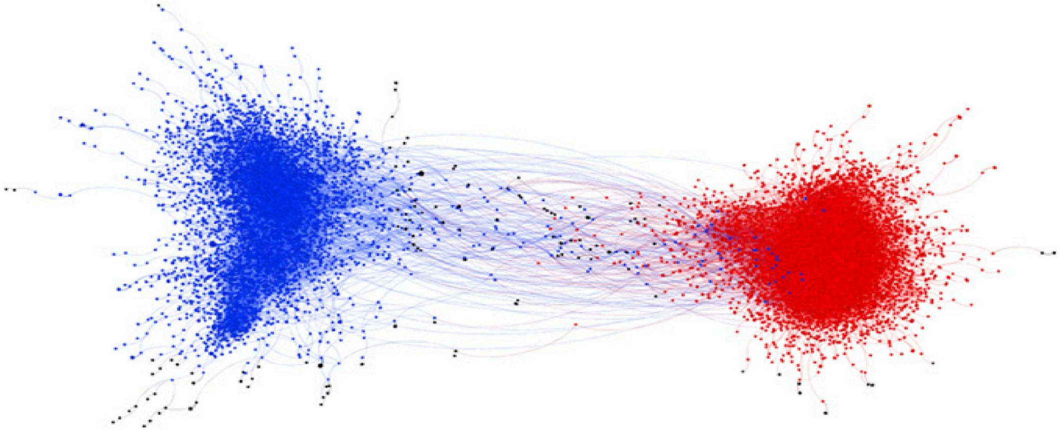
Community Detection Lecture 1

E0: 259

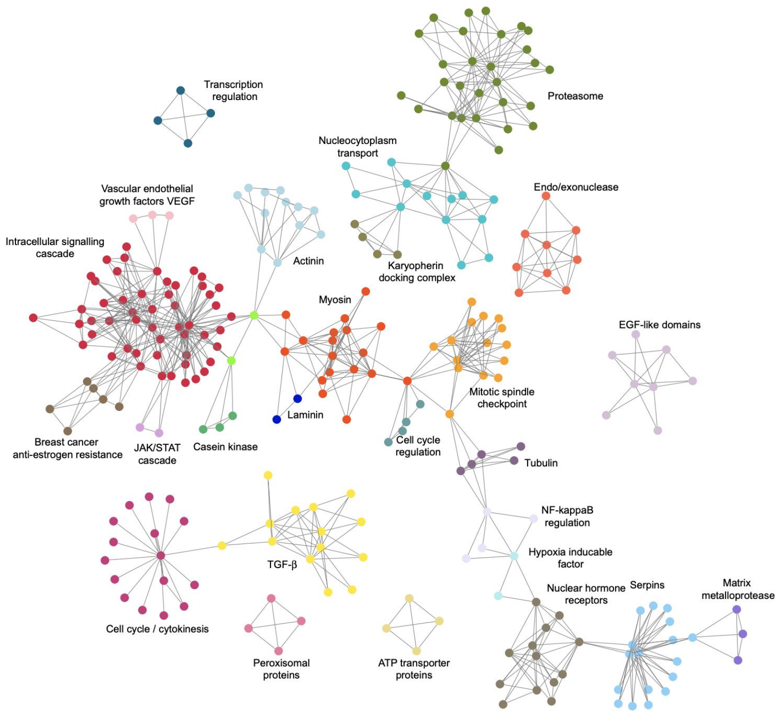
What is a Community?

- Intuitions:
  - People who have similar sets of beliefs - political, religious affiliations etc.
  - People who interact with each other on a regular basis - co-workers, students in a class, neighbors etc.
  - People in the same professions - lawyers, software developers, Professors etc.
  - People who have the same tastes - similar movie genres, similar kinds of books, similar types of food etc.



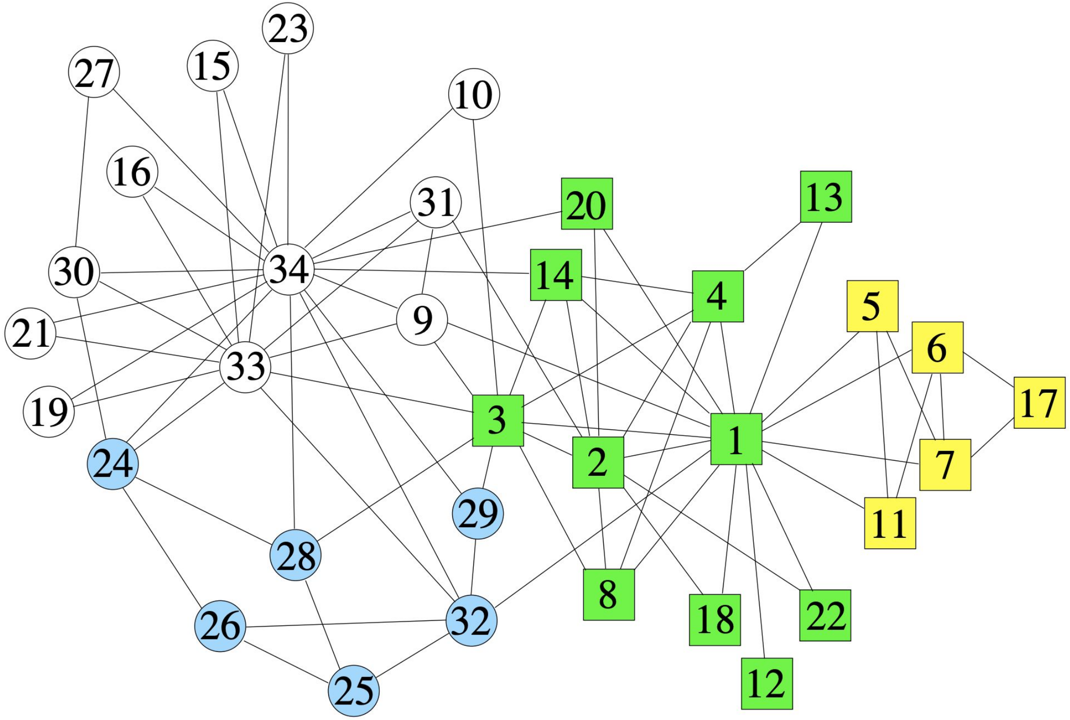
Brady et. al., PNAS 2017

Microscopic scale



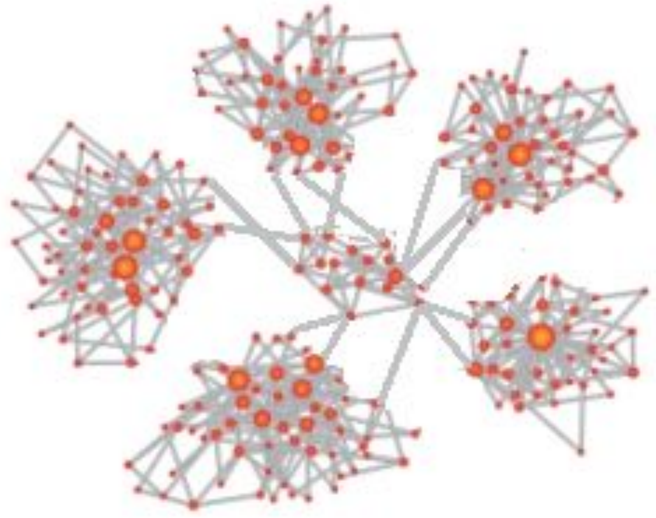
- Protein interactions in cancerous rats
- Johnson et. al., 2006

Small Communities - Zachary Karate Club



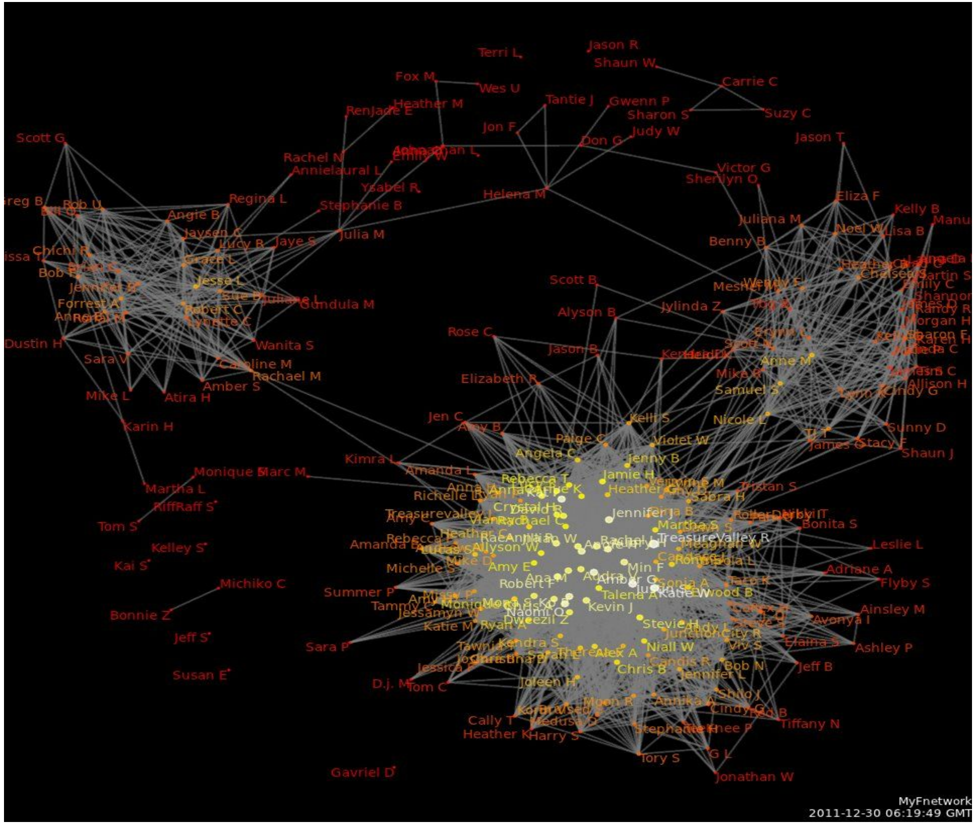
- Social Network Analysis
- Zachary Karate Club
- 34 members
- Fight between owner and trainer
- Split into 2

Ecological Networks



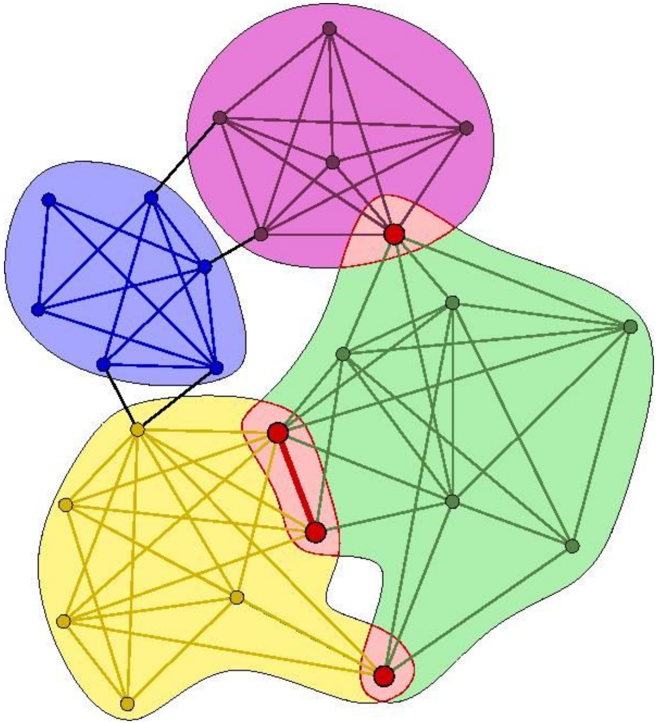
- Most plant and animal species interact only with a few other species and form tight inter species communities
- Ref: https://blogs.cornell.edu/info2040/2012/09/26/7720/

Planetary Scale - Online Social Network



- Online social network
- Planetary scale - spans several billion people

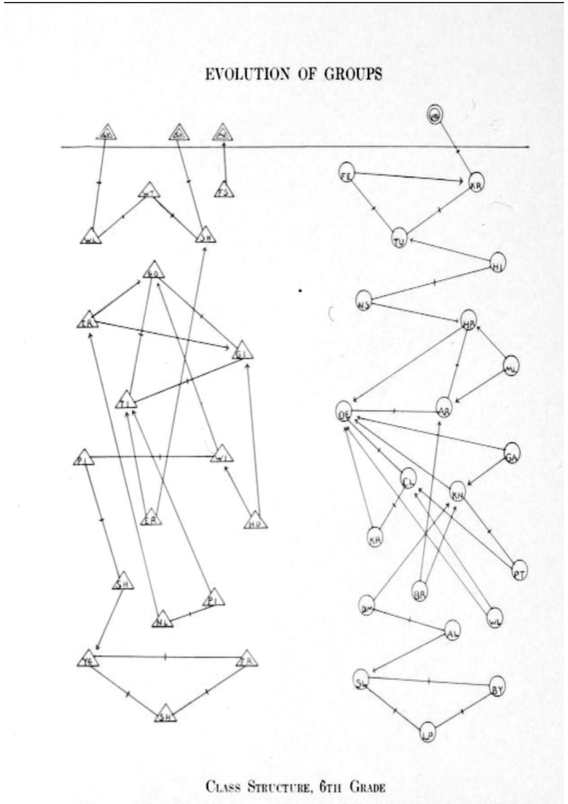
Overlapping Communities

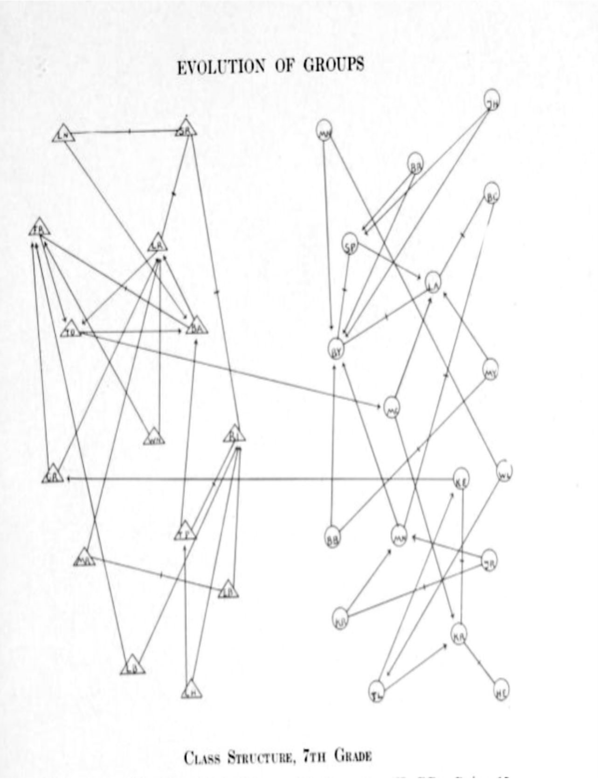


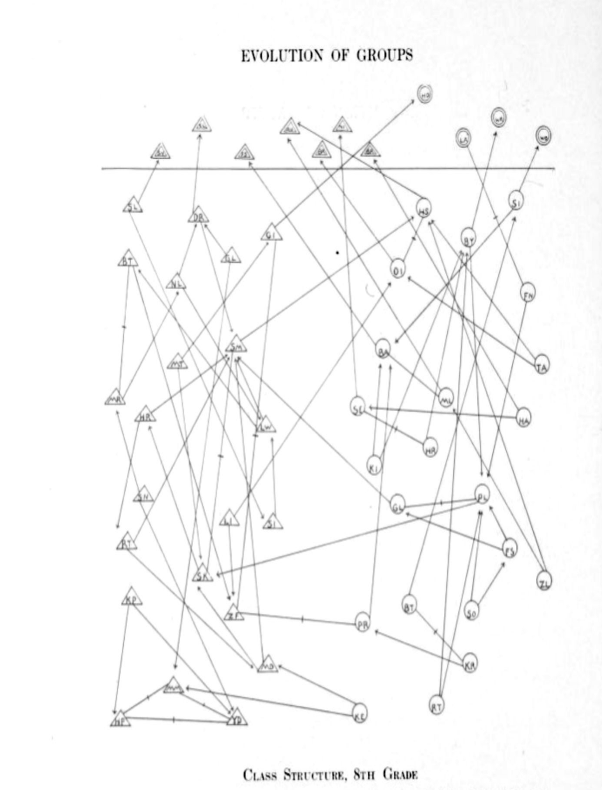
- Individuals have multiple identities - religion, politics, professional etc.
- Networks associated with each of these could be different.

Time Varying Communities

Moreno, 1938

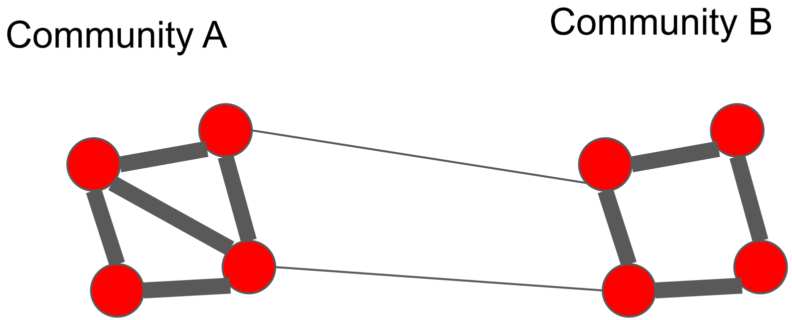






Intuitions to Definitions

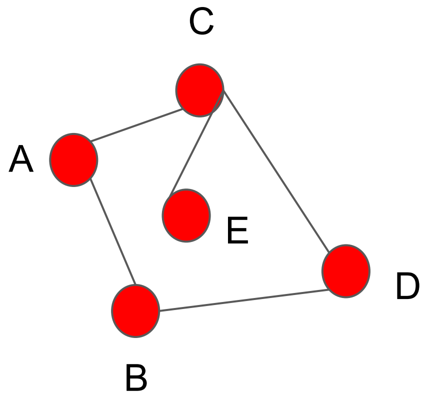
- Represent as a graph
- A community is set of nodes:
  - That have mostly strong relationships - or high edge weights
  - Most nodes in a community have an edge between them
  - Between communities there are few edges
  - They have low edge weights



How to Look for a Job? - Granovetter 1960

- You desperately need a job - who do you reach out to?
  - Option 1: Close friends
  - Option 2: Distant acquaintances
- Intuition: Information in tightly knit communities percolate very fast (think gossip in your family)
- If you want to gather new information, you need to go outside your community
  - Weak links or long range dependencies allow you to gather more information

Triadic Closure



- Given this graph, which edges are more likely?
- A->D or A->E
- A->D: intuitively if you and another person have several friends in common, you two are also likely to be friends.
- Use such intuitions to develop formal definition of a community and design algorithms around them

Definitions of Community - Degree based

$$C \subset G$$

$$\delta_{int}(C) = \frac{number\ of\ \ intra\ custer\ edges\ in\ C}{\frac{n_c\left(n_c - 1\right)}{2}}$$

$$\delta_{ext}(C) = \frac{number\ of\ \ inter\ cluster\ edges\ of\ \ C}{n_c\left(n - n_c\right)}$$

$$Objective: max \sum_{C \in G}\left(\delta_{int}(C) - \delta_{ext}(C)\right)$$

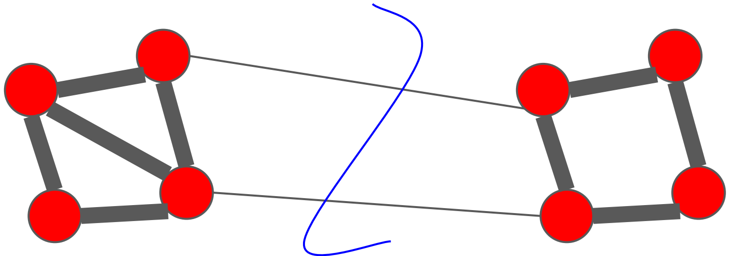
Cut based Partition Definitions

$$G = (V, E),\ V = V_1 + V_2, \qquad cut(V_1, V_2) = \sum_{i \in V_1,\ j \in V_2} e_{ij}, \qquad Vol(V_1) = \sum_{i \in V_1} k_i$$

$$Ratio\ Cut: \qquad Q = \frac{cut(V_1, V_2)}{||V_1||} + \frac{cut(V_1, V_2)}{||V_2||}$$

$$Normalized\ cut: \qquad Q = \frac{cut(V_1, V_2)}{Vol(V_1)} + \frac{cut(V_1, V_2)}{Vol(V_2)}$$

$$Conductance: \qquad Q = \frac{cut(V_1, V_2)}{min(Vol(V_1), Vol(V_2))}$$



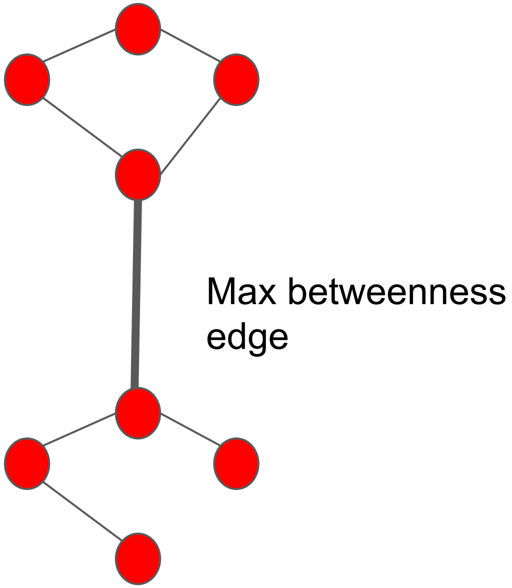
Betweenness Centrality

σ_{st} = *number of shortest paths between node s and node t in a graph*

$\sigma_{st}(e)$ = *number of shortest paths between node s and node t that pass through edge* $e \in E$

Edge Betweenness Measure

$$C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$$

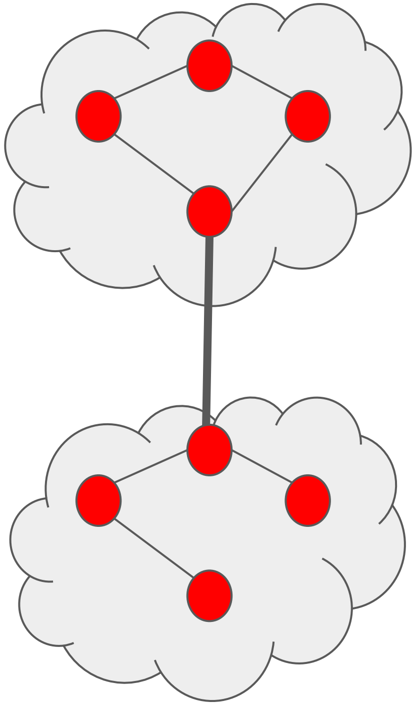


Intuitively removing max betweenness edges recursively should give good partitions

Modularity

$$\textit{Let } P \textit{ be set of partitions of } G$$

$$\textit{Modularity } Q = \sum_{p \in P} (\ \# \textit{ of edges in } p - \textit{expected } \# \textit{ of edges in } p)$$



Erdos-Renyi Graph

- Random graph model that matches number of nodes and expected number of edges in the graph.
- G(n,p)
- Construct a graph with n nodes.
- Pick every pair of nodes and assign an edge with probability p

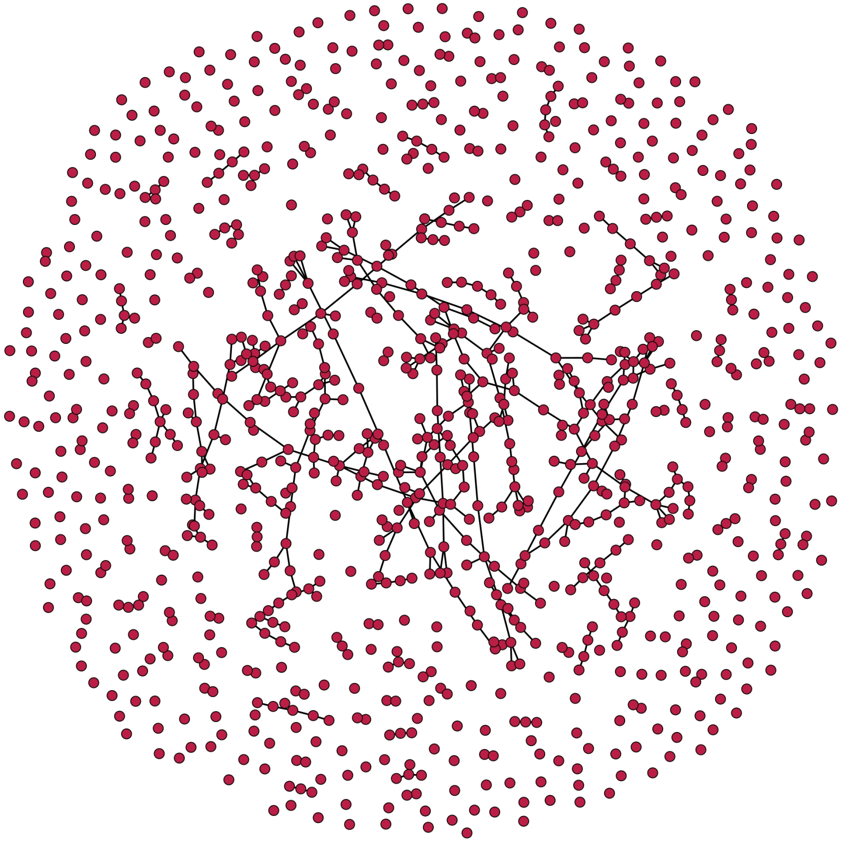
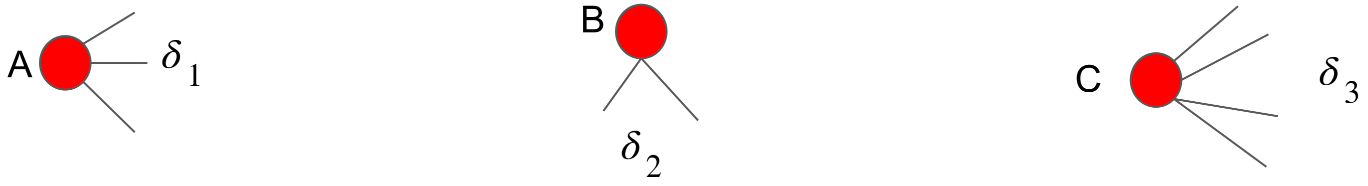
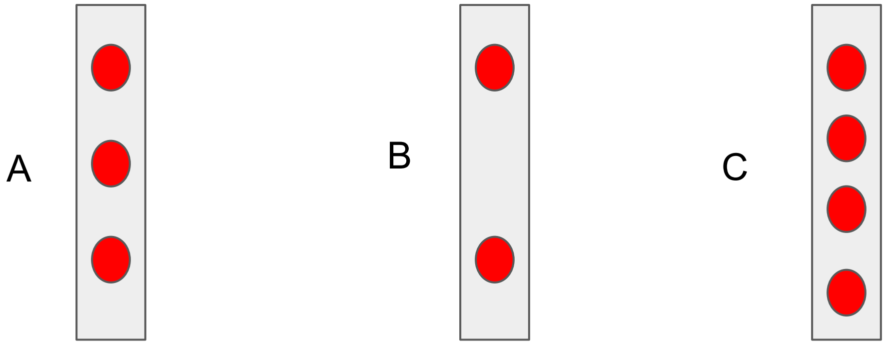


Image from Wikipedia

Configuration Model

- Ideally would like to preserve additional structural properties of the actual subgraph.
    - E.g. the node degrees are identical to the original graph.





- Make δ_i copies of each node i
- Pick 2 nodes from set of eligible nodes
- Put an edge between them
- Remove these 2 nodes from set of eligible nodes
- For large graphs guarantees that expected number of edges is same as original graph

Configuration Model

Consider a graph $G = (V, E)$, *let* $n = ||V||$ *and* $m = ||E||$,

let δ_i *be degree of node* i

then in the configuration model, expected number of edges between node i *and node* j *is* $\delta_i \frac{\delta_j}{2m}$

Expected number of edges with configuration model:

$$\sum_{i \in V} \sum_{j \in V} \delta_i \frac{\delta_j}{2m} = \frac{1}{2m} \sum_{i \in V} \delta_i \sum_{j \in V} \delta_j = 2m$$

With the configuration model, number of nodes, degree of each node and total number of edges are all preserved

Modularity with Configuration Model

$$\text{Let } A \text{ be the adjacency matrix of the graph. } A_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ have an edge} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Modularity } Q = \sum_{p \in P} (\#\text{ of edges in } p - \text{expected } \#\text{ of edges in } p)$$

$$\text{Modularity } Q = \frac{1}{2m} \sum_{p \in P} \sum_{i \in p} \sum_{j \in p} \left(A_{ij} - \frac{\delta_i \delta_j}{2m} \right)$$