Advanced Human Machine Interaction
Interaction Data Analysis

Social Network Analysis

Alexandre Pauchet

alexandre.pauchet@insa-rouen.fr - BO.B.RC18

Normandie Université

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

litis

Scope: interactions in social networks



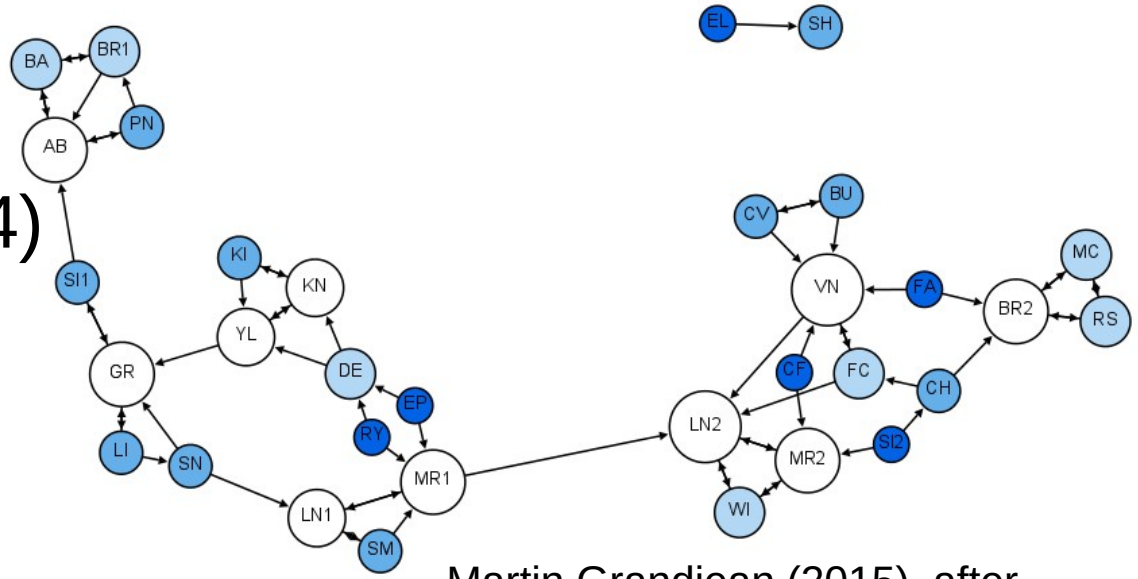


- **Interactions** in social networks:
  - Messages, actions (Friend, Like, …), …
- Interactions in **social** networks:
  - Human-to-human mediated interaction, bots
  - Open world
- Interactions in social **networks**:
  - Graph representation
  - Dynamic network

Social networks before online platforms

- Jacob L. Moreno sociometry, 1930s

  Who Shall Survive? (1934)

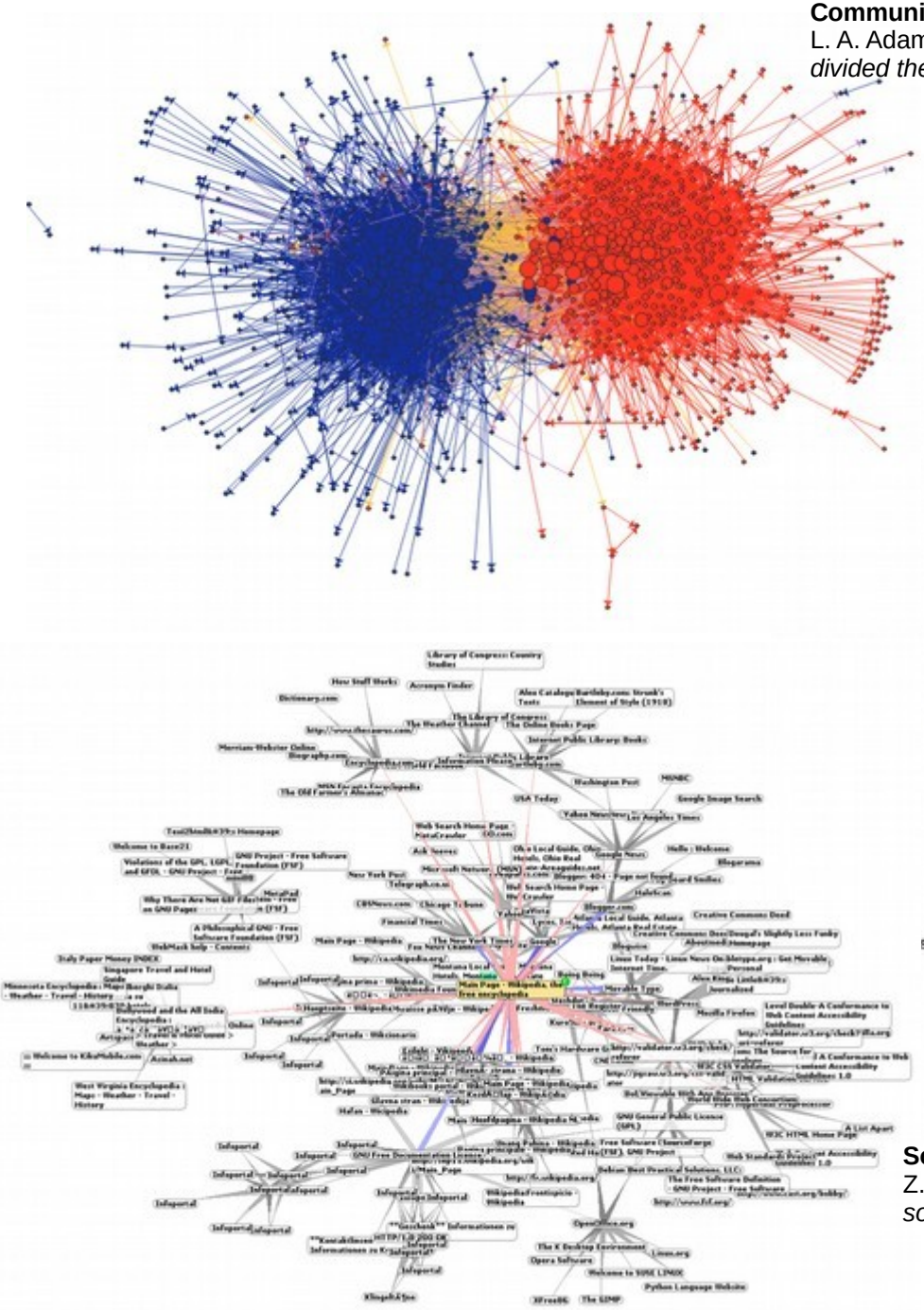


Martin Grandjean (2015), after Moreno (1934), Who Shall Survive? - CC BY-SA 4.0

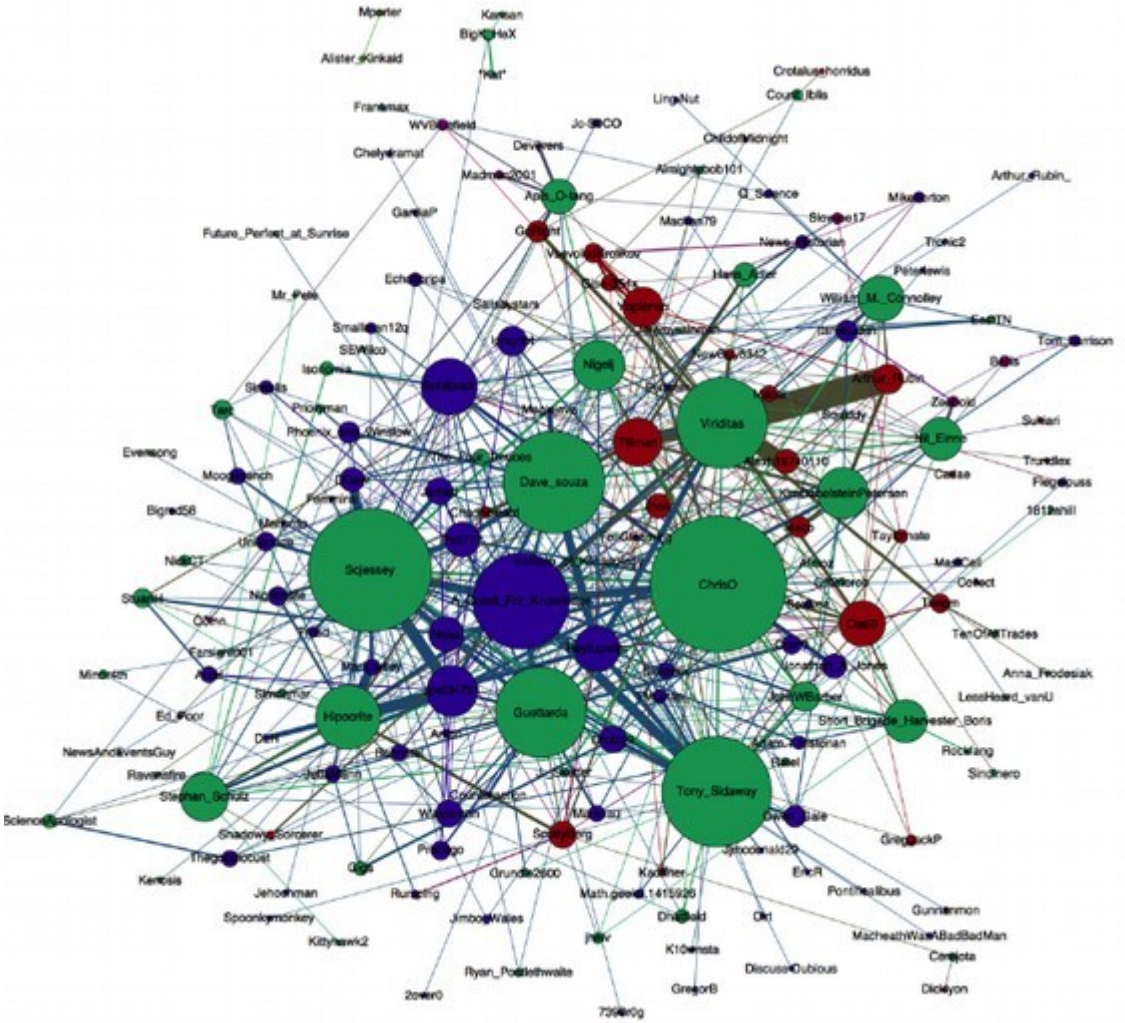
- Nodes: pupils
- Directed edges: an arrow indicates a pupil's choice of a preferred seatmate

- What can we observe?
  - Reciprocal choices;
  - isolated or frequently chosen pupils;
  - groups with few connections between them.

Examples



Community structure of political US blogs.
L. A. Adamic and N. Glance, "*The political blogosphere and the 2004 U.S. election: divided they blog*", In Proceedings of Link discovery (LinkKDD '05), ACM, pp. 36-43, 2005.

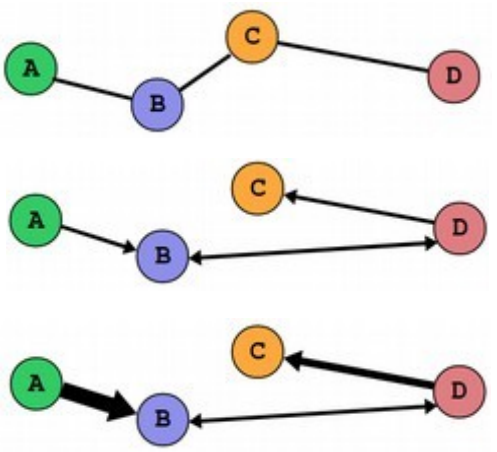


Social Sybil detection / false identities.
Z. Yamak, J. Saunier, L. Vercouter, "*Automatic detection of multiple account deception in social media*", in proceedings of Web Intelligence, pp. 219-231, 2017.

Process

1) Dataset collection

- Define the question and collect the data (population, interaction type, observation period, ...)
- Anonymization and preprocessing

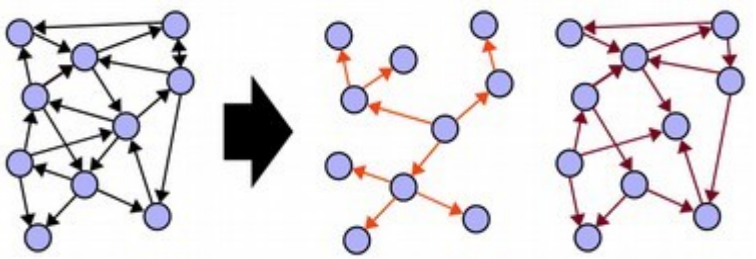


2) Graph construction

- Nodes (vertices) / Links (edges)
- Directed or undirected graph
- Weights
- Self-loops

3) Inspection and filtering

- Missing data (nodes or edges)
- Node/edge removing (isolated nodes, thresholds)
- Network decomposition

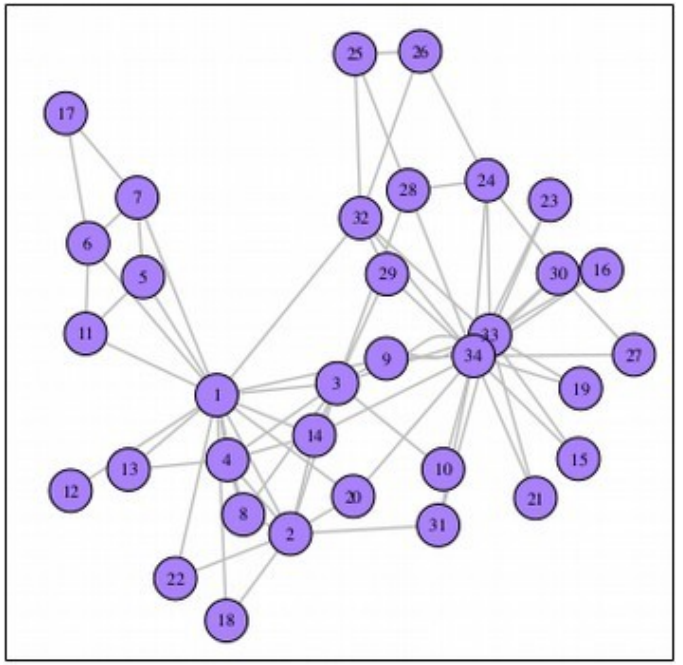


4) Interpret and assess robustness

- What do the results mean (objective function)?
- How sensitive are they to modelling choices

Graph: definition

- A **graph G = (V, E)** can be represented with
  - V, the set of nodes (vertices);
  - E, the set of edges (links);



Zachary's Karate Club Network
W. W. Zachary, "*An information flow model for conflict and fission in small groups*", Journal of Anthropological Research 33, pp. 452-473 (1977).

- **Directed graph**
  an edge (u,v) is an ordered pair

- **Undirected graph**
  an edge {u,v} is an unordered pair

- **Weighted graph**
  nodes and edges may carry attributes
  a weighted graph associates a numerical weight w(e) with each edge

Graph construction

Graph construction



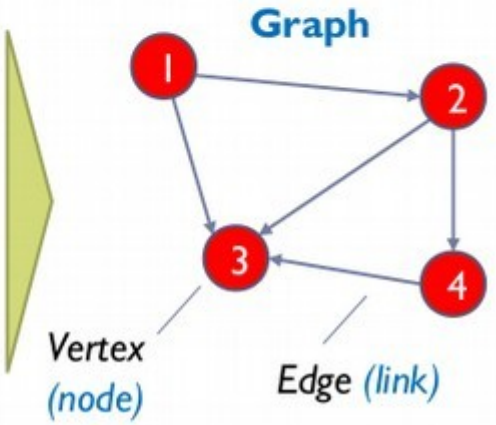


Can we study their interactions as a network?

$$A_{uv} = \begin{cases} 1 & \text{if } u \text{ addresses } v, \\ 0 & \text{otherwise.} \end{cases}$$

Communication

Anne: Jim, tell the Murrays they're invited
Jim: Mary, you and your dad should come for dinner!
Jim: Mr. Murray, you should both come for dinner
Anne: Mary, did Jim tell you about the dinner? You must come.
John: Mary, are you hungry?
…



Adjacency matrix

senders

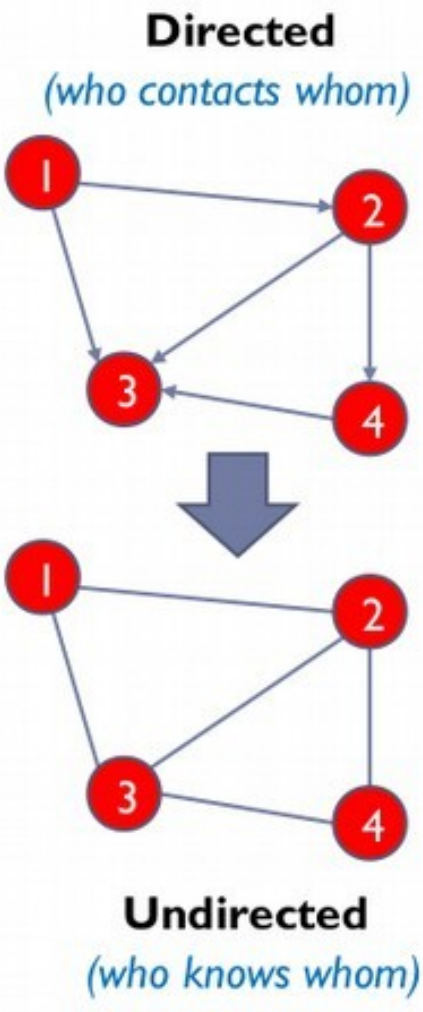
recipients

Vertex	1	2	3	4
1	-	1	1	0
2	0	-	1	1
3	0	0	-	0
4	0	0	1	-

Edge list

Vertex	Vertex
1	2
1	3
2	3
2	4
4	3

Directed *versus* undirected graphs



Edge list

Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

Adjacency matrix becomes symmetric

Vertex	1	2	3	4
1	-	1	1	0
2	1	-	1	1
3	1	1	-	1
4	0	1	1	-

Exercise

Directed or undirected?

The observed interactions are:

$1 \rightarrow 2$; $2 \rightarrow 1$; $1 \rightarrow 3$

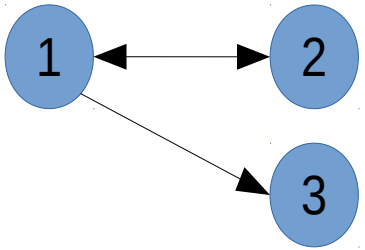
Write the edge list:

- when direction matters;
- when only the existence of an interaction matters.

Store each undirected edge once.

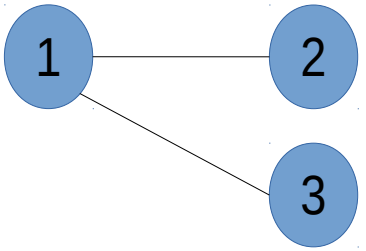
Exercise

Directed Graph



Vertex	Vertex
1	2
1	3
2	1

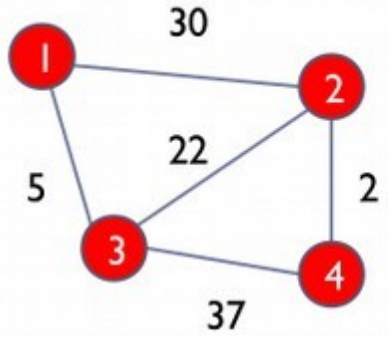
Undirected Graph



Vertex	Vertex
1	2
1	3

A symmetric adjacency matrix represents each undirected edge in two entries.
The edge list stores it once.

Weighted edges



Weights could be:

- *Frequency of interaction in period of observation*
- *Number of items exchanged in period*
- *Individual perceptions of strength of relationship*
- *Costs in communication or exchange, e.g. distance*
- *Combinations of these*

Edge list: add column of weights

Vertex	Vertex	Weight
1	2	30
1	3	5
2	3	22
2	4	2
3	4	37

Adjacency matrix: add weights instead of 1

Vertex	1	2	3	4
1	-	30	5	0
2	30	-	22	2
3	5	22	-	37
4	0	2	37	-

Interaction strength and path distance are different quantities:
a stronger connection does not necessarily mean a longer path.

Exercise

Construct a directed, weighted interaction graph

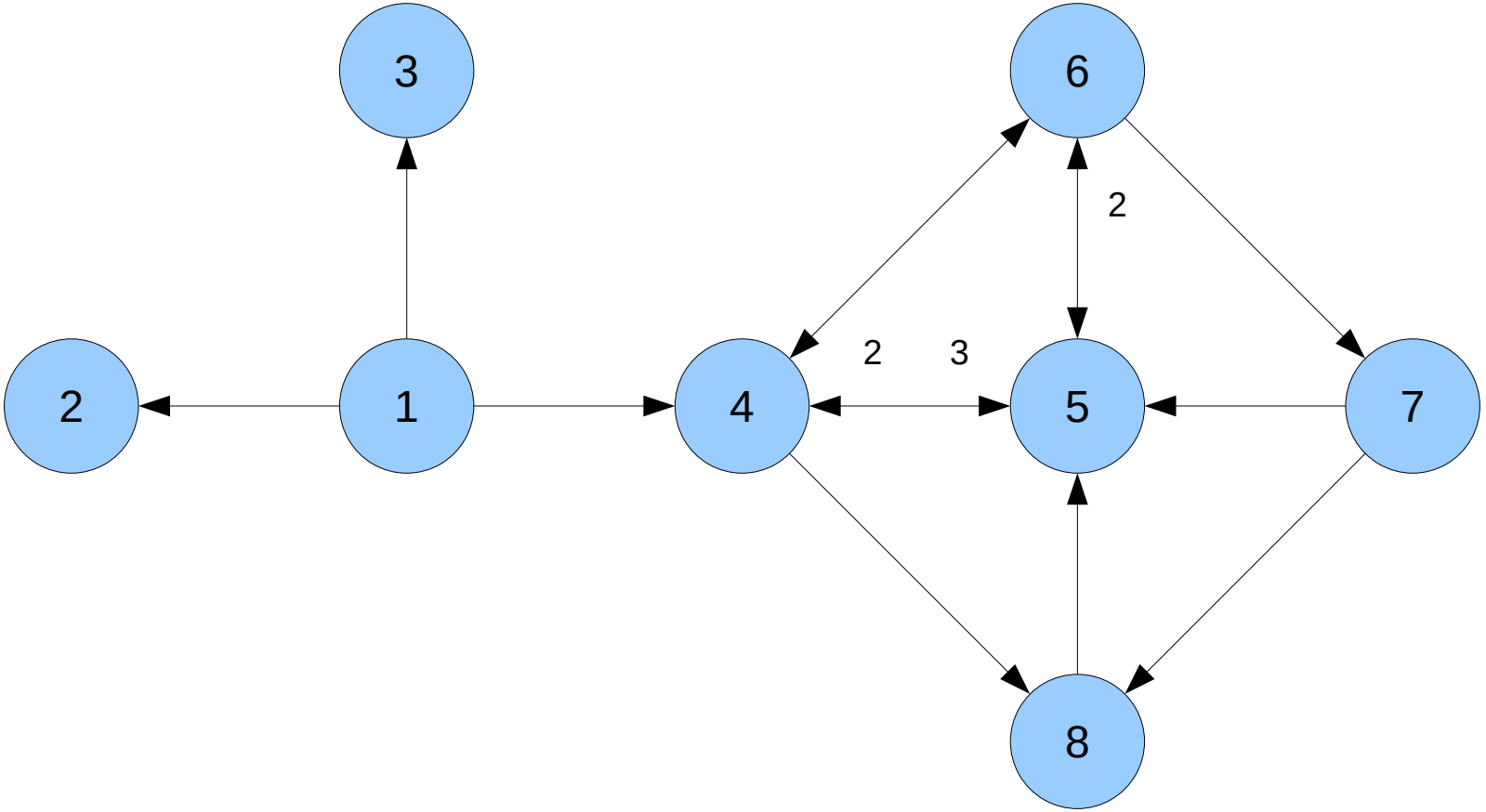
- One node per author
- One edge u → v when u explicitly addresses v
- Each qualifying message contributes 1 to the edge weight
- Ignore messages without an identifiable recipient
- Ignore message topics for this graph
- Retweet: point at the original sender

Auteur	Message	Interaction	Thème
1	hello @2 viens sur mon site : http :url	ME	Autre
1	hello @3 viens sur mon site : http :url	ME	Autre
1	hello @4 viens sur mon site : http :url	ME	Autre
6	un cinéma @5 ?	ME	Cinéma
5	@6 : quel film ? #Film, #Film1 ou photo_Film_2 ?	RE	Cinéma
5	allons au cinéma @4 ! @6 vient aussi	ME	Cinéma
4	@5 : ok	RE	Autre
4	super #Film, à la prochaine @6	ME	Cinéma
5	RT @4 : super #Film, à la prochaine @6	RT, ME	Cinéma
6	RT @4 : super #Film, à la prochaine @6	RT, ME	Cinéma
5	la bande-annonce de photo_Film_2 :) ! #Film2 dès demain !		Cinéma
4	RT @5 : la bande-annonce de photo_Film_2 :) ! #Film2 dès demain !	RT	Cinéma
4	au fait, @8, tu devrais aller voir #Film	ME	Cinéma
5	le théâtre c'est bien aussi		Théâtre
4	RT @5 : le théâtre c'est bien aussi	RT	Théâtre
7	RT @5 : le théâtre c'est bien aussi	RT	Théâtre
8	RT @5 : le théâtre c'est bien aussi	RT	Théâtre
6	@7 : cinéma :-) theââââatre :-(	RE	Cinéma
7	viens @8, allons au théâtre	ME	Théâtre

Exercise

Auteur	Message	Interaction	Thème
1	hello @2 viens sur mon site : http :url	ME	Autre
1	hello @3 viens sur mon site : http :url	ME	Autre
1	hello @4 viens sur mon site : http :url	ME	Autre
6	un cinéma @5 ?	ME	Cinéma
5	@6 : quel film ? #Film, #Film1 ou photo_Film_2 ?	RE	Cinéma
5	allons au cinéma @4 ! @6 vient aussi	ME	Cinéma
4	@5 : ok	RE	Autre
4	super #Film, à la prochaine @6	ME	Cinéma
5	RT @4 : super #Film, à la prochaine @6	RT, ME	Cinéma
6	RT @4 : super #Film, à la prochaine @6	RT, ME	Cinéma
5	la bande-annonce de photo_Film_2 :) ! #Film2 dès demain !		Cinéma
4	RT @5 : la bande-annonce de photo_Film_2 :) ! #Film2 dès demain !	RT	Cinéma
4	au fait, @8, tu devrais aller voir #Film	ME	Cinéma
5	le théâtre c'est bien aussi		Théâtre
4	RT @5 : le théâtre c'est bien aussi	RT	Théâtre
7	RT @5 : le théâtre c'est bien aussi	RT	Théâtre
8	RT @5 : le théâtre c'est bien aussi	RT	Théâtre
6	@7 : cinéma :-) theââââatre :-(	RE	Cinéma
7	viens @8, allons au théâtre	ME	Théâtre

Exercise



What would change if we keep only replies, or analyse each topic separately?

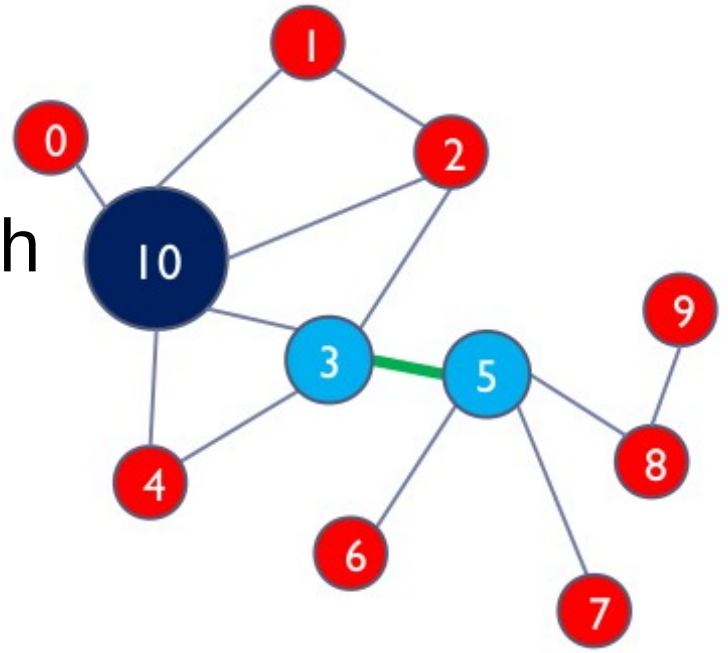
Identification of Key-users Roles in Social Networks

Node importance and structural roles

Identifying key users

Which nodes are (the most) important?

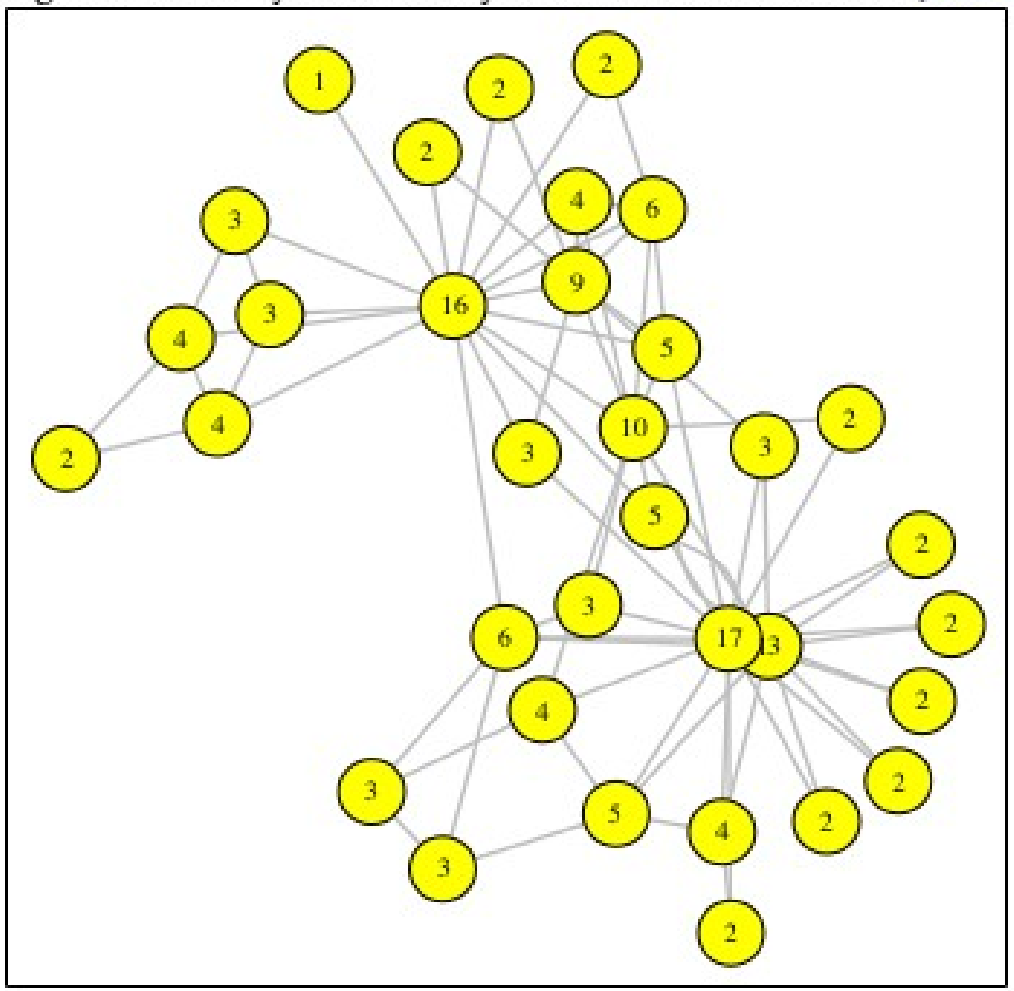
- Node 10 has the largest number of direct neighbours
- Nodes 3 and 5 reach more nodes with paths of length 2
- Nodes 3 and 5 connect the two parts of the network
- Removing the edge between 3 and 5 disconnects the network



Which node would you choose to reach many direct contacts?
Which nodes would you monitor to understand communication between groups?

Degree centrality

Degree Centrality for Zachary's Karate Club Network (Zachary, 1977)



Undirected graph

- Degree *k(v)*: number of neighbours of node *v*
- Normalised degree centrality:

$$C_D(v) = \frac{k(v)}{n-1}$$

Directed graph

- In-degree: number of incoming edges
- Out-degree: number of outgoing edges

Weighted graph

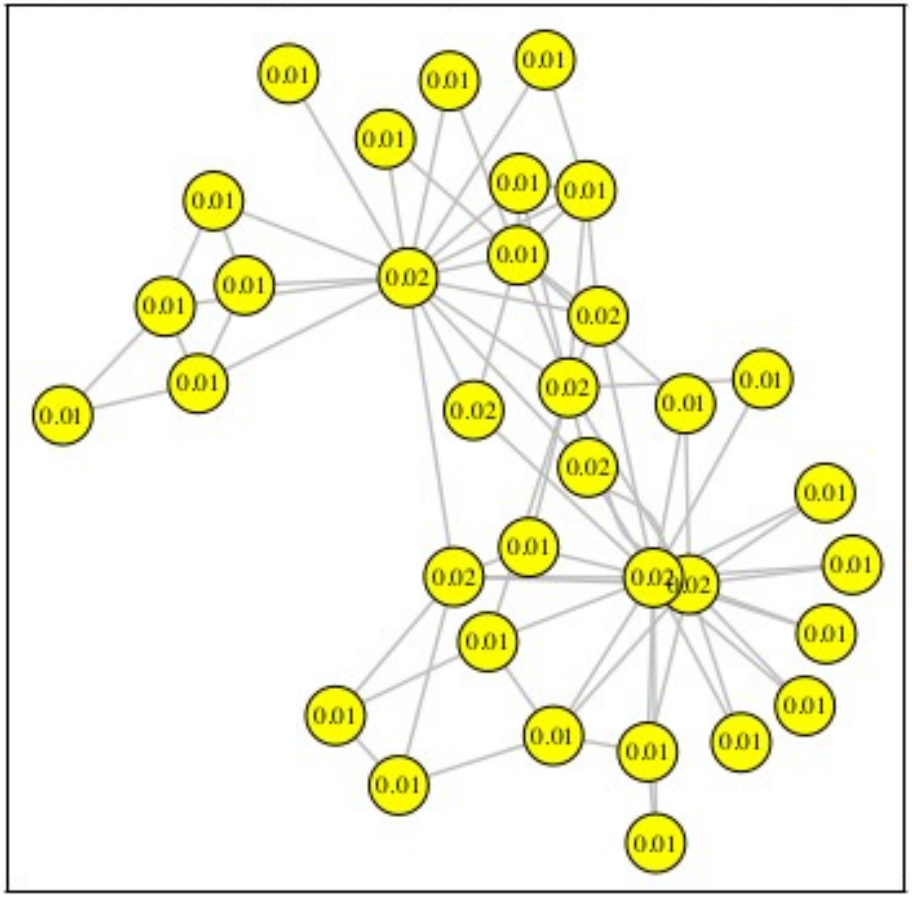
- Strength: sum of the weights of incident edges
- Interpretation depends on what the edges represent

Shortest paths and closeness centrality

- **Shortest-path distance**
  - *d(u,v)*: minimum number of edges needed to reach *v* from *u* in an unweighted graph

NB: Shortest paths are desirable when speed of communication or exchange is desired.

Closeness Centrality for Zachary's Karate Club Network (Zachary, 1977)



- **Closeness centrality**:
  - A node has high closeness when it can reach all other nodes through short paths
  - For a connected, undirected graph with *n* nodes:

$$C_C(v) = \frac{n-1}{\sum_{u \neq v} d(v,u)}$$

  - High values indicate a central position

Interpretation

  - How many steps are needed to reach the other participants?
  - Short paths describe structural accessibility

For disconnected graphs, unreachable nodes require an adapted definition!

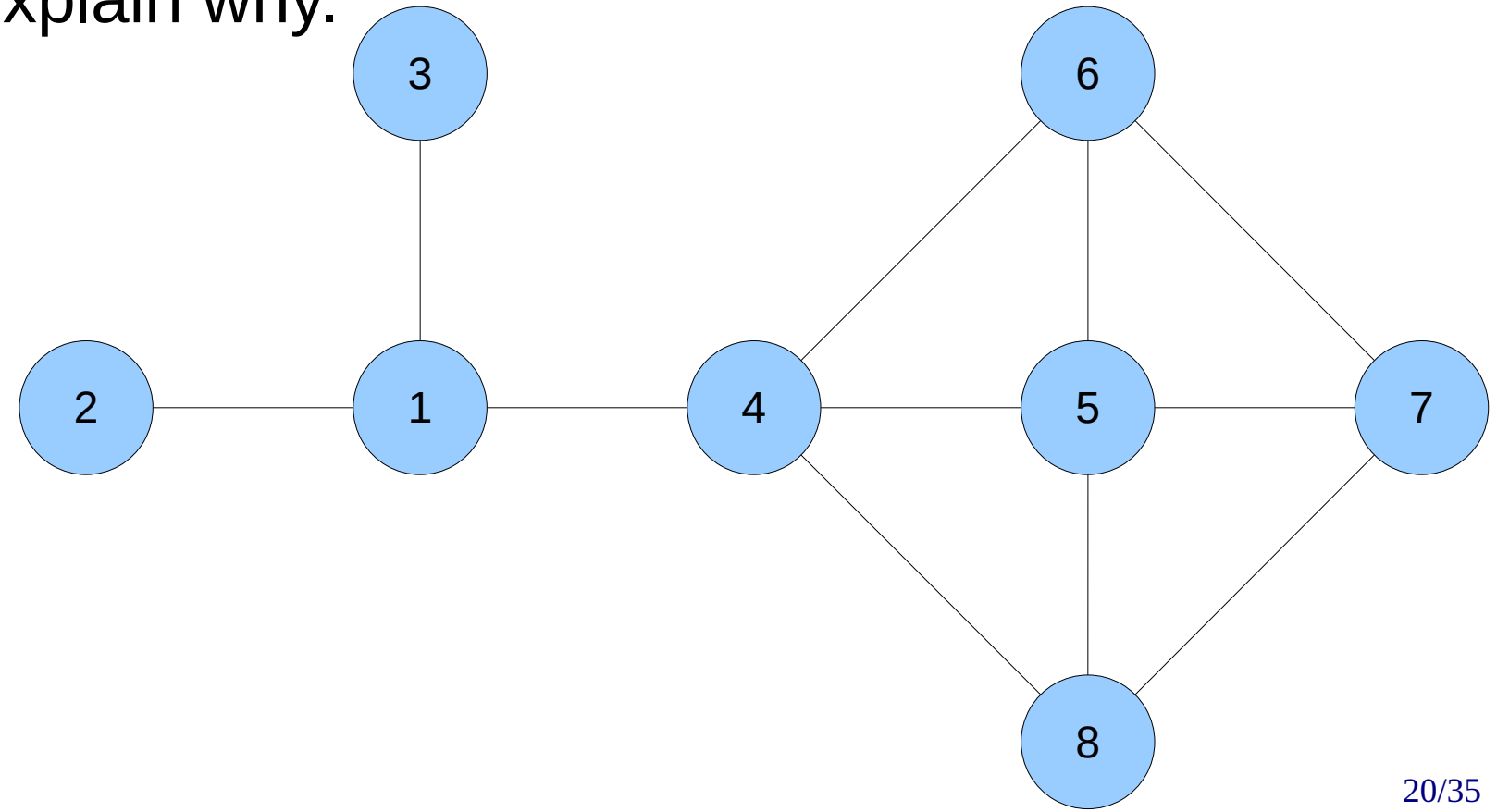
Exercise

Consider this undirected, unweighted graph

1. Identify the nodes with the highest degree

2. Compute closeness centrality for nodes 4 and 5

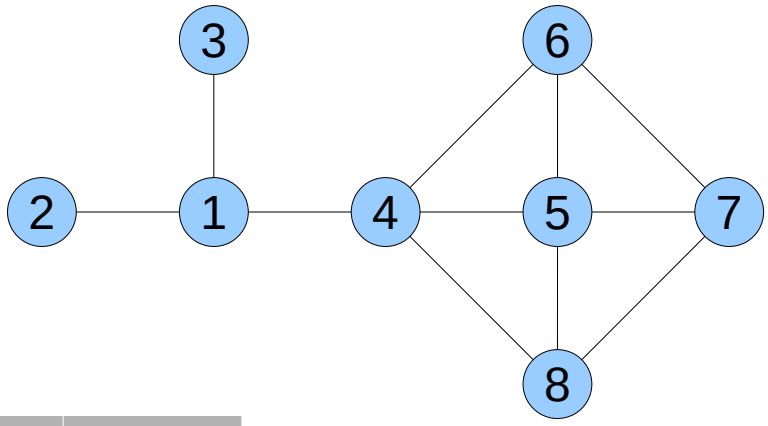
3. They have the same degree. Do they have the same closeness? Explain why.



Exercise

- Degree centrality

N1	N2	N3	**N4**	**N5**	N6	N7	N8
3	1	1	**4**	**4**	3	3	3



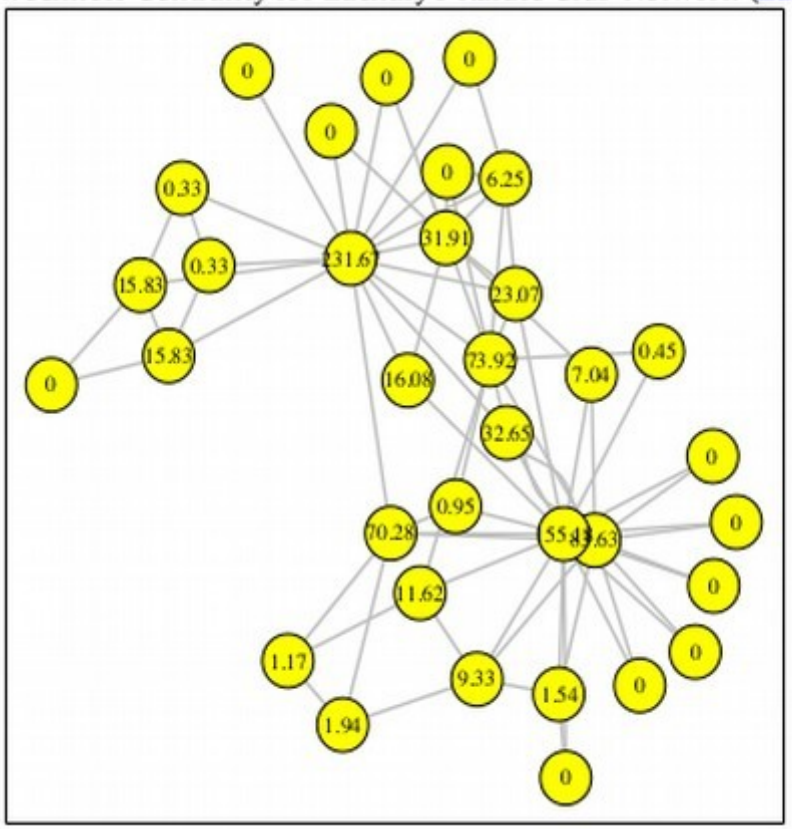
- Closeness

N1	N2	N3	**N4**	N5	N6	N7	N8
0,583	0,389	0,389	**0,700**	0,583	0,538	0,438	0,538

Nodes 4 and 5 have the same number of neighbours,
but node 4 is closer to the network as a whole

Betweenness centrality

Betweenness Centrality for Zachary's Karate Club Network (Zachary, 1977)



Betweenness centrality:
Measures how often a node lies on shortest paths between other nodes

$$C_B(v) = \sum_{\substack{s<t \\ s,t \neq v}} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

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

$\sigma_{st}(v)$: number of these paths passing through *v*

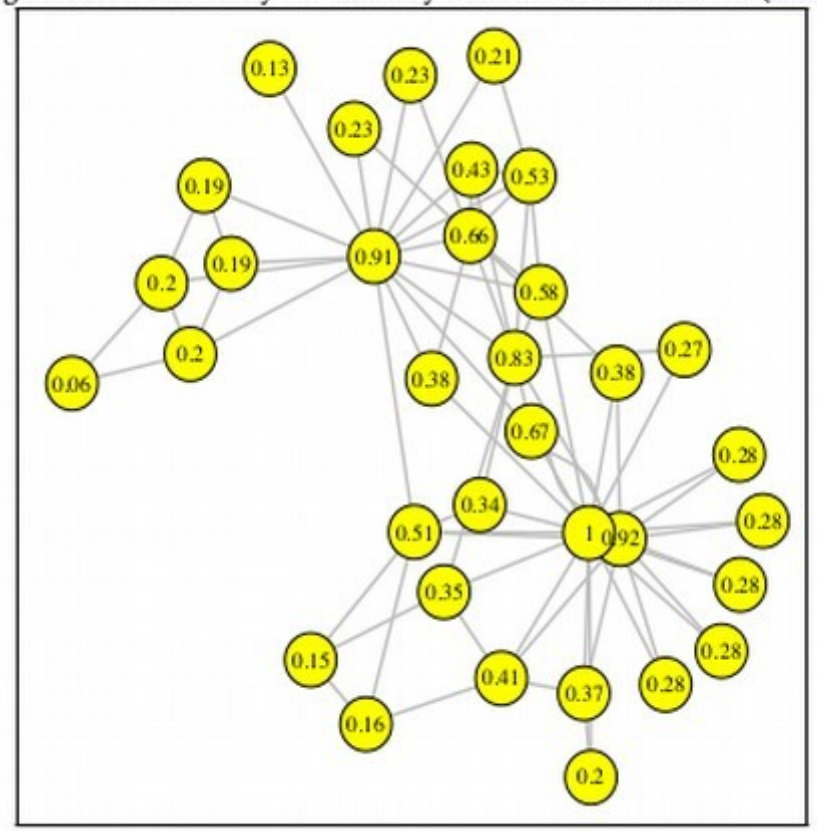
For an undirected graph, count each unordered pair once

1) Consider each pair of distinct nodes other than v
2) Compute the fraction of their shortest paths passing through v
3) Sum these fractions over all pairs

Can a node with few neighbours still have high betweenness?

Eigenvector centrality (or eigencentrality)

Eigenvector Centrality for Zachary's Karate Club Network (Zachary, 1977)



Eigenvector centrality:

- A node has high eigenvector centrality when it connects to other nodes with high eigenvector centrality
- For an undirected, unweighted graph:

$$x_v = \frac{1}{\lambda} \sum_{u \in N(v)} x_u \qquad \text{or equivalently} \qquad Ax = \lambda x$$

  - A: adjacency matrix.
  - x_v: centrality score of node v
  - $N(v)$: neighbours of v
  - Use a non-negative eigenvector associated with the largest eigenvalue

Can two nodes with the same degree have different eigenvector centralities?

Interpretation

Connections to highly central neighbours contribute more to the score

Interpretating centrality measures

Measure	Question it answers	Possible interpretation
Degree	How many direct neighbours does this participant have?	Extent of direct contacts
Closeness	How few steps separate this participant from everyone else?	Structural accessibility
Betweenness	How often does this participant lie on shortest paths between others?	Potential brokerage
Eigenvector	How central are this participant's neighbours?	Connection to central participants

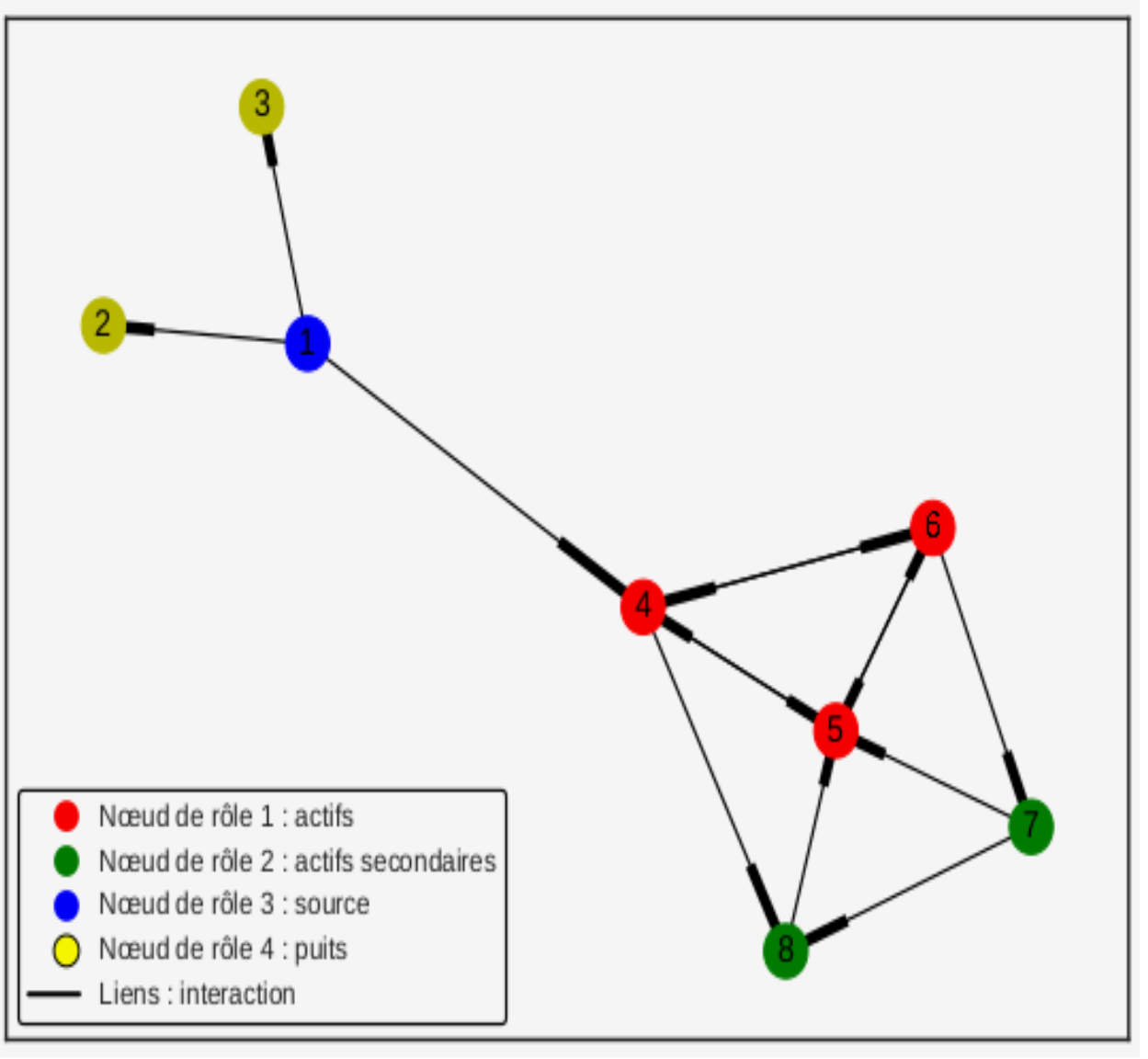
Interpretation depends on the meaning and direction of edges, their weights, and the observation period

Structural roles: the RolX example

- Structural roles describe how nodes connect to their surroundings
- Nodes with similar structural profiles may share a role, even if they are far apart in the network.

RolX

- Extract structural features from each node and its neighbourhood
- Use non-negative matrix factorisation to discover latent roles
- Represent each node as a mixture of roles
- The number of roles can be selected automatically using a model-selection criterion (MDL)



Community Detection

Identifying groups of densely connected nodes

Homophily, triadic closure and bridges

Homophily

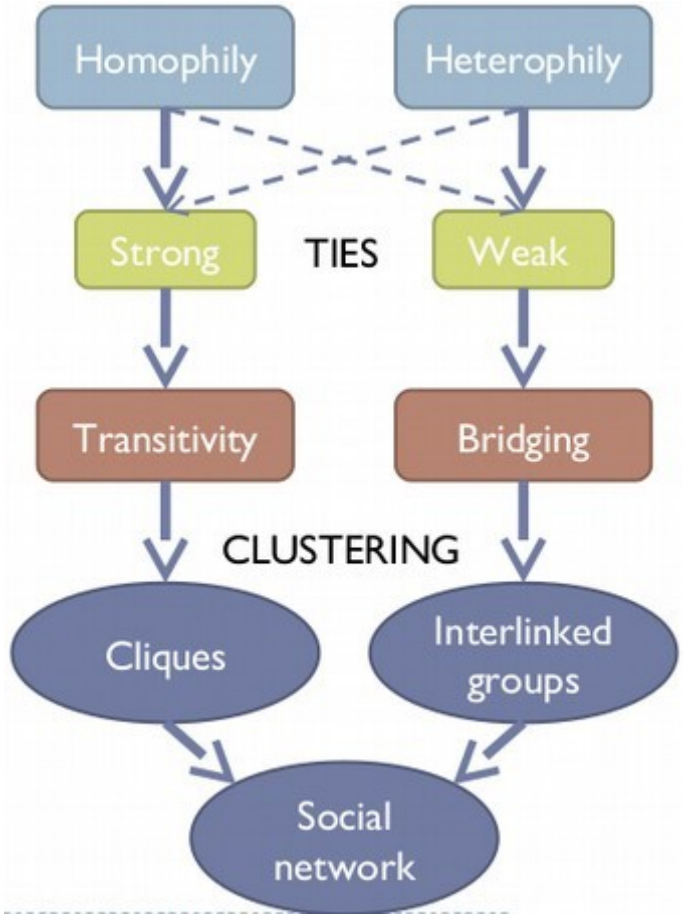
- People tend to connect with others who share similar characteristics
- These similarities may concern interests, beliefs or social background

Triadic closure

- Two people who share a contact may become connected
- This process creates triangles and contributes to local clustering

Connections between groups

- Some edges connect otherwise weakly connected groups
- They provide potential routes for communication between groups



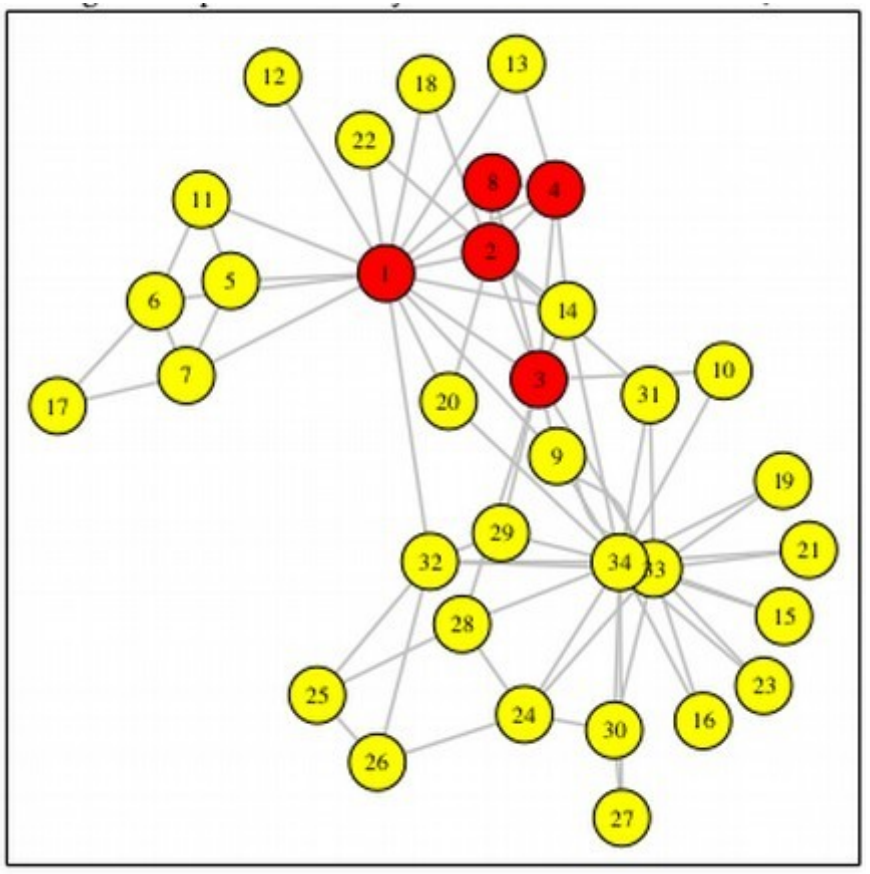
A **bridge** is an edge whose removal disconnects a connected component.
An **articulation point** is a node with this property.

Cliques and communities

A **clique** is a set of nodes in which every pair of distinct nodes is connected by an edge

A **k-clique** contains k nodes. Each node has k − 1 neighbours within the clique.

A **community** may contain missing internal edges: it does not need to be a clique.



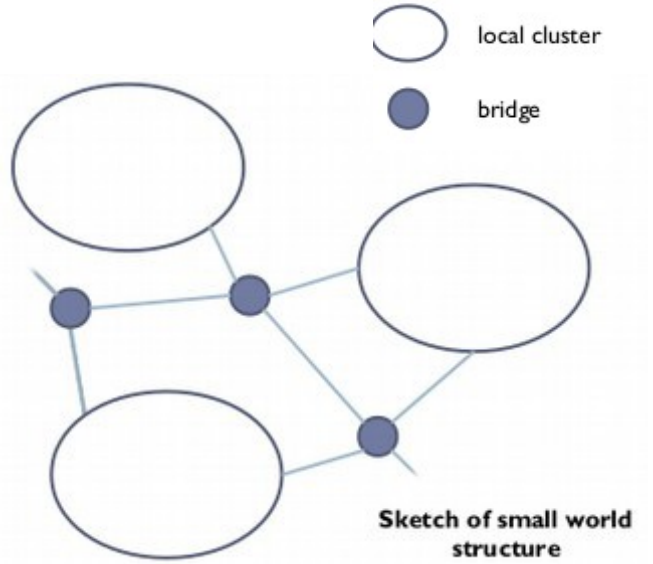
If we remove one edge from a clique, is it still a clique? Could it still be a community?

Small-world networks

Many social networks combine:

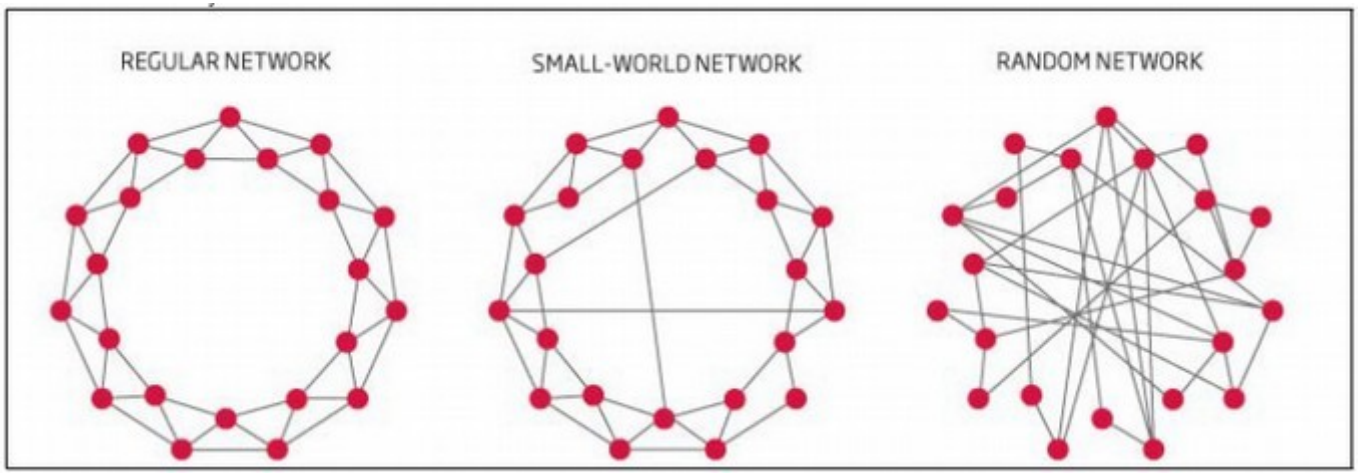
- High local clustering:
  neighbours of a node often connect to each other
- Short average path lengths:
  relatively few steps connect most pairs of nodes

Connections between distant parts of the network can substantially shorten paths.



Sketch of small world structure

Small-world structure and community structure describe different properties of a network



Modularity: evaluating a partition

A partition divides the nodes into non-overlapping communities

Modularity compares:

- the observed weight of edges within communities
- the expected weight under a random-network model that preserves node strengths in expectation

A high modularity indicates more internal connectivity than expected under this reference model

For an undirected graph, modularity is

$$Q = \sum_{C \in \mathcal{P}} \left[\underbrace{\frac{w_C}{m}}_{\text{Observed fraction}} - \underbrace{\left(\frac{s_C}{2m}\right)^2}_{\text{Expected fraction}} \right]$$

- $\mathcal{P}$: partition of the nodes
- m: total edge weight, counting each edge once
- w_C: total weight of edges internal to community C, counted once
- s_C: sum of the strengths of nodes in C

Community detection algorithms

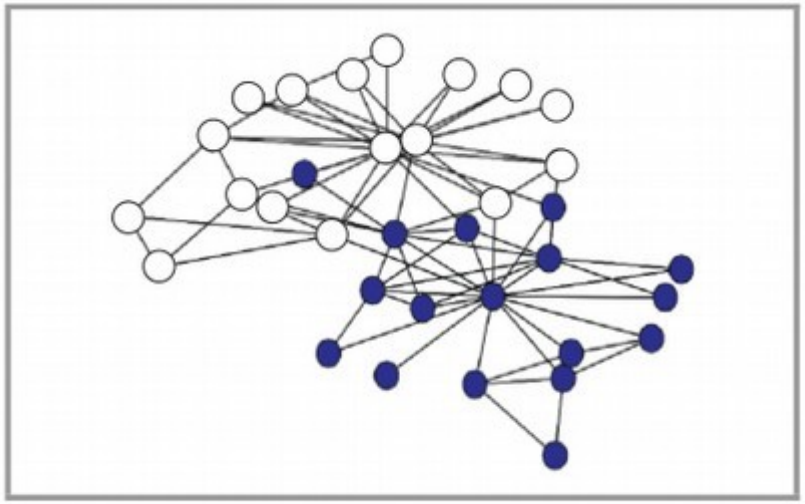


Fig. 1. Detected communities (Clauset&Newman aglorithm).

FastGreedy
Clauset, A., Newman, M. E., and Moore, C. : "*Finding community structure in very large networks",* Physical review E, 70(6), 2004.

Progressively merge communities to increase modularity

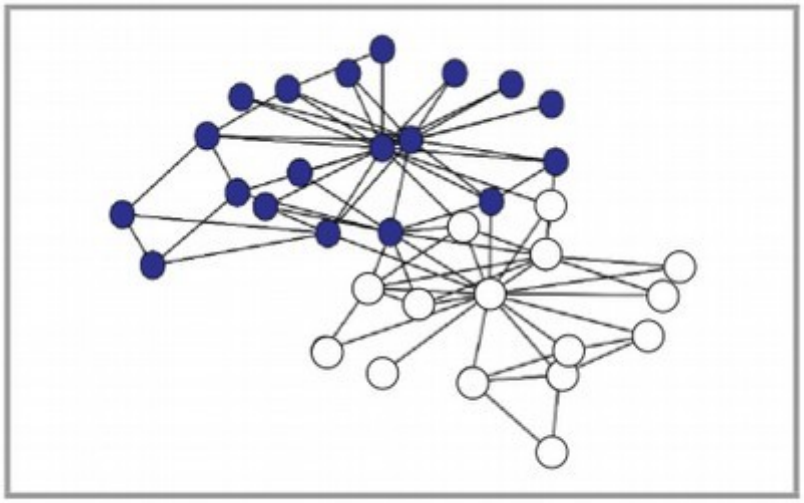


Fig. 2. Detected communities (MLC aglorithm).

Markov CLustering for graphs (MCL)
Van Dongen, S. M., *"Graph clustering by flow simulation*", PhD thesis, 2000.

Simulate flow in the graph to identify regions where it remains concentrated.

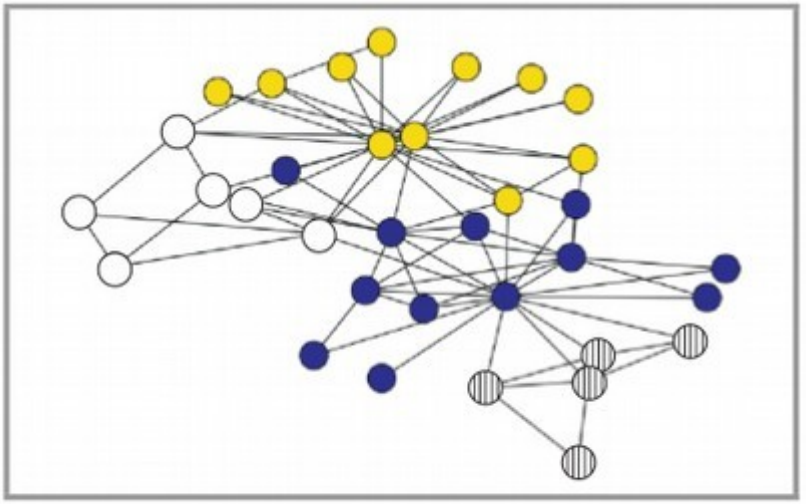
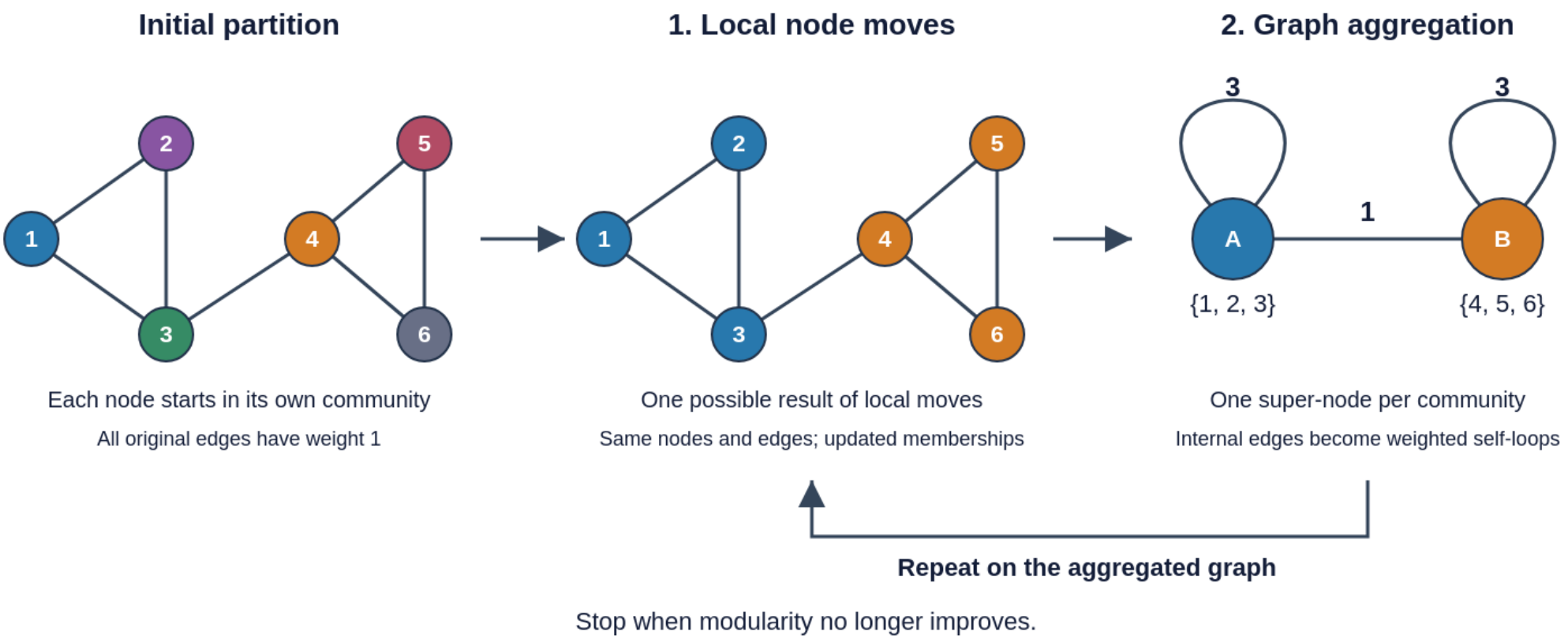


Fig. 3. Detected communities (Blondel aglorithm).

Blondel/Louvain
Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E., "*Fast unfolding of communities in large networks*", Journal of statistical mechanics : theory and experiment, volume 10, 2008.

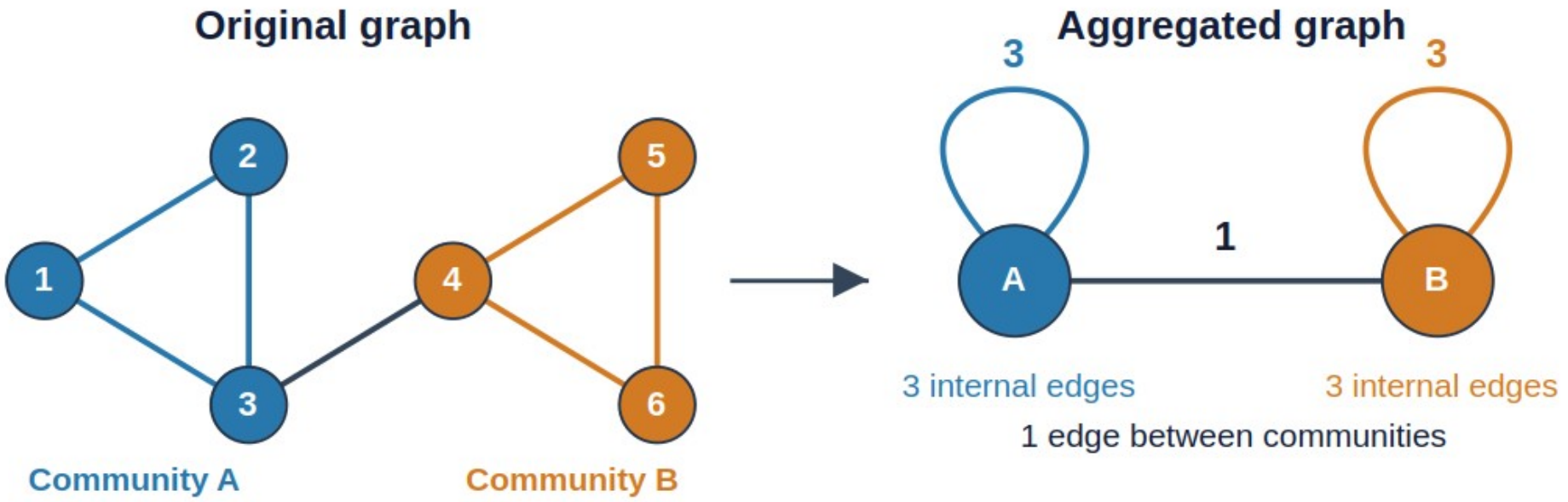
Alternate local node moves and graph aggregation to increase modularity.

Louvain: two alternating phases



Louvain is a heuristic: it does not guarantee the global maximum of modularity!

Louvain: aggregating edge weights



Each original edge has weight 1.

Sum weights; count each original edge once.

Total edge weight

$$m = 3 + 3 + 1 = 7$$

The loop weight is counted once.

Strength of super-node A

$$s(A) = 2 \times 3 + 1 = 7$$

A loop contributes twice to strength.

The same calculation gives s(B) = 7.

Same partition, same modularity before and after aggregation

$$Q = 2 \times [\, 3/7 - (7/14)^2 \,] = 5/14 \approx 0.357$$

Community detection stability

Illustrative partitions from three runs with the same parameters

Run 1		Run 2		Run 3	
{A, B}	{C, D}	{A, B}	{C, D}	{A, B, C}	{D}

Which pairs consistently belong to the same community?

	A	B	C	D
A	1	1	1/3	0
B	1	1	1/3	0
C	1/3	1/3	1	2/3
D	0	0	2/3	1

C_{ij} = fraction of runs in which nodes i and j belong to the same community.

A and B always stay together.

C changes its community membership across runs.

Darker cells indicate more consistent co-membership.

The matrix reveals stable associations and uncertain memberships.

A single consensus partition requires an additional step.

Additionnal measures of communities

G=(V, E) with n=|V| nodes and m=|E| edges

A community S have n_S=|S| nodes, m_S internal edges, c_S external edges

- **Internal density**: ratio between the actual number of edges over the maximum number of edges.

$$d_{directed}(S)=\frac{2m_s}{n_s(n_s-1)} \qquad d_{undirected}(S)=\frac{m_s}{n_s(n_s-1)}$$

- **Triad participation ratio**: number of nodes belonging to a triad in the community.

$$TPR(S)=\frac{\left|\{u:u\in S,\{(v,w):v,w\in S,(u,v)\in E,(u,w)\in E,(v,w)\in E\}\neq\emptyset\}\right|}{n_s}$$

- **Conductance**: proportion of edges that are directed to other communities.

$$C(S)=\frac{c_s}{2m_s+c_s}$$

- **Modularity**: number of internal edges to the groups, compared with a graph model with random edges

$$M(S)=\frac{1}{4}(m_s-E(m_s))$$