Community structures

Slides modified from Huan Liu, Lei Tang, Nitin Agarwal

Community Detection

- A community is a set of nodes between which the interactions are (relatively) frequent

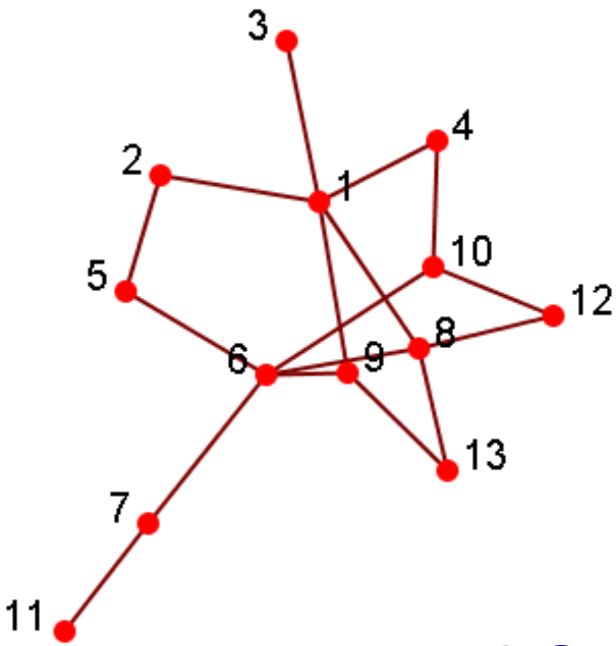
  a.k.a. group, subgroup, module, cluster

- Community detection

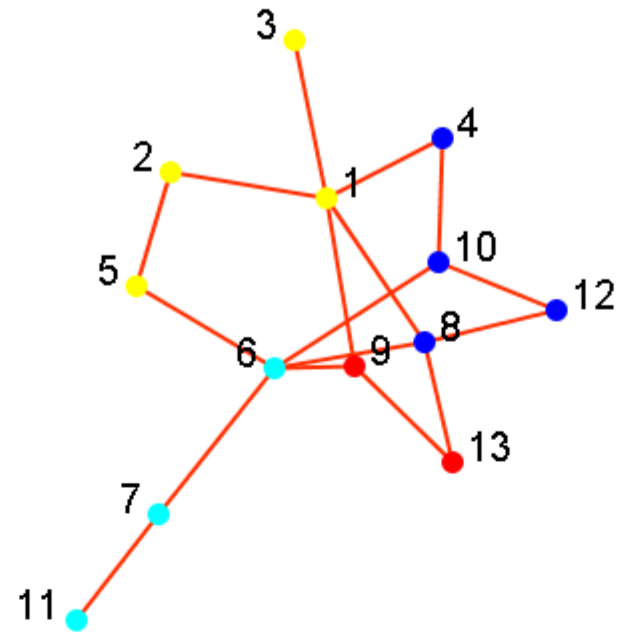
  a.k.a. grouping, clustering, finding cohesive subgroups
  - Given: a social network
  - Output: community membership of (some) actors

- Applications
  - Understanding the interactions between people
  - Visualizing and navigating huge networks
  - Forming the basis for other tasks such as data mining

Visualization after Grouping



4 Groups:
{1,2,3,5}
{4,8,10,12}
{6,7,11}
{9,13}



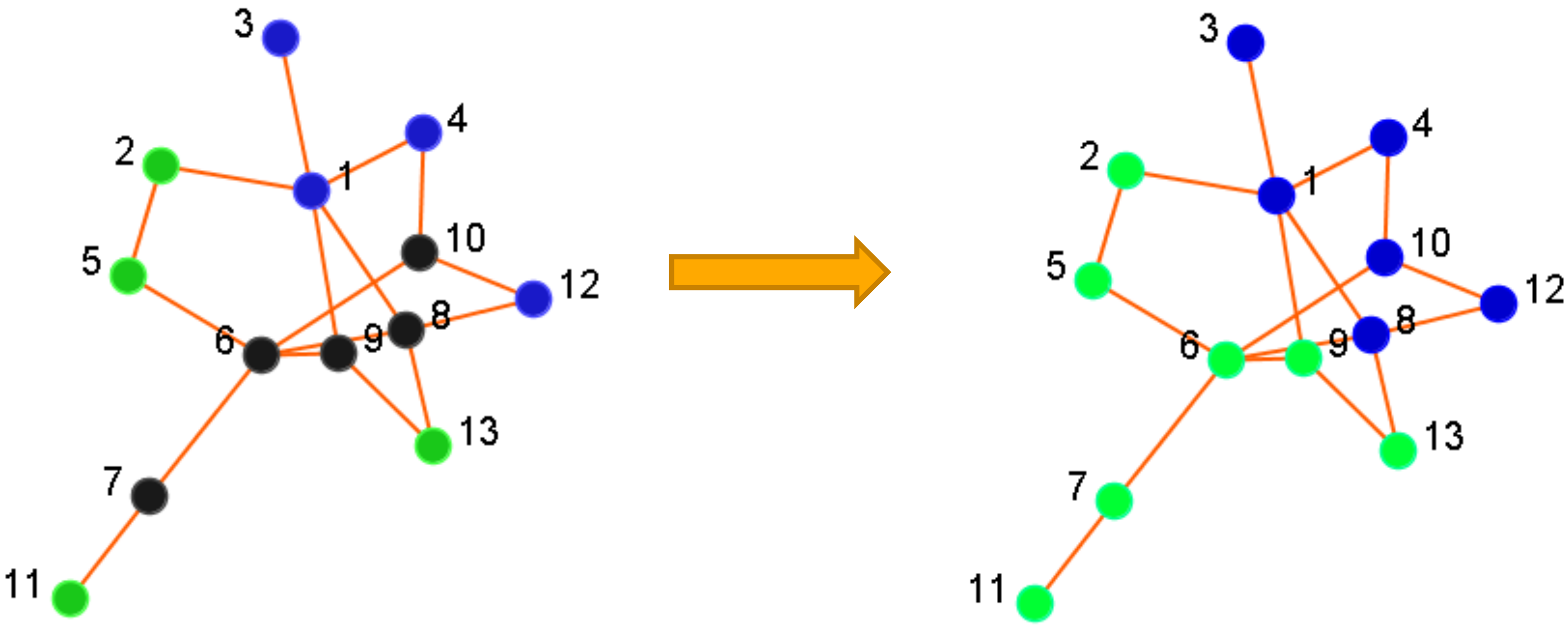
(Nodes colored by Community Membership)

Classification

- User Preference or Behavior can be represented as class labels
  - Whether or not clicking on an ad
  - Whether or not interested in certain topics
  - Subscribed to certain political views
  - Like/Dislike a product

- Given
  - A social network
  - Labels of some actors in the network
- Output
  - Labels of remaining actors in the network

Visualization after Prediction



: Smoking

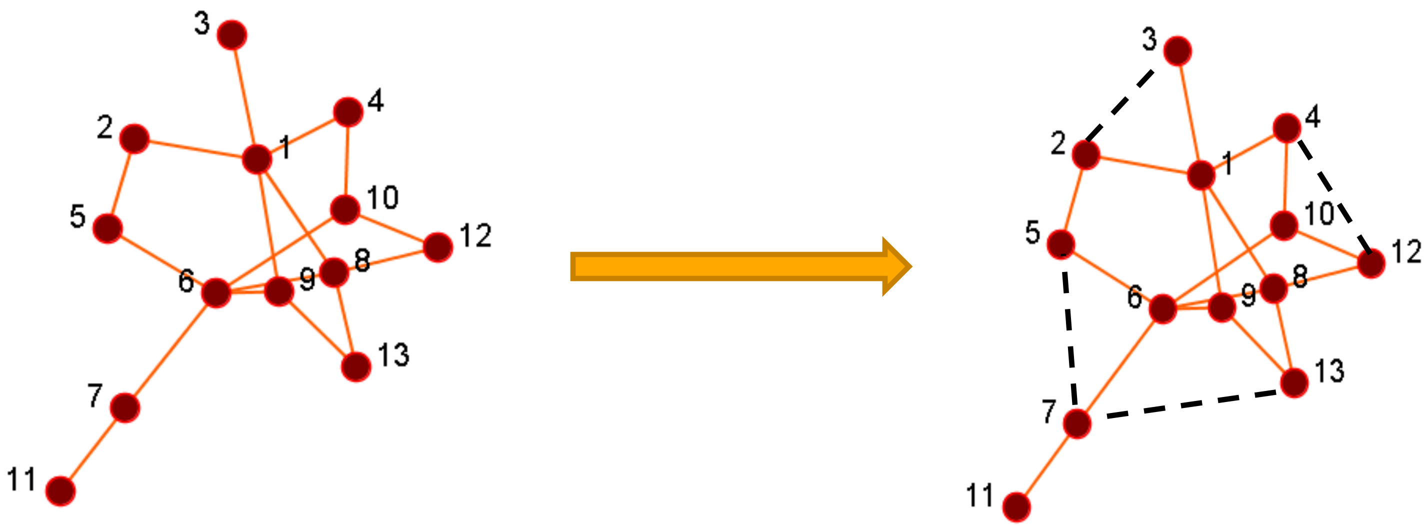
: Non-Smoking

: ? Unknown

Predictions
6: Non-Smoking
7: Non-Smoking
8: Smoking
9: Non-Smoking
10: Smoking

Link Prediction

- Given a social network, predict which nodes are likely to get connected
- Output a list of (ranked) pairs of nodes
- Example: Friend recommendation in Facebook



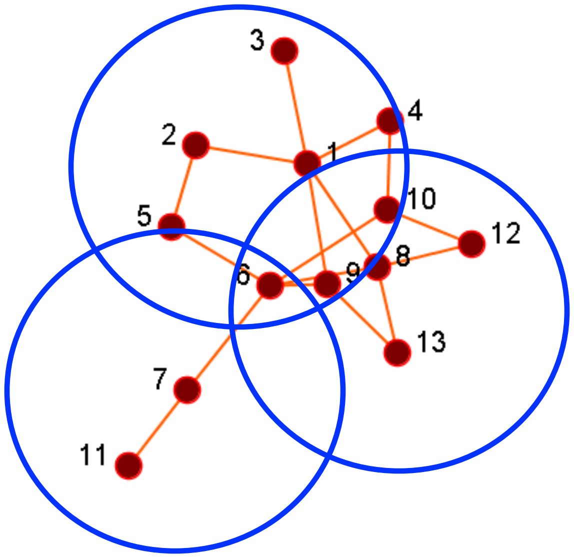
(2, 3)
(4, 12)
(5, 7)
(7, 13)

Viral Marketing/Outbreak Detection

- Users have different social capital (or network values) within a social network, hence, how can one make best use of this information?
- Viral Marketing: find out a set of users to provide coupons and promotions to influence other people in the network so my benefit is maximized
- Outbreak Detection: monitor a set of nodes that can help detect outbreaks or interrupt the infection spreading (e.g., H1N1 flu)
- Goal: given a limited budget, how to maximize the overall benefit?

An Example of Viral Marketing

- Find the coverage of the whole network of nodes with the minimum number of nodes
- How to realize it – an example
  - Basic Greedy Selection: Select the node that maximizes the utility, remove the node and then repeat



- Select Node 1
- Select Node 8
- Select Node 7

Node 7 is not a node with high centrality!

PRINCIPLES OF COMMUNITY DETECTION

Communities

- Community: “subsets of actors among whom there are relatively strong, direct, intense, frequent or positive ties.”

  -- Wasserman and Faust, *Social Network Analysis, Methods and Applications*

- Community is a set of actors interacting with each other *frequently*

- A set of people without interaction is NOT a community
  - e.g. people waiting for a bus at station but don’t talk to each other

Example of Communities

Communities from Facebook

Name: **Social Computing**
Type: Organizations
Members: 14 members

Name: **Social Computing**
Type: Internet & Technology
Members: 12 members

Name: **Social Computing Magazine**
Type: Internet & Technology
Members: 34 members

Name: **Trustworthy Social Computing**
Type: Internet & Technology
Members: 28 members

Name: **Social Computing for Business**
Type: Internet & Technology
Members: 421 members

Name: **UCLA Social Sciences Computing**
Type: Internet & Technology
Members: 22 members

Name: **Social Media and Computing**
Type: Organizations
Members: 6 members

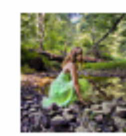
Communities from Flickr

! * Urban LIFE in Metropolis ////
4,286 members | 31 discussions | 89,645 items | Created 46 months ago | Join?
UrbanLIFE, People, Parties, Dance, Musik, Life, Love, Culture, Food and Everything what we could imagine by hearing that word URBANLIFE! Have some FUN! Please add... (more)

Islam Is The Way Of Life (Muslim World)
619 members | 13 discussions | 2,685 items | Created 23 months ago | Join?
The word islām is derived from the Arabic verb aslama, which means to accept, surrender or submit. Thus, Islam means submission to and acceptance of God, and believers must... (more)

*** THE CELEBRATION OF ~LIFE~ (Post1~Award1) [only living things]**
4,871 members | 22 discussions | 40,519 items | Created 21 months ago | Join?
WELCOME to THE CELEBRATION OF ~LIFE~ (Post1~Award1) PLEASE INVITE & COMMENT USING only THE CODES FOUND BELOW! ☆ ☆ This group is for sharing BEAUTIFUL, TOP QUALITY images... (more)

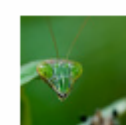
"Enjoy Life!"
2,027 members | 10 discussions | 39,916 items | Created 23 months ago | Join?
There are lovely moments and adorable scenes in our lives. Some are in front of you, and some are just waiting to be discovered. A gaze from someone we love, might touch the... (more)



Baby's life
2,047 members | 185 discussions | 30,302 items | Created 32 months ago | Join?
This group is designed to highlight milestones and important events in your baby's life (ie 1st time smiling/crawling/sitting in a high chair/reading/playing etc). It can also be... (more)

Only group members s pool

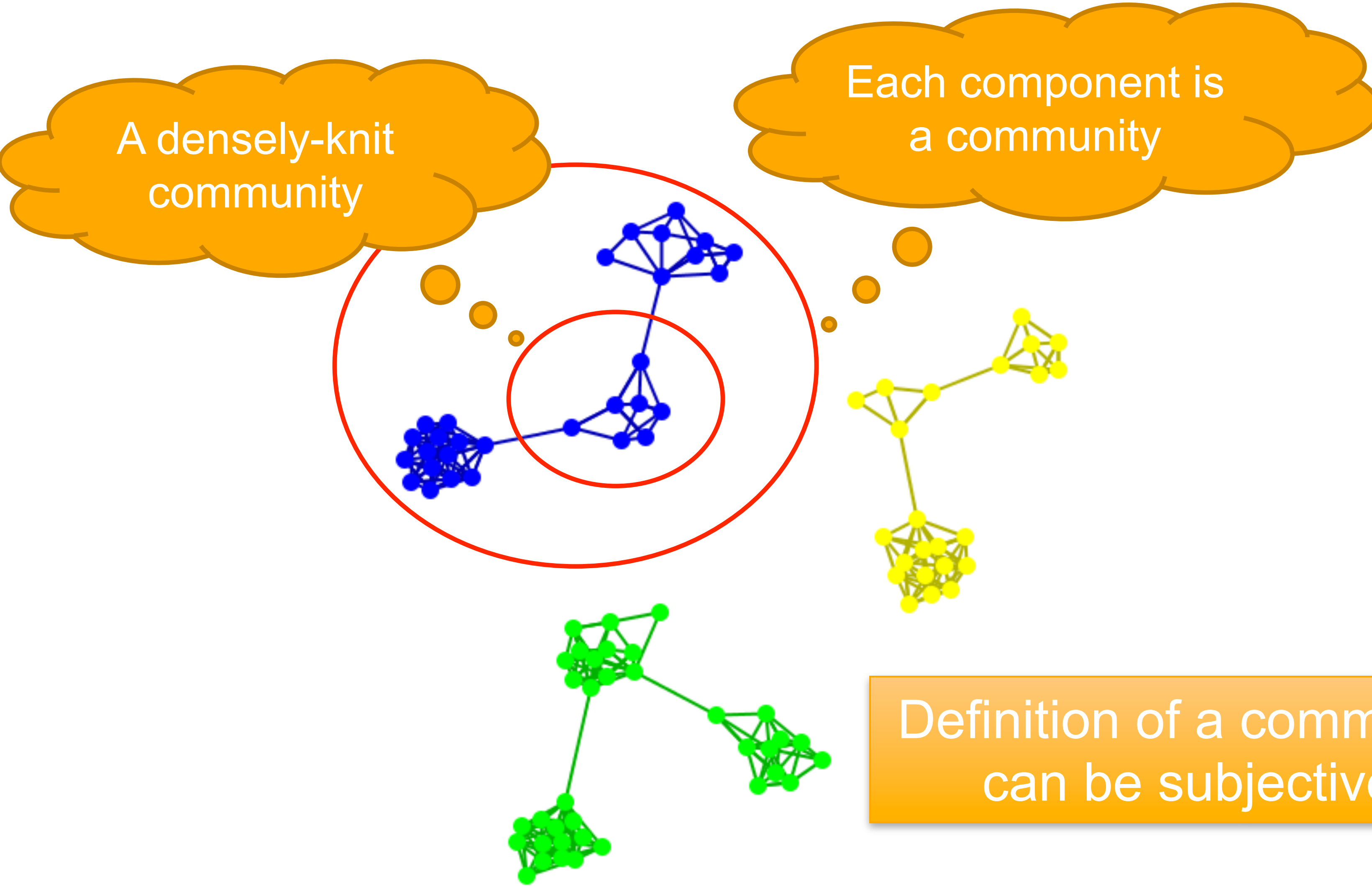
Pond Life
903 members | 20 discussions | 6,877 items | Created 32 months ago | Join?
Pic of the week: chosen from the pool by the group admins. Nuphar by guus timpers Pond Life is a group for all aquatic flora and fauna. Koi ponds, wildlife ponds, garden ponds,... (more)



Community Detection

- Community Detection: "formalize the strong social groups based on the social network properties"
- Some social media sites allow people to join groups
  - Not all sites provide community platform
  - Not all people join groups
- Network interaction provides rich information about the relationship between users
  - Is it necessary to extract groups based on network topology?
  - Groups are *implicitly* formed
  - Can complement other kinds of information
  - Provide basic information for other tasks

Subjectivity of Community Definition

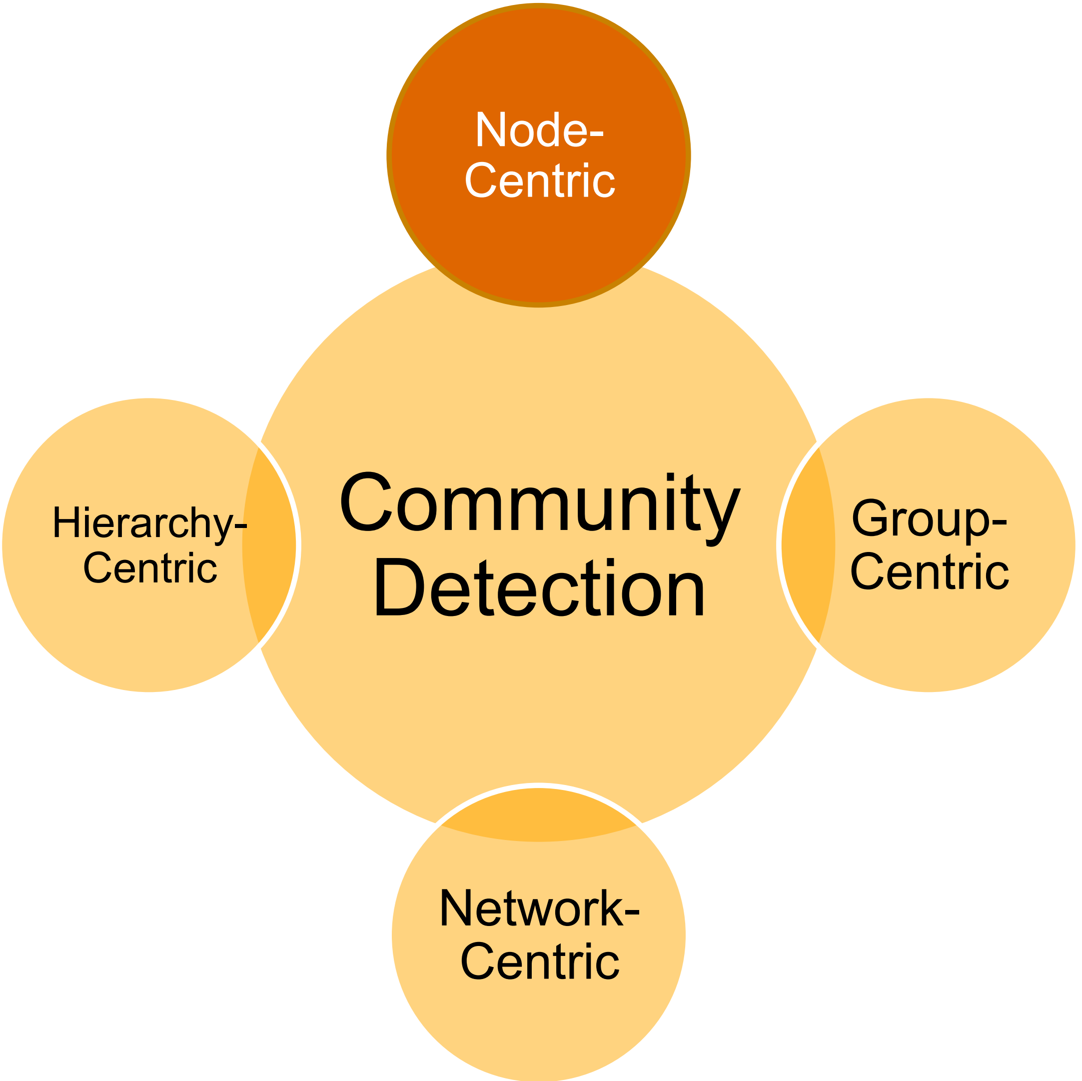


Definition of a community can be subjective.

Taxonomy of Community Criteria

- Criteria vary depending on the tasks
- Roughly, community detection methods can be divided into 4 categories (not exclusive):
- Node-Centric Community
  - Each node in a group satisfies certain properties
- Group-Centric Community
  - Consider the connections within a group as a whole. The group has to satisfy certain properties without zooming into node-level
- Network-Centric Community
  - Partition the whole network into several disjoint sets
- Hierarchy-Centric Community
  - Construct a hierarchical structure of communities

Node-Centric Community Detection

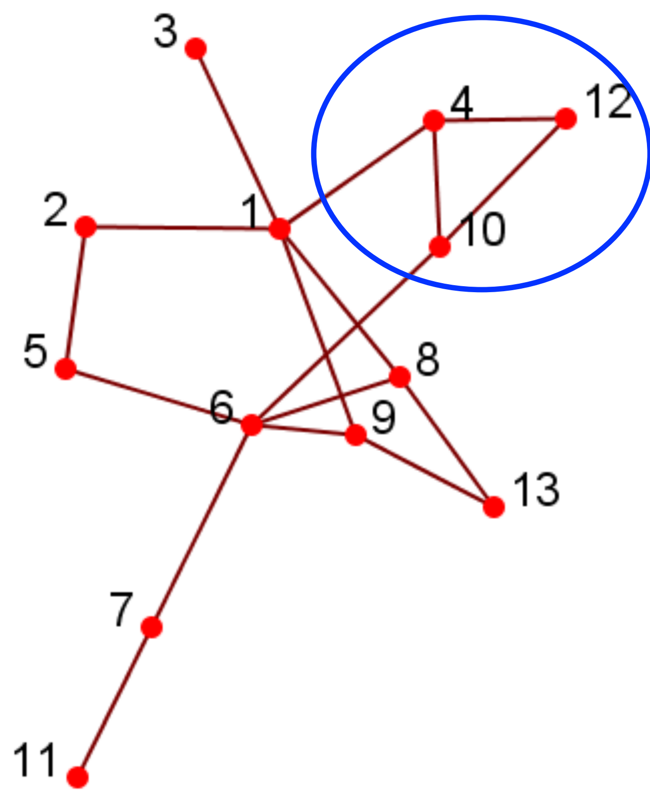


Node-Centric Community Detection

- Nodes satisfy different properties
  - Complete Mutuality
    - cliques
  - Reachability of members
    - k-clique, k-clan, k-club
  - Nodal degrees
    - k-plex, k-core
  - Relative frequency of Within-Outside Ties
    - LS sets, Lambda sets
- Commonly used in traditional social network analysis

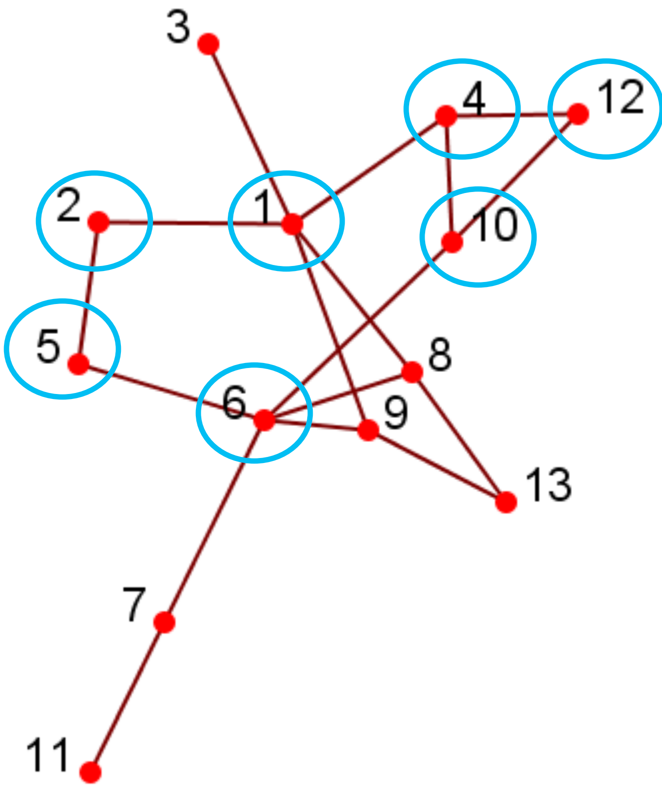
Complete Mutuality: Clique

- A maximal complete subgraph of three or more nodes all of which are adjacent to each other
- NP-hard to find the maximal clique
- *Recursive pruning*: To find a clique of size k, remove those nodes with less than k-1 degrees
- Normally use cliques as a core or seed to explore larger communities



Geodesic

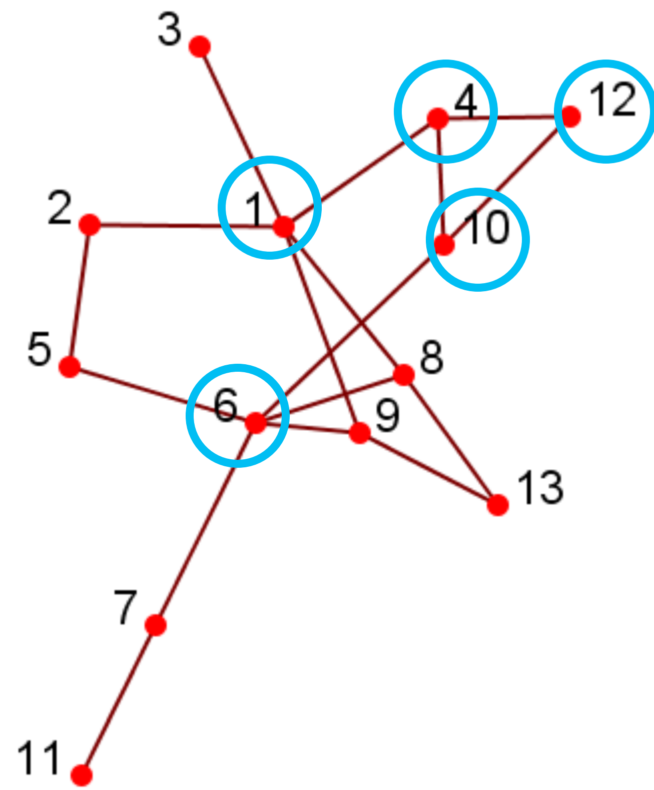
- Reachability is calibrated by the Geodesic distance
- Geodesic: a shortest path between two nodes (12 and 6)
  - Two paths: 12-4-1-2-5-6, 12-10-6
  - 12-10-6 is a geodesic
- Geodesic distance: #hops in geodesic between two nodes
  - e.g., d(12, 6) = 2, d(3, 11)=5
- Diameter: the maximal geodesic distance for any 2 nodes in a network
  - #hops of the longest shortest path



Diameter = 5

Reachability: k-clique, k-club

- Any node in a group should be reachable in k hops
- k-clique: a maximal subgraph in which the largest geodesic distance between any nodes <= k
- A k-clique can have diameter larger than k within the subgraph
  - e.g., 2-clique {12, 4, 10, 1, 6}
  - Within the subgraph d(1, 6) = 3



- k-club: a substructure of diameter <= k
  - e.g., {1,2,5,6,8,9}, {12, 4, 10, 1} are 2-clubs

Nodal Degrees: k-core, k-plex

- Each node should have a certain number of connections to nodes within the group
  - k-core: a substracture that each node connects to at least k members within the group
  - k-plex: for a group with n_s nodes, each node should be adjacent no fewer than n_s-k in the group
- The definitions are complementary
  - A k-core is a (n_s-k)-plex

Within-Outside Ties: LS sets

- LS sets: Any of its proper subsets has more ties to other nodes in the group than outside the group

- Too strict, not reasonable for network analysis

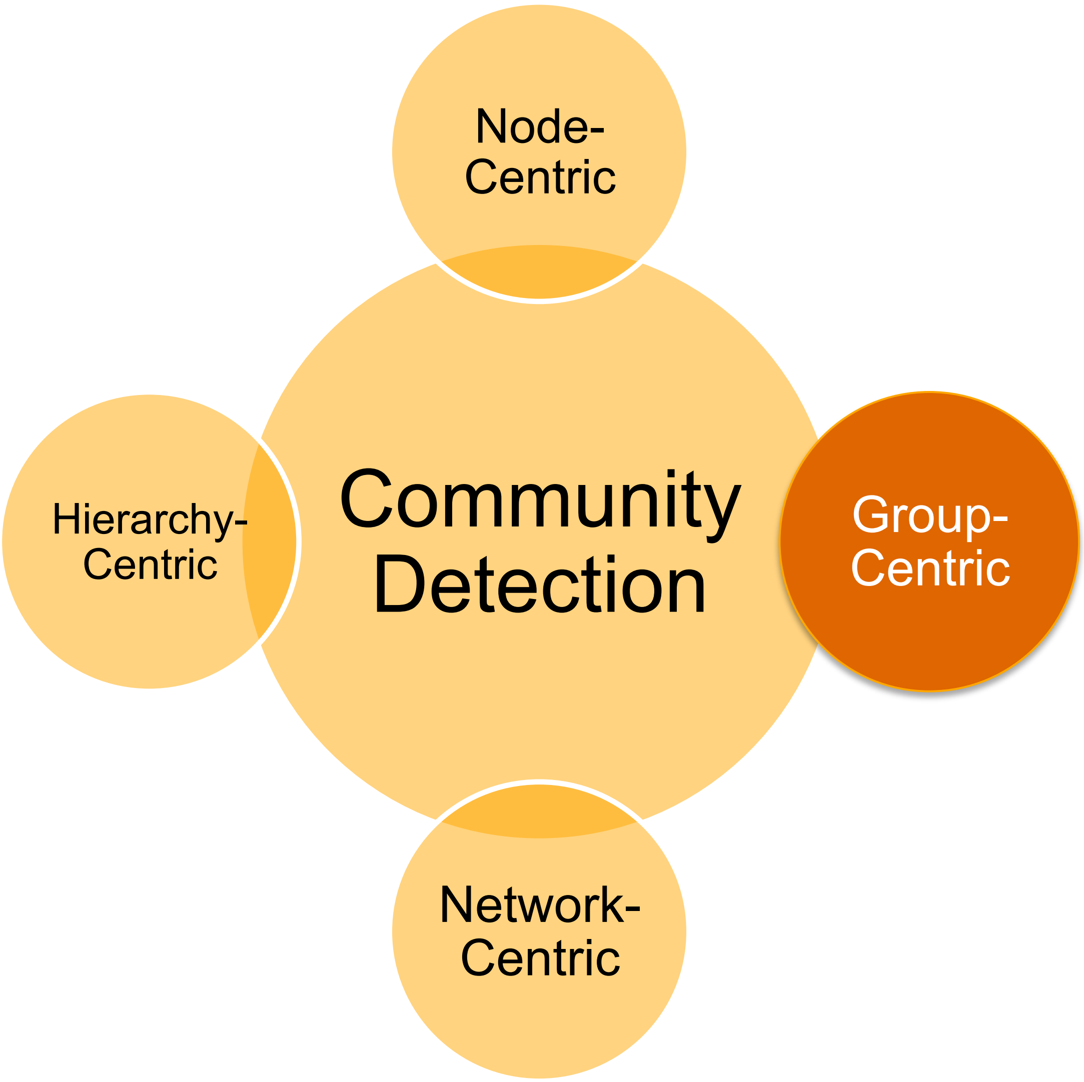
- A relaxed definition is Lambda sets
  - Require the computation of edge-connectivity between any pair of nodes via minimum-cut, maximum-flow algorithm

Recap of Node-Centric Communities

- Each node has to satisfy certain properties
  - Complete mutuality
  - Reachability
  - Nodal degrees
  - Within-Outside Ties

- Limitations:
  - Too strict, but can be used as the core of a community
  - Not scalable, commonly used in network analysis with small-size network
  - Sometimes not consistent with property of large-scale networks
    - e.g., nodal degrees for scale-free networks

Group-Centric Community Detection



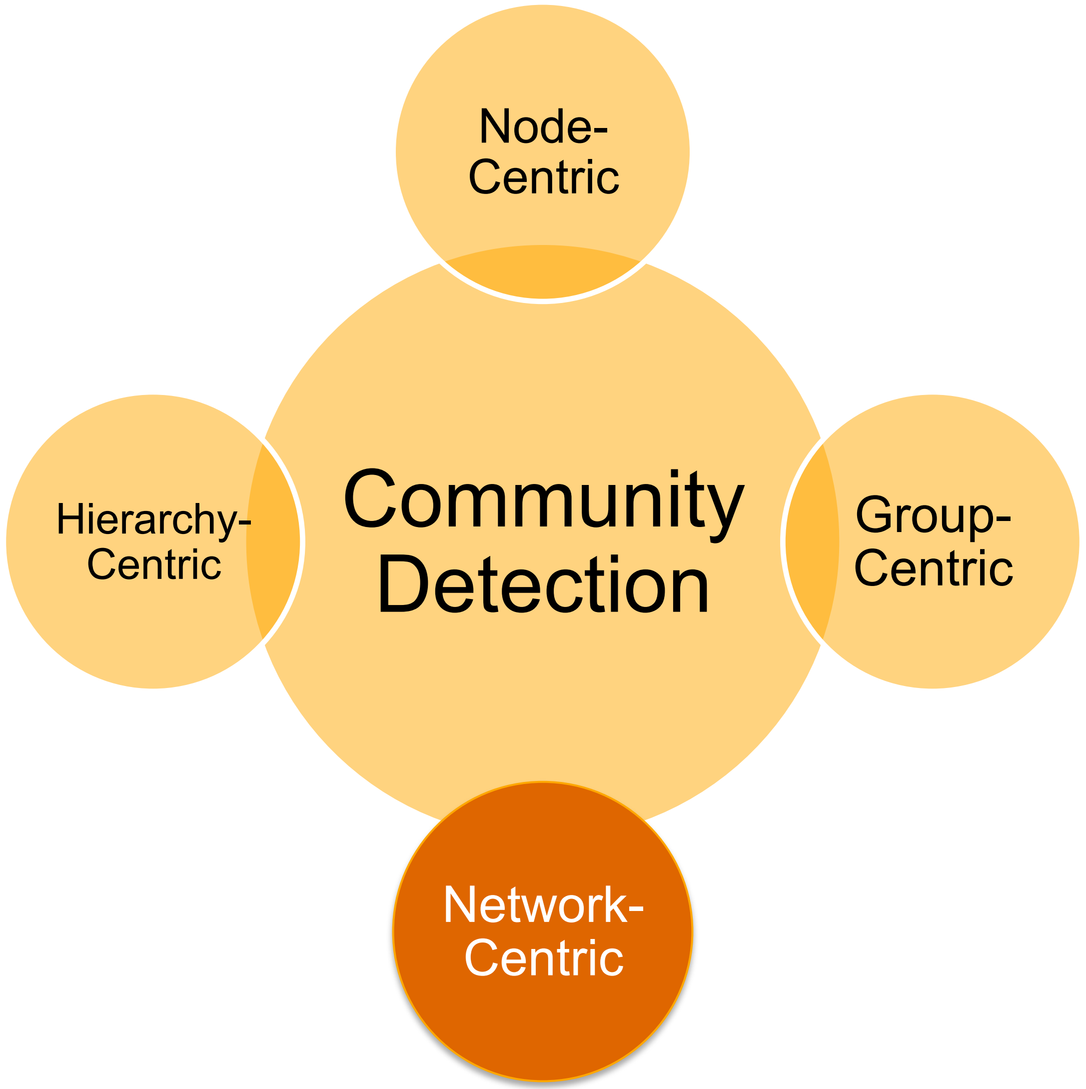
Group-Centric Community Detection

- Consider the connections within a group as whole,
- Some nodes may have low connectivity
- A subgraph with V_s nodes and E_s edges is a γ -dense quasi-clique if

$$\frac{E_s}{V_s(V_s-1)/2} \geq \gamma$$

- Recursive pruning:
  - Sample a subgraph, find a maximal γ -dense quasi-clique
    - the resultant size = k
  - Remove the nodes that
    - whose degree < $k\gamma$
    - all their neighbors with degree < $k\gamma$

Network-Centric Community Detection



Network-Centric Community Detection

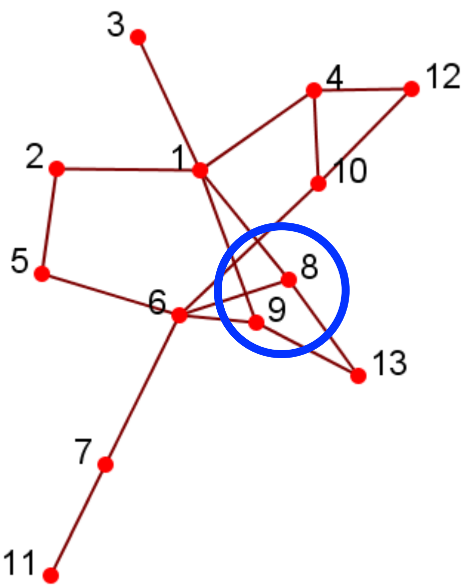
- To form a group, we need to consider the connections of the nodes globally.

- Goal: partition the network into disjoint sets

- Groups based on
  - Node Similarity
  - Latent Space Model
  - Block Model Approximation
  - Cut Minimization
  - Modularity Maximization

Node Similarity

- Node similarity is defined by how similar their interaction patterns are
- Two nodes are structurally equivalent if they connect to the same set of actors
  - e.g., nodes 8 and 9 are structurally equivalent
- Groups are defined over equivalent nodes
  - Too strict
  - Rarely occur in a large-scale
  - Relaxed equivalence class is difficult to compute
- In practice, use vector similarity
  - e.g., cosine similarity, Jaccard similarity



Vector Similarity

a vector

structurally equivalent

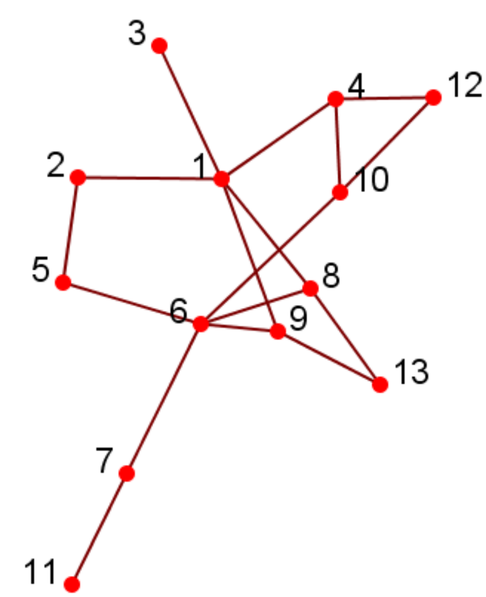
	1	2	3	4	5	6	7	8	9	10	11	12	13
5		1				1							
8	1					1							1
9	1					1							1

Cosine Similarity: $$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

$$sim(5,8) = \frac{1}{\sqrt{2} \times \sqrt{3}} = \frac{1}{\sqrt{6}}$$

Jaccard Similarity: $$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

$$J(5,8) = \frac{|\{6\}|}{|\{1,2,6,13\}|} = 1/4$$



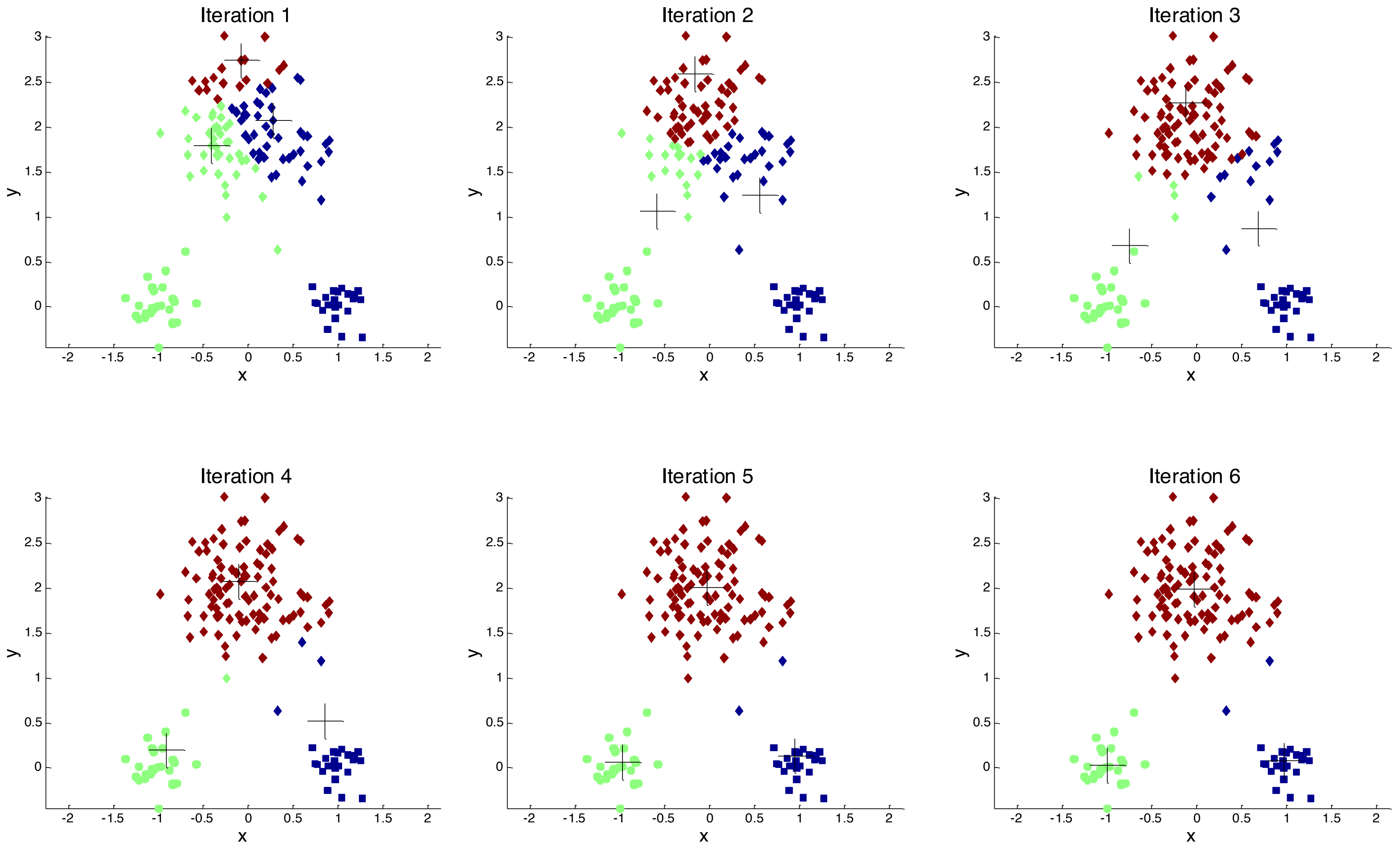
Clustering based on Node Similarity

- For practical use with huge networks:
  - Consider the connections as features
  - Use Cosine or Jaccard similarity to compute vertex similarity
  - Apply classical k-means clustering Algorithm
- K-means Clustering Algorithm
  - Each cluster is associated with a centroid (center point)
  - Each node is assigned to the cluster with the closest centroid

Algorithm 1 Basic K-means Algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3:     Form K clusters by assigning all points to the closest centroid.
4:     Recompute the centroid of each cluster.
5: **until** The centroids don't change

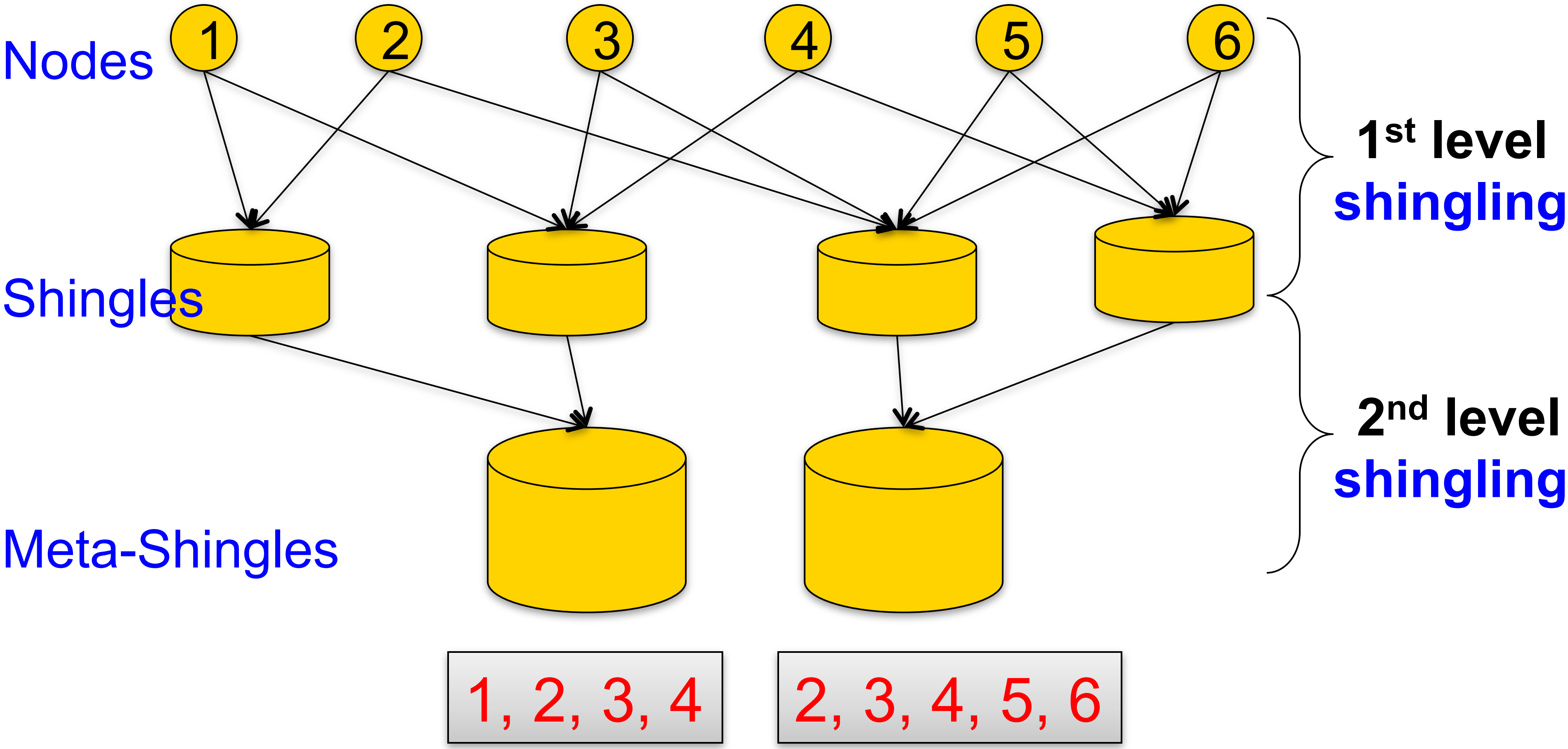
Illustration of k-means clustering



Shingling

- Pair-wise computation of similarity can be time consuming with millions of nodes
- Shingling can be exploited
  - Mapping each vector into multiple shingles so the Jaccard similarity between two vectors can be computed by comparing the shingles
  - Implemented using a quick hash function
  - Similar vectors share more shingles after transformation
- Nodes of the same shingle can be considered belonging to one community
- In reality, we can apply 2-level shingling

Fast Two-Level Shingling



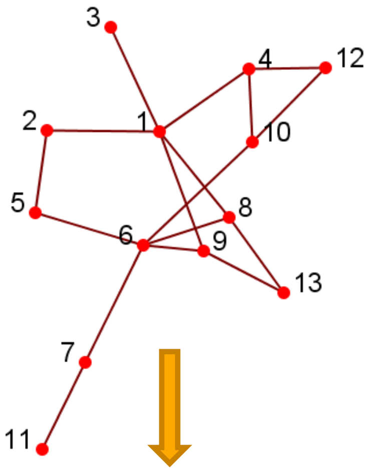
Groups on Latent-Space Models

- **Latent-space models**: Transform the nodes in a network into a lower-dimensional space such that the distance or similarity between nodes are kept in the Euclidean space
- Multidimensional Scaling (MDS)
  - Given a network, construct a proximity matrix to denote the distance between nodes (e.g. geodesic distance)
  - Let D denotes the *square distance* between nodes
  - $S \in R^{n \times k}$ denotes the coordinates in the lower-dimensional space

$$SS^T = -\frac{1}{2}(I - \frac{1}{n}ee^T)D(I - \frac{1}{n}ee^T) = \Delta(D)$$

  - Objective: minimize the difference $\min \| \Delta(D) - SS^T \|_F$
  - Let $\Lambda = diag(\lambda_1, \cdots, \lambda_k)$ (the top-k eigenvalues of Δ), V the top-k eigenvectors
  - Solution: $S = V\Lambda^{1/2}$
- Apply k-means to S to obtain clusters

MDS-example



1, 2, 3, 4, 10, 12

5, 6, 7, 8, 9, 11, 13

k-means

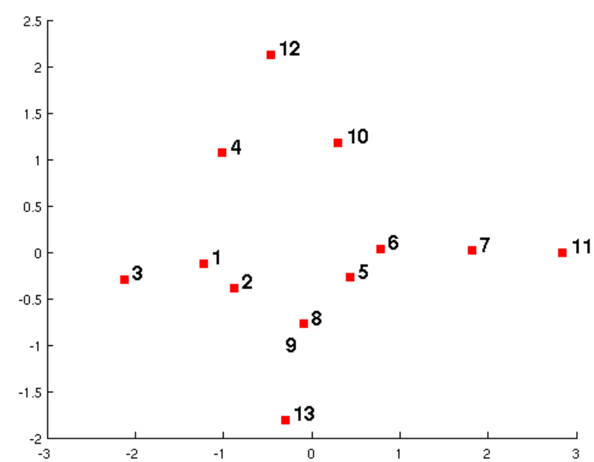
Geodesic Distance Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	1	1	2	2	3	1	1	2	4	2	2
2	1	0	2	2	1	2	3	2	2	3	4	3	3
3	1	2	0	2	3	3	4	2	2	3	5	3	3
4	1	2	2	0	3	2	3	2	2	1	4	1	3
5	2	1	3	3	0	1	2	2	2	2	3	3	3
6	2	2	3	2	1	0	1	1	1	1	2	2	2
7	3	3	4	3	2	1	0	2	2	2	1	3	3
8	1	2	2	2	2	1	2	0	2	2	3	3	1
9	1	2	2	2	2	1	2	2	0	2	3	3	1
10	2	3	3	1	2	1	2	2	2	0	3	1	3
11	4	4	**5**	4	3	2	1	3	3	3	0	4	4
12	2	3	3	1	3	2	3	3	3	1	4	0	4
13	2	3	3	3	3	2	3	1	1	3	4	4	0

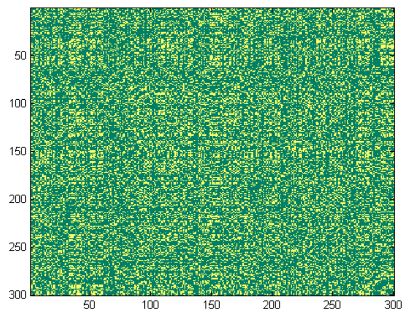
MDS

S

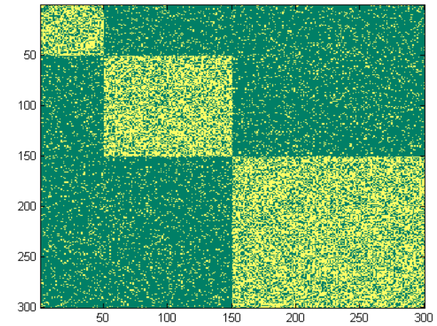
-1.22	-0.12
-0.88	-0.39
-2.12	-0.29
-1.01	1.07
0.43	-0.28
0.78	0.04
1.81	0.02
-0.09	-0.77
-0.09	-0.77
0.30	1.18
2.85	0.00
-0.47	2.13
-0.29	-1.81



Block-Model Approximation



After
Reordering



Network Interaction Matrix

Block Structure

- Objective: Minimize the difference between an interaction matrix and a block structure

$$\min_{S,\Sigma} \|A - S\Sigma S^T\|_F$$

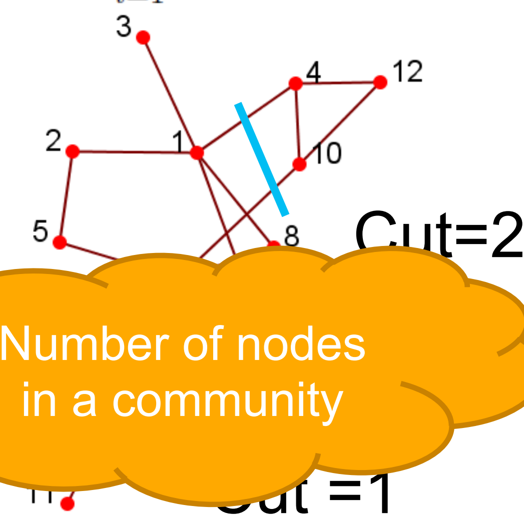
$$s.t. \quad S \in \{0,1\}^{n\times k}, \Sigma \in R^{k\times k} \text{ is diagonal}$$

S is a community indicator matrix

- Challenge: S is discrete, difficult to solve
- Relaxation: Allow S to be continuous satisfying $S^T S = I_k$
- Solution: the top eigenvectors of A
- Post-Processing: Apply k-means to S to find the partition

Cut-Minimization

- Between-group interactions should be infrequent
- Cut: number of edges between two sets of nodes
- Objective: minimize the cut $cut(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} cut(C_i, \overline{C_i})$



  - Limitations: often find communities of only one node
  - Need to consider the group size

Number of nodes in a community

- Two commonly-used variants:

$$\text{Ratio-cut}(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C_i})}{|V_i|}$$

$$\text{Normalized-cut}(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C_i})}{vol(V_i)}$$

Number of within-group Interactions

Graph Laplacian

- Cut-minimization can be relaxed into the following min-trace problem

$$\min_{S \in R^{n \times k}} Tr(S^T L S) \quad s.t.\ S^T S = I$$

- L is the (normalized) Graph Laplacian

$$\begin{aligned} L &= D - A \\ \text{normalized-}L &= I - D^{-1/2} A D^{-1/2} \end{aligned} \qquad D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}$$

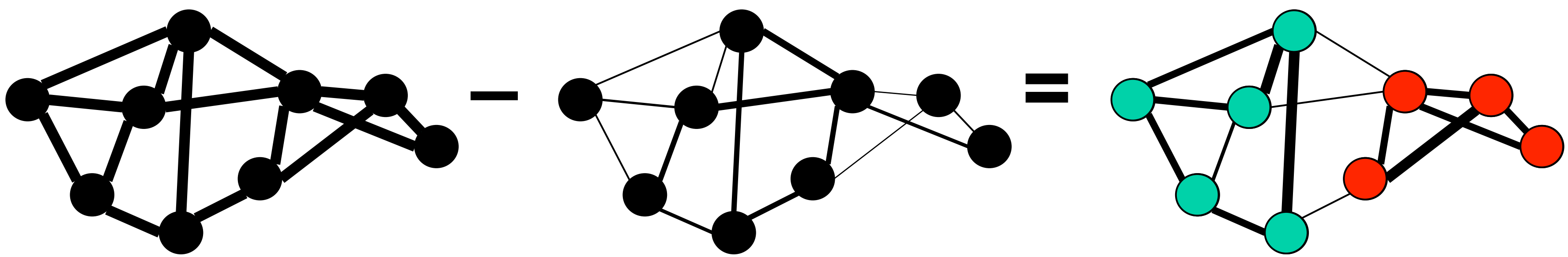
- Solution: S are the eigenvectors of L with smallest eigenvalues (except the first one)
- Post-Processing: apply k-means to S
  - a.k.a. Spectral Clustering

Graph Modularity

- Relational network given by G = (V, A)

  V : set of n vertices A : n x n adjacency matrix, m total edges

- Newman-Girvan (2006) graph modularity



Original A Null Model P Modularity (A-P)

– Measures the global community structure of G:

$$Q(C) = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(C_i, C_j) \qquad P_{ij} = \frac{d_i d_j}{2m}$$

Kronecker delta

– Foundation for a large number of methods (Fortunato, 2010)

Modularity Maximization

- Modularity measures the group interactions compared with the expected random connections in the group
- In a network with m edges, for two nodes with degree d_i and d_j, expected random connections between them are

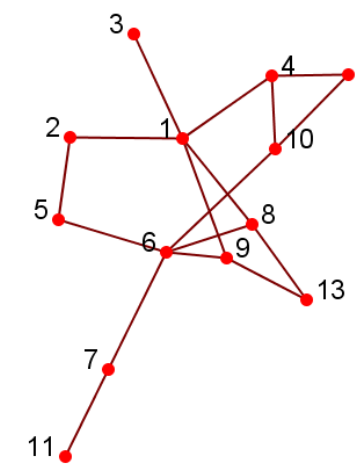
$$d_i d_j / 2m$$

- The interaction utility in a group:

$$\sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m$$

- To partition the group into multiple groups, we maximize

$$\frac{1}{2m} \sum_{C} \sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m$$



Expected Number of edges between 6 and 9 is
5*3/(2*17) = 15/34

Modularity Matrix

- The modularity maximization can also be formulated in matrix form

$$Q = \frac{1}{2m} Tr(S^T B S)$$

- B is the modularity matrix

$$B_{ij} = A_{ij} - d_i d_j / 2m$$

- Solution: top eigenvectors of the modularity matrix

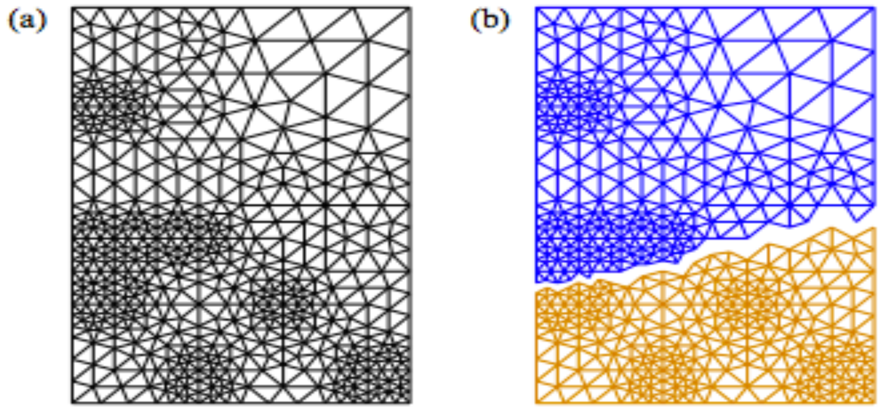
Properties of Modularity

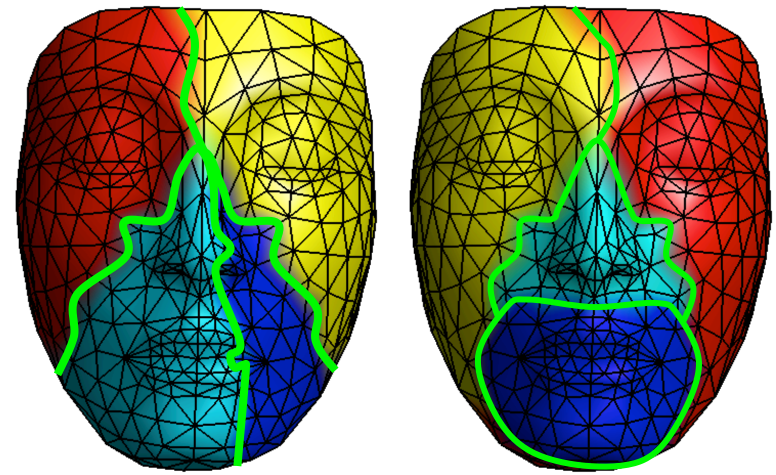
- Properties of modularity:
  - Between (-1, 1)
  - Modularity = 0 If all nodes are clustered into one group
  - Can automatically determine optimal number of clusters

- Resolution limit of modularity
  - Modularity maximization might return a community consisting multiple small modules

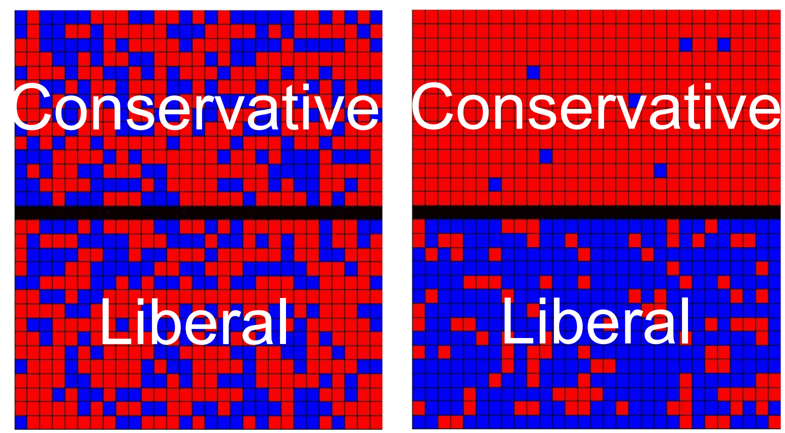
Graph Laplacian vs Graph Modularity

Mesh Network by Bern et al.
partitioned by the Laplacian



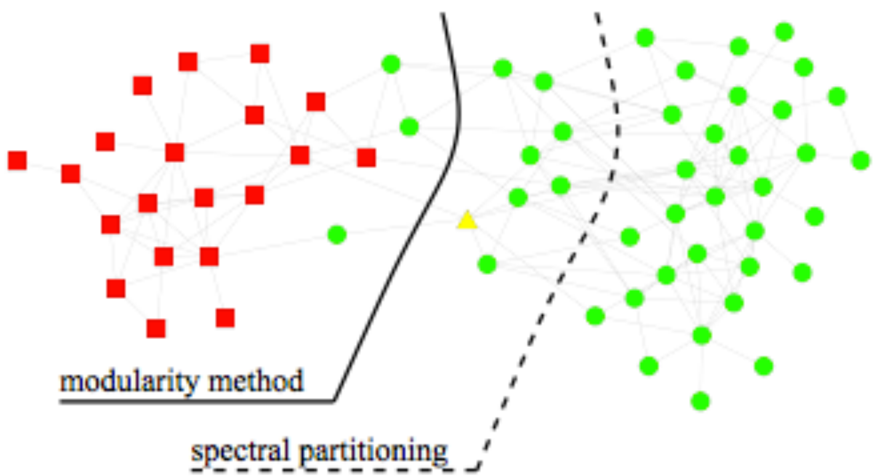


Laplacian Modularity



Political Blogs from 2004 U.S. Election,
data set from Adamic & Glance (2005)

Dolphin social network



Matrix Factorization Form

- For latent space models, block models, spectral clustering and modularity maximization
- All can be formulated as

$$\max(\min)_S \quad Tr(S^T X S)$$
$$s.t. \quad S^T S = I$$

$$X = \begin{cases} \Delta(D) & \text{(Latent Space Models)} \\ \text{Sociomatrix} & \text{(Block Model Approximation)} \\ \text{Graph Laplacian} & \text{(Cut Minimization)} \\ \text{Modularity Matrix} & \text{(Modularity maximization)} \end{cases}$$

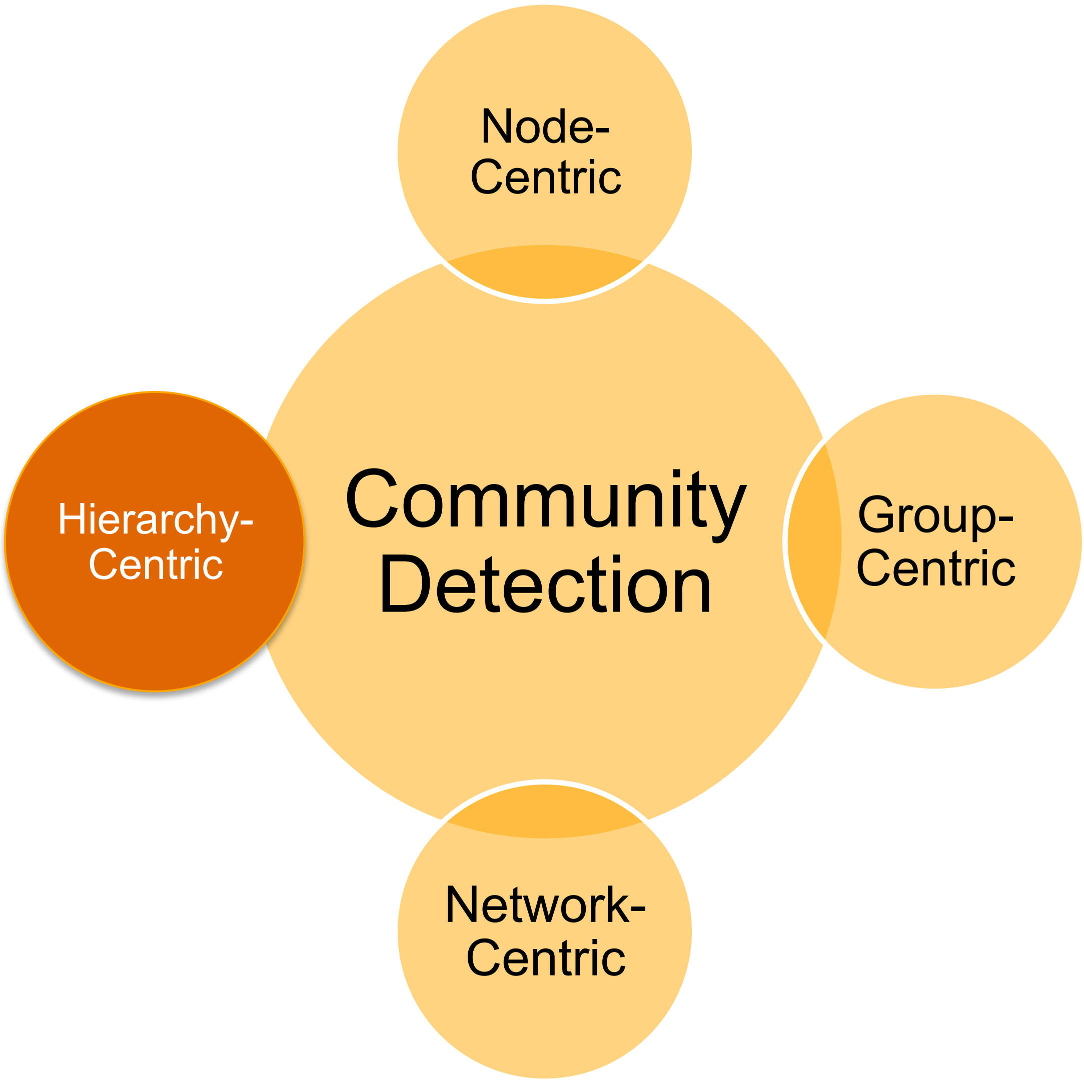
Recap of Network-Centric Community

- Network-Centric Community Detection
  - Groups based on
    - Node Similarity
    - Latent Space Models
    - Cut Minimization
    - Block-Model Approximation
    - Modularity maximization

- Goal: Partition network nodes into several disjoint sets

- Limitation: Require the user to specify the number of communities beforehand

Hierarchy-Centric Community Detection



Hierarchy-Centric Community Detection

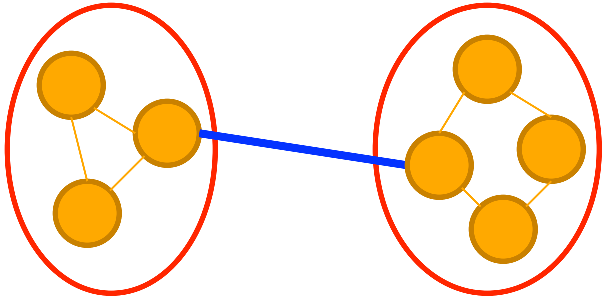
- Goal: Build a hierarchical structure of communities based on network topology

- Facilitate the analysis at different resolutions

- Representative Approaches:
  - Divisive Hierarchical Clustering
  - Agglomerative Hierarchical Clustering

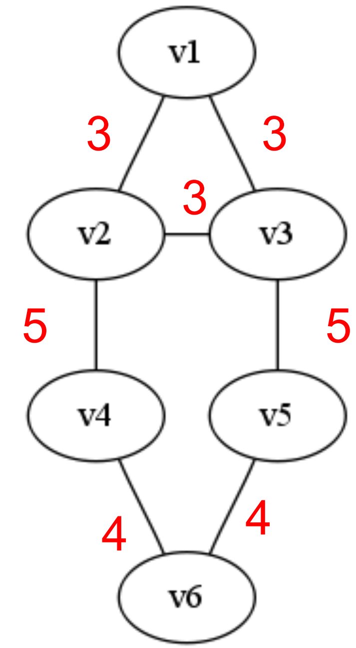
Divisive Hierarchical Clustering

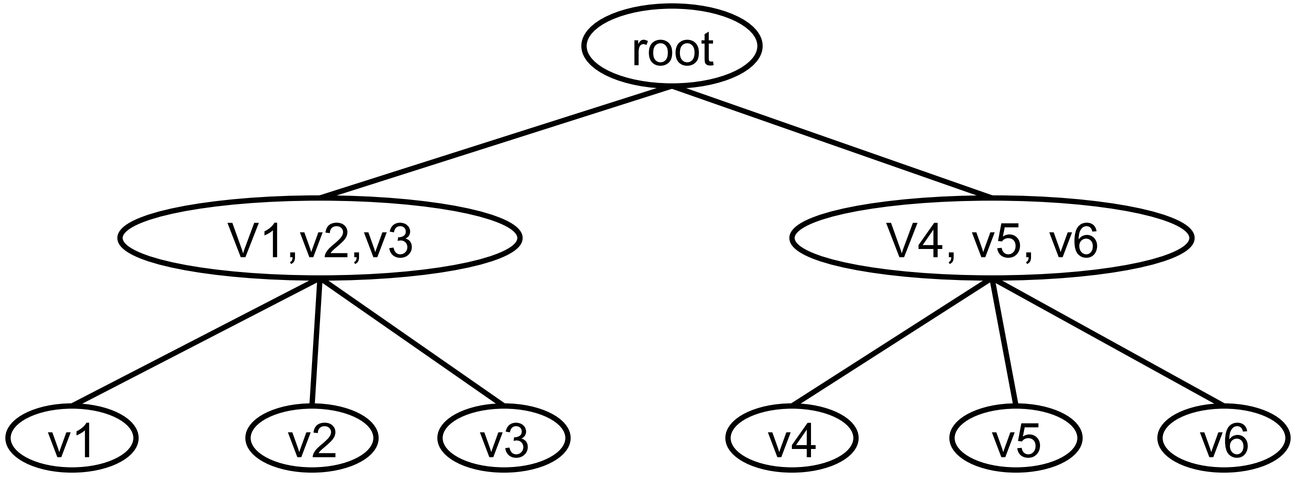
- Divisive Hierarchical Clustering
  - Partition the nodes into several sets
  - Each set is further partitioned into smaller sets
- Network-centric methods can be applied for partition
- One particular example is based on edge-betweenness
  - Edge-Betweenness: Number of shortest paths between any pair of nodes that pass through the edge
- Between-group edges tend to have larger edge-betweenness



Divisive clustering on Edge-Betweenness

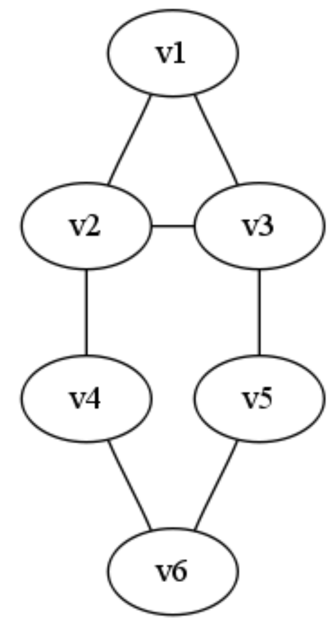
- Progressively remove edges with the highest betweenness
  - Remove e(2,4), e(3, 5)
  - Remove e(4,6), e(5,6)
  - Remove e(1,2), e(2,3), e(3,1)

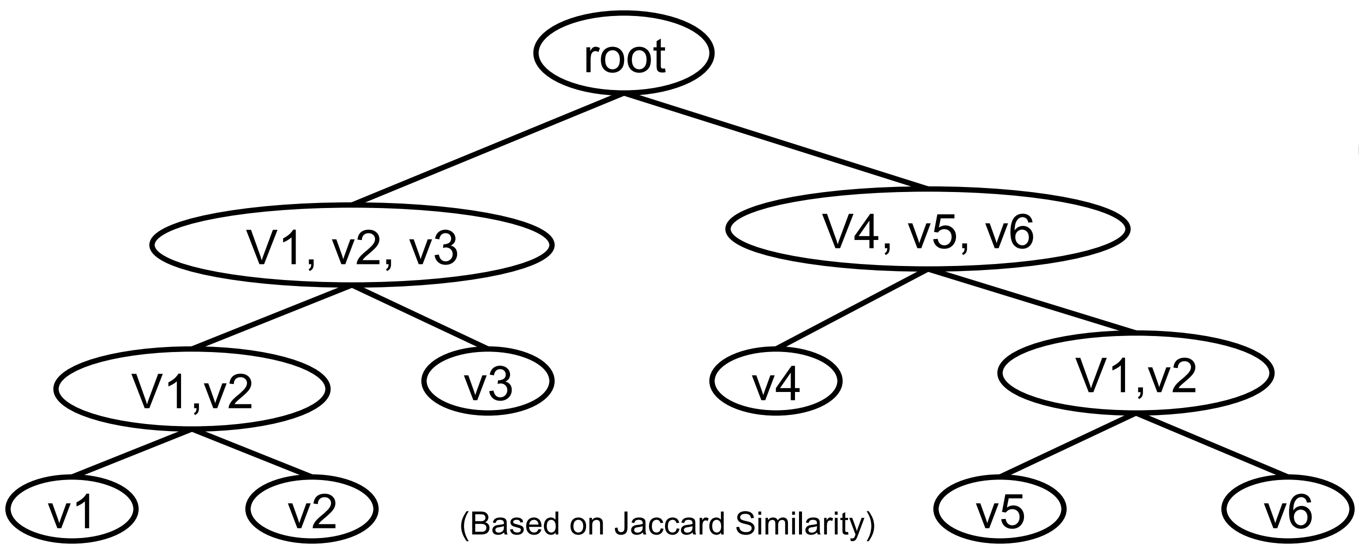




Agglomerative Hierarchical Clustering

- Initialize each node as a community
- Choose two communities satisfying certain criteria and merge them into larger ones
  - Maximum Modularity Increase
  - Maximum Node Similarity

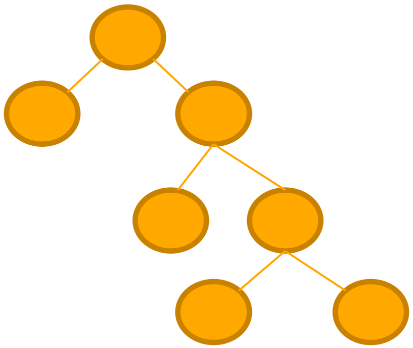




(Based on Jaccard Similarity)

Recap of Hierarchical Clustering

- Most hierarchical clustering algorithm output a binary tree
  - Each node has two children nodes
  - Might be highly imbalanced

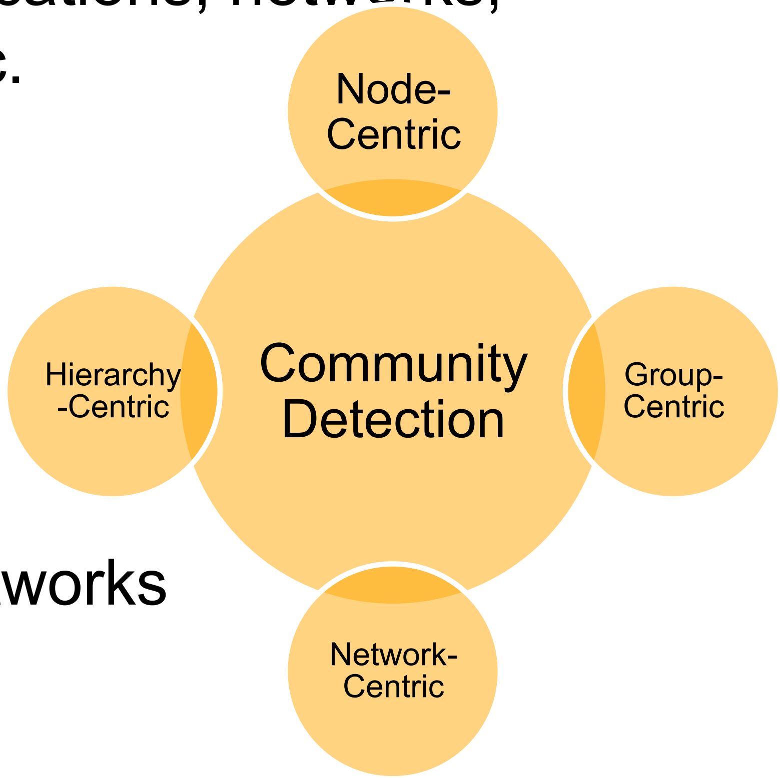


- Agglomerative clustering can be very sensitive to the nodes processing order and merging criteria adopted.

- Divisive clustering is more stable, but generally more computationally expensive

Summary of Community Detection

- The Optimal Method?
  - It varies depending on applications, networks, computational resources etc.



- Other lines of research
  - Communities in directed networks
  - Overlapping communities
  - Community evolution
  - Group profiling and interpretation