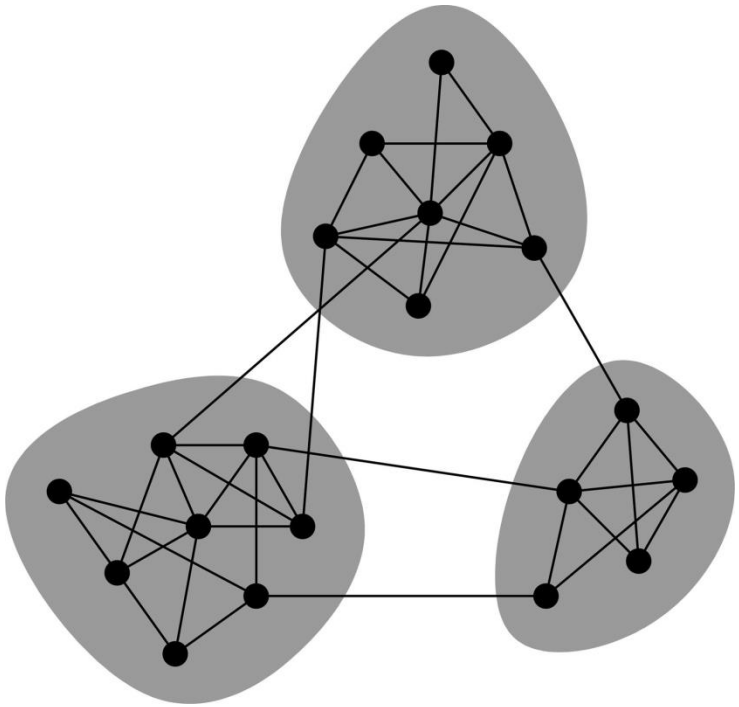
Subgroups

- also called clusters, communities

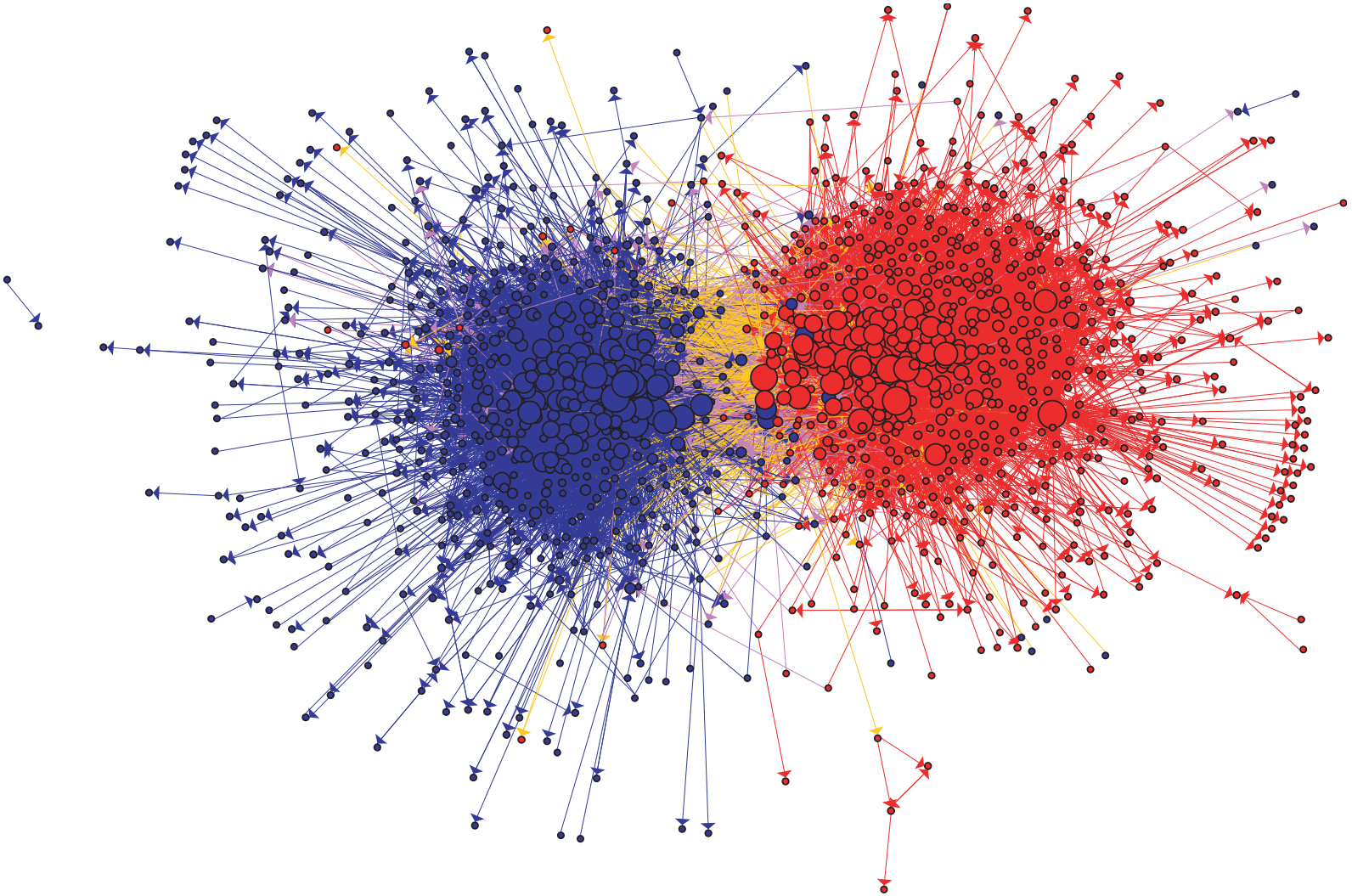
Groups of actors who are more closely related to each other within a group than actors outside the group.



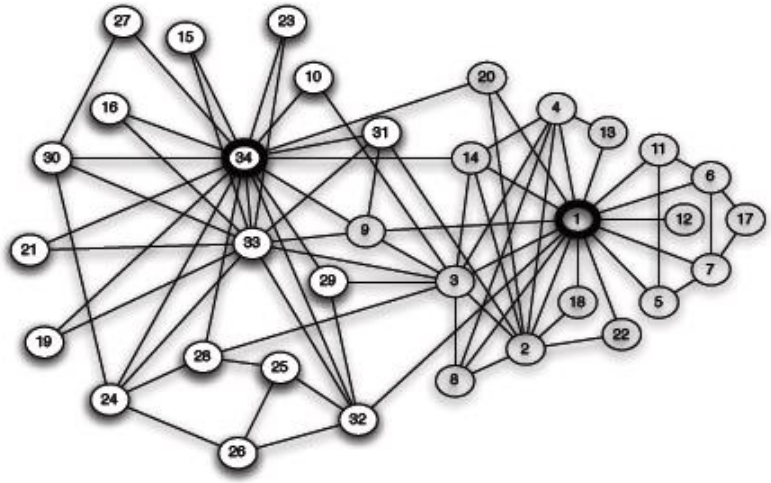
Why do we care about subgroups?

- Suggest certain social processes (homophily…)
- Explain certain outcomes or predict (Karate club)
- Useful for practical purposes (group leaders/bridges)
- Simplify large networks for visualization/further analysis

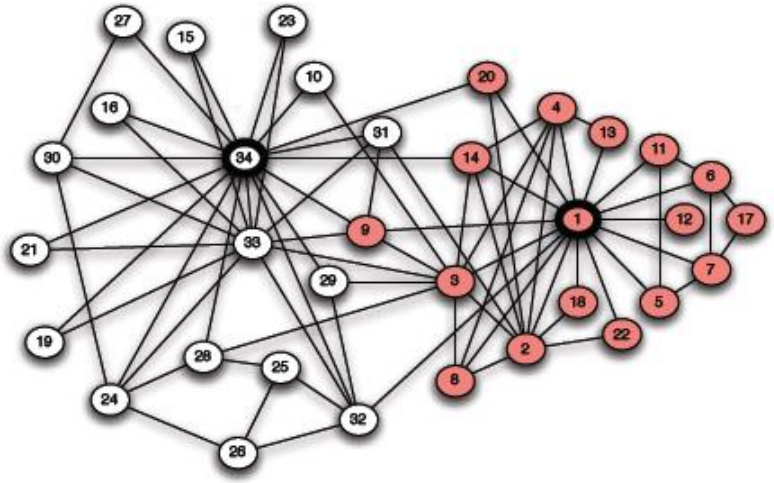
Political blogs – Lada Adamic and Natalie Glance



Zachary Karate Club: Easley/Kleinberg

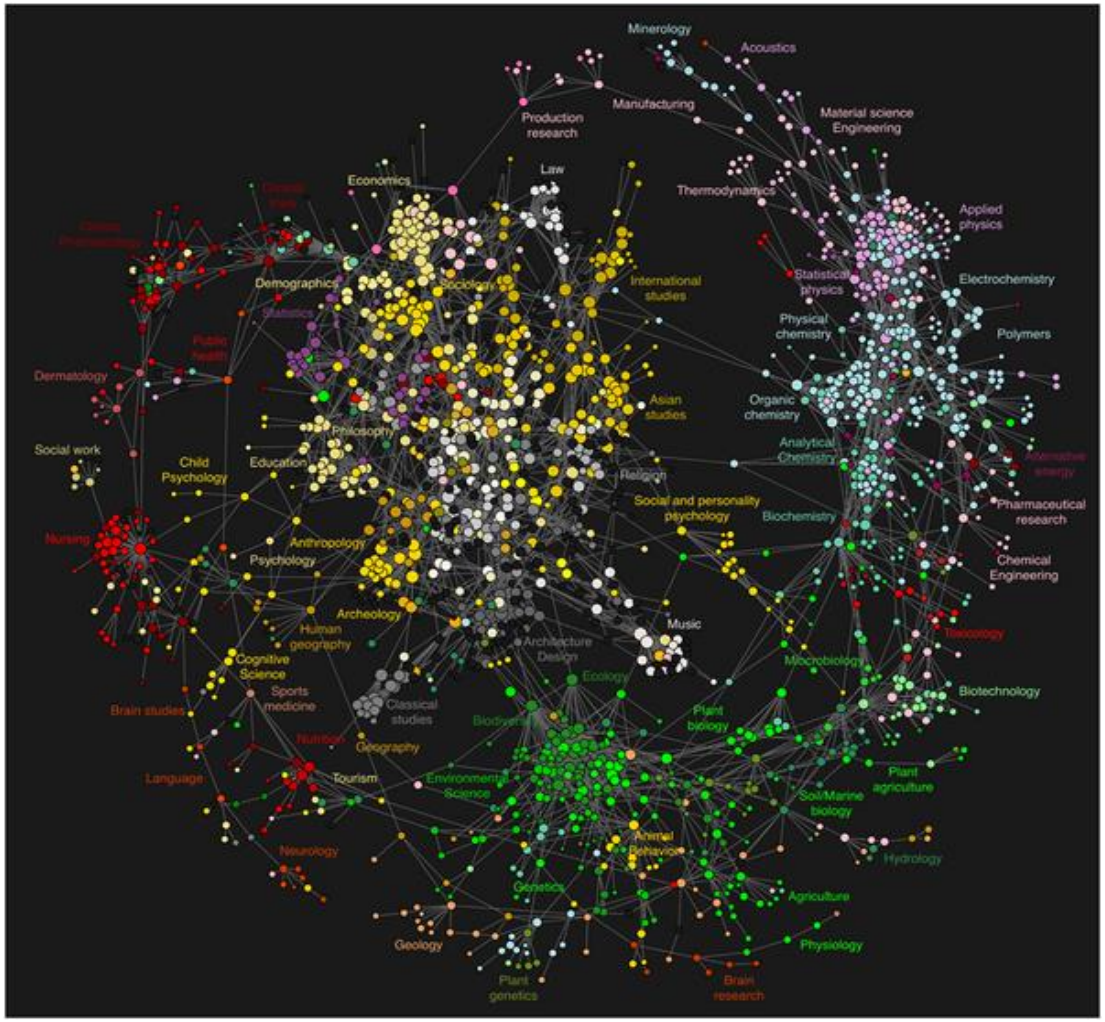


(a) *Karate club network*



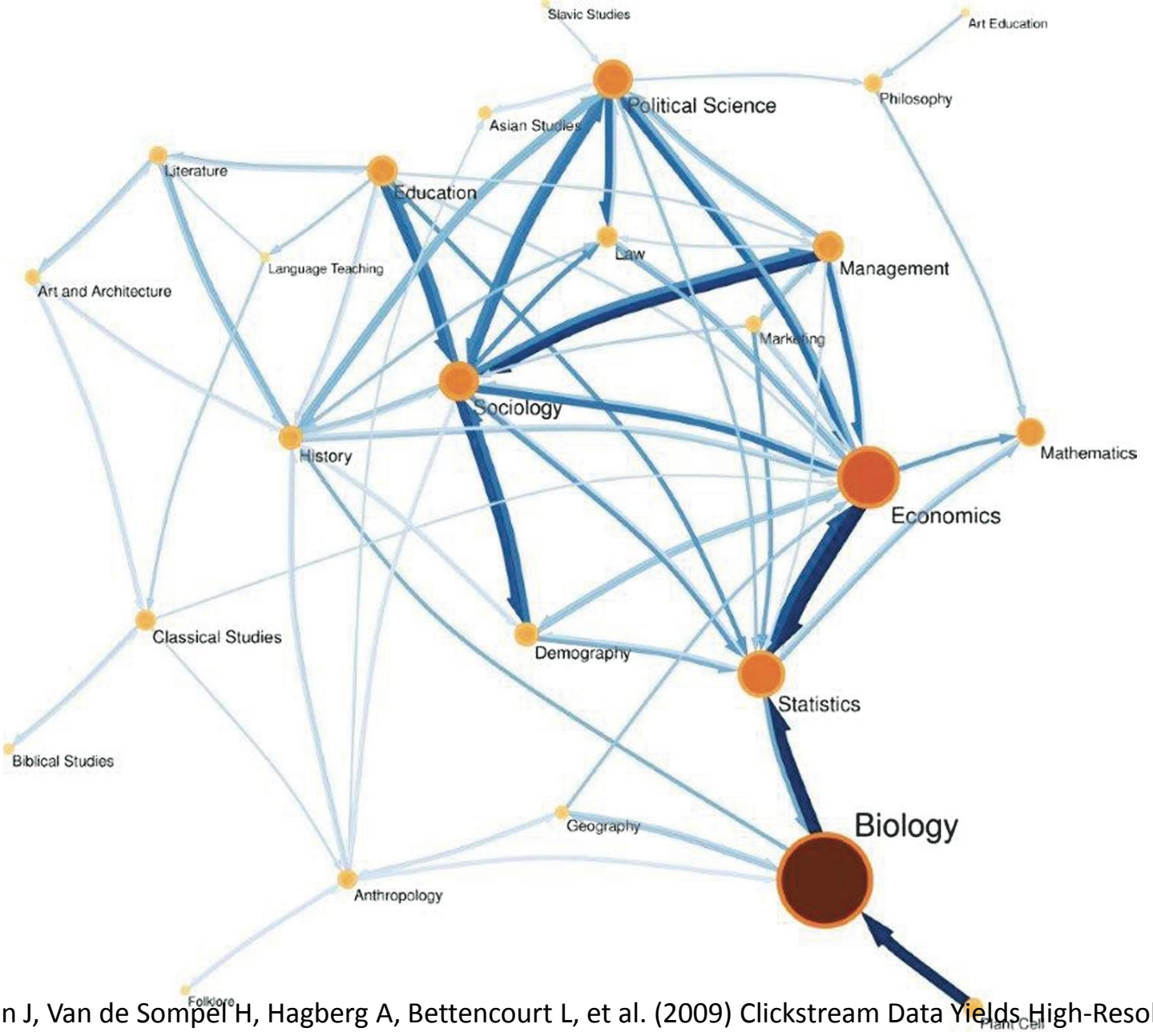
(b) *After a split into two clubs*

Map of science derived from clickstream data



Bollen J, Van de Sompel H, Hagberg A, Bettencourt L, et al. (2009) Clickstream Data Yields High-Resolution Maps of Science. PLoS ONE 4(3): e4803. doi:10.1371/journal.pone.0004803
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0004803

PLOS | ONE



Bollen J, Van de Sompel H, Hagberg A, Bettencourt L, et al. (2009) Clickstream Data Yields High-Resolution Maps of Science. PLoS ONE 4(3): e4803. doi:10.1371/journal.pone.0004803
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0004803

Goal

- Find meaningful, robust groups that represent the network structure inherent in the data.

Approaches to identify subgroups

- **Clique-based approach** that starts with finding cliques and often requires further analysis of overlapping cliques.
  - hierarchical clustering
  - Two-mode analysis
- **Use algorithms** that capture subgroup properties
  - Girvan-Newman method (based on edge betweenness)
  - Modularity-based algorithms
  - Factions

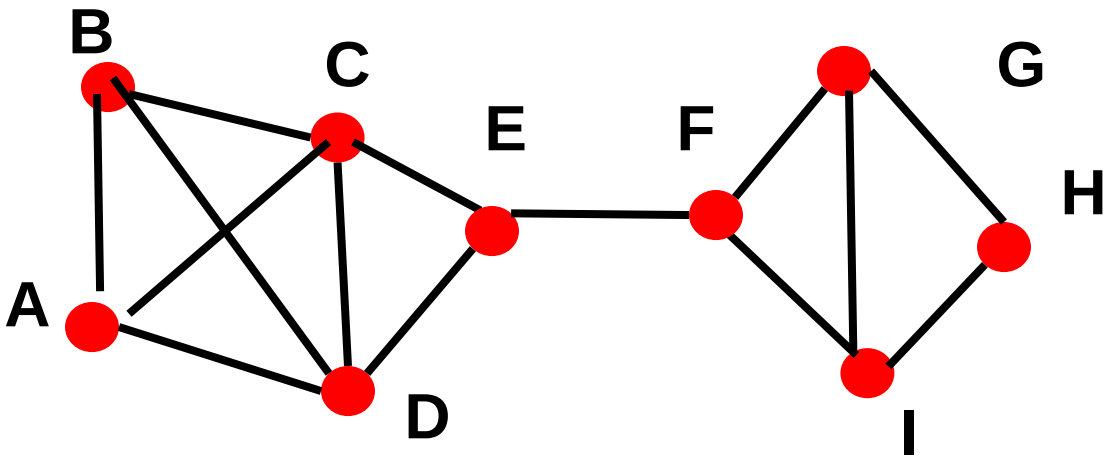
Cliques

- Maximal complete subgraph

  - A sub-network in which every node is **directly** connected every other node

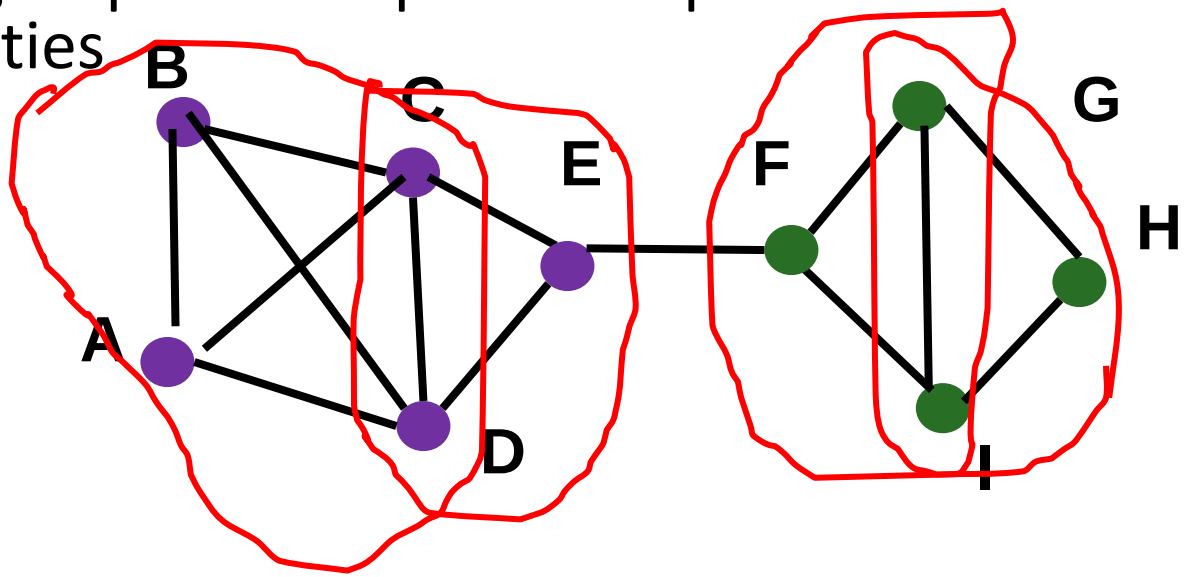
  - This would no longer hold if adding another node

- cliques have at least three members



Issues with cliques

- Two strict, not very interesting
  - everybody is connected to everybody else
  - no core-periphery structure
- Not robust
  - one missing link can disqualify a clique
- Do not represent natural communities
- Can overlap
- Analyzing clique overlaps can help find natural communities



Secondary analyses of overlapping cliques

Hierarchical clustering

Look at “Clique co-membership matrix”

- A matrix of actors which records how many cliques each pair of actors have been together in. Implemented in UCInet.

Two-mode analysis

Look at “Clique participation matrix”

- A two-mode matrix which records the extent to which an actor participates in a clique.

Analyzing Cliques: the bank wiring room game social network

Cliques:

1.I1 W1 W2 W3 W4
2.W1 W2 W3 W4 S1
3.W1 W3 W4 W5 S1
4.W6 W7 W8 W9
5.W7 W8 W9 S4

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
		I1	I3	W1	W2	W3	W4	W5	W6	W7	W8	W9	S1	S2	S4
1	I1	0	0	1	1	1	1	0	0	0	0	0	0	0	0
2	I3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	W1	1	0	0	1	1	1	1	0	0	0	0	1	0	0
4	W2	1	0	1	0	1	1	0	0	0	0	0	1	0	0
5	W3	1	0	1	1	0	1	1	0	0	0	0	1	0	0
6	W4	1	0	1	1	1	0	1	0	0	0	0	1	0	0
7	W5	0	0	1	0	1	1	0	0	1	0	0	1	0	0
8	W6	0	0	0	0	0	0	0	0	1	1	1	0	0	0
9	W7	0	0	0	0	0	0	1	1	0	1	1	0	0	1
10	W8	0	0	0	0	0	0	0	1	1	0	1	0	0	1
11	W9	0	0	0	0	0	0	0	1	1	1	0	0	0	1
12	S1	0	0	1	1	1	1	1	0	0	0	0	0	0	0
13	S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	S4	0	0	0	0	0	0	0	0	1	1	1	0	0	0

Analyzing Cliques: the bank wiring room game social network

Clique participation matrix: reflects the extent to which an actor participates in a clique.

- A node in a clique => 1
- A node has no edge to a clique => 0
- Other cases: divide the number of edges to the nodes in a clique by the number of edges it'd require for a node to become a member of the clique.

	1	2	3	4	5
I1	1.000	0.800	0.600	0.000	0.000
I3	0.000	0.000	0.000	0.000	0.000
W1	1.000	1.000	1.000	0.000	0.000
W2	1.000	1.000	0.800	0.000	0.000
W3	1.000	1.000	1.000	0.000	0.000
W4	1.000	1.000	1.000	0.000	0.000
W5	0.600	0.800	1.000	0.250	0.250
W6	0.000	0.000	0.000	1.000	0.750
W7	0.000	0.000	0.200	1.000	1.000
W8	0.000	0.000	0.000	1.000	1.000
W9	0.000	0.000	0.000	1.000	1.000
S1	0.800	1.000	1.000	0.000	0.000
S2	0.000	0.000	0.000	0.000	0.000
S4	0.000	0.000	0.000	0.750	1.000

Cliques:

1. I1 W1 W2 W3 W4
2. W1 W2 W3 W4 S1
3. W1 W3 W4 W5 S1
4. W6 W7 W8 W9
5. W7 W8 W9 S4

Analyzing Cliques: the bank wiring room game social network

Visualization of clique participation matrix

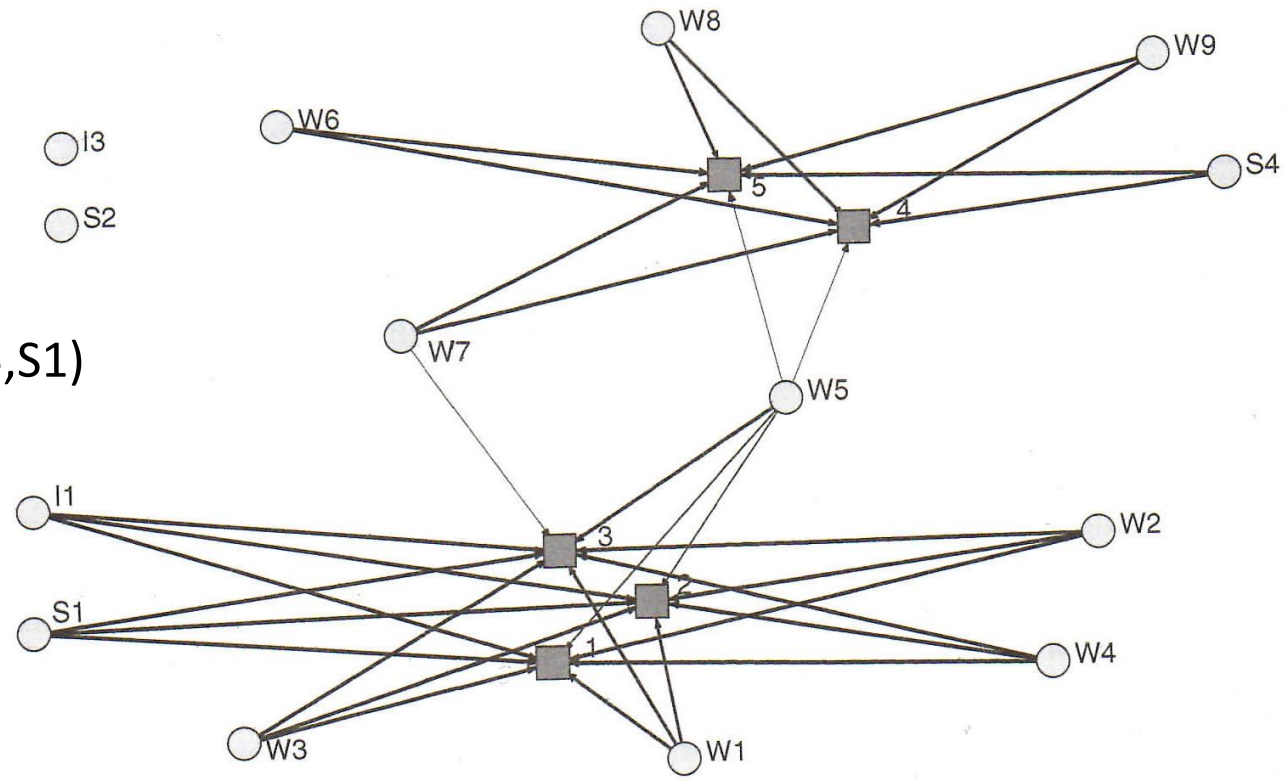
- Two large groups:

(I1,W2,W1,W3,W4,W5,S1)

(W6,W7,W8,W9,S4)

- Two isolates:

S2, I3



Analyzing Cliques: the bank wiring room game social network

Clique co-membership matrix: records how many cliques each pair of actors have been together in.

Cliques:

1.I1 W1 W2 W3 W4
2.W1 W2 W3 W4 S1
3.W1 W3 W4 W5 S1
4.W6 W7 W8 W9
5.W7 W8 W9 S4

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
		I1	I3	W1	W2	W3	W4	W5	W6	W7	W8	W9	S1	S2	S4
1	I1	1	0	1	1	1	1	0	0	0	0	0	0	0	0
2	I3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	W1	1	0	3	2	3	3	1	0	0	0	0	2	0	0
4	W2	1	0	2	2	2	2	0	0	0	0	0	1	0	0
5	W3	1	0	3	2	3	3	1	0	0	0	0	2	0	0
6	W4	1	0	3	2	3	3	1	0	0	0	0	2	0	0
7	W5	0	0	1	0	1	1	1	0	0	0	0	1	0	0
8	W6	0	0	0	0	0	0	0	1	1	1	1	0	0	0
9	W7	0	0	0	0	0	0	0	1	2	2	2	0	0	1
10	W8	0	0	0	0	0	0	0	1	2	2	2	0	0	1
11	W9	0	0	0	0	0	0	0	1	2	2	2	0	0	1
12	S1	0	0	2	1	2	2	1	0	0	0	0	2	0	0
13	S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	S4	0	0	0	0	0	0	0	0	1	1	1	0	0	1

Analyzing Cliques: the bank wiring room game social network

Hierarchical clustering of clique co-membership matrix

At the 0.381 level:

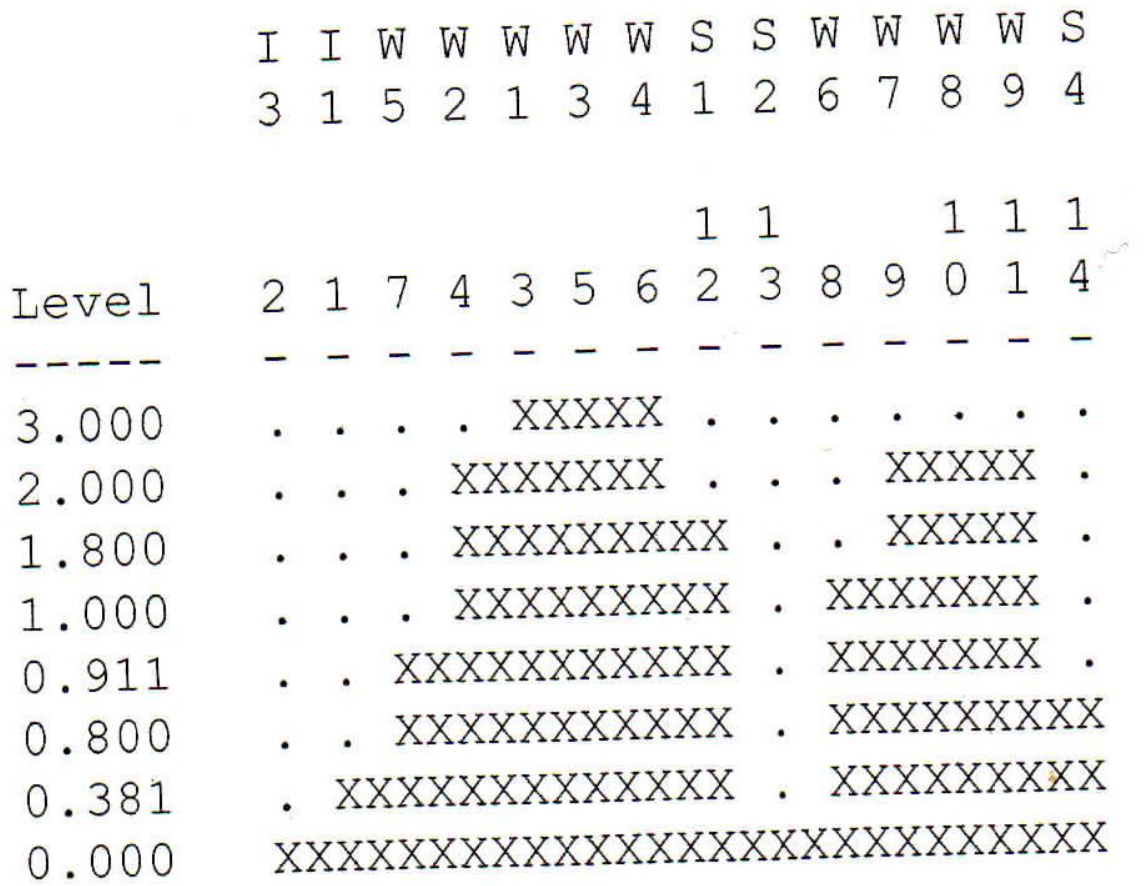
- Two large groups:

(I1,W2,W1,W3,W4,W5,S1)

(W6,W7,W8,W9,S4)

- Two isolates:

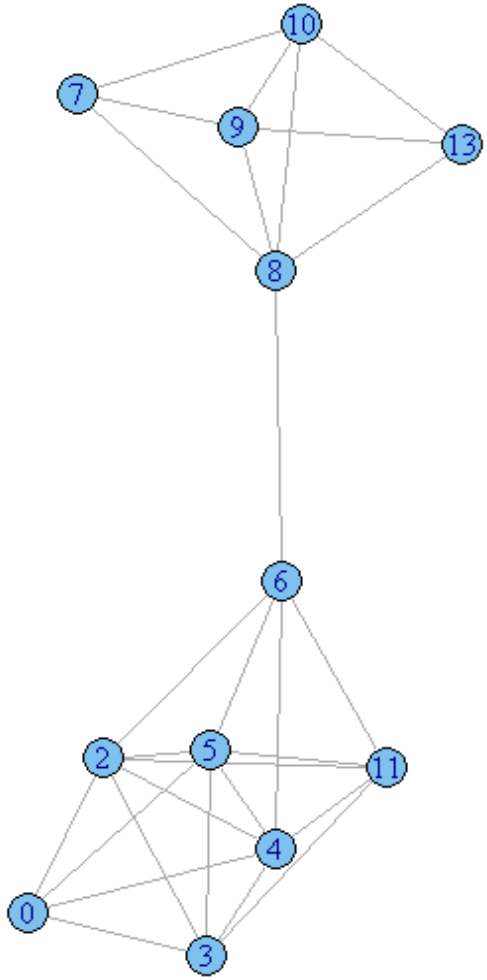
S2, I3



Analyzing Cliques: the bank wiring room game social network

Cliques:

1. 0 2 3 4 5
2. 2 3 4 5 11
3. 2 4 5 6 11
4. 7 8 9 10
5. 8 9 10 13



Find Cliques in iGraph

cliques(graph, min=NULL, max=NULL)
maximal.cliques(graph, min=NULL, max=NULL, subset=NULL, file=NULL)
maximal.cliques.count(graph, min=NULL, max=NULL, subset=NULL)

Other Definitions

- Many relaxations of the clique concept
  - k-clique, n-clan, n-club, K-plex, k-core
- In real data most are of no use

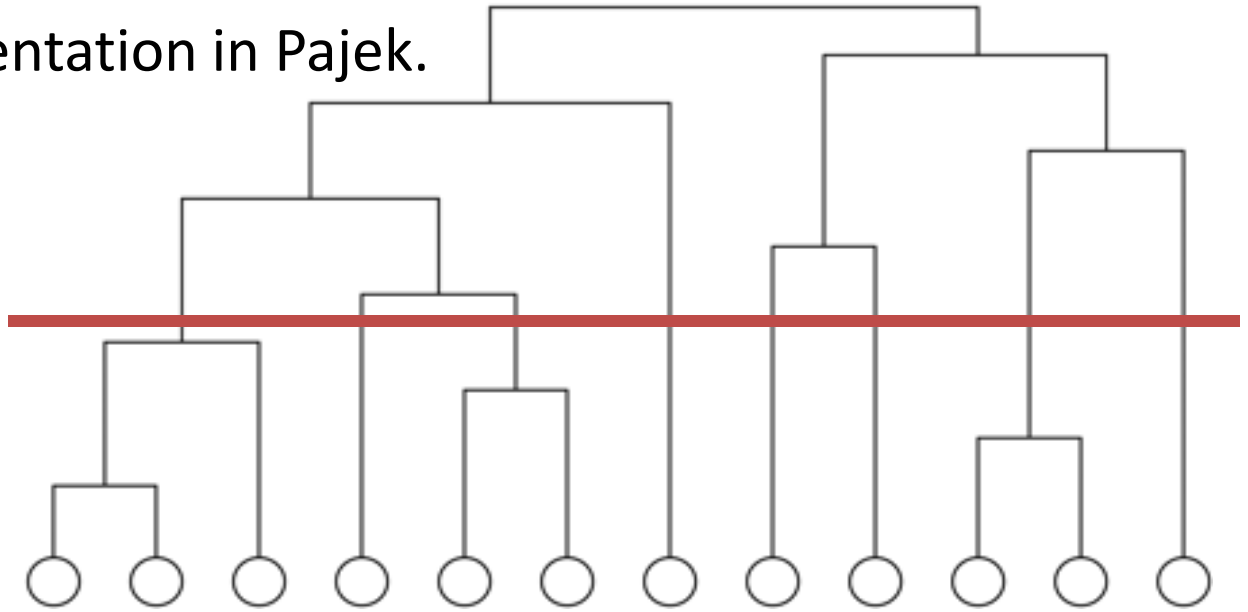
Hierarchical Clustering

- Clique co-membership matrix is one way to characterize similarity/closeness between nodes in a network.
- More generally, we can calculate closeness for all pairs of nodes using some kind of definition to produce a similarity/distance matrix and then apply hierarchical clustering.

Hierarchical Clustering

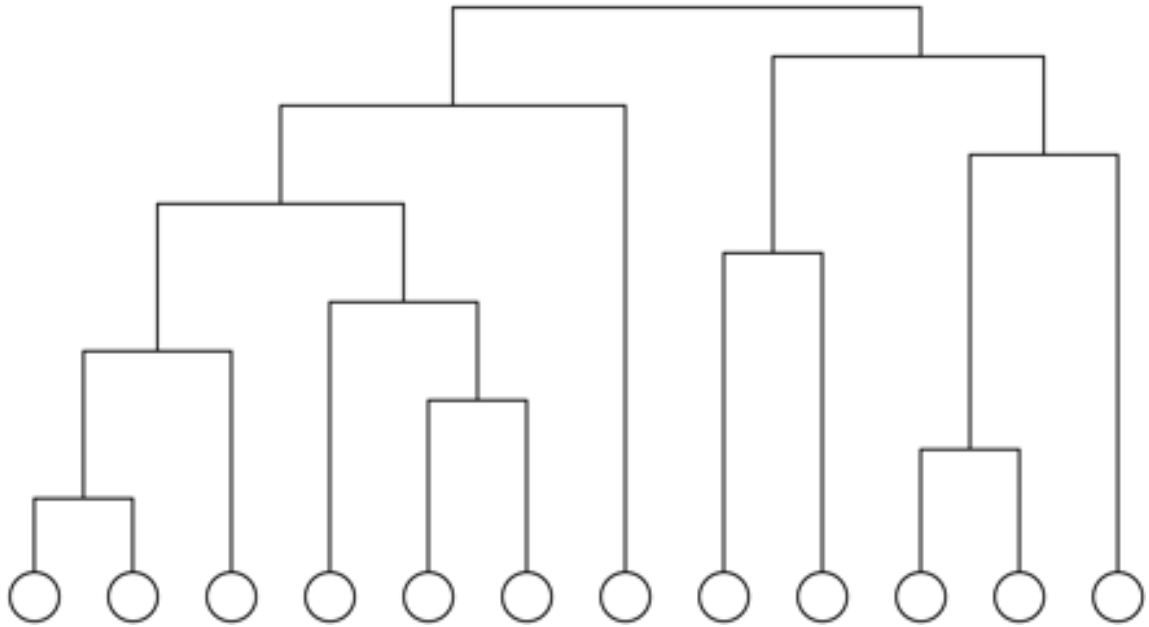
- Start with all nodes as individual clusters
- Join clusters each time based on similarity/closeness between nodes/clusters
- This process creates a hierarchy of nested clusters
- We can then obtain partitions, either by specifying the number of clusters or a level of similarity/closeness desired.

Nice implementation in Pajek.



Something to think about

What does it mean if a node joins a cluster late in the process?



Girvan-Newman Algorithm

- Based on edge betweenness centrality:

  The number of times an edge lies on a geodesic path between a pair of nodes.

- General Idea: The edges that connect sub-groups will have high betweenness. By removing these edges, the groups will be separated from one another and the underlying community structure of the network can be revealed.

- Use a measure “modularity” to evaluate how well the detected communities are.

  - M Newman and M Girvan: Finding and evaluating community structure in networks, Physical Review (2004).

Modularity

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w)$$

- a property of a network and a partition of the network.
- compares the number of numerical links in the groups to how many you would expect to see if they were distributed at random to groups.
- for a random network, Q = 0: the number of edges within a community is no different from what you would expect.
- higher values => the algorithm has found more significant groupings.
- a value above 0.3 is a good indicator of significant community structure in a network.

- M Newman and M Girvan: Finding and evaluating community structure in networks, Physical Review (2004).

Girvan-Newman Algorithm

In iGraph:

- edge.betweenness.community(): create a hierarchy of clusters based on edge betweenness.
- community.to.membership(): obtain a partition given the number of clusters.
- modularity(): to find the modularity of a given partition.

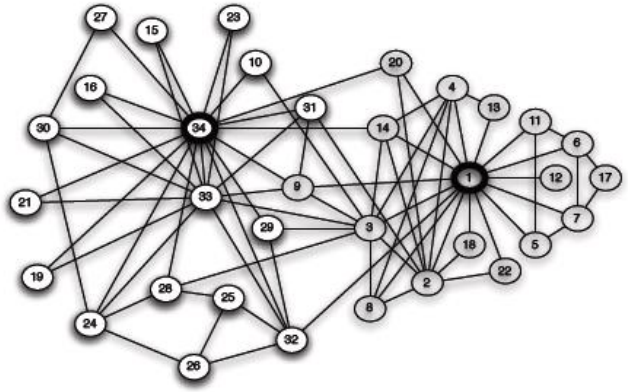
M Newman and M Girvan: Finding and evaluating community structure in networks, Physical Review (2004).

Girvan-Newman Algorithm

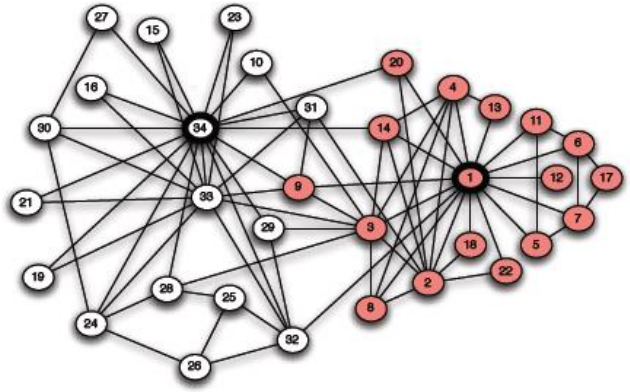
How to decide the number of communities?

- Create different partitions and compare Q scores.

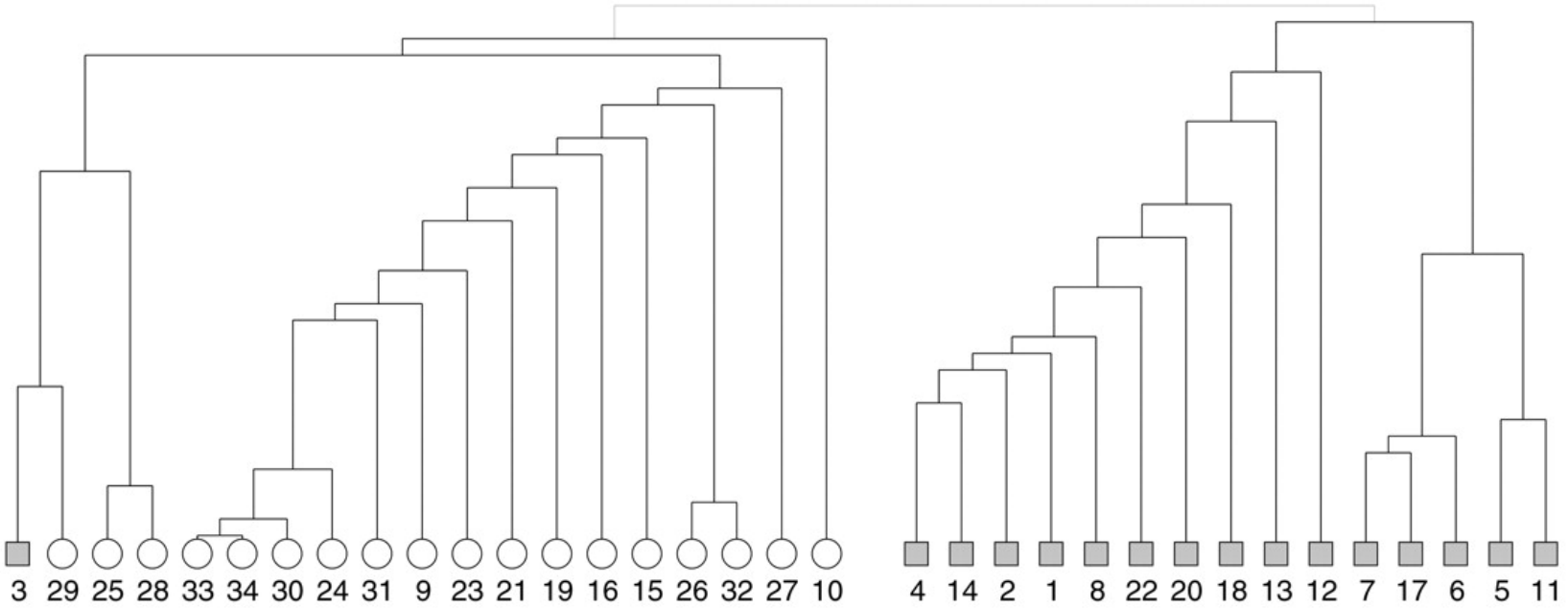
Applying Girvan-Newman Algorithm to the karate club



(a) *Karate club network*



(b) *After a split into two clubs*



- Girvan and Newman, PNAS June 11, 2002 99(12):7821-7826

Something to think about

Can Betweenness clustering algorithms be applied to directed networks?

Betweenness clustering algorithm

- Computationally expensive.
- Up to a few hundred nodes.

Modularity as a Framework for Community Detection

"Greedy" Algorithm

- start with all nodes as individual groups
- Repeatedly join groups so to achieve the maximum increase DQ in modularity
- Till DQ can no more increase from joining any two groups

fastgreedy.community() in iGraph

- A Clauset, MEJ Newman, C Moore - Finding community structure in very large networks. **Phys. Rev. E 70, (2004)**

- Determine the optimal number of natural communities.
- Used to find communities in large networks.
- Has a tendency to produce super-communities that contain a large fraction of the nodes, even on synthetic networks that have no significant community structure.
- May produce values of modularity that are significantly lower than what can be found by using other methods.

Modularity as a Framework for Community Detection

Modularity in Gephi

- Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre - Fast unfolding of communities in large networks. *J. Stat. Mech.* (2008)

- A heuristic method that is based on modularity optimization.
- Determine the optimal number of natural communities.
- Extremely fast and used to find communities in very large networks.
- Mitigate the issue of modularity optimization approach that fails to identify communities smaller than a certain scale.

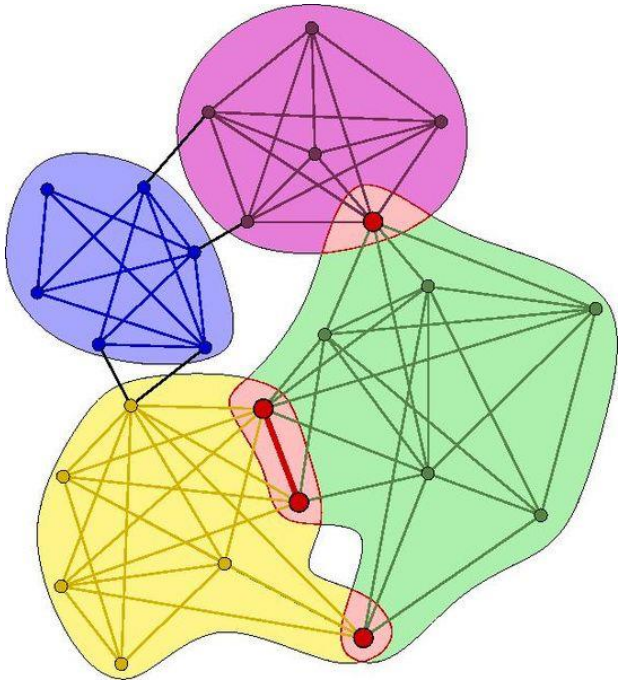
Overlapping communities

In real world networks, communities are often overlapping.

- Statistical Properties of Community Structure in Large Social and Information Networks by J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. International World Wide Web Conference (WWW), 2008.

A free software package for finding overlapping communities of large networks: **http://cfinder.org**

- G. Palla, I. Derényi, I. Farkas, and T. Vicsek - Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818 (2005)



Directed Data

For directed data, the maximum cohesion can only be achieved when every tie is reciprocated. The approach to directed data is therefore often to simply symmetrize the data first to produce an undirected network of reciprocated ties - consider undirected network of reciprocated ties.

Valued Data

Cliques	In principle, can be reformulated, but in practice no generally accepted definition. Usually dichotomize the data.
Clustering	Can be directly applied
Girvan-Newman	Can be applied
Other modularity based algorithms	Some can; some cannot

Summary

To detect meaningful, robust subgroups, you may start with visually examining the structure and components of a network and try different methods.

- **Clique-based approach** that starts with finding cliques (maximal complete subgraphs) and often requires further analysis of overlapping cliques.

- **Hierarchical clustering**
- **Edge betweenness** clustering algorithms
- **Modularity** based algorithms

Assignment – due on Feb 12 before class

In 1948, American sociologists executed a large field study in the Turrialba region, which is a rural area in Costa Rica (Latin America). They were interested in the impact of formal and informal social systems on social change. Among other things, they investigated visiting relations between families living in haciendas (farms) in a neighborhood called Attiro. The network of visiting ties is a simple directed graph: each arc represents "frequent visits" from one family to another.

The investigators proposed an ethnographic classification of the families into six family-friendship groupings on substantive criteria. In rural areas where there is little opportunity to move up and down the social ladder social groups are usually based on family relations.

We would like to reconstruct the way the families were assigned to family-friendship groupings. You may use different community detection methods (at least try Girvan-Newman algorithm in iGraph and the modularity community detection algorithm in Gelphi) to find communities. Discuss which methods find groups that best match the family-friendship groupings. Do you think researchers used additional information to assign families to the groupings?

Download the dataset from the course website:

- Attiro.net: The network of visiting ties.
- Attiro_grouping.csv: ethnographic family-friendship groupings.

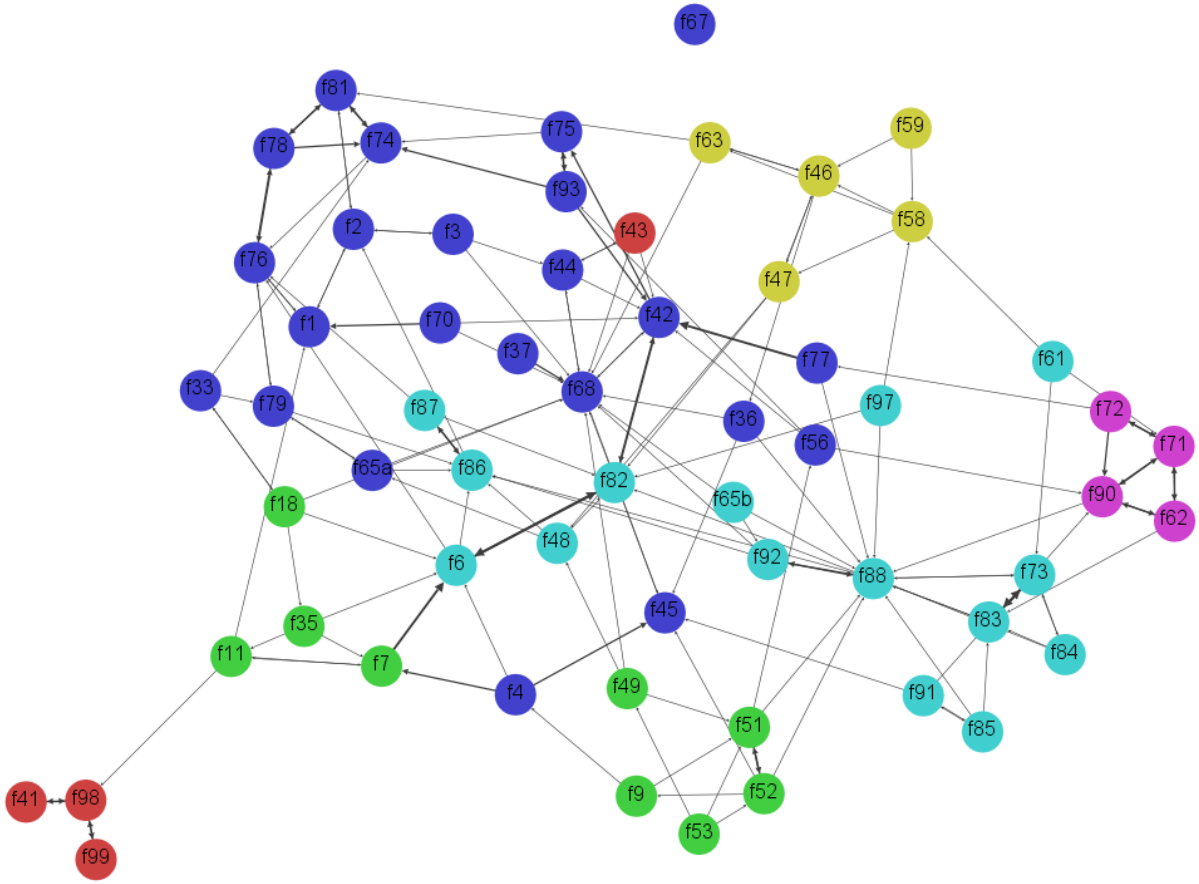
You can also download a file that includes **useful iGraph commands**

Tutorial on modularity-based community detection in Gephi:

https://www.youtube.com/watch?v=7LMnpM0p4cM

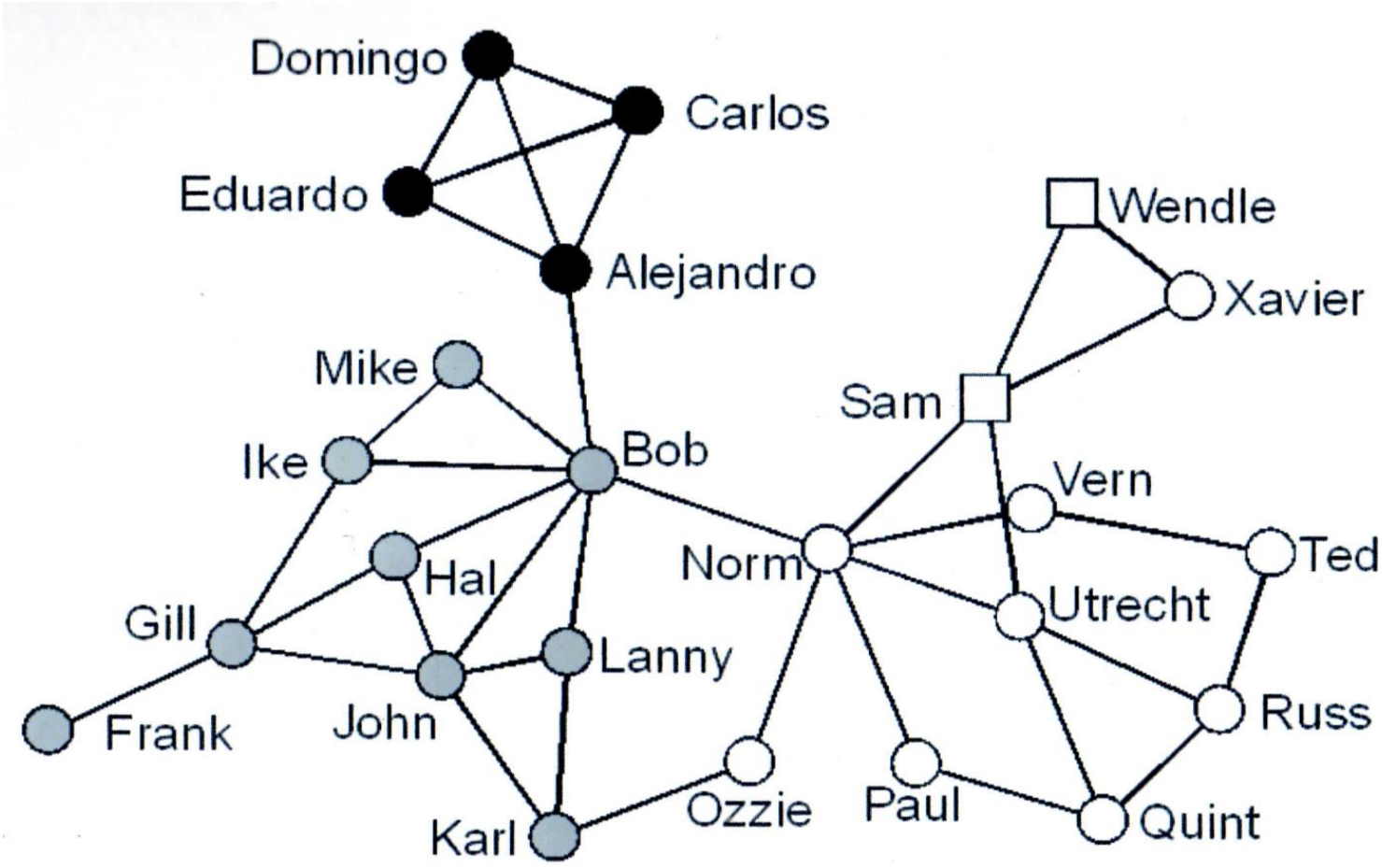
- Exploratory Social Network Analysis with Pajek by Wouter de Nooy, Andrej Mrvar, Vladimir Batagelj.

Attiro family visiting network



Next Week

Brokers and Bridges



Extra Stuff

Factions

Partition the data into a pre determined number of no-overlap groups.

1. Arbitrarily assign nodes to one of the groups and calculate the fit using a fit function which measures how good the partition is.
2. Move some nodes from one group to another and calculate the fit again to see if there is any improvement.
3. Repeat step 2 till no more improvement can be made.

Implemented in UCInet.

Factions

- Difficult to decide what is a good move
  - applies combinatorial optimizations
- User has to define the number of groups
  - Try different numbers
- There will be always be a solution even if data contains no real group.
- Multiple forms of the fit function to choose from.
- Can be computationally difficult. Most combinatorial optimization search routines best used to fewer than 100 nodes.