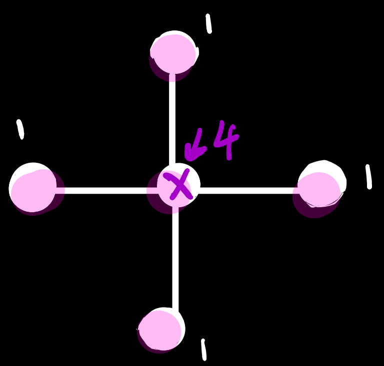
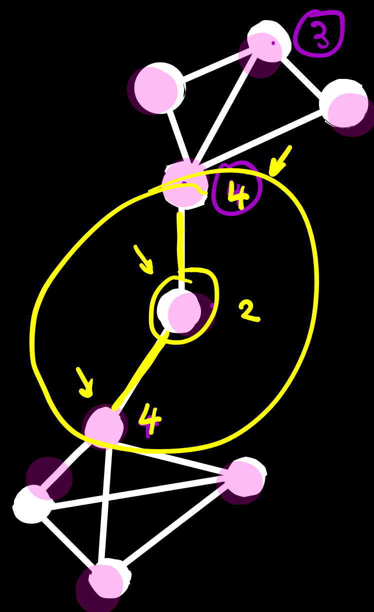
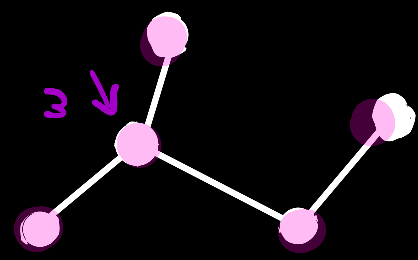
$\mathcal{G}raph\mathcal{T}_2$



Graph 1
G_1

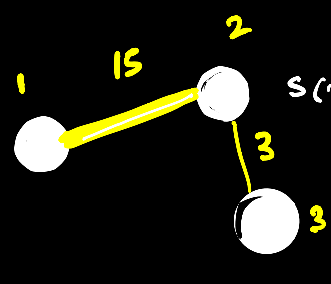


Graph 2
G_2



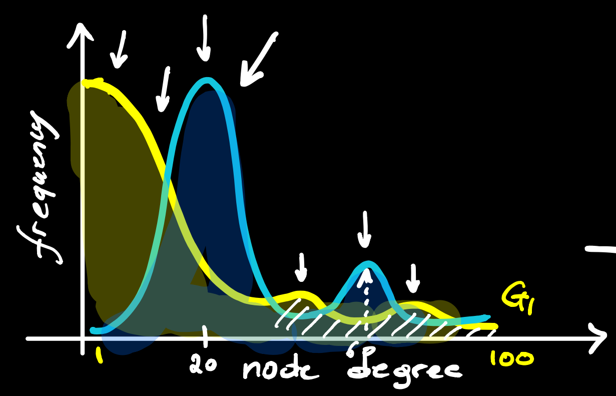
Graph 3
G_3

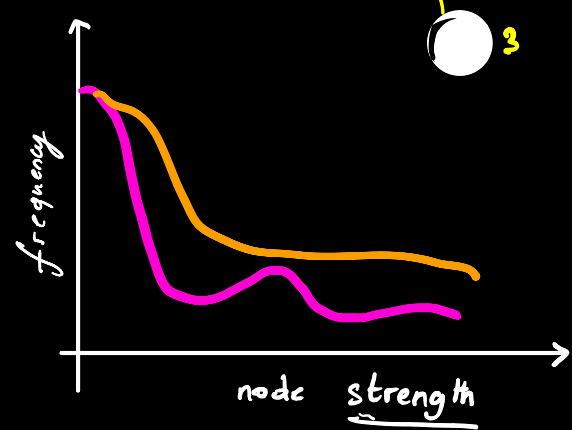
$d(v_2) = 2$



$s(v_2) = 18$

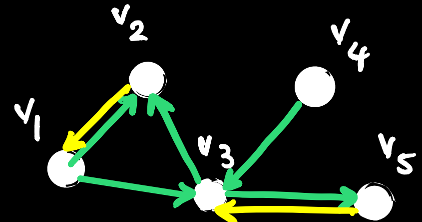
* Node degree distribution

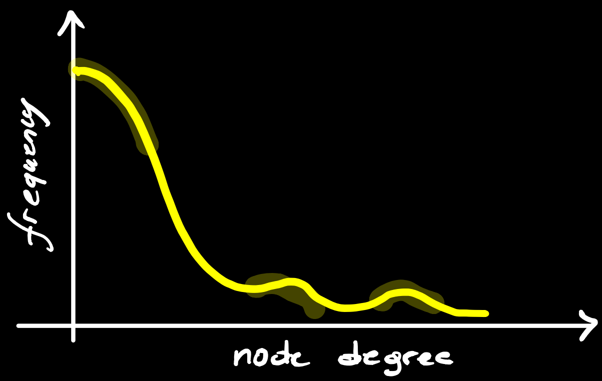




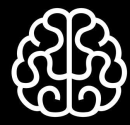
* Directed graphs

- in-degree : $d^{in}(v_i) = \sum_{j \neq i} A^b_{ij}$
  (number of edges from nodes $j \to i$)
- out-degree : $d^{out}(v_i) = \sum_{j \neq i} A^b_{ji}$
  (number of edges incident to node i)





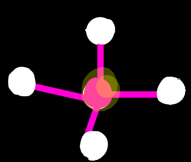
Non-uniform degree distribution across graph nodes.



The heterogeneous distribution of node degree implies that different nodes serve different topological roles.

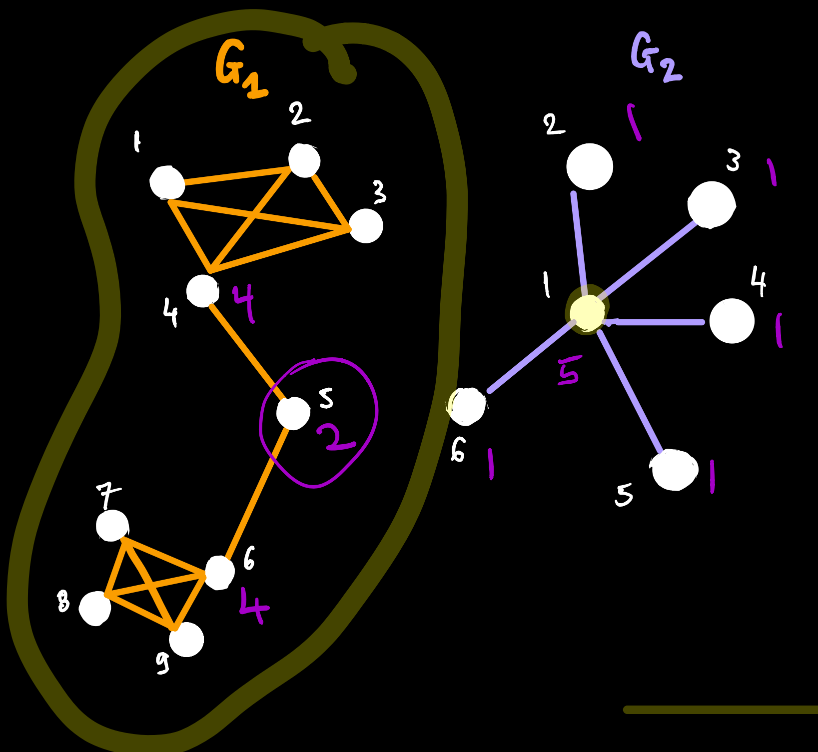
GraphT

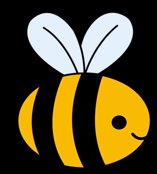
① centrality → degree, closeness, betweeness



② identifying hub nodes & other centrality measures

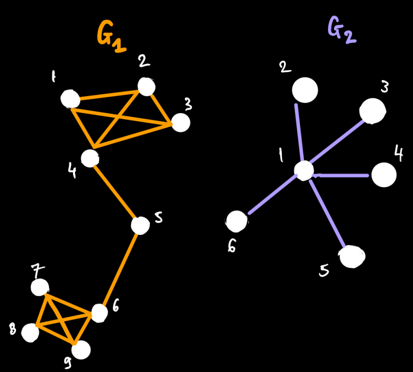
CENTRALITY





"What is the most important node in each graph? what about its degree?"

- Node degree or strength provide only partial account of the role or significance of a node in a graph.
- Node degree does not provide a comprehensive characterization of individual nodes (limited to a local scale)



→ More broader measures can help examine the role of a node w.r.t. other nodes.

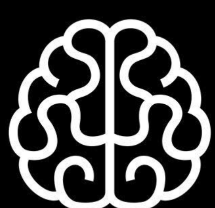
→ How to examine the capacity of a node to influence other nodes?

⇓

we examine its connection topology

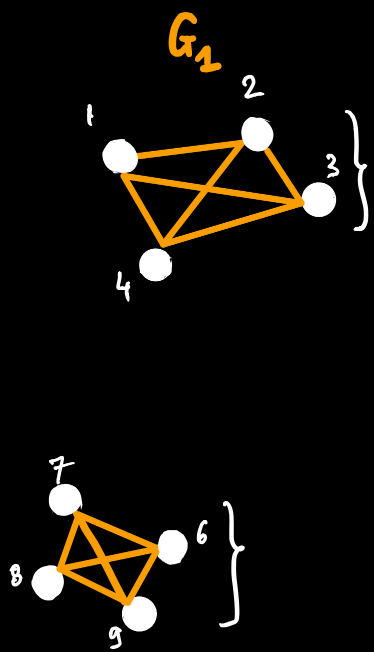
↓

Any ideas?

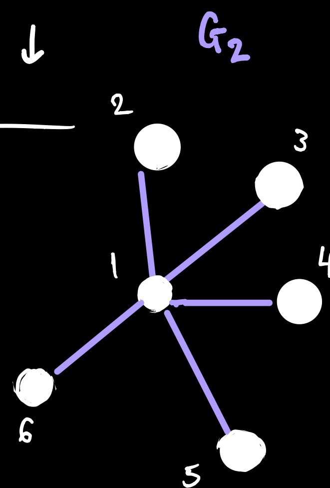


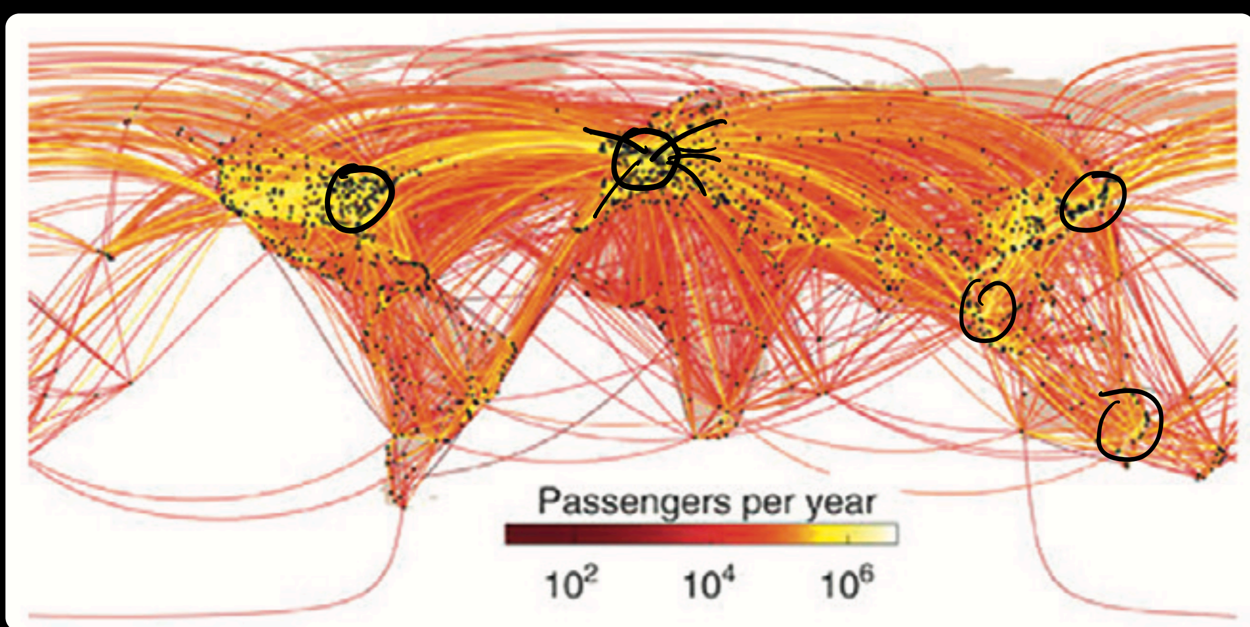
The definition of an important/influencial or central node depends on the kind of influence we aim to measure.

➡ different definitions of central / hub nodes.



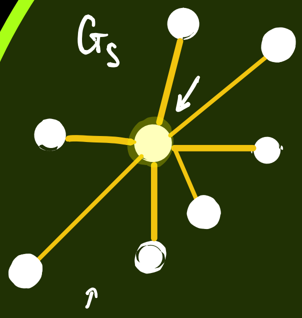
$$\frac{\text{aver}\,(C_B) - \downarrow}{\text{aver}\,(C_B)}$$





- A geographical representation of the worldwide airline network.
- black dots = airports ; edges = passenger traffic between airports

① Emergence of the concept of centrality



* The concept of centrality was introduced by Bavelas (1948) to understand the structural position of an individual within a social network and determine the influence of that person in group-wide processes.

* Freeman (1979) showed that all centrality measures identify the same central node in a star-like (called hub-and-spoke) graph → any communication through G_S must pass through its topological center.

Freeman (1979)

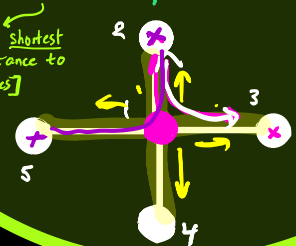
3 fundamental properties can be ascribed to a central node:

① it has the maximum possible degree, since it is connected to all other nodes

② falls on the shortest possible topological path between all pairs of nodes.

③ it is "maximally close" to all other nodes.

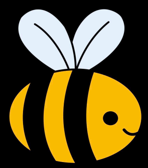
[falls on the shortest topological distance to all other nodes]



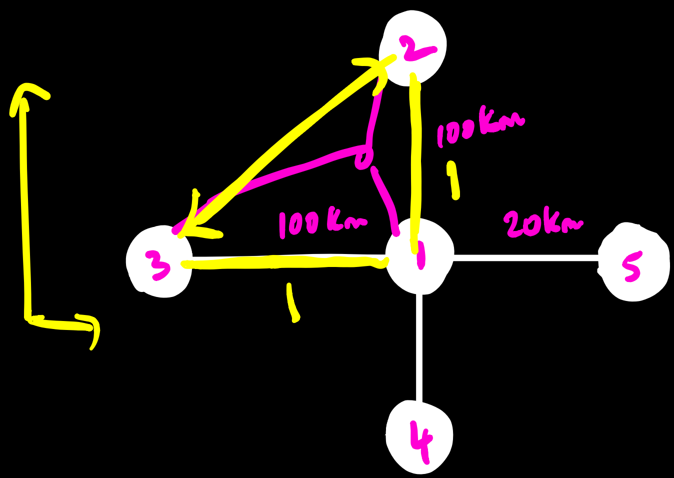
- degree ✓
- betweenness ✓
- closeness

↓

cardinal aspects of topological centrality



"Is the topological space a Euclidean Space ?"



Graphs are Non-Euclidean objects.

Degree-based measures of centrality

1. DEGREE CENTRALITY

Simplest measure of centrality.
≈
equivalent to the node degree

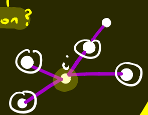
→ for undirected graphs, we define it as:

$C_D(v_i) = d(v_i) = \sum_{i \neq j} A^b_{ij}$

- What is the underlying assumption about central nodes when using this definition?

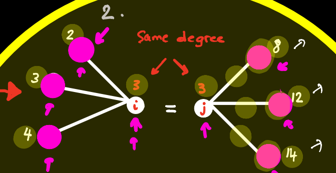
→ **local structure**

⇒ immediate neighbors of a node i



One limitation: all connections are treated equally.

2.



Which node is most influential?
(with maximum reach to other nodes?
Node i or j?

* The influence of each node is determined by the degree of its neighbors.

degree centrality counts quantity, but does not consider quality.

→ ANY IDEAS?

3. EIGEN CENTRALITY

→ accounts for both the quantity and quality.

quantity ↓ degree of a node

quality ↓ degree of its neighbors

- Eigen decomposition of the graph adjacency matrix A. an eigen vector x of A is a nonzero vector, when multiplied by A satisfies: $Ax = \lambda x$

eigenvector of A — x; λ — eigen value, scaling factor

- $C_E(v_i)$ of node v_i is defined as the i^{th} entry in the eigenvector with the largest eigenvalue of A.

→ it is equivalent to the summed centrality of v_i's neighbors:

$$C_E(v_i) = x_i = \frac{1}{\lambda_1} \times \sum_{j=1}^{N} A_{ij} x_j$$

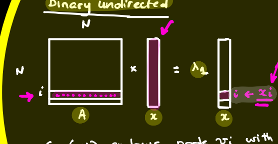
x v_i

[Bonacich, 1987]

4. EIGEN CENTRALITY

$$C_E(v_i) = x_i = \frac{1}{\lambda_1} \sum_{j=1}^{N} A_{ij} x_j$$

binary undirected



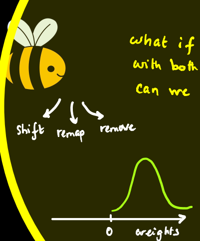
- $C_E(v_i)$ endows node v_i with high centrality if:
  1. it has many neighbors
  2. its neighbors are highly connected
  3. or both.

weighted undirected

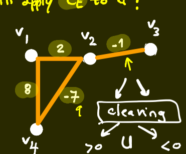
- C_E also applies to weighted adjacency matrices with strictly positive entries.

5. EIGEN CENTRALITY

$$C_E(v_i) = x_i = \frac{1}{\lambda_1} \sum_{j=1}^{N} A_{ij} x_j$$



what if we have a weighted graph with both positive and negative weights, can we still apply C_E to G?



- This depends on how we choose to interpret the polarity of edges.

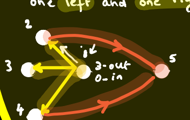
6. EIGEN CENTRALITY

$$C_E(v_i) = x_i = \frac{1}{\lambda_1} \sum_{j=1}^{N} A_{ij} x_j$$

directed graphs (directionality)

- A is asymmetric → it has two leading eigenvectors, one left and one right.

$$\begin{cases} A x_r = \lambda_1 x_r \\ (x_l^T A)^T = A^T x_l = \lambda_1 x_l \end{cases}$$



	1	2	3	4	5
1	0	0	0	0	0
2	1	0			0
3	1		0		
4	1			0	0
5	0	1		1	0

= A

out-degree — in-degree

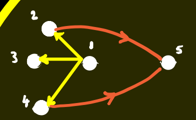
A_{ij}: edge from $j \rightarrow i$

$A_{21} = 1$

left eigenvector centrality wrt out-degree d^{out}

right eigenvector centrality wrt in-degree d^{in}

7. EIGEN CENTRALITY



$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & & & 0 \\ 3 & 1 & & 0 & & \\ 4 & 1 & & & 0 & 0 \\ 5 & 0 & 1 & & 1 & 0 \end{array} = A$$

A_{ij} : edge from $j \to i$

E.g. Twitter users (which one to choose C_E^r or C_E^ℓ?)
↳ someone with many followers (high d^{in}) will be more central than someone who follows many people (high d^{out}) and (low d^{in}).

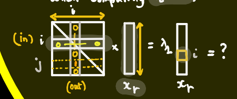
- C_E^r (in-degree) → identifies central sources of information whose efferent connections can influence many other nodes.
- C_E^ℓ (out-degree) → identifies sinks of info that receive a large amount of afferent connections.

8. EIGEN CENTRALITY

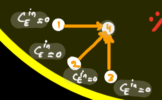
E.g., we are interested in centrality related to incoming (efferent) connections.

↓ $C_E^{in}(v_i) = 0$ ($d^{out}(v_i) > 0$ and $d^{in} = 0$)

with no incoming connections, $C_E^{in}(v_i) = 0$!
↳ This is because we sum over all nodes which project to i when computing $C_E(v_i)$



$$\begin{cases} A x_r = \lambda_1 x_r \\ (x_\ell^T A)^T = A^T x_\ell = \lambda_1 x_\ell \end{cases}$$



nodes with incoming edges from nodes with zero centrality will also have zero centrality.

EIGEN DECOMPOSITION

① GENERAL

$$\underbrace{A}_{N \times N} \; \underbrace{x}_{N \times 1} = \underbrace{\lambda}_{\in \mathbb{R}} \, x$$

eigenvalue = eigenmode of system A (magnitude)

$N \times N$: describes a system that affects some linear, geometric transformation of a vector.

x: eigenvector changes only in magnitude and not direction

- Eigenvalues are sometimes thought of as the "natural modes" of a system A.
- Example: the eigenvalues of an oscillating system are its natural or resonant frequencies of its oscillations and the eigenvectors correspond to their shapes.

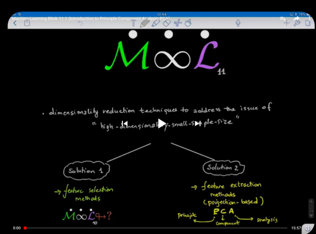
② COVARIANCE MATRIX

$$\underbrace{\Sigma}_{\text{data covariance matrix}} x = \lambda x$$

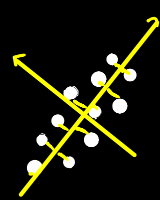
λ: amount of variance in each direction

x: orthogonal directions of maximum data variance

→ The eigen decomposition of Σ is equivalent to a principle component analysis (PCA).



[BASIRA Lab, Machine Learning Blink II]



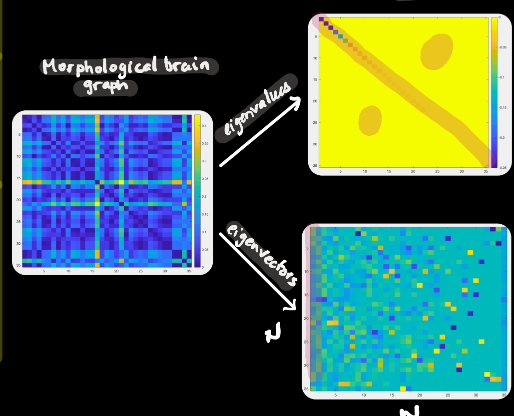
③ ADJACENCY MATRIX

$$\underbrace{A}_{\text{adjacency matrix}} x_1 = \overbrace{\lambda}^{\text{largest eigenvalue}} \underbrace{x_1}$$

largest eigenvalue $\lambda_1 > \lambda_2 > \ldots$

adjacency matrix of a graph with positive values

principle eigenvector with largest eigenvalue gives a topological measure of the centrality of each node



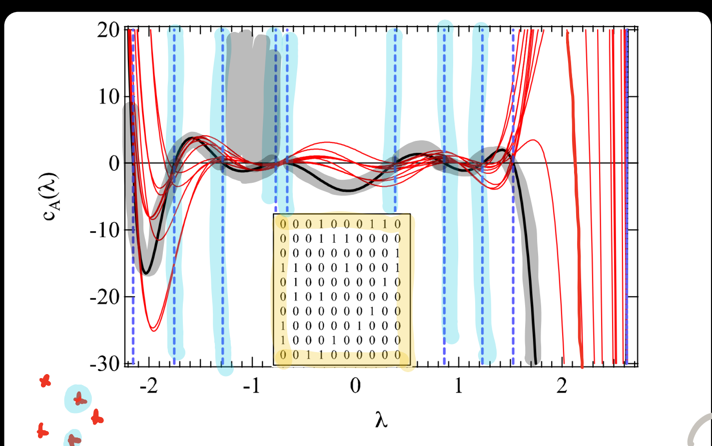
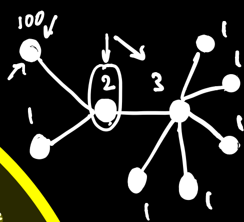


Figure 1: The characteristic polynomials $c_{A_{G\setminus\{n\}}}(\lambda)$ for $1 \le i \le N$ in red and $c_{A_G}(\lambda)$ in black for an Erdős-Rényi graph $G_{0.2}(10)$, whose adjacency matrix is also shown. The blue vertical lines denote the eigenvalues of A (zeros of $c_A(\lambda)$).

P. Van Mieghem 2018 [reference below]

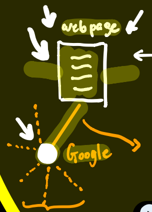
randomly generated graph
with probability $p = 0.2$ for creating edges.



9. EIGENVECTOR CENTRALITY

C_E can assign high centrality to low-degree nodes simply because they are connected to a hub node.

Why this is problematic?



assign a high centrality to an individual page because it receives a link from the search engine Google (hub)

this link is one of many millions of links that Google make.

" How to solve this?
We don't want to create 'fake' hubs / control nodes "

10. PAGERANK CENTRALITY

- Proposed by Google founders (Brin and Page, 1998)
- PageRank is one of the popular algorithms which ranks websites.

Sol°: scale the contributions that the neighbors of node i make to its centrality by the degree of those neighbors.

→ accounts for potential biases associated with links to highly connected nodes (~ Normalization)

$$C_{PR}(v_i) = x = (I - \alpha A D^{-1})^{-1} \mathbb{1} \quad \text{[Newman, 2010]}$$

$$= D (D - \alpha A)^{-1} \mathbb{1}$$

D: $N \times N$ diagonal matrix, $D_{ii} = \max(d^{out}(v_i), 1)$

α: damping factor

$\alpha \uparrow$ → network topology will highly influence C_{PR}

$\alpha \in [0, \frac{1}{\lambda_1}[$ → largest eigenvalue of AD^{-1}

C_D
C_E
C_{PR}
⋮
"learning centrality measures"

In this paper, we have taken on the audacious task of condensing every page on the World Wide Web into a single number, its PageRank. PageRank is a global ranking of all web pages, regardless of their content, based solely on their location in the Web's graph structure.

Using PageRank, we are able to order search results so that more important and central Web pages are given preference. In experiments, this turns out to provide higher quality search results to users. Th intuition behind PageRank is that it uses information which is exernal to the Web pages themselves - their backlinks, which provide a kind of peer review. Furthermore, backlinks from "important" pages are more significant than backlinks from average pages. This is encompassed in the recursive definition of PageRank (Section 2.4).

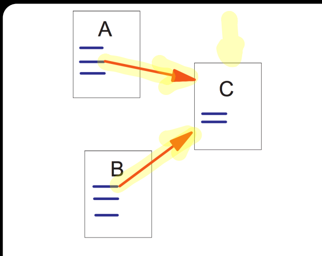
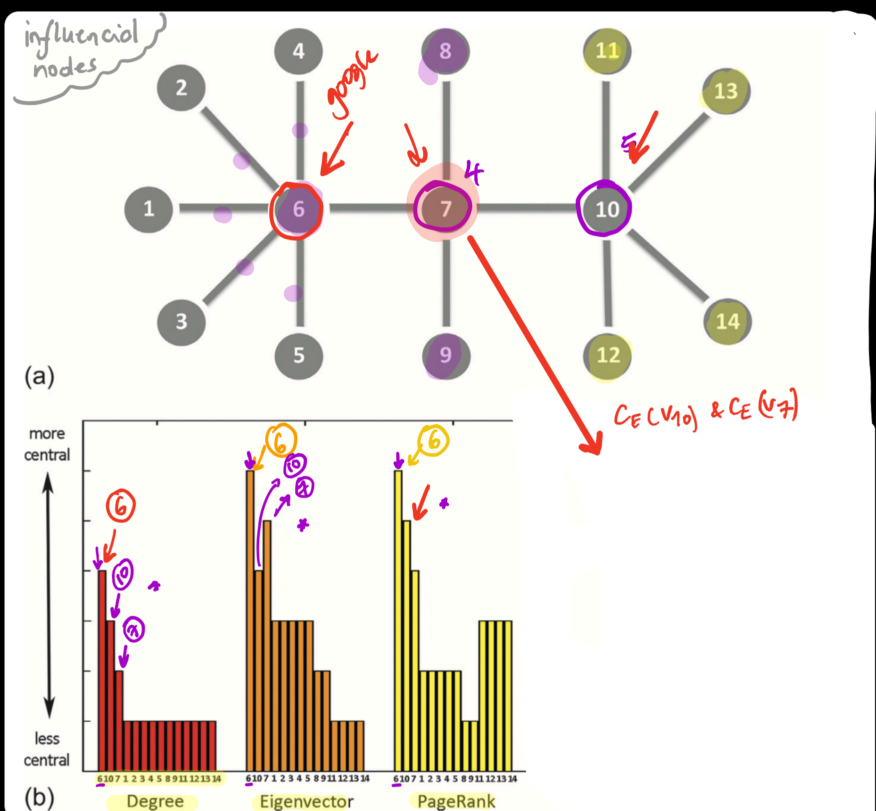


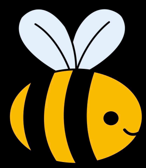
Figure 1: A and B are Backlinks of C

[Brin & Page, 1998]

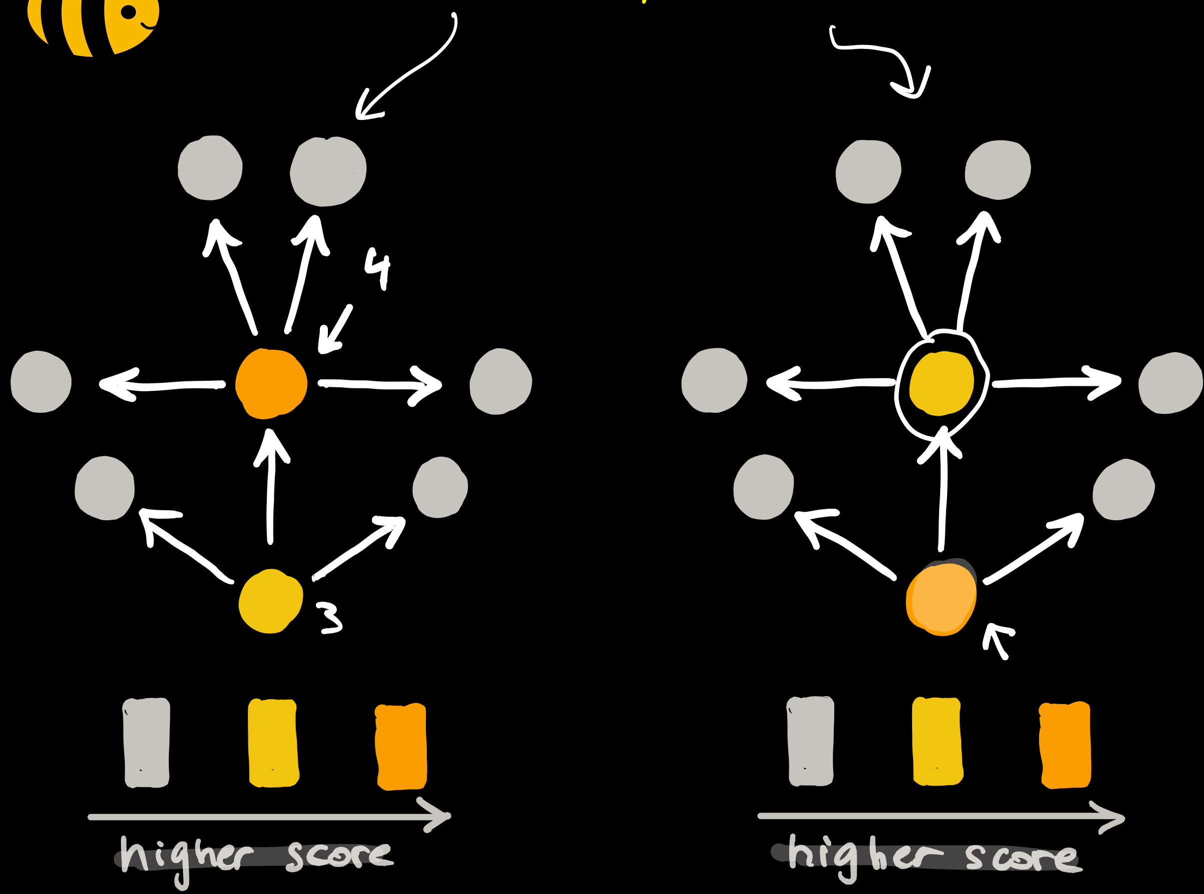


① why $C_E(v_{10}) < C_E(v_7)$?

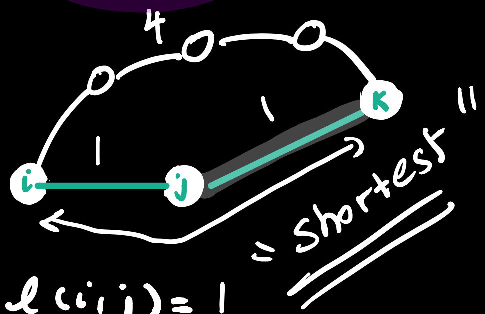
② why using C_{PR}, we get $C_{PR}(v_{10}) > C_{PR}(v_7)$?



" Degree centrality or eigencentrality ? "



CLOSENESS CENTRALITY



$l(i,j) = 1$

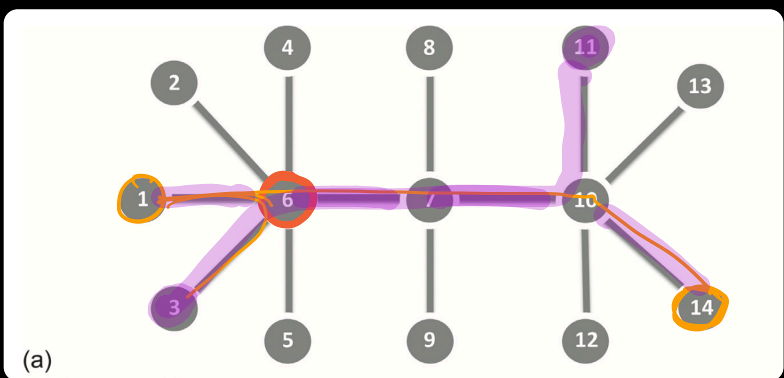
$l(i,k) = 2$

- Two nodes are maximally close in a topological sense if they share a direct connection → adjacent / neighbors.

Def : topological distance between nodes i and j : l_{ij}

l_{ij} = length of the shortest path in G between i and j.

e.g., count the number of edges in an unweighted undirected graph.



l_{ij} = shortest path length or geodesic distance

$l_{1,6} = 1$

$l_{1,14} = 4$

$l_{1,3} = 2$

- A node with short average path length is able to interact with many network elements via only a few links → topologically central.

- Messages (information flow) emanating from such a central node will spread to other nodes in a relatively short period of time.

  Conversely, signals originating from other non-central nodes will only take a short time to reach the central node.

DEFINITION

(first articulated by Bavelas 1950 & formulated by Beauchamp, 1965)

- Closeness centrality of a node is defined as the inverse of its average shortest path length to all other nodes:

$$C_C(v_i) = \frac{N-1}{\sum_{j \neq i} l_{ij}}$$

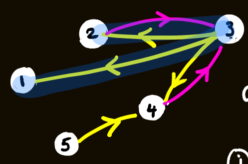
l_{ij} : shortest path length, or topological distance between nodes i and j

+ physical proximity

two nodes are "close" if they share a direct connection.

$$C_c(v_i) = \frac{N-1}{\sum_{j \neq i} l_{ij}}$$

11. DIRECTED GRAPHS



$$C_c^{in}(v_1) = \frac{4}{l_{12} + l_{13} + l_{14} + l_{15}} = \frac{4}{2+1+2+3} = \frac{4}{8}$$

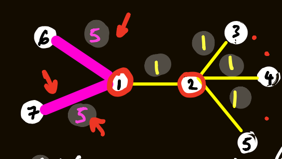
j → i

- l_{ij} : shortest paths that are incoming to node i.
- we can also compute C_c based on paths outgoing from node i (sum over all paths from i to j, $j = 1..N$)

$$\begin{cases} C_c^{out}(v_i) = \dfrac{(N-1)}{\sum_{j \neq i} l_{ji}} \\ C_c^{in}(v_i) = \dfrac{(N-1)}{\sum_{j \neq i} l_{ij}} \end{cases}$$

12. WEIGHTED GRAPHS

- C_c can be estimated by computing the shortest weighted path length between regions.

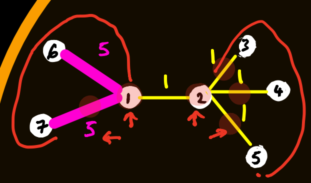


$$\begin{cases} C_c(V_2) = 6 / (\underbrace{1+1+1+1+6+6}) = 6/16 \\ C_c(V_1) = 6/16 \end{cases}$$

v_2: node with many low weight edges

v_1: node with a few high weight edges.

13. WEIGHTED GRAPHS



$C_c(V_2) = C_c(V_1)$

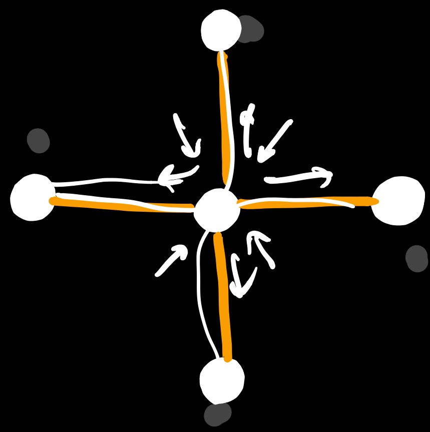
" How to solve this ambiguity? "

Hint: emphasize the contribution of the number of edges (also called hop count) - not only weights.

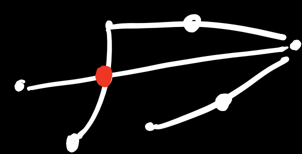
Think of new closeness centrality measures and justify your intuition and choice.

check sol° in Box 5.3 by Opsahl et. al (2010)

BETWEENNESS CENTRALITY



- degree ⇒ local neighborhood
- closeness ⇒ effective outreach via closest path
- betweenness ⇒ best mediators of high proportion of traffic



- Betweeness centrality is a popular measure proposed by Freeman (1977) and Anthonisse (1971).

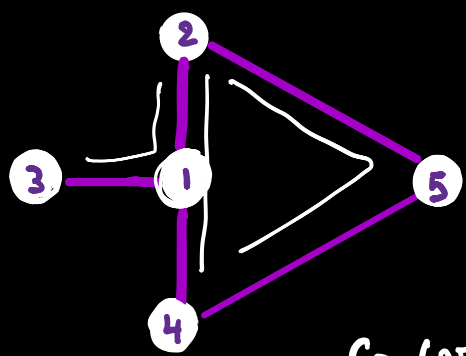
$C_B(v_i)$ measures the proportion of shortest paths between all node pairs in G that pass through node v_i.

→ C_B indexes the extent to which a node lies "between" other pairs of nodes.

→ If information travels through a graph along the shortest path, then nodes that lie on many shortest paths will mediate a high proportion of traffic.

$$C_B(v_i) = \underbrace{\frac{1}{(N-1)(N-2)}}_{\text{number of node pairs that does not include node } i} \sum_{\substack{h \neq i \\ h \neq j \\ j \neq i}} \frac{\overbrace{\rho_{hj}(v_i)}^{\text{number of shortest paths between } v_h \text{ and } v_j \text{ passing by } v_i}}{\underbrace{\rho_{hj}}_{\text{number of shortest paths between } h \text{ and } j}}$$

this normalization accounts for multiple possible shortest paths between any pair of nodes.



2 3	3 4	4 5
2 4	3 5	
2 5		

$$C_B(v_1) = \cancel{\frac{1}{1/2}} \left[\underbrace{\frac{1}{1} + \frac{1}{2} + \frac{1}{1} + \cdots +}_{?} \right]$$

① C_B can be computed similarly for {weighted / directed} graphs using appropriate methods for finding shortest paths.

② C_B can be computed for each individual {node / edge}.

FOOD FOR THOUGHT ("YOUR ASSUMPTIONS")

Both closeness and betweenness centralities are rooted in the assumption that information is routed along the shortest paths of the graph.

→ This assumption might not be appropriate for biological graphs (e.g., brain graphs)

$\exists$

"How to relax this strong assumption ?"

[Hint : use hinting strategy]

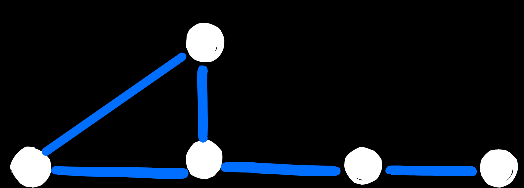
Sol^o by Estrada and Hatano, 2008 :

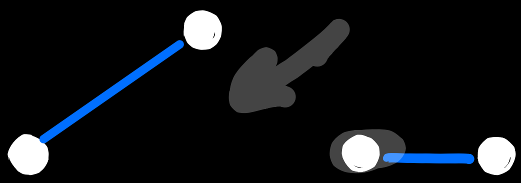
communicability = accounts for all possible paths between nodes, with shorter paths given higher weights.

DELTA CENTRALITY

More ideas ...

- Define centrality without specifically depending on { degree, closeness, betweenness }





(Latora and Marchiori, 2007)

- How does the removal of a specific node affect the structure and function of a graph?

Intuition: remove of highly central nodes will exert a disproportionate impact on other graph nodes.

- Delta centrality:

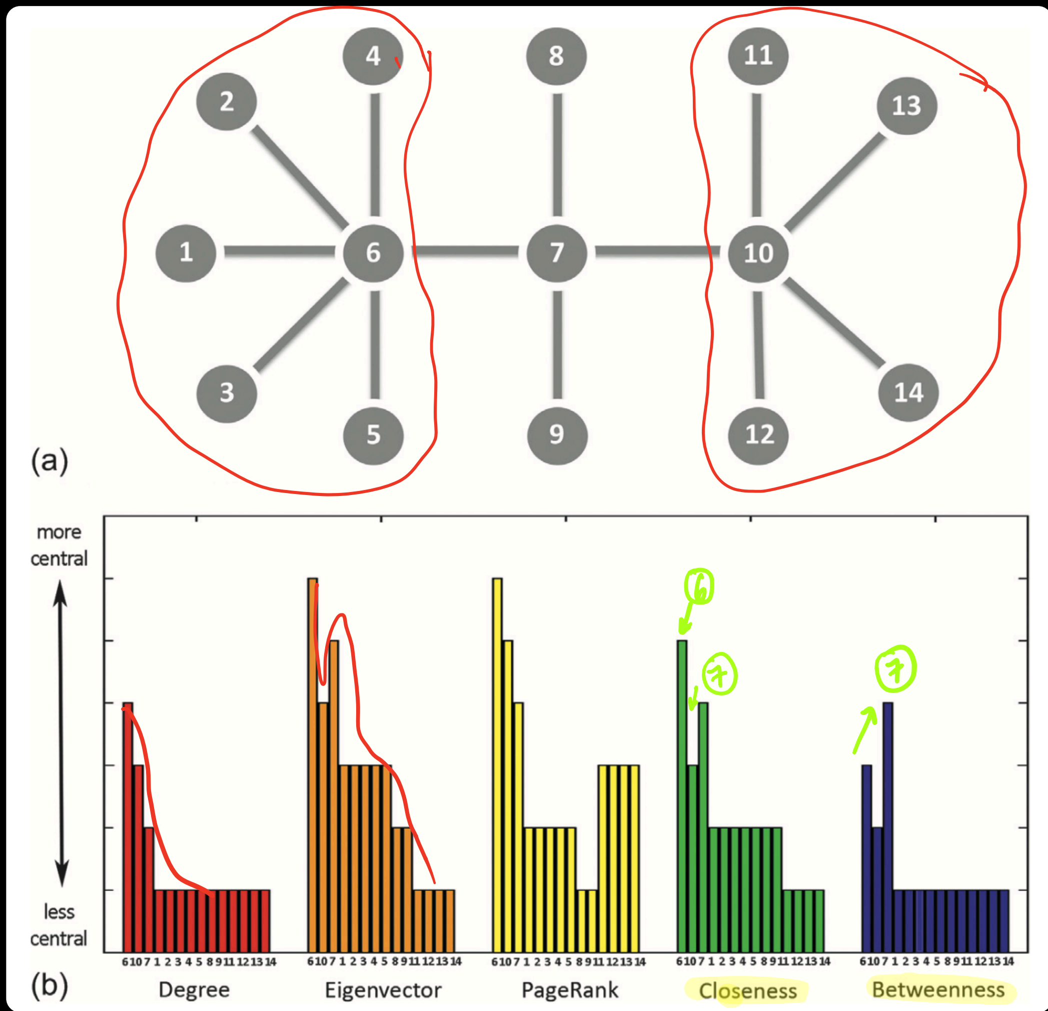
edges   nodes

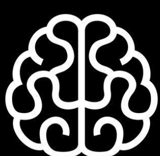
$$C_\Delta = \frac{\Delta M_i}{M} = \frac{M(G) - M(G \setminus \{v_i\})}{M(G)}$$

M: a measure to index the functional or topological integrity of a graph. ⇐ $\underline{\underline{G}}$

ΔM_i : change observed after the removal of node v_i.

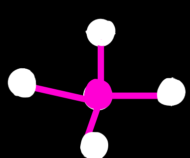
$G \setminus \{v_i\}$: graph induced by removing v_i.





- Different measures of centrality make different assumptions about how information flows on a graph G.
  → These assumptions should be borne in mind when interpreting the results of a centrality analysis.
- Centrality measures are important for finding hubs.

① centrality → degree, closeness, betweeness



② identifying hub nodes & other centrality measures