Session II

Community Detection and Centrality

Community Detection

Based on the slides for Data Driven SNA
by Kaltenbrunner & Gómez

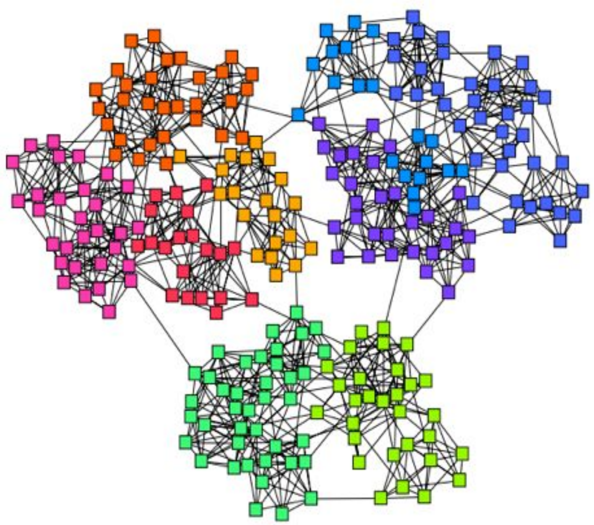
Community structure

Definition

Vertices in networks are often found to cluster into tightly-knit groups with a high density of within-group edges and a lower density of between-group edges.

Applications

- Identify groups of users who are more likely to interact with each other
- Identify customer with similar interests (purchasing history)
- Graph Compression
- Classification of vertices



Finding Communities

Given a graph G=<V,E>

- Community detection problem: find modules and their hierarchical organization

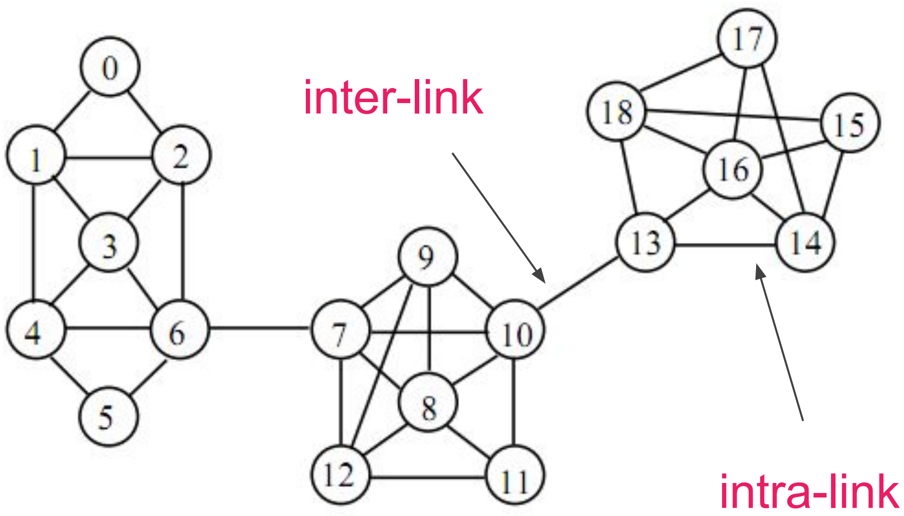
- What do we miss?
    - Define what is “a community”
    - Design algorithms that will find set of nodes which lead to “good communities”
    - Why just “good”??
    - Evaluate different results

Community

Definition

- There is no universally accepted definition of community

- Informally, a community C is a subset of nodes of V such that there are more edges inside the community than edges linking vertices of C with the rest of the graph



Community

Properties

Intra cluster density:

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

Inter cluster density:

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

$\partial_{ext}(C) \ll 2m/n(n-1) \ll \partial int(C)$

Community detection makes only sense on sparse graphs

Notation

V	set of nodes
E	set of edges
n	|V|
m	|E|
C	Subset of V
n_c	|C|

@elaragon

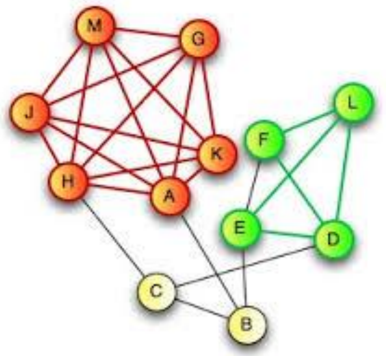
Local community: Clique

Definition

Subset of V such that all the vertices are adjacent to each other

Properties

- Triangles are really frequent in real networks
- Finding cliques in a graph is NP Complete
- However efficient algorithms for sparse graphs exist (e.g. Bron-Kerbosch algorithm)
- Too strict definition for Communities



Local community: Pseudo-clique

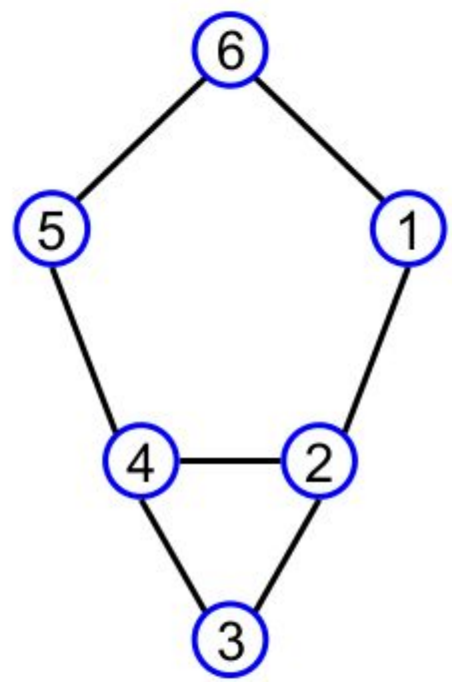
---

k-clique: Subset of vertices C: for every node i, j d(i,j)≤ k in G

k-club: Subset of vertices C: diam(G[C]) ≤ k

Examples

- 1-clique and 1-club (2,3,4)
- 2-clique:{1,2,3,4,5},{1,2,4,5,6}
- 2-club:{1,2,4,5,6}



@elaragon

Global community

Properties

- A graph has a community structure if it is different from a random graph

- A random graph is not expected to have any community structure: any two vertices have the same probability to be adjacent

- We can define a null model and use it to investigate whether a graph under consideration exhibits a community structure

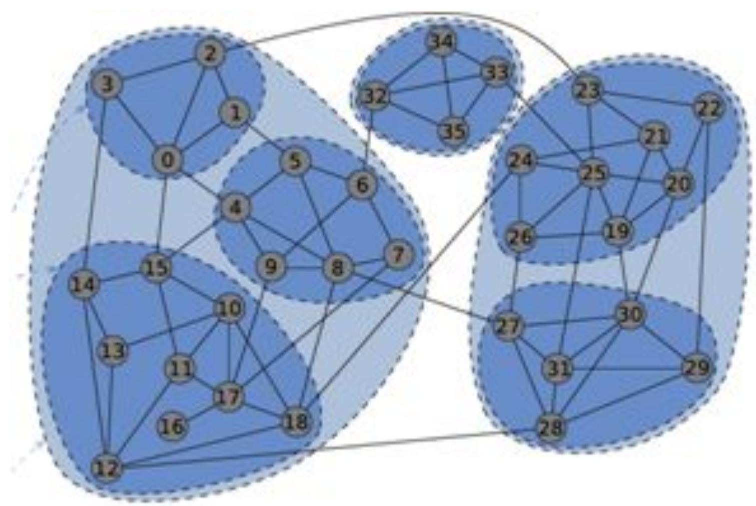
Graph partition

Definition

Division of a graph in clusters (communities), such that each vertex belongs to (at least)one cluster

Applications

- Hierarchical organization (communities can be embedded within other communities)
- Nodes can be shared between different communities (overlapping communities)



Modularity

Definition

The most popular quality function $$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j),$$

(density of edges in a subgraph vs density in a null model graph)

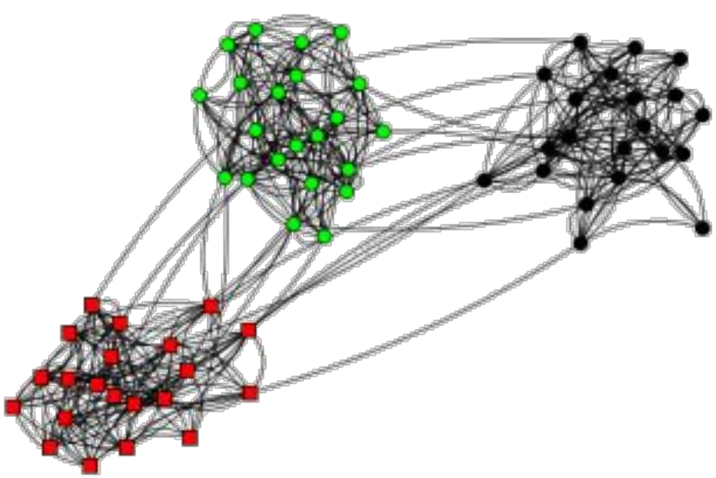
Applications

- The δ-function yields one if vertices i and j are in the same community, zero otherwise.
- P_{ij} represents the expected number of edges between vertices i and j in the null model (which is arbitrary)
- A_{ij} is the actual number of edges.

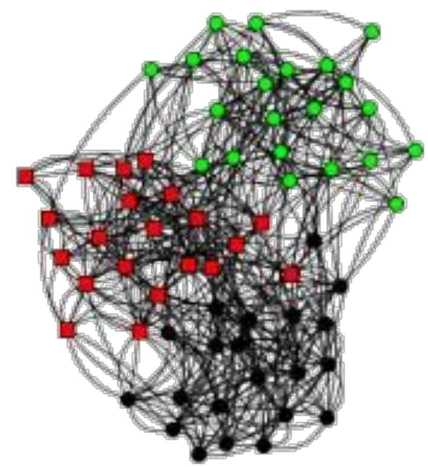
Modularity

Example

- In a random graph (Erdős-Rényi model), we expect that any possible partition would lead to Q=0
- Typically, in non-random graphs modularity takes values between 0.3 and 0.7.



Q = 0.60 clear community structure



Q = 0.37 fuzzy communities

Hierarchical Clustering

- Widely used in social network analysis
  - No need to specify the number of clusters
  - Graph may have a hierarchical structure
- Hierarchical Clustering aim at identifying groups of vertices with high similarity (not focusing on connectedness)
  - Define a similarity measure between vertices
  - Compute the n x n similarity matrix
  - **Agglomerative algorithms:** (bottom up) clusters are merged if their similarity if sufficiently high
  - **Divisive algorithms:** (top-down) clusters are iteratively split by removing edges connecting vertices with low similarity

Hierarchical Clustering: Merging Clusters

Merging Clusters criteria:

- Minimum od single linkage

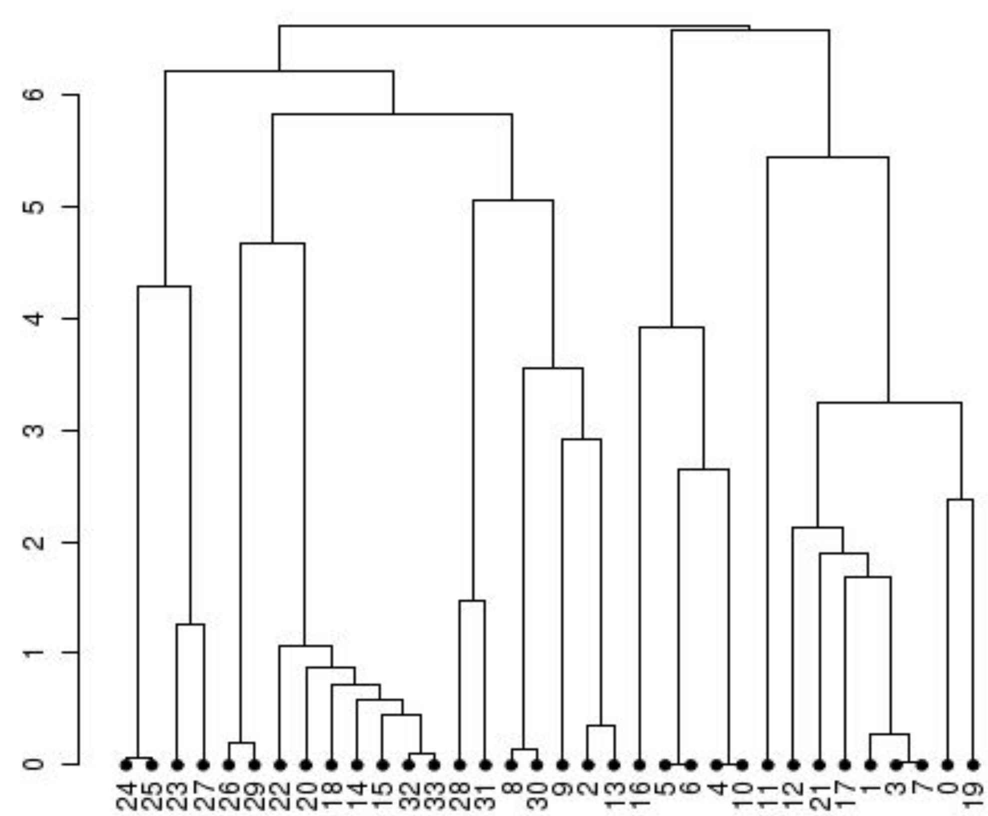
$$\max \{ d(a,b) : a \in A,\, b \in B \}.$$

- Maximum or Complete linkage

$$\min \{ d(a,b) : a \in A,\, b \in B \}.$$

- Average linkage

$$\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b).$$



Drawback of the hierarchical procedure: it does not provide a way to discriminate which level better represents the community structure of the graph

Girvan-Newman Method

Definition [Girvan 2002]

Divisive method that detect edges that connect different communities and remove them until clusters are disconnected

Steps

1. Compute Edge centrality
2. Remove the edge with the highest centrality
3. Update Centralities
4. If number of edges |E|>0, go to step 2

Girvan-Newman Method

- Instead of trying to construct a measure which tells us which edges are most central to communities, we focus instead on those edges which are least central

- If a network contains communities or groups that are only loosely connected by a few inter-group edges, then all shortest paths between different communities must go along one of these few edges

Girvan-Newman Method

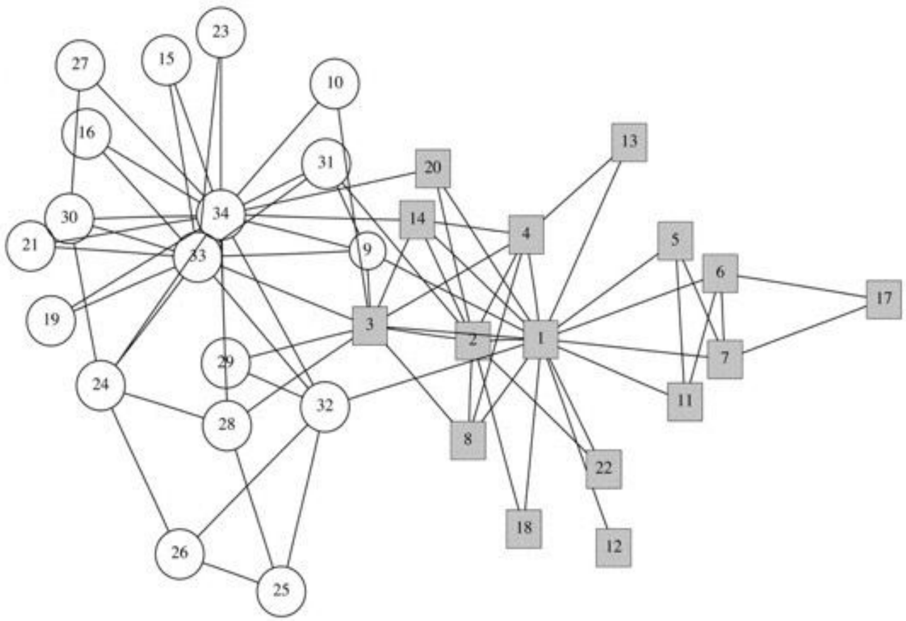
- Edge Betweenness O(mn)

  number of shortest path between all vertex pair that run along the considered edge

- The edges connecting communities will have high edge betweenness

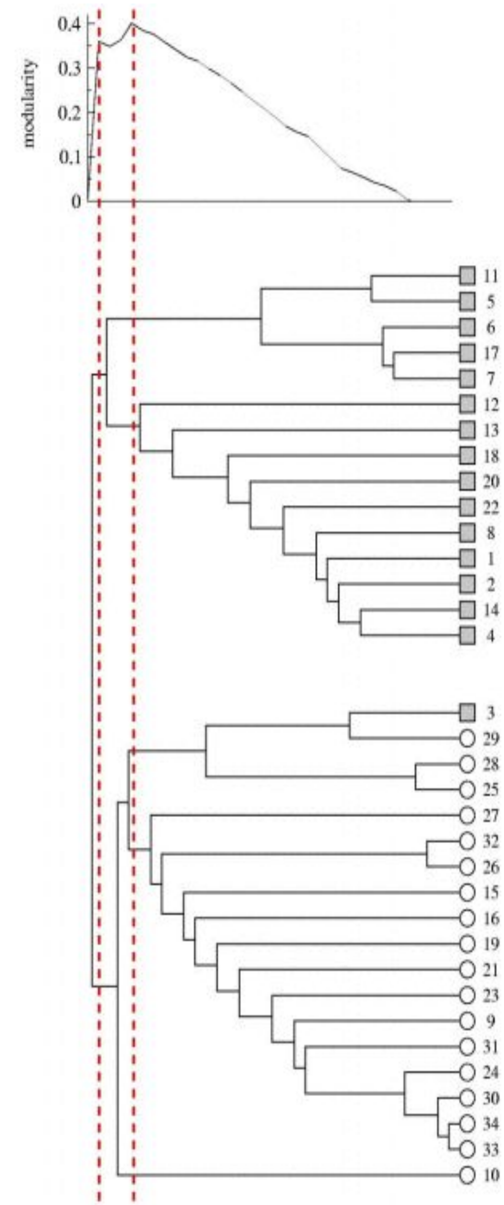
- Which partition is the best?
- Answer: Compute modularity

Girvan-Newman Method



Optimal community structure for Zachary's Karate club

Modularity without recalculation



Modularity optimization

If high modularity indicate goods partition, why not simply optimize Modularity over all partitions to find the best one?

$$Q = \sum_i (e_{ii} - a_i^2)$$

- e_{ii} is the fraction of edges in the network that connect vertices in the group i.
- a_i is the fraction of edges that connect vertices in the group i with every other group (including group i).

Answer: The search-space is exponential in |V|

Modularity optimization: Approximate Solution

Greedy algorithm [Newman 2004]

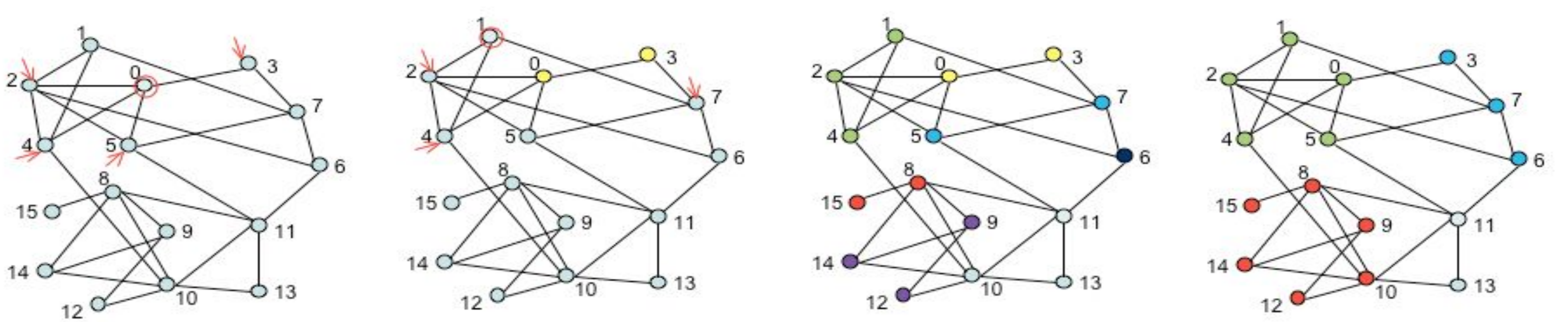
- **Agglomerative clustering:** we repeatedly join communities together in pairs, choosing at each step the join that results in the greatest increase (or smallest decrease) in Q

- Note: joining communities that are not connected cannot result in an increase in Q. => This limits the number of tentative joins to (m)

Louvain Method

Steps [Blondel 2008]

1. Initially, each node belongs to its own community (n nodes => n communities)
2. Pass through each node with a standard order. To each node, assign the community of their neighbor as long as this leads to an increase in modularity.
3. This step is repeated many times until a local modularity maximum is found.

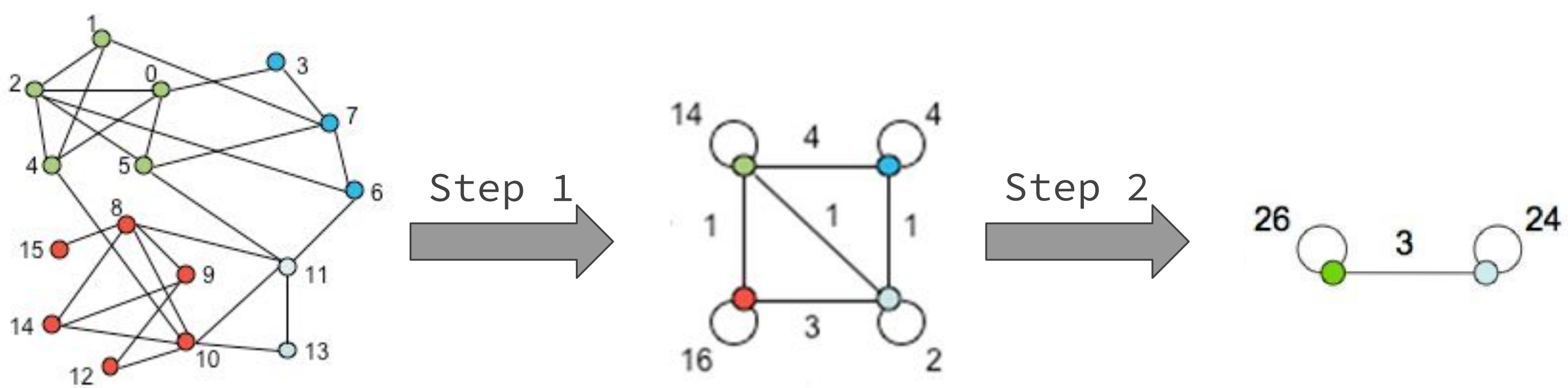


After 1 iteration　　After 4 iterations

Louvain Method

Folding

- Create new graph in which nodes correspond to the communities detected in the previous step.
- Edge weights between community nodes are defined by the number of inter-community edges.
- Folding ensures rapid decrease in the number of nodes that need to be examined and thus enables large-scale application of the method.



@elaragon

Louvain Method

Observations

- The output is also a hierarchy

- The method works for weighted graphs, and so modularity has to be generalized to

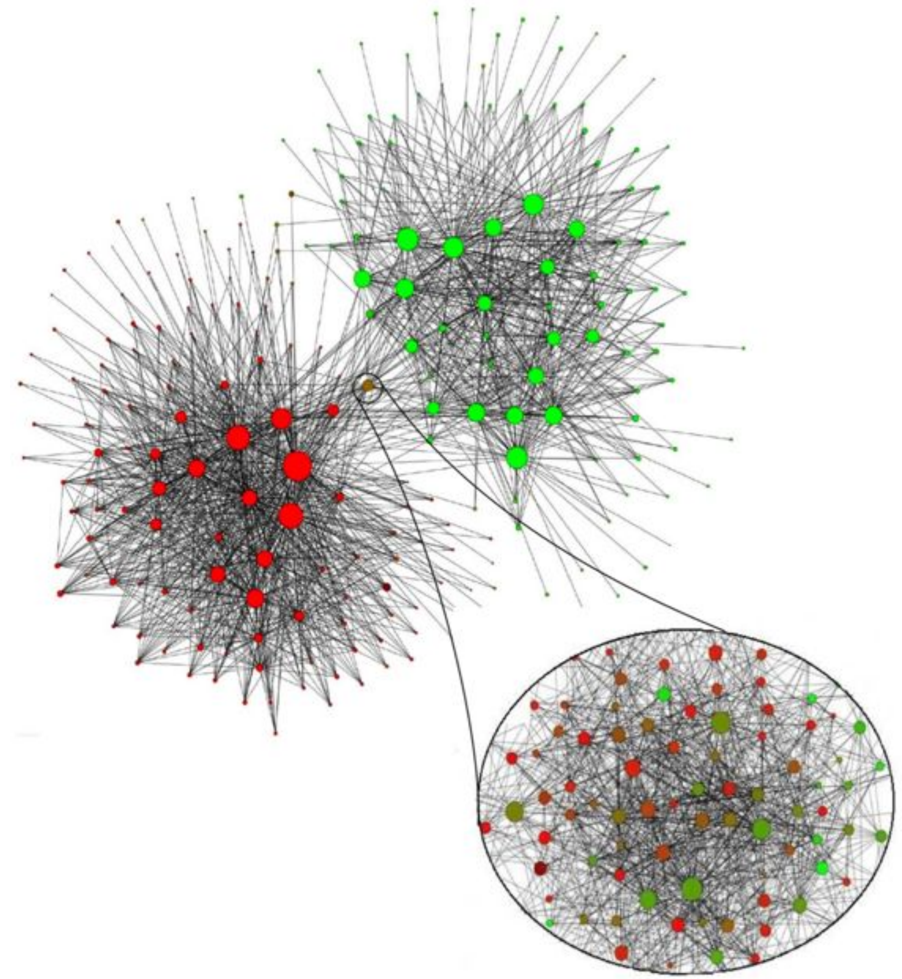
$$Q^w = \frac{1}{2W} \sum_{ij} \left(W_{ij} - \frac{s_i s_j}{2W} \right) \delta(C_i, C_j)$$

where W_{ij} is the weight of undirected edge (i, j),
$W = \sum_{ij} W_{ij}$ and $s_i = \sum_k W_{ik}$.

Louvain Method

Example

- Cell phone operator from Belgium
- 2.6 million customers
- 260 assemblages with over 100 customers, 36 with over 10,000
- 6 assemblage levels
- French and Dutch segments are almost independent



Conclusions

- Social networks are typically formed by communities of nodes.

- A community is a group of nodes with many edges between them and few edges with the rest of the nodes of the network.

- There are methods to detect communities, in this course we will use the **Louvain Method**:
  - Good results
  - Very fast

Centrality

Motivation

- People influence each other
- Interactions among individuals affect the thoughts, feelings and actions of others
- Can you measure the potential of a person in a social network to influence others?

Source: Mashable

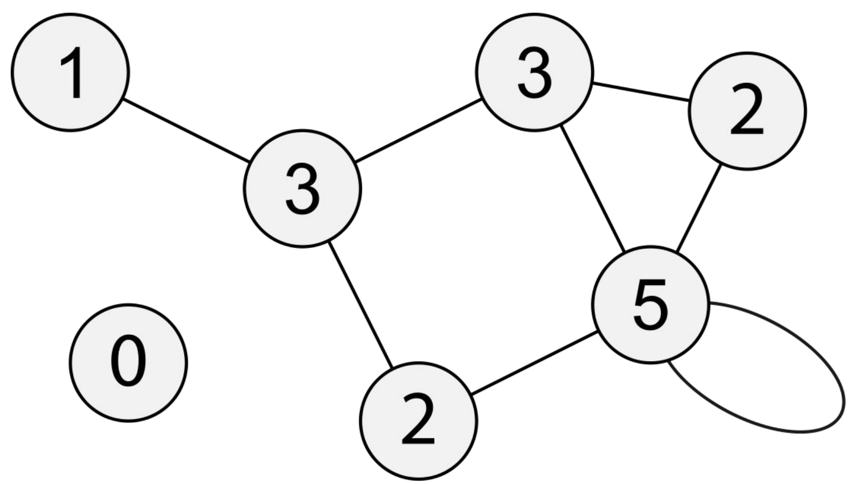
Degree centrality

Motivation

Identify the nodes with the highest number of links to other nodes

Method

- Node degree

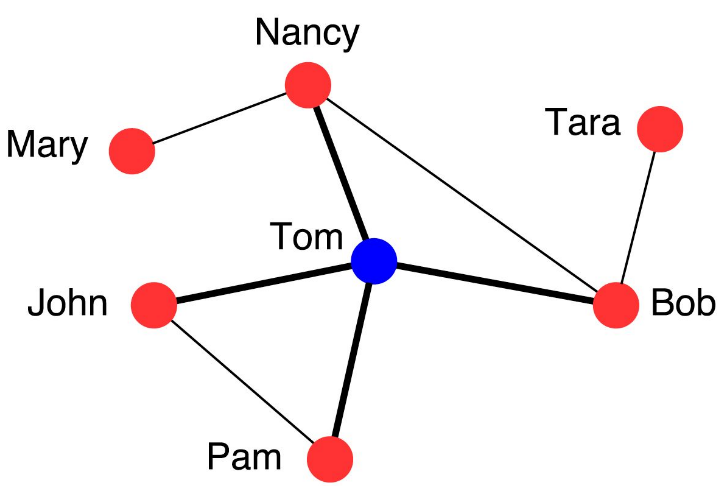


Source: Wikipedia

Degree centrality

Friendship paradox

- By choosing nodes at random, Tom has the same chance to be picked as everybody else
- By choosing links at random, Tom has a higher chance to be picked than everybody else
- By following links, the chance to hit a hub increases
- Avg. degree of a node = 2.29
- Avg. degree of the neighbors of a node = 2.83 > 2.29
- Our friends have more friends than we do, on average



Closeness centrality

Motivation

- In a diffusion model, it is often interpreted as the arrival time of something flowing through the network
- It measures the accessibility of one node to another.

Method

- It is the sum of the distances in a network from all the nodes in the network, where the distance from one node to another is defined as the length of the shortest path from one node to another.

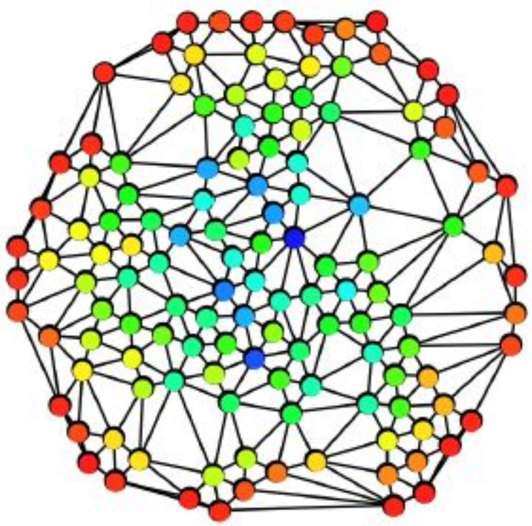
Betweenness centrality

Motivation

- Frequency that a node occurs on the shortest path between two others

Method

- Node i: $C_B(i)=\sum_{s \neq i \neq t \in V} \sigma_{st}(i)/\sigma_{st}$
- $\sigma_{st}(i)$ number of different shortest paths between nodes s,t
- σ_{st} number of different shortest paths between nodes s,t containing i

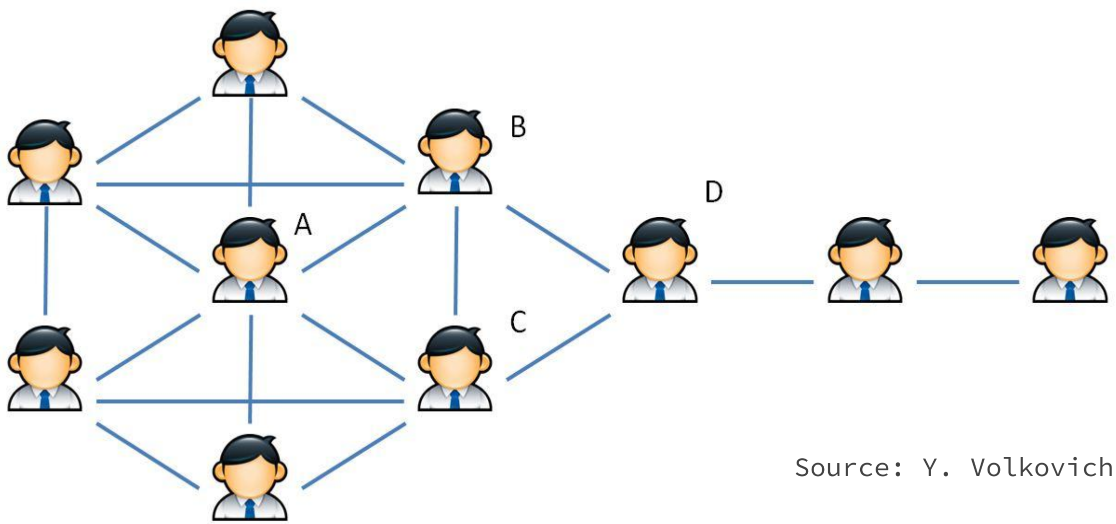


Source: Wikipedia

Comparison

Central nodes

- Degree centrality:
- Closeness centrality:
- Betweenness centrality:

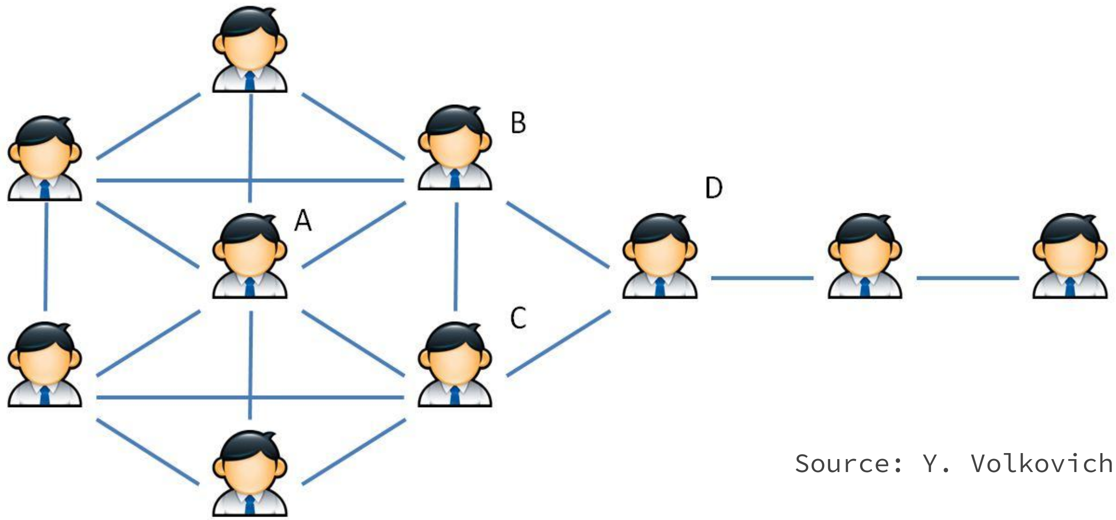


Source: Y. Volkovich

Comparison

Central nodes

- Degree centrality: USER A
- Closeness centrality: USERS B,C
- Betweenness centrality: USER D

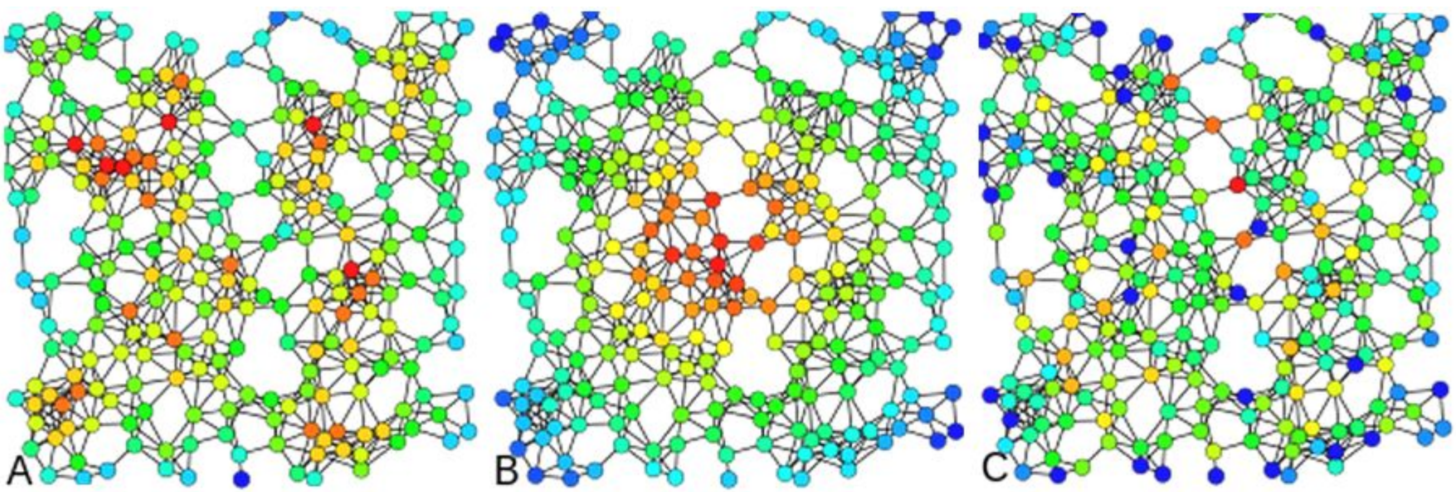


Source: Y. Volkovich

Comparison

Central nodes

- Degree centrality: Graph A
- Closeness centrality: Graph B
- Betweenness centrality: Graph C



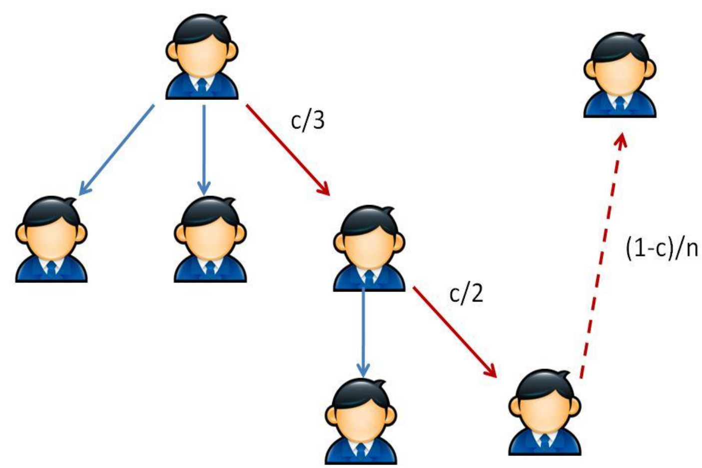
Source: Wikipedia

Pagerank [Brin 1998]

Motivation

- Google-defined popularity metrics for web ranking
- A random walk is simulated where at each step a jump is made to a random node with a probability (1-c)

$$PR^*(i) = c \sum_{j \to i} \frac{1}{d_j^*} PR^*(j) + \frac{1-c}{N^*},$$



Source: Y. Volkovich

- *PR*(i)* PageRank
- *d*j* Outdegree node j
- *N** Number of nodes

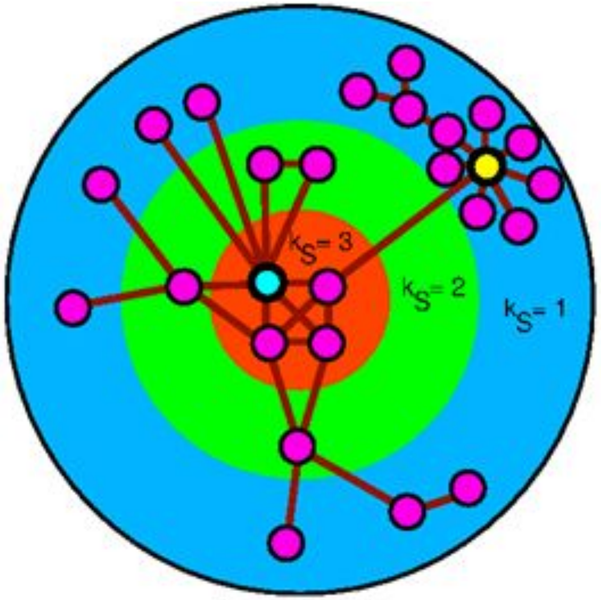
K-core decomposition

Motivation

- Detect nodes that are globally efficient to infect other nodes
- Discard local hubs (with many isolated contacts)

Method

- Larger sub-graph where each node has at least k direct neighbours



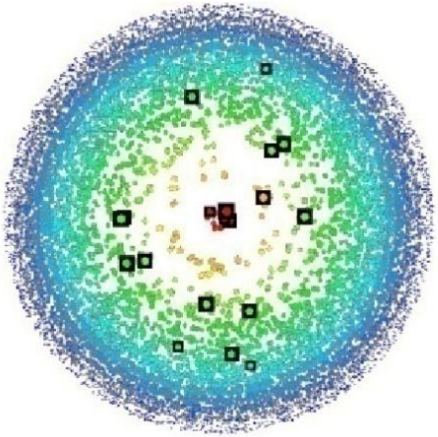
Source: Wikipedia

K-core decomposition

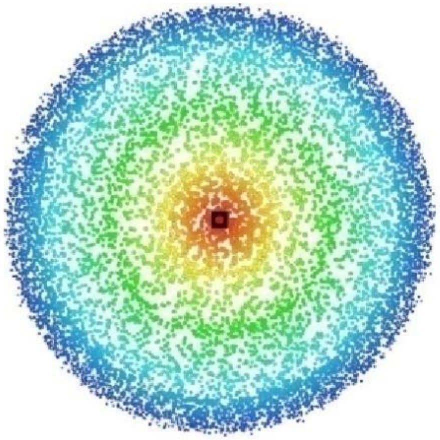
K-index

- Maximum k-shell that a node belongs to contacts)

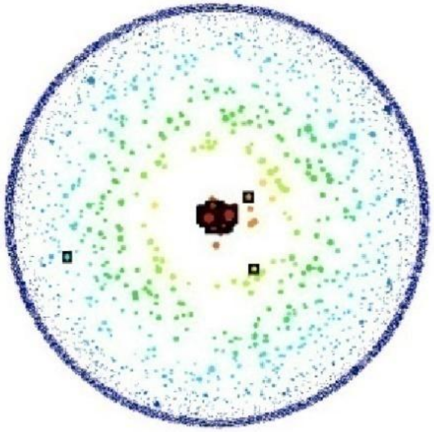
Examples



Physicists



Actors



Mails

Conclusions

Duncan Watts. Challenging the influential hypothesis

- The detection of influencers always happens a posteriori
- Influence might be based non-repeatable anecdotal data
- Influence might occur by accident
- Anyone can be influential
- Someone can be influential on one issue but not on another
- Influence exploitation probably leads to loss of influence

In short...

- There are nodes with more potential for influence than others
- But there's no guarantee they will exploit their capabilities

References

- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117.
- Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826.
- Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. Physical review E, 69(6), 066133.

Homework

Read the following paper:

Grandjean, M., & Jacomy, M. (2019). Translating Networks: Assessing correspondence between network visualisation and analytics. In Digital Humanities.

https://reticular.hypotheses.org/1745