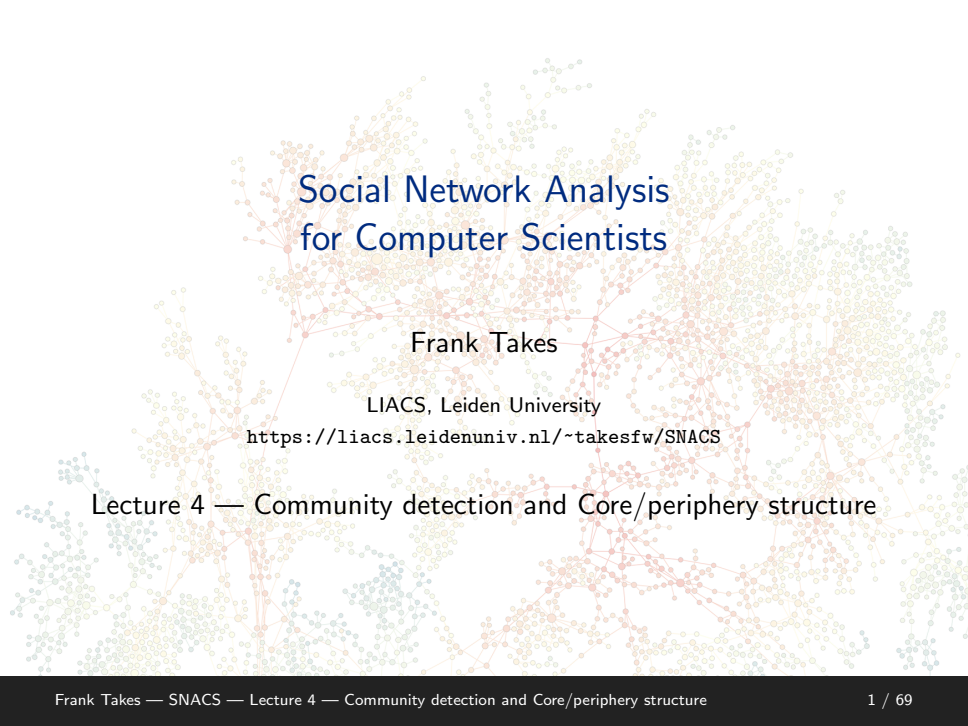




Social Network Analysis for Computer Scientists

Frank Takes

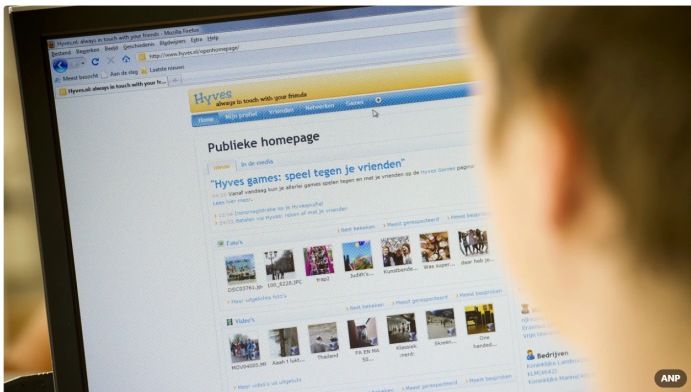
LIACS, Leiden University

<https://liacs.leidenuniv.nl/~takesfw/SNACS>

Lecture 4 — Community detection and Core/periphery structure

Today

- Recap
- Subgraphs and cliques
- Community detection
- Core/periphery structure
- LIACS REL Compute infrastructure



NOS Nieuws • Gisteren, 08:02

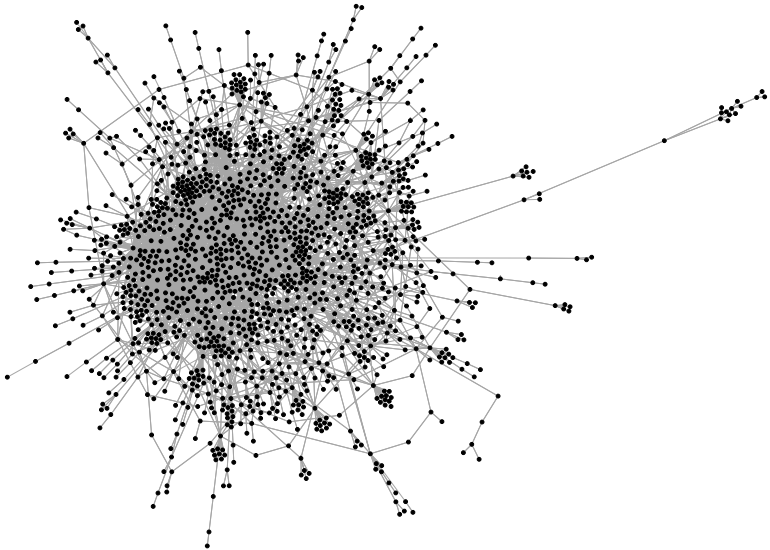


Pimp je profiel en 'krabbel' elkaar: Hyves maakt na 13 jaar een comeback

Hyves, ooit de grootste en populairste socialenetwerksite van Nederland, keert na 13 jaar terug. Mediahuis Nederland, krantenuitgever van onder meer De Telegraaf, brengt de voor velen nostalgische vriendenwebsite terug.

Recap

Networks



Notation

Concept

- Network (graph)
- Nodes (objects, vertices, ...)
- Links (ties, relationships, ...)
 - Directed — $E \subseteq V \times V$ — “links”
 - Undirected — “edges”
- Number of nodes — $|V|$
- Number of edges — $|E|$
- Degree of node u
- Distance from node u to v

Symbol

$G = (V, E)$

V

E

n

m

$\deg(u)$

$d(u, v)$

Real-world networks

- 1 Sparse networks density
- 2 Fat-tailed power-law degree distribution degree
- 3 Giant component components
- 4 Low pairwise node-to-node distances distance
- 5 Many triangles clustering coefficient
- Many examples: social networks, communication networks, citation networks, collaboration networks, protein interaction networks, information networks, webgraphs, financial networks, ...

Advanced concepts

- Reciprocity
- Complete graph, complement of a graph
- Planar graph, lattice, regular graph
- Tree and dag
- Subgraphs
- Ego network
- Spanning trees
- Diameter
- Bridges
- Power law exponent

Centrality measures

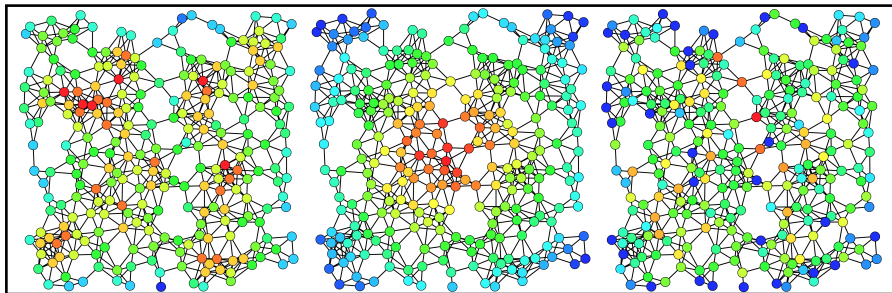
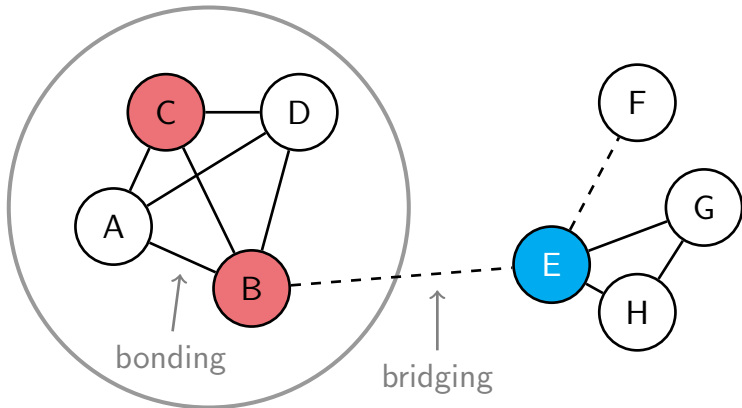


Figure: Degree, closeness and betweenness centrality.

Source: "Centrality" by Claudio Rocchini, Wikipedia File:Centrality.svg

Weak ties and strong ties



After F. van Tubergen, Utrecht University

Attribute assortativity

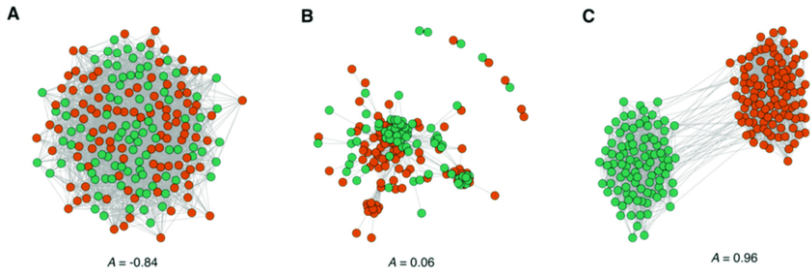


Figure: Attribute assortativity.

Image: Moya-García, A. et al. Identification of New Toxicity Mechanisms ... *Genes*, 13(7), 1292, 2022.

Projection

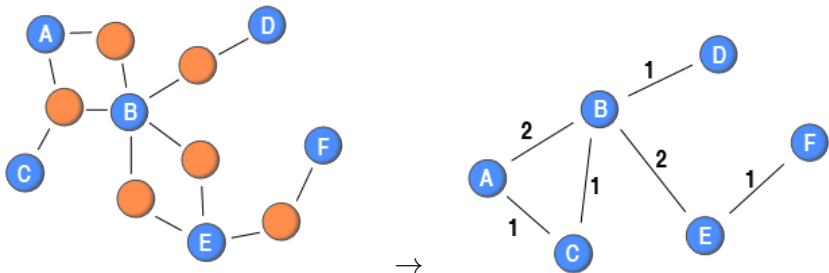


Image: <http://toreopsahl.com>

Subgraphs and cliques

Subgraphs

- Given a graph $G = (V, E)$, there are two ways in which subgraphs are commonly studied:
 - 1 **Subgraph** $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq (E \cap (V' \times V'))$
(subset of the nodes and edges of the original network, commonly used when defining communities or clusters)

Subgraphs

- Given a graph $G = (V, E)$, there are two ways in which subgraphs are commonly studied:
 - 1 **Subgraph** $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq (E \cap (V' \times V'))$
(subset of the nodes and edges of the original network, commonly used when defining communities or clusters)
Induced subgraph: subgraph with a set of nodes $V' \subseteq V$ and *all* edges between these nodes $E' = (E \cap (V' \times V'))$

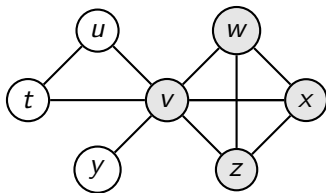
Subgraphs

- Given a graph $G = (V, E)$, there are two ways in which subgraphs are commonly studied:
 - 1 **Subgraph** $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq (E \cap (V' \times V'))$
(subset of the nodes and edges of the original network, commonly used when defining communities or clusters)
Induced subgraph: subgraph with a set of nodes $V' \subseteq V$ and *all* edges between these nodes $E' = (E \cap (V' \times V'))$
 - 2 **Subgraph** $G' = (V, E')$ with $E' \subseteq E$
(alternative definition where only edges are left out, commonly used when modelling network evolution; not today)

Subgraphs and cliques

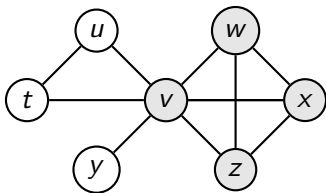
- Complete graph: graph with all edges present
- **Clique**: complete subgraph
- **Maximal clique**: complete subgraph of maximal size (cannot be extended with another node)
- **Maximum clique**: largest possible clique in the graph
- **Clique problem**: find the maximum clique(s) of a graph (one of Karp's 21 NP-complete problems introduced in 1972)

Subgraphs and cliques



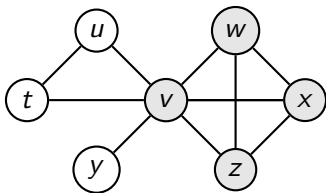
- Subgraph $V' = \{t, v, y\}$ is not complete

Subgraphs and cliques



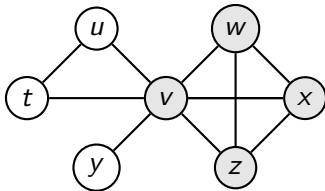
- Subgraph $V' = \{t, v, y\}$ is not complete
- Subgraph $V'' = \{t, u\}$ is complete and thus a clique of size 2

Subgraphs and cliques



- Subgraph $V' = \{t, v, y\}$ is not complete
- Subgraph $V'' = \{t, u\}$ is complete and thus a clique of size 2
- Subgraph $V''' = \{t, u, v\}$ is a maximal clique of size 3

Subgraphs and cliques



- Subgraph $V' = \{t, v, y\}$ is not complete
- Subgraph $V'' = \{t, u\}$ is complete and thus a clique of size 2
- Subgraph $V''' = \{t, u, v\}$ is a maximal clique of size 3
- Subgraph with $V'''' = \{v, w, x, z\}$ is a maximum clique of size 4

k -cliques

- A **k -clique** is a subgraph of k nodes with all $\frac{1}{2} \cdot k \cdot (k - 1)$ edges present
- An edge is a 2-clique, a triangle is a 3-clique
- In bipartite graphs, we can have a **biclique**: all nodes of one type connected to all nodes of the other

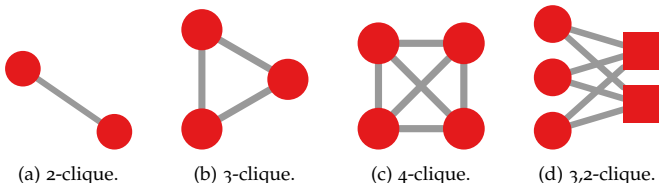


Figure: Different cliques.

Independent sets

- An **independent set** is the opposite of a clique: a set of nodes *none* of which are connected to each other
- Covering a graph with independent sets is the classical problem of **graph coloring**
- For example: how many colors do you need for a map such that no two neighboring countries share a color?

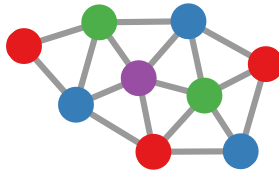


Figure: A graph divided into independent sets, one per node color.

Community detection

Related: clustering

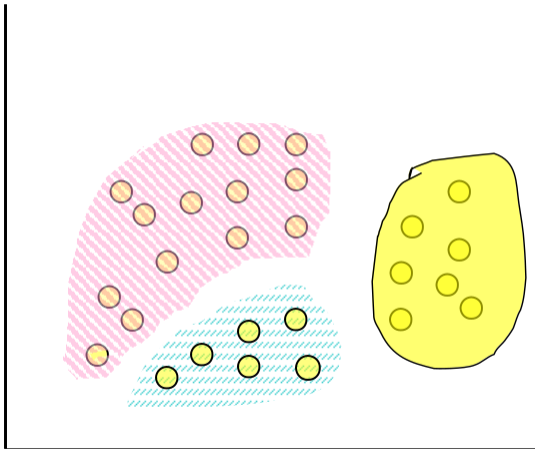
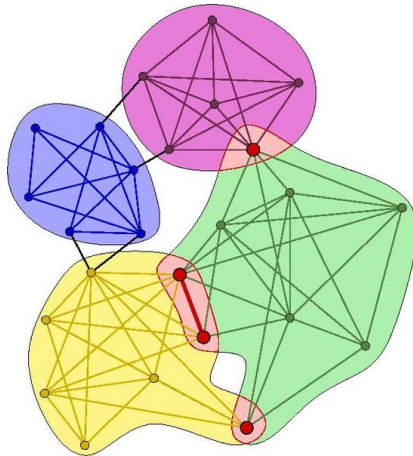


Image: KDnuggets - Clustering, 2009.

From cliques to community detection



Source: J. Leskovec, Community detection workshop (1), 2014.

Communities as partitions

- **Community**: groups of nodes densely connected to each other and sparsely connected to nodes outside the group
- A **partition** assigns every node to exactly one community
- In the adjacency matrix, communities are dense blocks on the diagonal (if the matrix is ordered in a smart way)

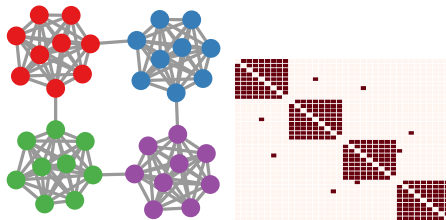
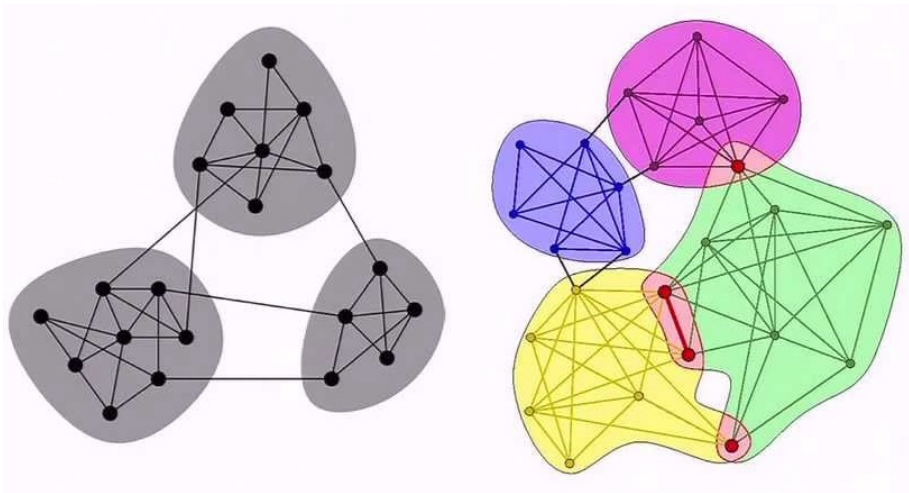


Figure: Community structure (left) and corresponding adjacency matrix (right).

Partitions vs. communities



J. Leskovec, Affiliation Network Models for Densely Overlapping Communities, MMDS 2012.

Community detection

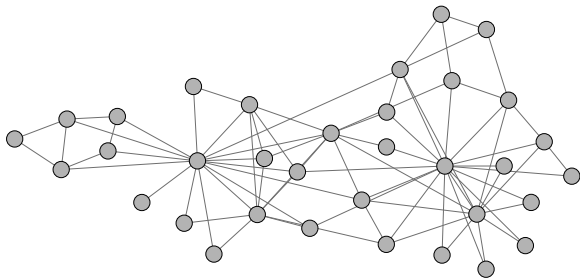


Figure: Communities: node subsets connected more strongly with each other.

Community detection

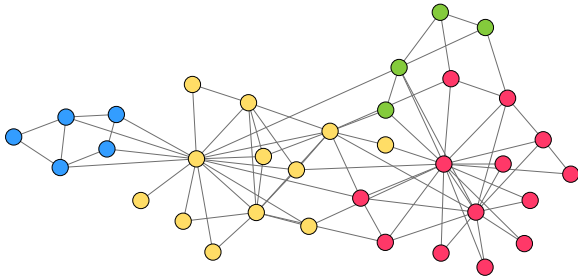


Figure: Communities: node subsets connected more strongly with each other.

The Belgian call graph

- Nodes are cellphones, connected when they call each other often
- Belgium: half the population speaks French and half speaks Dutch
- In the capital, Brussels, both languages are spoken

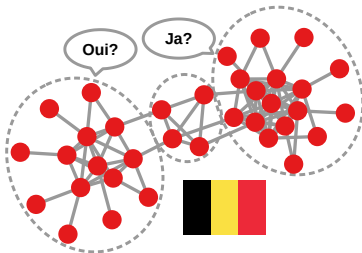


Figure: Belgian cellphone call graph.

Community detection algorithms

- Clique-based methods (inefficient; not covered)
- Divisive algorithms
- Modularity maximization algorithms
- Label propagation
- Stochastic Block Model (SBM; covered later)
- Random walk-based approaches

Remember from last week?

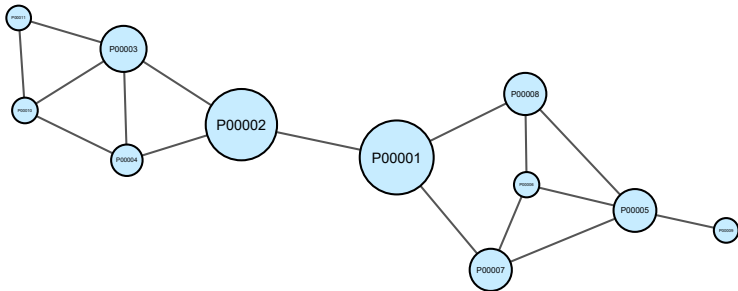


Figure: Betweenness centrality reveals globally important criminals.

Remember from last week?

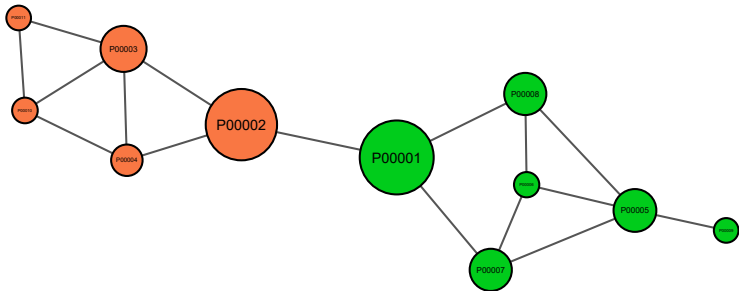


Figure: Community detection finds groups of tightly connected criminals.

Divisive algorithm: Girvan–Newman

- **Edge betweenness**: fraction of shortest paths running through an edge

Divisive algorithm: Girvan–Newman

- **Edge betweenness**: fraction of shortest paths running through an edge
- The algorithm is **divisive** because it starts from the whole network and repeatedly:
 - 1 Computes the edge betweenness of every edge
 - 2 Removes the edge with the highest value
 - 3 **Recomputes** the betweenness of all remaining edges

Divisive algorithm: Girvan–Newman

- **Edge betweenness**: fraction of shortest paths running through an edge
- The algorithm is **divisive** because it starts from the whole network and repeatedly:
 - 1 Computes the edge betweenness of every edge
 - 2 Removes the edge with the highest value
 - 3 **Recomputes** the betweenness of all remaining edges
- The result is a dendrogram, cut at desired level, e.g., using a quality measure Q (or stop early as Q diminishes)
- One betweenness computation is $O(mn)$, repeated for each of the m removals makes $O(m^2n)$

M. Girvan and M.E.J. Newman, Community structure in social and biological networks, in *PNAS* 99(12): 7821–7826, 2002.

Girvan–Newman on a small example

- Removal 1 deletes the edge (5, 6)
- Removal 2 deletes the edge (10, 11)
- etc.

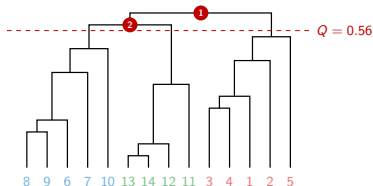
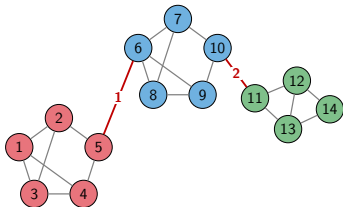


Figure: First two iterations (left) and corresponding dendrogram (right).

Quality measure: Modularity

- **Modularity** Q weighs the edges inside the communities against the number expected if the edges were rewired at random:

$$Q = \frac{1}{2m} \sum_{u,v \in V} \left(A_{uv} - \frac{\deg(u) \cdot \deg(v)}{2m} \right) \delta(c_u, c_v)$$

- $A_{uv} = 1$ if u and v are connected, and 0 otherwise
- $\frac{\deg(u) \cdot \deg(v)}{2m}$: how often u and v would be connected in a random network with those same degrees
- $\delta(c_u, c_v) = 1$ if u and v are in the same community, and 0 otherwise, so only pairs inside a community count

Modularity

- **Modularity** compares the observed network with that of a network with the same node and edge count and degrees, rewired at random

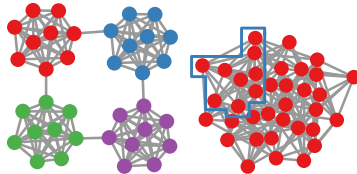


Figure: Network with communities (left) and rewired network (right).

- Resolution parameter r indicating how “tough” the algorithm should look for communities
- Algorithms for community detection optimize (maximize) modularity score Q given some r (using local search, heuristics, hill climbing, genetic algorithms or other optimization techniques)

The domain of modularity

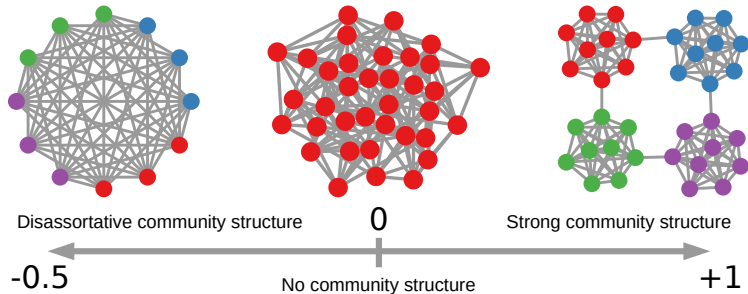


Figure: Modularity runs from -0.5 for disassortative communities, through 0 for no community structure, to $+1$ for a strong community structure.

Modularity maximization: Louvain

- The Louvain algorithm **greedily** maximizes Q in two repeated alternating phases:
 - 1 **Vertex moving**: put every node in its own community. For each node u , evaluate the gain ΔQ of moving u into the community of each of its neighbors and apply the best move if it is positive. Sweep over all nodes until no move improves Q
 - 2 **Graph coarsening**: collapse every community into a single node. Edges between two communities become one weighted edge, edges inside a community a weighted self-loop

Modularity maximization: Louvain

- The Louvain algorithm **greedily** maximizes Q in two repeated alternating phases:
 - 1 **Vertex moving**: put every node in its own community. For each node u , evaluate the gain ΔQ of moving u into the community of each of its neighbors and apply the best move if it is positive. Sweep over all nodes until no move improves Q
 - 2 **Graph coarsening**: collapse every community into a single node. Edges between two communities become one weighted edge, edges inside a community a weighted self-loop
- Output is a **hierarchy** of communities, pick the one with highest Q

Modularity maximization: Louvain

- The Louvain algorithm **greedily** maximizes Q in two repeated alternating phases:
 - 1 **Vertex moving**: put every node in its own community. For each node u , evaluate the gain ΔQ of moving u into the community of each of its neighbors and apply the best move if it is positive. Sweep over all nodes until no move improves Q
 - 2 **Graph coarsening**: collapse every community into a single node. Edges between two communities become one weighted edge, edges inside a community a weighted self-loop
- Output is a **hierarchy** of communities, pick the one with highest Q
- In practice, $O(m)$
- Greedy and order-dependent: two runs on the same network can return different partitions

V.D. Blondel, J-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, in *Journal of Statistical Mechanics: Theory and Experiment* 10: P10008, 2008.

Modularity maximization: Louvain

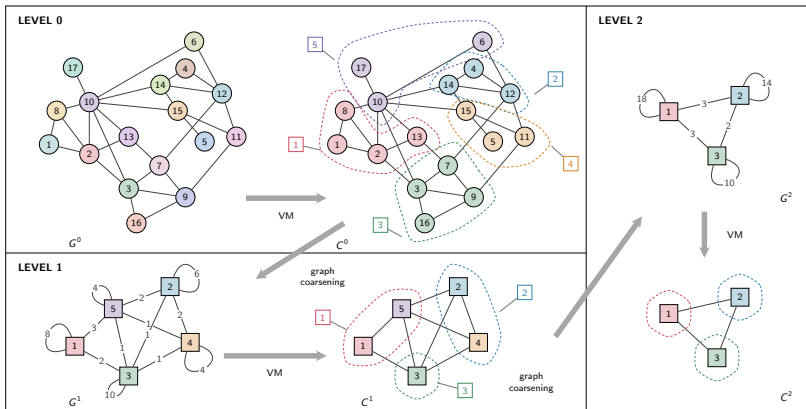


Figure: Louvain algorithm: vertex moving (VM) and graph coarsening.

Resolution parameter in modularity maximization

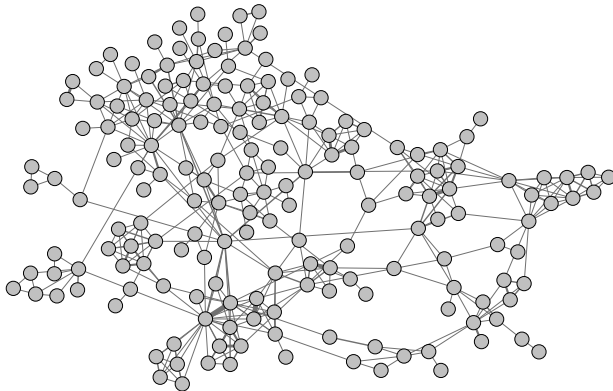


Figure: Do you see communities?

Resolution = 1.0

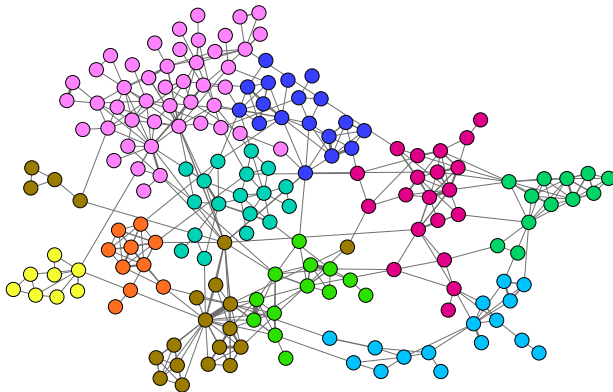


Figure: Modularity = 0.747; 10 communities.

Resolution = 2.0

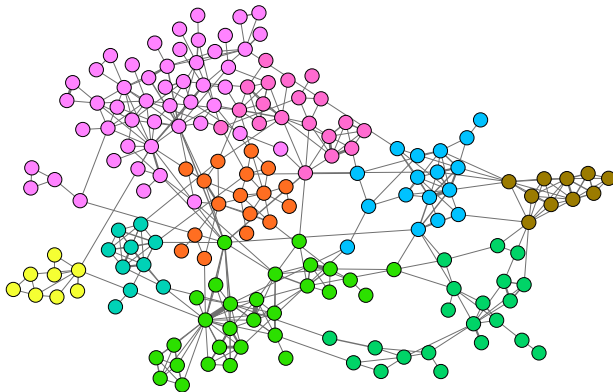


Figure: Modularity = 0.732; 8 communities.

Resolution = 4.0

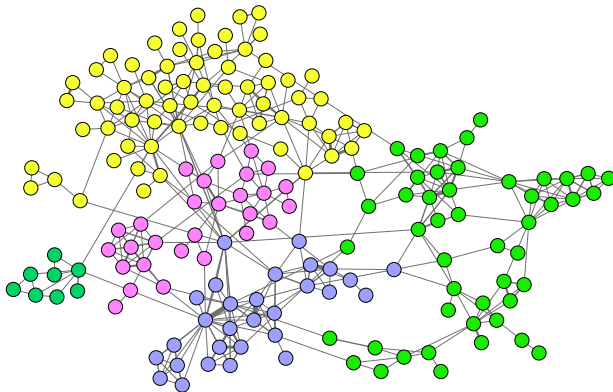


Figure: Modularity = 0.641; 5 communities.

Alternative quality measure: Conductance

- **Conductance** weighs the edges leaving a community C against the edges inside it (should be *minimized* over all communities):

$$f(C) = \frac{|E_{B,C}|}{2|E_C| + |E_{B,C}|}$$

- $E_C = \{(u, v) \in E : u \in C \wedge v \in C\}$: the edges **inside** the community
- $E_{B,C} = \{(u, v) \in E : u \in C \wedge v \notin C\}$: the edges on its **boundary**
- Left network: $|E_C| = 21$ and $|E_{B,C}| = 4$, so $f(C) \approx 0.09$
- Right network: $|E_C| = 10$ and $|E_{B,C}| = 7$, so $f(C) \approx 0.26$

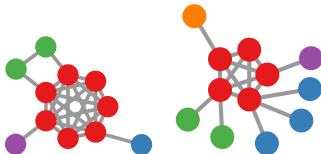


Figure: A community with low conductance (left) and high conductance (right).

Detection by label propagation

- Start with a label per node, distributed at random
- Every node repeatedly adopts the label most common among its neighbors, breaking ties at random
- Labels percolate until nothing changes any more; nodes in a community end up surrounded by their own label

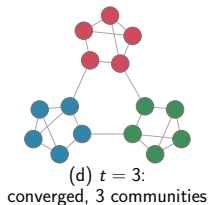
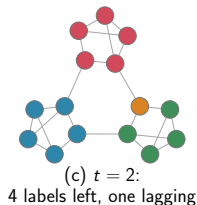
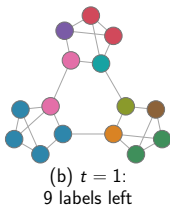
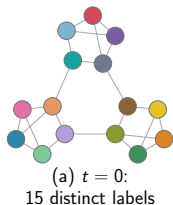


Figure: Label propagation.

Community detection by random walk

- Idea: a random walker rarely leaves the community it is visiting
- Nodes that keep appearing together in a walk are likely to be in the same community
- **Infomap**: the best partition is the one that describes such a walk in the fewest bits
- See the Coscia book for beautiful details

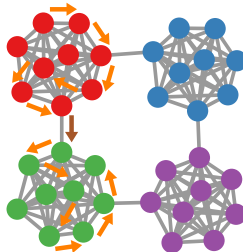


Figure: Random walk of 13 steps; the brown edge crosses communities.

Consensus clustering

- Two problems:
 - 1 One community detection method may produce different partitions
 - 2 Different methods produce different community divisions

Consensus clustering

- Two problems:
 - 1 One community detection method may produce different partitions
 - 2 Different methods produce different community divisions
- **Consensus clustering**: combine the answers of multiple runs, possibly of different algorithms. Given a weighted graph:
 - 1 Run the algorithm(s) n_p times, giving n_p partitions
 - 2 Build the **consensus matrix** D : D_{uv} is the fraction of those partitions in which u and v share a community
 - 3 Set entries below a threshold τ to zero, dropping accidental agreement
 - 4 Set D as the graph and go to step 1, until the partitions no longer change

A. Lancichinetti and S. Fortunato, Consensus clustering in complex networks, in *Scientific Reports* 2: 336, 2012.

Consensus clustering

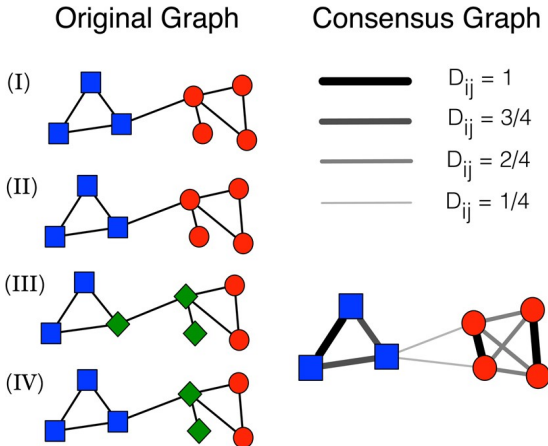


Figure: Consensus clustering.

Evaluating communities and partitions

- If node attributes carry information about the true communities, the partition and the attributes should agree
- Human interpretation by hand can suffer from subjective bias
- **Normalized mutual information** (NMI) measures that agreement on a scale from 0 to 1

Evaluating communities and partitions

- If node attributes carry information about the true communities, the partition and the attributes should agree
- Human interpretation by hand can suffer from subjective bias
- **Normalized mutual information** (NMI) measures that agreement on a scale from 0 to 1
- **Adjusted mutual information** (AMI) “subtracts” the expected mutual information that two random partitions with the same group sizes already share



Figure: Two vectors drawn independently, yet with a positive NMI.

Communities in corporate networks

Corporate board interlocks

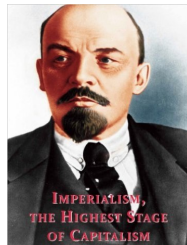
- **Nodes** are organizations/firms/companies/corporations

Corporate board interlocks

- **Nodes** are organizations/firms/companies/corporations
- **Edges** are **board interlocks**: relationships between firms because they share a board member or director

Corporate board interlocks

- **Nodes** are organizations/firms/companies/corporations
- **Edges** are **board interlocks**: relationships between firms because they share a board member or director
- Vladimir I. Lenin, *Imperialism, The Highest Stage of Capitalism*, 1916.
- "... a personal union, so to speak, is established between the banks and the biggest industrial and commercial enterprises, the merging of one with another through the acquisition of shares, through the appointment of bank directors to the Supervisory Boards (or Boards of Directors) of industrial and commercial enterprises, and vice versa."



Board interlocks

- **Causes** of interlocks:

- Collusion
- Cooptation and monitoring
- Legitimacy
- Career advancement
- Social cohesion

- **Consequences** of interlocks:

- Corporate control
- Economic performance
- Access to resources

Board interlocks

■ Causes of interlocks:

- Collusion
- Cooptation and monitoring
- Legitimacy
- Career advancement
- Social cohesion

■ Consequences of interlocks:

- Corporate control
- Economic performance
- Access to resources



M. Mizruchi, What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates, *Annual Review of Sociology* 22: 271–298, 1996.

Corporate city networks

- Nodes are cities
- Edges between cities are based on firms sharing directors
- Weights on edges denote the number of connections
- Each city has an associated country
- Provides insight in geographical orientation of global economy

Corporations



Corporate network



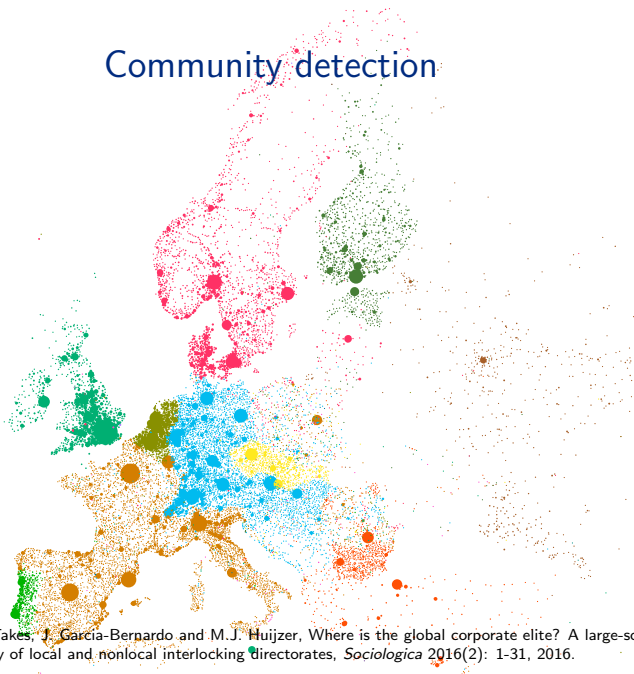
F.W. Takes and E.M. Heemskerk, Centrality in the Global Network of Corporate Control, *Social Network Analysis and Mining* 6(1): 1-18, 2016.

Community detection



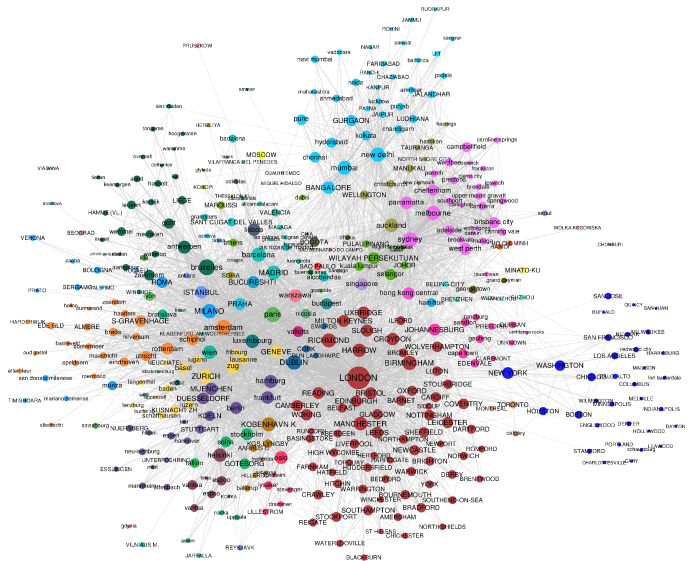
E.M. Heemskerk and F.W. Takes, The Corporate Elite Community Structure of Global Capitalism, *New Political Economy* 21(1): 90-118, 2016.

Community detection



E.M. Heemskerk, F.W. Takes, J. Garcia-Bernardo and M.J. Huijzer, Where is the global corporate elite? A large-scale network study of local and nonlocal interlocking directorates, *Sociologica* 2016(2): 1-31, 2016.

Nodes colored by country (sample)



Community composition

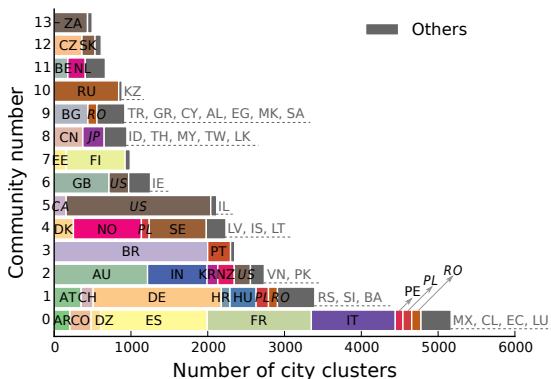


Figure: Country involved in each community.

Communities in scientific co-citation networks

Co-citation network

- Nodes are scientific publications
- Edges indicate that papers cite the same previous work
- Each node has an associated scientific field
- Network provides insight in how scientific fields interact

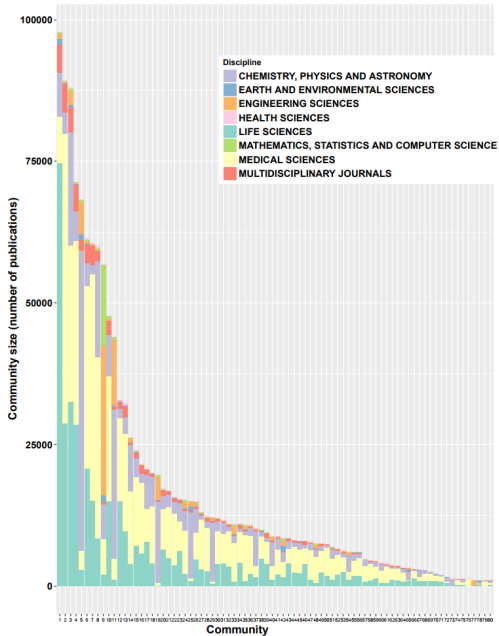
Co-citation network

- Nodes are scientific publications
- Edges indicate that papers cite the same previous work
- Each node has an associated scientific field
- Network provides insight in how scientific fields interact
- Size: 1.6 million nodes and 44 million edges
- 99% in giant component, scale-free, small-world

Co-citation network

Discipline	Number of publications
MEDICAL SCIENCES	550672.68
LIFE SCIENCES	403633.85
CHEMISTRY, PHYSICS AND ASTRONOMY	293971.77
ENGINEERING SCIENCES	66186.33
MULTIDISCIPLINARY JOURNALS	55394.00
MATHEMATICS, STATISTICS AND COMPUTER SCIENCE	23192.52
EARTH AND ENVIRONMENTAL SCIENCES	10596.43
HEALTH SCIENCES	5043.42

Figure: Categories of publications (weighting applied if multiple apply).



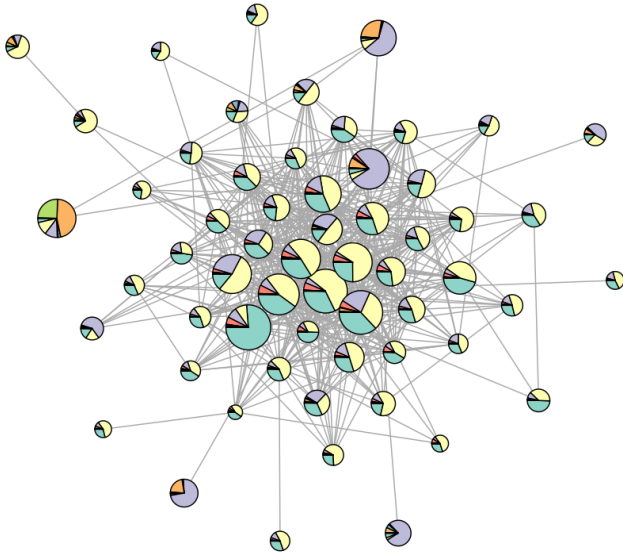


Figure: Community composition and connections.

Core/periphery structure

Core/periphery structure

- Many large networks have a densely connected **core** and a **periphery** of nodes attaching to only a few neighbors
- The core is also called the **rich club**

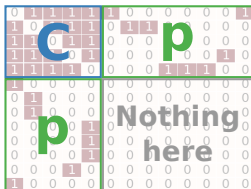


Figure: Adjacency matrix of the **discrete model**: a dense core (blue), a sparse periphery (green), and nothing at all between peripheral nodes.

Two common structures

- **Core/periphery structure**: there is **one** dense area
- **Community structure**: there are **several**, with little periphery
- Nodes deep inside a community have a high degree, so a core/periphery algorithm wants to put them all in one core . . .

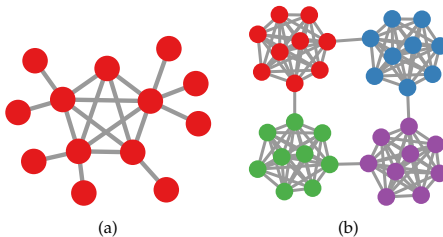


Figure: (a) Core/periphery structure and (b) Community structure.

Communities, core/periphery, or both?

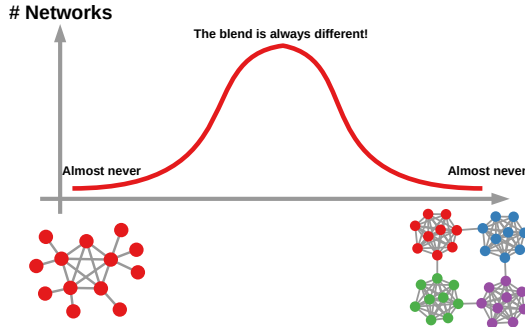


Figure: Almost no network is a pure core/periphery or a pure community structure.

Hotelling's law; a possible explanation for core formation

- Two ice cream vendors of equal quality on a beach; every customer walks to the nearest one
- Each vendor moves to capture the customers in the middle, and the equilibrium has both of them in the center
- Nobody sells more, but every customer walks further: cores emerge from ordinary rational behavior

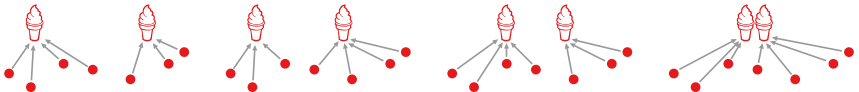


Figure: As time goes by, the stands move closer and closer together.

LIACS REL Compute
<https://rel.liacs.nl/compute>

Upcoming lab session and “Homework”

- Work on Assignment 1
- Ask questions!

Credits

- Several images are gratefully reused from Michele Coscia, The Atlas for the Aspiring Network Scientist, v3 of <https://arxiv.org/abs/2101.00863>