Massive Graph Management & Analytics

Centrality Measures

Nacéra Seghouani

Computer Science Department, CentraleSupélec
Laboratoire Interdisciplinaire des Sciences du Numérique, LISN
nacera.seghouani@centralesupelec.fr

2024-2025

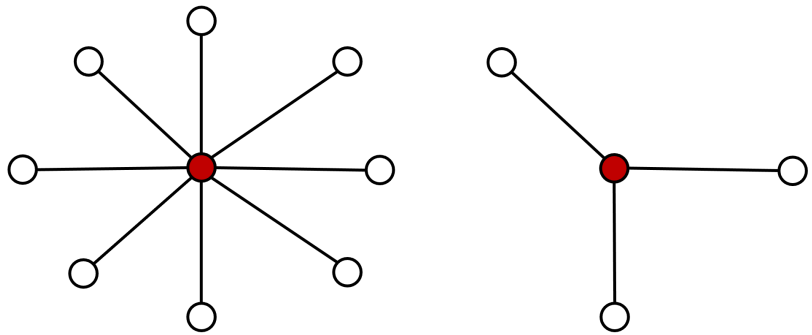
Centrality Measures

Centrality Measures

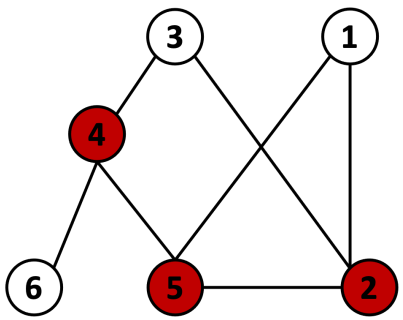
- ☞ Centrality indices are answers to the question **"What characterizes an important vertex?"** Need to define a real-valued function on the vertices of a graph, to provide a ranking which identifies **the most important nodes**.
- ☞ Centrality has a wide number of meanings, leading to different definitions of centrality:
  $\longrightarrow$ cohesiveness, ability to transfer information across the network, to affect/influence the others, to control information flow, ...
- ☞ Many centrality measures count the number of **paths (or walks)** through a given vertex; the measures differ in how the relevant walks are defined and counted: from walks of length one (degree centrality) to infinite walks (eigenvector centrality).
- ☞ Other centrality measures, such as betweenness centrality focus not just on overall connectedness but occupying **positions that are pivotal** to the network's connectivity.
  $\longrightarrow$ Influencers in social networks, vehicles for disease spreading, hubs in road networks, key infrastructures on the Internet, ...

Degree Centrality

- More neighbors (connections) a node has high potential communication activity, more important it is!



- The matrix form: $\mathbf{c} = \mathbf{A1}$ where is the all one vector



$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 3 \\ 3 \\ 1 \end{pmatrix}$$

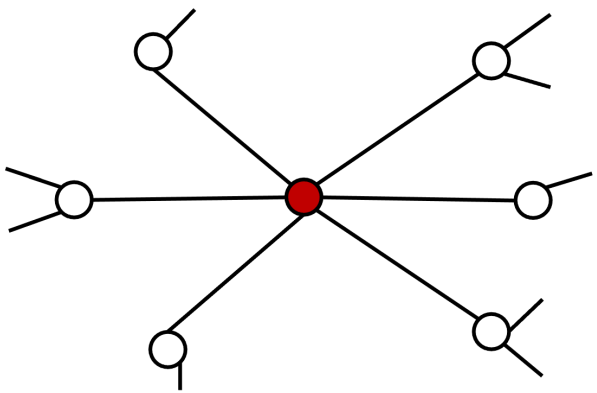
- Very likely that more than one vertex has the same degree $\rightarrow$ difficult to uniquely rank the vertices

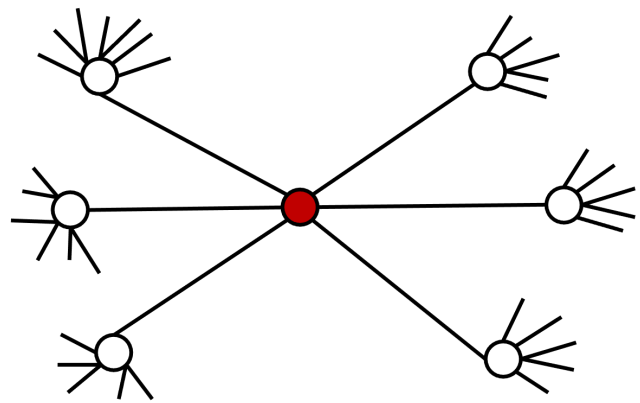
Neighborhood connectivity

☞ The average degree of v's neighbors, more this average is high for a node more it is important

$$c_v = \frac{\sum_{u \in \mathcal{N}_v} d_u}{d_v}$$

$$\mathbf{c} = \mathbf{D}^{-1}\mathbf{A}\mathbf{1}$$





Eigenvector centrality

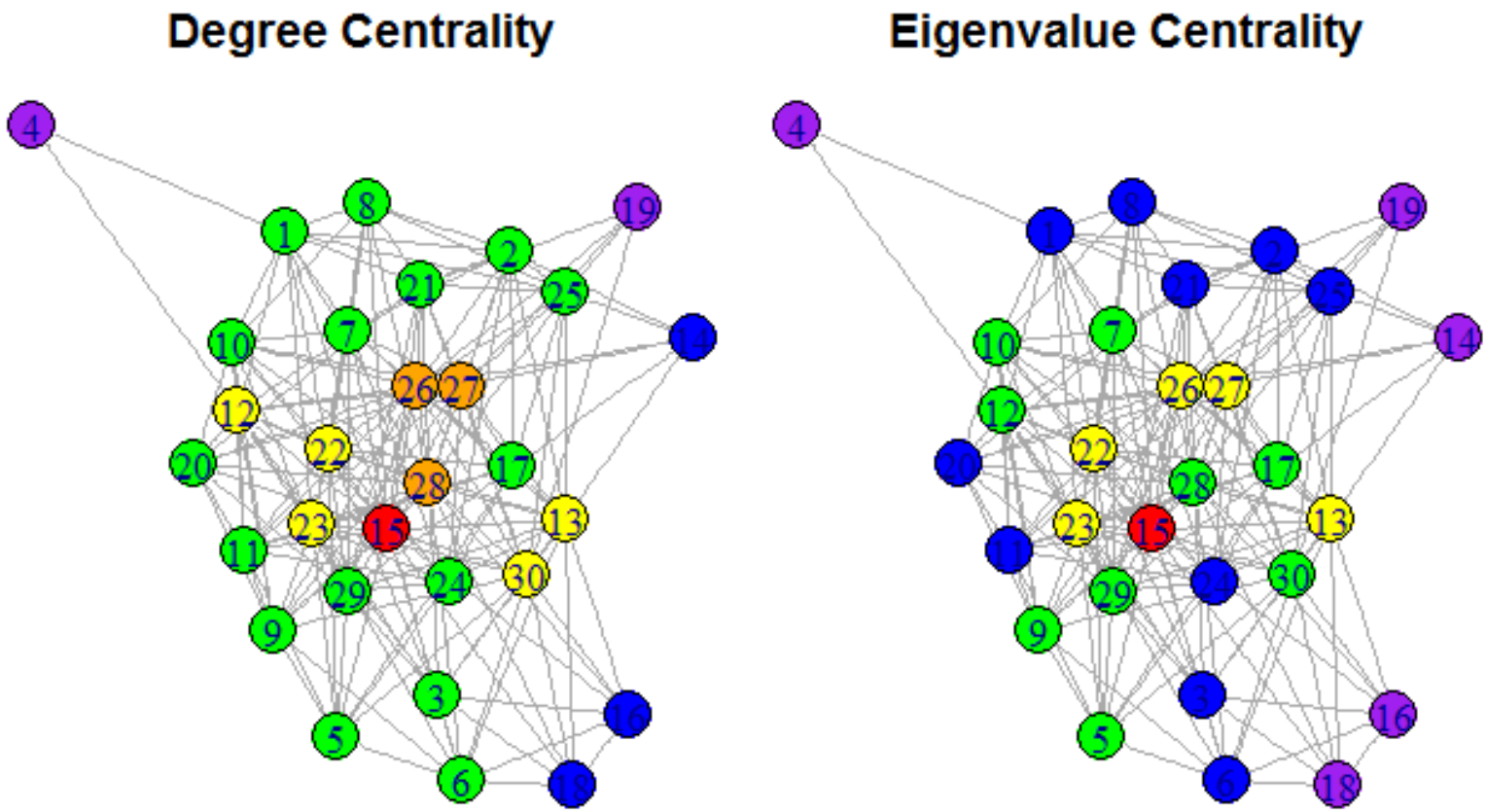
- A natural extension of degree centrality is eigenvector centrality
  ✓ Eigenvector centrality measures a node's importance while giving consideration to the importance of its neighbors. A high eigenvector score means that a node is connected to many nodes which themselves have high scores.
- The eigenvector centrality c_{v_i} of v_i is the sum of the centralities of its neighbors:
  $c_{v_i} = \frac{1}{\lambda} \sum_{v_j \in \mathcal{N}_i} \mathbf{A}_{ij}\, c_{v_j}$

$$\mathbf{Ac} = \lambda \mathbf{c}$$

- Introduced by the sociologist Bonacich (1972) suggested that the eigenvector of the largest eigenvalue of an adjacency matrix could make a good network centrality measure
- The eigenvector **c** must be non-negative, according to P-F theorem only the largest λ_{max}, the dominant value, results in the desired centrality measure.
- Example: $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$; $|\mathbf{A} - \lambda \mathbf{I}| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 3 - 4\lambda + \lambda^2 = 0$

  $\mathbf{E}_{\lambda=3} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

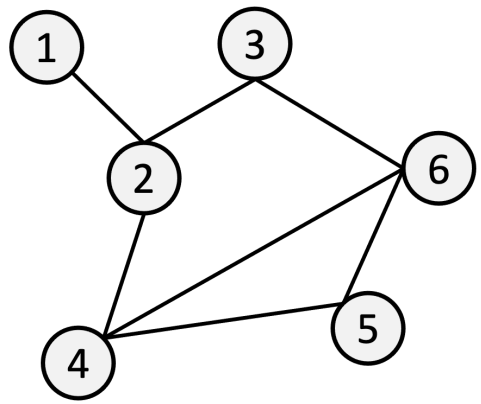
Eigenvector Centrality *versus* Degree centrality



from "purple","blue","green","yellow","orange","red"

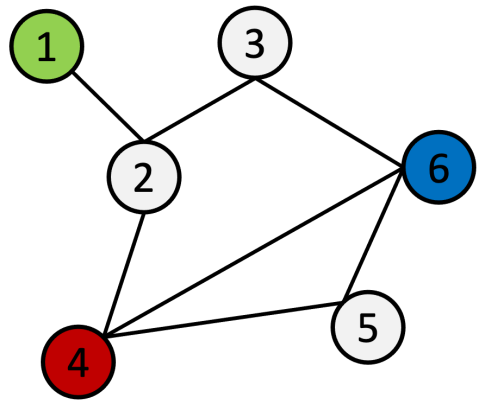
Eigenvector centrality

☞ Power method $\mathbf{c}^{(k+1)} = \mathbf{A}\mathbf{c}^{(k)}$. Initially the centralilty of each node is 1.



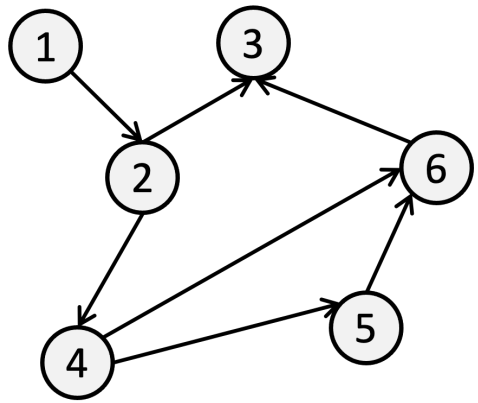
Eigenvector centrality

☞ Power method $\mathbf{c}^{(k+1)} = \mathbf{A}\mathbf{c}^{(k)}$
4, 6, 2, 5, 3, 1



Eigenvector centrality

☞ Power method $\mathbf{c}^{(k+1)} = \mathbf{A}\mathbf{c}^{(k)}$. Initially the centralilty of each node is 1.



Katz Centrality

- The main problem with eigenvector centrality is that it works well only if the graph is (strongly) connected. Real networks, especially directed networks, do not. The vertices that are not in (strongly) connected components have zero eigenvector (null) centrality.
- A way to work around this problem has been proposed by Leo Katz (1953). Give each node a minimum, positive amount of centrality that it can transfer to other nodes.

$$c_{v_i} = \alpha \sum_j \mathbf{A}_{ij} c_{v_j} + \beta$$

c_{v_i} denotes Katz centrality of a node v_i, where β is a vector whose elements are all equal a given positive constant (generally 1) and $\alpha \in (0, 1)$.

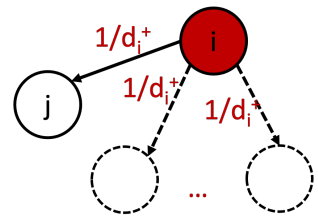
$$\mathbf{c} = \alpha \mathbf{A}\mathbf{c} + \mathbf{1}$$

$$\mathbf{c} = (\mathbf{I} - \alpha \mathbf{A})^{-1}\mathbf{1}$$

- $(\mathbf{I} - \alpha\mathbf{A})$ invertible $\left|(\mathbf{I} - \alpha\mathbf{A})\right| \neq 0 \implies \left|(\mathbf{A} - \frac{1}{\alpha}\mathbf{I})\right| \neq 0$
  This is $\mathbf{A}$ characteristic equation for eigenvalues, pick $0 < \alpha < \frac{1}{\lambda}$ where λ is the highest eigenvalue.
- Iterative method: $\mathbf{c}^{(k+1)} = \alpha\mathbf{A}\mathbf{c}^{(k)} + \mathbf{1}$ The strength of α decreases at each iteration (attenuation factor) more k (number of paths of length k) increases more the centrality decreases

PageRank Centrality

- Google's PageRank is a variant of the eigenvector centrality. PageRank uses in-degree to award one centrality point (popularity) for every link a node receives.



$$\mathbf{P}_{ij} = \begin{cases} \frac{1}{d_i^+}, & \text{if } v_j \in \mathcal{N}_i^+ \\ 0, & \text{otherwise} \end{cases};$$

$\frac{1}{d_i^+}$ represents the probability for a surfer to jump randomly from page v_i to a page v_j (independent events). $\mathbf{P}$ is a row stochastic $\sum_j \mathbf{P}_{ij} = 1$.

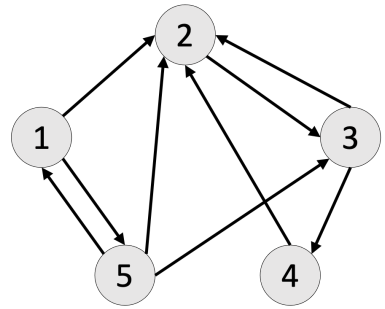
- The matrix form is defined as follows:

$$\lambda \mathbf{c} = \mathbf{c}\mathbf{P} \text{ with } \lambda = 1$$

$$\mathbf{c} = (\mathbf{D}^+)^{-1}\mathbf{A}$$

PageRank Centrality

- We use power iteration method: $\mathbf{c}^{(k+1)} = \mathbf{c}^{(k)}\mathbf{P}$; $\mathbf{c}^{(k)} = \mathbf{c}\mathbf{P}^k$ with $\mathbf{c} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{bmatrix}$



$$\mathbf{P} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \end{bmatrix} \text{ and } \mathbf{c} = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{bmatrix}$$

$$\mathbf{c}^{10} = \begin{bmatrix} 0 & 0.4 & 0.4 & 0.2 & 0 \end{bmatrix}$$

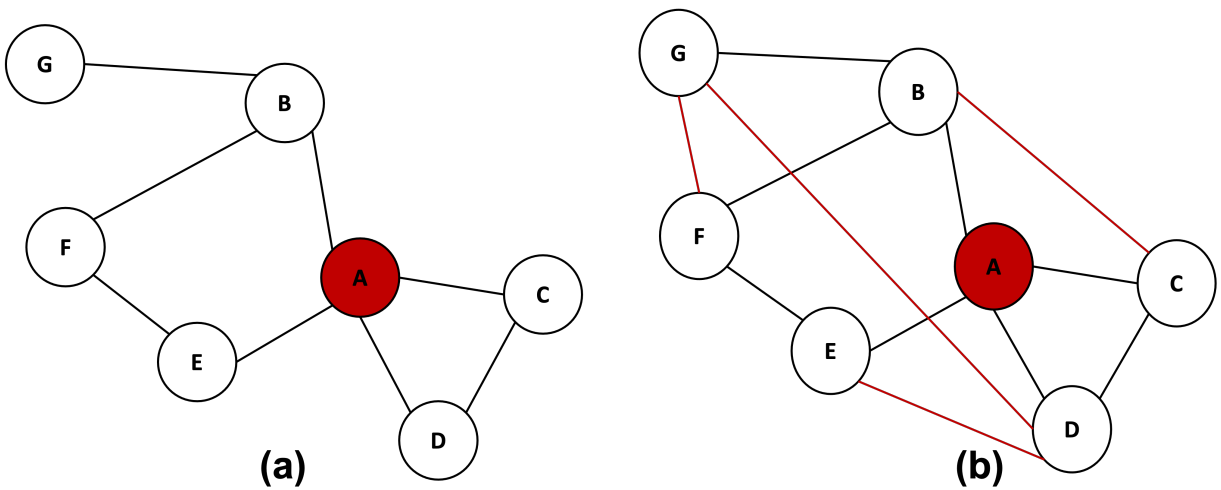
PageRank Centrality

- PageRank algorithm holds that a surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the surfer will continue is a damping factor α Damping factor d

$$\mathbf{c} = \left[\frac{(1-\alpha)}{n} \quad \frac{(1-\alpha)}{n} \quad \dots \quad \frac{(1-\alpha)}{n}\right] + \alpha \mathbf{cP}$$

Clustering Coefficient Centrality

- Triadic Closure: *If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves*.
- The probability that two randomly selected friends of node *A* are friends with each other:
  - ✓ Figure (a) = $\frac{1}{6}$ because there is only the single (C, D) edge among the six pairs of (B, C), (B, D), (B, E), (C, D), (C, E), and (D, E) increased to $\frac{1}{2}$ in Figure (b) three edges (B, C), (C, D), and(D, E) among the same six pairs.

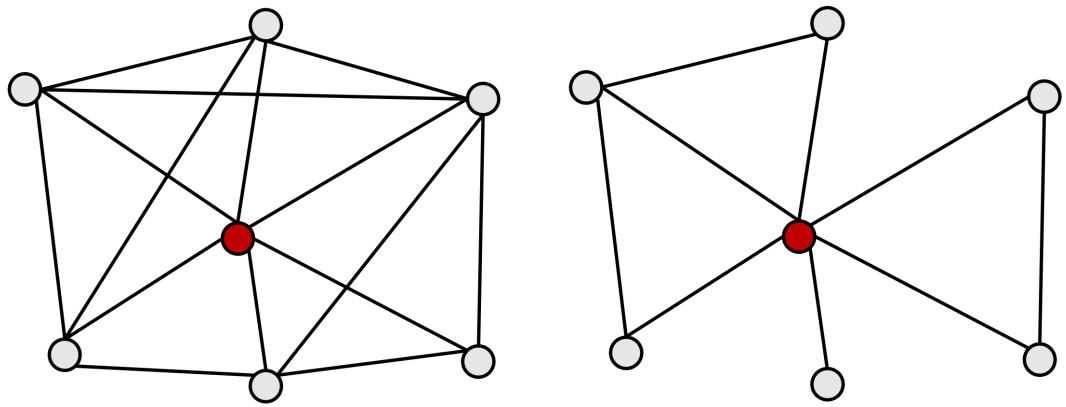


Clustering Coefficient Centrality

- Clustering Coefficient cc_v captures how the neighbors of v are linked to each other.

$$c_v = \frac{\left|\{(v,u),(u,w),(w,v) \in E\}\right|}{\frac{1}{2} d_v(d_v - 1)}$$ where $\frac{1}{2} d_v(d_v - 1)$ is the total number of links between neighbors

$\left|\{(v,u),(u,w),(w,v) \in E\}\right|$ is the number of all triangles involving v and v's neighbors



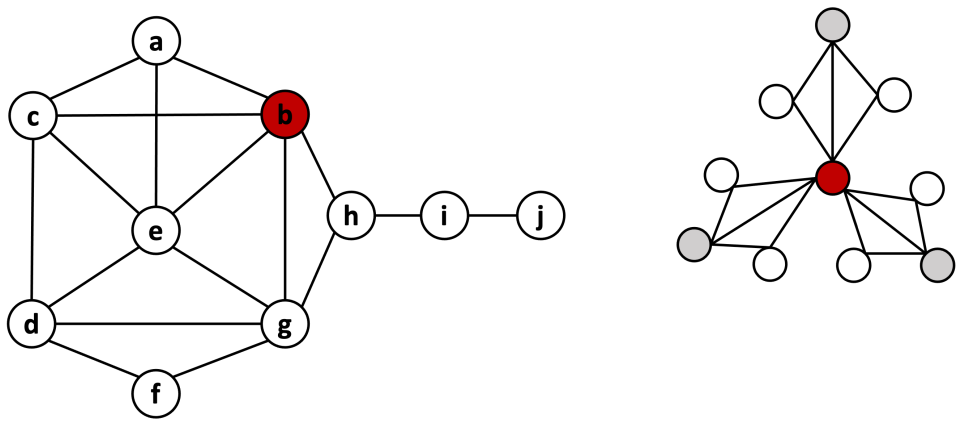
- The more densely interconnected v's neighborhood, the higher is its clustering coefficient. $c_v = 0$ v's neighborhood are not connected, $c_v = 1$ v's neighborhood are all connected.

Closeness centrality

- Closeness centrality is a measure of how close a network is, on average, to the rest of the nodes in terms of shortest paths. It measures the average distance between a node v and all other nodes in the network. Thus, the more central a node is, the closer it is to all other nodes.

$$c_V = \frac{1}{\sum_{r \neq v} dist(v, r)}$$

it can be normalised by $\frac{\sum_r dist(v,r)}{|V|-1}$



- Harmonic centrality: $c_V = \sum_{r \neq v} \frac{1}{dist(v, r)}$. The inverse distance to an unreachable vertex is considered to be zero $\frac{1}{dist(v,r)} = 0$

Betweenness centrality

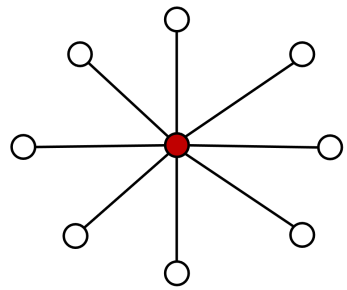
- A family of betweenness measures are defined to capture a node's importance as a conduit of information flow in a network
- Wide applications in network theory: in a telecommunications network, a node with higher betweenness centrality would have more control over the network, because more information will pass through that node.
- The most well known measures the number of times a node is on a shortest path between two nodes.

$$c_v = \sum_{s \neq v \neq t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$$

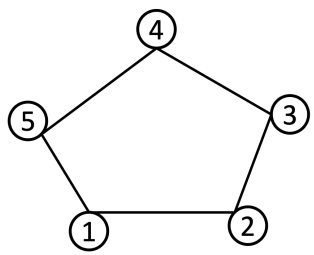
  where $\sigma_{s,t}$: number of shortest paths between source node s and target node t, while $\sigma_{s,t}(v)$: number of shortest paths between source s and target t nodes that pass through v node.
- It may be normalised by the number of ordered pairs of vertices not including v (combination of 2 over $(n-1)$).
  - ✓ for directed graphs $c_v = \frac{1}{(n-1)(n-2)} \sum_{s \neq v \neq t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$
  - ✓ for undirected graphs $c_v = \frac{2}{(n-1)(n-2)} \sum_{s \neq v \neq t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$

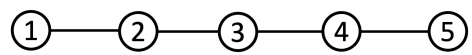
Betweenness centrality

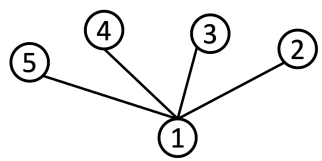
- undirected star graph, the center vertex has a betweenness of 1, while the leaves have a betweenness of 0.

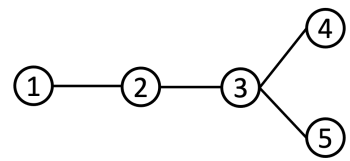


- What about the following graphs?



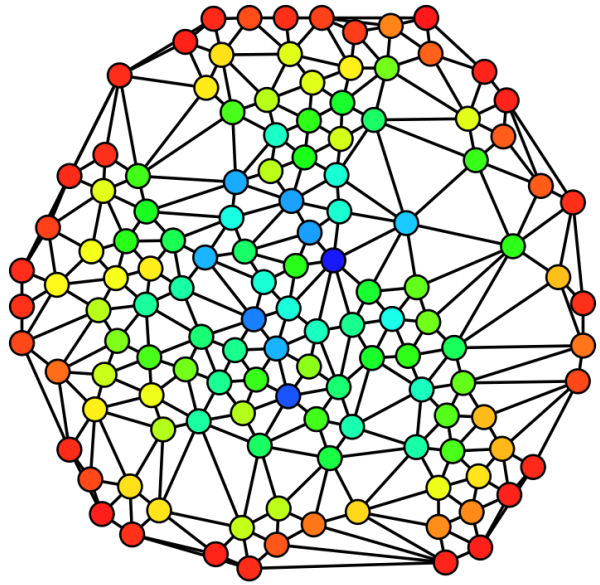






Betweenness centrality

- This graph shows the node betweenness, from red nodes to blue nodes = max,



- Betweeness assumes that information flow is through the shortest path. In a transport, the traffic will likely go through alternatives paths. Also in the case of rumours or infection.
- Katz centrality takes into account all paths of length k
- PageRank centrality takes into account the most probable walks.

1st Mini-Project

- ☞ Study/analyse the different centralities on different kind of graphs ( )
- ☞ Other centralities such as Hyperlink-Induced Topic Search (HITS).