

Advanced Human Machine Interaction

Interaction Data Analysis

Social Network Analysis

Alexandre Pauchet

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



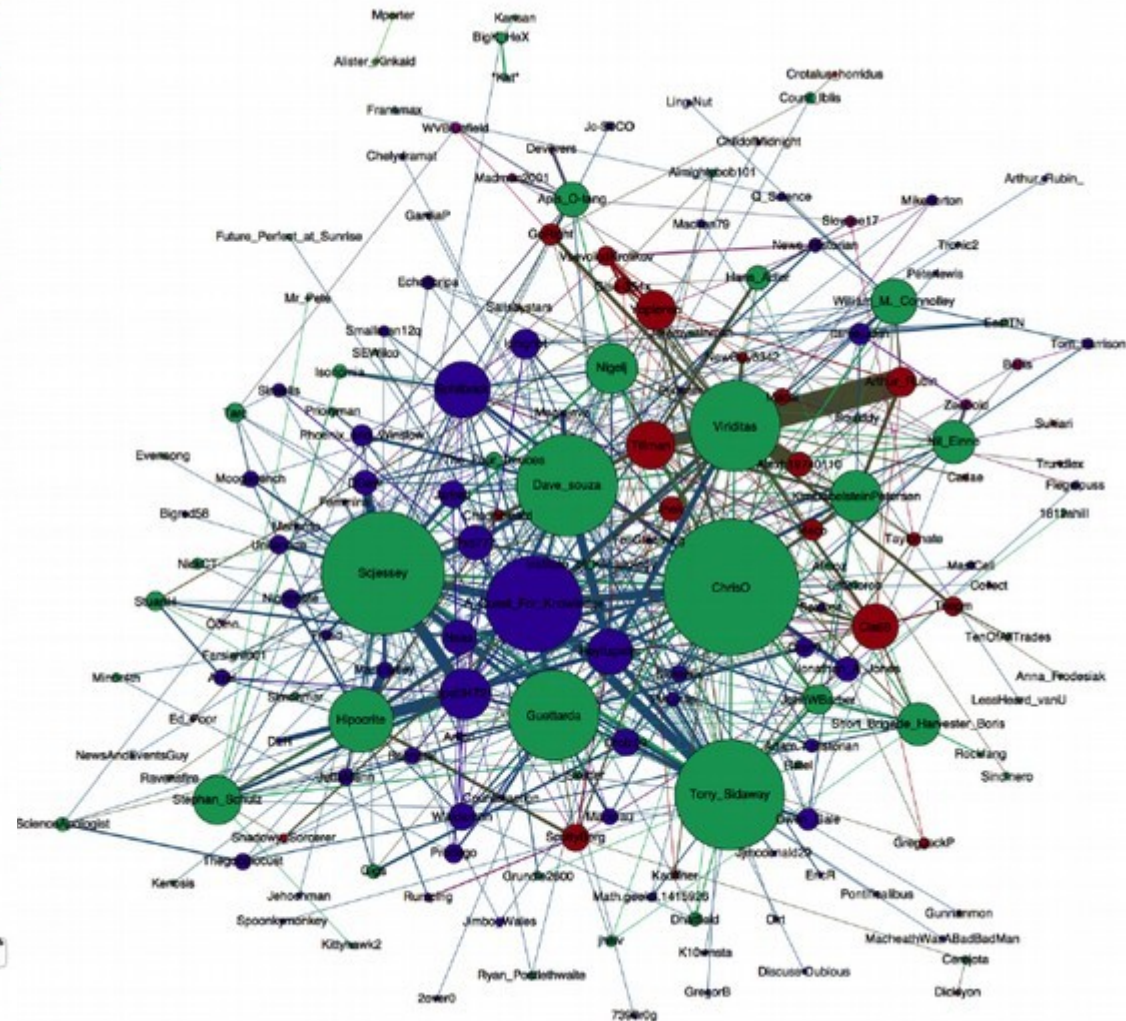
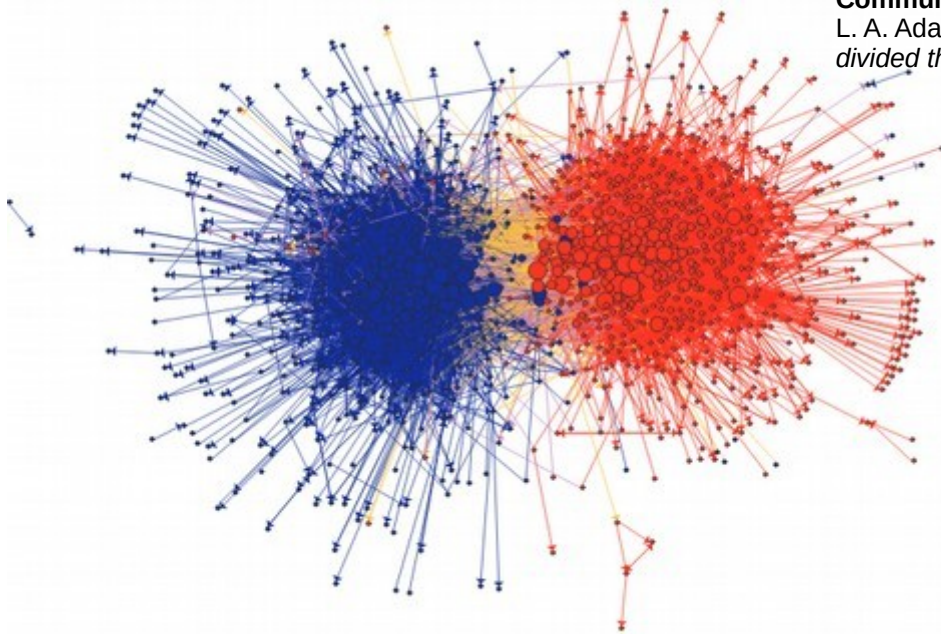
Normandie Université



Examples

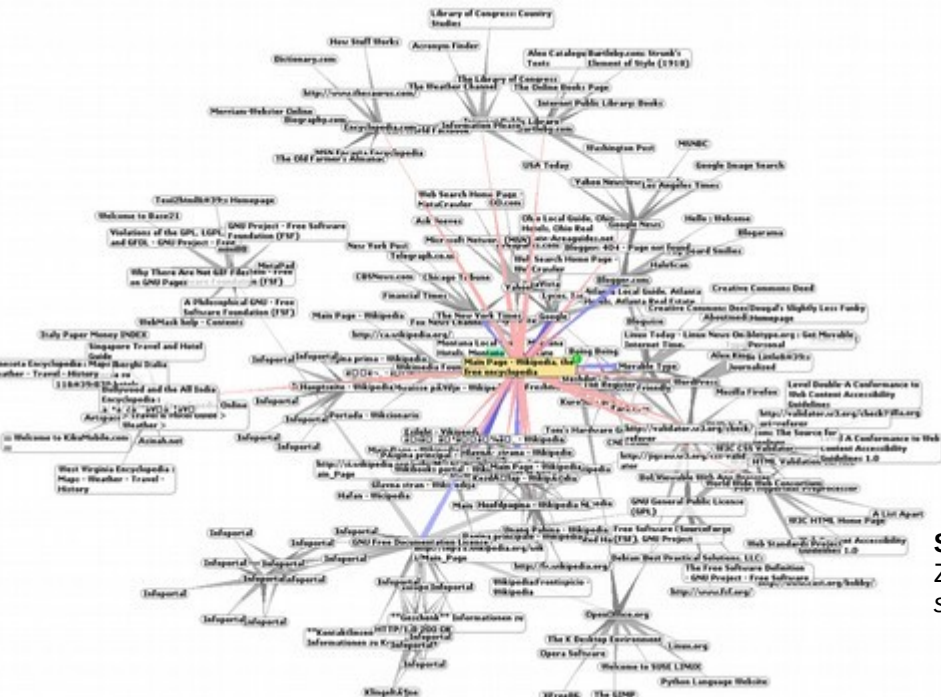
Community structure of political US blogs.

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.



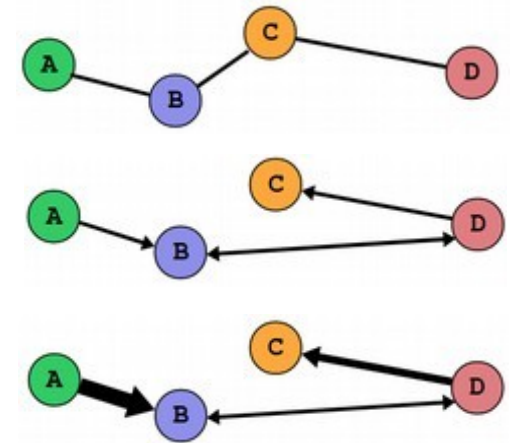
Process

1) Dataset collection

- Anonymization and processing

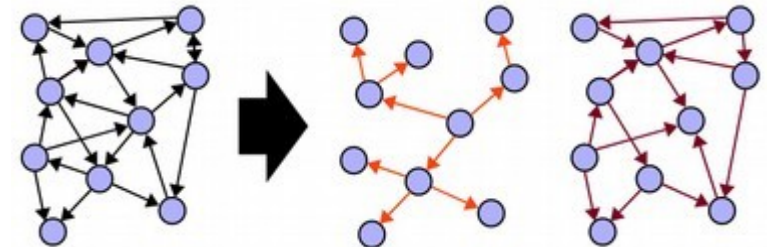
2) Graph construction

- Nodes (vertices)? Links (edges)?
- Directed or undirected graph ?
- Weighted graph ?



3) Filtering

- Node/edge removing
- Network decomposition



4) Community detection

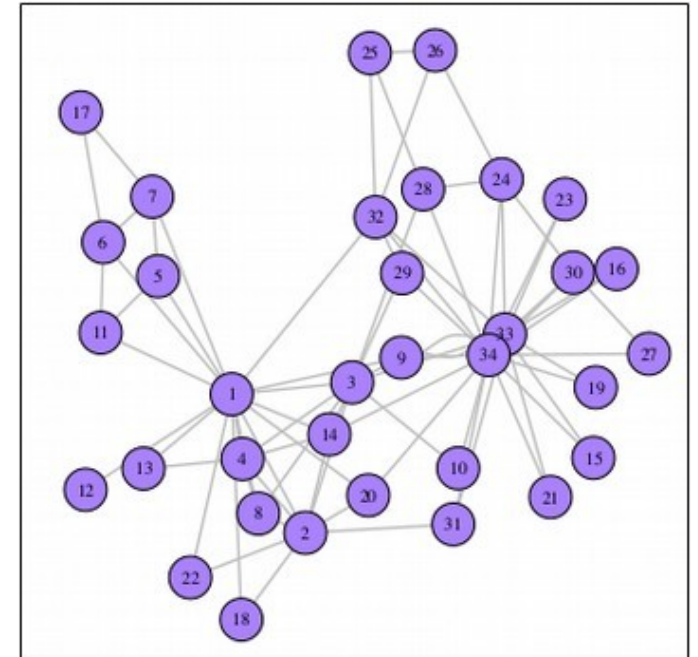
- Objective function?

Graph: definition

A graph G can be represented with:

$$G(V, E, \Psi, v, \omega)$$

- V is the set of nodes (vertices);
- E is the set of edges (links);
- Ψ is an incidence function $\Psi:E \rightarrow V \times V$;
- v is the node-labelling function;
- ω is the edge-labelling function;



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).

Graph construction

Graph construction



Can we study their interactions as a network?

Communication

Anne: Jim, tell the Murrays they're invited

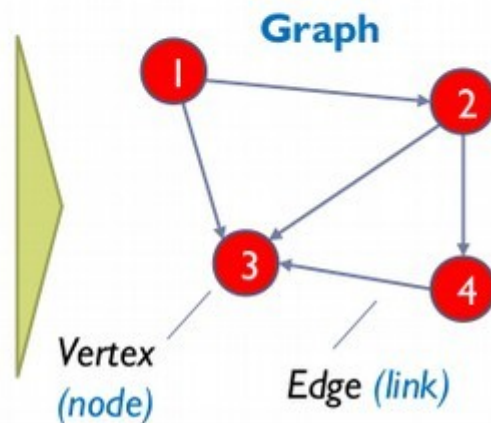
Jim: Mary, you and your dad should come for dinner!

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?

...



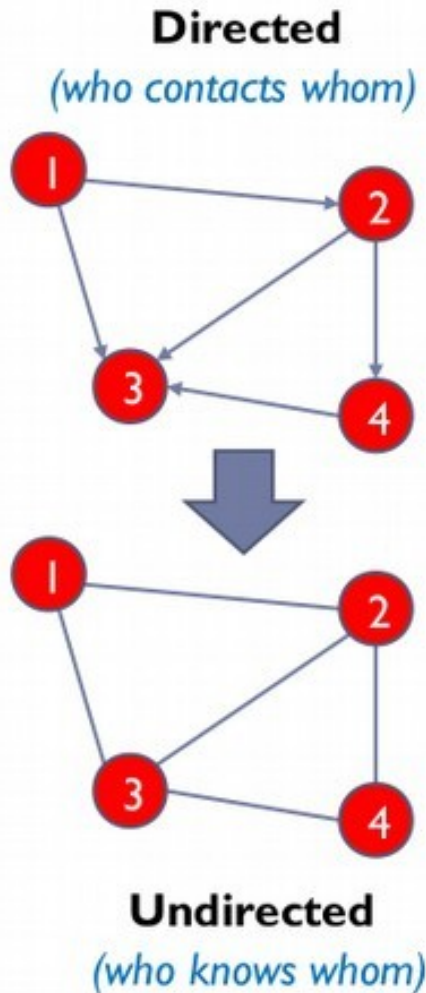
Edge list

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

Adjacency matrix

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

Directed *versus* undirected graphs



Edge list

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

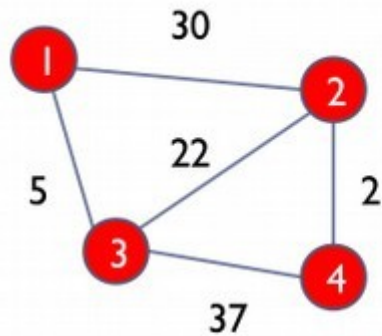
Adjacency matrix becomes symmetric

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

Quizz

- When are edge lists different between directed graphs and undirected graphs?

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	-

Quizz

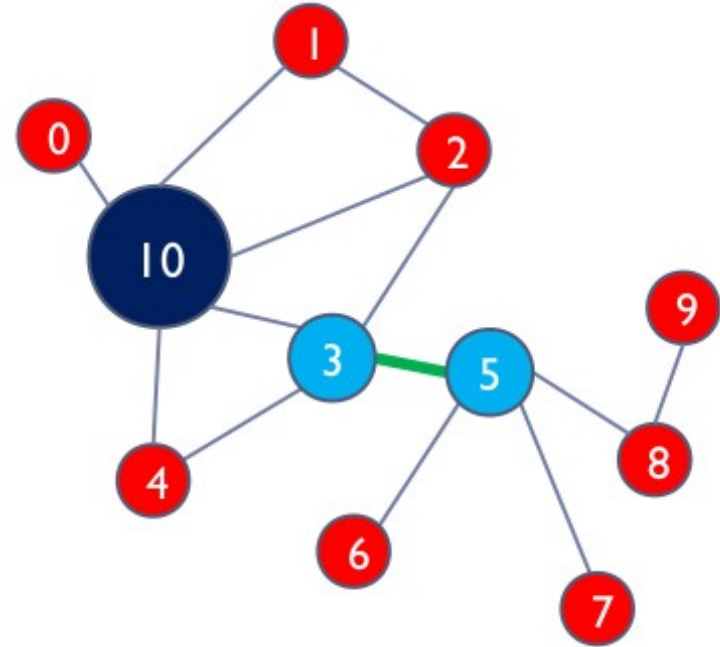
Construct the interaction graph of these messages

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

Identification of Key-users Roles in Social Networks

Identifying key users

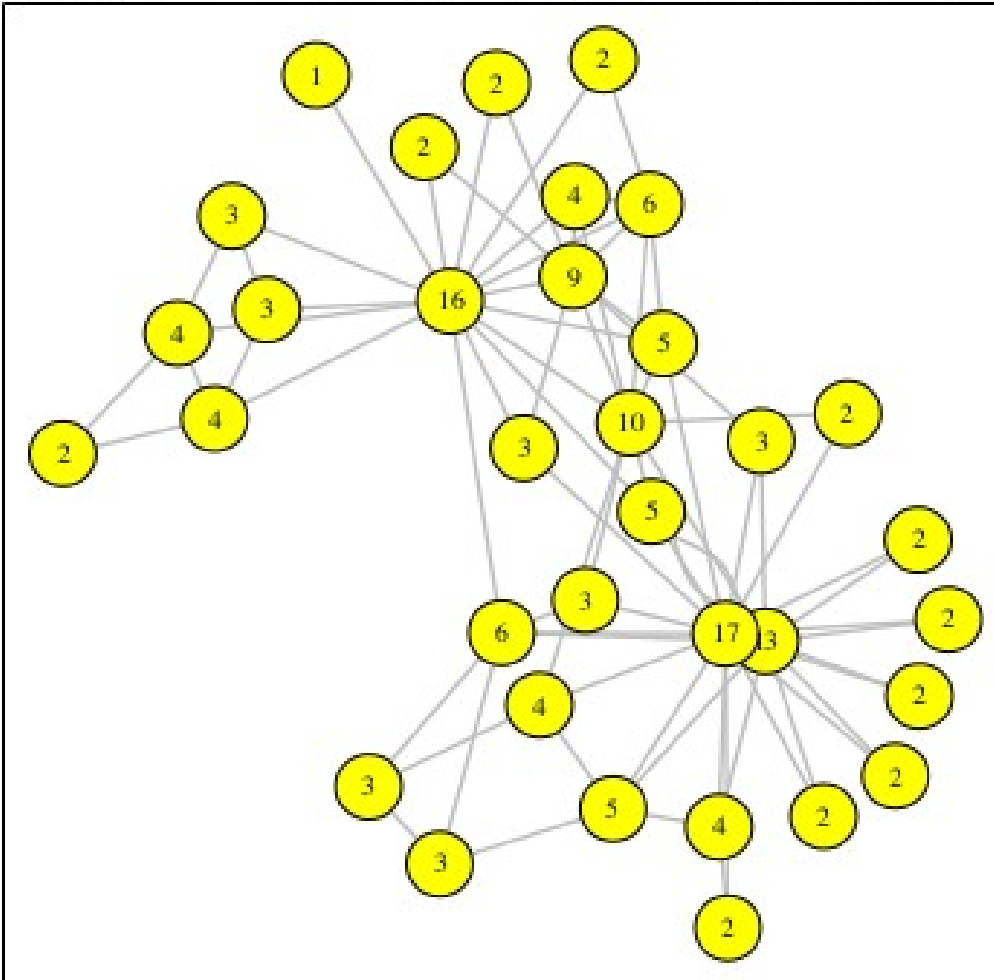
- In this network, node 10 seems the most important according to Degree
- Nodes 3 and 5 together reach more nodes with paths of length 2
- The edge between 3 and 5 is critical: if removed, the network is broken into two isolated sub-networks
- All in all, **nodes 3 and 5** seems more 'important' than 10



(Degree) Centrality
is the number of links that lead into and out of a node

Degree centrality

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



- A node (**in-**) or (**out-**)**degree** is the number of (weighted) links that lead into or out of the node
- In an undirected graph: $d_{in} = d_{out}$
- Often used as measure of a node connectedness and hence of influence and/or popularity
- Useful for assessing which nodes are central in terms of information spread and influence upon others in their immediate 'neighborhood'

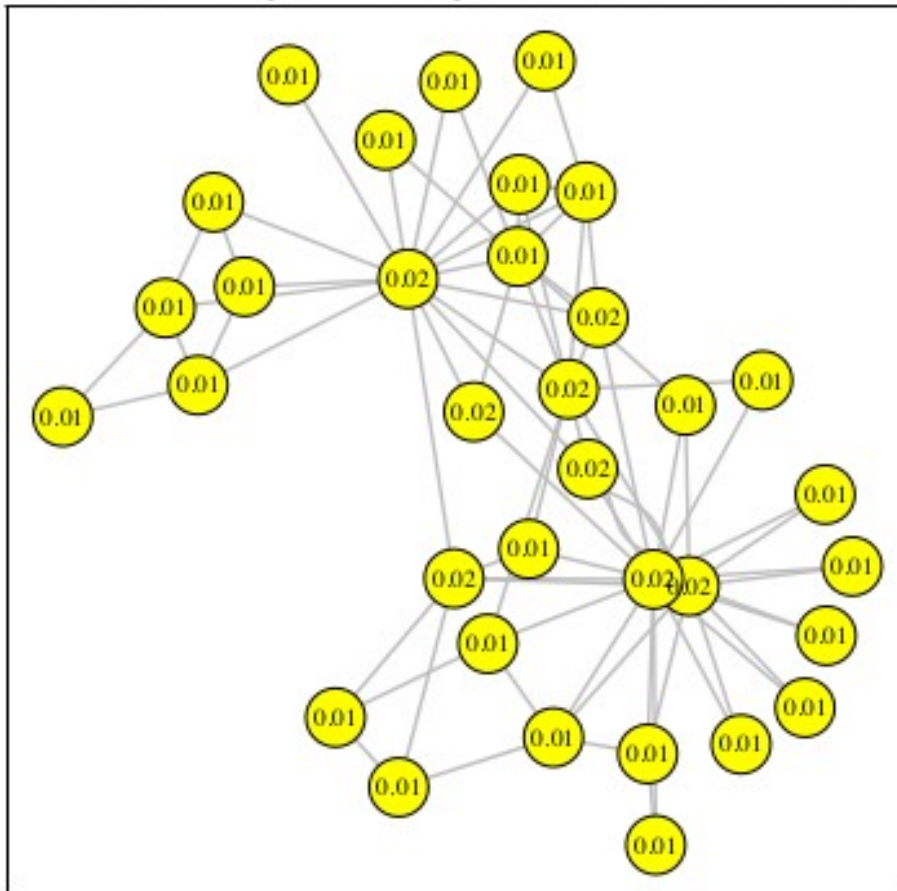
Shortest paths and closeness centrality

- **Shortest paths:**

- A *path* between 2 nodes is a sequence of non-repeating edges connecting those 2 nodes
- The *shortest path* between 2 nodes is the path that connects those 2 nodes with the shortest number of edges (also called the distance between the nodes).

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

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



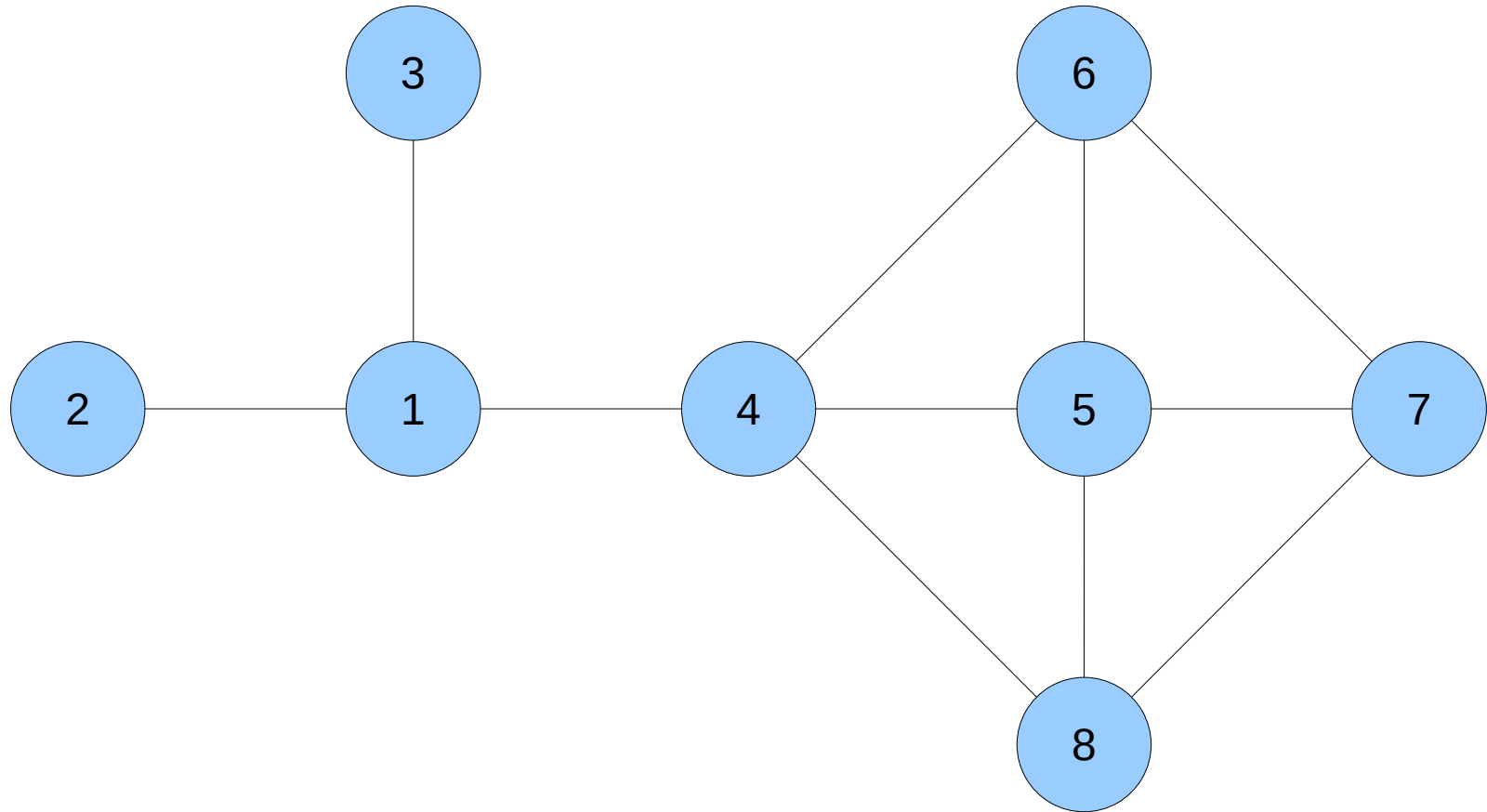
- **Closeness centrality:**

- Average length of the shortest path between a node and all other nodes in the graph. The more central a node is, the closer it is to all other nodes.
- It is a measure of the speed with which information can reach other nodes from a given starting node.

$$C(x) = \frac{\sum_y d(x, y)}{N}$$

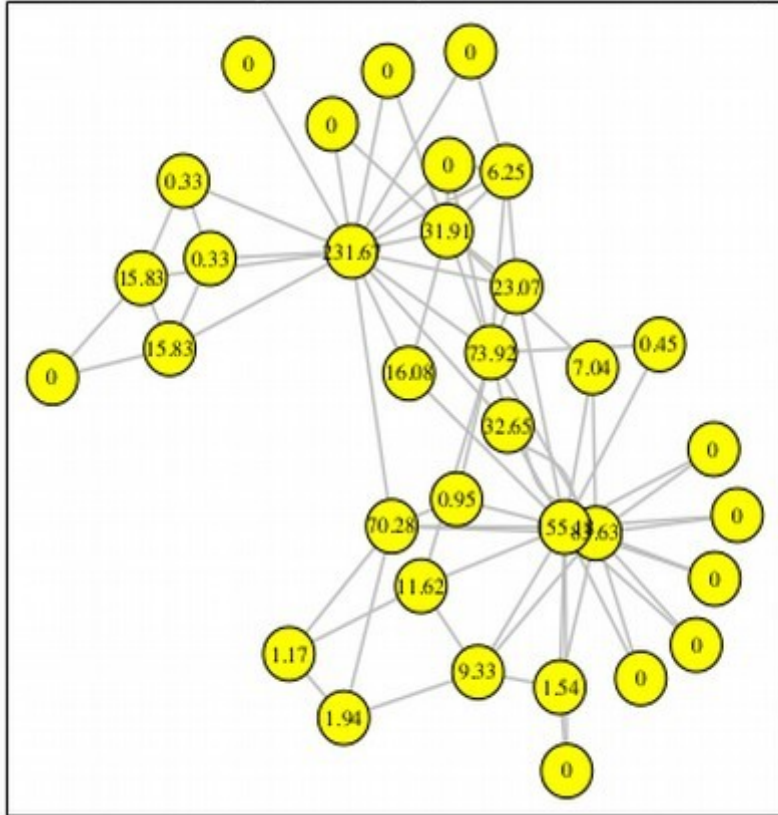
Quizz

Identify the key-users according to degree centrality and closeness



Betweenness centrality

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



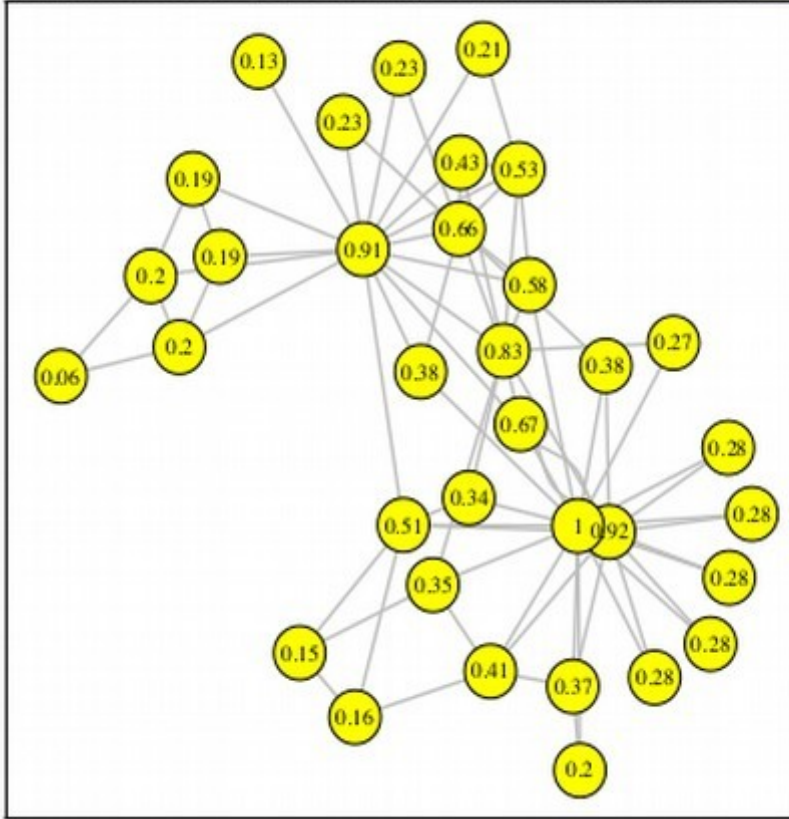
- **Betweenness:**
 - Number of times a node acts as a bridge along the shortest path between 2 nodes
 - Shows which nodes are:
 - more likely to be in communication paths between other nodes
 - breaking points of a network

$$C_B(v) = \frac{\sum_{s \neq v \neq t \in V} \sigma_{st}(v)}{\sigma_{st}}$$

- 1) For each pair of nodes, compute the shortest paths between them;
- 2) For each pair of nodes, determine the fraction of shortest paths that pass through this very node (here, node/vertex v);
- 3) Sum these fractions over all pairs of nodes.

Eigenvector centrality (or eigencentrality)

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



- **Eigenvector centrality:**
 - Eigencentrality is a measure of the influence of a node in a network; a node with a high eigencentrality is connected to other nodes with high eigenvector centrality
 - This is similar to how Google ranks web pages: links from highly linked-to pages count more (pagerank)
 - Useful in determining who is connected to the most connected nodes

Interpretations

Centrality measure

Interpretation in social networks

► Degree

How many people can this person reach directly?

► Betweenness

How likely is this person to be the most direct route between two people in the network?

► Closeness

How fast can this person reach everyone in the network?

► Eigenvector

How well is this person connected to other well-connected people?

► Degree

In network of music collaborations: how many people has this person collaborated with?

► Betweenness

In network of spies: who is the spy though whom most of the confidential information is likely to flow?

► Closeness

In network of sexual relations: how fast will an STD spread from this person to the rest of the network?

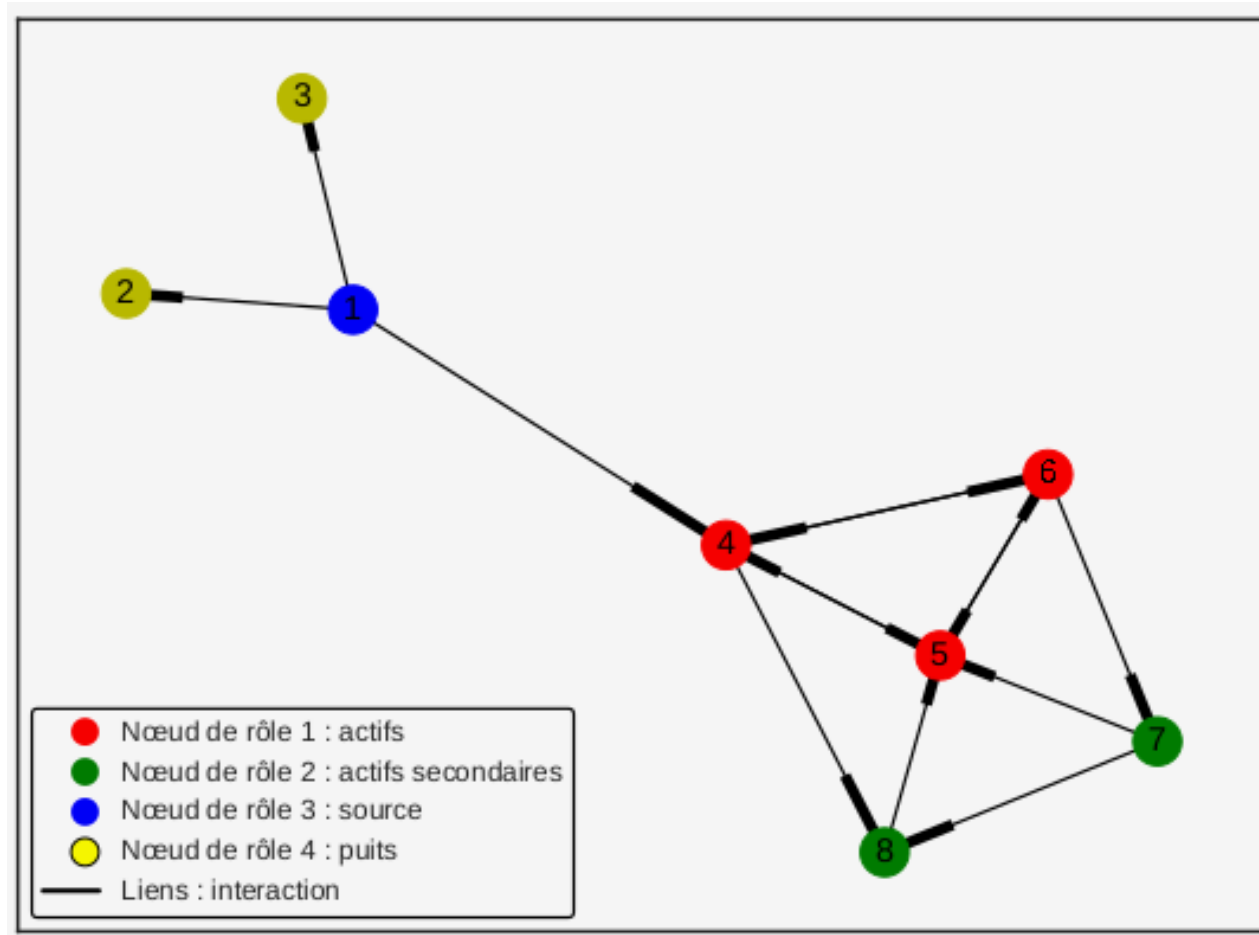
► Eigenvector

In network of paper citations: who is the author that is most cited by other well-cited authors?

RolX [Henderson et al., 2012]

- Topological measures (degrees, centralities, ...)
- Selection of characteristics
- Non-negative factorization to extract characteristic roles

NB: the number of roles must be fixed

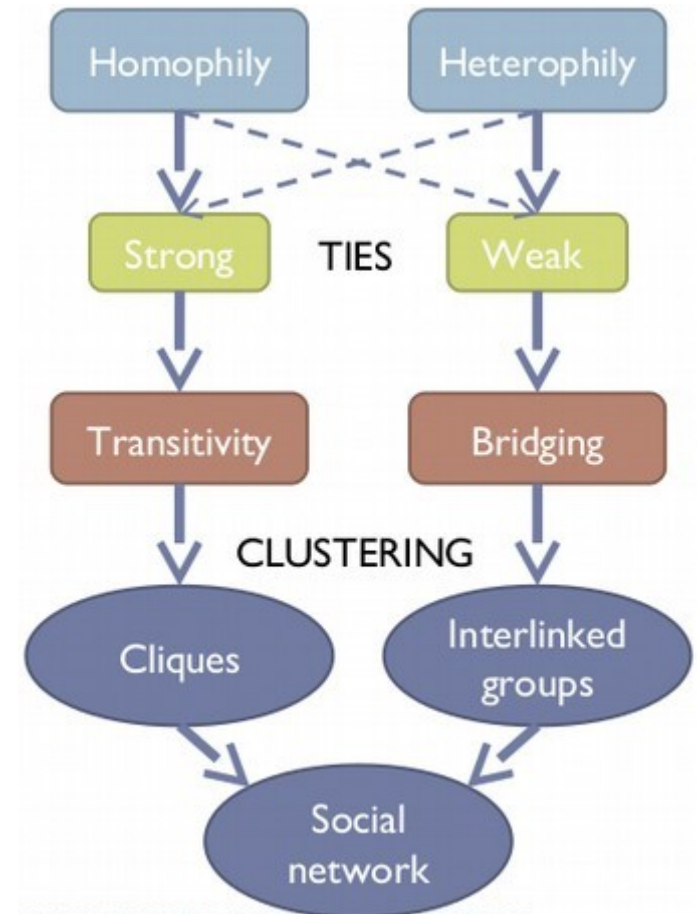


Groups of users in Social Networks

Community Detection

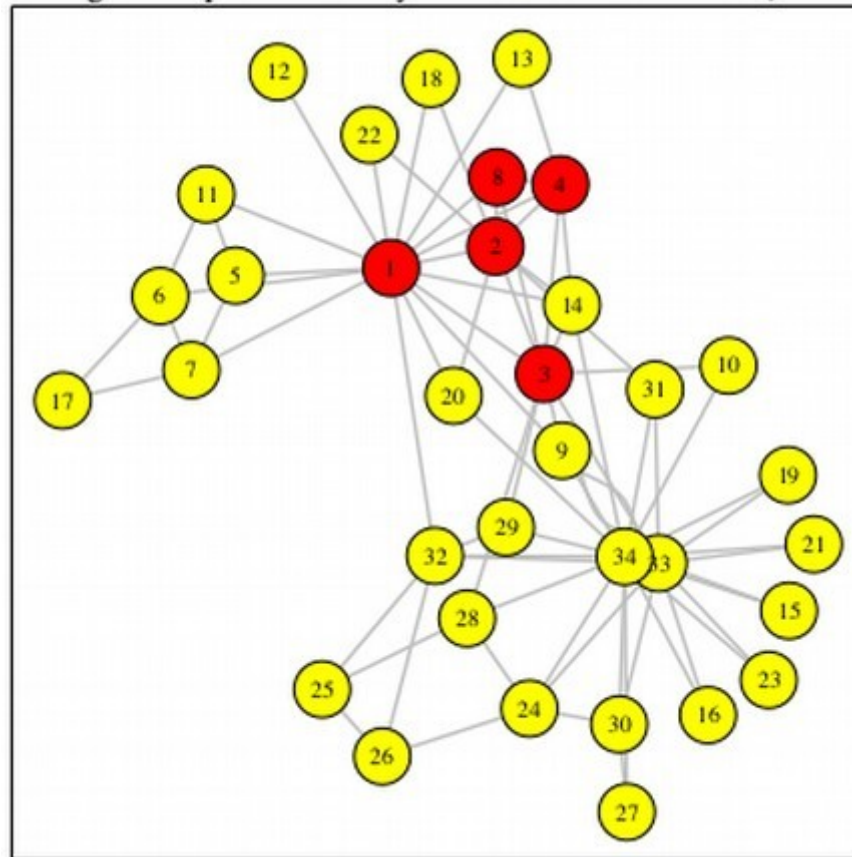
Homophily, transitivity, and bridging

- **Homophily** is the tendency to relate to people with similar characteristics (status, beliefs, etc.)
 - It leads to the formation of homogeneous groups (clusters) where relations are easier
 - Homophilous ties can be strong or weak
- **Transitivity** in SNA is a property of ties: if there is a tie between A and B and one between B and C, then in a transitive network A and C will also be connected
 - Strong ties are more often transitive than weak ties; transitivity is a hint of strong ties (but not a necessary or sufficient condition)
 - Transitivity and homophily together lead to the formation of cliques (fully connected clusters)
- **Bridges** are nodes and edges that connect across groups
 - Facilitate inter-group communication, increase social cohesion, and help stimulating innovation
 - They are usually weak ties, but not every weak tie is a bridge

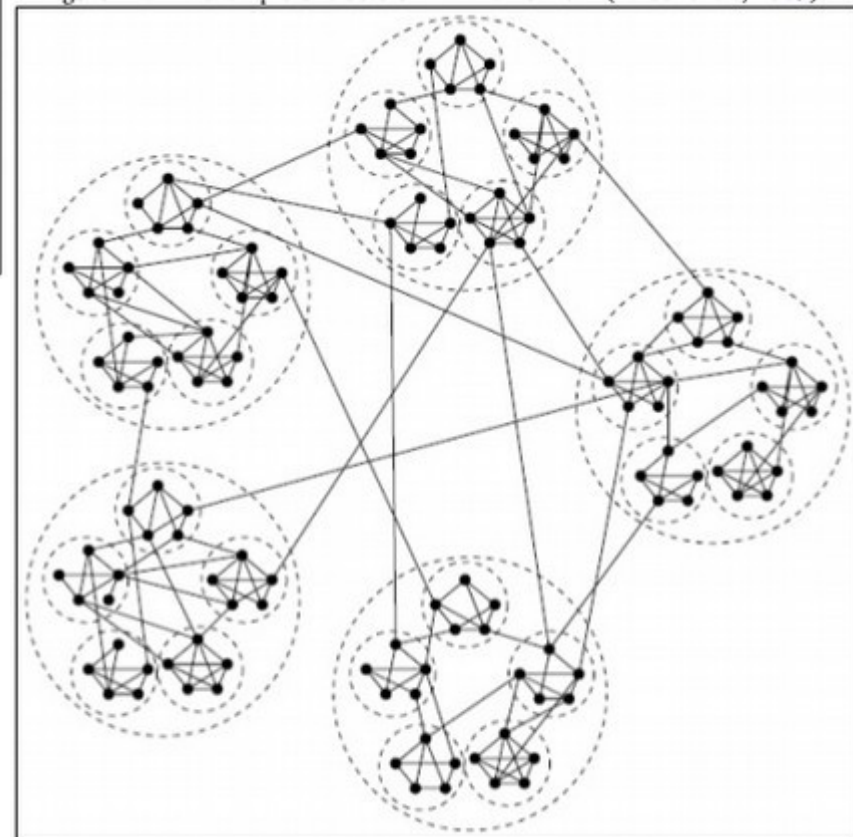
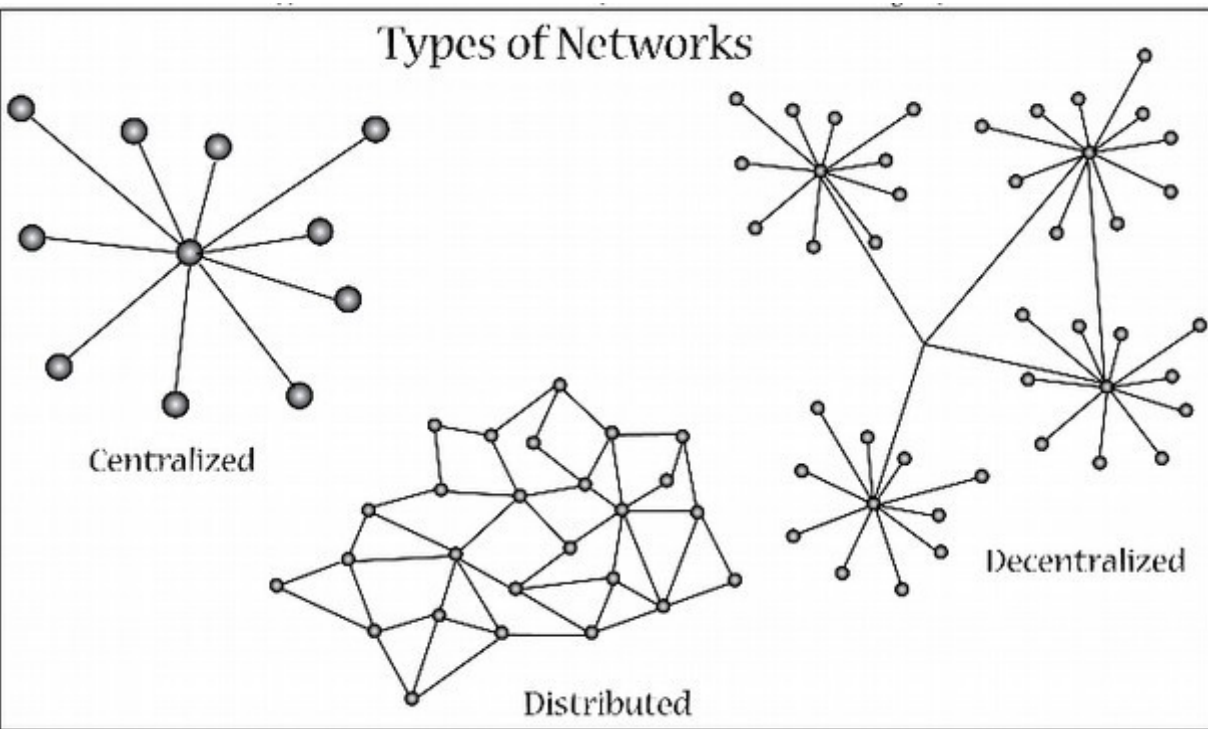


Strong homophily: cliques of users

A **clique** is a subset of nodes in a network such that every two nodes in the subset are connected by a tie. A **k-clique** is a clique of degree k .



Types of networks

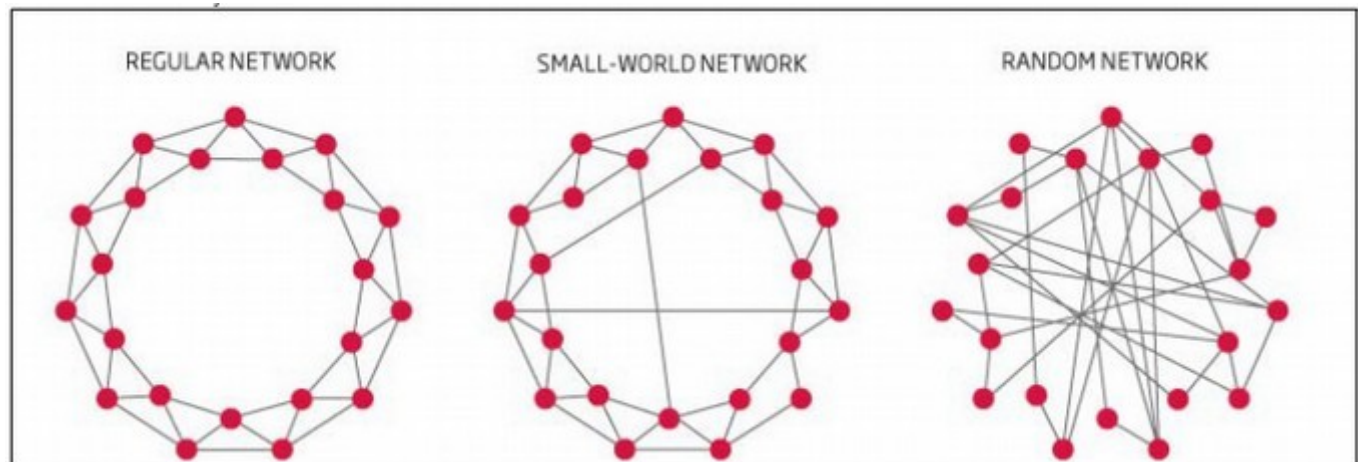
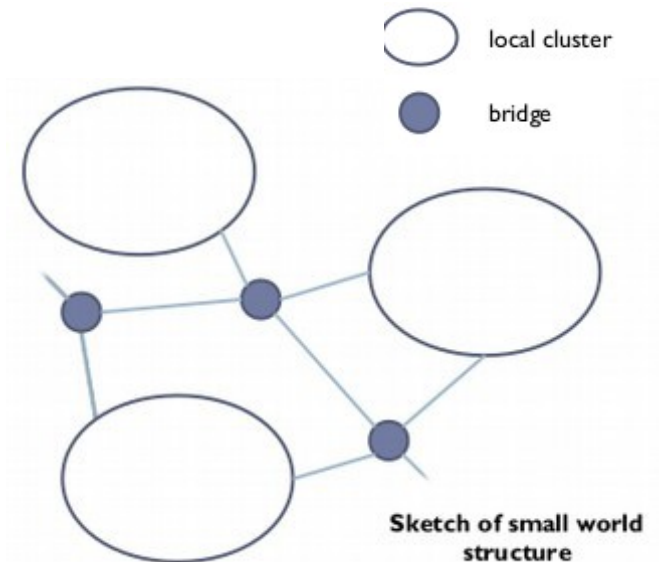


Small words: communities in SN

- A **small world** is a network that looks almost random but exhibits a significantly high clustering coefficient (nodes tend to cluster locally) and a relatively short average path length (nodes can be reached in a few steps)

- It is a very common structure in social networks because of transitivity in strong social ties and the ability of weak ties to reach across clusters

- Such a network has many clusters but also many bridges between clusters that help shorten the average distance between nodes



Community detection algorithms

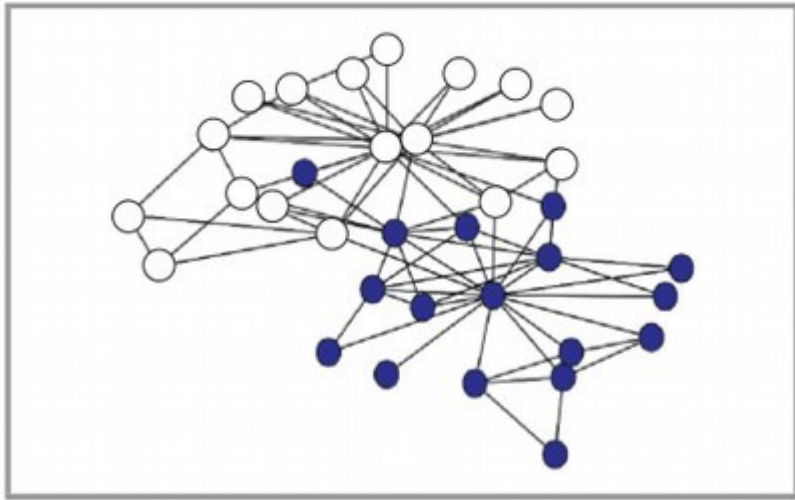


Fig. 1. Detected communities (Clauset&Newman algorithm).

FastGreedy

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

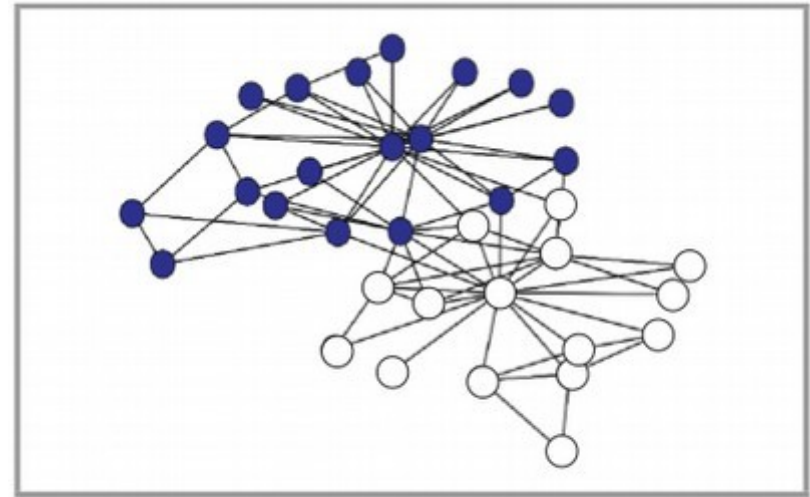


Fig. 2. Detected communities (MLC algorithm).

Markov CLustering for graphs (MCL)

Van Dongen, S. M., “Graph clustering by flow simulation”, PhD thesis, 2000.

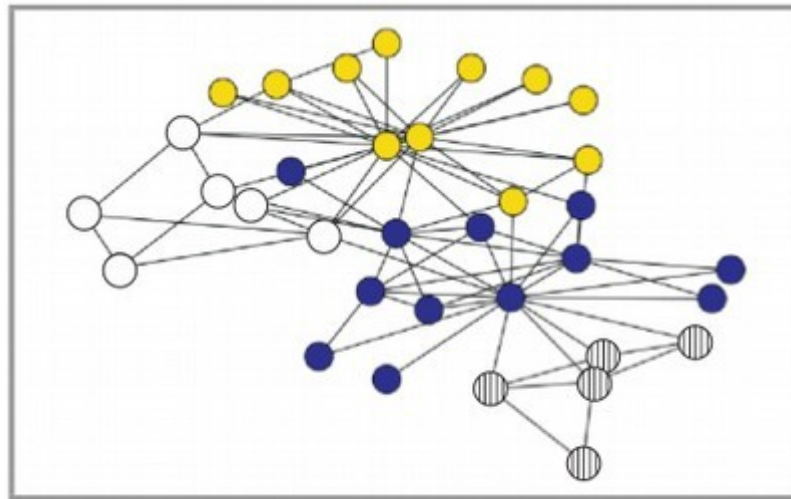


Fig. 3. Detected communities (Blondel algorithm).

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.

Topological 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.

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

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

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

- **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

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