Quantifying Network Structure: Clustering and Modularity

IV124

Josef Spurný & Eva Výtvarová

Faculty of Informatics, Masaryk University

March 17, 2023

Network Cohesion

a measure of the connectedness and togetherness among nodes within a network

- network density – a measure of how many links between nodes exist compared to how many links between nodes are possible $D = \frac{2*|E|}{|V|*(|V|-1)}$
- component – a group of nodes where a path exists between any two nodes of the component
- number of components as an important part of network description
- giant component – a component of a network with the majority of nodes
- isolated nodes, isolated pairs of nodes, isolated n-tuples

Edge Weights

Analysis and interpretation of measures extracted from weighted networks depend on weight semantics.

Weight captures a distance between two nodes

- euclidean distance
- Manhattan distance
- Chebyshev distance
- Hamming distance
- ...

Weight captures a similarity between two nodes

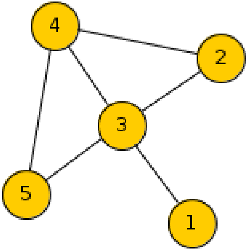
- Pearson correlation
- Spearman correlation
- Jaccard coefficient
- mutual information
- ...

Most algorithms use distances, a similarity is usually converted as $w_D(i,j) = 1/w_S(i,j)$. Watch out for $w_S(i,j) = 0$!

Clustering Coefficient

Clustering coefficient C_i of a node i:

- how are the neighbors of a node i connected?
- $C_i = \frac{L_i}{k_i(k_i-1)}$,
- where L_i are links connecting neighbors of a node i



- $C_3 = 1/6$

Clustering Coefficient

Average clustering coefficient $\overline{C}$

- $\overline{C} = \frac{1}{N} \sum_{i=1}^{N} C_i$
- can be read as a probability that two neighbors of a random node are connected

What does it tell about a network

- *local* transitivity: friends of my friends are also my friends
- regularity in a network structure: triangles

Identifying Subgroups

- identification of nodes that are densely connected with each other but loosely connected with the rest of the network

bottom-up approach

- nodes form subgroups
- overlaps of subgroups constitute a network
- cliques, n-cliques, k-cores
- from strict to more benevolent criteria

top-down approach

- fragmenting a network to subgroups by removing edges or nodes
- communities, components, clusters

Community Structure

Sociological detour

Homophily

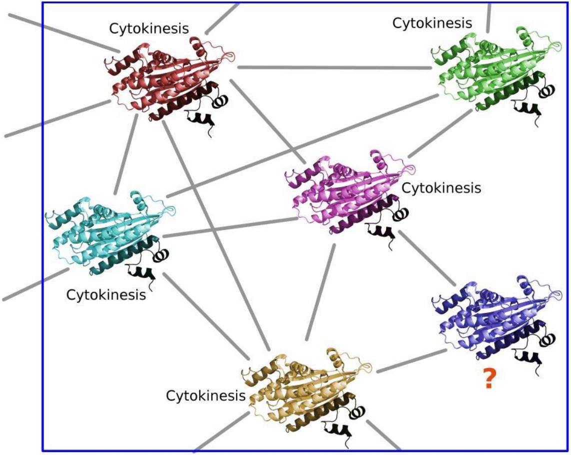
- a tendency of nodes to connect to nodes with similar attributes
- gender, age, social rank

Motivation

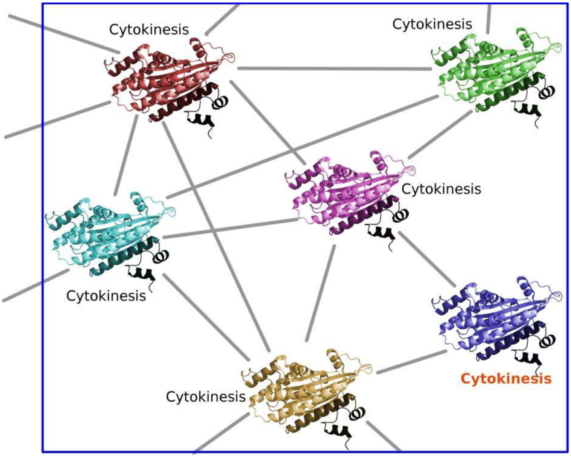
- in real networks, we often observe the emergence of clusters
- norms emerge in the clusters with peer pressure to follow them
- clusters are a result of the self-organization of a network

An exact definition of a community/cluster depends on the nature of the observed system.

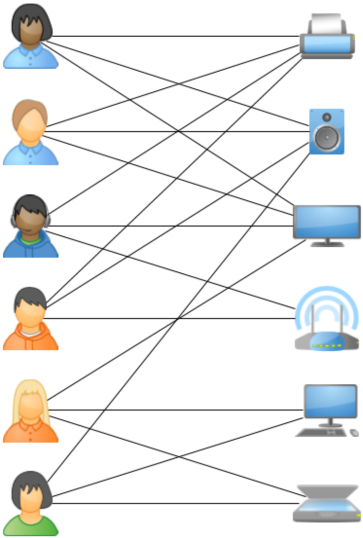
Motivational Example: Proteins Function



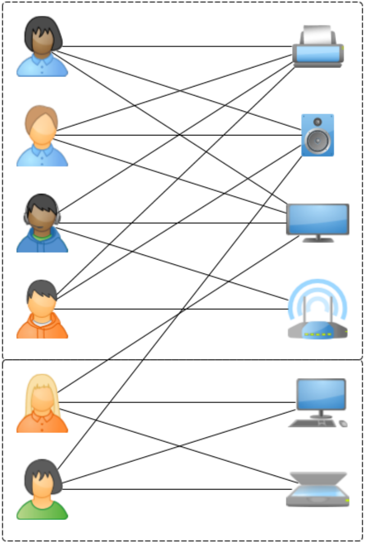
Motivational Example: Proteins Function



Motivational Example: Recommender Systems



Motivational Example: Recommender Systems



Community Structure Detection

1. we have a network with a particular semantics (social, transport, biological, ...)
2. we identify clusters
3. we interpret clusters either as functional units or as real communities

What is the Issue?

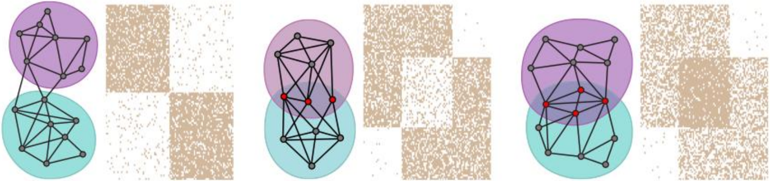
Unclear definition of a problem

- the quality of distribution into clusters is not unambiguous
- the interpretation is not necessarily straightforward
- for most networks, we do not have a control sample to compare the result

Complicating features of networks

- directed links
- weighted links
- hierarchic structure
- overlapping communities

Overlapping Communities

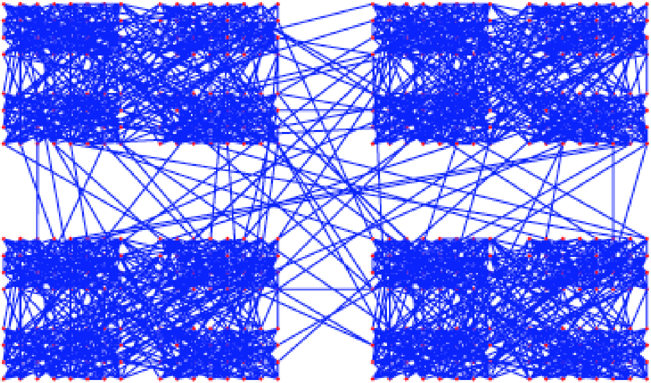


(a) No overlaps (b) Sparse overlaps (c) Dense overlaps [1]

Dense overlaps cause problems for most algorithms.

[1] Yang & Leskovec (2014)

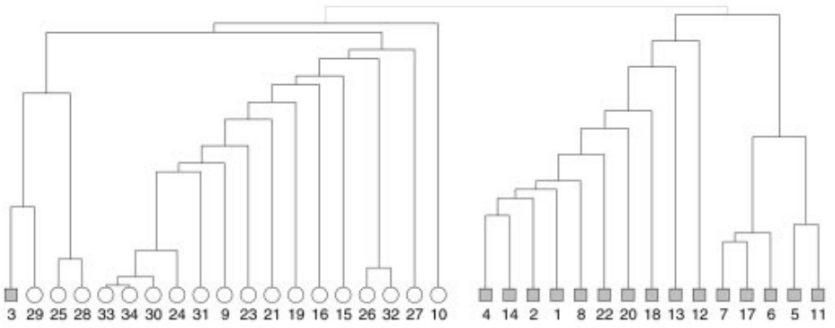
Hierarchic Structure



Community Detection Approaches

- Hierarchical clustering approaches
  - agglomerative clustering procedures
  - k-means clustering
- Betweenness clustering
- Modularity
- Block modeling

Hierarchical Cluster Analysis



Hierarchical Cluster Analysis

A general method for classification into groups

- hierarchical system of subsets
- similarity function (distance)
- members of a set are more similar to one another than to the rest
- represented by a dendrogram

Approaches:

- agglomerative: unification from individual members (bottom-up)
- divisive: division towards members (top-down)

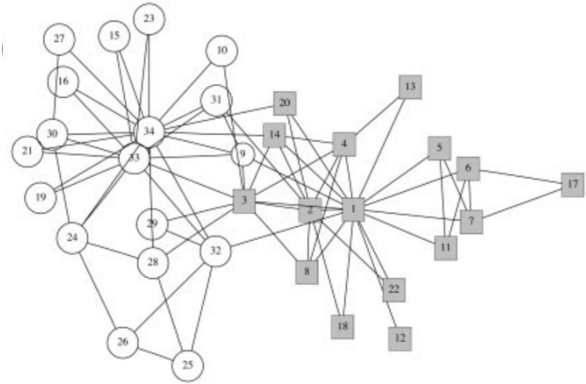
Hierarchical Cluster Analysis

In a network context, it is important to define similarity W_{ij}

Common options:

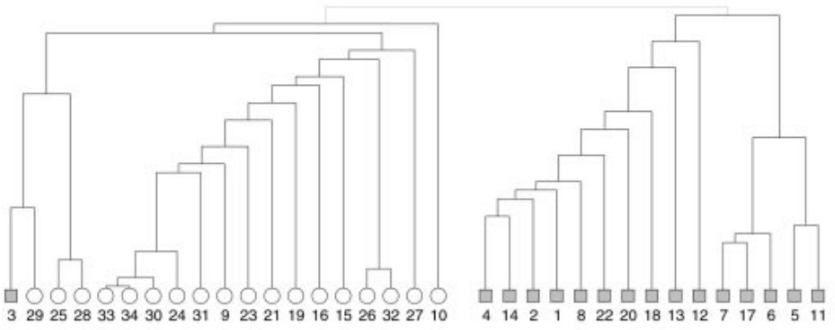
- number of node-independent paths between nodes i and j
  - must not share any other than terminal nodes
- number of link-independent paths between nodes i and j
  - every link must be included in no more than one path

Example: Zachary Karate Club [2]



[2] Girvan, M., & Newman, M. E. (2002)

Example: Zachary Karate Club [3]



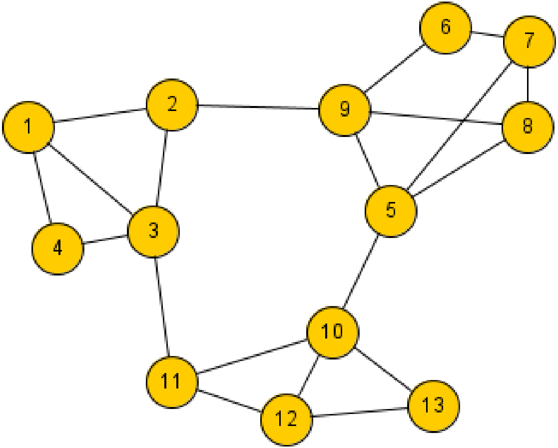
[3] Girvan, M., & Newman, M. E. (2002)

Betweenness Clustering

Core concept

- edges with high betweenness are considered bridges between communities
- progressively, edges with the highest betweenness are removed
- components obtained this way are considered to be communities

Betweenness Clustering



Modularity

Main idea:

- to create a division of nodes into groups C
- evaluate the division using function $Q(C)$
- find the maximum for Q

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

- where $P_{ij} = \frac{k_i k_j}{2m}$ is the probability of an edge between i and j
- $\delta(a, b) = 1 \iff a = b$
- $m = |E|$

Modularity: Properties

- Q indicates the degree of separation between communities
- for a random network, $Q = 0$
- computationally expensive, NP-complete problem
- optimization heuristics (such as simulated annealing, Louvain algorithm, Potts, Infomap)

Modularity: Efficient Algorithm[4]

Greedy approach:

- starts with isolated nodes
- gradually merges pairs of clusters to maximize ΔQ
- stop if merging any two clusters does not improve Q

Successfully applied to networks with $|V| > 400k$ (e.g. related items on Amazon).

[4] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical Review E, 70(6), 066111.

Modularity: resolution limit

Main problem:

- the null model is global: $\frac{k_i k_j}{2m}$
- in large networks, communities tend to have a more local character
- problems with communities of vastly different sizes

Solution: resolution limit

- $Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \gamma P_{ij}) \delta(C_i, C_j)$
- small γ favors more small communities
- large γ favors fewer larger communities

Local optimization[5]

Cluster evaluation function

- $f(C) = \frac{k_{int}}{(k_{ext}+k_{int})^{\alpha}}$
- k_{int} is the sum of internal degrees within the cluster
- k_{ext} is the sum of external degrees of the cluster
- α is the resolution parameter

[5] Lancichinetti et al., Detecting the overlapping and hierarchical community structure in complex networks, New Journal of Physics, 2009

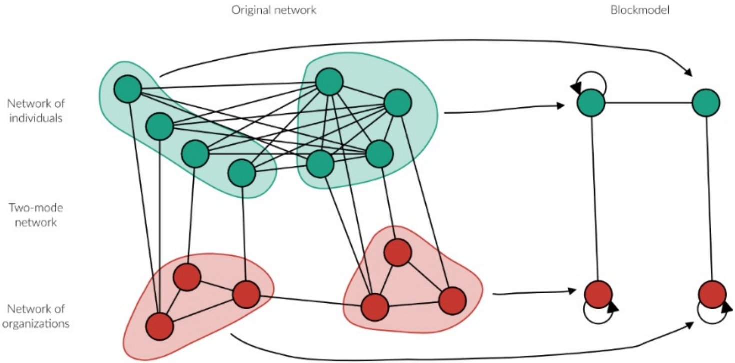
Local optimization[6]

Detection procedure:

- Start with a single node
- Add neighbors such that Δf is maximized
- At each step, test if removing a node could increase f
- Cluster is closed when adding a neighboring node does not increase f
- Start again with an unclassified node

[6] Lancichinetti et al., Detecting the overlapping and hierarchical community structure in complex networks, New Journal of Physics, 2009

Block Modeling

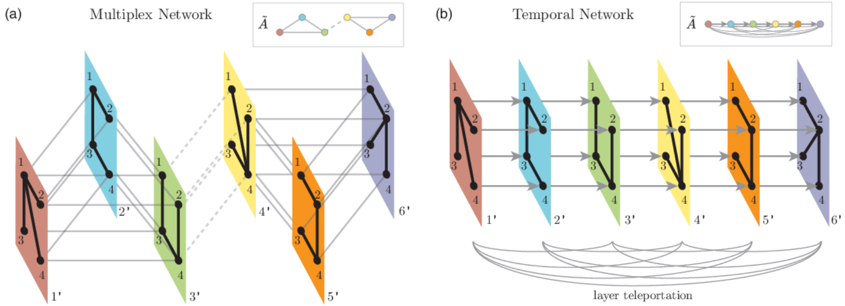


https://youngstats.github.io/post/2020/10/01/cugmas/

Multilayer Networks

Community detection successfully implemented in multilayer networks.

- multislice (multiplex) networks
- temporal network



Dane et al., Tunable Eigenvector-Based Centralities for Multiplex and Temporal Networks, ArXiv abs/1904.02059, 2019.

Testing of Clustering Algorithms

Assessing the quality of a specific algorithm is challenging[7]

- trade-off between generalizability and accuracy in a specific case
- obtaining training data with known community structure is difficult

[7] Yang & Leskovec, *Defining and evaluating network communities based on ground-truth.* Knowledge and Information Systems 42.1 (2015): 181-213.

Testing of Clustering Algorithms

LFR Benchmark[8]

- a set of synthetic networks with community structure
- various distributions of cluster sizes, degrees, and other network properties
- allows for comparing different algorithms on general networks

Case-specific surrogate benchmark networks

[8] Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008), *Benchmark graphs for testing community detection algorithms*. Physical Review E, 78(4), 046110.

MUNI
FACULTY
OF INFORMATICS