CIFF

Centro Internacior
de Formación Fina

CIFF Trustees:

Santander

Universidad de Alcalá

Social Network Analysis

Lecture 3: Communities in Networks

David Martín-Corral Calvo
dmartincc@smartvel.com

Enero / Marzo

MASTER EN BA & BD

Recap:

Many measures:

degree, betweenness, closeness, eigenvector

In indirected networks:

indegree, outdegree, page rank

Outline:

- Communities in networks
- Clustering
- Community structures
- Modularity
- Overlapping communities

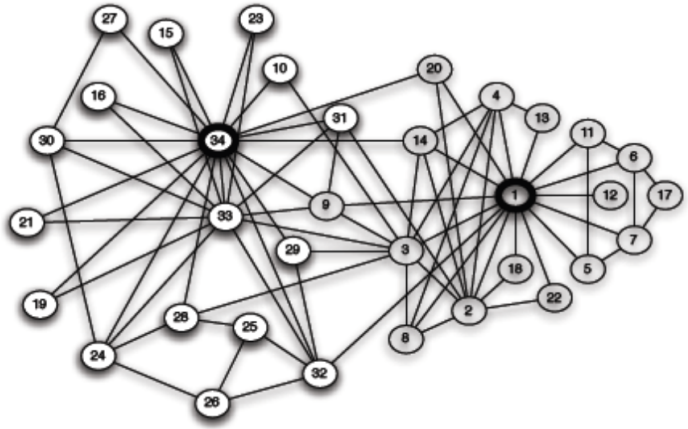
Why do we look for communities structures?

- We need to define it in order to find it
- Approaches to find it

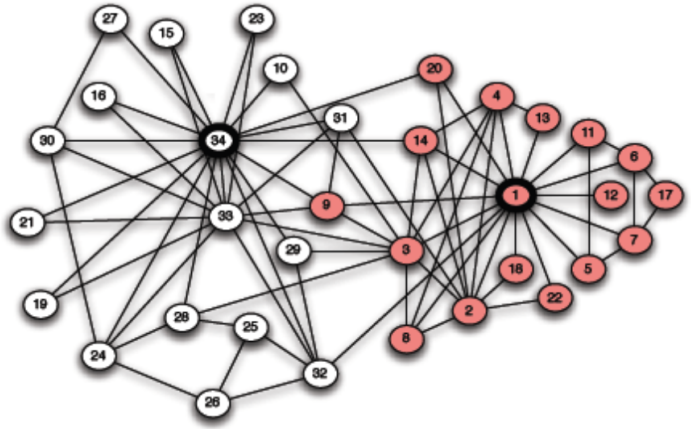
Why do it?

- Discover communities of practice
- Measure isolation of groups
- Understand opinion dynamics/adoption

Zachary Karate Club

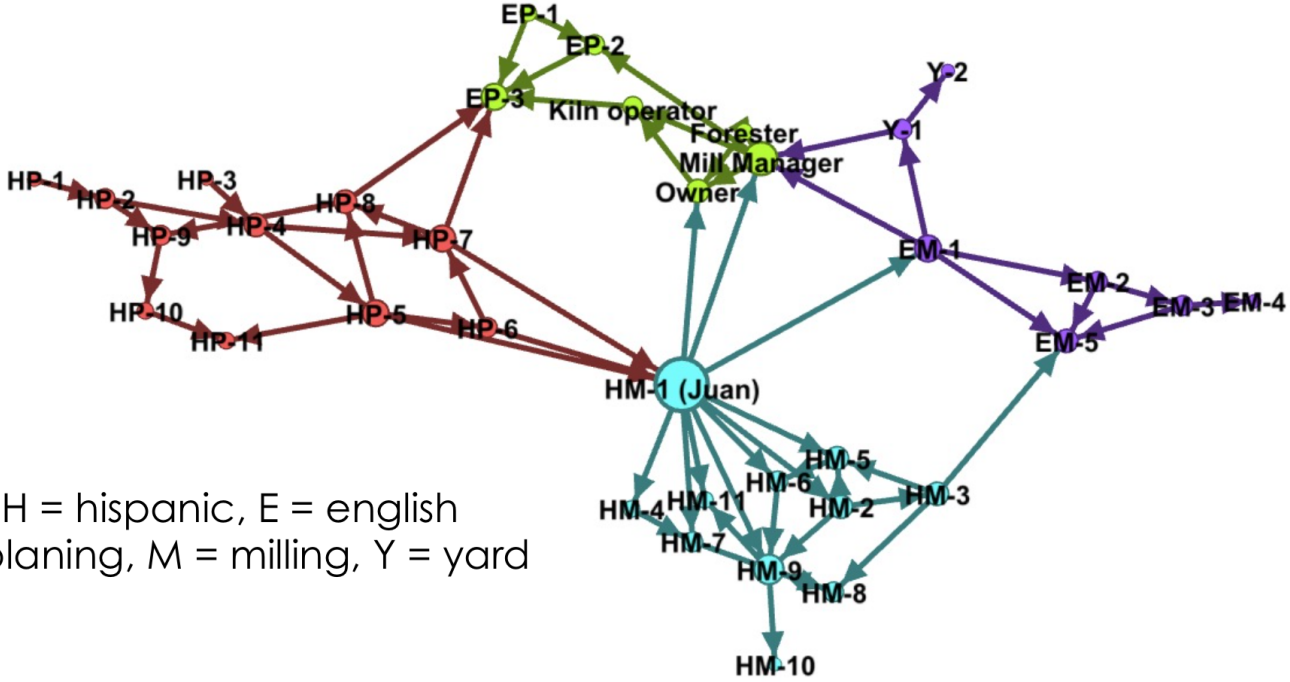


(a) *Karate club network*



(b) *After a split into two clubs*

Why look for community structure?



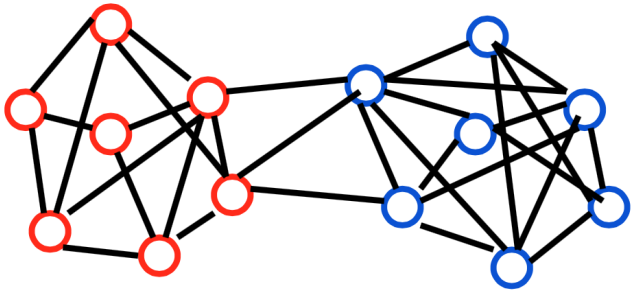
Key, H = hispanic, E = english
P = planing, M = milling, Y = yard

Quiz:

The management at the sawmill was having difficulty persuading the workers to adopt a new plan, even though everyone would benefit. In particular the Hispanic workers (H) were reluctant to agree. The management called in a sociologist who mapped out who talked to whom regularly. Then they suggested that the management talk to Juan and have him talk to the Hispanic workers. It was a success, promptly everyone was on board with the new plan. Why?

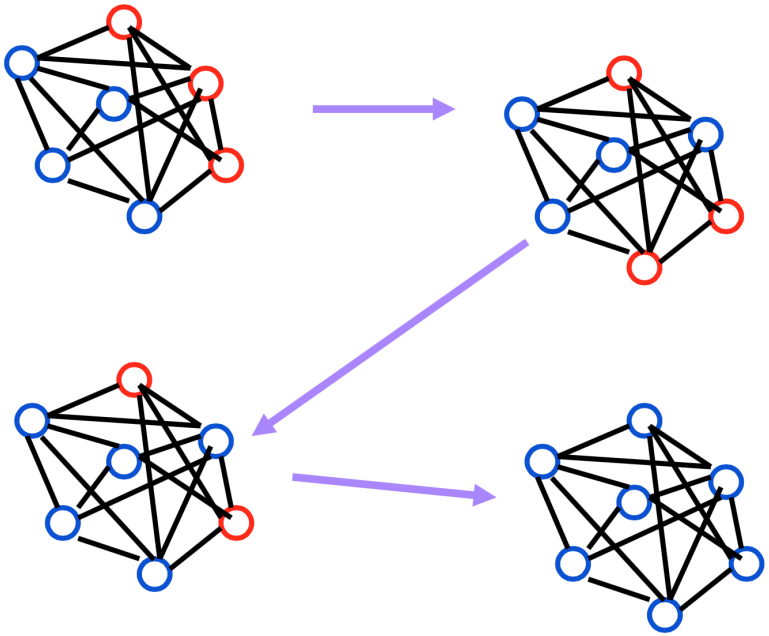
Why care about group cohesion?

Opinion formation and uniformity



- if each node adopts the opinion of the majority of its neighbors, it is possible to have different opinions in different cohesive subgroups

Within a cohesive subgroup, greater uniformity

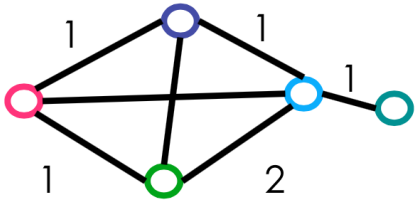


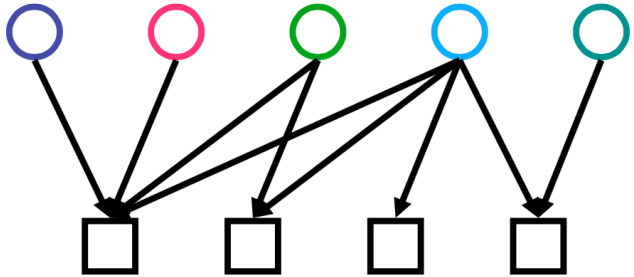
What makes a community?

- mutuality of ties
  - everybody in the group knows everybody else
- frequency of ties among members
  - everybody in the group has links to at least k others in the group
- closeness or reachability of subgroup members
  - individuals are separated by at most n hops
- relative frequency of ties among subgroup members compared to nonmembers

Affiliation networks

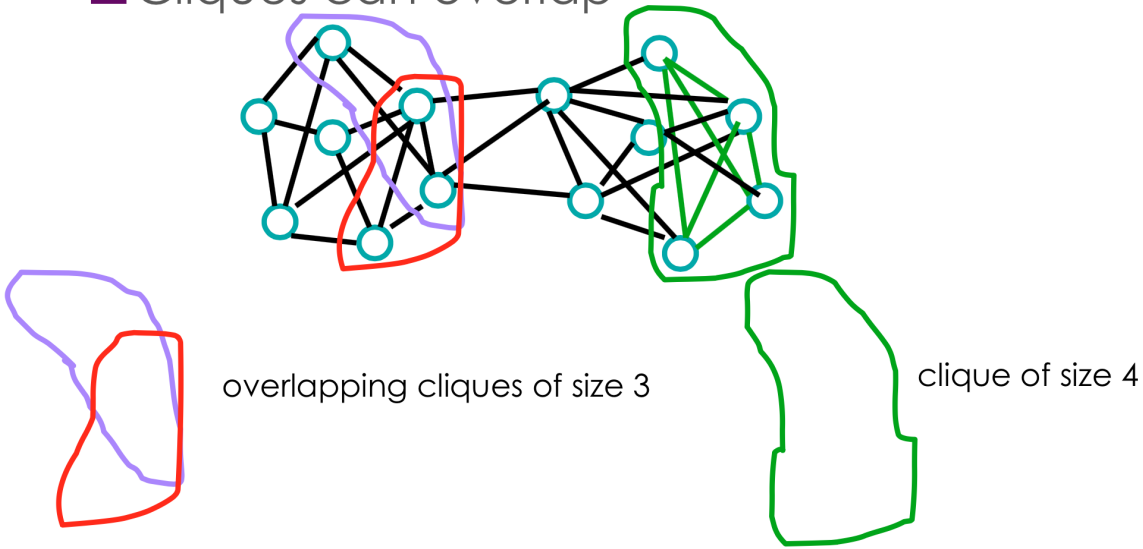
- otherwise known as
  - membership network
    - e.g. board of directors
  - hypernetwork or hypergraph
  - bipartite graphs
  - interlocks





Cliques

- Every member of the group has links to every other member
- Cliques can overlap

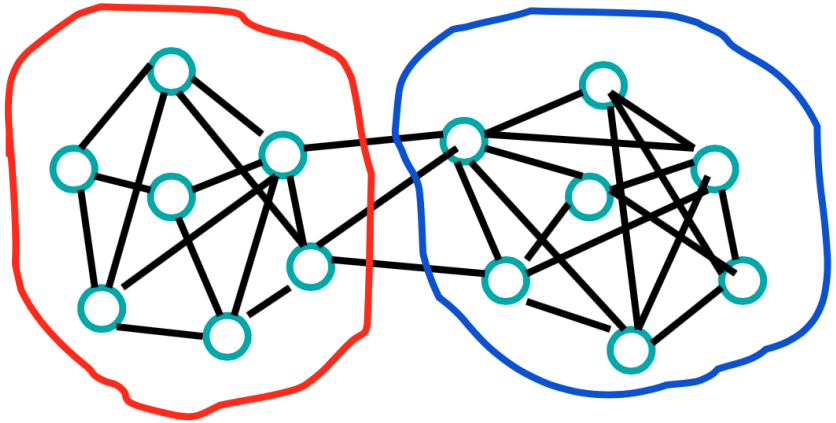


Meaningfulness of Cliques

- Not robust
  - one missing link can disqualify a clique
- Not interesting
  - everybody is connected to everybody else
  - no core-periphery structure
  - no centrality measures apply
- How cliques overlap can be more interesting than that they exist

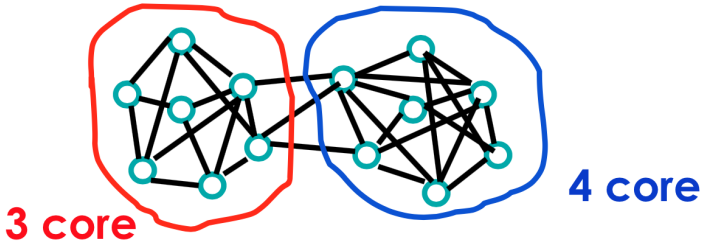
K-cores: Similar idea, less stringent

- Each node within a group is connected to k other nodes in the group

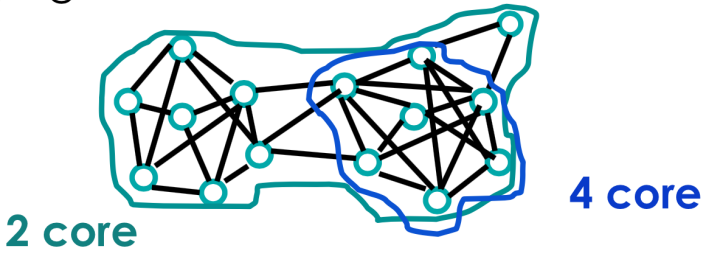


K-cores

- Each node within a group is connected to k other nodes in the group



- but even this is too stringent of a requirement for identifying natural communities



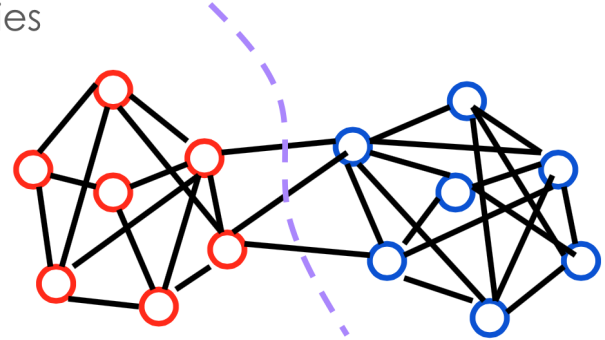
Cohesion in directed & weighted networks

- something we've already learned how to do:
  - find strongly connected components

- keep only a subset of ties before finding connected components
  - reciprocal ties
  - edge weight above a threshold

Community finding vs. other approaches

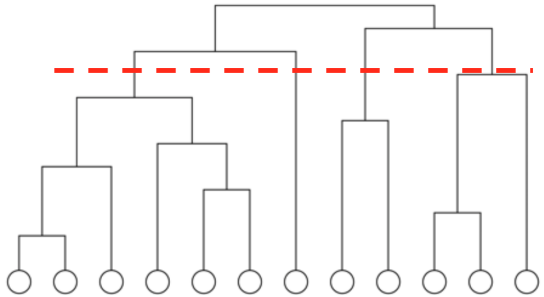
- Social and other networks have a natural community structure
- We want to discover this structure rather than impose a certain size of community or fix the number of communities



- Without "looking", can we discover community structure in an automated way?

Hierarchical clustering

- Process:
  - after calculating the "distances" for all pairs of vertices
  - start with all n vertices disconnected
  - add edges between pairs one by one in order of decreasing weight
  - result: nested components, where one can take a 'slice' at any level of the tree



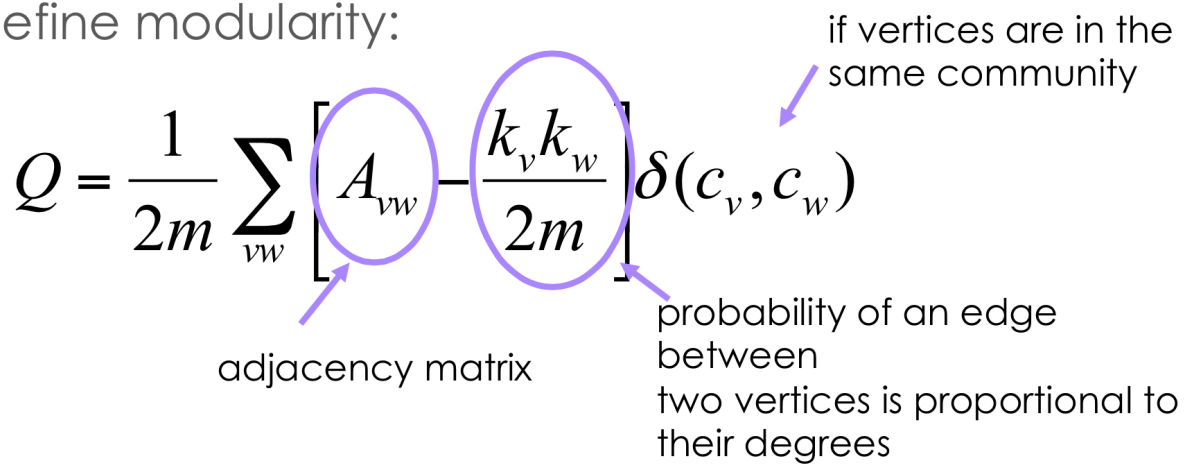
Betweenness clustering

- Algorithm
  - compute the betweenness of all edges
  - while (betweenness of any edge > threshold):
    - remove edge with highest betweenness
    - recalculate betweenness

- Betweenness needs to be recalculated at each step
  - removal of an edge can impact the betweenness of another edge
  - very expensive: all pairs shortest path – $O(N^3)$
  - may need to repeat up to N times
  - does not scale to more than a few hundred nodes, even with the fastest algorithms

Modularity

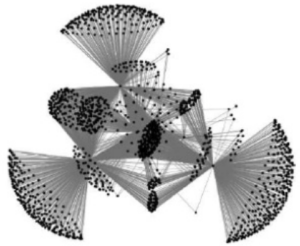
- Consider edges that fall within a community or between a community and the rest of the network
- Define modularity:



- For a random network, Q = 0
  - the number of edges within a community is no different from what you would expect

Modularity

- Algorithm
  - start with all vertices as isolates
  - follow a greedy strategy:
    - successively join clusters with the greatest increase ΔQ in modularity
    - stop when the maximum possible ΔQ <= 0 from joining any two
  - successfully used to find community structure in a graph with > 400,000 nodes with > 2 million edges
    - Amazon's people who bought this also bought that…
  - alternatives to achieving optimum ΔQ:
    - simulated annealing rather than gre



Recap:

- Community structure is a way of ‘x-raying’ the network, finding out what it’s made of.
- You can look for a specific structure:
  k-cliques, k-cores, etc.
- But most popular is to discover the “natural” community boundaries.

Activity: Communities

Ingredients complements

Football Transfers

1. The same as previous weeks (layout and metrics)
2. Run Modularity
3. Visualize nodes by modularity class

Activity: R

R script

1. Number of cliques
2. Number of communities

Assignment II

A network should be discovered somewhere in the wild and extracted from its source, modelled, visualized and drawn conclusions from it.

Must have:

- Explanation of your network origin and what it represents
- Data pipeline definition
- Data source analysis
- Data graph model
- Network analysis (metrics and visualizations)
- What questions can be answered from your network

Document: (5 - 10 pages) + annexes

Presentation: 5 slides

Reading list:

Statistical Properties of Community Structure in Large Social and Information Networks

Clique Finder

Uncovering the overlapping community structure of complex networks in nature and society

Suggested links:

igraph documentation

CIFF

Centro Internacional
de Formación Financiera

CIFF Trustees:

Santander

Universidad de Alcalá