Community Detection

- Definitions
- Applications
- Measures for identifying communities (clusters)
  - Vertex similarity
  - Cluster fitness
- Hierarchical clustering
  - Hierarchical clustering: similarity-based, edge betweenness, modularity optimization
  - K-means
- Graph partitioning
  - Spectral bisection
  - Kernighan-Lin algorithm
- Other methods: spectra algorithms, dynamic algorithms, etc.

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Community Detection

- Belgium appears to be the model bi-cultural society: 59 % of its citizens are Flemish, speaking Dutch and 40 % are Walloons who speak French
- How did this country foster the peaceful coexistence of these two ethnic groups since 1830? Is Belgium a densely knitted society, where it does not matter if one is Flemish or Wallon? or we have two nations within the same borders, that learned to minimize contact with each other?
- The answer was provided by Vincent Blondel and his students in 2007, who developed an algorithm to identify the country's community structure
- They started from the mobile call network, placing individuals next to whom they regularly called on their mobile phone
- The algorithm revealed that Belgium's social network is broken into two large clusters of communities and that individuals in one of these clusters rarely talk with individuals from the other cluster

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Communities in Belgium Social Network

The origin of this separation became obvious once they assigned to each node the language spoken by each individual: French speakers and Dutch speakers clusters

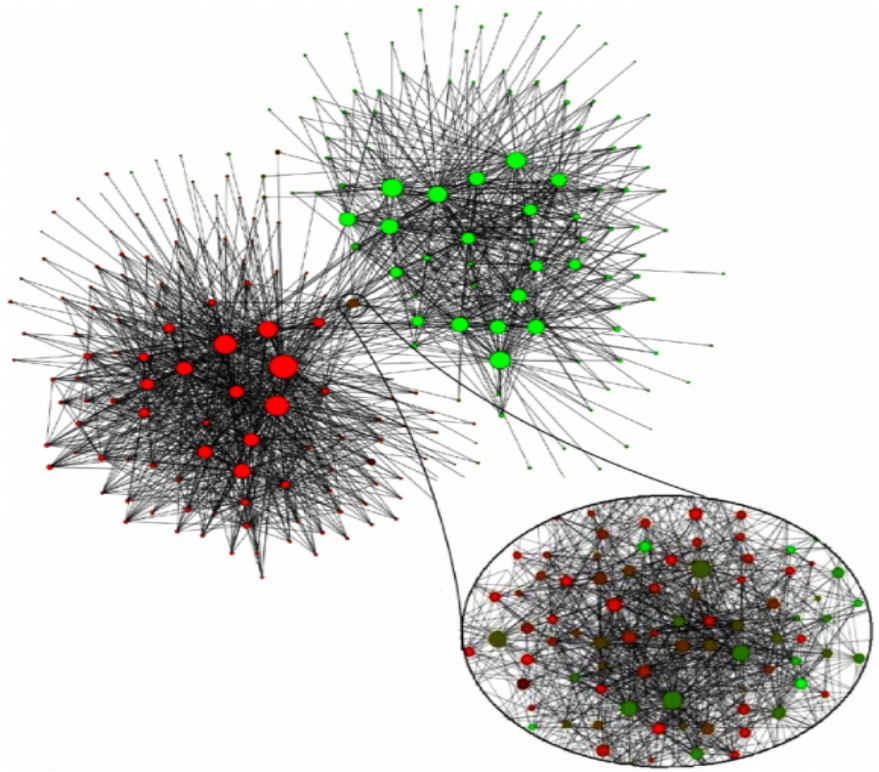
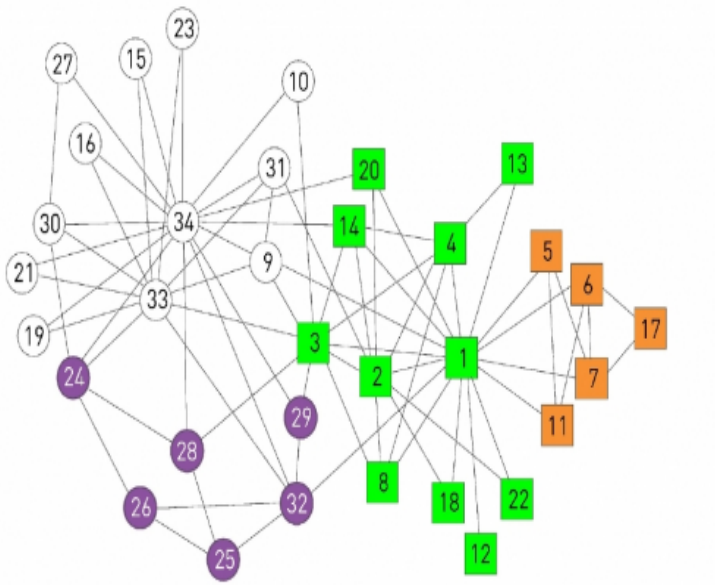


Figure: Communities extracted from the call pattern of the consumers of the largest Belgian mobile phone company.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

A social network that has received particular attention in community detection

- Zachary's Karate Club captures the links between 34 members of a karate club
- To uncover the true relationships between club members, sociologist Wayne Zachary documented 78 pairwise links between members who regularly interacted outside the club
- The interest in the dataset is driven by a singular event: A conflict between the club' s president and the instructor split the club into two.



Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

A social network that has received particular attention in the context of community detection

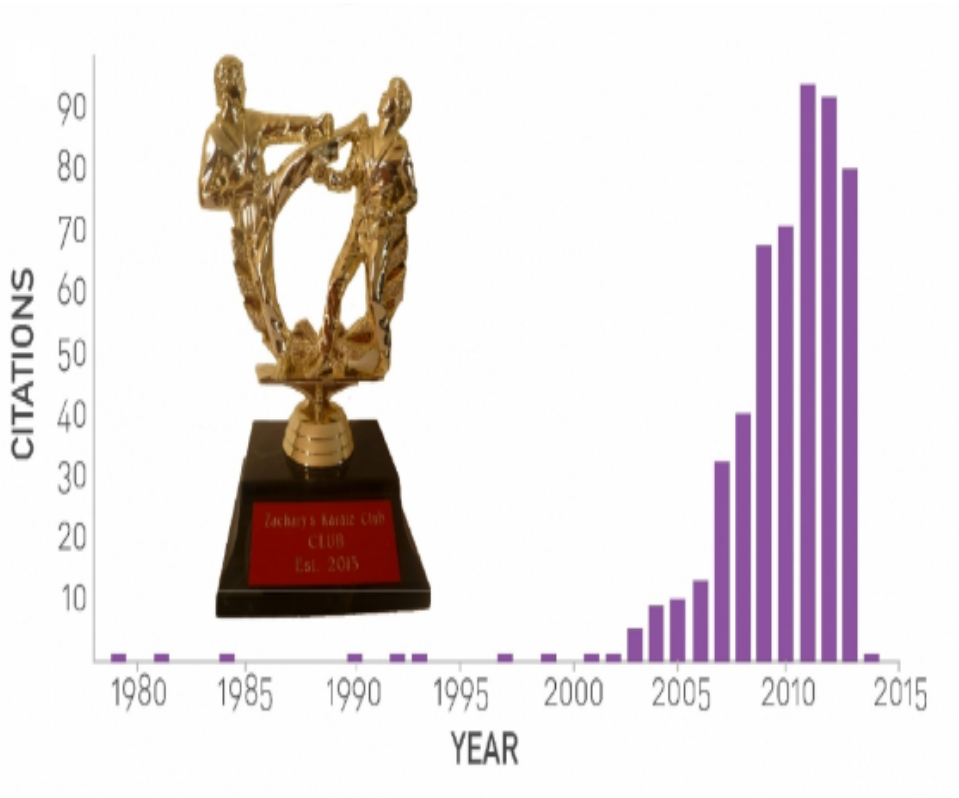


Figure: The citation history of the Zachary karate club paper mirrors the history of community detection in network science. Indeed, there was virtually no interest in Zachary' s paper until Girvan and Newman used it as a benchmark for community detection in 2002. Since then the number of citations to the paper exploded, reminiscent of the citation explosion to Erdos and Renyi's work following the discovery of scale-free networks

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Community Detection in Biological Networks

- Communities play a particularly important role in our understanding of how specific biological functions are encoded in cellular networks.
- Two years before receiving the Nobel Prize in Medicine, Lee Hartwell argued that biology must move beyond its focus on single genes.
- It must explore instead how groups of molecules form functional modules to carry out a specific cellular functions
- Communities play a particularly important role in understanding human diseases. Indeed, proteins that are involved in the same disease tend to interact with each other

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

The purpose of this lecture

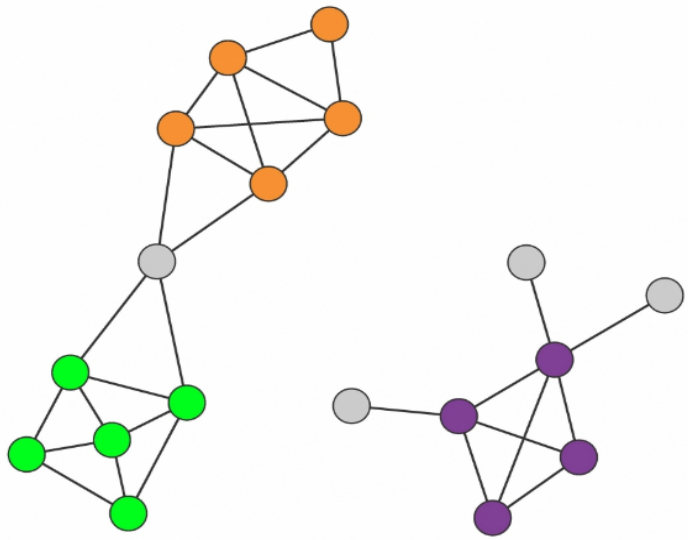
- The purpose of this chapter is to introduce the concepts necessary to understand and identify community structure of a complex network
- We will describe how communities are defined
- We will explore the various community characteristics
- We will introduce a couple of algorithms, relying on different principles, for community identification
- So we need to answer
  - what do we really mean by a community?
  - how many communities are in a network?
  - how many different ways can we partition a network into communities?

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Defining Communities : Connectedness and Density Hypothesis

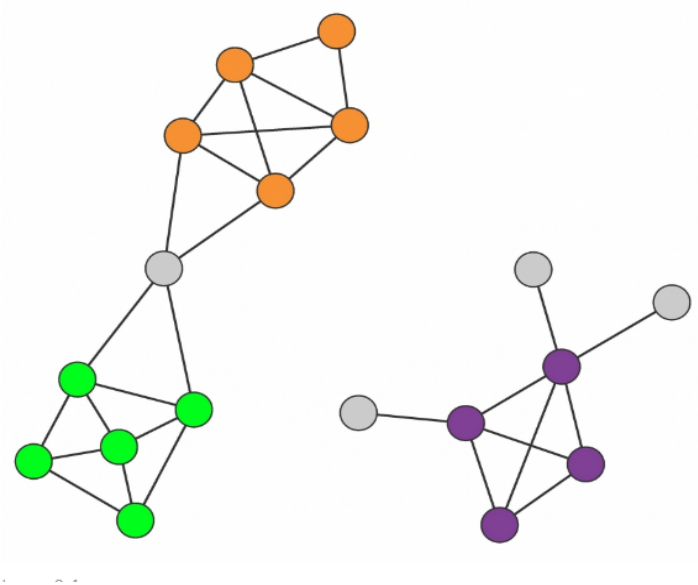
- Connectedness Hypothesis
  - each community corresponds to a connected subgraph, like the subgraphs formed by the orange, green or the purple nodes
  - if a network consists of two isolated components, each community is limited to only one component
  - this hypothesis also implies that on the same component a community cannot consists of two subgraphs that do not have a link to each other



Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Defining Communities : Connectedness and Density Hypothesis

- Density Hypothesis
  - nodes in a community are more likely to connect to other members of the same community than to nodes in other communities
  - the orange, the green and the purple nodes satisfy this expectation



Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Defining Communities : Connectedness and Density Hypothesis

- A community is a locally dense connected subgraph in a network
  - all members of a community must be reached through other members of the same community (connectedness)
  - at the same time we expect that nodes that belong to a community have a higher probability to link to the other members of that community than to nodes that do not belong to the same community (density)
  - While this hypothesis considerably narrows what would be considered a community, it does not uniquely define it

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Defining Communities : Maximum Cliques

- Luce and Perry in 1949 defined a community as a group of individuals whose members all know each other
- In a graph theoretic terms this means that a community is a complete subgraph, or a clique
- A clique automatically satisfies - it is a connected subgraph with maximal link density
- Viewing communities as cliques has several drawbacks
  - while triangles are frequent in networks, larger cliques are rare
  - requiring a community to be a complete subgraph may be too restrictive, missing many other legitimate communities

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Defining Communities : Strong and Weak Communities

- To relax the rigidity of cliques, consider a connected subgraph C of nodes N_C nodes in a network
- The internal degree k_i^{int} of node i is the number of links that connect i to other nodes in C
- The external degree k_i^{ext} is the number of links that connect i to the rest of the network
  - if $k_i^{ext} = 0$, each node i should be assigned to a different community
- These definitions allow us to distinguish two kinds of communities
  - **Strong community**: C is a strong community if each node within C has more links within the community than with the rest of the graph. Specially, a subgraph C forms a strong community if for each node $i \in C$, $k_i^{int}(C) > k_i ext(C)$
  - **Weak community**: C is a weak community if the total internal degree of a subgraph exceeds its total external degree. Specially, a subgraph C forms a weak community if $\sum_{i \in C} k_i^{int}(C) > \sum_{i \in C} k_i^{ext}(C)$

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Defining Communities : Strong and Weak Communities

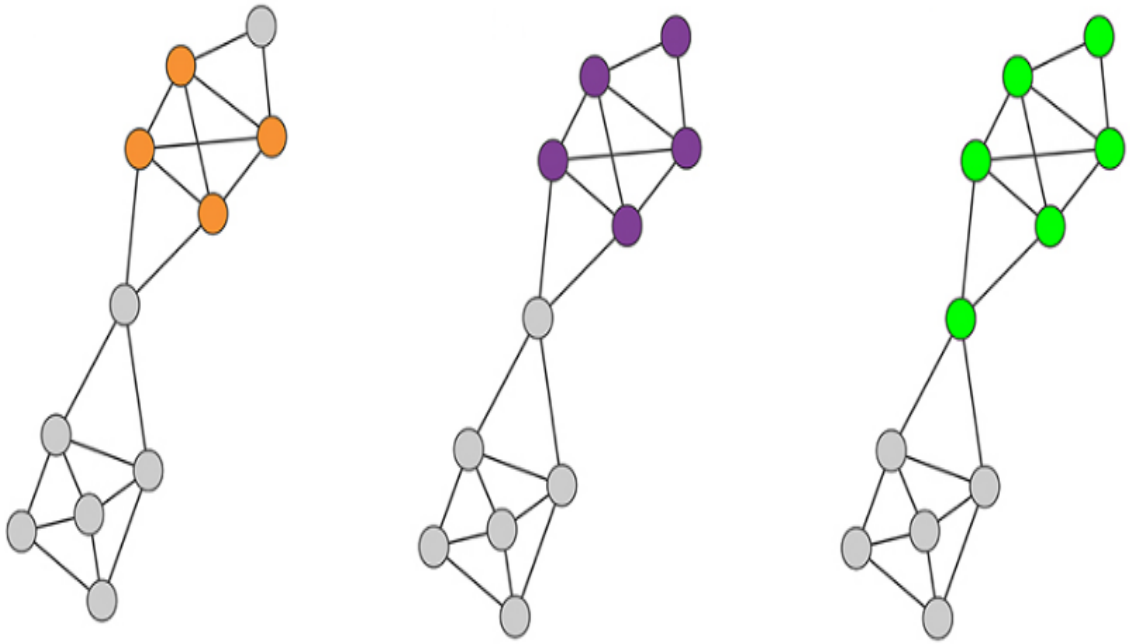
- Each clique is a strong community, and each strong community is a weak community - the converse is generally not true
- The community definitions discussed above (cliques, strong and weak communities) refine our notions of communities
- At the same time they indicate that we do have some freedom in defining communities
- The higher order clique of this network is a square, shown in orange - there are several three-node cliques on this network, can you find them?
- A strong community is shown in purple; there are additional strong communities on the graph, can you find at least two more?
- The green nodes represent one of the several possible weak communities of this network

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Defining Communities

- The higher order clique of this network is a square, shown in orange - there are several three-node cliques on this network, can you find them?
- A strong community is shown in purple; there are additional strong communities on the graph, can you find at least two more?
- The green nodes represent one of the several possible weak communities of this network



Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Communities of Users for Twitter

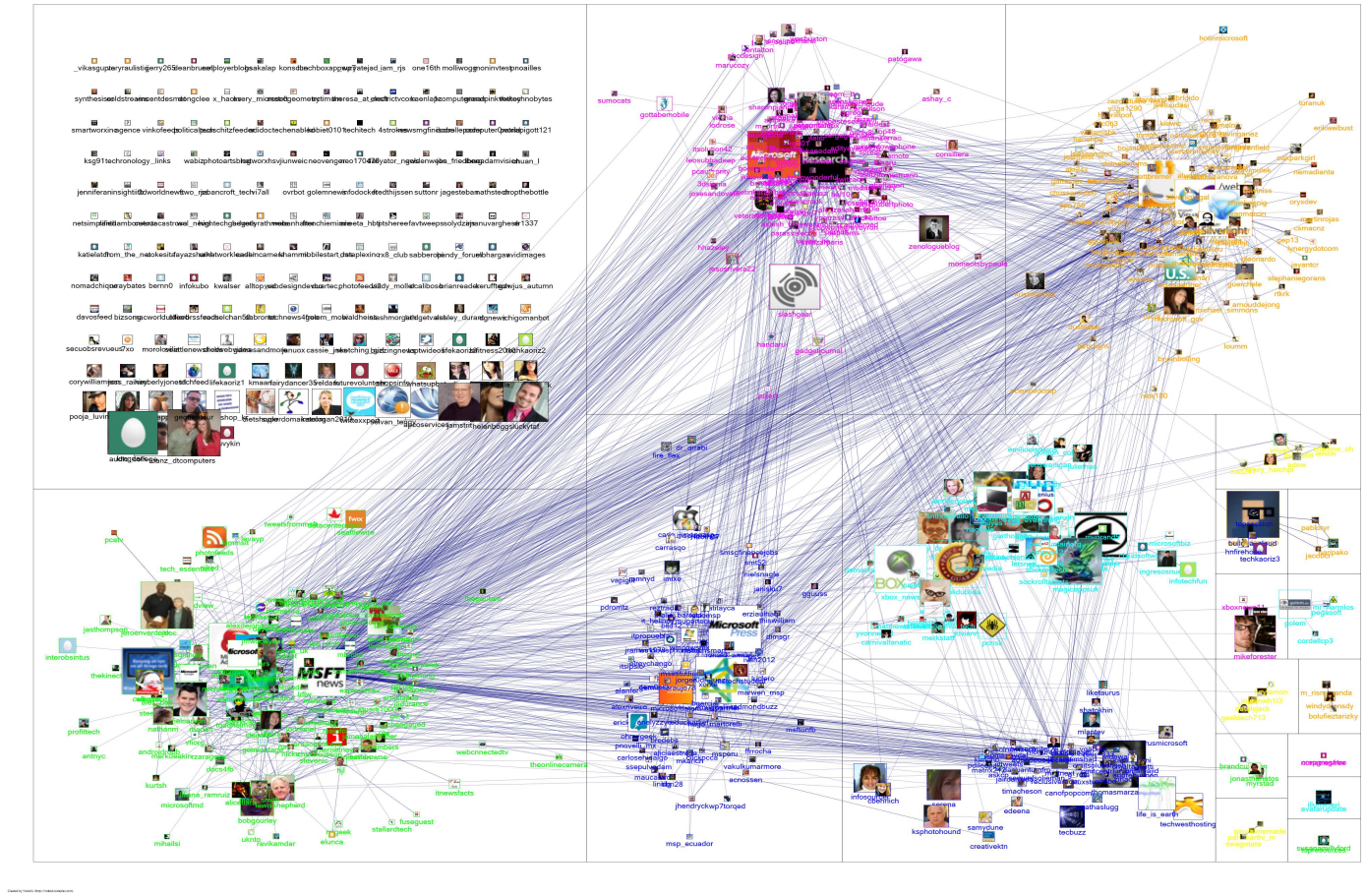


Figure: Clustered network of Twitter users. The clusters are distributed in a tree map structure. Harel-Koren layout is used for the cluster visualization inside each box.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Applications

- In social network analysis
  - Detect the communities
  - Analysis of the evolution of communities
- In biology
  - Group genes having the same functions
  - Study of diseases spreading and counter solutions
- In WWW:
  - Group similar pages
  - Identify topics
- In recommender systems
  - Improve the performance and accuracy by user's communities
- Other applications in business and market analysis, social science, transportation, etc.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Definitions

- Data clustering in data mining: find a partition of the given data set into subsets which share some common properties, e.g. some distance measures
- Network clustering: given a network $\Gamma = (V, E)$, group vertices into clusters $V_1, V_2, ..., V_k$ such that the clustering satisfies some properties
- Non-overlapping clustering
  - Each node only belongs to one cluster, e.g. $V_i \cap V_j = \oslash$, for any $i \neq j$ and $i, j \in \{1, ..., k\}$
- Overlapping clustering
  - Nodes can belong to several clusters, e.g. $V_i \cap V_j \neq \oslash$, for some $i \neq j$ and $i, j \in \{1, ..., k\}$

cc BY SA

AC5 RWTH
Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Community Definitions (1)

- Community: There are many edges *within* communities (larger density) and relatively few edges *between* communities
- Local definitions: consider the vertices of the subgraph and its immediate neighborhood
  - Mutual cohesion of vertices of the community is compared with the their cohesion with the external neighbors
  - Consider some classes of subgraphs:
    - Cliques: complete subgraphs
    - n-cliques: the distance of each pair of vertices is not larger than n
    - k-plex: each vertex is connected to all other vertices except k of them
    - k-core: each vertex is connected to at least k vertices
    - Strong community: a subgraph where each vertex has more neighbors inside then outside the subgraph
    - Weak community: total degree of vertices inside is greater than the external degree

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Community Definitions (2)

- Global definitions
  - Subgraphs are analyzed with respect to the whole graph
  - *Null model*: the subgraph which has the same topological features, but has no community structure
  - The linking properties of the community is compared with the corresponding *null models*
  - Popular null model: the modularity by Newman and Girvan used randomness as a null model
- Vertex similarity
  - Communities are groups of similar vertices
  - Feature-based measures, path-based measures, etc.
  - Basic of hierarchical clustering

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Community Structure in Network Models

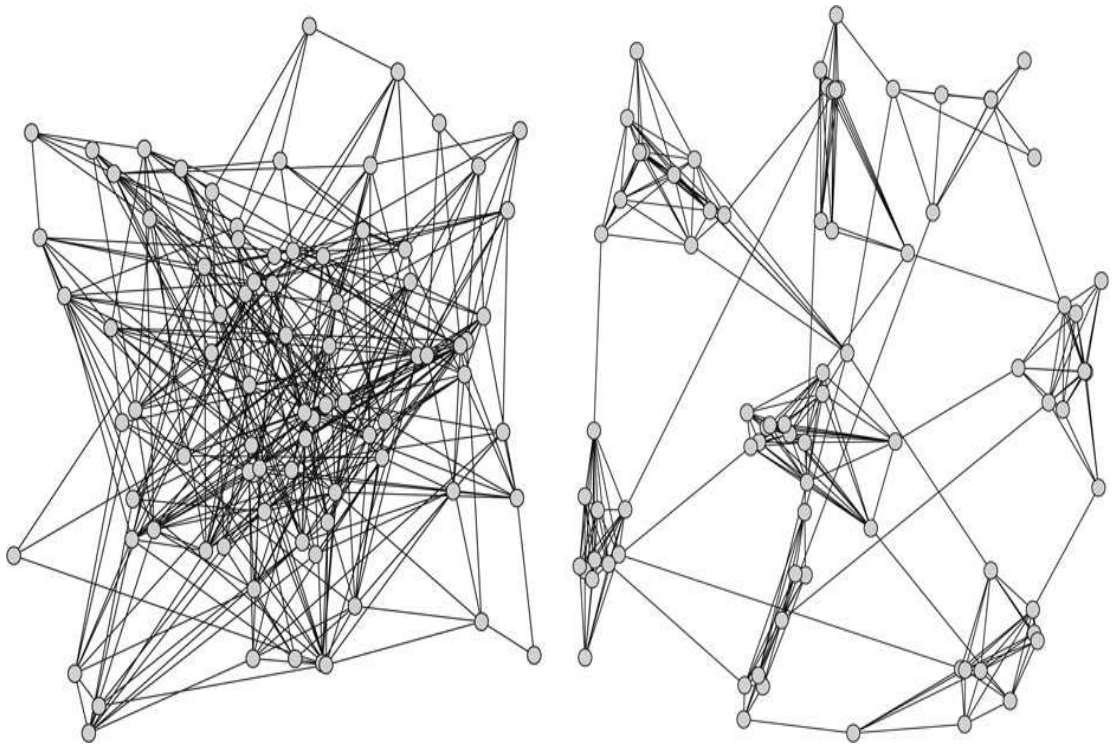


Figure: Two networks with 84 vertices and 358 edges. Network on the left is a uniform random network of Erdös-Rényi model and network on the right is generated by relaxed caveman model that captures the clustering properties of the networks

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Measures for Identifying Communities

- Vertex similarity
  - Feature-based similarity: based on vertex properties, e.g. a document is characterized by terms
  - Structural similarity: based on the connectivity of a vertex, e.g. neighbors
  - Can be used directly for clustering, e.g. hierarchical clustering, k-means
- Cluster fitness
  - Measures the quality of a clustering over a set of possible clusterings
  - Select the clustering that meets or optimizes a certain criteria
  - Modularity, density measure, cut-based measure, etc.

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Vertex Similarity
Feature-based Measures

- Vertices are represented as vectors of n features: $v = (f_{v,1}, f_{v,2}, ..., f_{v,n})$
- Euclidean distance

$$dist(u, v) = \sqrt{\sum_{k=1}^{n} (f_{u,k} - f_{v,k})^2}$$

- Cosine similarity

$$sim(u, v) = cos(u, v) = \frac{u \bullet v}{\sqrt{\sum_{k=1}^{n} f_{u,k}^2} \sqrt{\sum_{k=1}^{n} f_{v,k}^2}}$$

- Other feature-based measures: term-frequency inverse-document-frequency (tf-idf), edit distance, etc. for document similarity

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Vertex Similarity
Neighborhood-based Measures

- Given the network $\Gamma = (N, L)$ represented as adjacency matrix A
- Jaccard index

$$J(u,v) = \frac{\mathcal{N}(u) \bigcap \mathcal{N}(v)}{\mathcal{N}(u) \bigcup \mathcal{N}(v)}$$

- Pearson correlation of vertices u and v as column u and column v in the adjacency matrix A of the network

$$Pearson(u,v) = \frac{\frac{1}{n}(\sum_{k=1}^{n}(A_{u,k} - \mu_u)(A_{v,k} - \mu_v)}{\sigma_u \sigma_v}$$

where $A_{u,k}$ is an element in adjacency matrix A, n is the number of nodes, and

$$\mu_u = \frac{1}{n}\sum_{j}^{n} A_{u,j}, \sigma_u = \frac{1}{n}\sum_{j}^{n}(A_{u,j} - \mu_u)^2$$

- Euclidean distance:

$$dist(u,v) = \sqrt{\sum_{k=1}^{n}(A_{u,k} - A_{v,k})^2}$$

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Vertex Similarity
Path-based Measures

- Vertices are similar if they are connected by short paths
- Shortest path distance

$$dist(u, v) = d(u, v)$$

  This measure does not work well with small-world networks (every pair of nodes is connected with short path)
- Katz measure

$$Katz(u, v) = \sum_{l=1}^{\infty} \beta^l . |paths_{u,v}^{\langle l \rangle}|$$

  where $|paths_{u,v}^{\langle l \rangle}|$ is the set of all paths of length l from u to v
- Threshold path length based algorithm: finds induced subgraphs that are k-cliques

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Cluster Fitness
Cut-based Measures

- Given a network $\Gamma = (V, E)$ represented as an adjacent matrix A where $A_{i,j}$ represents the edge weight between nodes i and j
- Normalized cut of the group of vertices $S \subset V$

$$Ncut(S) = \frac{\sum_{i \in S, j \in \overline{S}} A_{i,j}}{\sum_{i \in S} deg(i)} + \frac{\sum_{i \in S, j \in \overline{S}} A_{i,j}}{\sum_{j \in \overline{S}} deg(j)}$$

  where $\overline{S} = V \setminus S$
- Conductance of the group of vertices $S \subset V$

$$Conductance(S) = \frac{\sum_{i \in S, j \in \overline{S}} A_{i,j}}{min(\sum_{i \in S} deg(i), \sum_{j \in \overline{S}} deg(j))}$$

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Cluster Fitness
Modularity Measure

- Proposed by Newman and Girvan (2004)
- Assumption: a partition is regarded good if
  - There are many edges within the communities
  - Only few among them
- The modularity Q for a partition into q communities of arbitrary size is defined as

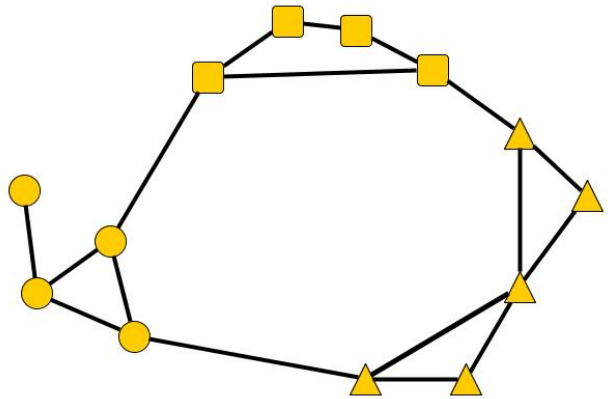
$$Q = \sum_{i=1}^{q} (e_{ii} - a_i^2) \text{ with } a_i = \sum_{j=1}^{q} e_{ji}$$

  where:

  - e_{ji} is the fraction of edges between nodes from community j and i
  - a_i is the overall fraction of internal links in i
  - a_i^2 corresponds to the expected fraction of internal edges
  - $Q \approx 0$ if no more internal edges are expected in a community
- $Q \geq 0.3$ implies significant community structure

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Modularity Example



We can construct a representation, in which rows and columns represent communities and each cell show fraction of the links among them. So we have $C = \begin{bmatrix} \frac{4}{17} & \frac{1}{17} & \frac{1}{17} \\ \frac{1}{17} & \frac{6}{17} & \frac{1}{17} \\ \frac{1}{17} & \frac{1}{17} & \frac{4}{17} \end{bmatrix}$.

$$\text{Modularity} = \frac{4}{17} - (\frac{6}{17})^2 + \frac{6}{17} - (\frac{8}{17})^2 + \frac{4}{17} - (\frac{6}{17})^2 =$$

$$\frac{4}{17} - \frac{36}{289} + \frac{6}{17} - \frac{64}{289} + \frac{4}{17} - \frac{36}{289} = \frac{68 - 36 + 102 - 64 + 68 - 36}{289} \approx 0.35.$$

Lehrstuhl Informatik 5 (Information Systems) Prof. Dr. M. Jarke

Community Detection Approaches

- Computer science approaches: graph partitioning
  - Applied in computer science and related fields, e.g. parallel computing and VLSI design
  - Number and size of the communities are fixed in advance
  - Minimize the number of edges lying between clusters
  - Well-known methods: spectral bisection, Kernighan-Lin algorithm
- Sociological approaches
  - Applied by sociologists, physicists and applied mathematicians, e.g. for social and biological networks
  - Based on vertex similarity (structural equivalence)
  - Methods: hierarchical clustering, k-means
- Recent methods
  - Un-supervised: number of clusters and their size are not known in advance
  - Popular methods: divisive methods based on edge betweenness, modularity optimization

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

A Brief Survey of Community Detection Algorithms

- Regarding classification, time complexity and basic algorithms

Type	Author	Year	Time Complexity	Short Description
Hierarchical Clustering	Fortunato [16]	2004	$O(N^3)$	Information centrality
	Zhou and Lipowsky [39]	2004	$O(N^4)$	Brownian particles
	Pons and Latapy [24]	2004	$O(MN^2)(O(N^2 \log N))$	Random walks
	Newman [29]	2004	$O((M+N))(O(N^2))$	Greedy optimization of modularity
	Newman and Girvan [31]	2004	$O(M^2N)(O(N^3))$	Greedy optimization of modularity
	Girvan and Newman [17]	2002	$O(M^2N)(O(N^3))$	Edge Betweenness
	Duch and Arenas [13]	2005	$O(N^2 \log N)$	Extremal optimization (of modularity)
	Radicchi et al. [33]	2004	$O(N^2)$	Edge-clustering coefficient
	Donetti and Mu noz [12]	2004	$O(N^3)$	Spectral analysis
	Clauset et al. [9]	2004	$O(Md \log N)(O(N \log^2 N))$	improved version of Newman
	Wakita and Tsurumi [36]	2007	$O(Md \log N)(O(N \log^2 N))$	improved version of Clauset et al.
Graph partitioning	Karypis and Kumar [22]	1998	polyn.	separator via minimization of # edges cut
	Brandes et al. [6]	2003	polyn.	separator via spectral clustering
Others	Flake et al. [15]	2000	polyn.	Max-flow min-cut
	Eckmann and Moses [14]	2002	$O(M < k^2 >)$	"curvature" and "reciprocity"
	Guimera et al. [18]	2005	parameter-dep.	Network as spin system
	Bagrow and Bollt [3]	2005	$O(N^3)$	local detection based on fitness-measure
	Wu and Huberman [37]	2004	$O(M+N)$	Kirchhoff's Law
	Palla et al. [32]	2005	$O(exp(N))$	Clique Percolation
	Reichardt and Bornholdt[34]	2004	parameter-dep.	q-state Potts-model
	Chakrabarti [8]	2004	$O(M)$	Reorganisation of the adjacency-matrix
	Boccaletti et al. [5]	2007	$O(MN)$	Synchronization process of phase oscillators
	Tasgin and Bingol [35]	2006	$O(M)$	Genetic Algorithm
	Newman [30]	2006	$O(N^2 \log N)$	Opt. of modularity by spectral analysis

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Sociological Approaches

- Hierarchical clustering
  - Measure similarity between vertices based on network structure
  - Agglomerative method: start with empty network of n vertices and no edges, join pair of vertices with highest similarity
  - Divisive method: start with the given network, remove edges that connect least similar pairs of vertices
  - Network structure is used only to compute vertex similarity
  - Result: a *dendrogram* or a tree
  - New methods: network structure is used to compute clustering quality, e.g. modularity
- k-means clustering
  - From data mining
  - Based on vertex distance measure

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Agglomerative Algorithm Based on Vertex Similarity

- Steps
  1. Start with an empty network of n vertices and no edges
  2. Add edges between pairs of vertices in order of decreasing similarity, starting with strongest similarity
  3. Result: a dendrogram where leaves are individual vertex, root is the network after all vertices are joined. A cut through the dendrogram represents the communities found if process stops at that level
- Extract communities: two methods
  - Single linkage: communities are components formed as edges are added
  - Complete linkage: communities are maximum cliques formed as edges are added
- Running time: $O((n^2 \log n)$ for sparse graphs
- Pros. and cons.: user does not need to provide the number of clusters in advance, but it does not tell where to cut the dendrogram to have the best clustering

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Agglomerative Algorithm Example

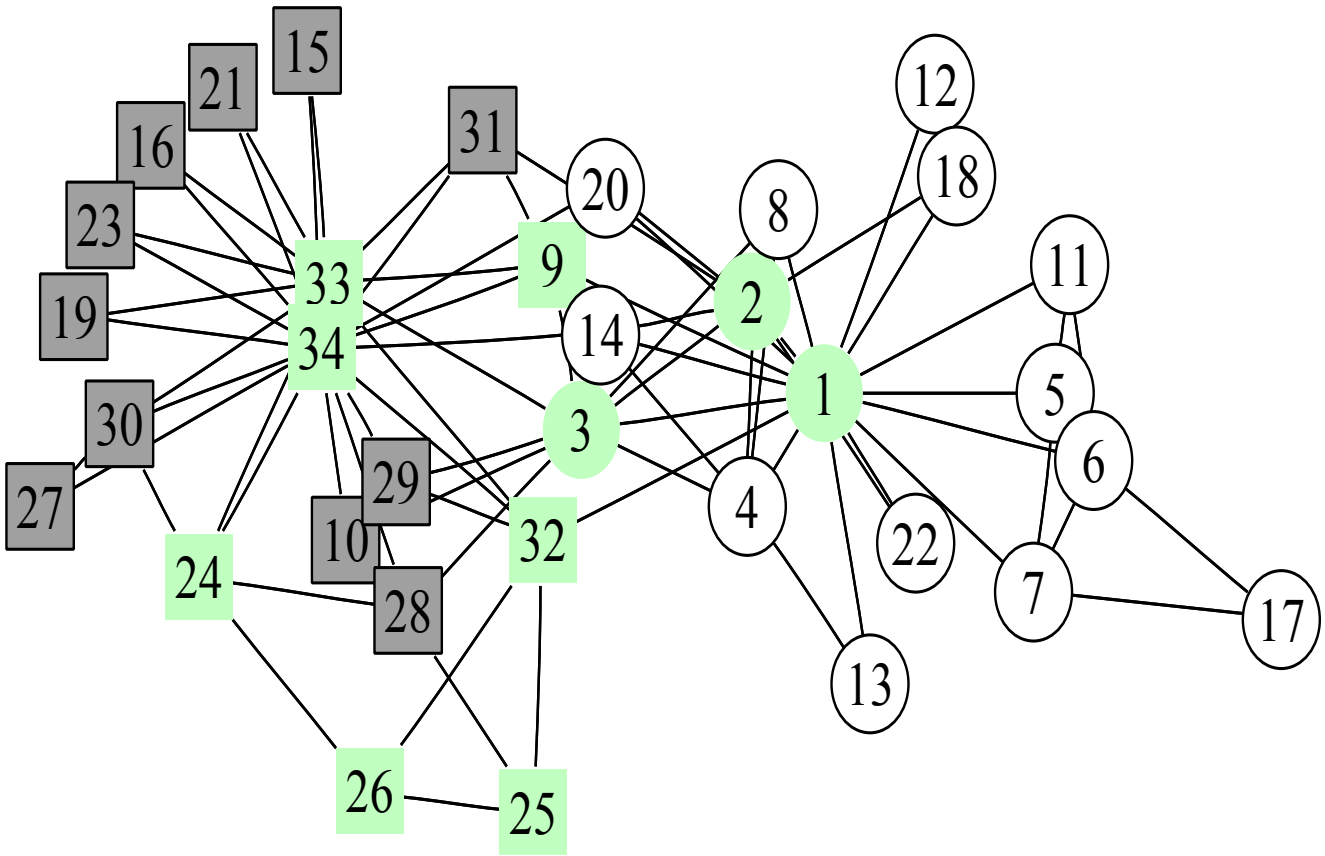


Figure: The resulting communities found by a single linkage hierarchical clustering of the karate club based on neighborhood Euclidean distance.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Agglomerative Algorithm Example
The Dendrogram

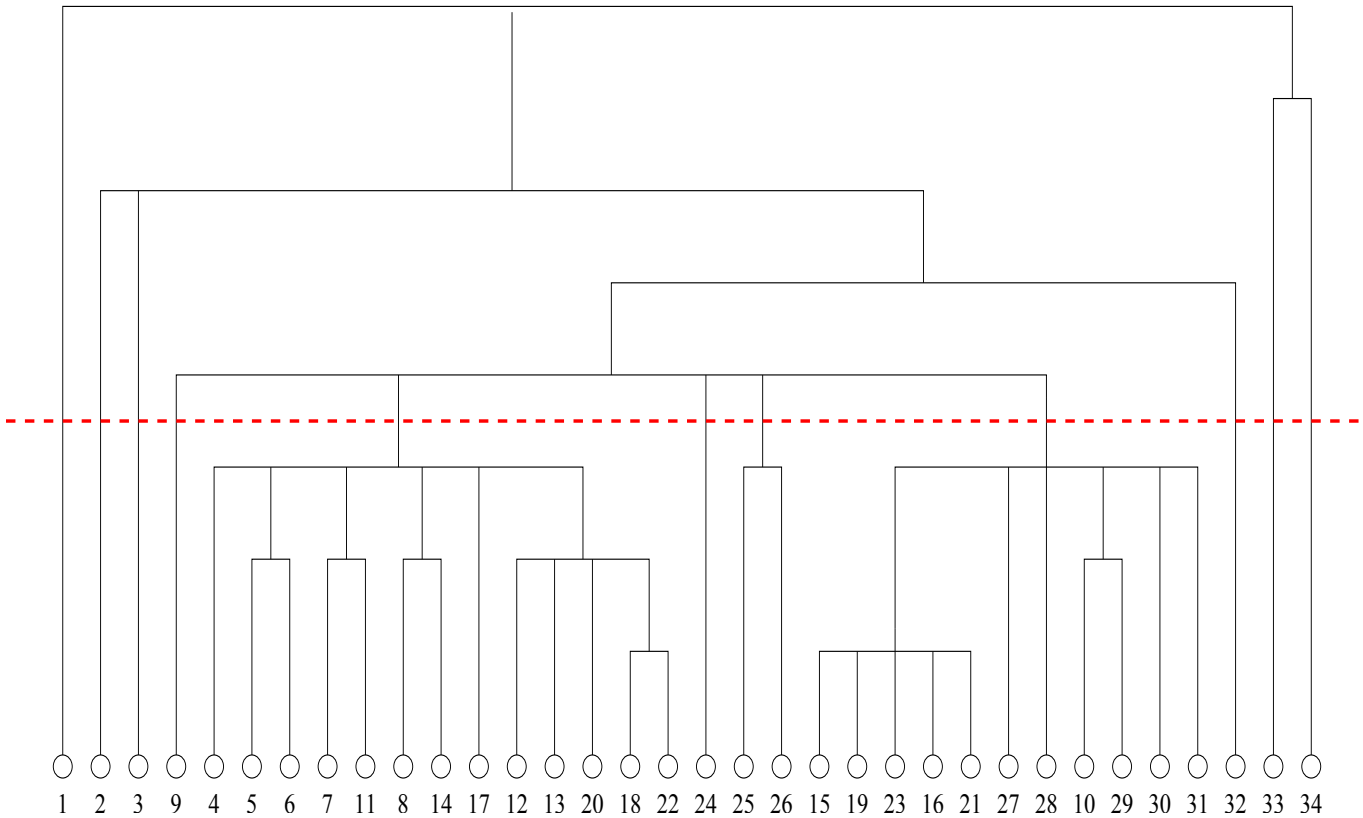


Figure: Dendrogram depicting the communities found by a single linkage hierarchical clustering in previous slide.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

K-means Clustering

- Each vertex is a point in metric space and represented as a vector with n dimensions
- Need to provide the number of clusters k
- K-means
  1. Choose k vertices (e.g. at random) as clusters' *centroid*
  2. Assign each vertex to nearest centroid
  3. Relocate the centroids of k clusters based on the current members (e.g. by the *mean* of coordinates of cluster members)
  4. Repeat from step 2, stop when centroids are stable and clusters do not change
- Clustering quality depends much on the initial choices of centroids

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Divisive Algorithm Based on Edge Betweenness

- Proposed by Newman and Girvan (2004)
- Idea: removal of edges that lie between communities
- Edge betweenness:
  - Shortest path: number of shortest paths that pass along the given edge
  - Random walk: the expected number of times a random walk between a pair of vertices will pass through the edge and sum over all vertex pairs
  - Current-flow betweenness: value of current along the edge summed over all vertex pairs
- Algorithm:
  1. Calculate the betweenness of all edges
  2. Find the edge with the highest betweenness and remove it (choose randomly if more than one edge have highest score)
  3. Recalculate the betweenness for the remaining edges
  4. Repeat from step 2
- Running time: $O(m^2 n)$, or $O(n^3)$ for sparse graph, where m is the number of edges, n is the number of nodes

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Divisive Algorithm Example
The Dendrogram

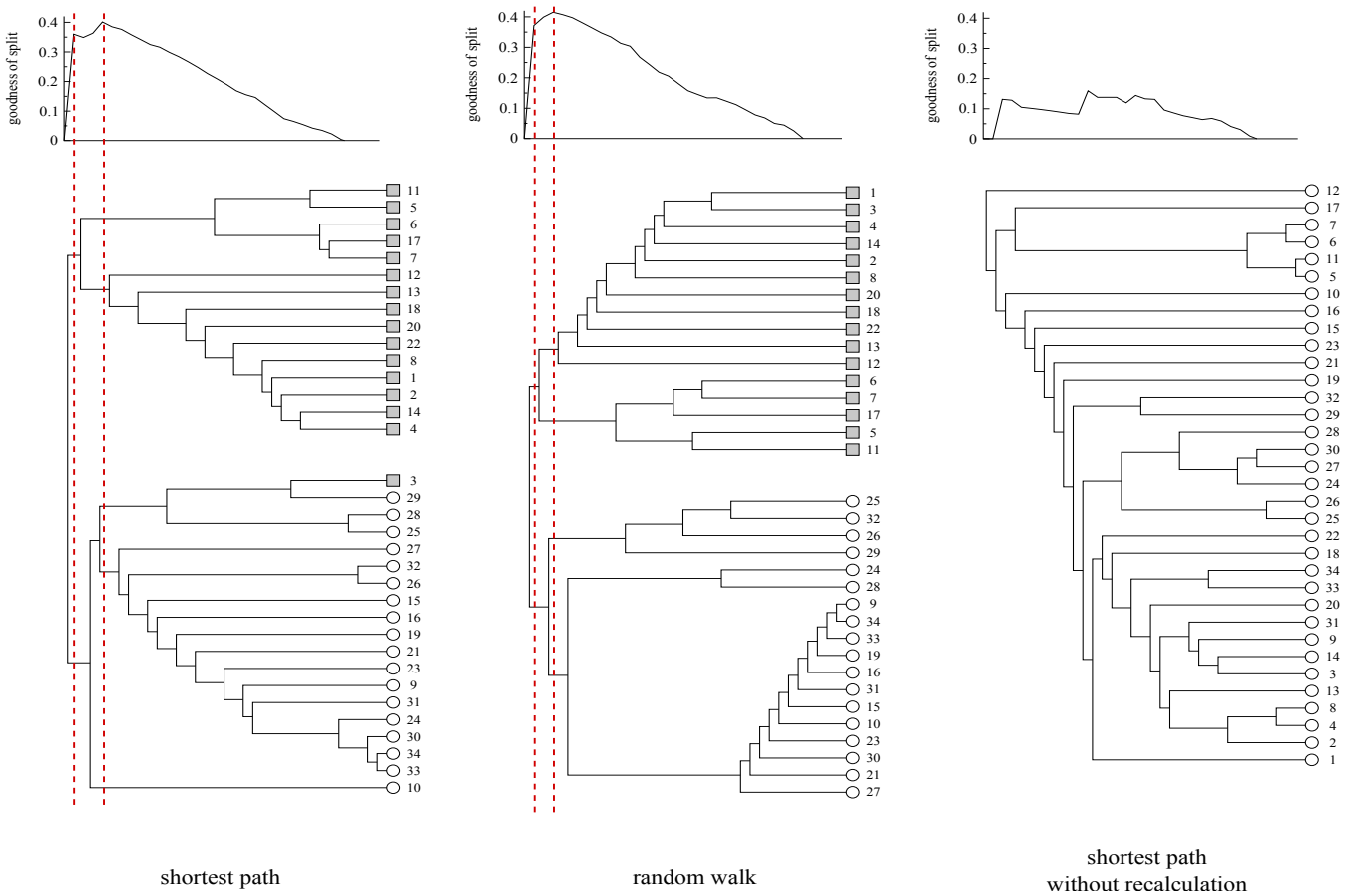


Figure: Dendrogram depicting the communities of karate club found by edge betweenness clustering using different betweenness measures. Shortest path and random walk found high value for modularity

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Divisive Algorithm Based on Edge Clustering Coefficient

- Proposed by Radicchi et al. (2004)
- Idea: (short) loops are easily found in the edges within communities
- Edge clustering coefficient:

$$C_{ij} = \frac{z_{ij} + 1}{\min(k_i - 1, k_j - 1)}$$

  where k_i is the degree of node i, z_{ij} is the number of triangles to which the edge (i, j) belongs to
- Algorithm:
  1. Calculate the clustering coefficient of all edges
  2. Find the edge with the lowest clustering coefficient and remove it (choose randomly if more than one edge have lowest score)
  3. Recalculate the clustering coefficient for the remaining edges
  4. Repeat from step 2, until all clusters are *strong* or *weak communities*
- Running time: $O(n^2)$ for sparse graph, where n is the number of nodes
- Gives poor result if the graph has few loops

Lehrstuhl Informatik 5 (Information Systems) Prof. Dr. M. Jarke

Modularity Optimization Algorithm

- Proposed by Newman (2004)
- Idea: Greedy optimization of modularity
- Steps:
  - Start with each vertex in one of n communities
  - Repeat joining communities in pairs, choosing at each step the join that results in the greatest increase (or smallest decrease) of modularity Q
  - Result: a dendrogram that shows the order of the joins
  - Cuts through this dendrogram at different levels give divisions of network into larger or smaller number of communities
  - Select the best cut by looking for the maximum value of Q
- Running time: $O((m+n)n)$, or $O(n^2)$ for sparse graph, where m is the number of edges, n is the number of nodes

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Modularity Optimization Example

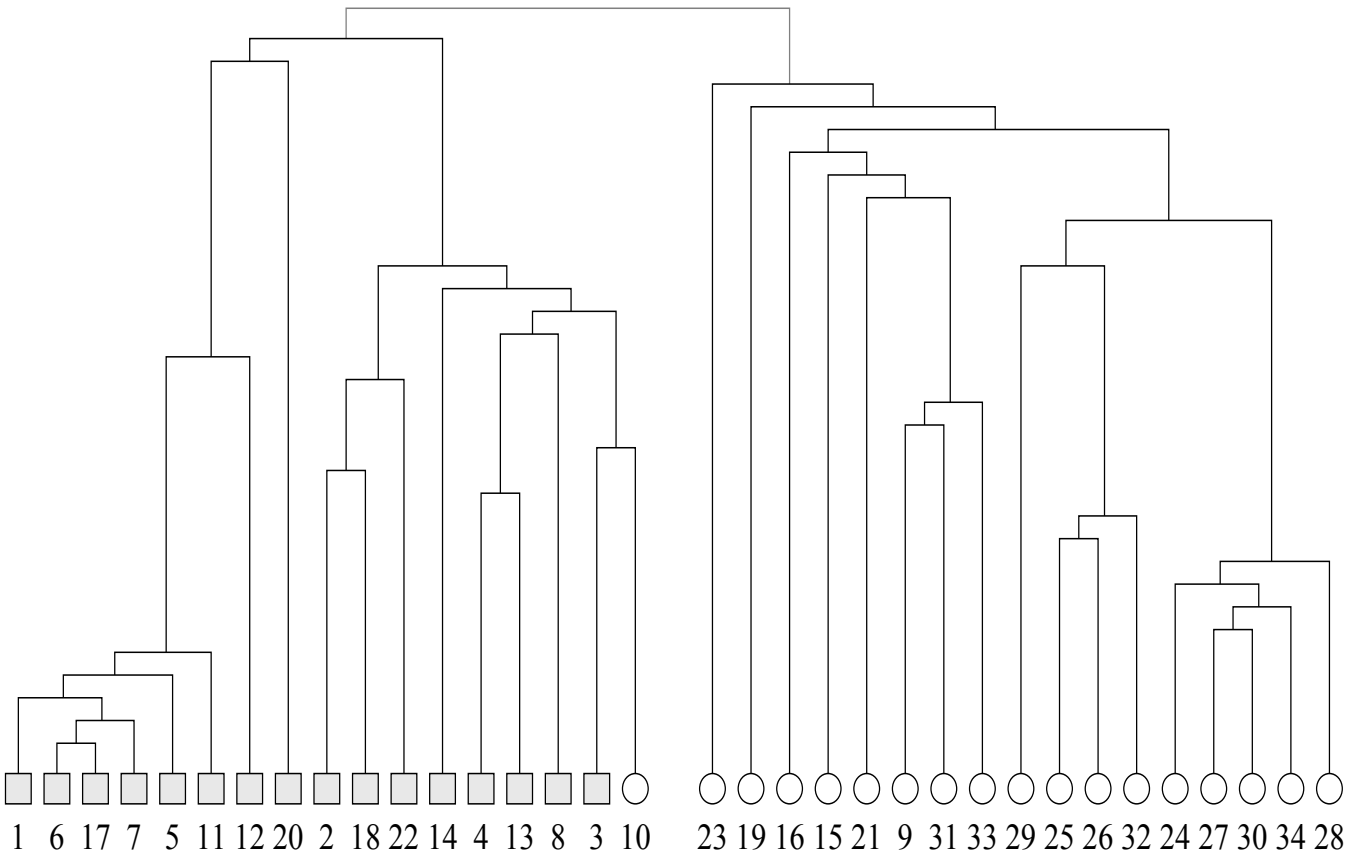


Figure: Dendrogram of the communities found by modularity optimization algorithm in the karate club of Zachary. The shapes of the vertices represent the two groups into which the club split as the result of an internal dispute

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Graph Partitioning

- Objectives:
  - Dividing the vertices into k clusters with roughly equal size
  - Minimizing the number of edges running between vertices in different clusters
- Based on iterative bisection:
  - Find the best division of the graph into two clusters
  - Further subdivide those two clusters until having the required number k of clusters
- Two well-known algorithms: spectral bisection and Kernighan-Lin algorithms

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Kernighan-Lin Algorithm

- Idea: divide the network into two clusters based on greedy optimization
  - Assign a benefit function Q to divisions of network
  - Optimize function Q over possible divisions
- Benefit function

$$Q = E_{in} - E_{out}$$

  where E_{in} and E_{out} are the number of edges within two clusters and number of edges between them
- Steps:
  - Choose the size of two clusters and a starting configuration for the clusters
  - For all pairs of vertices in which vertex is chosen from each of clusters
    - Compute the change ΔQ that would result from swapping the two vertices
    - Select the pair that maximizes ΔQ and perform the swap
    - Repeat until all vertices in one of the clusters are swapped
  - Go back to the sequence of swaps and find the point where Q is highest

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Kernighan-Lin Algorithm
Advantages and Disadvantages

- Advantages
  - Gives good results
  - Quite fast: running time is $O(n^2)$ in worst case, where n is the number of vertices
- Disadvantages
  - Need to choose the size of two clusters and the starting configuration for the clusters
  - Result depends much on the specified cluster size: can run the algorithm on different choices of cluster sizes, but running time is then $O(n^3)$
  - Only divides network into two clusters: repeat bisection to divide into more clusters, but no guarantee about clustering quality
- Not suitable for social network clustering where there is no knowledge about the number of clusters and cluster sizes

CC BY SA

Spectral Algorithms
Eigenvector-based Clustering

- Proposed by Donetti and Munoz (2004)
- Idea: values of the eigenvector components of the Laplacian matrix are close for the vertices in the same community and can be used to represent vertices in metric space
- Algorithm:
  - Compute M eigenvectors of Laplacian matrix of the input network
  - Vertices are represented as points in *M-dimensional* space
  - Vertices are clustered using hierarchical clustering methods
- Running time is $O(n^3)$
- Number of eigenvectors M need to be provided
- Similar idea by Capocci et al. (2004)

cc BY SA

Dynamic Algorithms
Random-walk Clustering (1)

- Given a network $\Gamma = (V, E)$ represented as an adjacent matrix A where $A_{i,j}$ represents the edge weight between nodes i and j
- Transition probability matrix P

$$P = \{P_{i,j} = \frac{A_{i,j}}{deg(i)}\}$$

  if $(i, j) \in E$ and $P_{i,j} = 0$ otherwise
- A *random walk* is a Markov chain with transition probabilities specified by transition matrix P
- *Hitting time* h_{ij} is defined as the number of steps a random walk starting from node i before node j is visited for the first time
- *Commute time* $c_{ij} = h_{ij} + h_{ji}$

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Dynamic Algorithms
Random-walk Clustering (2)

- Proposed by Zhou (2003)
- Idea: a random walker spends a long time inside a community because of the high density
- The distance between nodes i and j: the average steps needed for a random walk going from node i to node j
- *Local attractor* of node i: its closest neighbor
- *Global attractor* of node i: the closest vertex to node i
- Local-attractor based (L-community) and global attractor based (G-community) communities: each node has to be put into the same cluster of its attractors and of all other vertices for which it is an attractor
- For small networks, L-community and G-community are identical
- For large networks: each G-community contains several L-communities

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Community Detection Summary

- Different approaches from different areas: computer science, social science, physics, etc.
- When comparing algorithms, we need to consider:
  - The input parameters: do users need to provide and control input parameters (e.g. number of clusters, clusters size, etc.)?
  - Complexity
  - Clustering quality: depends on many factors (e.g. input parameters, the characteristics of the graph, etc.)
  - Is the algorithm deterministic? (e.g. it gives hint to have good clustering)
- Most of algorithms has running time $O(n^2)$: not applicable for very large networks
- Current trends: parallelized algorithms for grid and cloud computing

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Reading List to Community Detection

- *Graph Clustering* by Satu Elisa Schaeffer, 2007
- *Detecting Community Structure in Networks* by M. E. J Newman, 2004
- *Community Structure in Graphs* by Santo Fortunato and Claudio Castellano, 2007
- *Finding and Evaluating Community Structure in Networks* by M. E. J. Newman and M. Girvan, 2004
- *Fast Algorithm for Detecting Community Structure in Networks* by M. E. J. Newman, 2004
- *Community Structure Identification, A Modern Review* by Leon Danon and Albert Diaz-Guilera, 2007
- *Community Detection Algorithms: A Comparative Analysis* by Andrea Lancichinetti and Santo Fortunato, 2009

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Overlapping Communities

- Definitions
- Applications
- Overlapping Community Detection Algorithms
  - Clique percolation
  - Link Partitioning
  - Local Expansion
  - Fuzzy Detection
  - Agent-Based/Dynamical Algorithms

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Definitions

- Overlap is a significant characteristic in many real-world social networks
- Notions
  - a *cover* $C = c_1, c_2, ..., c_k$ is the set of clusters found, where a node may belong to more than one cluster
  - *belonging factor* $[a_{i1}, a_{i2}, ..., a_{ik}]$ associates a node to a community, where a_{ic} is the strength of association between node i and community c
  - *crisp* assignment represents the community detection where the overlapping is binary
  - *fuzzy* assignment represents the community detection where a node is associated to communities according to a belonging factor
- Overlapping communities can output more extra-group similarity than intra-group similarity
- Detection of truly overlapping communities is a challenging problem, especially in large-scale networks

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Overlapping Communities: Ego and Global Levels

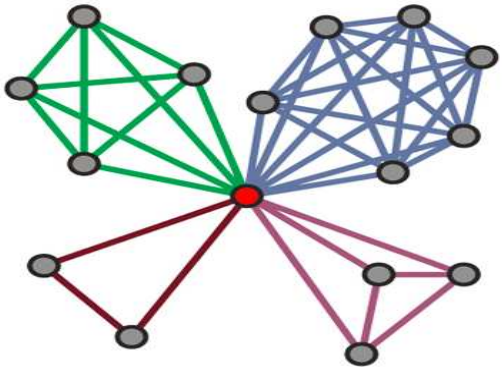


Figure: An individual with his belonging communities (Ahn et al. 2010) (e.g. family, friends, colleagues, acquaintances, etc.)

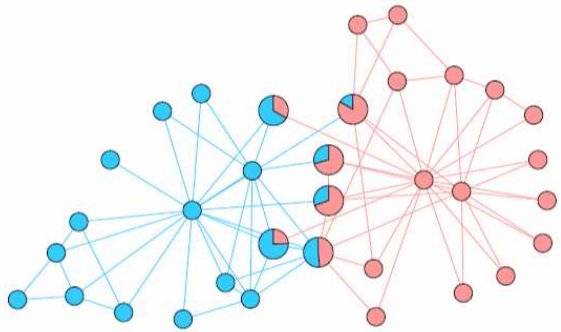


Figure: Results of an overlapping community algorithm (Ball et al. 2009) performed on the network of the karate club studied by Wayne Zachary - the colors show the division of both vertices and edges

Note: Different resolutions may apply for overlapping communities

Lehrstuhl Informatik 5 (Information Systems) Prof. Dr. M. Jarke

Overlapping Community

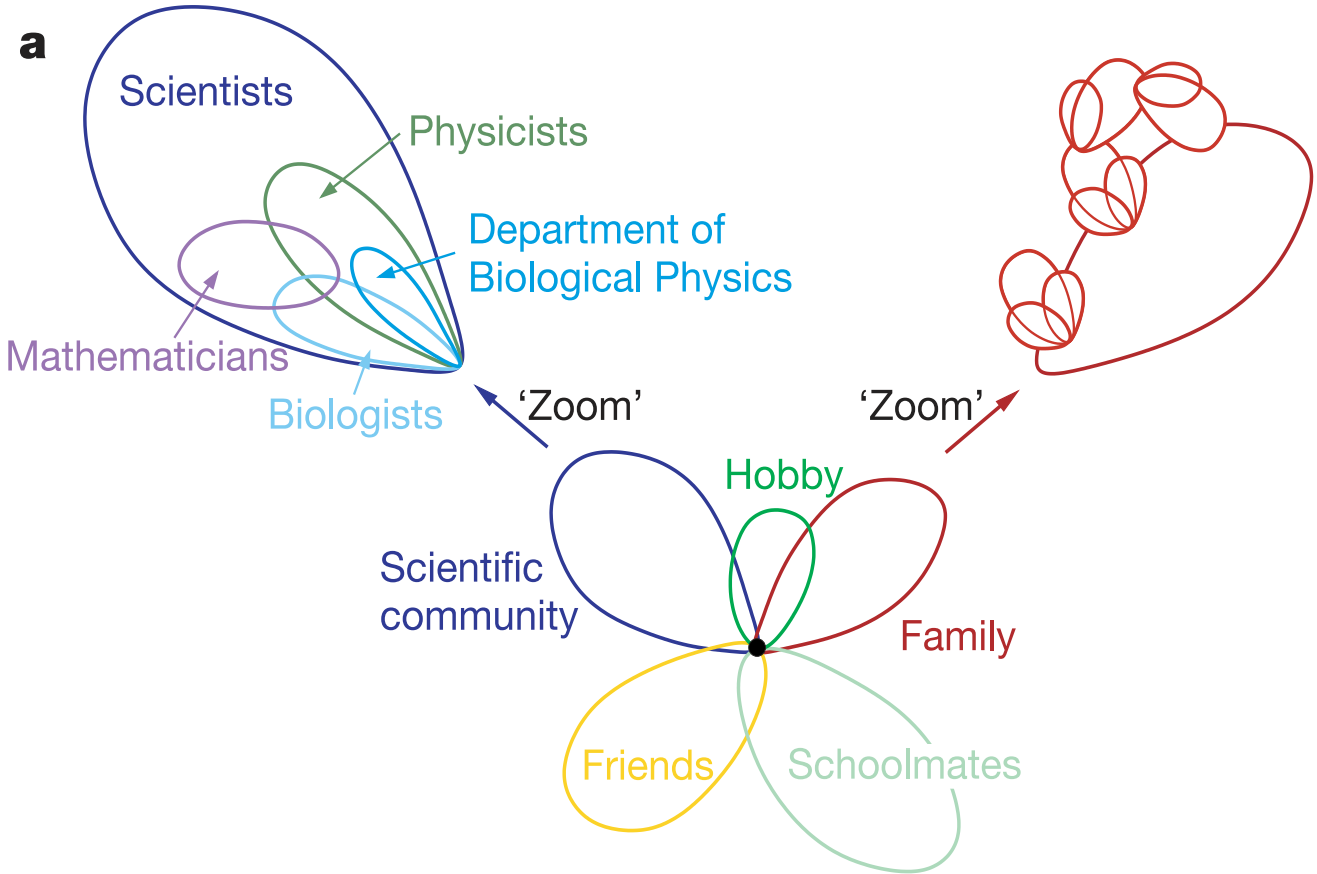


Figure: The black dot in the middle represents a person with several of his communities around. Zooming in on the scientific community demonstrates the nested and overlapping structure of the communities.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Overlapping Community Detection Algorithms

Type	Algorithm	Year	Complexity	Description
Clique Percolation	CPM, Palla et al.	2005	polynomial	good results in networks with dense connected subgraphs
	CPMw, Farkas et al.	2007	–	CPM for weighted networks; introduces a link weight threshold
Link Partitioning	Evans et al.	2009	–	partition a line graph constructed from the original network
	Evans et al.	2009	$O(< k^2 > / < k >)$	generalization of previous algorithm with weighted/directed graphs
	Ahn et al.	2010	$O(nk^2_{max})$	hierarchical clustering edge partitioning using edge similarity
	Ball et al.	2011	$O(mK)$	based on the link communities
	Kim et al.	2011	–	modified version of map equation method
Local Expansion and Optimization	Baumes et al.	2005		iterative scanning to locally maximize a density function
	COCD, Du et al.	2008	$O(C_M \times Tri^2)$	maximal clique based detection
	EAGLE, Du et al.	2008	$O(n^2 + (h+n)s)$; $O(n^2 s)$	maximal cliques, agglomerative framework for dendrogram
	GCE	2009	$O(n^2 \log n)$	identify cliques as seeds and greedily optimize local metric
	LFM, Lancichinetti et al.	2009	$O(n^2 \log n)$	based on local optimization of a fitness function
	OCA, Sureda et al.	2010	$O(n^2 + (h+n)s)$; $O(n^2 s)$	each node is mapped to a d-dimensional vector; fitness function
	MONC, Havemann et al.	2011	$O(n^2)$	optimally grows natural or partial communities
Fuzzy Detection	Zhang et al.	2006	$O(mKh + nK^2h + K^3h)$; $O(nk^2)$	based on modularity function and fuzzy c-means clustering
	Wang et al.	2009	$O(m) + X$; $O(n) + X$	based on local community strength and the overlapping extent
	NMF, Psorakis et al.	2011	$O(Kn^2)$	based on Bayesian non-negative matrix factorization
	MOSES, McDaid et al.	2010	–	uses the greedy maximization of a global objective function
Agent-based/Dynamical	COPRA	2010	$O(v^3 n)$; $O(vn \log v)$	based on label propagation
	Chen et al.	2010	$O(\lvert L_i \rvert \cdot \lvert L(N_i) \rvert \cdot \Delta_i)$	uses game theory notions
	SLPA, Xie et al.	2011	$O(Tn)$	based on label propagation, allows nodes with multiple labels
Other	Li et al.	2005	–	overlapping community of named entities based on triangle expansion and content similarity clustering
	CONGA	2007	$O(m^3)$	extension of Girvain and Newman's divisive clustering, based on edge and split betweenness
	Nepusz et al.	2008	$O(N^2 ch)$	based on similarity function and optimization of a fitness function
	BiTector, Du et al	2008	$O(M^2)$	bi-clique community identification for large-scale sparse bipartite networks
	CONGO, Gregory	2008	$O(m \log m + m^{2h+2}/n^{2h+1})$; $O(n \log n)$	improvement of CONGA algorithm
	GA-NET+, Pizzuti C.	2009	–	uses a genetic algorithm on the line graph to extract dense communities
	DOCS, Wei et al.	2009	–	based on spectral partition and random walk expansion
	Peacock, Gregory	2009	$O(n^2 \log n) + X$	similar to CONGA; based on edge splitting and disjoint clustering
	Shen et al.	2009	–	based on a maximal clique network partitioning using modularity optimization
	CIS, Goldberg et al.	2010	–	uses iterative scan for community detection

Table: Overlapping Community Detection Algorithms

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Applications

- Sociology and social networks
  - People belong to more than one circle
- Biology
  - Protein complexes can belong to multiple classes (protein-protein interaction)
  - Metabolic networks
- Bibliometric research
- Recommender systems
- Mobile networks
  - Worm containment in cellular networks
  - Sensor reprogramming

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Clique Percolation Method (CPM)
Clique Percolation Algorithm

CPM

assumes that a community consists of overlapping sets of fully connected subgraphs and detects communities by searching for adjacent cliques

- *CFinder* (implementation) - Proposed by Palla et al. (2005)
- Idea: internal edges of community are likely to form cliques, while intercommunity edges are not
- Notions:
  - *k-clique*: complete graph with k vertices
  - *Adjacent cliques*: two k-cliques are adjacent if they share $k-1$ vertices
  - *k-clique chain*: the union of adjacent k-cliques
  - Two cliques are connected if they are part of a k-clique chain
  - A *k-clique community* is the largest connected subgraph obtained by the union of a k-clique and of all k-cliques which are connected to it

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

K-clique Community Example

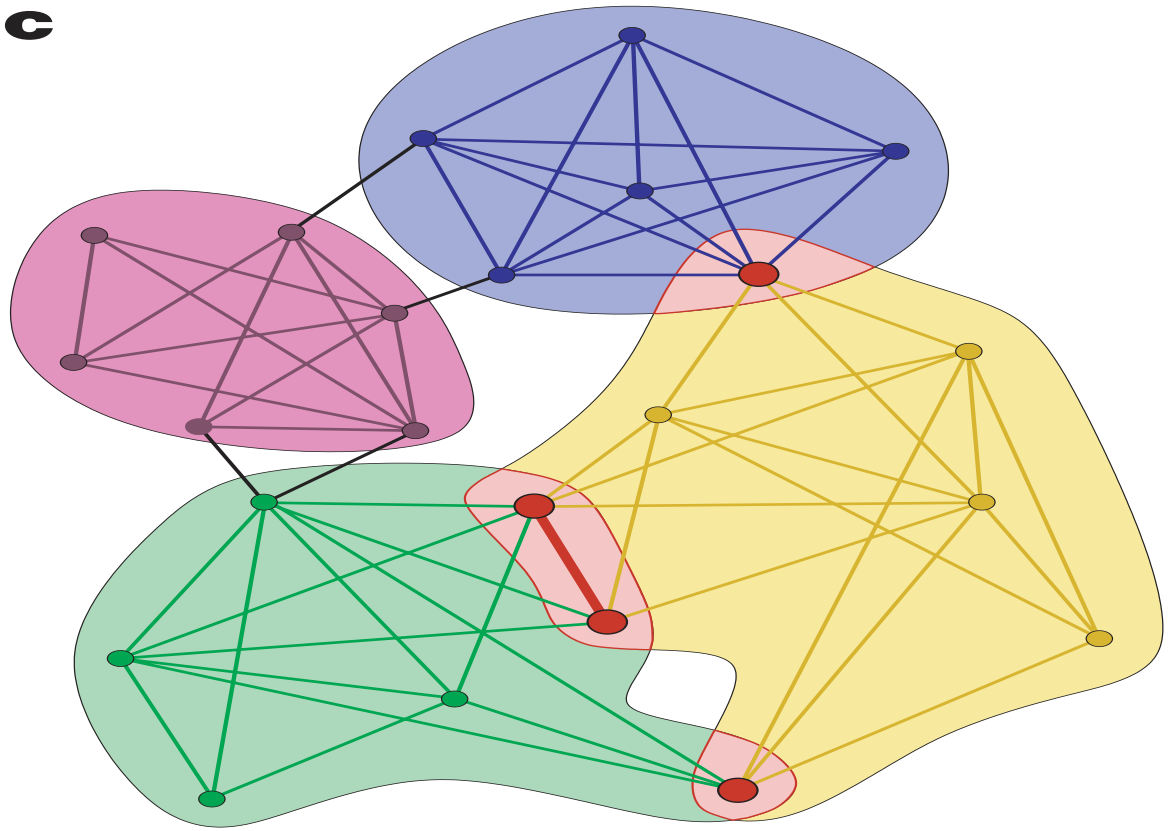


Figure: An example of overlapping k-clique communities at k = 4. These overlapping regions are emphasized in red. Notice that any k-clique (complete subgraph of size k) can be reached only from the k-cliques of the same community through a series of adjacent k-cliques

Lehrstuhl Informatik 5 (Information Systems) Prof. Dr. M. Jarke

Clique Percolation Algorithm II

- Process:
  - Find k-cliques ($k = 3, 4, ..., n$)
  - A k-clique community is identified by rotating a k-clique about the $k - 1$ vertices it shares with any adjacent k-clique
  - Vertices which belong to non-adjacent k-cliques are in different clusters
- Running time grows exponentially with graph size, but it is quite fast for real world networks
- Limitations: it assumes that the graph has a large number of cliques

cc BY SA

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Link Partitioning Algorithms

Link Partitioning Algorithms

consider a community to be a set of closely interrelated links

- *Link communities reveal multiscale complexity in networks*, Ahn et al. (2010)
- Idea: uses hierarchical clustering for dendrogram creation, where the branches are link communities
- Has to determine the best level cut for relevant community detection, using the partition density D (objective function)
- For m_c the nr. of edges in a community c and n_c the corresponding nodes $D = \frac{2}{M} \sum m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}$
- D measures the quality of a link partition (1 when every community is a fully connected clique, 0 when every community is a tree)

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Fuzzy Detection Algorithms
MOSES

Fuzzy Detection Algorithms - a belonging factor is calculated for each node in the network; need to determine the number of communities

> *quantify the strength of association between all pairs of nodes and communities*

- A Model-based Overlapping Seed Expansion algorithm(MOSES) - Proposed by McDaid and Hurley (2010)
- Idea: greedily maximize a global objective function by creating/deleting communities and adding/removing nodes to/from those communities
- Suitable for unweighted, undirected networks with no self loops and highly overlapping community structure
- Aim: Maximize the joint distribution over (z, p_{in}, p_{out}) where z is the proposed set of communities, p_{in}, p_{out} are the probabilities that two nodes share or not a community

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

MOSES II

- Steps:
  - randomly select edges and expand a community for each of the edges (by adding directly connected nodes)
  - greedily add nodes that maximize the global function
  - continue until a local maximum is found (with a small look ahead, e.g. 2 expansions)
  - periodically scan the communities to determine if deleting one community can output a positive change in the objective
  - after the edge expansion, perform a fine-tuning by reconsidering each node and its immediately surrounding communities for maximizing the function

cc BY SA

Dynamical Algorithms
SLPA

Dynamical algorithms - e.g., label propagation, allow a node to have multiple labels and use them to detect the communities

- Speaker-listener Label Propagation Algorithm(SLPA) - Proposed by Xie et al. (2011, 2012)
- Extension of label propagation algorithm
- Idea: speaker-listener information propagation process (mimics human communication)
- Nodes can store updated labels
- Pros. and cons.:
  - performs well and it is stable
  - can implement different speaker-listener rules
  - detects both crisp and fuzzy overlapping communities
  - requires a threshold variable (varying from 0.05 to 0.5), for specifying the amount of overlapping nodes; 0.05 converts it to crisp clustering
  - not deterministic (can output different results on same data)

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

SLPA II

- Steps:
  1. node's memory is initialized with unique label
  2. do until a user defined iteration number T is reached:
     - select one node as listener
     - each neighbor randomly selects a label with a probability proportional to the label's occurrence frequency and sends it to the listener
     - listener accepts one of the propagated labels according to a rule (e.g., most popular label)
  3. post-processing phase for identifying the communities (uses a threshold r)
- Running time is $O(Tm)$, where m is the total number of edges

cc BY SA

AC 5 RWTH
Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Local Expansion Algorithms
MONC

Local expansion algorithms - optimally grow partial or natural communities, starting from single nodes or core (disjoint) communities

- *Merging of Overlapping Natural Communities (MONC)* - Proposed by Havemann et al. (2010)
- expands a community from random seed node to locally maximize a fitness function
  $$f(G, \alpha) = \frac{k_{in}^G + 1}{(k_{in}^G + k_{out}^G)^\alpha}$$
  where k_{in}^G and k_{out}^G are total internal and external degree of the natural community G, α is a resolution parameter
- Running time is $O(n^2)$
- Detects the hierarchical structure of the network
- Builds entire networks starting from single nodes (very useful in settings such as publication networks)

Lehrstuhl Informatik 5 (Information Systems) Prof. Dr. M. Jarke

MONC Expansion Example

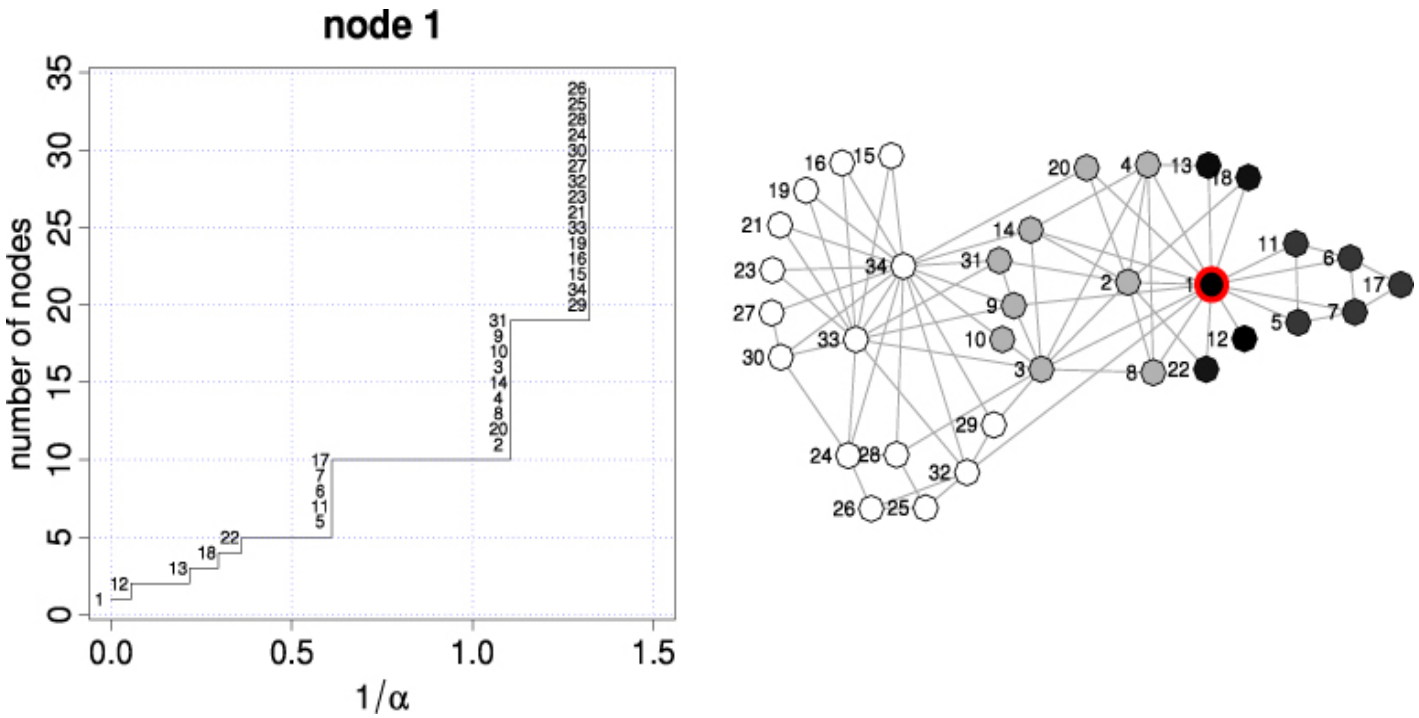


Figure: Growing a natural community for node 1 using the MONC algorithm in Zachary's karate club network (Havemann et al., 2011)

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

DMID: A Two Phase Approach

- Disassortative Degree Mixing and Information Diffusion (DMID) - Proposed by Shahriari, Krott and Klamma (2015)
- Identifying most influential nodes
- Working based on two well-known social dynamics
  - Identifying leaders using disassortative degree mixing and degree
  - Using of cascading behavior named network coordination game for computing degree membership of nodes to communities

Properties of DMID

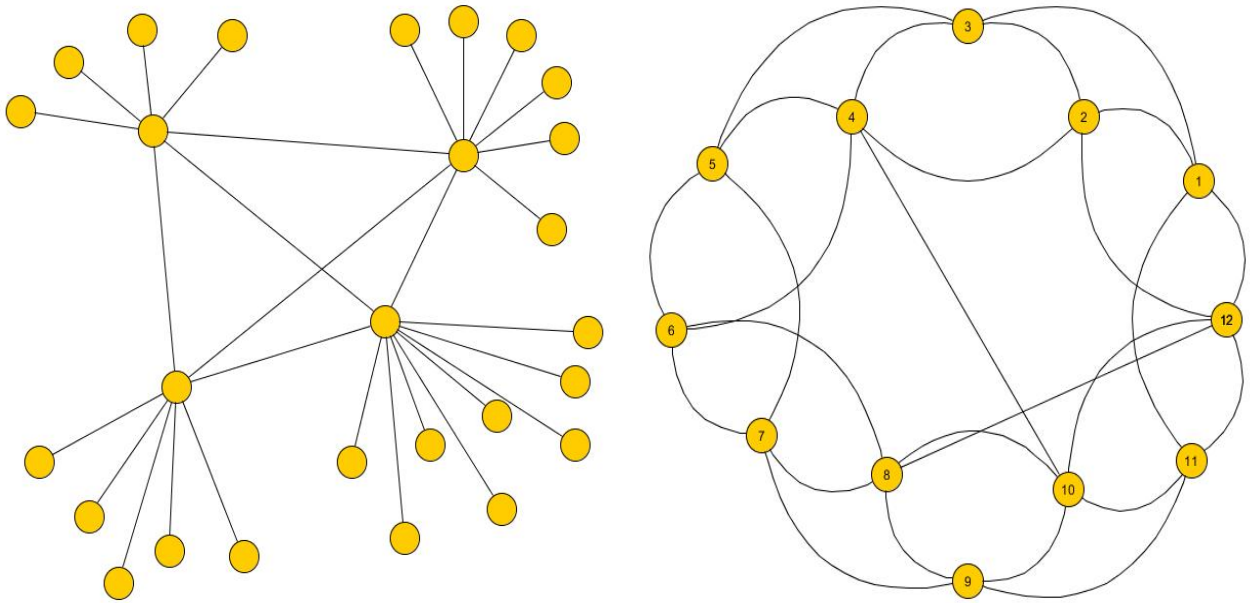
- performs well and it is competitive
- suitable for disassortative degree mixing networks
- detects both crisp and fuzzy overlapping communities
- requires a threshold variable for changing the behavior
- identifies local leaders and hierarchy of the network

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

(Dis)assortative Degree Mixing

Disassortative degree mixing

- Sign of dissimilarity
- Large degree nodes tend to connect to low degree nodes and low degree nodes tend to connect with large degree nodes



(a) (b)

Figure: (a) a is toy example of disassortative networks and (b) is a toy example of assortative networks.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

DMID: Leader Identification

Discovering influential nodes

- Using of disassortative degree mixing property
  - $AS_{ij} = |deg(i) - deg(j)|$
- Row normalize disassortative matrix
  - $T_{ij} = \frac{AS_{ij}}{\sum_{k=1}^{|N|} AS_{ik}}$
- Performing a random walk
  - $DA^{t+1} = DA^t \times T$
- Computing local leadership value
  - Combining degree (DRN) and disassortative value (DA) of node i
  - $LS_i = DA_i \times DRN_i$
- Finding local leaders
  - For each node i in the network, local leadership value should be compared with all of its neighbours $N(i)$
  - $LS_i > LS_j \quad \forall j \in N(i)$
- Finding leaders using average follower degree (AFD)
  - Local leaders which have enough followers (bigger than AFD value) can be considered global leaders

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Cascading Behaviour: Network Coordination Game

- How new opinions, interests, ideas, innovations, practices, conventions and etc are diffused over the networks
- Friends influence you in social networks
- In network study, it is sometimes known as **diffusion of innovations**
- One of the approaches for modelling diffusions is network coordination games
  - Consider a social network in which each node has possibility of adopting only two behaviours A and B
  - When two nodes are directly connected to each other then the tendency of matching their behaviour would be higher.
  - One can model network coordination game using u and v as the players of this game, A and B as possible strategies
  - Each node adjust its behaviour based on the pay-off received from its neighbors

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Cascading Behaviour: Network Coordination Game

The pay-off can be defined as follows

- If v adopt behavior of u then it is getting payoff a
  - If v does not adopt behaviour of u then it is getting pay-off b
- v has k neighbors
  - p fraction of its neighbors adopt A, then Payoff of $v = a.p.k$
  - $1 - p$ fraction of its neighbors adopt B, then payoff of $v = b.(1 - p).k$
- Behaviour B is preferred if $b.(1 - p).k > a.p.k$

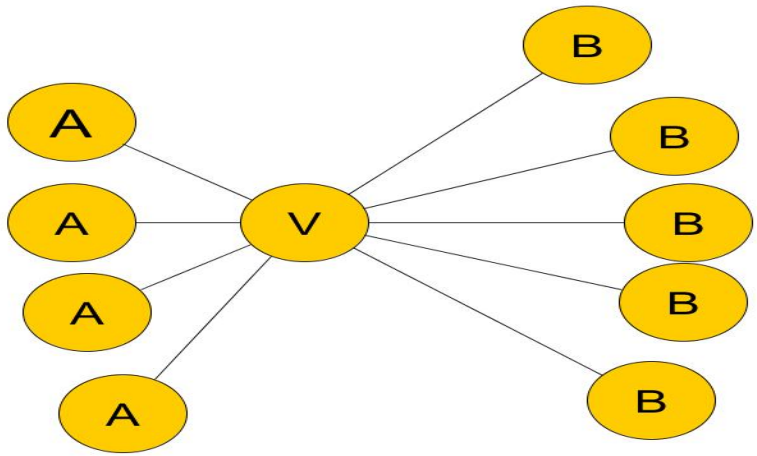


Figure: In this example, if $a = 1$ and $b = 1$ then the payoff in which node v receives for behavior A is 1*0.4*10=4 and the payoff in which it receives for behavoir B would be 1*0.6*10=6.

Lehrstuhl Informatik 5 (Information Systems) Prof. Dr. M. Jarke

DMID II: Cascading behavior

Network coordination game

- Cascades initiates by the identified leaders
- Different cascades can overlap
- Threshold for nodes as the input parameter for this section
  - e.g, local leadership value can be as the threshold
- Causing different cascade sizes
- Cascades are independent, therefore, they can be executed in parallel
- $p_A(i) = \frac{|\{j \in N(i) : j \text{ has behavior } A\}|}{|N(i)|}$
  - where $N(i)$ is the set of neighbours of node i
  - $p_A(i)$ is the received pay off of node i from its neighbours

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

DMID II: Cascading behavior

Network coordination game

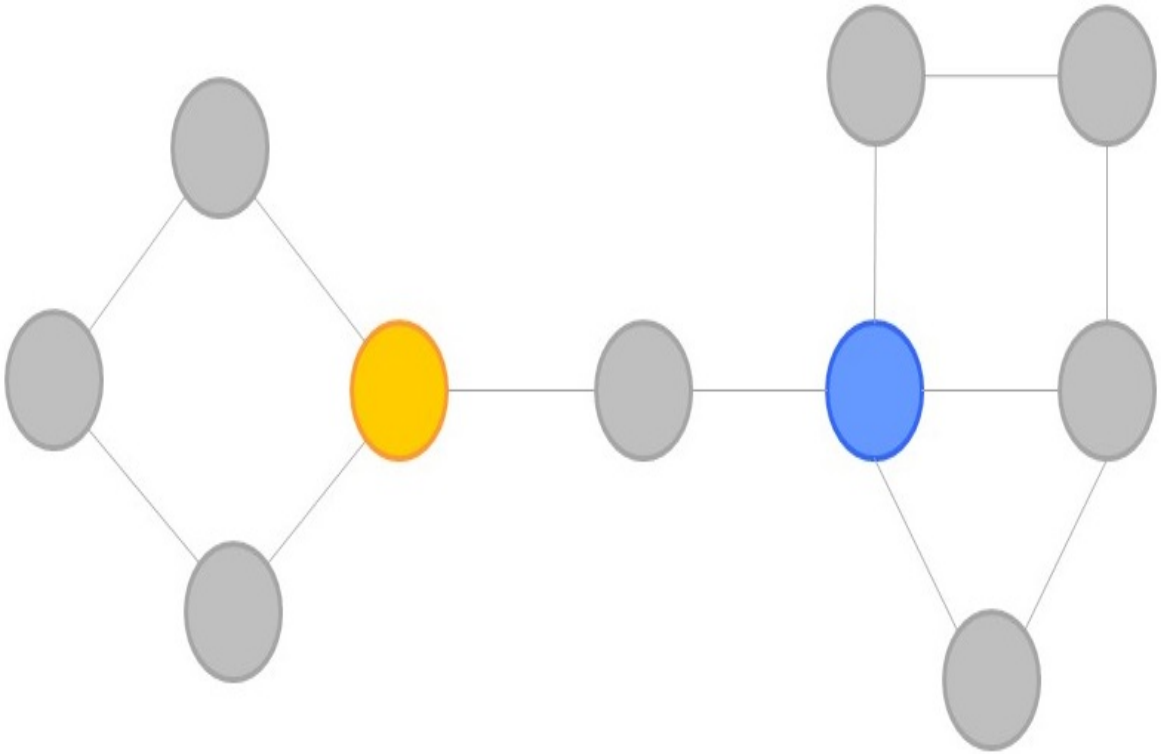


Figure: Two leaders identified with blue and yellow color. Each node as threshold of 0.5.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

DMID II: Cascading behavior

Network coordination game

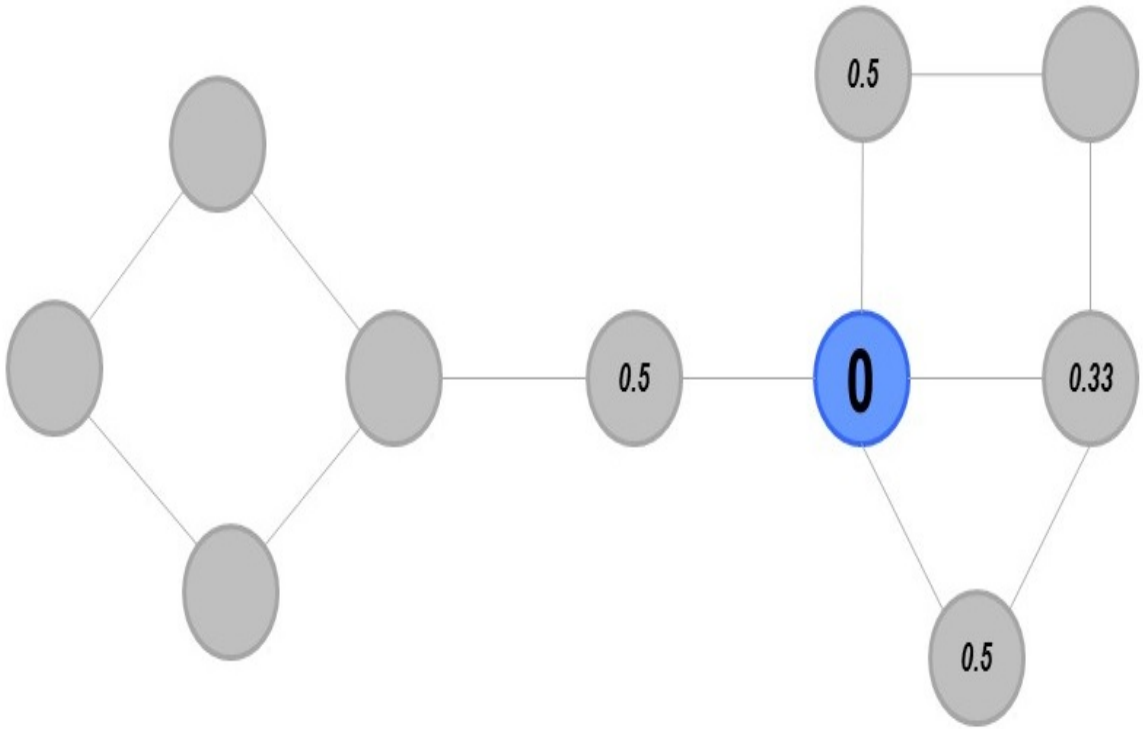


Figure: Leader blue has blue behaviour and other nodes have behaviour gray. Numbers in the circles are the received pay-off value.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

DMID II: Cascading behavior

Network coordination game

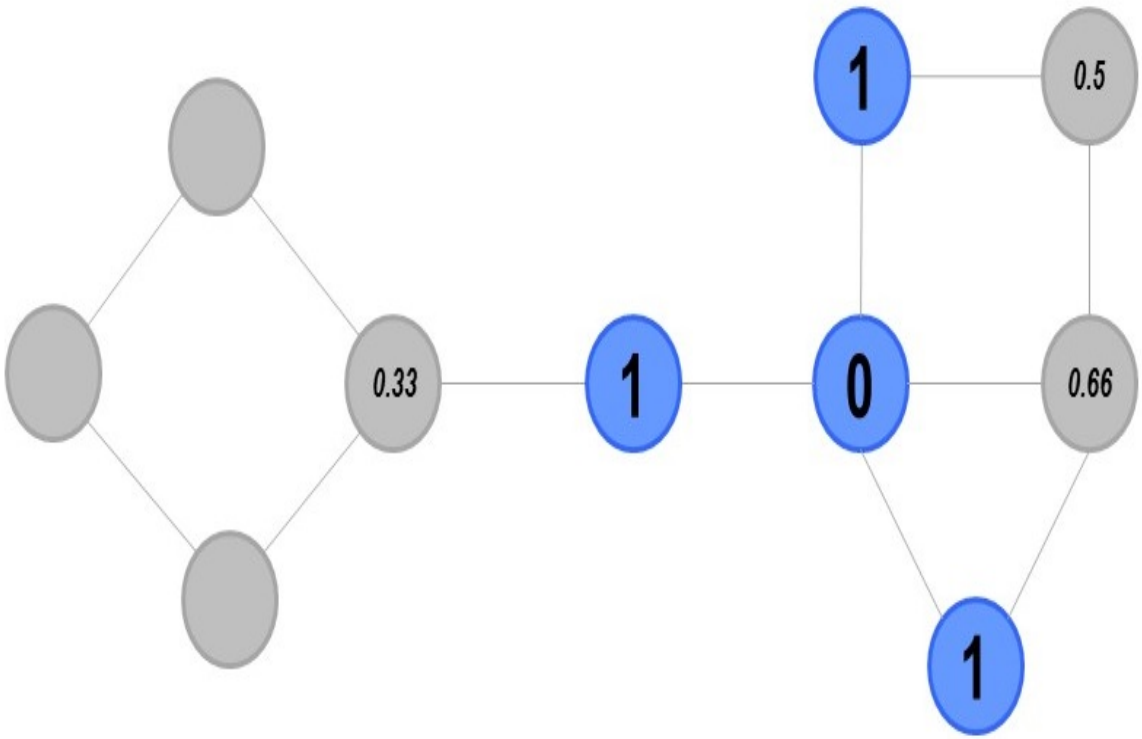


Figure: First set of nodes in which change their behaviour to blue. They would have higher membership to the leader

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

DMID II: Cascading behavior

Network coordination game

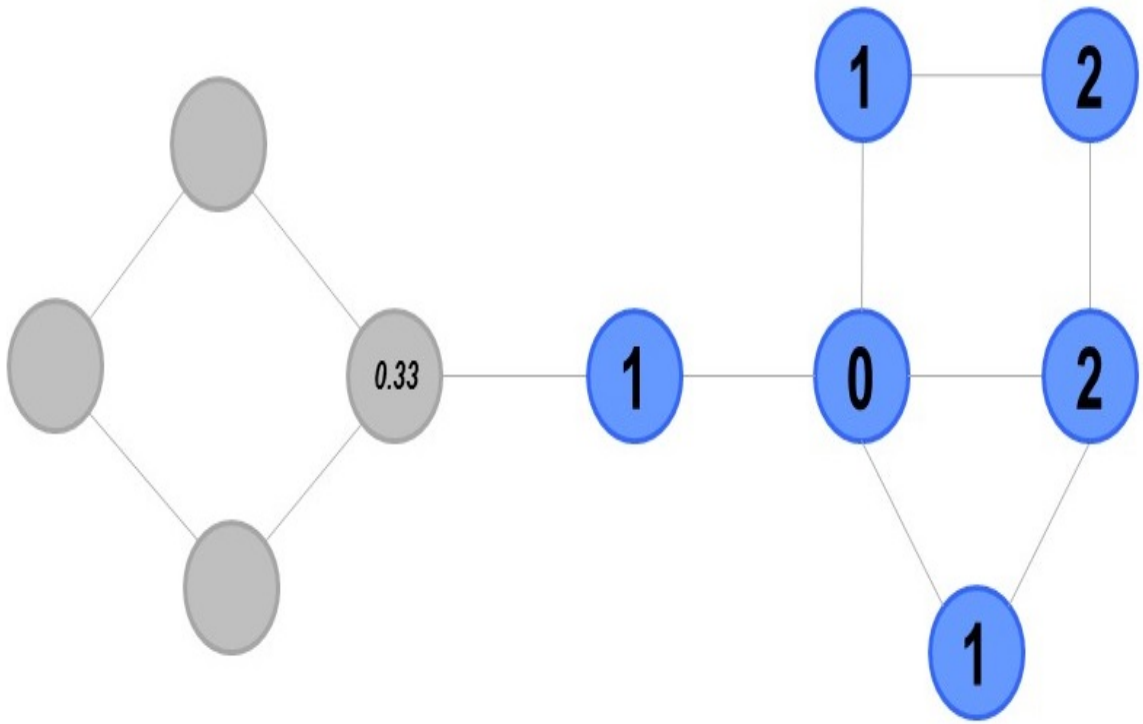


Figure: Cascades stops and other nodes do not accept the blue behaviour

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

DMID II: Cascading behavior

Network coordination game

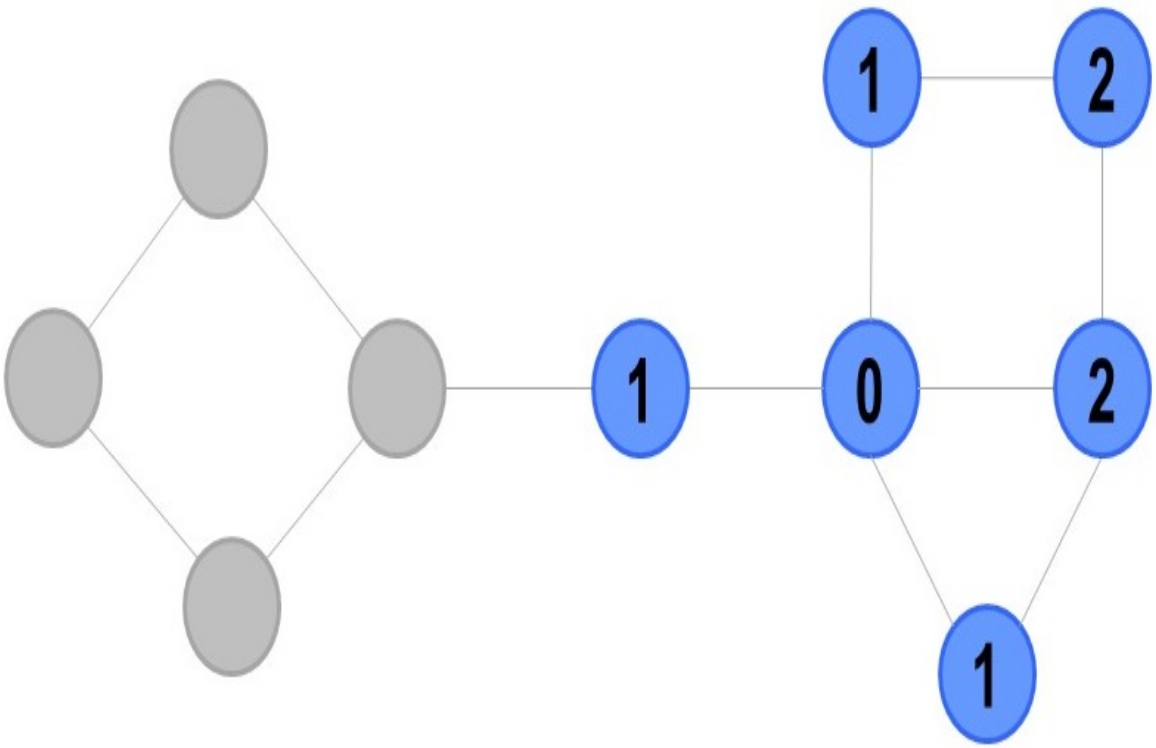


Figure: Node which receives 0.33 does not accept the blue behaviour because it is less than its threshold (0.5).

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

DMID II: Cascading behavior

Network coordination game

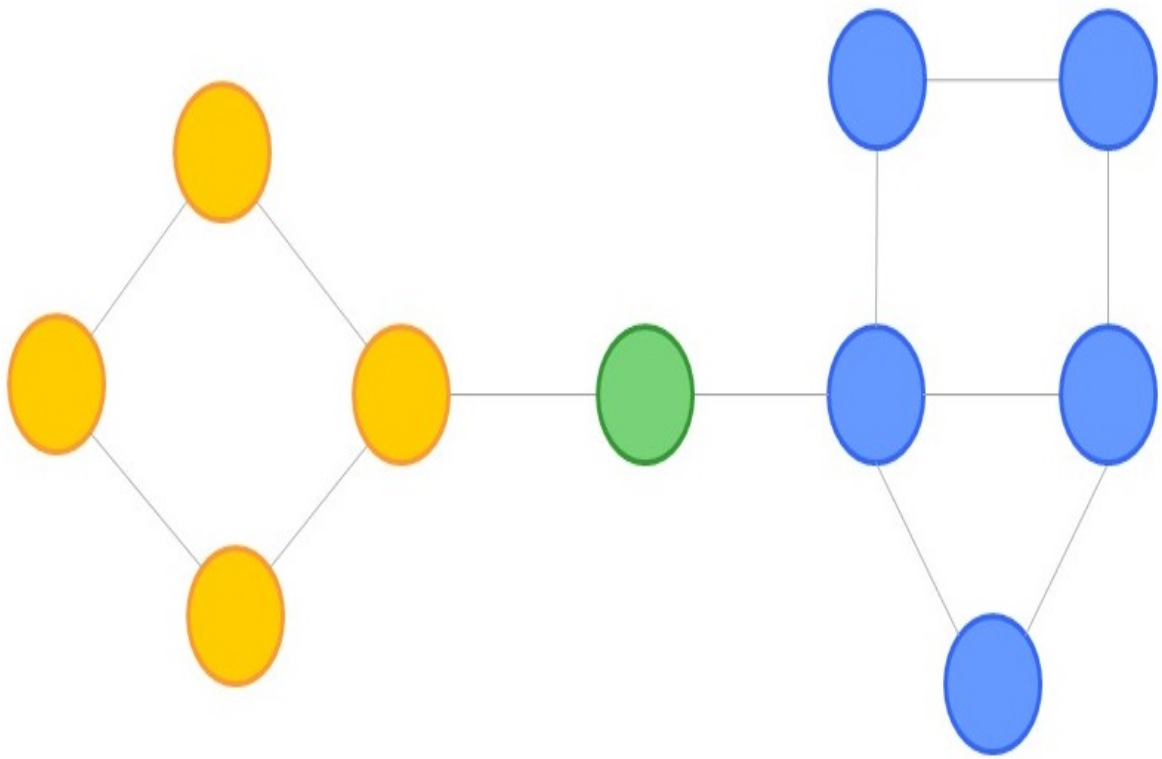


Figure: Similar process can be shown with yellow leader. DMID identifies two communities and the green overlapping node.

Lehrstuhl Informatik 5
(Information Systems)
Prof. Dr. M. Jarke

Overlapping Communities Summary

- Developing research topic
- Multiple approaches exist, suitable for different types of networks - the design must be considered before the network analysis phase
- Can represent better the underlying network structure
- An overlapping structure of a network is not always associated with a hierarchical one
- Multiple benchmarks exist for evaluation, on real and synthetic networks
- Most of the algorithms have a high complexity, but efforts to linearize the time complexity are made

cc BY SA

Reading List to Overlapping Communities

- *Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society* by Palla,G. et al., 2005
- *Link communities reveal multiscale complexity in networks* by Ahn,Y. et al., 2005
- *Overlapping Community Detection in Networks: the State of the Art and Comparative Study* by Xie,J. et al., 2012
- *SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process* by Xie,J. et al., 2011
- *Detecting the Overlapping and Hierarchical Community Structure in Complex Networks* by Lancichinetti,A. et al., 2009
- *Identification of Overlapping Communities and Their Hierarchy by Locally Calculating Community-Changing Resolution Levels* by Havemann,F. et al., 2011

cc BY SA