COMP0123

Complex Networks and Web

14. Network Communities

Network Communities

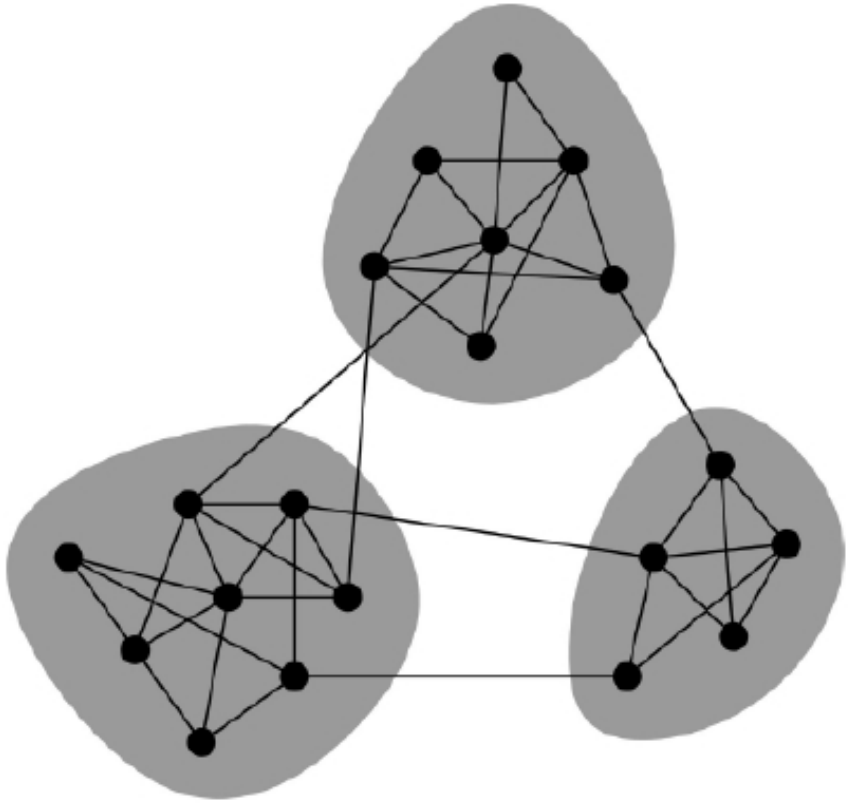
- **Network communities:**

Groups of nodes with

lots of connections **inside**

and

few connections to **outside.**



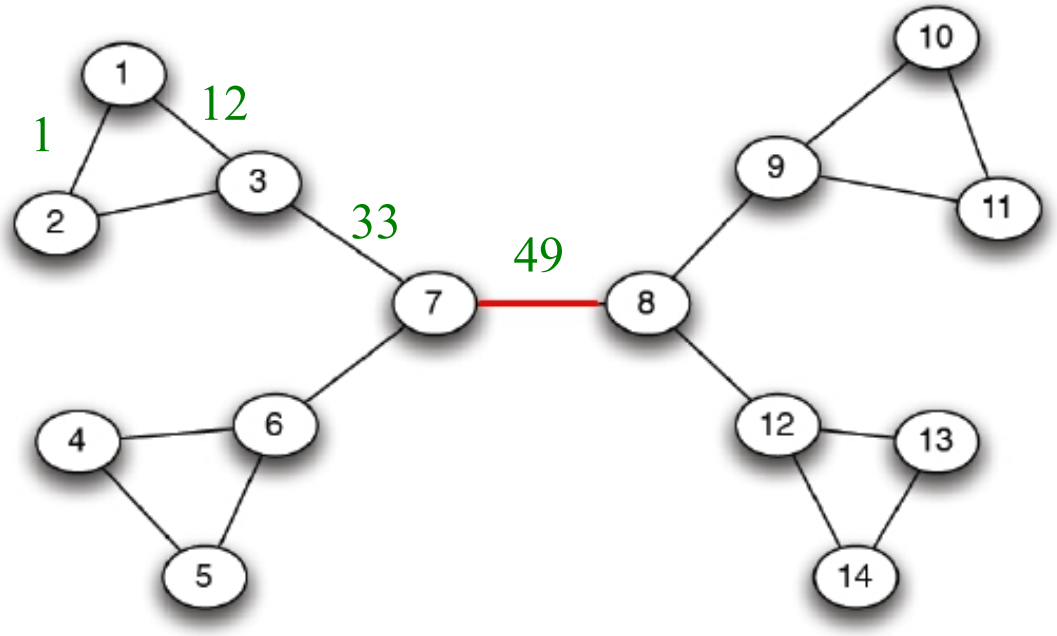
Communities, clusters, groups, modules

Community Detection

- Why do we want to find communities?
  - Customers with similar interests could be clustered to help recommendation systems.
  - Identification of clusters in WWW can improve page ranking.
  - Study hierarchical organisation in a network.
- How to find densely connected communities of nodes in a network?
  - We consider undirected, unweighted networks
  - A method based on **edge betweenness**

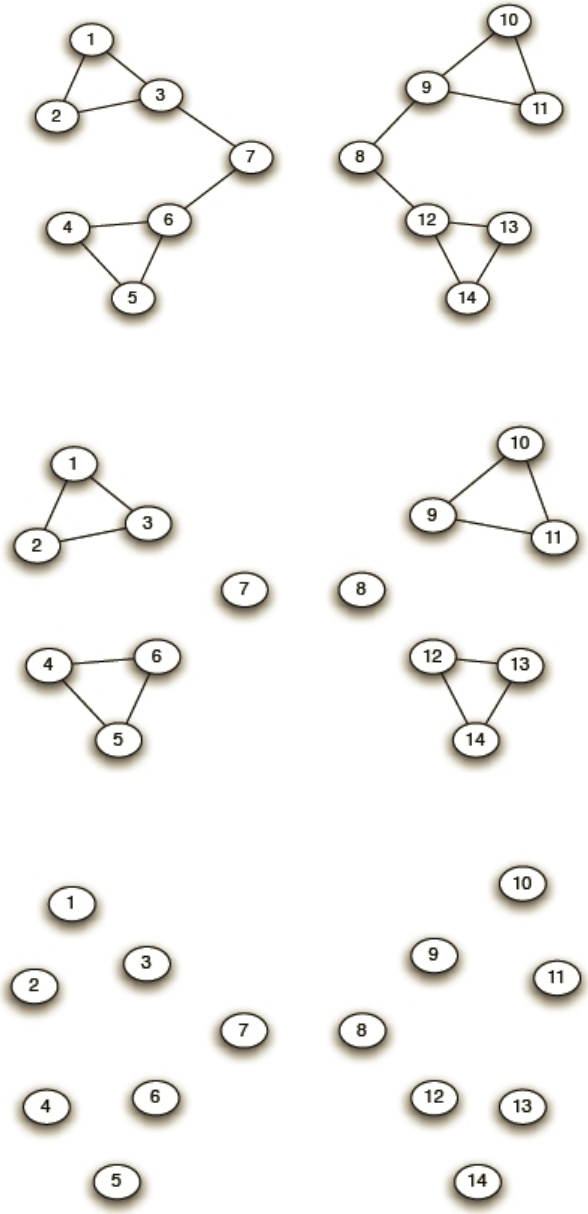
Edge Betweenness

- The number of shortest paths through an edge



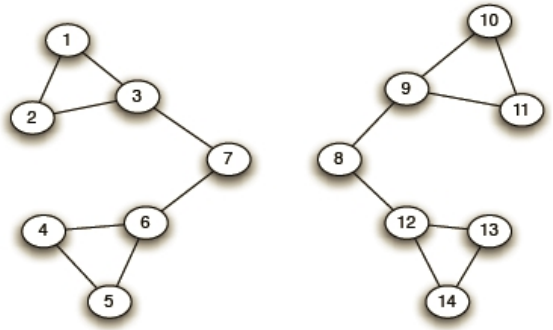
Community Detection

- **Girvan-Newman Algorithm**
  - Calculate betweenness of edges
  - Remove edges with highest betweenness
  - Repeat the above two procedures
  - Until no edges are left

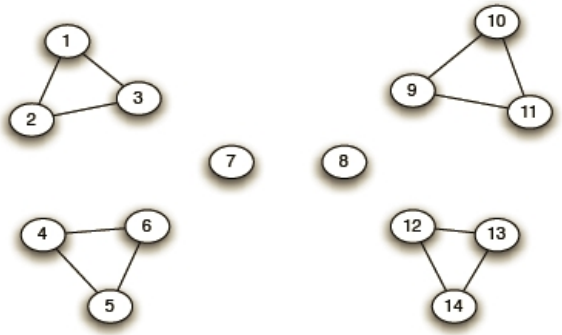


Girvan-Newman: Example

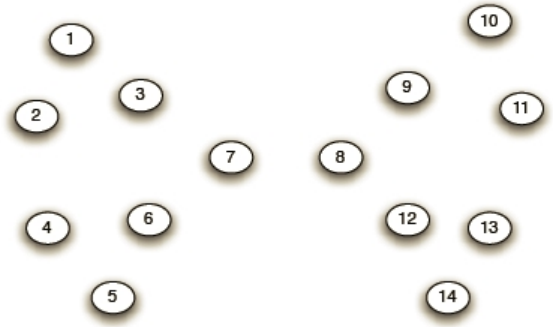
Step 1:



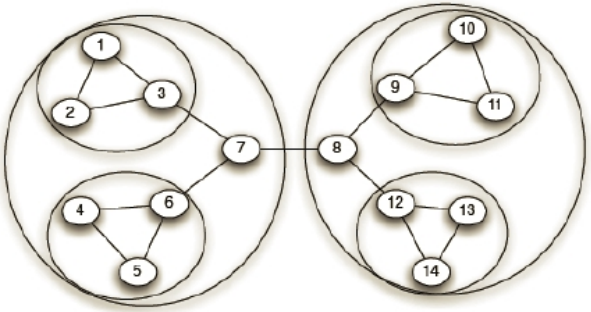
Step 2:



Step 3:

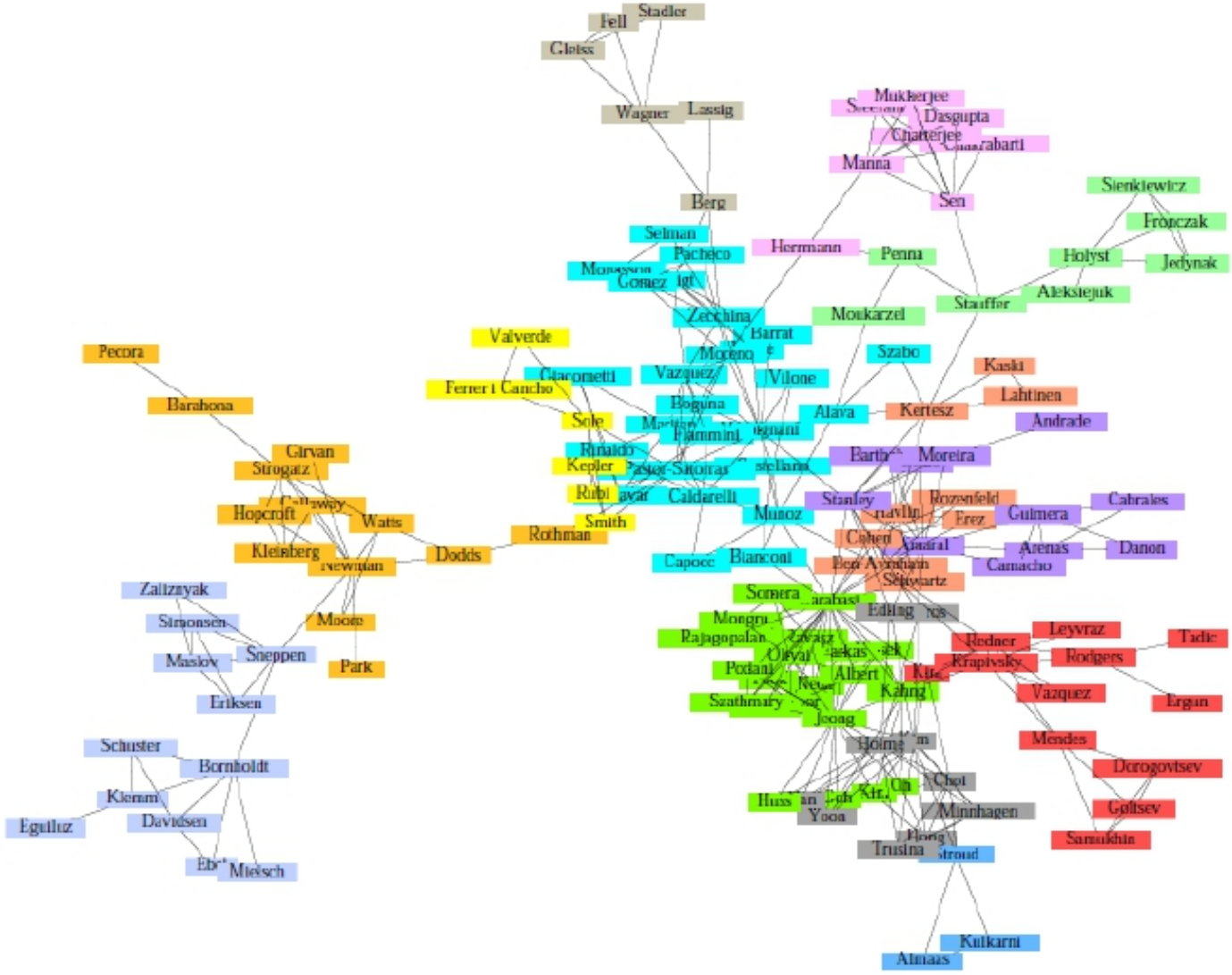


Hierarchical decomposition:



- At **each** step, connected components are communities
- Gives a hierarchical decomposition of the network

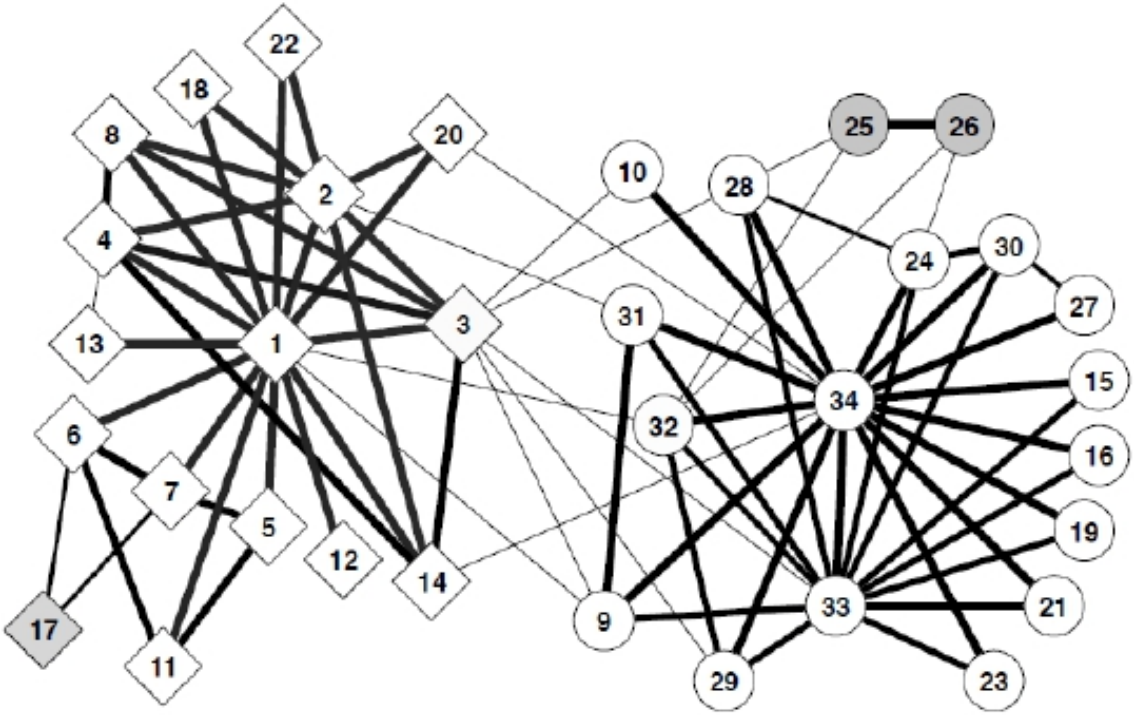
Girvan-Newman: Results



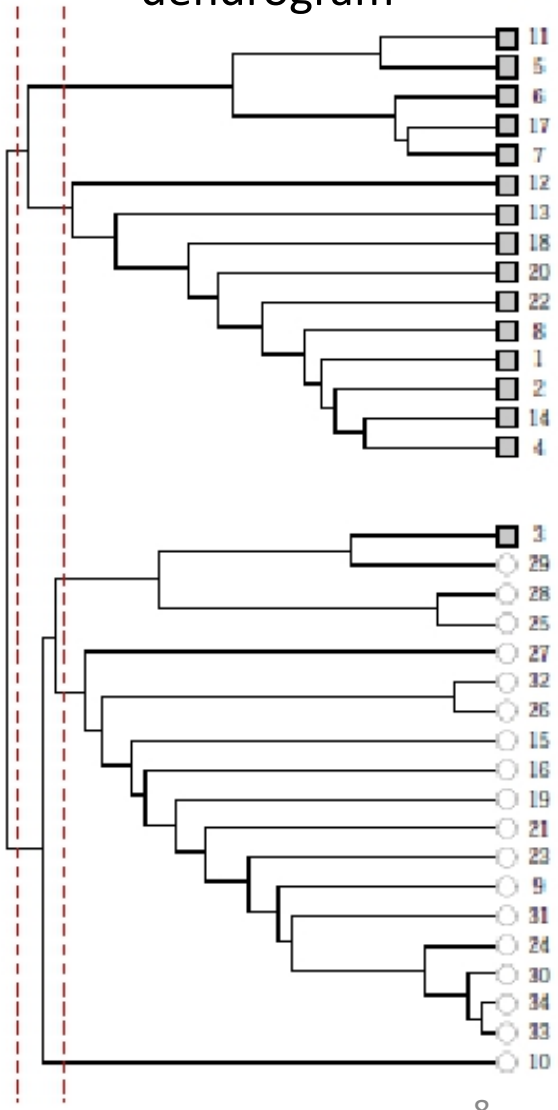
Communities in **physics collaborations**

Girvan-Newman: Results

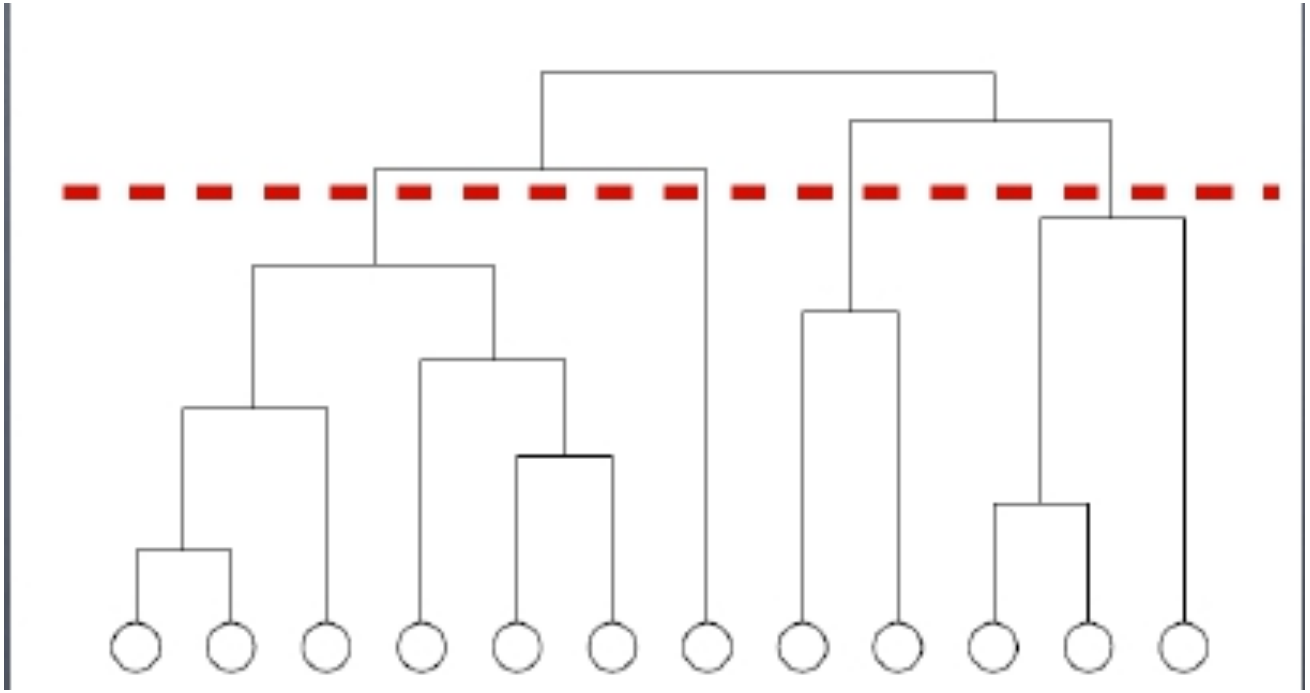
- **Sport club:**



dendrogram



How to select the number of communities?



- What is the **best** decomposition?
  - For each community, the cohesion within the community should be higher than outside.

The concept of Modularity

- A **measure** of how well a network is partitioned into communities.
- When a network is partitioned into **S** groups, the modularity, **Q,** is defined as:

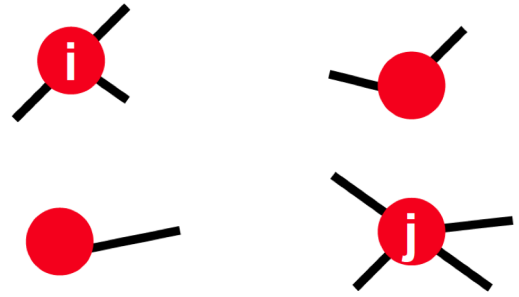
$$Q \propto \sum_{s \in S} [\ (\#\text{ edges within group } s) - (\text{expected } \#\text{ edges within group } s)\ ]$$

- We need a null model to determine the "expected" number of edges.

The Null Model

- Consider the Configuration model
  - It constructs a random surrogate network with the same degree distribution as a given real network.
- For nodes **i** and **j** with degrees $\mathbf{k_i}$ and $\mathbf{k_j}$, the probability of a link existing between them is

$$\frac{k_i k_j}{2m}$$



where **m** is the total number of edges.

Calculation of modularity

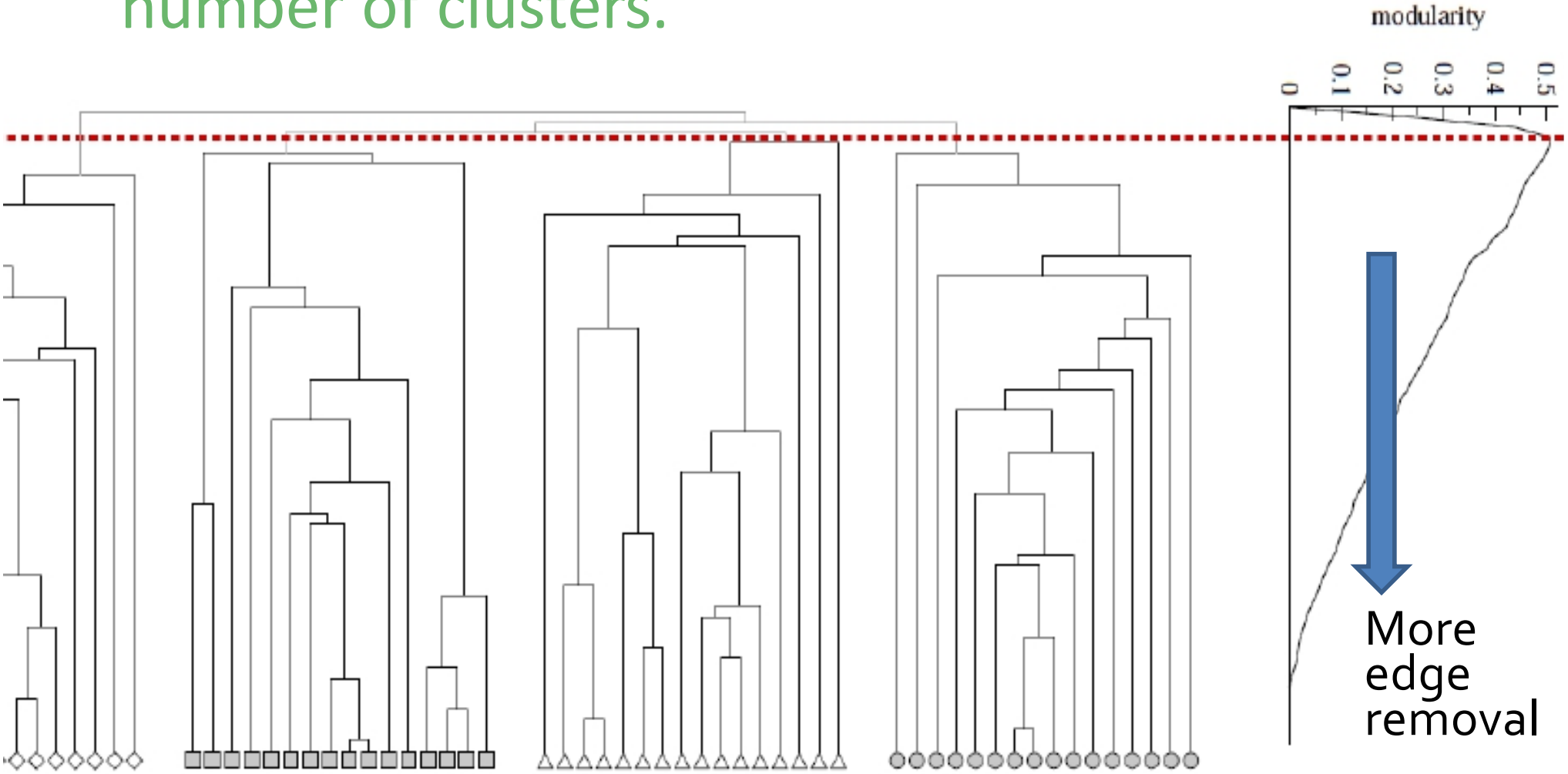
- Modularity of partitioning graph G into S communities:

$$Q(G,S) = \frac{1}{2m}\sum_{s\in S}\sum_{i\in s}\sum_{j\in s}\left(A_{ij} - \frac{k_i k_j}{2m}\right)$$

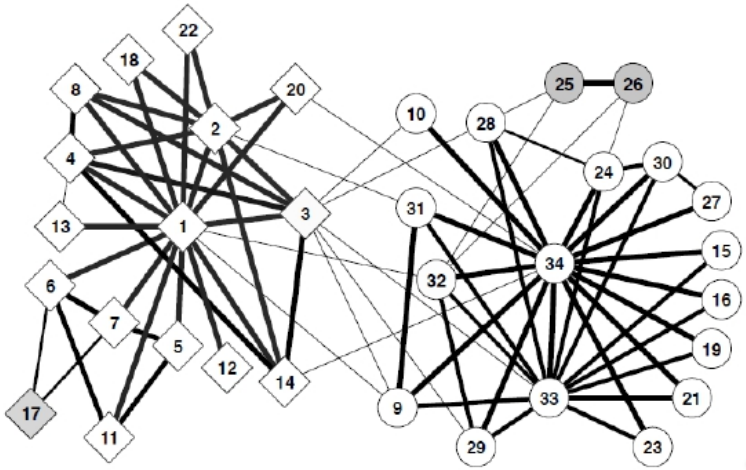
  - The adjacency matrix element $\mathbf{A_{ij}}$ = 1, if there is an actual edge between i and j; otherwise it is 0.
- Modularity lies in the range [-1, 1]
  - It is positive if the number of edges within groups exceeds the expected number.
  - Q >0.3 means significant community structure.

Modularity: Number of clusters

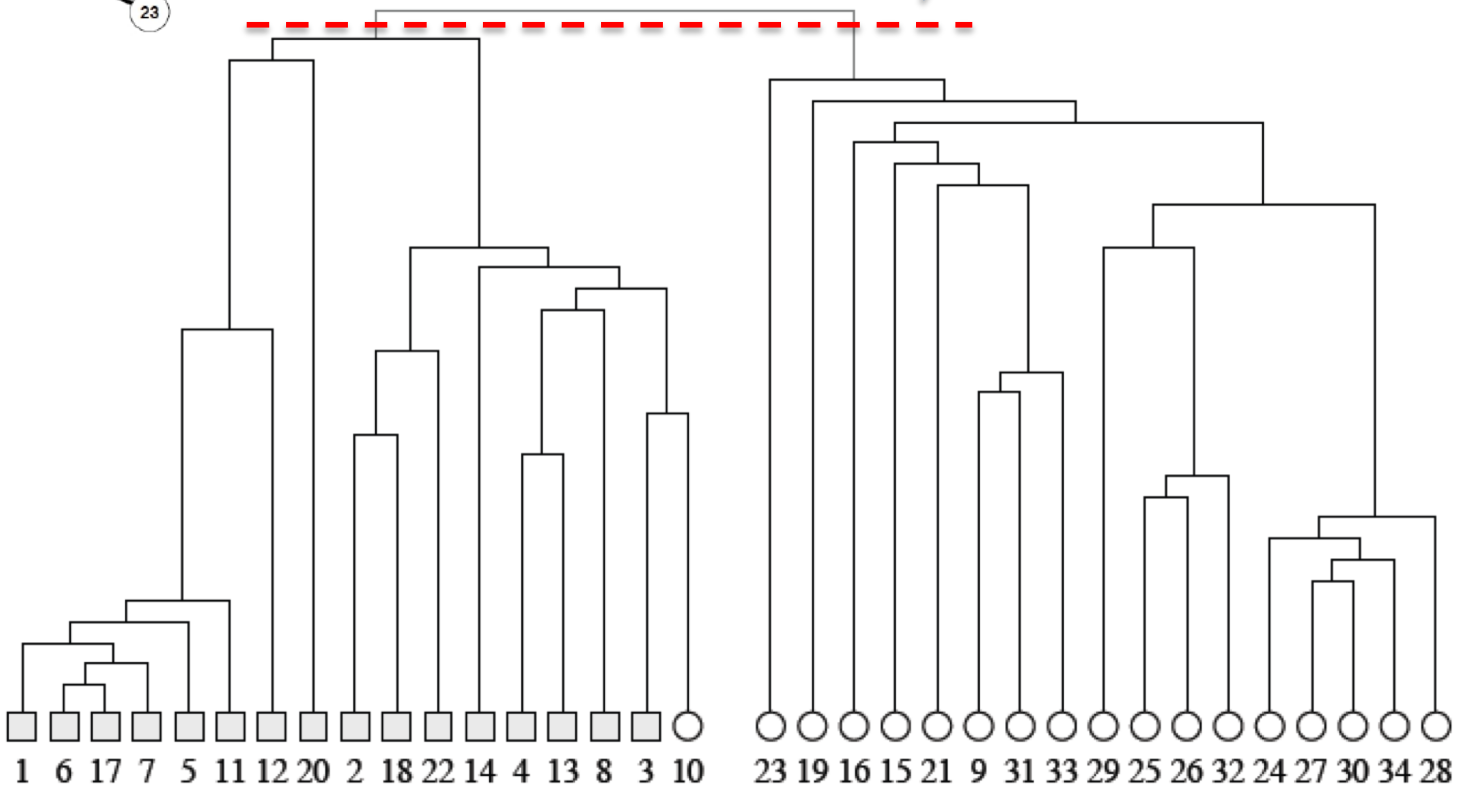
- Modularity is useful for selecting the “best” number of clusters.



Application to the Sport Club example



Mmaximum value of Q when
S=2, thus there are 2 communities



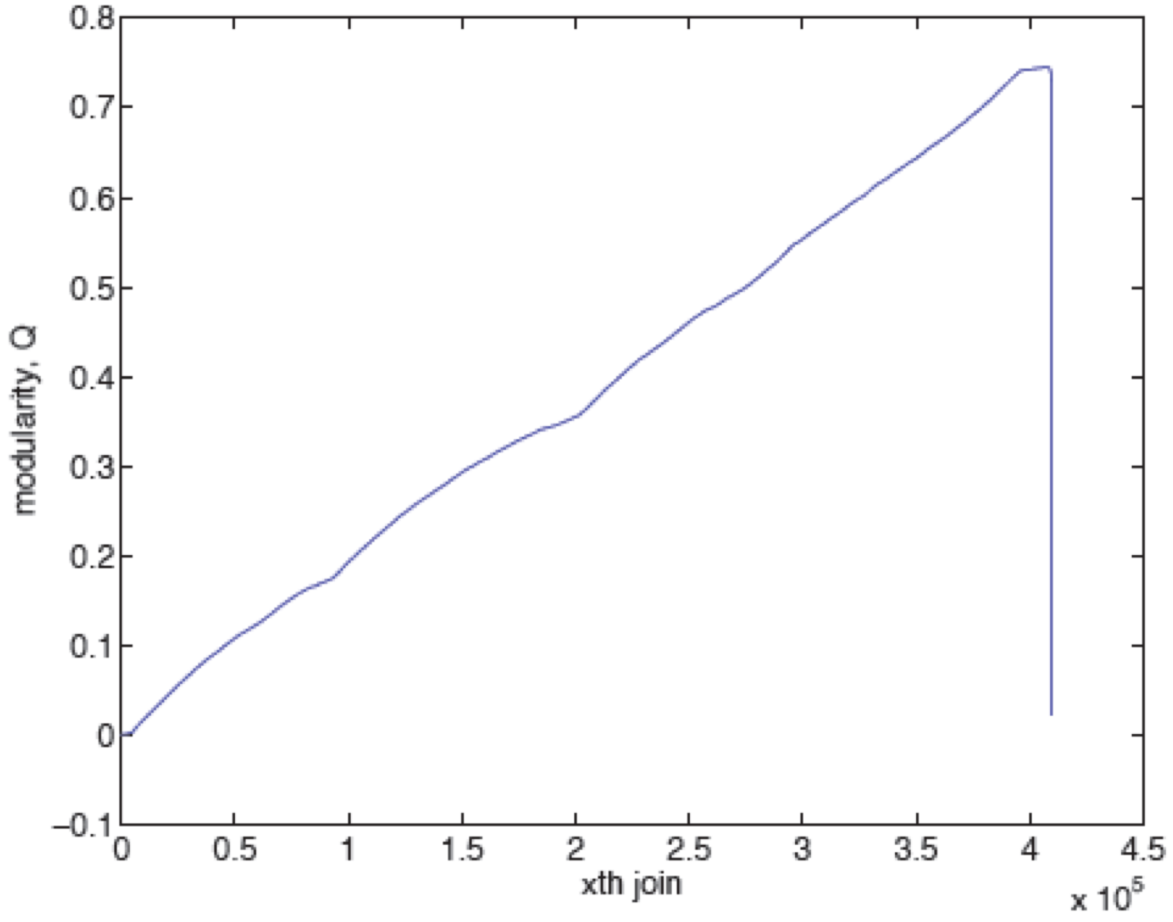
Fast Modularity

- The complexity is of calculating the Q(G, S) is O(m2n).
- Fast Modularity
  1. Start with a network of **n** communities of 1 node.
  2. Calculate ΔQ for all possible community pairs.
  3. Merge the pair of the largest increase in Q.
  4. Repeat (2)&(3) until one community remains.
  5. Cross cut the dendrogram where Q is maximum.

  – The complexity of this algorithm is O((m+ n)n).

Application to Amazon Recommendations

- Network of products: 200k nodes (products) and 2m links.
  - A link between products A and B if B was frequently purchased by buyers of A.
- Max Q=0.745, where the partition consists of 1,684 communities.



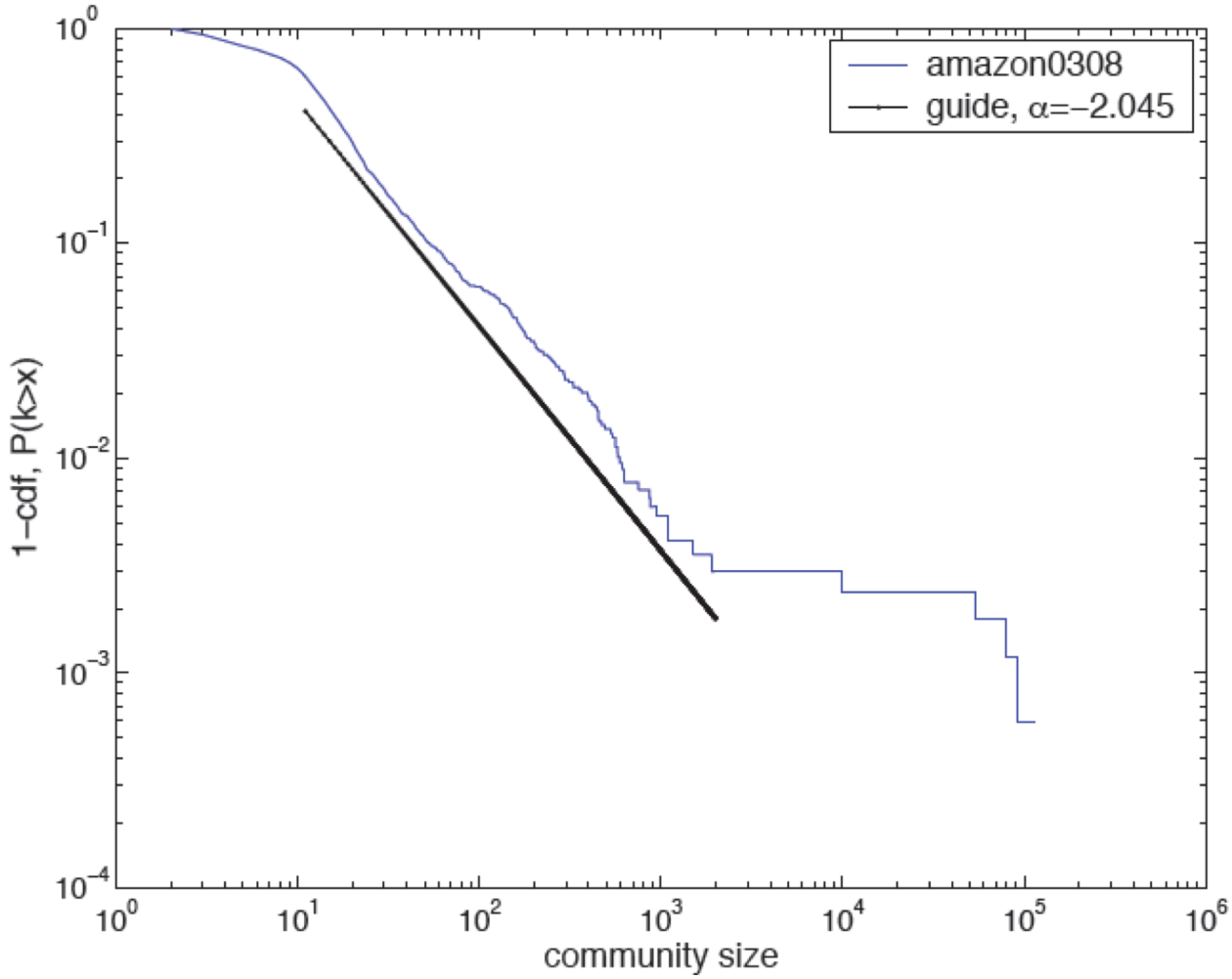
Amazon: Top Communities

- The 10 largest communities account for 87% of the nodes.

Rank	Size	Description
1	114538	General interest: politics; art/literature; general fiction; human nature; technical books; how things, people, computers, societies work, etc.
2	92276	The arts: videos, books, DVDs about the creative and performing arts
3	78661	Hobbies and interests I: self-help; self-education; popular science fiction, popular fantasy; leisure; etc.
4	54582	Hobbies and interests II: adventure books; video games/comics; some sports; some humor; some classic fiction; some western religious material; etc.
5	9872	classical music and related items
6	1904	children's videos, movies, music and books
7	1493	church/religious music; African-descent cultural books; homoerotic imagery
8	1101	pop horror; mystery/adventure fiction
9	1083	jazz; orchestral music; easy listening
10	947	engineering; practical fashion

Amazon community size distribution

- Power-law distribution.
  - 1,683 communities
  - Mean size of 243 nodes.

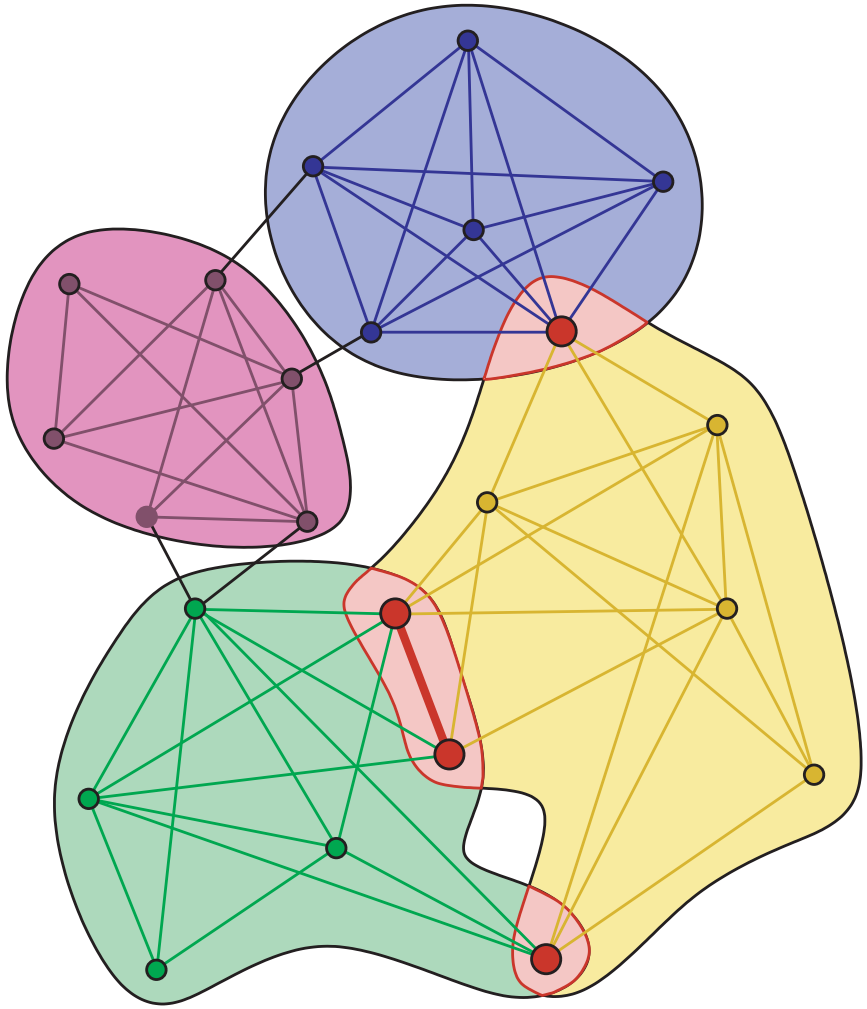


Limitations of Modularity

- Not a perfect measure
  - Depends on the number of links in network
  - Depends on links from a community to the rest of the network.
- What about weighted, directed network?
- What about overlapping community?

Overlapping Community

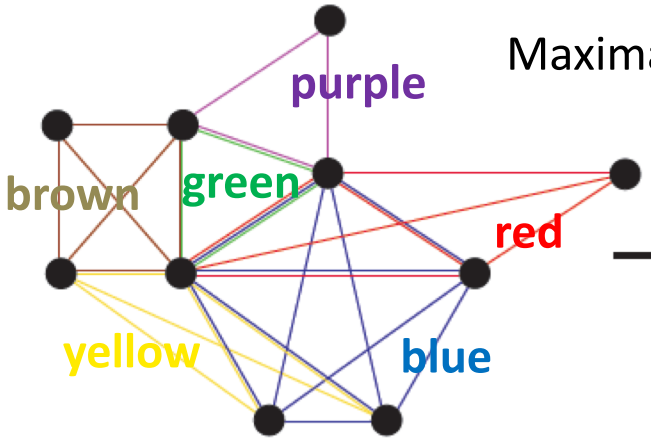
- Community membership could overlap
  - A node could be part of more than one community
- How to find them?



Clique Percolation Method - Algorithm

- Find all the maximal cliques
  - A maximal clique is a clique that cannot be extended by including one more adjacent node.
  - This is complex but real networks are sparse.
- Build ***clique-overlap-matrix***
  - Each clique is a node
  - Connect two k-cliques if they overlap at k-1 nodes
- Communities:
  - Connected components of the clique-overlap-matrix

Example



Maximal cliques

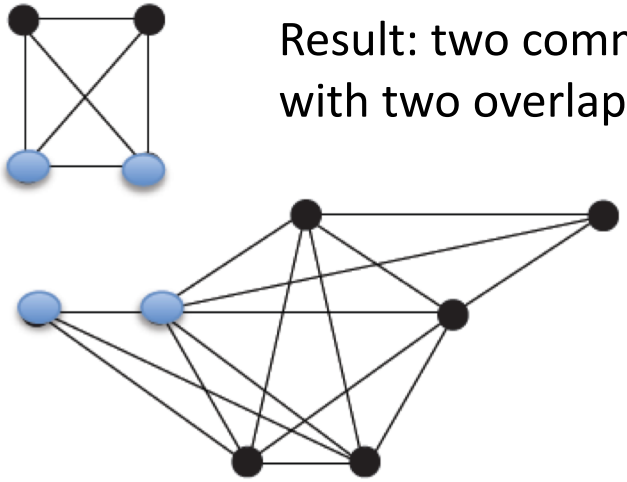
	5	3	2	1	3	1
	3	4	2	1	1	1
	2	2	3	2	1	2
	1	1	2	3	0	1
	3	1	1	0	4	2
	1	1	2	1	2	4

Clique-overlap-matrix: elements are number of overlapping nodes.

k=4 if we want communities With 4 or more nodes

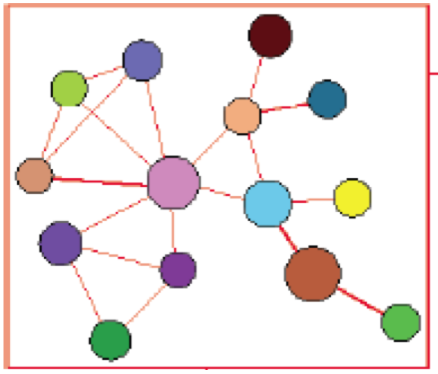
Erase elements <4 on diagonal and <3 elsewhere

	1	1	0	0	1	0
	1	1	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	1	0	0	0	1	0
	0	0	0	0	0	1

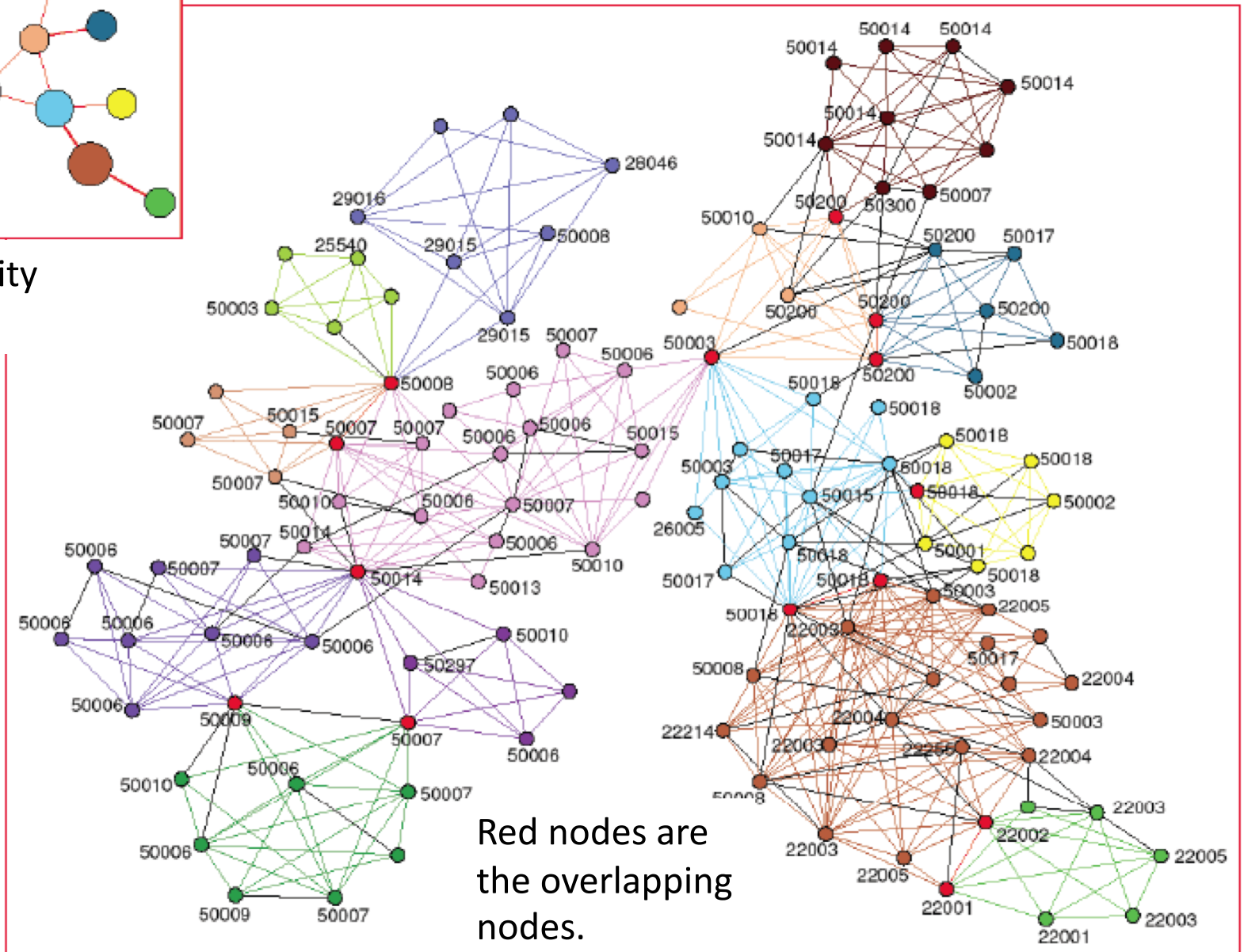


Result: two communities with two overlapping nodes

Application: phone call network



Community structure



Red nodes are the overlapping nodes.

Research on network communities

- A very hot research area in network science
- Researchers have proposed MANY
  - Definitions of different types of communities
  - Methods to detect communities faster
  - Algorithms to calculate modularity
  - Applications of network communities
  - …

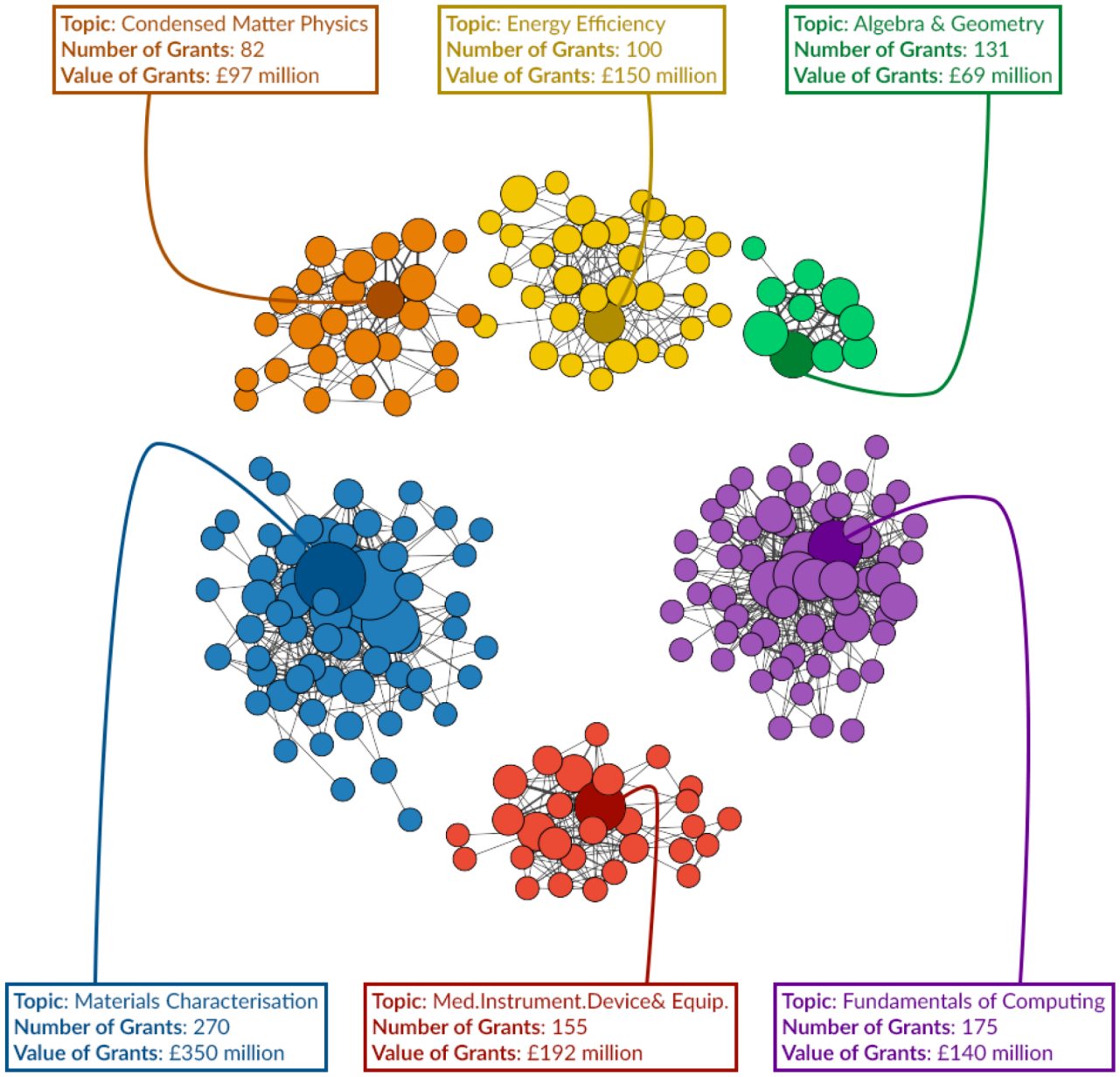
Reference

- **Community structure in social and biological networks.**
  - Michelle Girvan and Mark Newman. PNAS 99 (12) 7821–7826, 2002.
- **Finding and evaluating community structure in networks.**
  - Mark Newman and Michelle Girvan. Phys. Rev. E 69 (026113), 2004.
- **Social Features of Online Networks: The strength of intermediary ties in online social media**
  - A. Przemyslaw et al. PLoS ONE 7 (1) e29358, 2012.

Examples: Two MSc WSBDA projects related to network community detection

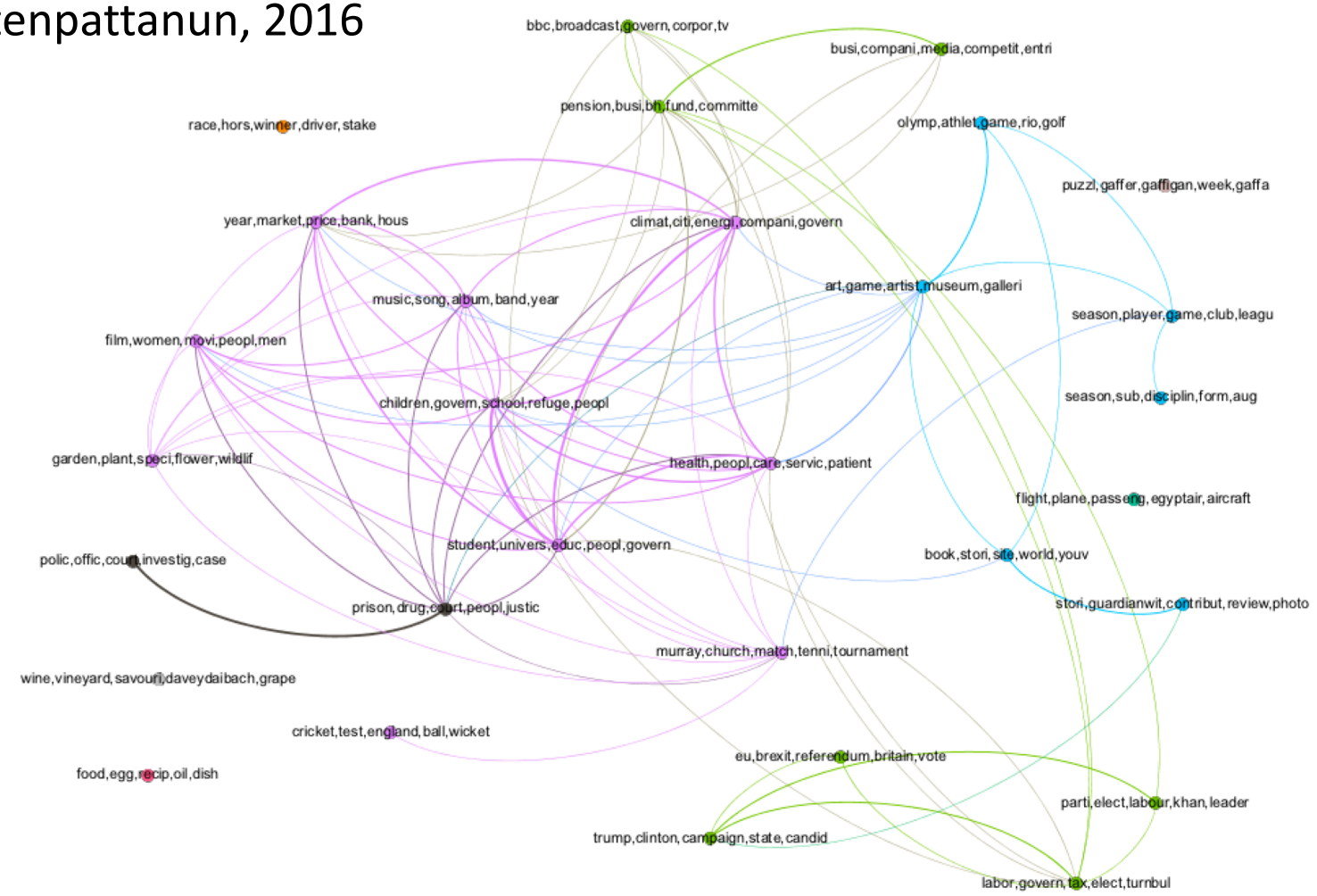
MSc project 1

- Clustering of EPSRC research topics and researchers: using a network analysis approach based on grant data
  - S. Tripon, 2016



MSc project 2

- Detection of the structure of news story topics of The Guardian Newspaper website
  - P. Utenpattanun, 2016




Thank you!