Network communities

Leonid E. Zhukov

School of Data Analysis and Artificial Intelligence
Department of Computer Science
National Research University Higher School of Economics

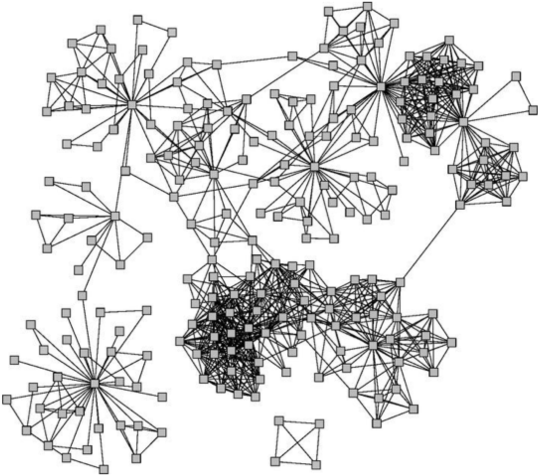
Structural Analysis and Visualization of Networks

NATIONAL RESEARCH
UNIVERSITY

Lecture outline

1. Cohesive subgroups
   - Graph cliques
   - k-plex, k-core
2. Network communities
   - Similarity based clustering
   - Graph partitioning

Network communities



Connected and undirected graphs

Network communities

What makes a community (cohesive subgroup):

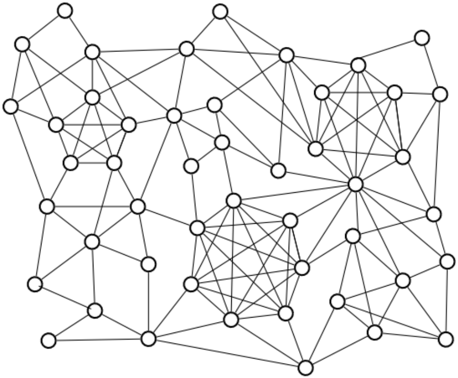
- Mutuality of ties. Everyone in the group has ties (edges) to one another
- Compactness. Closeness or reachability of group members in small number of steps, not necessarily adjacency
- Density of edges. High frequency of ties within the group
- Separation. Higher frequency of ties among group members compared to non-members

Wasserman and Faust

Graph cliques

Definition

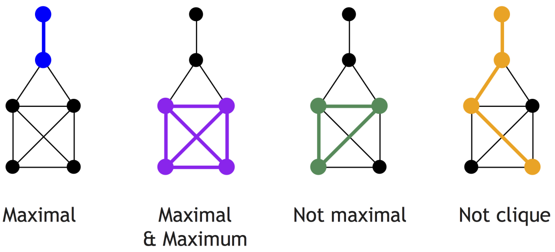
A *clique* is a complete (fully connected) subgraph, i.e. a set of vertices where each pair of vertices is connected.



Cliques can overlap

Graph cliques

- A **maximal clique** is a clique that cannot be extended by including one more adjacent vertex (not included in larger one)
- A **maximum clique** is a clique of the largest possible size in a given graph



- Graph clique number is the size of the maximum clique

image from D. Eppstein

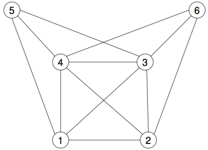
Graph cliques

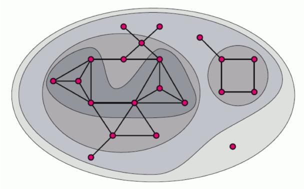
Computational issues:

- Finding click of fixed given size k - $O(n^k k^2)$
- Finding maximum clique $O(3^{n/3})$
- But in sparse graphs...

Relaxation of a clique

- k-**plex** of size n is a maximal subset of n vertices such that each vertex is connected to at least $n-k$ others in the subset (any vertex can be lacking ties with no more than k members).
- k-**core** is a maximal subset of vertices such that each is connected to at least k others in the subset (degree of every vertex in k-core $k_i \geq k$) . (k+1) core is always a subgraph of $k - core$





K-core

- The core number of a vertex is the highest order of a core that contains this vertex

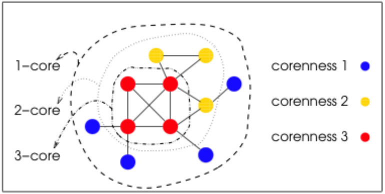
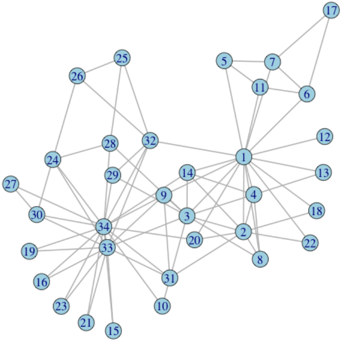


image from Alvarez-Hamelin et.al., 2005

Graph cliques

Zachary Karate Club, 1977

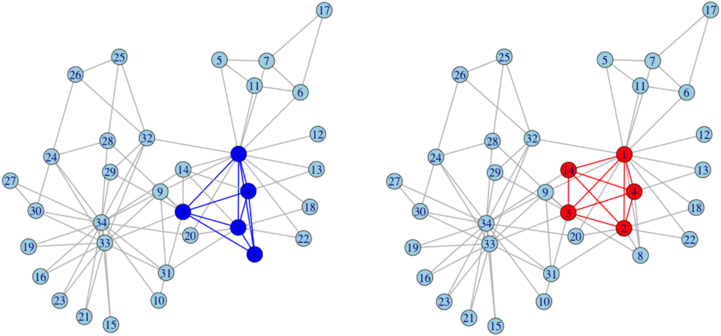


Maximal cliques:

Clique size:	2	3	4	5
Number of cliques:	11	21	2	2

Graph cliques

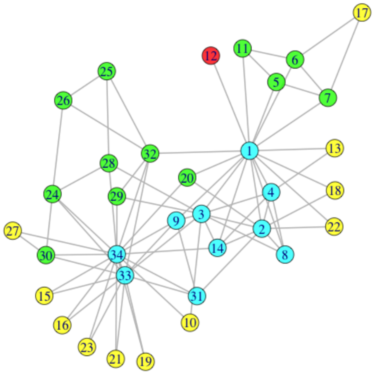
Zachary karate club 1,2,3,4 - cores



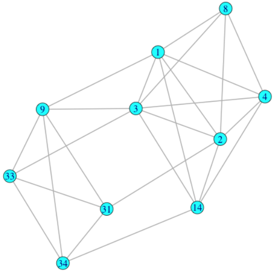
Maximum cliques

K-cores

Zachary karate club: 1,2,3,4 - cores



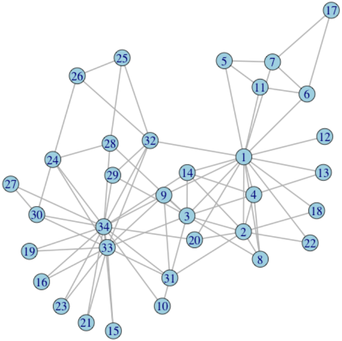
$\xrightarrow{4-core}$



Network communities

Definition

Network communities are groups of vertices similar to each other.



- Community detection is an assignment of vertices to communities.
- Non-overlapping communities (every vertex belongs to a single group)

Similarity based clustering

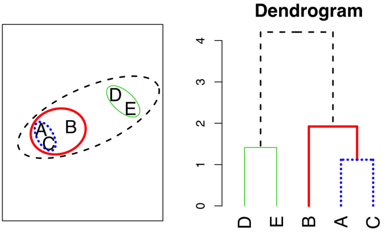
Similarity based vertex clustering:

- Define similarity measure between vertices based on network structure
  - Jaccard similarity
  - Cosine similarity
  - Pearson correlation
  - Eucledian distance (dissimilarity)
- Calculate similarity between all pairs of vertices in the graph (similarity matrix)
- Group together vertices with high similarities

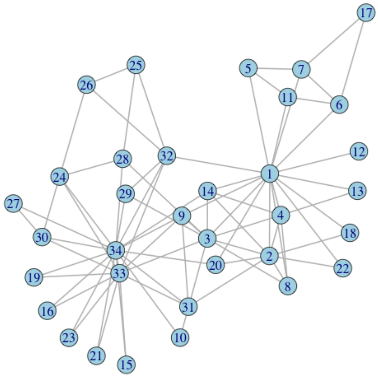
Hierarchical clustering

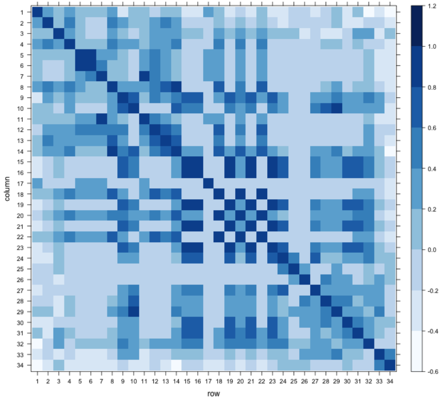
Agglomerative clustering:

- Assign each vertex to a group of its own
- Find two groups with the highest similarity and join them in a single group
- Calculate similarity between groups:
  - single-linkage clustering (most similar in the group)
  - complete-linkage clustering (least similar in the group)
  - average-linkage clustering (mean similarity between groups)
- Repeat until all joined into single group

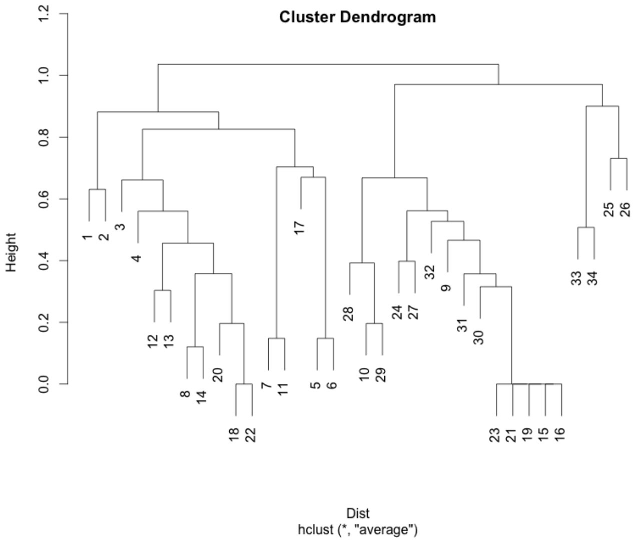


Zachary karate club

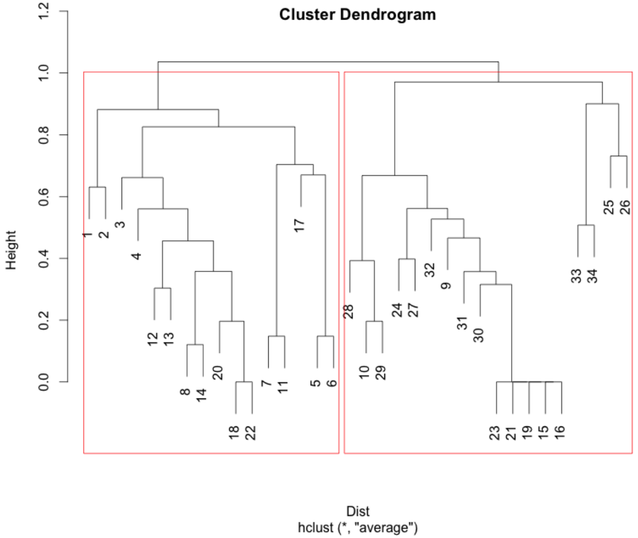




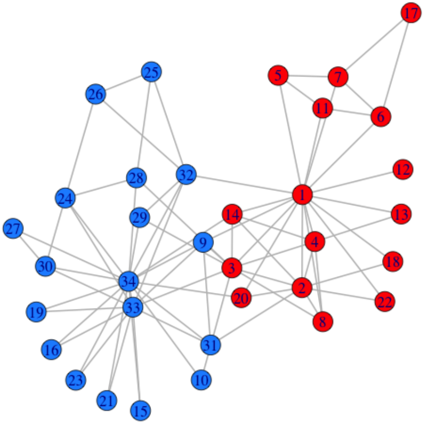
Hierarchical clustering



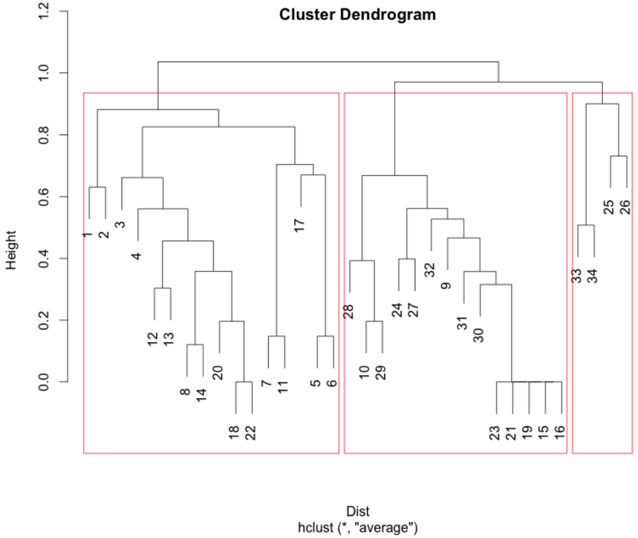
Hierarchical clustering



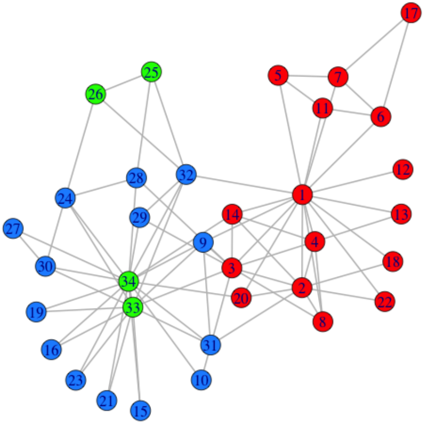
Hierarchical clustering



Hierarchical clustering



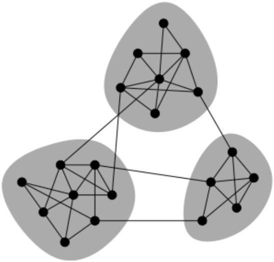
Hierarchical clustering



Network communities

Definition

Network communities are groups of vertices such that vertices inside the group connected with many more edges than between groups.



- Graph partitioning problem

Graph partitioning

Combinatorial problem:

- Number of ways to divide network of n nodes in 2 groups (bi-partition):

$$\frac{n!}{n_1!n_2!}, \qquad n = n_1 + n_2$$

- Dividing into k non-empty groups (Stirling numbers of the second kind)

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{n}(-1)^j C_k^j (k-j)^n$$

- Number of all possible partitions (n-th Bell number):

$$B_n = \sum_{k=1}^{n} S(n,k)$$

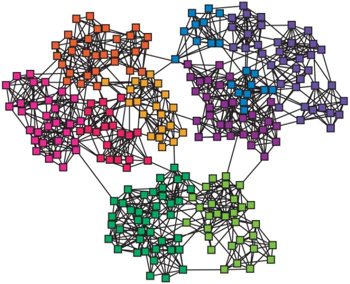
$B_{20} = 5,832,742,205,057$

Heuristic approach

Focus on edges that connect communities.

Edge betweenness -number of shortest paths $\sigma_{st}(e)$ going through edge e

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



Construct communities by progressively removing edges

Edge betweenness

Newman-Girvan, 2004

Algorithm: Edge Betweenness

Input: graph G(V,E)

Output: Dendrogram

repeat

  For all $e \in E$ compute edge betweenness $C_B(e)$;

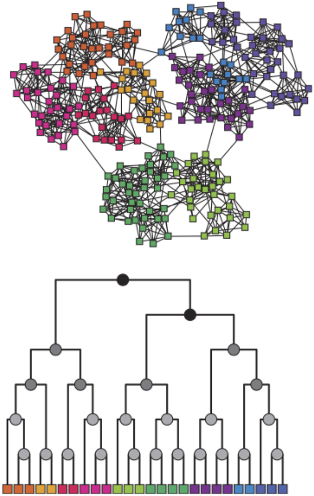
  remove edge e_i with largest $C_B(e_i)$;

until *edges left*;

If bi-partition, then stop when graph splits in two components (check for connectedness)

Edge betweenness

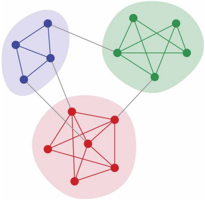
Hierarchical algorithm, dendrogram



Community "quality"

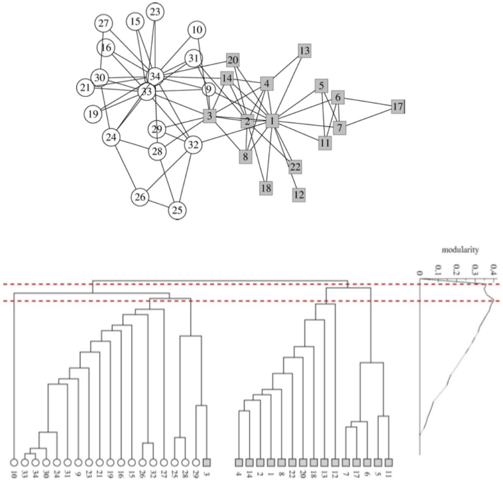
- Let n_c - number of classes, c_i - class label per node
- Compare fraction of edges within the cluster to expected fraction if edges were distributed at random
- Modularity:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \quad \delta(c_i, c_j)\text{- kronecker delta}$$



- The higher the modularity score - the better is community
- Modularity score range $Q \in [-1/2, 1)$
- Single class, $\delta(c_i, c_j) = 1$, $Q = 0$

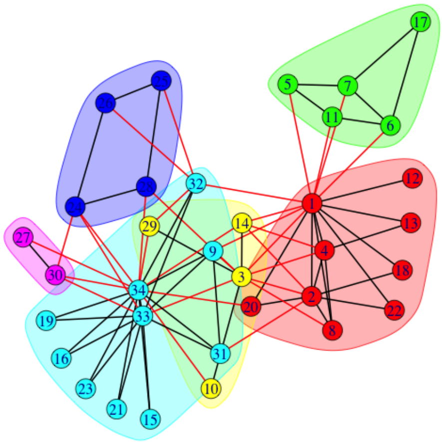
Dendrogram and modularity score



Newman and Girvan, 2004

Network communities

Zachary karate club



References

- Finding and evaluating community structure in networks, M.E.J. Newman, M. Girvan, Phys. Rev E, 69, 2004
- Modularity and community structure in networks, M.E.J. Newman, PNAS, vol 103, no 26, pp 8577-8582, 2006
- S. E. Schaeffer. Graph clustering. Computer Science Review, 1(1):2764, 2007.
- S. Fortunato. Community detection in graphs, Physics Reports, Vol. 486, Iss. 35, pp 75-174, 2010