Academic Year 2023/2024

Social Media Analytics (SMA)

Metrics for Social Network Analysis

Marco Viviani

University of Milano-Bicocca

Department of Informatics, Systems, and Communication

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

Classification of metrics

- Three main "families" **of metrics**
  - Connection
    - They have to do with the ways in which social network entities connect with each other.
  - Distribution
    - They have to do with the way in which information can flow within a social network.
  - Segmentation
    - They have to do with the ways of "clustering" the components of the social network.

Connection metrics

Connection metrics (1)

- **Homophily** (*Omofilia*): the extent to which actors form ties with similar versus dissimilar actors.
  - Similarity can be defined by gender, ethnicity, age, occupation, academic performance, social status, values, or any other salient characteristic;
  - Homophily is related to the concept of assortativity → Next lectures.
- **Multiplexity**: the number of relationship levels contained in a link
  - For example, two people who are friends and work together would have a multiplexity of 2.
  - Multiplexity can be associated with the strength of the tie.
- **Mutuality/reciprocity** (*Reciprocità*): the extent to which two actors mutually exchange friendship or other interaction.

Connection metrics (2)

- **Network closure** (*Chiusura di rete*): it measures the completeness of relational triads.
  - You can use various clustering coefficients to measure network closure.

- **Proximity/Propinquity** (*Prossimità*): the tendency for actors to have more links with others who are geographically close.

Homophily

- **Homophily**: from the Greek *ὁμοῦ* (homou, "together") and *φιλία* (philia, "friendship") is the tendency of individuals to associate and tie with other similar ones.

- Individuals in homophilic relationships share common characteristics (beliefs, values, upbringing, etc.) that facilitate communication and relationship formation.

- The concept opposite to homophily is heterophily.

Homophily in social media

- Social media favors the emergence of **homophilic relationships**.

- When a social media user likes or interacts with an article or post that relates to a certain idea or ideology (religious, political, etc.), social media tends to show similar posts (from the point of view of content and of the ideology treated).
  - Filtering algorithms (Information Filtering);
  - Personalization.

Positive effects of homophily

- The perception of **interpersonal similarity** improves coordination and increases the expected payoff of interactions, beyond the mere appreciation of others.

- Homophily helps people access information, innovations and widespread behaviors, and forms opinions and social norms.

- Homophily influences **diffusion patterns** on a social network in two ways:
  - Homophily affects the way a social network develops;
  - Individuals are more likely to successfully influence others when they are similar to them.

Negative effects of homophily

- Homophilic personal networks can translate into **limited social worlds**, with strong implications for how information is received and disseminated, and the attitude and form of interactions people experience.*

- Homophily can favor the division into closed communities through the so-called phenomenon of **filter bubbles** on social networking sites, where people with similar ideologies interact only with each other (also generating the phenomenon of **echo chambers**).

*McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. "Birds of a feather: Homophily in social networks." *Annual review of sociology* 27.1 (2001): 415-444

Network closure

- **Network closure** refers to the concept of **triadic closure** in social network theory.

- The triadic closure is the property that three nodes A, B and C can have, such that, if there is a strong tie between $A - B$ and $A - C$, there is a weak or strong tie between $B - C$.

Network closure (Cognitive balance)

- Mark Granovetter synthesized the theory of **cognitive balance** (*equilibrio cognitivo*).

- Cognitive balance refers to the propensity of two individuals to want to try the same things towards an entity that unites them.
  - If the triad of three individuals is not closed, then the persons connected to the same individual will want to close this triad to achieve closure in the network of relationships.

- The two most common measures to evaluate the triadic closure for a graph are:
  - The clustering coefficient (local or average);
  - Transitivity (global clustering coefficient) for that graph.

Network closure (Trust network)

- In a **trust network** (*rete di fiducia*), triadic closure is likely to develop due to the transitive property.

- If a node A trusts node B and node B trusts node C, node A will have the conditions to trust C.

Network closure (Social network)

- In a **social network**, a strong triadic closure occurs because there is a greater possibility for the nodes A and C that have B in common to meet and therefore to create a link (at least a weak link).

- While classical graph theory tends to analyze networks at a given time, the study of the principle of **triadic closure** in interaction networks can predict the development of ties within a network and show the progression of connectivity.

Proximity/Propinquity

- Proximity refers to the primarily **physical closeness** between people
  - Two people living on the same floor of a building, for example, have a higher propensity to establish relationships than those living on different floors.

- Distance measurement using geolocation information (when available).

Distribution metrics

Distribution metrics (1)

- **Bridges**: Identification of individuals whose ties (bridges) fill a structural hole, providing the only link between two individuals or clusters.
  - They allow us to evaluate the "strength" of a tie.
- **Centrality**: a group of metrics that aim to quantify the "importance" or "influence" (in a variety of senses) of a particular node (or group of nodes) within a network.
- **Density**: The percentage of effective links in a network out of the total possible number.

Distribution metrics (2)

- **Distance**: the minimum number of ties necessary to connect two particular actors (Stanley Milgram's experiment, the idea of the "six degrees of separation", the Erdős number, the Bacon number).

- **Structural holes**: absence of links between two parts of a network.
  - Finding and exploiting a structural hole can give an entrepreneur a competitive advantage.

- **Strength of the tie**: linear combination of time, emotional intensity, intimacy, reciprocity, etc.

Centrality

- In graph theory and network analysis, **centrality** indicators identify the most important vertices within a graph.

- Applications include identifying the most influential people in a social network, key infrastructure nodes on the Internet or urban networks, and disease 'super-spreaders'.

- The concept of centrality was first developed in the analysis of social networks and many of the terms used to measure centrality reflect their sociological origin.

Degree centrality (General case)

- Historically first and conceptually simpler is the **degree centrality** (*centralità di grado*), which is defined as the number of edges incident to a node.

- The degree can be interpreted in terms of the "immediate risk" of a node to catch whatever is flowing through the network (such as a virus or fake news).

Degree centrality (Directed relationships)

- In the case of a directed graph, two separate measures of degree centrality are defined, namely **in-degree** and **out-degree**.

- When ties are associated with some positive aspects such as friendship or collaboration, the in-degree is often interpreted as a form of **popularity**, and the out-degree as a propensity to follow the behavior of others (**heard**/**gregarious behavior**).

A Parenthesis: Heard Behavior

Comportamento gregario

- **Heard behavior** occurs when individuals observe the actions of all (or most) other individuals and act in a form aligned with them.

- **Main features**:
  - The network is observable;
  - Public information is available.

A Parenthesis: Herd behavior

Example: online auctions

- Individuals can observe the behavior of others by monitoring the offers that are made.
- Individuals are connected to each other via the auction platform where they can often view other buyers' profiles (and their reviews).
- It is possible to observe the people actively participating in the auctions and to “trust” them more.
- Such trust and the high number of offers received by an object as a strong signal of its value gives rise to gregarious behavior.

A Parenthesis: Herd behavior
Example: famous restaurants

- Suppose you are traveling in a metropolitan area with which you are not familiar.

- There is a restaurant A that was recommended to us by a friend.

- Once you arrive A is almost empty, while restaurant B, which is next door and serves the same cuisine, is full of customers.

- You will probably decide to try restaurant B!

Degree centrality (Formal definition)

- The **degree centrality** of a vertex v, for a given graph $G = (V, E)$ with $|V|$ vertices and $|E|$ edges, is defined as:

$$c_D(v) = d(v)$$

- In a directed graph, the **in-degree** and **out-degree** centralities of a vertex v can be indicated as:

$$c_D^{in}(v) = d^{in}(v) \qquad c_D^{out}(v) = d^{out}(v)$$

Normalized degree centrality

- The **normalized degree centrality** is obtained by dividing the degree centrality of a vertex by $n - 1$, where n is the number of vertices of the graph.

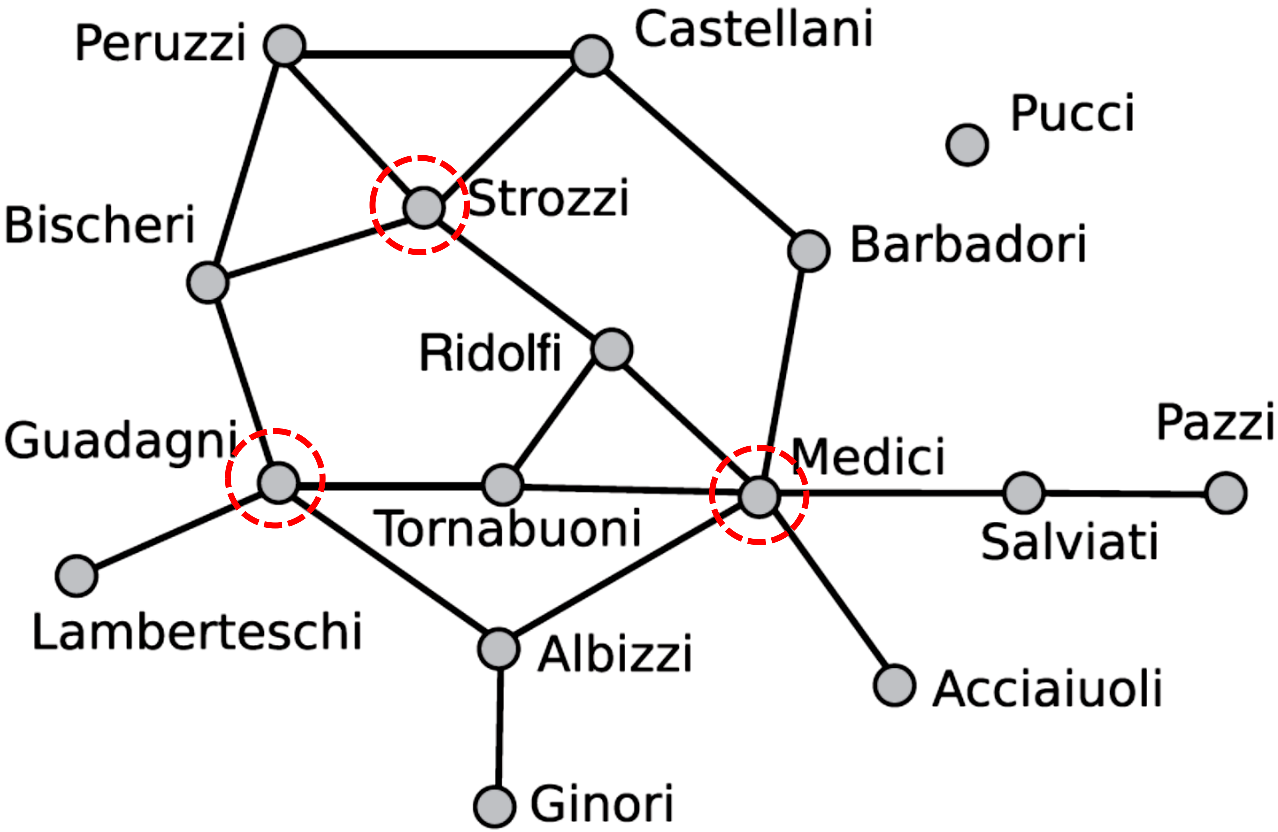
- Formally:

$$\overline{c_D}(v) = \frac{c_D(v)}{n - 1}$$

- In the same way: $\overline{c_D^{in}}(v) = \frac{c_D^{in}(v)}{n-1}$ e $\overline{c_D^{out}}(v) = \frac{c_D^{out}(v)}{n-1}$

Example

- Marriage relations between Florentine families in the Renaissance.



Degree centrality and prosperity

- Table illustrating the relationship between wealth, number of priorates and degree centrality.

Family	Wealth	Number of priorates	Node degree
Medici	103	53	6
Guadagni	8	21	4
Strozzi	146	74	4
Albizzi	36	65	3
Bischeri	44	12	3
Castellani	20	22	3
Peruzzi	49	42	3
Tornabuoni	48	..	3
Barbadori	55	..	2
Ridolfi	27	38	2
Salviati	10	35	2
Acciaiuoli	10	53	1
Ginori	32	..	1
Lamberteschi	42	0	1
Pazzi	48	..	1
Pucci	3	0	0

Centrality based on shortest (geodesic) paths
Cammini (più) brevi (Cammini geodetici)

- **Closeness centrality**
  - Based on the length of the shortest paths from a vertex.
- **Betweenness centrality**
  - Counts the number of shortest paths a vertex is part of.
- **Delta centrality**
  - Based on the concept of "performance" of the network to the removal of components.

Closeness centrality

- Based on the idea that an individual who is **closer** to other individuals than a social network is central because s/he can quickly interact with other actors.
- The simplest way to calculate this centrality is therefore to consider the sum of the distances to all the other nodes of the graph.
- A node with a low value of this sum is a node on average closest to all other nodes.

Closeness centrality (Formal definition)

- In a connected graph, the **closeness centrality** $c_C(v)$ of a node v is defined as the reciprocal of the sum of the distances from v to all the other nodes (a high value denotes closer proximity to the other nodes):

$$c_C(v) = \frac{1}{\sum_u d(v,u)}$$

- The normalized version of closeness centrality is expressed as:

$$\overline{c_C}(v) = (n-1)c_C(v)$$

Closeness centrality (Example)

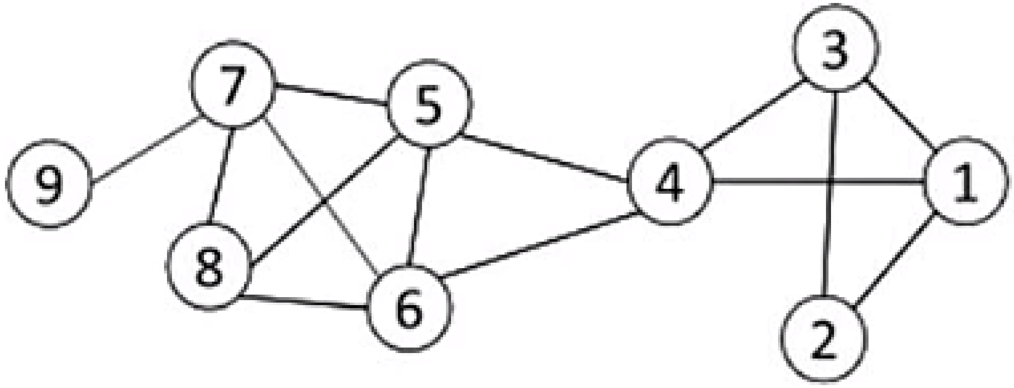


Tabella delle distanze									
Nodo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	1	1	1	2	2	3	3	4
2	1	0	1	2	3	3	4	4	5
3	1	1	0	1	2	2	3	3	4
4	1	2	1	0	1	1	2	2	3
5	2	3	2	1	0	1	1	1	2
6	2	3	2	1	1	0	1	1	2
7	3	4	3	2	1	1	0	1	1
8	3	4	3	2	1	1	1	0	2
9	4	5	4	3	2	2	1	2	0

$$c_C(v) = \frac{1}{\sum_u d(v,u)}$$

$$\overline{c_C}(v) = (n-1)c_C(v)$$

$$\overline{C_C}(3) =$$

$$\overline{C_C}(4) =$$

Closeness centrality (Example)

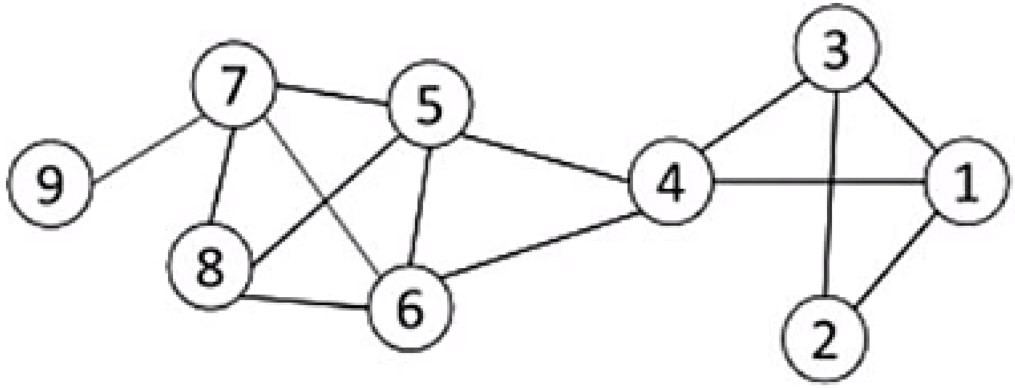


Tabella delle distanze									
Nodo	1	2	3	4	5	6	7	8	9
1	0	1	1	1	2	2	3	3	4
2	1	0	1	2	3	3	4	4	5
3	1	1	0	1	2	2	3	3	4
4	1	2	1	0	1	1	2	2	3
5	2	3	2	1	0	1	1	1	2
6	2	3	2	1	1	0	1	1	2
7	3	4	3	2	1	1	0	1	1
8	3	4	3	2	1	1	1	0	2
9	4	5	4	3	2	2	1	2	0

$$\overline{C_C}(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = 8/17 = 0.47,$$

$$\overline{C_C}(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = 8/13 = 0.62.$$

$$c_C(v) = \frac{1}{\sum_u d(v,u)}$$

$$\overline{c_C}(v) = (n-1)c_C(v)$$

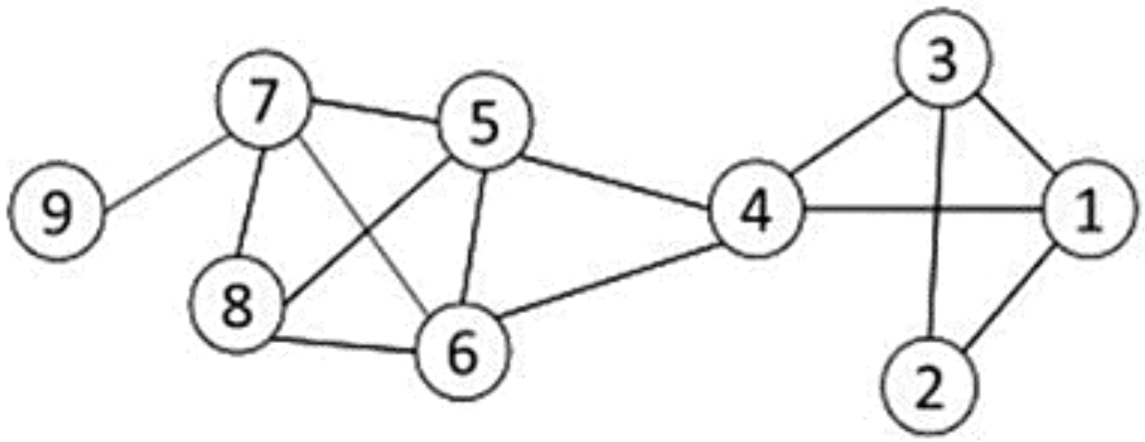
Betweenness centrality

- To calculate the **betweenness centrality** $c_B(v)$ of a node v it is necessary to calculate:
  - The number of shortest (geodesic) paths between each pair of vertices (s,t) in a graph G, denoted as: σ_{st}
  - The number of shortest (geodesic) paths between each pair (s,t) that cross the vertex v, denoted as: $\sigma_{st}(v)$

- Formally:

$$c_B(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

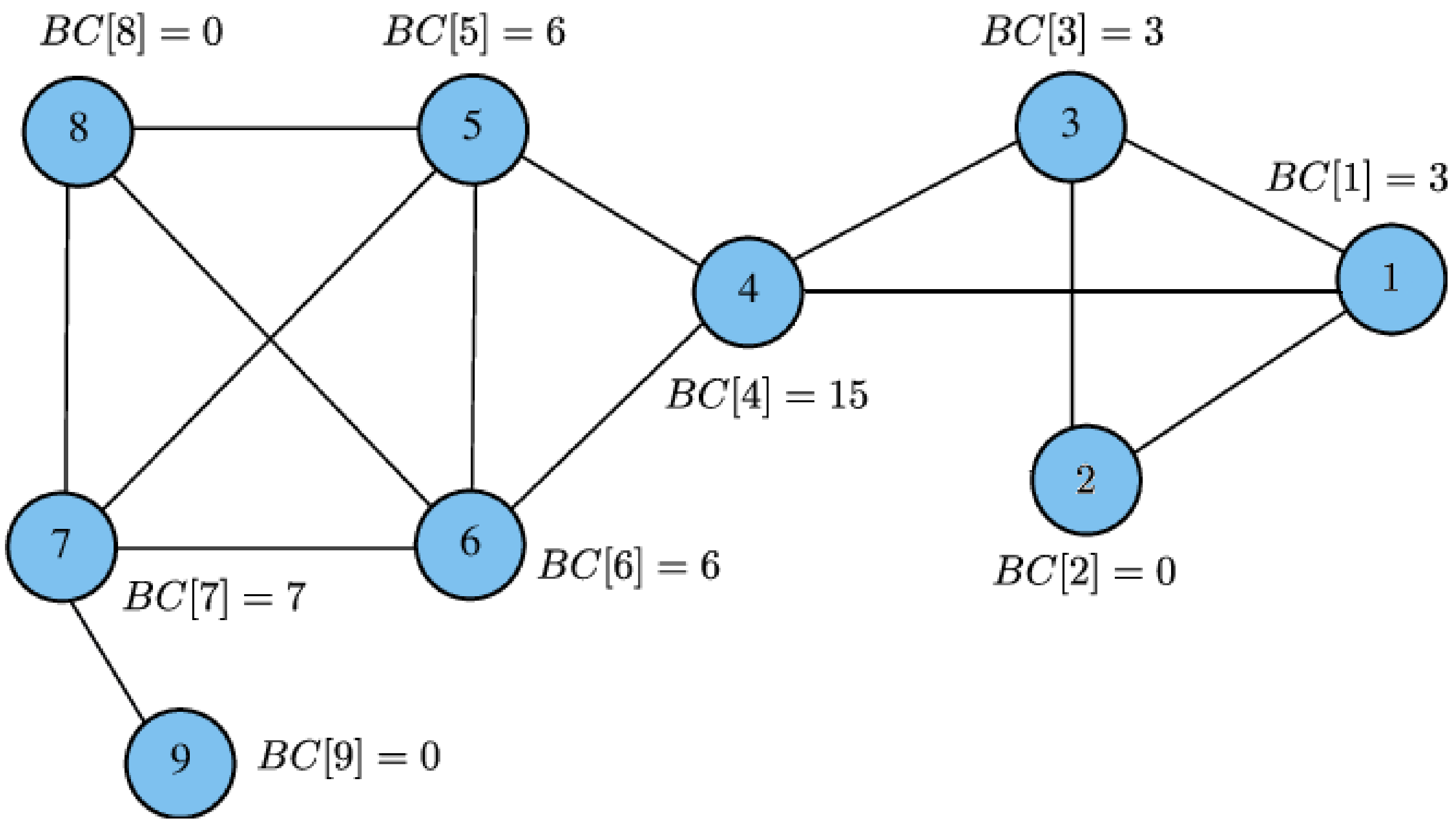
Betweenness centrality (Example)



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

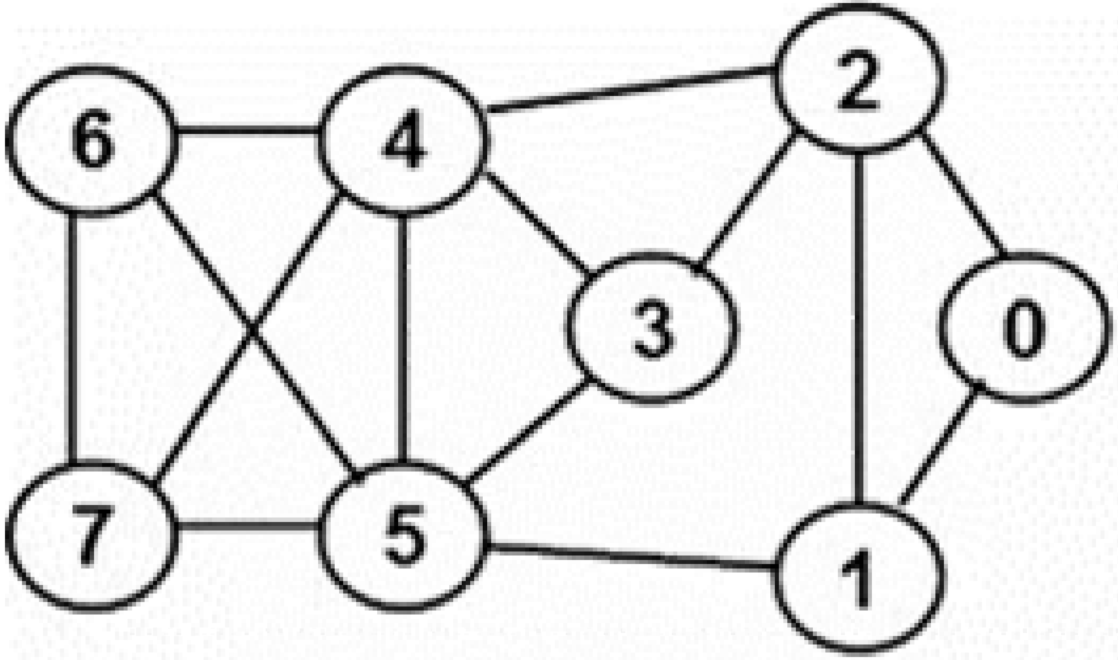
$$c_B(4) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v_i)}{\sigma_{st}} = \frac{1}{1} + \frac{1}{1} + \frac{2}{2} + \frac{2}{2} + \frac{2}{2} + \frac{2}{2} + \frac{2}{2} + \frac{4}{4} + \frac{4}{4} + \frac{4}{4} + \frac{1}{1} + \frac{1}{1} + \frac{2}{2} + \frac{2}{2} + \frac{2}{2} = 15$$

Betweenness centrality (Example)



Betweenness centrality (Exercise)

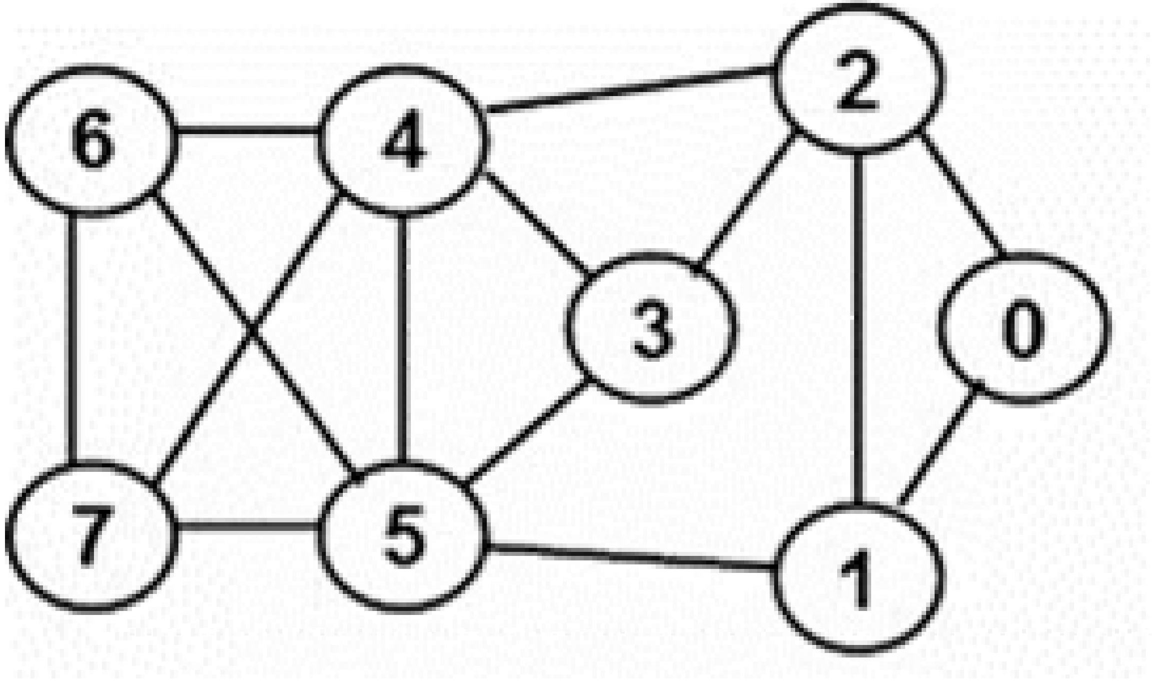
$$c_B(v_i) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$




```
betweeness for node 4
Pair (6,0) --->
Pair (7,0) --->
Pair (5,2) --->
Pair (6,2) --->
Pair (7,2) --->
Pair (6,3) --->
Pair (7,3) --->
Pair (2,5) --->
Pair (0,6) --->
Pair (2,6) --->
Pair (3,6) --->
Pair (0,7) --->
Pair (2,7) --->
Pair (3,7) --->
Betweenness of 4 :
```

Betweenness centrality (Exercise)

$$c_B(v_i) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$




```
betweeness for node 4
Pair (6,0) --->1 / 2
Pair (7,0) --->1 / 2
Pair (5,2) --->1 / 3
Pair (6,2) --->1 / 1
Pair (7,2) --->1 / 1
Pair (6,3) --->1 / 2
Pair (7,3) --->1 / 2
Pair (2,5) --->1 / 3
Pair (0,6) --->1 / 2
Pair (2,6) --->1 / 1
Pair (3,6) --->1 / 2
Pair (0,7) --->1 / 2
Pair (2,7) --->1 / 1
Pair (3,7) --->1 / 2
Betweenness of 4 : 8.66
```

Normalized betweenness centrality

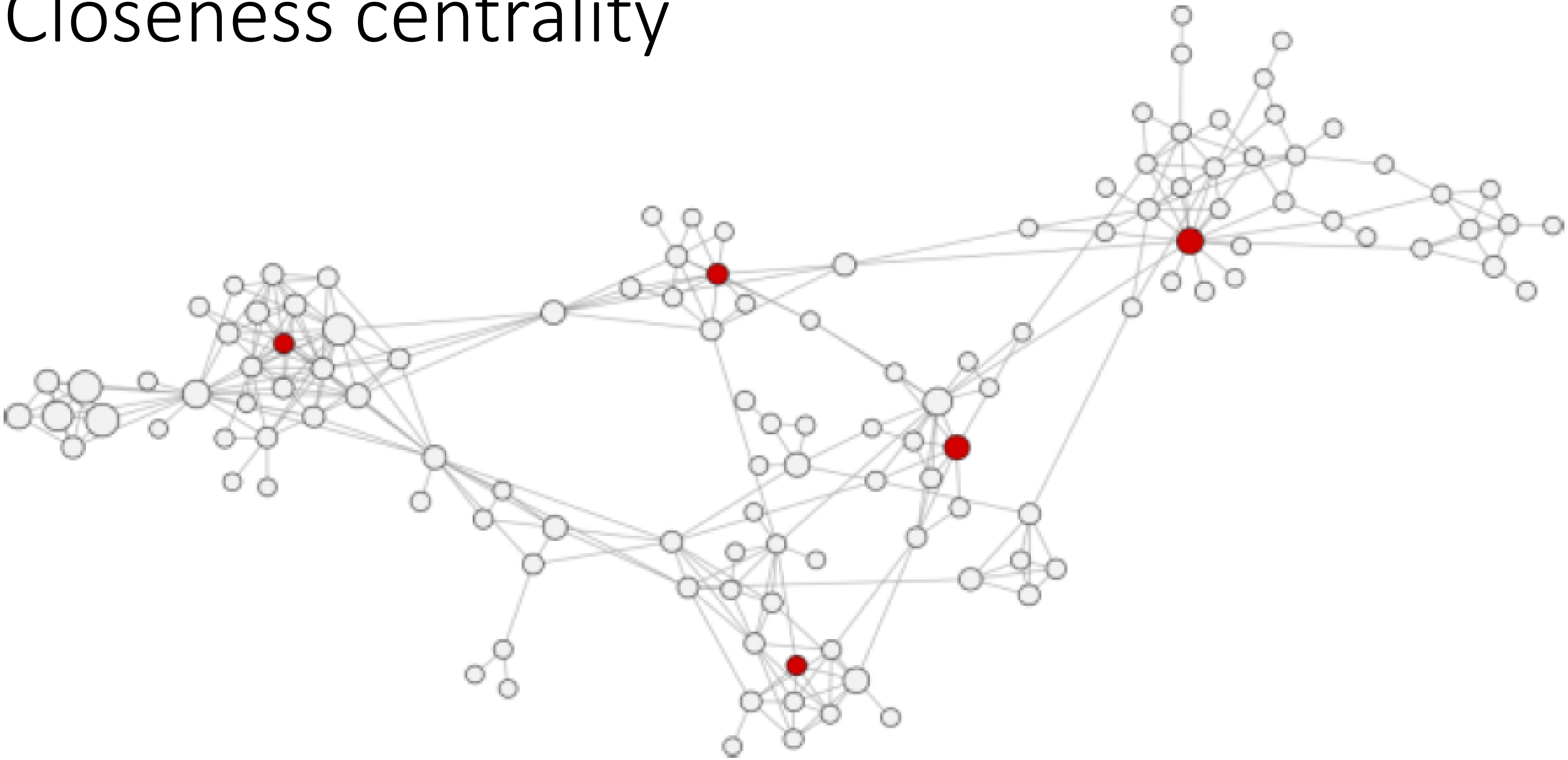
- Given the value of betweenness centrality $c_B(v)$, to calculate the **normalized** value it is necessary to divide this value by the maximum possible number of geodesics crossing node v.

- This maximum number is:
  - $(n-1)(n-2)$ for directed graphs
  - $\frac{(n-1)(n-2)}{2}$ for undirected graphs

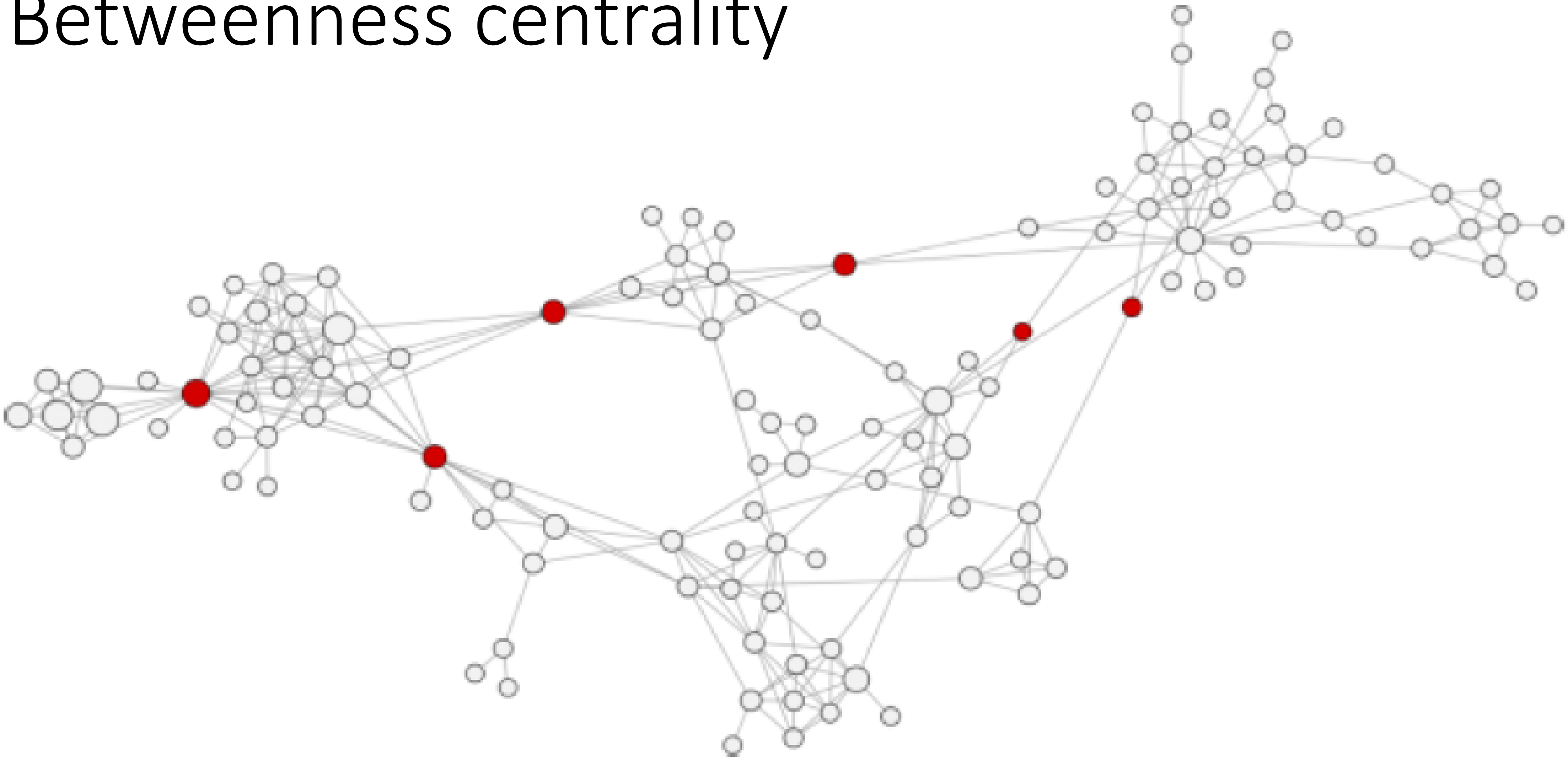
- Formally:

$$\overline{c_B}\,(v) = \frac{c_B(v)}{(n-1)(n-2)} \qquad \overline{c_B}\,(v) = \frac{2 * c_B(v)}{(n-1)(n-2)}$$

Closeness centrality



Betweenness centrality



Recap: Closeness VS betweenness centrality

- **Closeness centrality**
  - High closeness centrality indicates that a node is close to many other nodes in terms of the shortest paths, and it can efficiently communicate or spread information to other nodes.
  - Closeness centrality is particularly useful in situations where rapid information dissemination or interaction is important. For example, a node with high closeness centrality might quickly influence many other users.

- **Betweenness centrality**
  - High betweenness centrality indicates that a node acts as a bridge or intermediary between other nodes. It plays a crucial role in connecting different parts of the network.
  - Betweenness centrality is particularly useful in identifying nodes that control or mediate the flow of information or resources in a network, such as a person who acts as a bridge between two otherwise disconnected social groups.

Recap: Closeness VS betweenness centrality

- In summary, the **key difference** between closeness centrality and betweenness centrality is their focus:
  - Closeness centrality measures how quickly a node can reach all other nodes.
  - Betweenness centrality measures the extent to which a node controls the flow of information or resources between other nodes.

Delta centrality

- Idea: the importance of a node can be measured by its contribution to the "**performance**" of the entire network.
- The **delta centrality** of a node v in a graph G is calculated by comparing the performance $P(G)$ of graph G to the performance $P(G')$ of graph graph G' obtained by deactivating (removing) the node v.
- Formally:

$$c_{\Delta}(v) = \frac{\Delta P_{v_i}}{P} = \frac{P(G) - P(G')}{P(G)} \qquad \Delta P_{v_i} \geq 1, \forall v_i \in G$$

Delta centrality (The concept of “performance”)

- Fundamental: **how** to calculate $P(G)$?

- The meaning of the delta centrality depends on the way in which it is chosen to calculate $P(G)$.

- Easiest solution: $P(G) \equiv ||G||$.

Delta centrality (Performance as "efficiency")

- The **efficiency** in communication between two vertices v_i and v_j is calculated as the inverse of their distance:

$$\epsilon_{ij} = \frac{1}{d(v_i, v_j)}$$

If v_i and v_j are not connected, $d(v_i, v_j) = \infty \ \rightarrow \epsilon_{ij} = 0$

- Idea: $P(G) \equiv E(G)$

$$E(G) = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \frac{1}{d(v_i, v_j)}$$

Density

- The **density** of a graph measures the ratio of the number of actual edges in a graph to the potential number of edges in a graph.
  - A dense graph is a graph in which the number of effective edges approaches the number of potential edges.
  - A sparse graph is a graph in which the number of effective edges is much less than the number of potential edges.

Density (Directed and undirected graphs)

- For simple undirected graphs, the **density** of the graph is defined as:

$$D(G) = \frac{2|E(G)|}{n(n-1)}$$

- For simple directed graphs, the **density** of the graph is defined as:

$$D(G) = \frac{|E(G)|}{n(n-1)}$$

Segmentation metrics

Segmentation metrics

- **Counting** of:
  - Cliques, if each individual is directly related to each other individual;
  - "Social circles", groups of individuals less closely linked than in a clique.

- **Clustering** (or aggregation) **coefficient**.

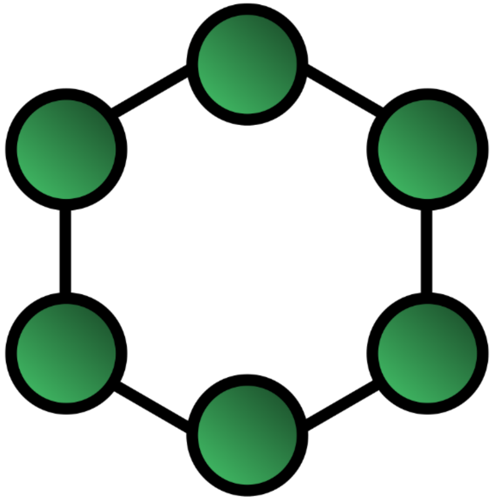
- **Cohesion**: degree to which actors are directly connected to each other by cohesive bonds.
  - Structural cohesion refers to the minimum number of members or ties that, if removed from a group, would disconnect the group.

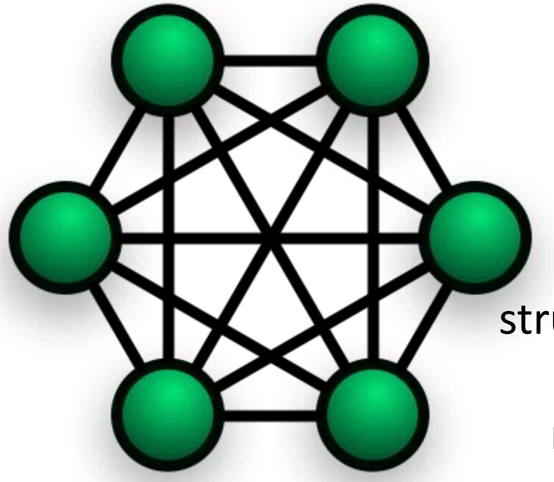
Segmentation metrics

Structural cohesion

- It is defined as the minimum number of actors or ties of a social network that must be removed to disconnect the group.
- It is therefore identical to the concept of connectivity of the nodes of a given graph.

The 6-node ring in the graph has connectivity 2 or a level 2 of structural cohesion because the removal of two nodes or two arcs is required to disconnect it.





A 6-node clique has structural cohesion equal to 5 as it disconnects with the removal of 5 arcs or nodes.

Neo4j

Neo4j Graph Platform

- The **Neo4j Graph Platform** supports transactional processing and analytical processing of graph data:
  - https://neo4j.com/

- The **Neo4j Graph Algorithms** library:
  - It includes parallel versions of algorithms supporting graph analytics and machine learning workflows:
  - https://neo4j.com/docs/graph-algorithms/current/

Useful Neo4j Graph Algorithms

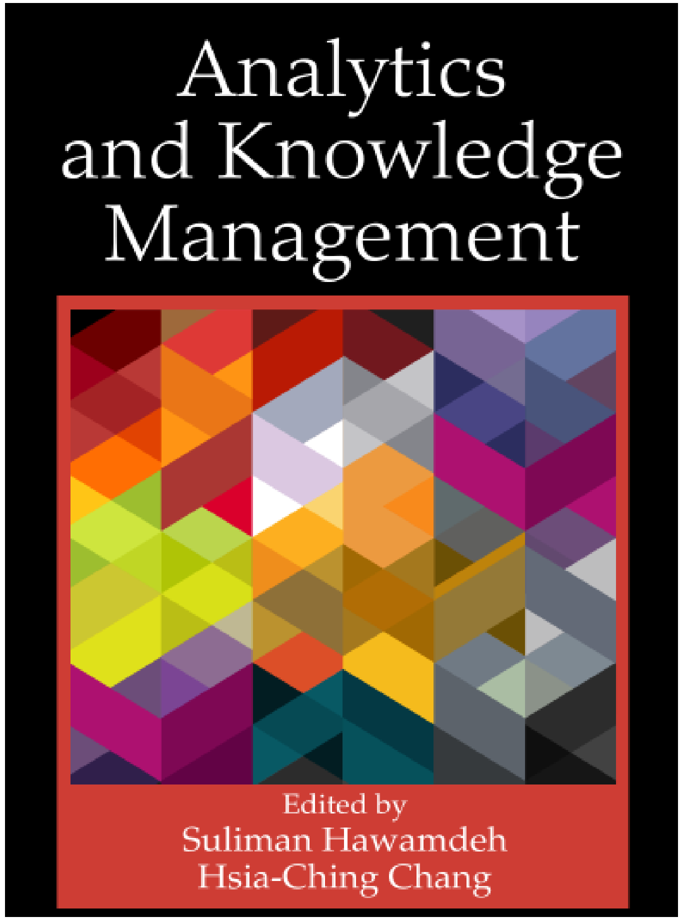
- **Shortest Path Algorithms**

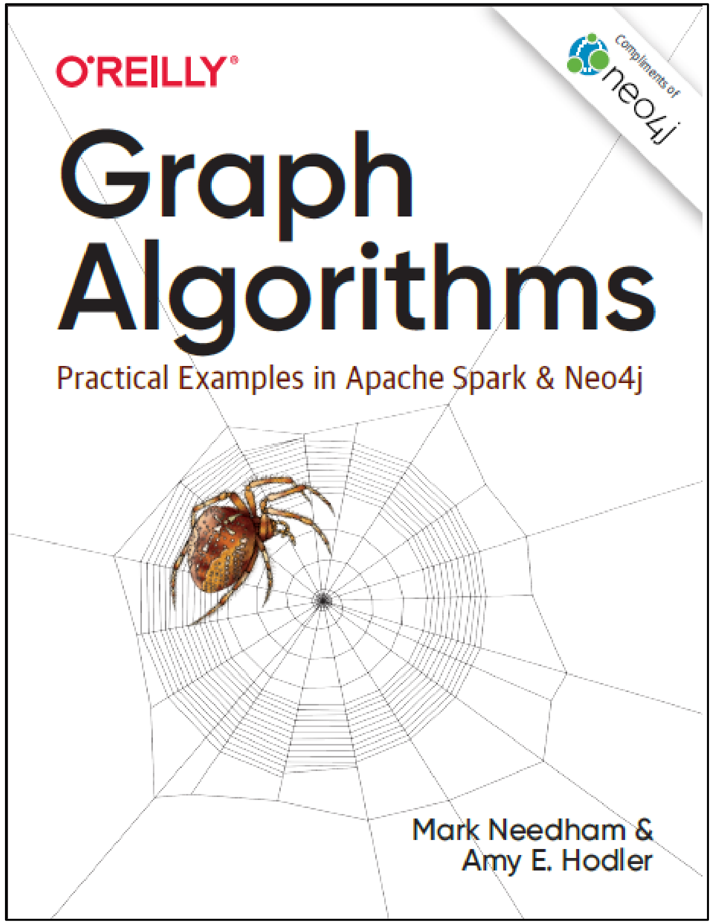
- **Centrality Algorithms**
  - Degree centrality
  - Closeness centrality
  - Betweenness centrality
  - PageRank

- **Community Detection Algorithms**
  - Clustering coefficients
  - Stongly connected components
  - Label propagation (In-depth lesson on the dissemination of information)
  - Modularity (Second part of the course)

Recommended reading

Link





Link