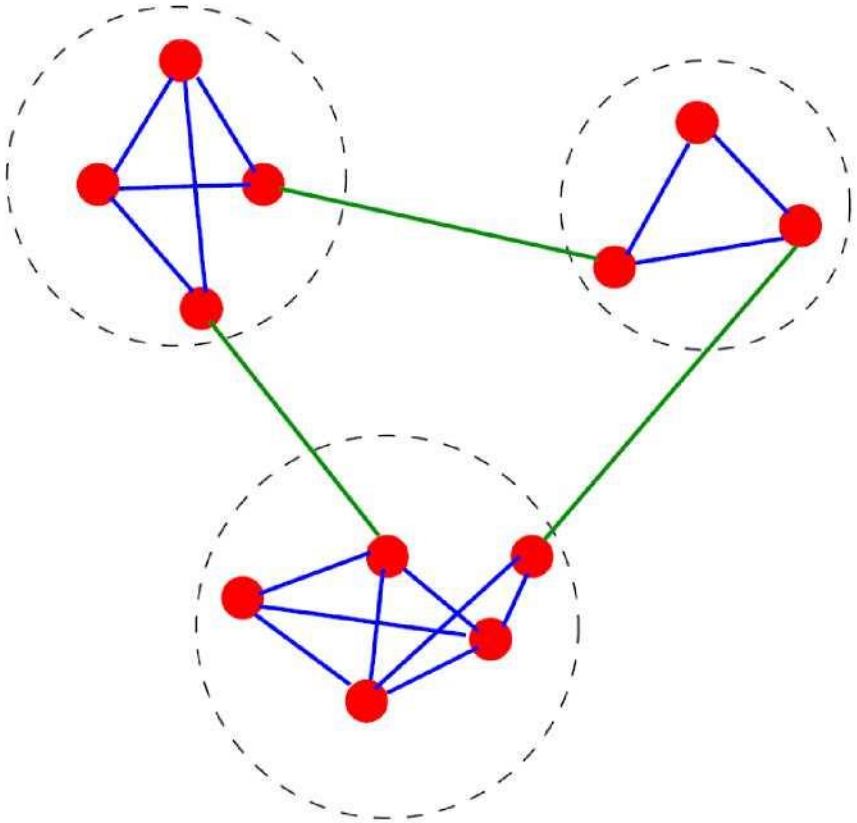
COMMUNITY DETECTION

Basic definition:

DISCOVERING IMPLICIT COMMUNITIES

COMPUTE SETS OF NODES BASED ON THEIR CONNECTIVITY

Examples of communities



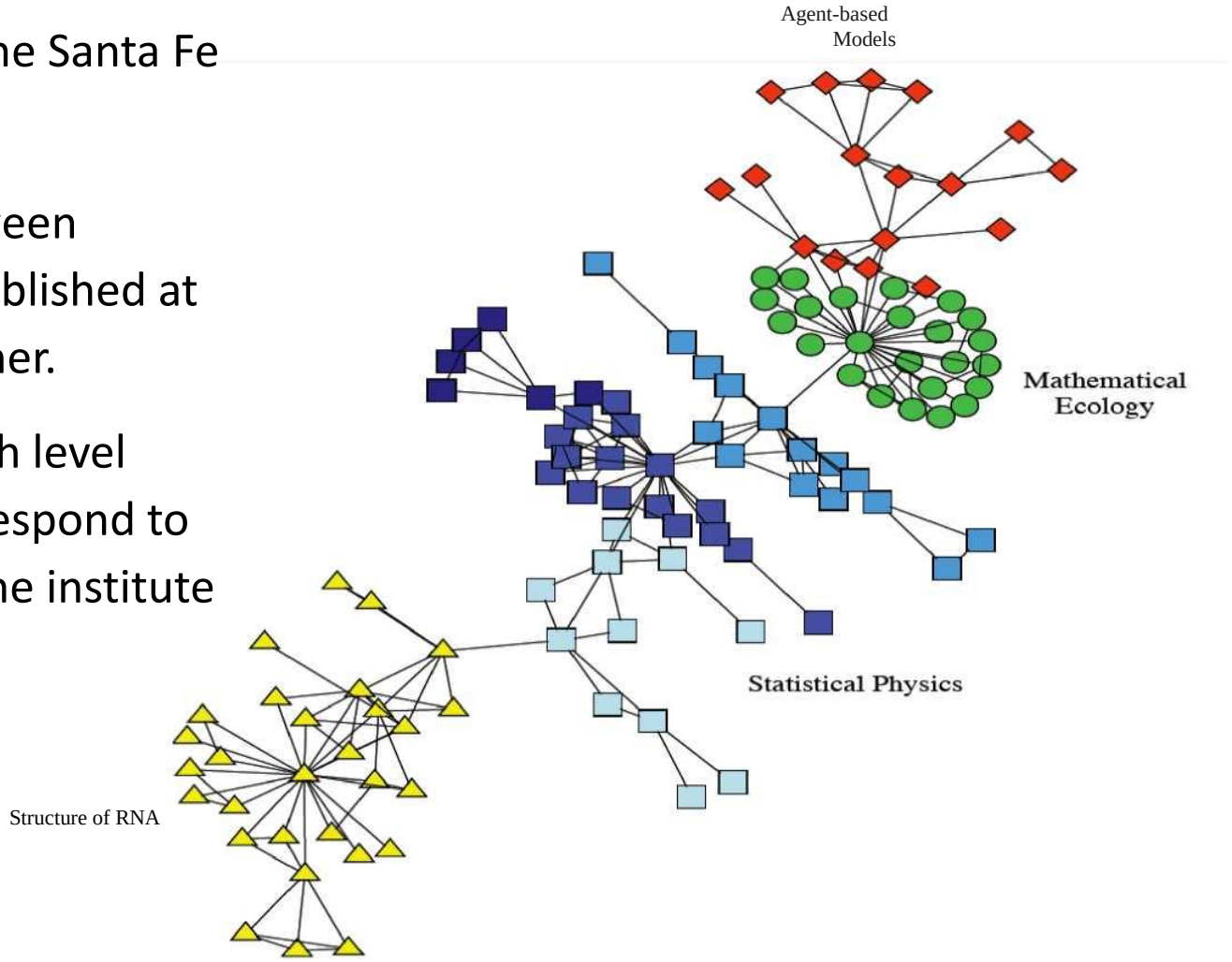
Real networks are not random. Weak ties seem to bridge groups of tightly coupled nodes (communities)

A simple graph with three communities, enclosed by the dashed circles

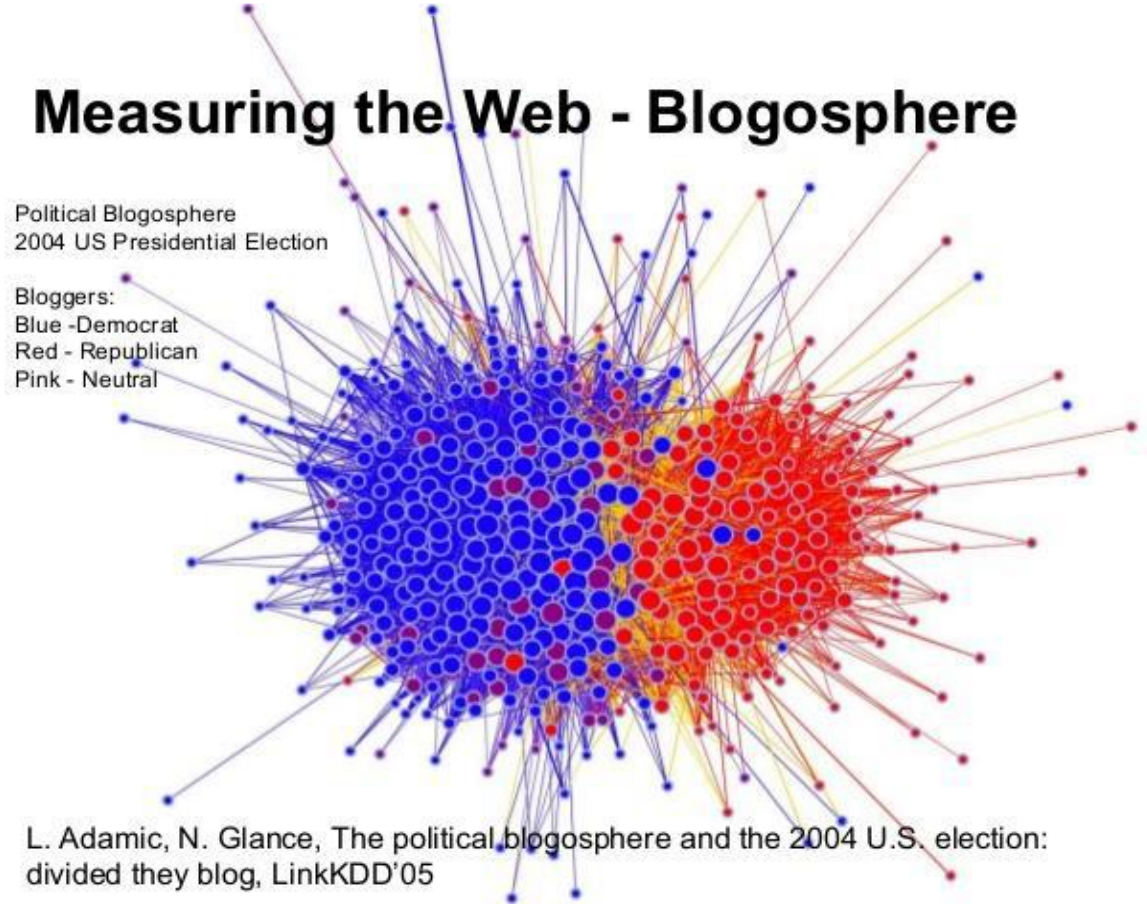
Source: S. Fortunato / Physics Reports 486 (2010) 75-174

Collaboration networks

- Collaboration network between scientists working at the Santa Fe Institute.
- Edges are placed between scientists that have published at least one paper together.
- The colors indicate high level communities and correspond to research divisions of the institute

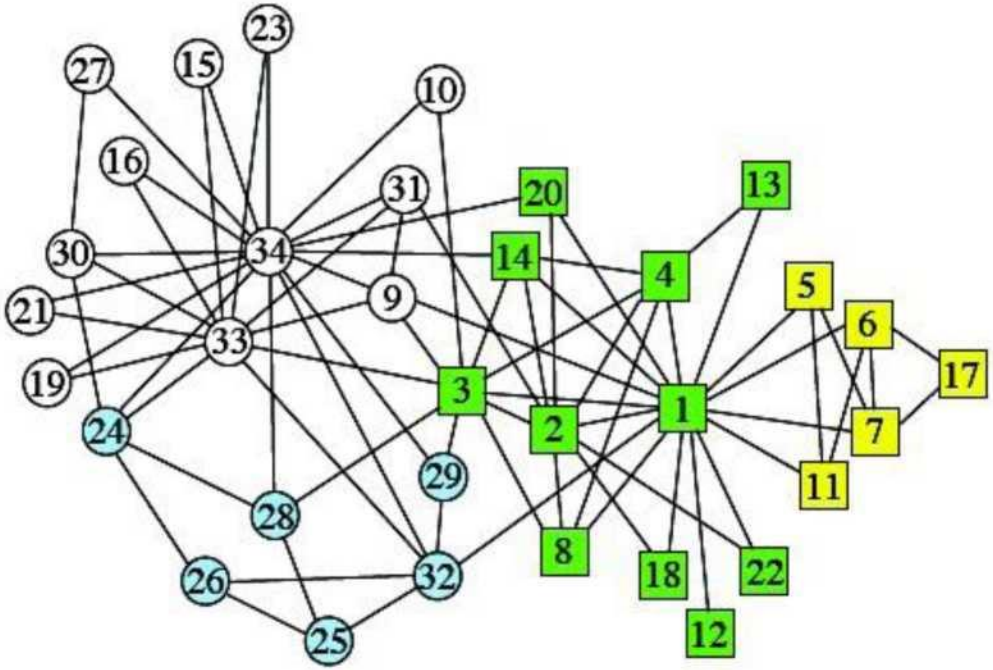


Political networks



Separating networks into disjoint subsets seems to make sense when communities are somehow “adversarial”

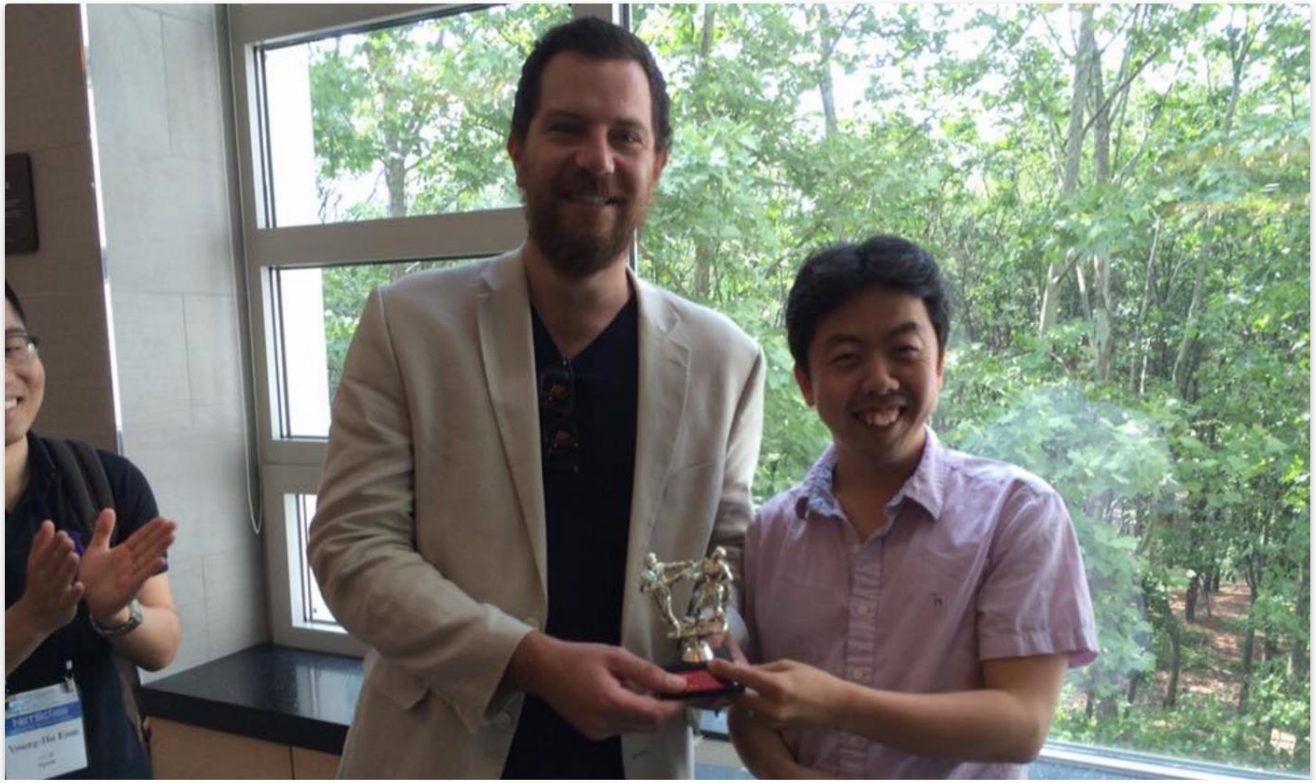
Zachary karate club example



Zachary observed 34 members of a karate club over two years. Edges connect individuals who were observed to interact outside the activities of the club.

During the course of the observation, the club members split into 2 groups because of the disagreement between the administrator of the club and the club's instructor (nodes 1 and 34), and the members of one group left to start their own club

Network scientists with Karate Trophies



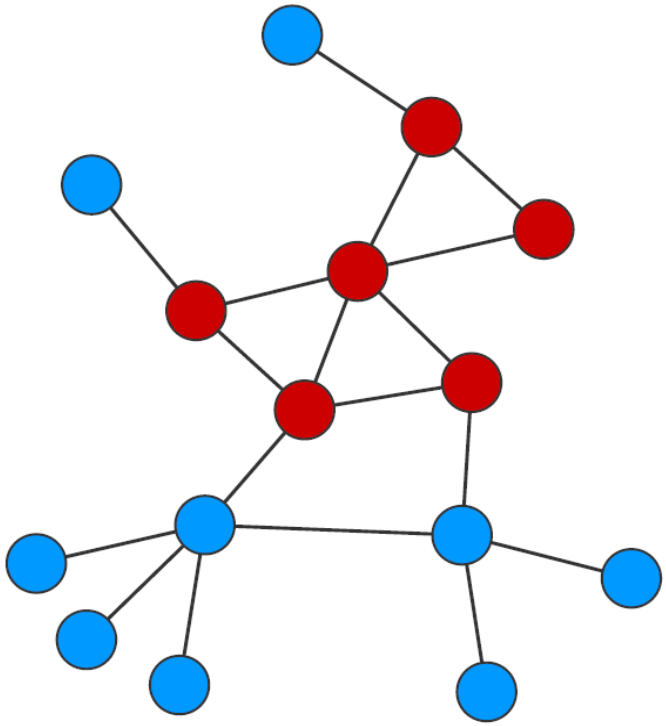
The first scientist at any conference on networks who uses Zachary's karate club as an example is inducted into the Zachary Karate Club Club, and awarded a prize. This tumblr records those moments.

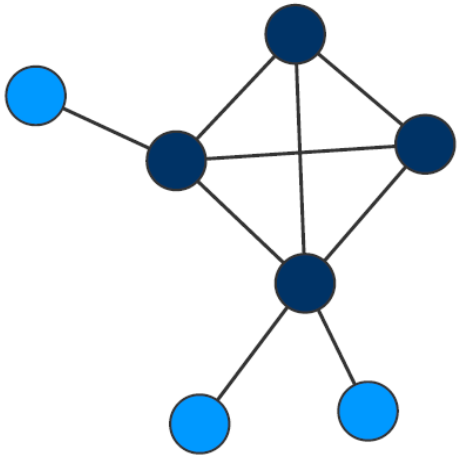
RSS
ARCHIVE

Hypotheses

Connectedness Hypothesis:
A community corresponds to a connected subgraph.

Density Hypothesis:
Communities correspond to locally dense neighborhoods of a network.





Connectedness

In contrast to a local perspective, and the search for small-scale subsets of cohesive vertices, we may also find it useful in some contexts to take a larger, global perspective.

A basic question of interest is whether a given graph separates into distinct subgraphs. If it does not, we might seek to quantify how close to being able to do so it is. Intimately related to such issues are questions associated with the flow of ‘information’ in the network.

Vertex/Edge-Connectivity

Somewhat more refined notion of connectivity derives from asking questions like, “If an arbitrary subset of k vertices (edges) is removed from a graph, is the remaining subgraph connected?”

Extreme case where G consists of two cliques joined by a single edge between a vertex in each.

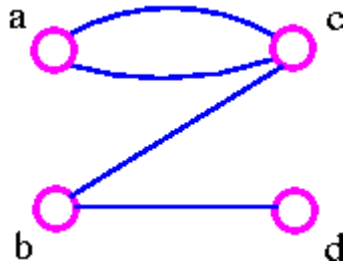
We would be inclined to think of G as consisting of ‘nearly two components.’

The concepts of vertex- and edge-connectivity, and the related concepts of vertex- and edge-cuts, help to make these notions precise.

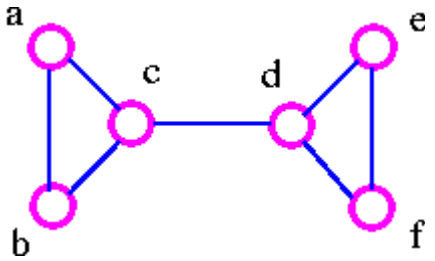
Vertex/Edge-Connectivity

If the removal of a particular set of vertices (edges) in *G* actually disconnects the graph, that set is called a *vertex-cut (edge-cut)*.

Cut Edge (Bridge) A bridge is a single edge whose removal disconnects a graph. Therefore, edge bc or bd is a bridge.



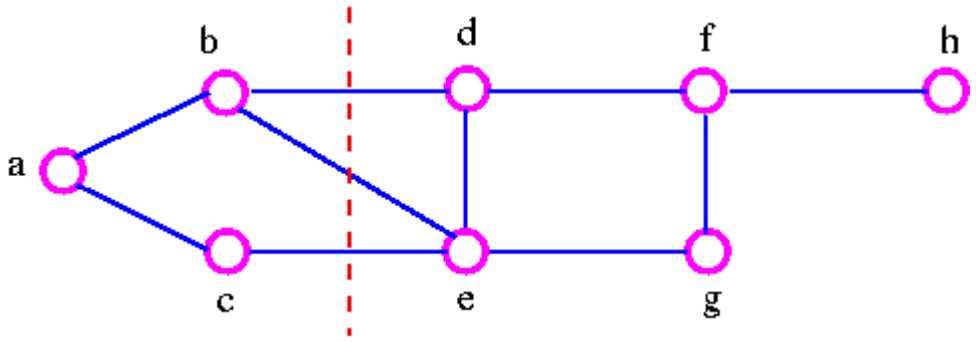
The graph can be disconnected by removing a single edge, cd. Therefore, edge cd is a bridge



Vertex/Edge-Connectivity

Cut Set: A cut set of a connected graph G is a set S of edges with the following properties

- The removal of all edges in S disconnects G.
- The removal of some (but not all) of edges in S does not disconnects G.



We can disconnect *G* by removing the three edges bd, be, and ce, but we cannot disconnect it by removing just two of these edges. Note that a cut set is a set of edges in which no edge is redundant.

Vertex/Edge-Connectivity

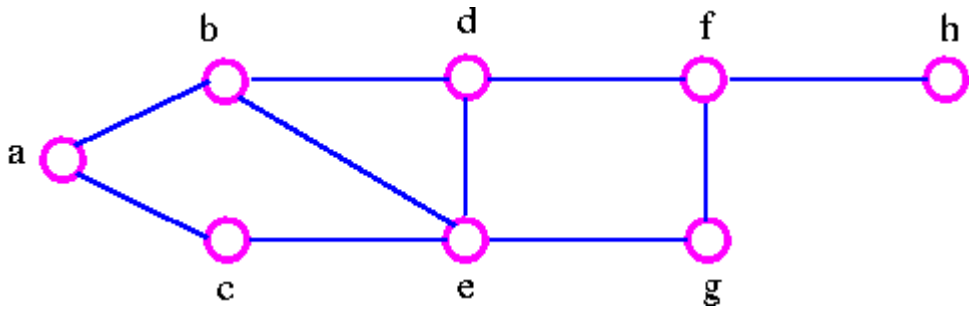
Cut Vertex: A cut-vertex is a single vertex whose removal disconnects a graph.

This definition breaks down if G is a complete graph, since we cannot then disconnects G by removing vertices.

Vertex-Cut set

A vertex-cut set of a connected graph G is a set S of vertices with the following properties.

- the removal of all the vertices in S disconnects G.
- the removal of some (but not all) of vertices in S does not disconnects G



We can disconnects the graph by removing the two vertices b and e, but we cannot disconnect it by removing just one of these vertices. the vertex-cutset of G is {b, e}.

Density based

we can define a measure of local density and then characterize the extent to which subsets of vertices are dense, according to this measure.

Such measures are commonly based on ratios of the number of edges among a subset of vertices to the total number of possible edges.

Similarly it is possible to use the notion of internal and external node or overall degrees

Density-based

Graph G, subgraph C have n and n_c vertices

k_v^{int}, k_v^{ext}: Internal and external degrees of $v \in C$

k_{int}^C, k_{ext}^C: Internal and external degrees of C

[Intra-cluster density:

$$\delta_{int}(C) = \frac{\#\ internal\ edges\ of\ C}{n_c(n_c - 1)/2}$$

Inter-cluster density:

$$\delta_{ext}(C) = \frac{\#inter-cluster\ edges\ of\ C}{n_c(n-n_c)}]$$