Community Analysis

Algorithms and Data Structures 2, 340300
Lecture – 2023W
Univ.-Prof. Dr. Alois Ferscha, teaching@pervasive.jku.at

Institute of
Pervasive Computing

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

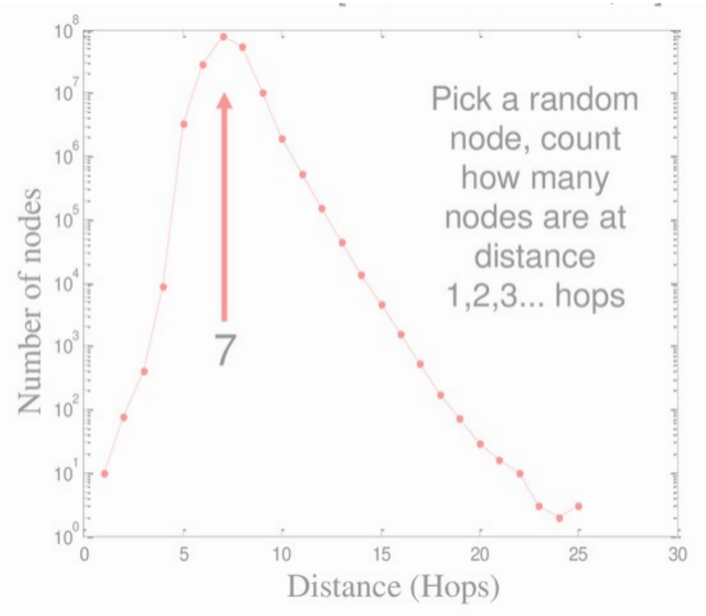
Community Analysis

Remember: Small-World Effect

- Six Degrees of Separation
- A famous experiment conducted by Travers and Milgram (1969)
  - Subjects were asked to send a chain letter to his acquaintance in order to reach a target person
  - The average path length is around **5.5**

Verified on a planetary-scale IM network (Microsoft Messenger) of
240 million people (Leskovec and Horvitz 2008) and 30 billion conversations

- 180 million nodes
- 1.3 billion undirected edges
- The average path length is 6.6



J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In Proceedings of the 17th international conference on World Wide Web, WWW ’08, pages 915–924, New York, NY, USA, 2008. ACM

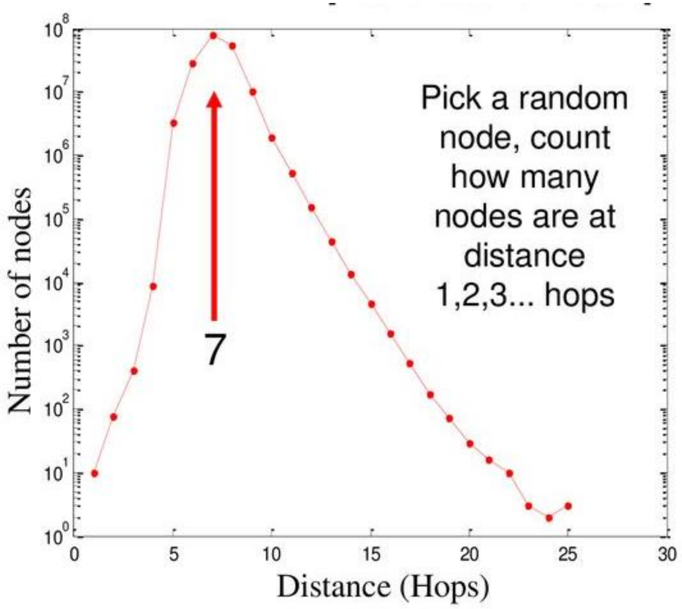
Community Analysis

Remember: Small-World Effect

- Six Degrees of Separation
- A famous experiment conducted by Travers and Milgram (1969)
  - Subjects were asked to send a chain letter to his acquaintance in order to reach a target person
  - The average path length is around **5.5**

Verified on a planetary-scale IM network (Microsoft Messenger) of **240 million people** (Leskovec and Horvitz 2008) and **30 billion conversations**

- 180 million nodes
- 1.3 billion undirected edges
- The average path length is **6.6**



J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In Proceedings of the 17th international conference on World Wide Web, WWW ’08, pages 915–924, New York, NY, USA, 2008. ACM

Communities in Social Media

Community

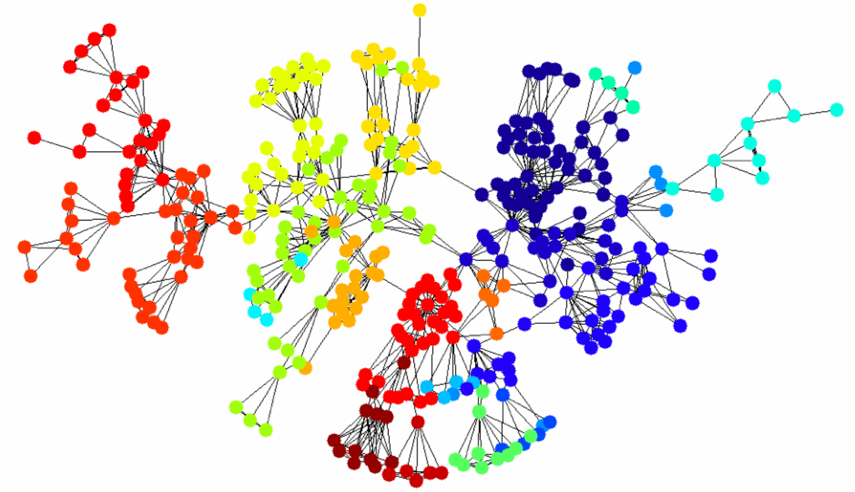
Community: It is formed by individuals such that those **within** a group interact with each other **more frequently** than with those **outside** the group

- a.k.a. group, cluster, cohesive subgroup, module in different contexts

Community detection: discovering groups in a network where individuals' group **memberships are not explicitly given**

Why **communities in social media?**

- Human beings are **social**
- **Easy-to-use social media** allows people to **extend** their **social life** in unprecedented ways
- Difficult to meet friends in the physical world, but **much easier** to find friend **online** with similar interests
- **Interactions between nodes** can help determine communities



Example: community network in social media

Communities in Social Media

Communities in Social Media

Two types of groups in social media

- **Explicit** Groups: formed by user **subscriptions**
- **Implicit** Groups: implicitly formed by **social interactions**

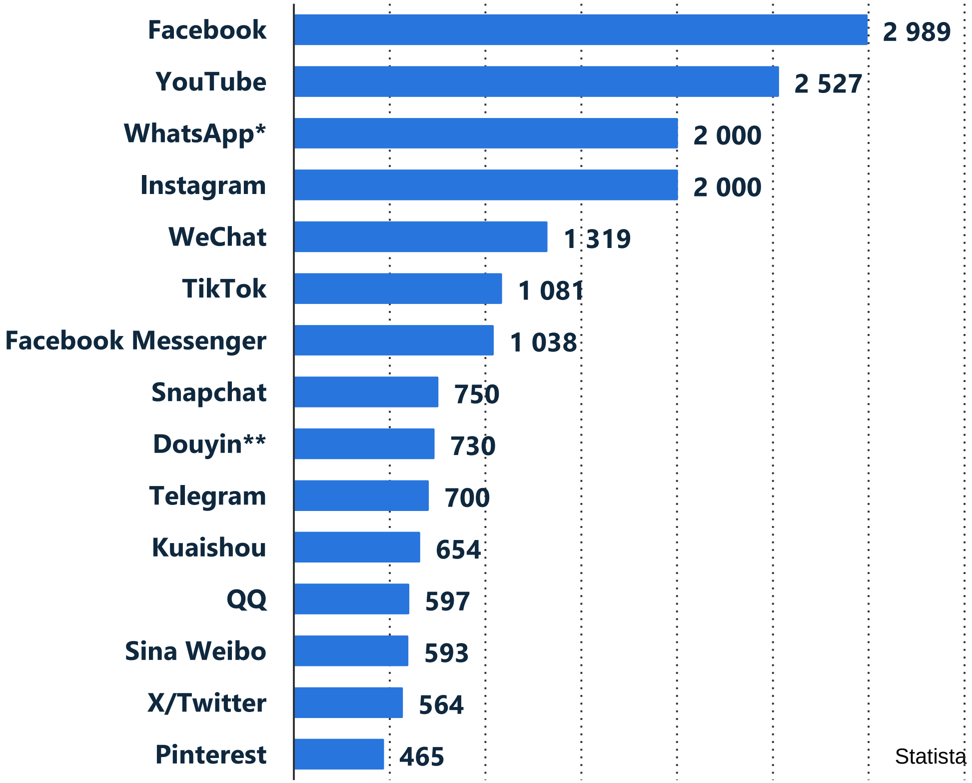
Some social media sites allow people to join groups,
- extract groups based on network topology?

- Not all sites provide a community platform
- Not all people want to make effort to join groups
- Groups can **change dynamically**

Network interaction provides rich information about the relationship between users

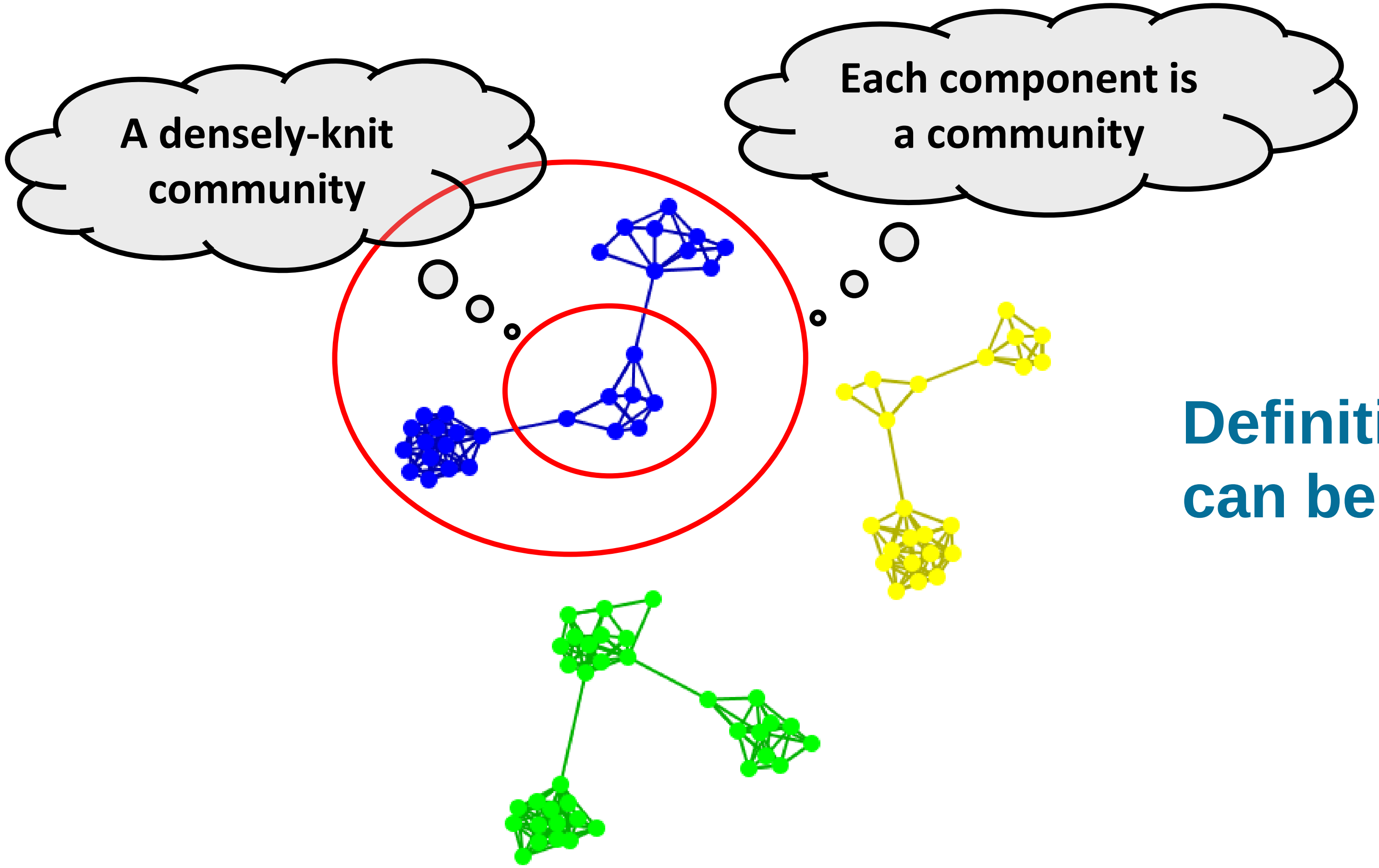
- Can complement other kinds of information
- Help network visualization and navigation
- Provide basic information for other tasks

Example of explicit groups:
Most popular social networks worldwide, July 2023
(number of active users, in millions)



Community Definition

Subjectivity of Community Definition



Definition of a community can be subjective

Community Detection Methods

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

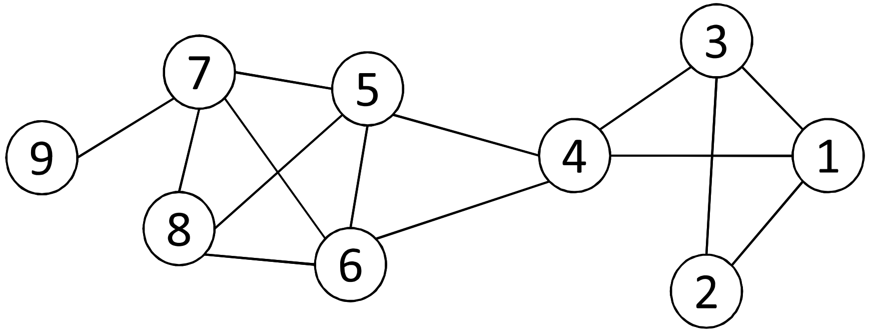
Community Metrics

Measures used to calibrate the small world effect:

Diameter

Measures used to calibrate the small world effect

- **Diameter**: the (maximum) longest shortest path in a network
- **Average shortest path length**

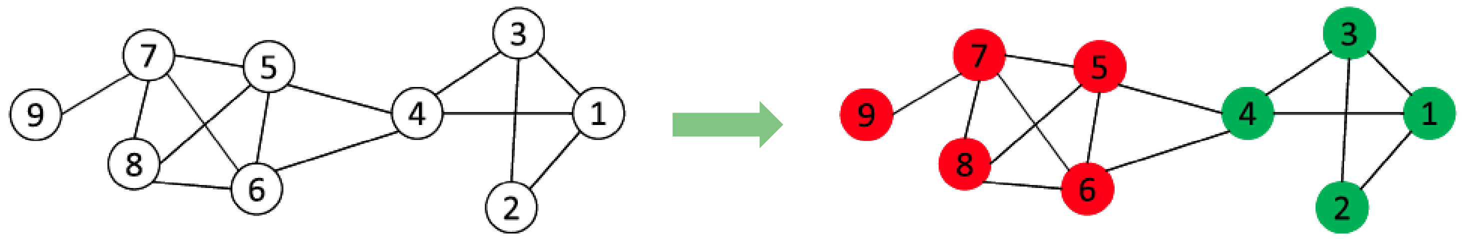


- The shortest path between two nodes is (also) called **geodesic**.
- The number of **hops** in the geodesic is the **geodesic distance**.

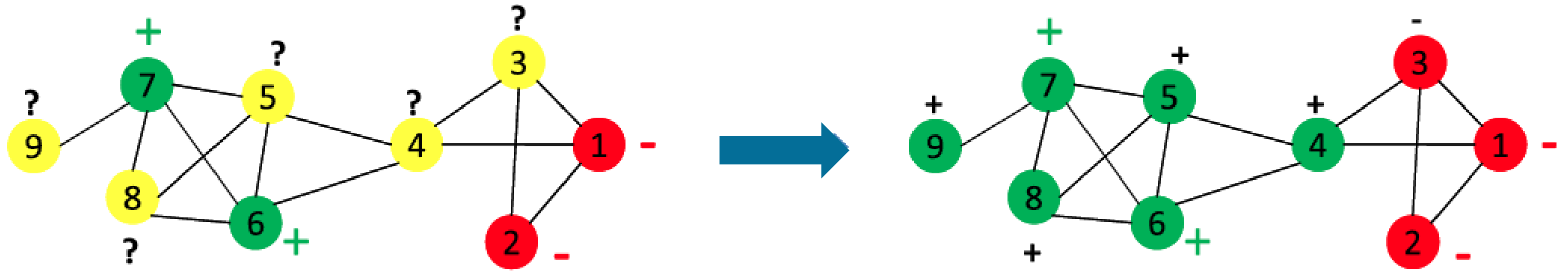
- The **geodesic distance** between node 1 and node 9 is **4**.
- The **diameter** of the network is **5**, corresponding to the geodesic distance between nodes 2 and 9.

Community Dynamics

A **community** are people in a **group interacting with** each other **more frequently than** those **outside** the group



Link prediction



Network-Based Classification

Community Analysis

Density of connections

- Friends of a friend are likely to be friends as well
  - density of connections among one's friends

$$C_i = \begin{cases} \frac{k_i}{d_i \times (d_i - 1)/2} & d_i > 1 \\ 0 & d_i = 0 \; or \; 1 \end{cases}$$

d … number of neighbours

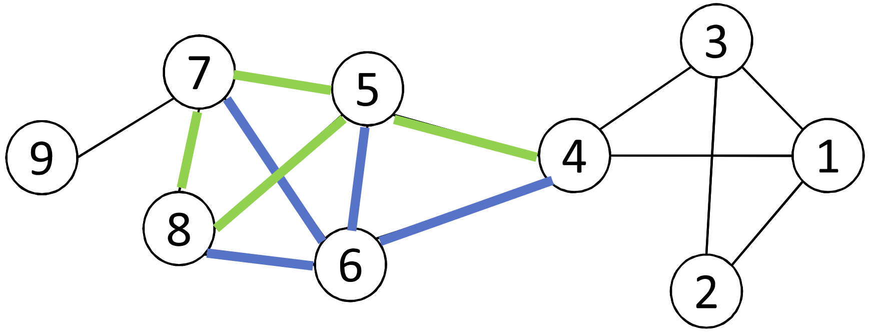
k … connections among neighboured friends

Example:

d_6=4, N_6= {4, 5, 7,8}

k_6=4 as e(4,5), e(5,7), e(5,8), e(7,8)

C_6 = 4/(4*3/2) = 2/3



Average clustering coefficient

$C = (C_1 + C_2 + \ldots + C_n)/n$

C = 0.61 for the given network

Centrality

Centrality in Graphs assign **numbers** or **rankings** to nodes within a graph **corresponding to their network position**.

Applications

- **identifying** the **most influential** person(s) in a social network
- **key infrastructure nodes** in the Internet or urban networks
- **super-spreaders** of disease in pandemics
- **cell networks** in brains

Degree Cetrality

- **number of links** incident to a node

Betweenness Cetrality (of a node)

- quantifies the **number of times** a node acts as a **connector/bridge along the shortest** path between two other nodes (the **number** of **shortest paths** that **pass** one node)

Eigenvector Cetrality (of a node)

- assigns **relative scores** to all nodes in the network based on the concept that **connections to high-scoring nodes contribute more** to the score of the node in question than equal connections to low-scoring nodes

Cross-Clique Cetrality

- determines the **connectivity** of a node to **different cliques**
- a node with **high cross-clique connectivity** facilitates the **propagation** of information (or diseases) in a graph
- (Cliques are subgraphs in which every node is connected to every other node in the clique).

Community Analysis

Degree of Centrality

The **importance** of a node is determined by the **number** of **nodes adjacent** to it

- **The larger the degree, the more import the node is**
- Only a small number of nodes have high degrees in many real-life networks

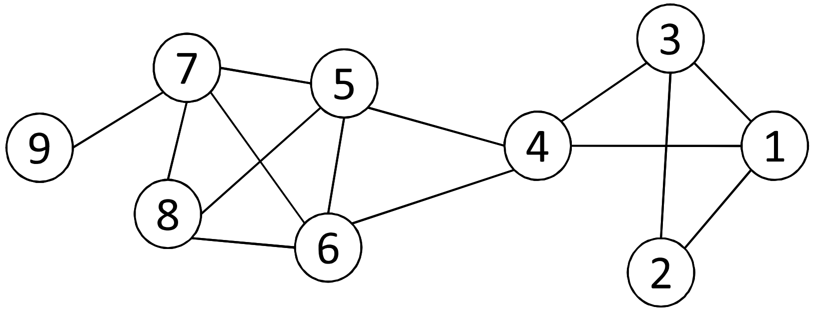
Degree Centrality

$$C_D(v_i) = d_i = \sum_j A_{ij}$$

Normalized Degree Centrality

$$C'_D(v_i) = d_i / (n-1)$$

- For node 1, degree centrality is 3;
- Normalized degree centrality is 3/(9-1)=3/8



Institute of Pervasive Computing

Community Analysis

Closeness Centrality

- "**Central**" nodes are **important**, as they can reach the whole network more quickly than non-central nodes

- Importance measured by how close a node is to other nodes

- **Average Distance**

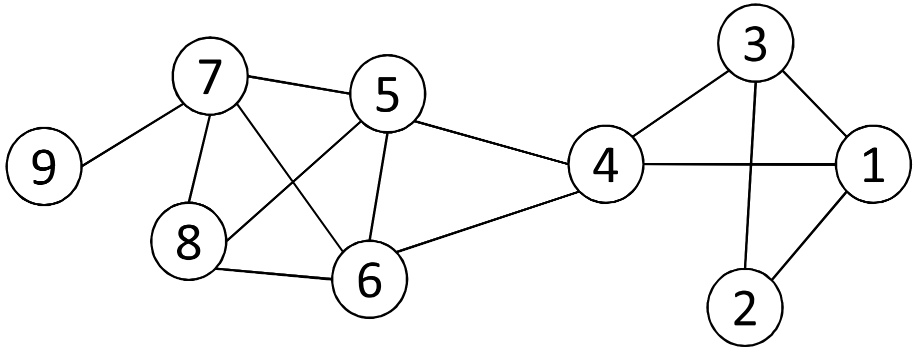
$$D_{avg}(v_i) = \frac{1}{n-1} \sum_{j \neq i}^{n} g(v_i, v_j)$$

- **Closeness Centrality**

$$C_C(v_i) = \left[\frac{1}{n-1} \sum_{j \neq i}^{n} g(v_i, v_j) \right]^{-1} = \frac{n-1}{\sum_{j \neq i}^{n} g(v_i, v_j)}$$

Community Analysis

Closeness Centrality: Example



$$C_c(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = \frac{8}{17} = 0.47$$

$$C_c(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = \frac{8}{13} = 0.62$$

Table 2.1: Pairwise geodesic distance

Node	1	2	3	4	5	6	7	8	9
1	0	1	1	1	2	2	3	3	4
2	1	0	1	2	3	3	4	4	5
3	1	1	0	1	2	2	3	3	4
4	1	2	1	0	1	1	2	2	3
5	2	3	2	1	0	1	1	1	2
6	2	3	2	1	1	0	1	1	2
7	3	4	3	2	1	1	0	1	1
8	3	4	3	2	1	1	1	0	2
9	4	5	4	3	2	2	1	2	0

Node 4 is more central than node 3

Community Analysis

Betweenness Centrality

- Node betweenness counts the **number** of **shortest paths that pass one node**

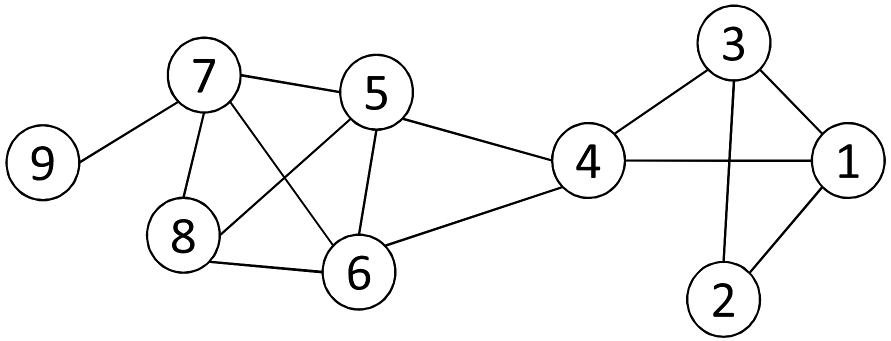
- Nodes with **high betweenness** are **important** in communication and information **diffusion**

- **Betweenness Centrality** $$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

σ_{st} : The number of shortest paths between *s* and *t*

$\sigma_{st}(v_i)$: The number of shortest paths between *s* and *t* that pass v_i

Community Analysis



$$C_B(4) = 15$$

	$\sigma_{st}(4)/\sigma_{st}$		
	$s = 1$	$s = 2$	$s = 3$
$t = 5$	1/1	2/2	1/1
$t = 6$	1/1	2/2	1/1
$t = 7$	2/2	4/4	2/2
$t = 8$	2/2	4/4	2/2
$t = 9$	2/2	4/4	2/2

What's the betweenness centrality for node 5? (= 6)

σ_{st} : The number of shortest paths between *s* and *t*

$\sigma_{st}(v_i)$: The number of shortest paths between *s* and *t* that pass v_i

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

Community Analysis

Eigenvalue (or Eigenvector) Centrality

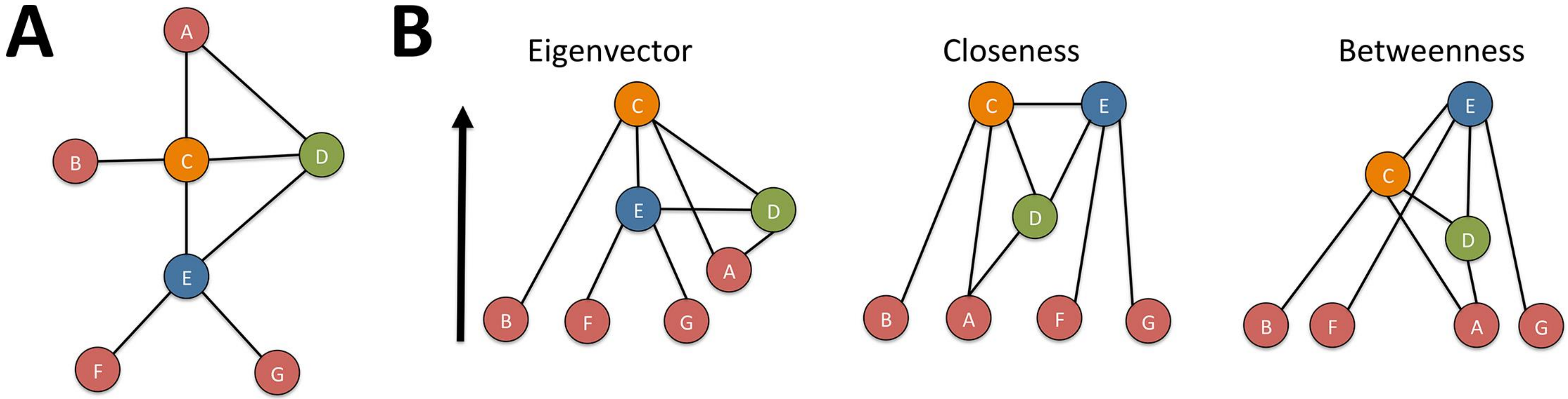
- **One's importance** is **determined** by his **friends**'
- If one has **many important friends**, he should be important as well.

$$C_E(v_i) \propto \sum_{v_j \in N_i} A_{ij} C_E(v_j)$$

$$\mathbf{x} \propto A\mathbf{x} \quad \Rightarrow \quad A\mathbf{x} = \lambda\mathbf{x}.$$

- The centrality corresponds to the top **eigenvector** of the **adjacency matrix A**.
- A variant of this eigenvector centrality is the PageRank **score.**

Community Analysis :: Importance (Closeness)



(A) Example of a hypothetical social network illustrating the individual level metrics. Letters label individuals in the network.

(B) Network is **restructured** in a **hierarchy** such that the **node with the highest relevant centrality measure** is on **top** following the arrow). For example, node **C** has the highest Eigenvector centrality, but node **E** had the highest betweenness centrality.

Community Analysis :: Ties and Influence

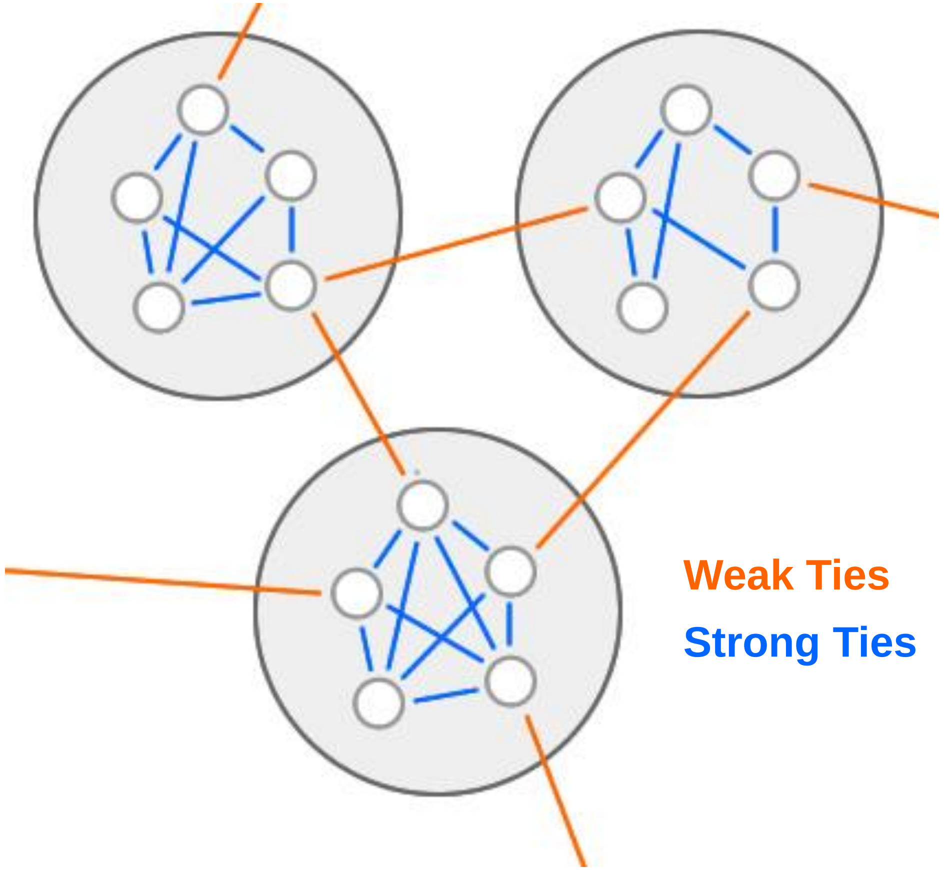
Weak and Strong Ties

- In practice, **connections** are **not** of the **same strength**

- **Interpersonal** social networks are composed of **strong ties** (close friends) and **weak ties** (acquaintances).

- Strong ties and weak ties play different roles for community formation and information diffusion
- **strong ties** are **transitive** with high probability
- **weak ties** are **not transitive** (or with low probability)

- ***The Strength of Weak Ties*** (*Granovetter, 1973*)
    - Occasional encounters with distant acquaintances can provide important information about new opportunities for job search



M. Granovetter. The Strength of Weak Ties. The American Journal of Sociology, 78(6):1360–1380, 1973.

Community Analysis

Connections in Social Media

Social Media allows users to **connect** to each other **more easily** than ever

- One user might have thousands of friends online
- Who are the most important ones among your Facebook friends?

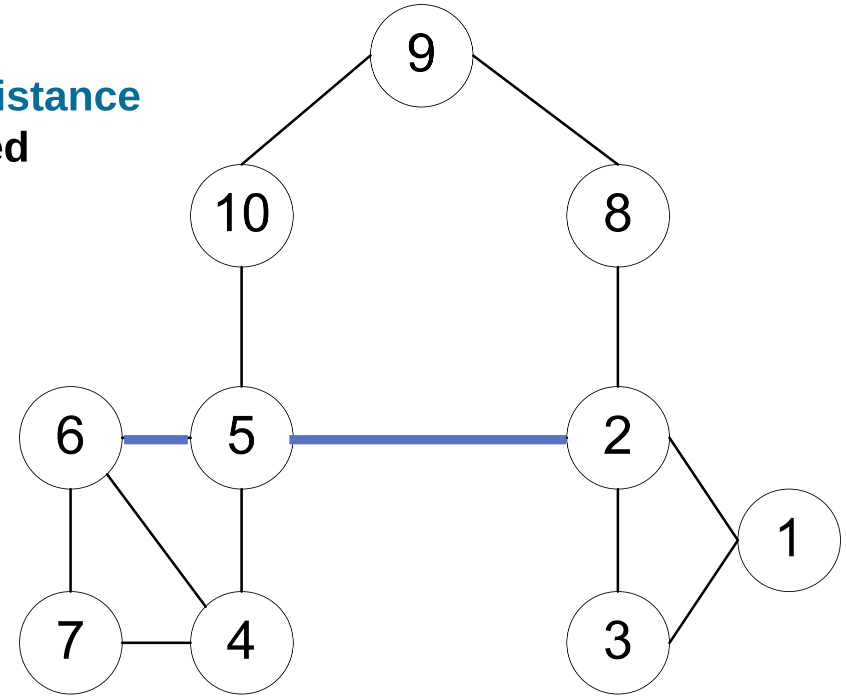
Imperative to **estimate** the **strengths** of **ties** for advanced analysis

- Analyze **network topology**
- Learn from **User Profiles and Attributes**
- Learn from **User Activities**

Community Analysis :: Network Topology

“shortcut” Bridge

- Bridges are **rare in real-life networks**
- Alternatively, one can relax the definition by checking if **the distance** between two terminal nodes **increases** if the edge is **removed**
- The **larger the distance**, the **weaker the tie** is



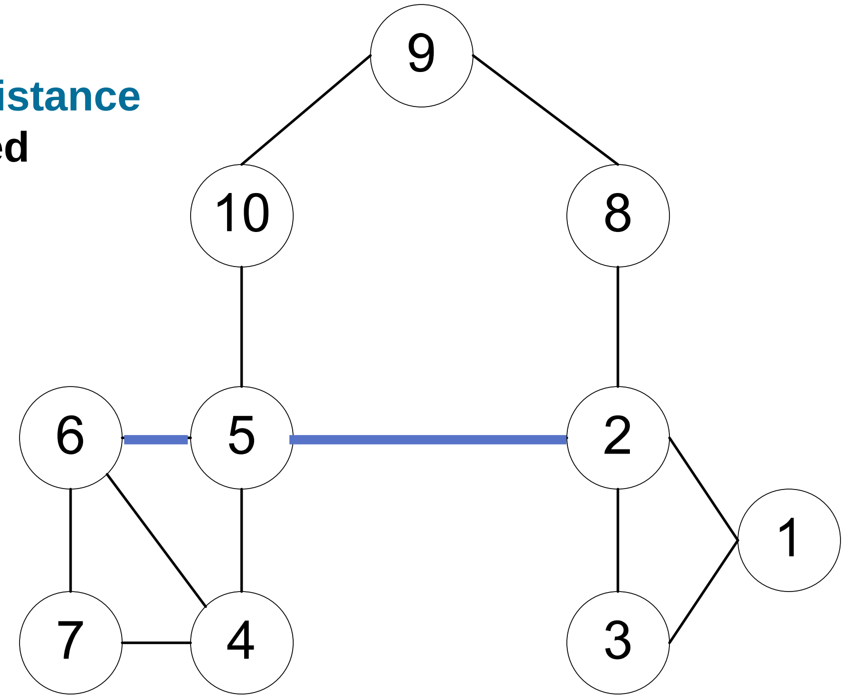
- d(2,5) = **4** if e(2,5) is removed
- d(5,6) = **2** if e(5,6) is removed

- e(5,6) is a **stronger** tie than e(2,5)

Community Analysis :: Network Topology

"shortcut" Bridge

- Bridges are **rare in real-life networks**
- Alternatively, one can relax the definition by checking if **the distance** between two terminal nodes **increases** if the edge is **removed**
- The **larger the distance**, the **weaker the tie** is



- d(2,5) = **4** if e(2,5) is removed
- d(5,6) = **2** if e(5,6) is removed

- e(5,6) is a **stronger** tie than e(2,5)

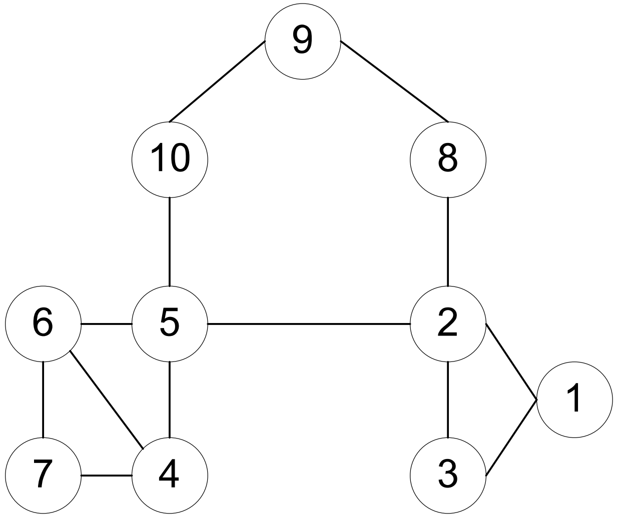
Community Analysis :: Network Topology

Neighbourhood Overlap

Tie strength can be measured based on **neighborhood overlap**; the **larger** the **overlap**, the **stronger** the **tie** is.

$$overlap(v_i, v_j) = \frac{\text{number of shared friends of both } v_i \text{ and } v_j}{\text{number of friends who are adjacent to at least } v_i \text{ or } v_j} = \frac{|N_i \cap N_j|}{|N_i \cup N_j| - 2}.$$

-2 in the denominator is to exclude v_i and v_j



$$overlap(2,5) = 0,$$
$$overlap(5,6) = \frac{|\{4\}|}{|\{2,4,5,6,7,10\}| - 2} = 1/4$$

Community Analysis :: User Activities

Learning from User Activities

- One might learn **how one influences his friends** if the user **activity log** is accessible
- Depending on the adopted influence model

  - **Independent cascading model**
  - **Linear threshold model**

Maximizing the likelihood of user activity given an influence model

Community Analysis :: Influence Modeling

Influence Modeling

Influence modeling is one of the fundamental questions in order to understand the information diffusion, *spread of new ideas*, *and* ***word-of-mouth*** *(**viral**)* **marketing**

Well known methods:

- **Linear threshold model (LTM)**
- **Independent cascade model (ICM)**

Common properties of influence modeling methods

- A social network is represented by a ***directed graph****, with each actor being one node;*
- Each node is started as **active** or **inactive**;
- A node, **once activated**, will **activate** his **neighboring** nodes;
- Once a node is activated, this node **cannot be deactivated**.

Community Analysis :: Influence Modeling

Linear Threshold Model (LTM)

An actor would take an action if the number of his friends who have taken the **action exceeds** (reaches) a certain **threshold**

- Each node ***v chooses*** *a* ***threshold*** $\boldsymbol{\Theta_v}$ ***randomly*** *from a uniform* distribution in an interval between 0 and 1.
- In each **discrete step**, all nodes that were **active** in the previous step **remain active**
- The nodes satisfying the following **condition** will be **activated**

$$\sum_{w \in N_v,\, w \text{ is active}} b_{w,v} \geq \theta_v$$

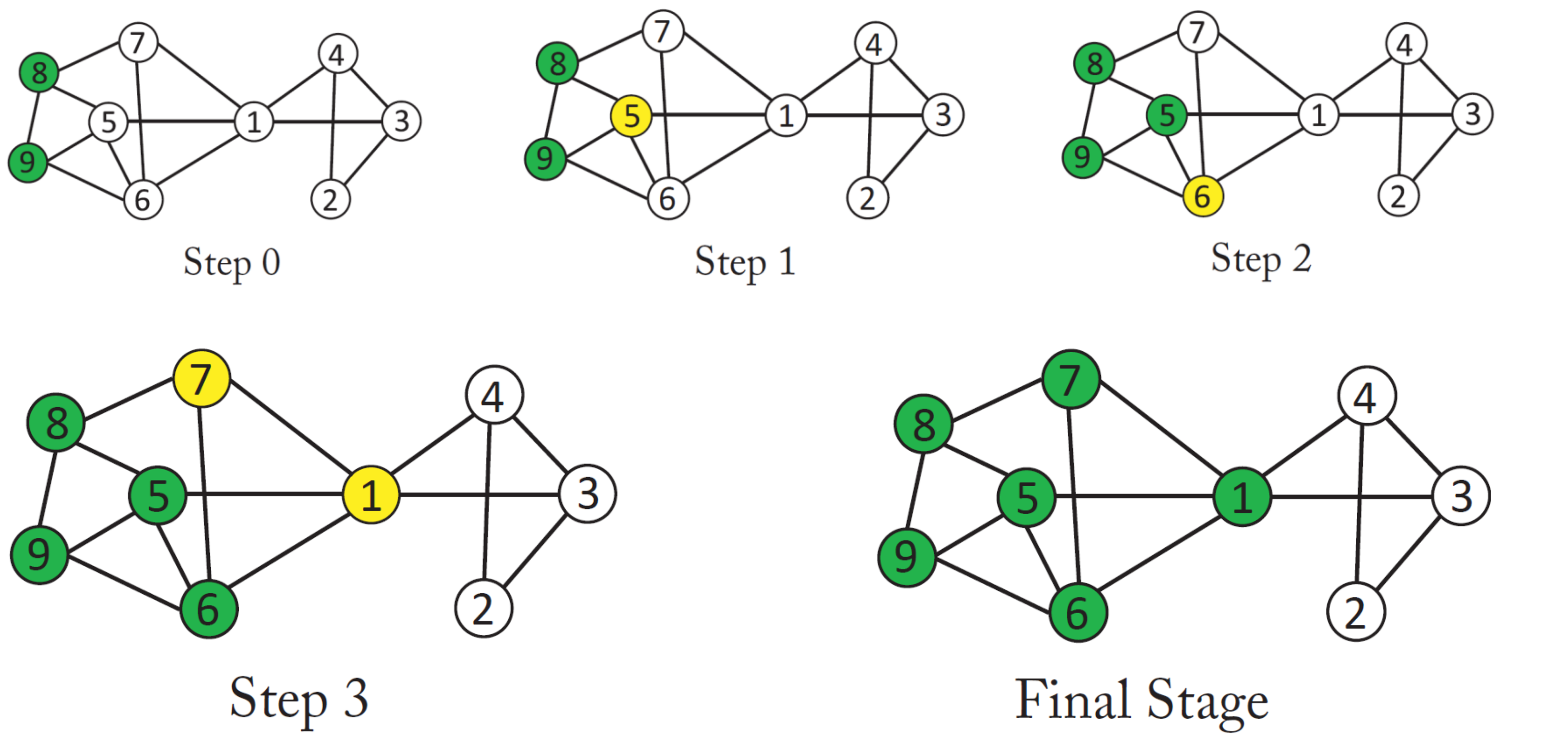
Community Analysis :: Influence Modeling

activation “receiver driven“

assume every v_i chooses 0.5

i.e. 50% of neighbours have to be active

Linear Threshold Model – Diffusion Process (Threshold = 50%)



Community Analysis :: Influence Modeling

Independent Cascade Model (ICM)

The independent cascade model focuses on the **sender's** rather than the receiver's **view**

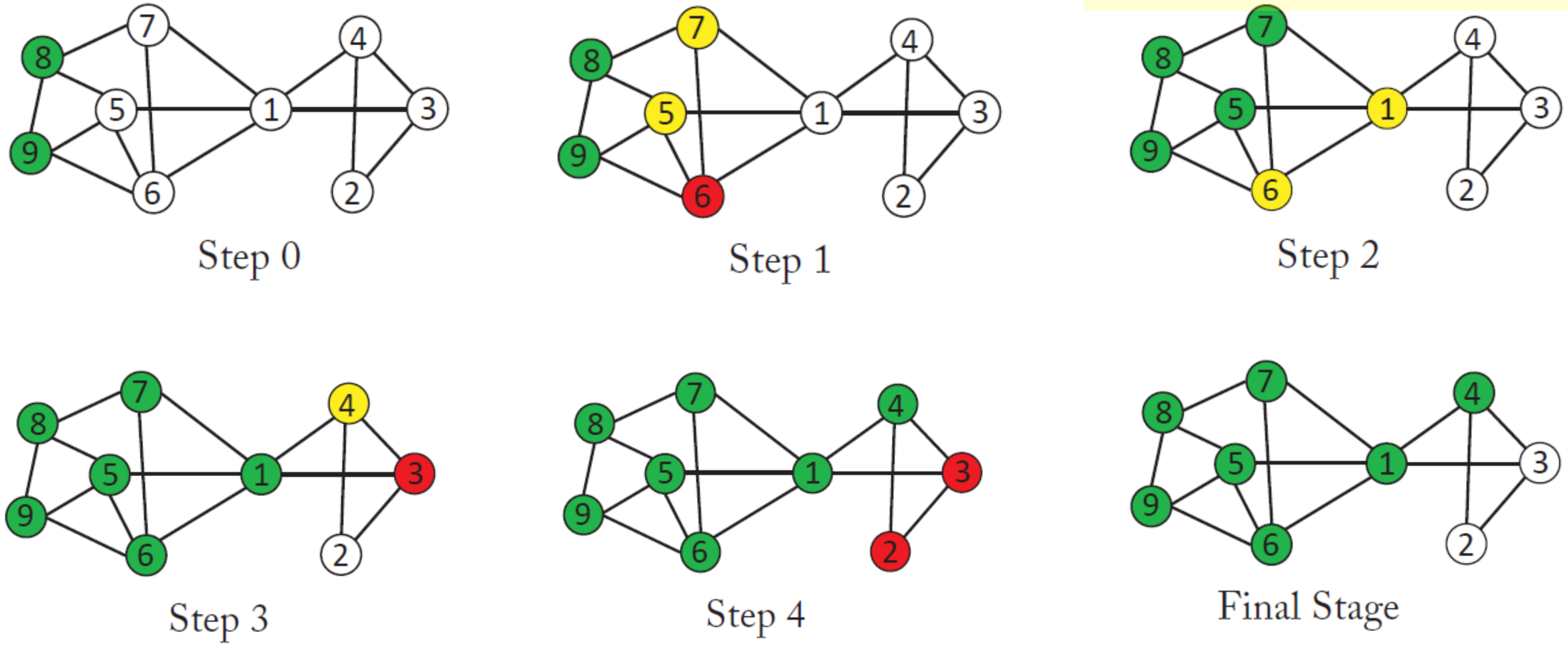
- A node ***w****, once activated* at step *t* , *has* ***one chance*** *to* ***activate each*** *of its* ***neighbors*** ***randomly***
  - For a neighboring node (say, *v), the* ***activation*** **succeeds** with **probability** $p_{w,v}$ (e.g. p = 0.5)
- If the activation succeeds, then ***v will become active*** *at* ***step t + 1***
- In the **subsequent** rounds, ***w*** *will* ***not attempt*** *to* ***activate v*** *anymore.*
- *The* ***diffusion process****,* starts with an **initial activated** set of nodes,
  then continues **until no** further **activation** is **possible**

Community Analysis :: Influence Modeling

activation “sender driven“

only active odes can set friends active …
with a certain probability
yellow … sucessful
red … not successful

Independent Cascade Model (ICM)



Community Analysis :: Influence Modeling

Influence Maximization

- Given a **network** and a **parameter *k:***
  *which **k nodes** should be selected to be **in the*** **activation** set ***B***
  *In order to **maximize** the **influence** in terms of active nodes at the* end?

- Let ***σ(B)*** *denote the **expected number** of nodes that* **can be *influenced*** *by* ***B***,
  the optimization problem can be formulated as follows:

$$\max_{B \subseteq V} \sigma(B) \ s.t. \ |B| \leq k$$

Community Analysis :: Influence Modeling

Influence Maximization – A greedy approach

- Maximizing the influence is an **NP-hard problem** but it is proven that the **greedy** approach gives a solution that is **63 % of the optimal.**

A greedy approach

- Start with B = Ø
- **Evaluate** σ(v) for **each node**, and **pick** the **node** with **maximum** σ as the **first node** v_1 to form $B = \{v_1\}$
- **Select** a node which will **increase σ(B) most** if the node is included in B

Essentially, we greedily find a node *v ∈ V \B such that*

$$v = \arg \max_{v \in V \setminus B} \sigma(B \cup \{v\})$$

Community Analysis :: Distinguishing Between Influence & Correlation

Correlation

It has been widely observed that **user attributes** and **behaviors** tend to **correlate** within their social networks

- Suppose we have a **binary attribute** with each node (say, whether or not being smoker)
- If the attribute is correlated with the network, we **expect actors sharing** the **same attribute value** to be **positively correlated** with **social connections**
- That is, **smokers** are **more likely** to **interact** with other **smokers**, and non-smokers with non-smokers

Community Analysis :: Distinguishing Between Influence & Correlation

Test For Correlation

If the **fraction** of **edges linking nodes** with **different attribute values** are **significantly less** than the expected probability, then there is **evidence** of **correlation**

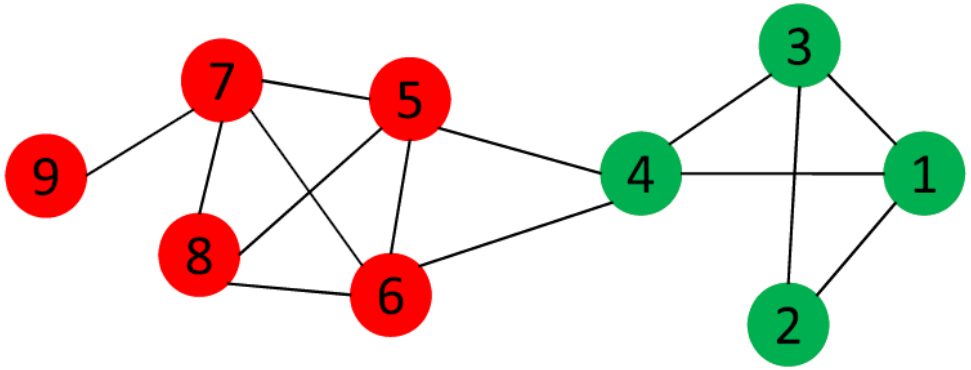
Example if connections are independent of the smoking behavior:

p fraction are smokers (1-p non-smoker)

- one edge is expected to **connect two smokers** with probability: $p \times p$
- ***two non-smokers*** *with* probability: $(1 - p) \times (1 - p)$
- *a* ***smoker*** *and a* ***non-smoker***: $2\,p \times (1-p)$

Community Analysis ::
Distinguishing Between Influence & Correlation

Test For Correlation



- **Red** nodes denote **non-smoke**rs, and **green** ones are **smokers**. If there is no correlation, then the probability of one edge **connecting** a **smoker** and a **non-smoker** is **2 × 4/9 × 5/9 = 49**%.
- In this example the fraction is **2/14** = 14% < 49%, so this network demonstrates some degree of **correlation** with respect to the smoking behavior.
- A more formal way is to conduct a **χ2 test** for independence of social connections and attributes [1]

[1] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In Proceedings of the 19th international conference on World wide web, WWW ’10, pages 601–610, New York, NY, USA, 2010. ACM.

Community Analysis :: Clustering

Node Centric clustering

Nodes satisfy different properties:

Complete Mutuality

- Cliques

Reachability of members

- k-clique, k-clan, k-club

Nodal degrees

- k-plex, k-core

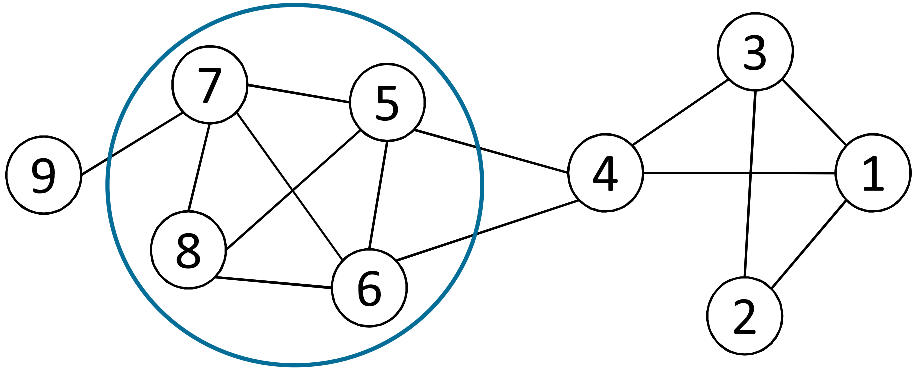
Relative frequency of Within-Outside Ties

- LS sets, Lambda sets
- Commonly used in traditional social network analysis
- Here, we discuss some representative ones

Community Analysis :: Clustering

Clique

A **maximum complete subgraph** in which all nodes are adjacent to each other



Nodes 5, 6, 7 and 8 form a clique

NP-hard to find the **maximum clique** in a network

Straightforward implementation to find cliques is very expensive in time complexity

Community Analysis :: Clustering

Finding the Maximum Clique

In a **clique of size k**, each node maintains **degree >= k-1**

Nodes with degree < k-1 will **not be included** in the maximum clique

Recursively apply the following **pruning** procedure

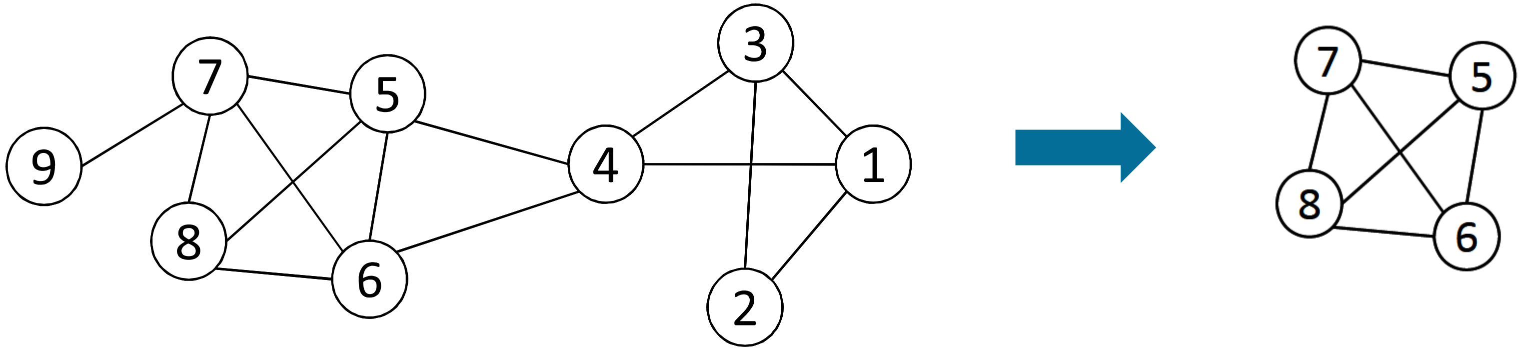
- **Sample** a **sub-network** from the given network, and **find** a **clique** in the **sub-network**, say, by a greedy approach
- Suppose the clique above is **size k**, in order to find out a ***larger*** **clique**, all nodes with **degree <= k-1** should be **removed**.

Repeat until the network is **small enough**

Many nodes will be **pruned** as social media networks follow a **power law** distribution for node degrees

Community Analysis :: Clustering

Maximum Clique Example



Suppose we sample a sub-network with nodes {1-5} and find a clique {1, 2, 3} of size 3

In order to find a clique >3, remove all nodes with degree <=3-1=2 (**prune recursively**)

- Remove nodes **2** and **9**
- Remove nodes **1** and **3**
- Remove node **4**

Community Analysis :: Clustering

Clique Percolation Method (CPM)

Clique is a very strict definition, unstable

Normally **use cliques** as a **core** or a **seed** to find **larger communities**

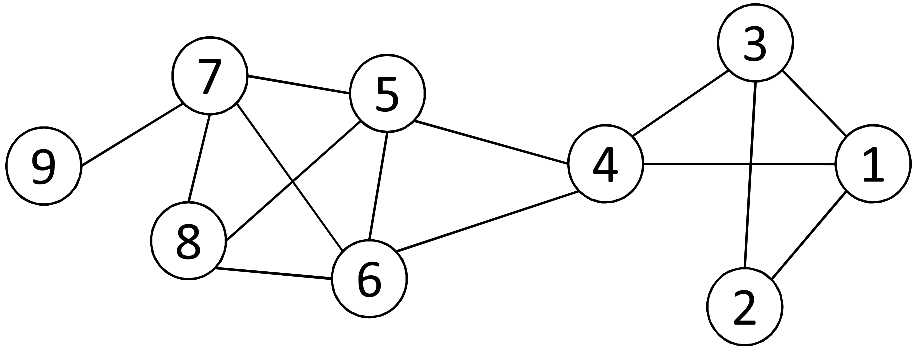
CPM is such a method to **find overlapping communities**

Input
- A parameter k, and a network

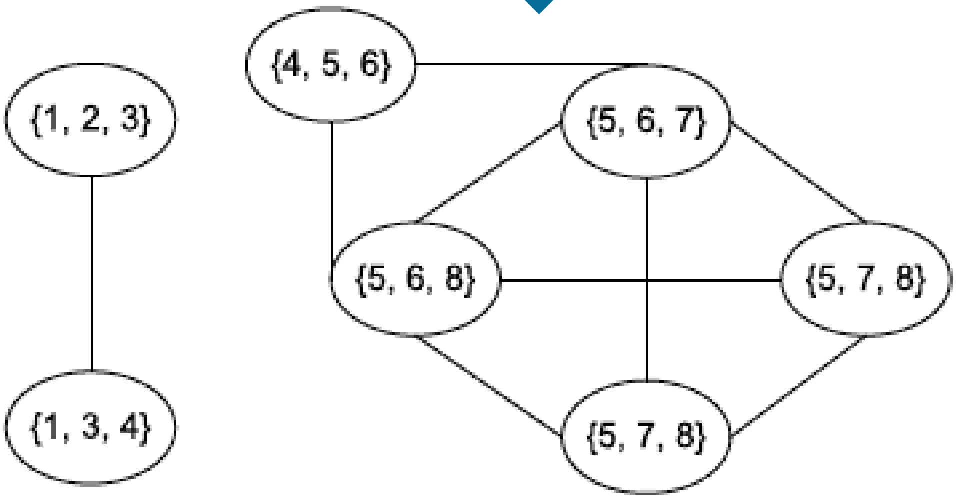
Procedure
- Find out **all cliques** of **size k** in a given network
- Construct a **clique graph**. Two cliques are **adjacent** if they **share k-1 nodes**
- Each **connected components** in the clique graph form a **community**

Community Analysis :: Clustering

CPM Example



Cliques of size **3**:
{1, 2, 3}, {1, 3, 4}, {4, 5, 6},
{5, 6, 7}, {5, 6, 8}, {5, 7, 8}, {6, 7, 8}

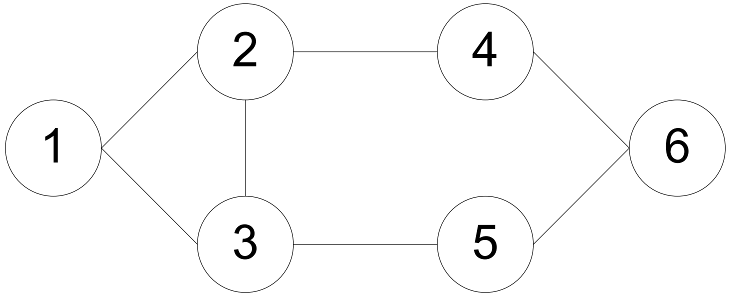


Communities:
{1, 2, 3, 4}
{4, 5, 6, 7, 8}

Community Analysis :: Clustering

Reachability: k-clique, k-club

Any node in a group should be **reachable in k hops**

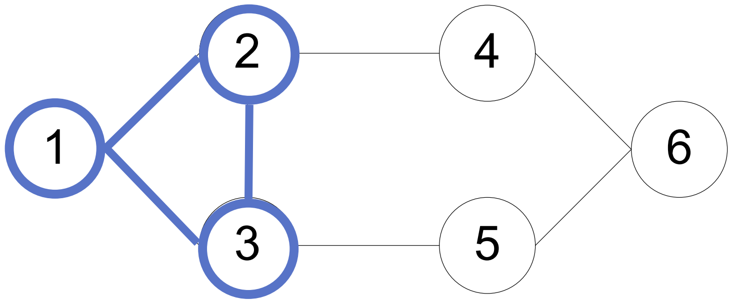


Cliques: {1, 2, 3}

Community Analysis :: Clustering

Reachability: k-clique, k-club

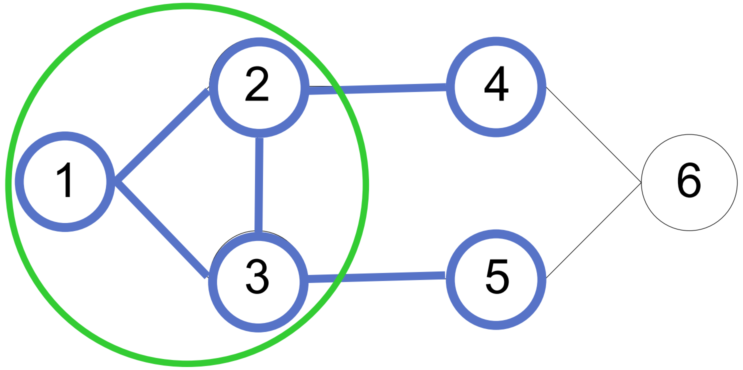
Any node in a group should be **reachable in k hops**



Cliques: {1, 2, 3}

Community Analysis :: Clustering

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**



Cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}

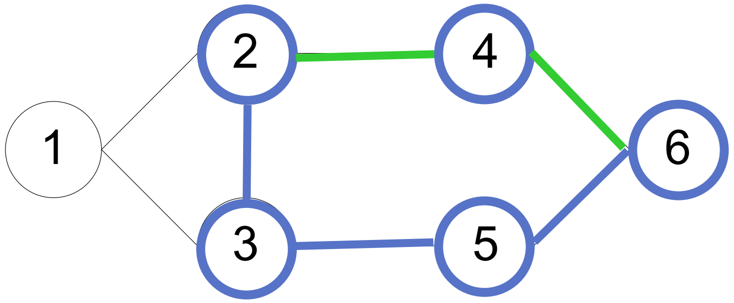
A **k-clique** might have **diameter larger than k** in the subgraph

Commonly used in traditional SNA

Often involves combinatorial optimization

Community Analysis :: Clustering

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**



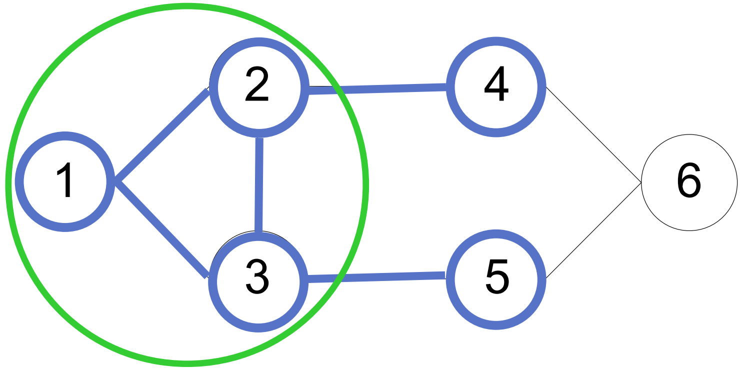
Cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5}, **{2, 3, 4, 5, 6}**

A **k-clique** might have **diameter larger than k** in the subgraph

Commonly used in traditional SNA

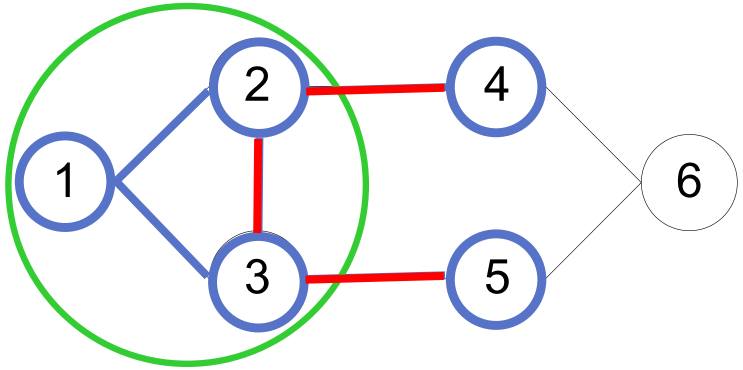
Often involves combinatorial optimization

Community Analysis :: Clustering



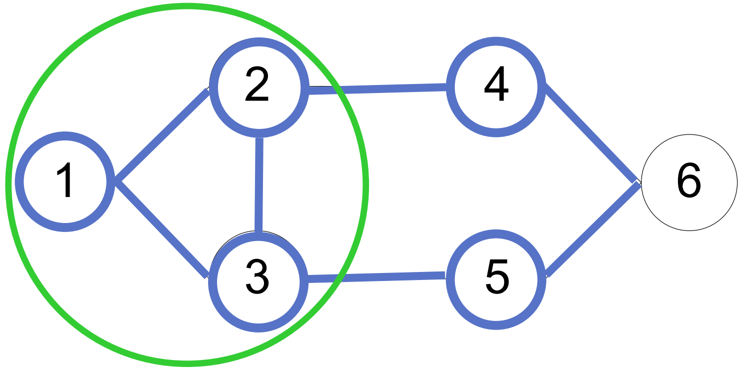
2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}

Community Analysis :: Clustering



2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}

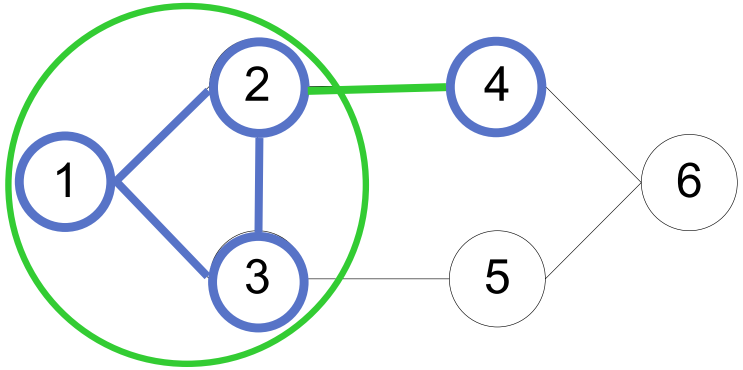
Community Analysis :: Clustering



2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}

Community Analysis :: Clustering

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**
k-club (also: **k-clan**): a **substructure** of **diameter <= k**



Cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}
2-clubs: {1,2,3,4}, {1, 2, 3, 5}, {2, 3, 4, 5, 6}

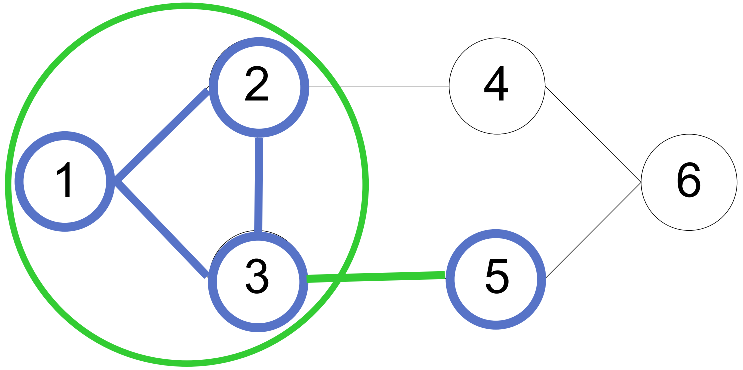
A **k-clique** might have **diameter larger than k** in the subgraph

Commonly used in traditional SNA

Often involves combinatorial optimization

Community Analysis :: Clustering

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**
k-club (also: **k-clan**): a **substructure** of **diameter <= k**



Cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}
2-clubs: {1,2,3,4}, **{1, 2, 3, 5},** {2, 3, 4, 5, 6}

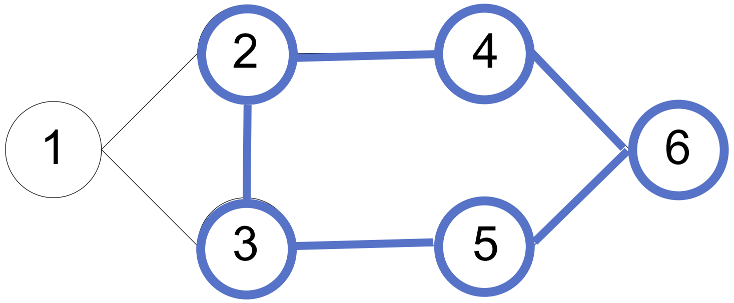
A **k-clique** might have **diameter larger than k** in the subgraph

Commonly used in traditional SNA

Often involves combinatorial optimization

Community Analysis :: Clustering

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**
k-club (also: **k-clan**): a **substructure** of **diameter <= k**



Cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}
2-clubs: {1,2,3,4}, {1, 2, 3, 5}, **{2, 3, 4, 5, 6}**

A **k-clique** might have **diameter larger than k** in the subgraph

Commonly used in traditional SNA

Often involves combinatorial optimization

Community Analysis :: Clustering

Group-Centric clustering: Density Based Groups

Quasi-cliques are **dense incomplete subgraphs of a graph that generalize the notion of cliques**.

The **group-centric criterion** requires the **whole group** to satisfy a certain condition

- E.g., the group density >= a given threshold

A **subgraph** $G_s(V_s, E_s)$ is a $\gamma - dense$ quasi-clique if

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

number of **edges in the quasi-clique**

number of **edges in a (full) clique of the same size**

A **similar strategy** to that of cliques can be used

- Sample a subgraph, and find a maximal $\gamma - dense$ quasi-clique (say, of size k)
- Remove nodes with degree $< k\gamma$

Community Analysis :: Clustering

Network-Centric clustering (as opposed to **Node-Centric clustering** so far)

Network-centric criterion needs to consider the **connections within** a **network globally**

Goal: **partition nodes of a network into disjoint sets**

Approaches:

- Clustering based on **vertex similarity**
- **Latent space models**
- **Block model** approximation
- **Spectral clustering**
- **Modularity maximization**

Community Analysis :: Clustering

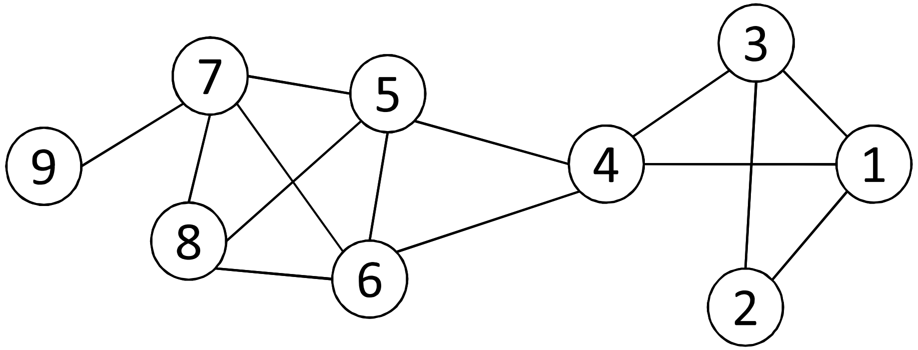
Clustering based on Vertex Similarity

Apply **k-means** or **similarity-based clustering** to nodes

Vertex similarity is defined in terms of the **similarity** of their **neighborhood**

Structural equivalence: two nodes are structurally equivalent if they are **connecting** to the **same set** of **actors**

Nodes 1 and 3 are structurally equivalent;



- Structural equivalence is **too restrictive** for practical use.

Community Analysis :: Clustering

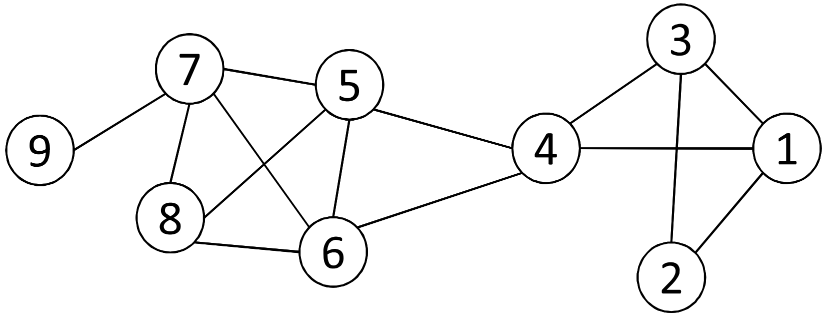
Vertex Similarity

Jaccard Similarity

$$Jaccard(v_i, v_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

Cosine similarity

$$cosine(v_i, v_j) = \frac{|N_i \cap N_j|}{\sqrt{|N_i| \cdot |N_j|}}$$



$$Jaccard(4, 6) = \frac{|\{5\}|}{|\{1, 3, 4, 5, 6, 7, 8\}|} = \frac{1}{7}$$

$$cosine(4, 6) = \frac{1}{\sqrt{4 \cdot 4}} = \frac{1}{4}$$

Community Analysis :: Clustering

Latent Space Models

Map nodes into a **low-dimensional space** such that the **proximity between nodes** based on network connectivity is **preserved** in the **new space**, **then** apply **k-means clustering**

Multi-dimensional scaling (MDS)

- Given a network, construct a proximity matrix P representing the pairwise distance between nodes (e.g., geodesic distance)
- Let $S \in R^{n \times \ell}$ denote the coordinates of nodes in the low-dimensional space

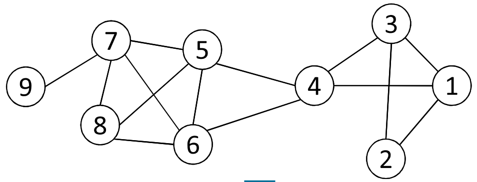
$$SS^T \approx -\frac{1}{2}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(P \circ P)(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = \widetilde{P}$$

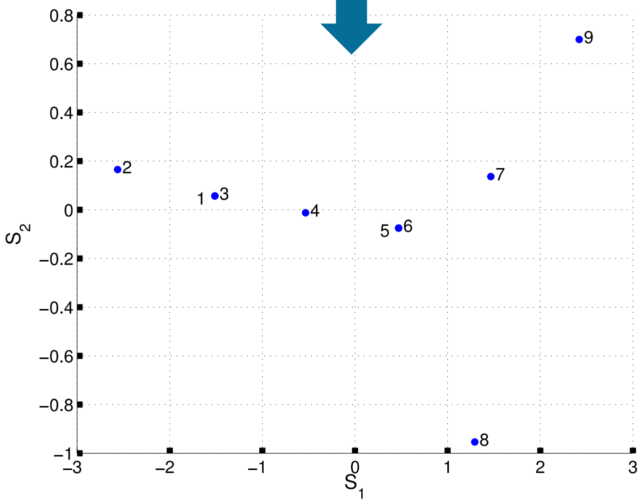
  - Objective function: $\min \|SS^T - \widetilde{P}\|_F^2$
  - Solution: $S = V\Lambda^{\frac{1}{2}}$
  - V is the top ℓ eigenvectors of $\widetilde{P}$, and Λ is a diagonal matrix of top eigenvalues $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_\ell)$

Sosa, J. and Buitrago, L., 2021. **A Review of Latent Space Models for Social Networks**.
Revista Colombiana de Estadística, *44*(1), pp.171-200.

Community Analysis :: Clustering

MDS Example





Two communities:
{1, 2, 3, 4} and {5, 6, 7, 8, 9}

geodesic distance

$$P = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 3 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 & 3 & 4 & 4 & 5 \\ 1 & 1 & 0 & 1 & 2 & 2 & 3 & 3 & 4 \\ 1 & 2 & 1 & 0 & 1 & 1 & 2 & 2 & 3 \\ 2 & 3 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \\ 2 & 3 & 2 & 1 & 1 & 0 & 1 & 1 & 2 \\ 3 & 4 & 3 & 2 & 1 & 1 & 0 & 1 & 1 \\ 3 & 4 & 3 & 2 & 1 & 1 & 1 & 0 & 2 \\ 4 & 5 & 4 & 3 & 2 & 2 & 1 & 2 & 0 \end{bmatrix}$$

$$\widetilde{P} = \begin{bmatrix} 2.46 & 3.96 & 1.96 & 0.85 & -0.65 & -0.65 & -2.21 & -2.04 & -3.65 \\ 3.96 & 6.46 & 3.96 & 1.35 & -1.15 & -1.15 & -3.71 & -3.54 & -6.15 \\ 1.96 & 3.96 & 2.46 & 0.85 & -0.65 & -0.65 & -2.21 & -2.04 & -3.65 \\ 0.85 & 1.35 & 0.85 & 0.23 & -0.27 & -0.27 & -0.82 & -0.65 & -1.27 \\ -0.65 & -1.15 & -0.65 & -0.27 & 0.23 & -0.27 & 0.68 & 0.85 & 1.23 \\ -0.65 & -1.15 & -0.65 & -0.27 & -0.27 & 0.23 & 0.68 & 0.85 & 1.23 \\ -2.21 & -3.71 & -2.21 & -0.82 & 0.68 & 0.68 & 2.12 & 1.79 & 3.68 \\ -2.04 & -3.54 & -2.04 & -0.65 & 0.85 & 0.85 & 1.79 & 2.46 & 2.35 \\ -3.65 & -6.15 & -3.65 & -1.27 & 1.23 & 1.23 & 3.68 & 2.35 & 6.23 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.33 & 0.05 \\ -0.55 & 0.14 \\ -0.33 & 0.05 \\ -0.11 & -0.01 \\ 0.10 & -0.06 \\ 0.10 & -0.06 \\ 0.32 & 0.11 \\ 0.28 & -0.79 \\ 0.52 & 0.58 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 21.56 & 0 \\ 0 & 1.46 \end{bmatrix}, \quad S = V\Lambda^{1/2} = \begin{bmatrix} -1.51 & 0.06 \\ -2.56 & 0.17 \\ -1.51 & 0.06 \\ -0.53 & -0.01 \\ 0.47 & -0.08 \\ 0.47 & -0.08 \\ 1.47 & 0.14 \\ 1.29 & -0.95 \\ 2.42 & 0.70 \end{bmatrix}$$

Community Analysis :: Clustering

Block Models

Table 3.1: Adjacency Matrix

-	1	1	1	0	0	0	0	0
1	-	1	0	0	0	0	0	0
1	1	-	1	0	0	0	0	0
1	0	1	-	1	1	0	0	0
0	0	0	1	-	1	1	1	0
0	0	0	1	1	-	1	1	0
0	0	0	0	1	1	-	1	1
0	0	0	0	1	1	1	-	0
0	0	0	0	0	0	1	0	-

$$\min ||A - S\Sigma S^T||_F^2$$

Table 3.2: Ideal Block Structure

1	1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0
0	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1

S is the community indicator matrix

Relax S to be numerical values, then the optimal solution corresponds to the top eigenvectors of A

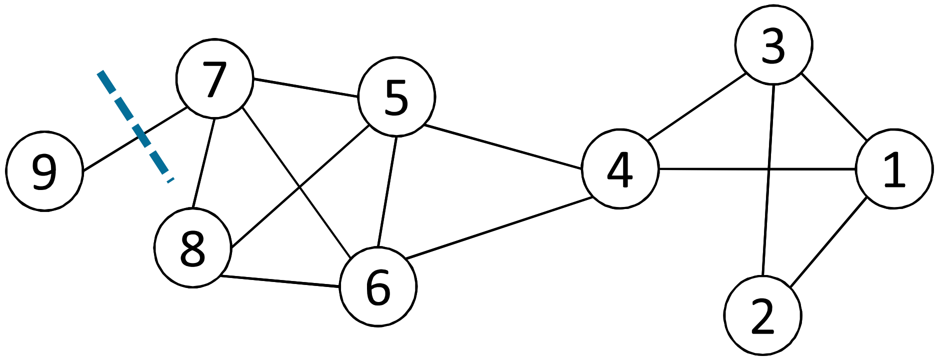
$$S = \begin{bmatrix} 0.20 & -0.52 \\ 0.11 & -0.43 \\ 0.20 & -0.52 \\ 0.38 & -0.30 \\ 0.47 & 0.15 \\ 0.47 & 0.15 \\ 0.41 & 0.28 \\ 0.38 & 0.24 \\ 0.12 & 0.11 \end{bmatrix}, \Sigma = \begin{bmatrix} 3.5 & 0 \\ 0 & 2.4 \end{bmatrix}.$$

Two communities:
{1, 2, 3, 4} and {5, 6, 7, 8, 9}

Community Analysis :: Clustering

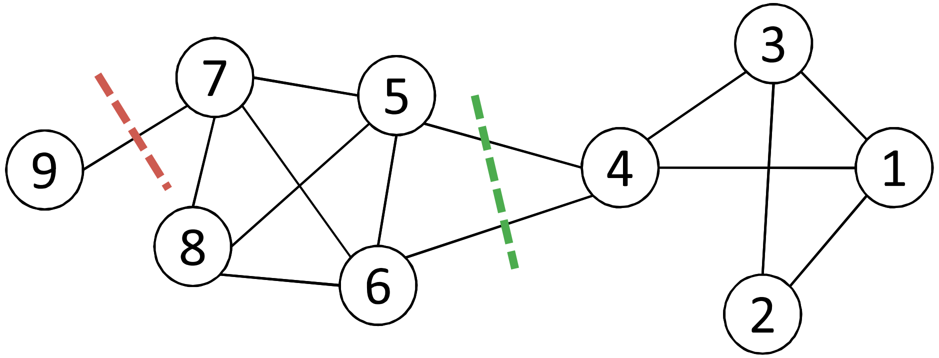
Cut

- **Most** interactions are **within** a **group,** whereas interactions **between groups** are **few**
- community detection (clustering) → **minimum cut problem**
- **Cut**: A partition of vertices of a graph into **two disjoint sets**
- **Minimum cut problem**: find a graph **partition** such that the **number** of **edges between** the two sets is **minimized**



Community Analysis :: Clustering

Ratio Cut & Normalized Cut



- **Minimum cut** often returns an **imbalanced partition**, with one set being a **singleton** (e.g. node 9)
- Change the objective function to consider **community size**

$$\text{Ratio Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i,\bar{C}_i)}{|C_i|},$$

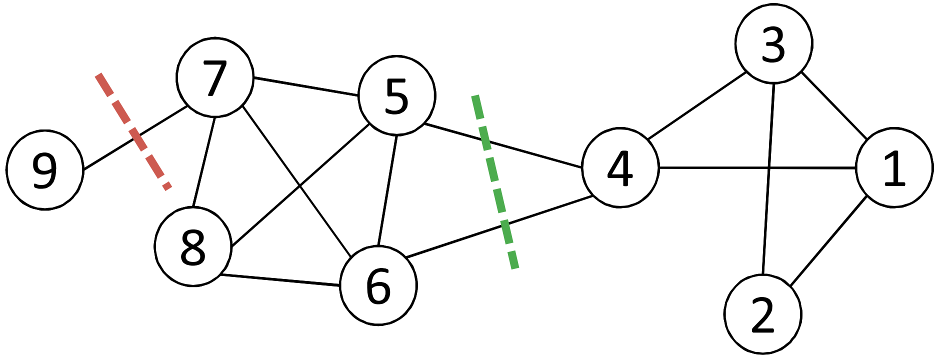
$$\text{Normalized Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i,\bar{C}_i)}{vol(C_i)}$$

$C_{i,}$: a community
cut(A, B): number of edges induced by cut
$|C_i|$: number of nodes in C_i (=community size)
vol(C_i): sum of degrees in C_i

Community Analysis :: Clustering

Ratio Cut & Normalized Cut Example

For partition in red: π_1



$$\text{Ratio Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{8}\right) = 9/16 = 0.56$$

$$\text{Normalized Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{27}\right) = 14/27 = 0.52$$

For partition in green: π_2

$$\text{Ratio Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{4} + \frac{2}{5}\right) = 9/20 = 0.45 < \text{Ratio Cut}(\pi_1)$$

$$\text{Normalized Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{12} + \frac{2}{16}\right) = 7/48 = 0.15 < \text{Normalized Cut}(\pi_1)$$

Both ratio cut and normalized cut prefer a balanced partition

Community Analysis :: Clustering

Spectral Clustering

Both **ratio cut** and **normalized cut** can be reformulated as

$$\min_{S \in \{0,1\}^{n \times k}} Tr(S^T \widetilde{L} S)$$

Where

$$\widetilde{L} = \begin{cases} D - A & \text{graph Laplacian for ratio cut} \\ I - D^{-1/2} A D^{-1/2} & \text{normalized graph Laplacian} \end{cases}$$

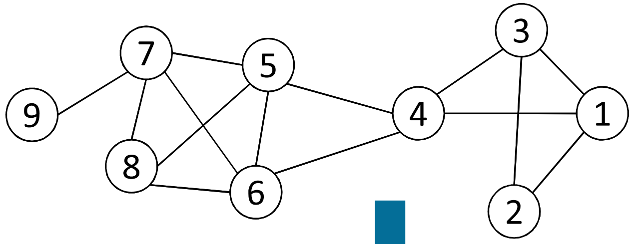
$$D = diag(d_1, d_2, \cdots, d_n) \quad \text{diagonal matrix of degrees}$$

Spectral relaxation: $\min_S Tr(S^T \widetilde{L} S) \quad s.t.\ S^T S = I_k$

Optimal solution: top eigenvectors with the smallest eigenvalues

Community Analysis :: Clustering

Spectral Clustering Example



$$D = diag(3, 2, 3, 4, 4, 4, 4, 3, 1)$$

$$\widetilde{L} = D - A = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 4 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 4 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

The 1st eigenvector means all nodes belong to the same cluster, no use

$$S = \begin{bmatrix} 0.33 & -0.38 \\ 0.33 & -0.48 \\ 0.33 & -0.38 \\ 0.33 & -0.12 \\ 0.33 & 0.16 \\ 0.33 & 0.16 \\ 0.33 & 0.30 \\ 0.33 & 0.24 \\ 0.33 & 0.51 \end{bmatrix}$$

k-means

Two communities:
{1, 2, 3, 4} and {5, 6, 7, 8, 9}

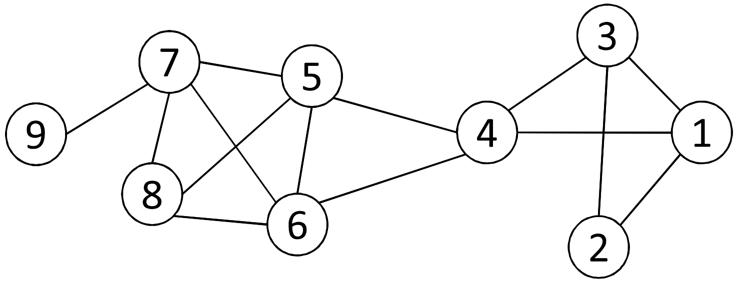
Community Analysis :: Clustering

Modularity Maximation

Modularity measures the **strength** of a **community partition** by taking into account the **degree distribution**

Given a network with ***m*** **edges**, the **expected number** of **edges** between two nodes with d_i and d_j is

$$d_i d_j / 2m$$



Strength of a community:

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

Modularity

$$Q = \frac{1}{2m} \sum_{\ell=1}^{k} \sum_{i \in C_\ell, j \in C_\ell} (A_{ij} - d_i d_j / 2m)$$

A larger value indicates a good community structure

Community Analysis :: Clustering

Modularity Maximation

Modularity matrix: $B = A - \mathbf{d}\mathbf{d}^T/2m \quad (B_{ij} = A_{ij} - d_i d_j/2m)$

Similar to spectral clustering, **Modularity maximization** can be reformulated as

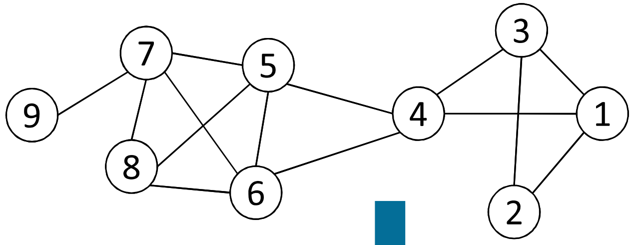
$$\max Q = \frac{1}{2m} Tr(S^T BS) \quad s.t.\ S^T S = I_k$$

Optimal solution: **top eigenvectors** of the **modularity matrix**

Apply k-means to S as a post-processing step to obtain community partition

Community Analysis :: Clustering

Modularity Maximation Example



Two communities:
{1, 2, 3, 4} and {5, 6, 7, 8, 9}

k-means

$$B = \begin{bmatrix} -0.32 & 0.79 & 0.68 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.79 & -0.14 & 0.79 & -0.29 & -0.29 & -0.29 & -0.29 & -0.21 & -0.07 \\ 0.68 & 0.79 & -0.32 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.57 & -0.29 & 0.57 & -0.57 & 0.43 & 0.43 & -0.57 & -0.43 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & -0.57 & 0.43 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & 0.43 & -0.57 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & -0.57 & 0.43 & 0.43 & -0.57 & 0.57 & 0.86 \\ -0.32 & -0.21 & -0.32 & -0.43 & 0.57 & 0.57 & 0.57 & -0.32 & -0.11 \\ -0.11 & -0.07 & -0.11 & -0.14 & -0.14 & -0.14 & 0.86 & -0.11 & -0.04 \end{bmatrix}$$

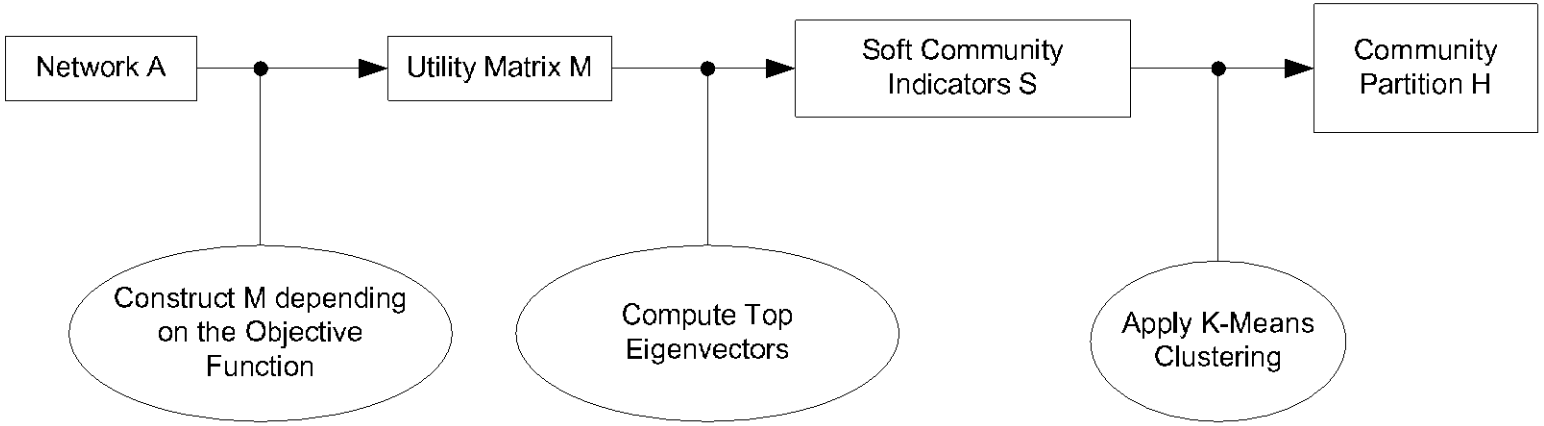
Modularity Matrix

$$S = \begin{bmatrix} 0.44 & -0.00 \\ 0.38 & 0.23 \\ 0.44 & -0.00 \\ 0.17 & -0.48 \\ -0.29 & -0.32 \\ -0.29 & -0.32 \\ -0.38 & 0.34 \\ -0.34 & -0.08 \\ -0.14 & 0.63 \end{bmatrix}$$

Community Analysis :: Clustering

A Unified View for Community Platform

Latent space models, **block** models, **spectral clustering**, and **modularity maximization** can be unified as



$$\text{Utility Matrix } M = \begin{cases} \text{modified proximity matrix } \widetilde{P} & \textit{if} \text{ latent space models} \\ \text{adjacency matrix } A & \textit{if} \text{ block models} \\ \text{graph Laplacian } \widetilde{L} & \textit{if} \text{ spectral clustering} \\ \text{modularity maximization } \; B & \textit{if} \text{ modularity maximization} \end{cases}$$

Community Analysis :: Clustering

Hierarchy-Centric Clustering

Goal: build a **hierarchical structure** of **communities**
based on **network topology**

Allow the analysis of a network at different resolutions

Representative approaches:

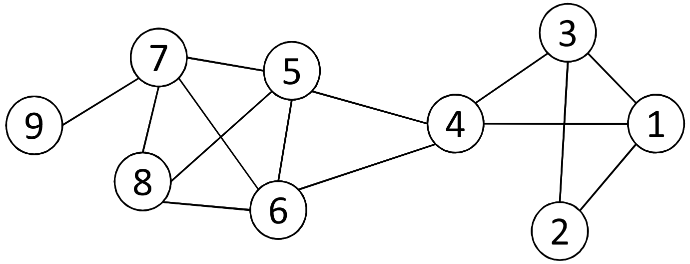
- **Divisive Hierarchical Clustering**
- **Agglomerative Hierarchical Clustering**

Community Analysis :: Clustering

Divisive Hierarchical Clustering

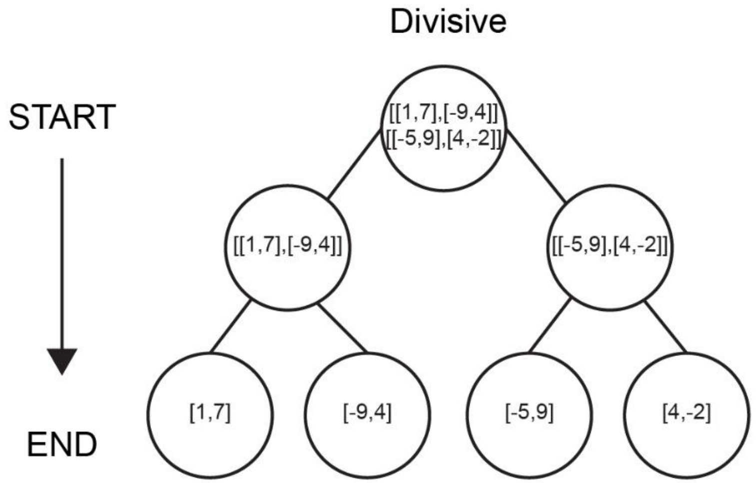
Divisive clustering

- Partition nodes into several sets
- Each set is further divided into smaller ones
- Network-partition can be applied for the partition



One particular example: recursively remove the "weakest" tie

- Find the edge with the least strength
- Remove the edge and update the corresponding strength of each edge



Recursively apply the above two steps until a network is discomposed into desired number of connected components.

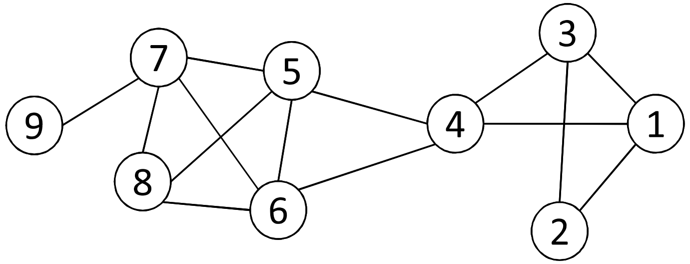
Each component forms a community

Community Analysis :: Clustering

Edge Betweeness

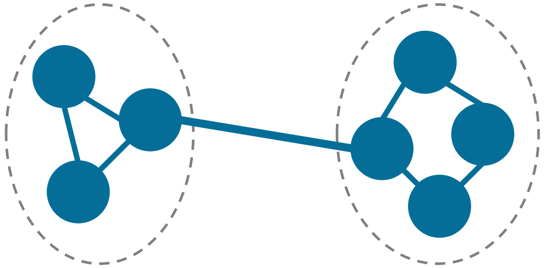
The **strength** of a **tie** can be measured by **edge betweenness**
Edge betweenness: the number of shortest paths that pass along with the edge

$$\text{edge-betweenness(e)} = \Sigma_{s<t} \frac{\sigma_{st}(e)}{\sigma_{s,t}}$$



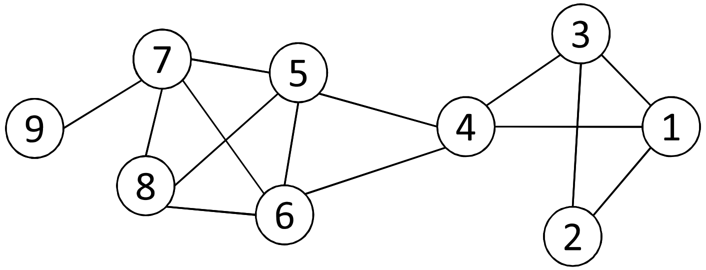
The edge betweenness of e(1, 2) is 4, as all the shortest paths from 2 to {4, 5, 6, 7, 8, 9} have to either pass e(1, 2) or e(2, 3), and e(1,2) is the shortest path between 1 and 2

The edge with **higher betweenness** tends to be the **bridge between two communities.**



Community Analysis :: Clustering

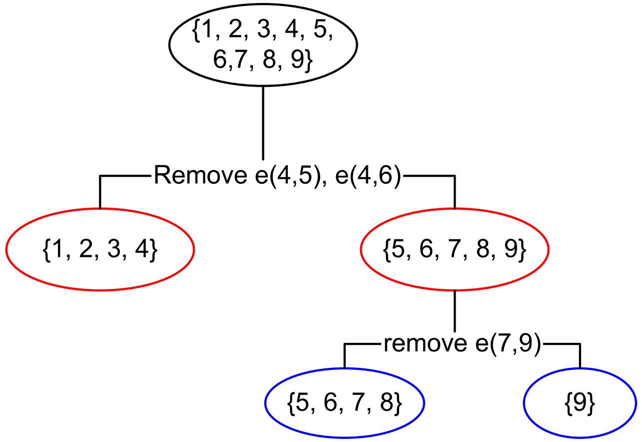
Divisive clustering based on Edge Betweenness



Initial betweenness value

Table 3.3: Edge Betweenness

	1	2	3	4	5	6	7	8	9
1	0	4	1	9	0	0	0	0	0
2	4	0	4	0	0	0	0	0	0
3	1	4	0	9	0	0	0	0	0
4	9	0	9	0	10	10	0	0	0
5	0	0	0	10	0	1	6	3	0
6	0	0	0	10	1	0	6	3	0
7	0	0	0	0	6	6	0	2	8
8	0	0	0	0	3	3	2	0	0
9	0	0	0	0	0	0	8	0	0



After remove e(4,5), the betweenness of e(4, 6) becomes 20, which is the highest;

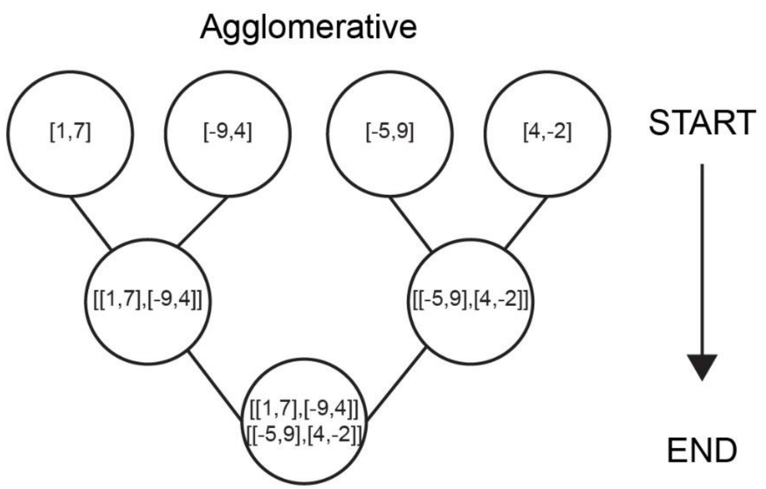
After remove e(4,6), the edge e(7,9) has the highest betweenness value 4, and should be removed.

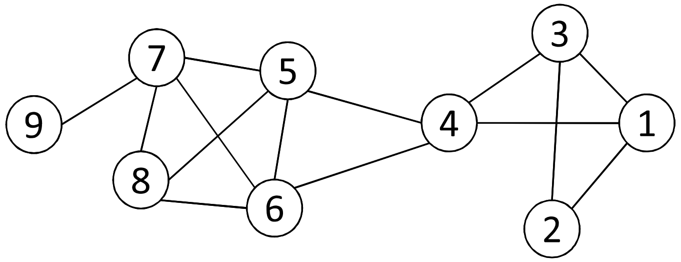
Community Analysis :: Clustering

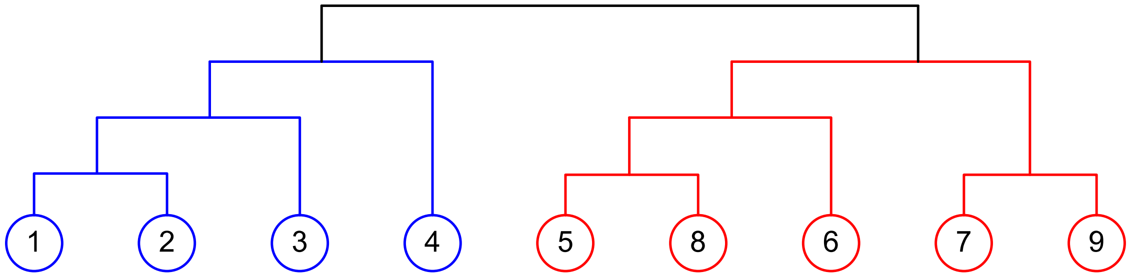
Agglomerative Hierarchical Clustering

- **Initialize** each node as a **community**
- **Merge communities** successively into larger communities following a certain criterion

  (E.g., based on modularity increase)







Social Computing :: Community Detection

Summary of Community Detection

Node-Centric Community Detection

- *cliques, k-cliques, k-clubs*

Group-Centric Community Detection

- *quasi-cliques*

Network-Centric Community Detection

- *Clustering based on vertex similarity*
- *Latent space models, block models, spectral clustering, modularity maximization*

Hierarchy-Centric Community Detection

- *Divisive clustering*
- *Agglomerative clustering*

Community Analysis

Algorithms and Data Structures 2, 340300
Lecture – 2023W
Univ.-Prof. Dr. Alois Ferscha, teaching@pervasive.jku.at

Institute of
Pervasive Computing

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at