



# Social Network Analysis

---

## NETWORK MEASURES

SLIDE CREDITS: TEACHING MATERIAL ON SOCIAL NETWORK ANALYSIS BY TANMOY CHAKRABORTY, WILEY, 2021

# Where's the similarity?



<https://hbr.org/2012/09/marketing-gangnam-style>

## Official Release

Jul 15, 2012      Nov 16, 2011

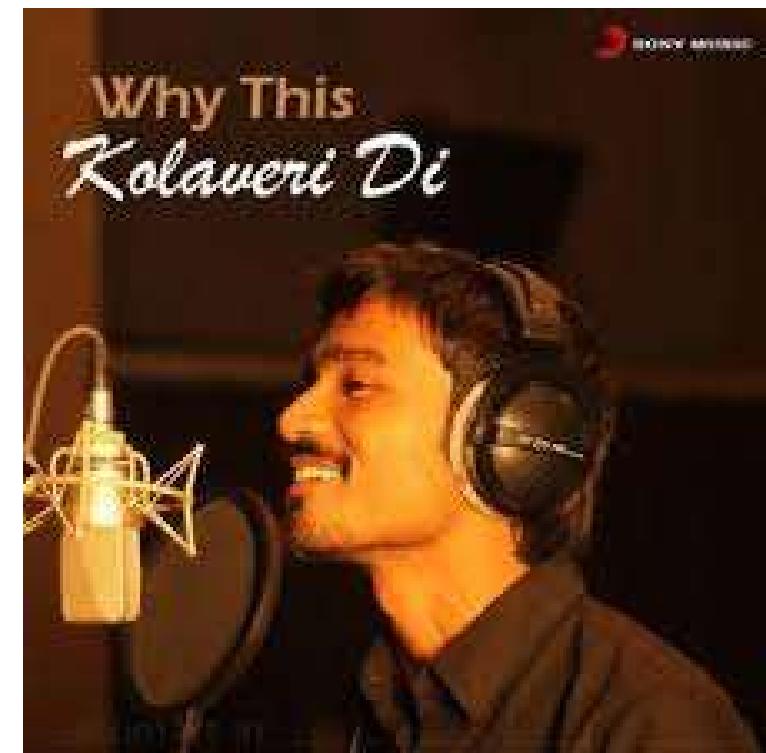
## Popularity

One billion views  
in 6 months      30 million views  
within 2 months

## Total YouTube Views

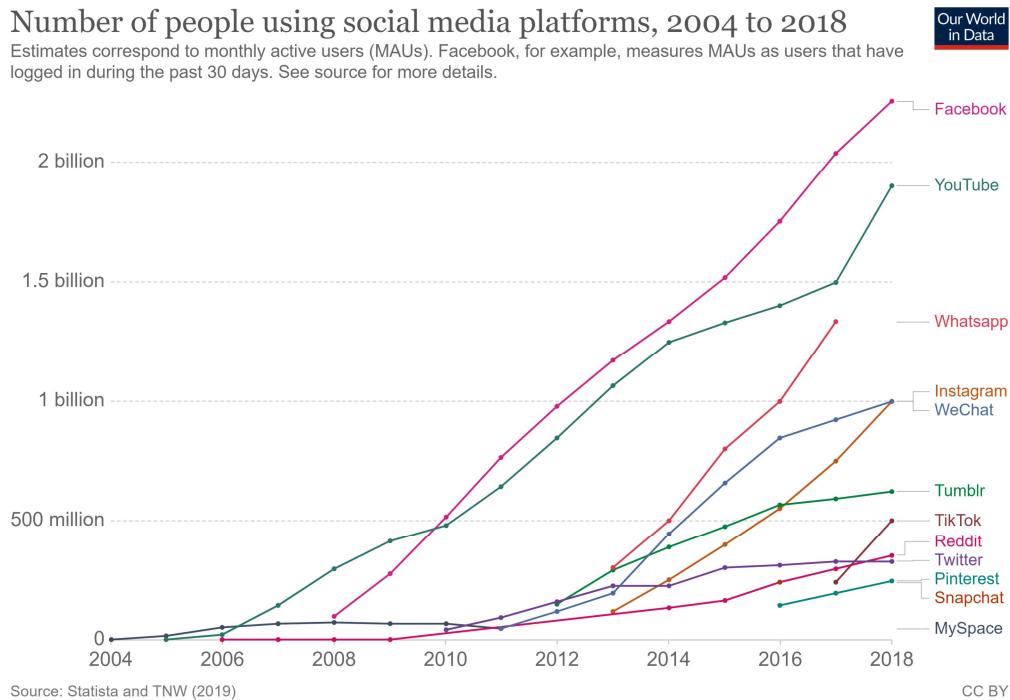
Over 3.9 billion  
views by 2021      Over 235 million  
views by 2020

## VIRAL MARKETING



<https://www.businesstoday.in/magazine/case-study/kolaveri-di-success-case-study/story/22957.html>

# The Rise of Online Social Media: Some Interesting Facts



- Online Social Media set in motion in early 2000
- First social networking and video website, Bolt, launched in 1996
- First social media site to reach a million monthly active users, MySpace in 2004
- 3.81 billion active social media users 2020
- 9.2% year-on-year increase from 3.48 billion in 2019
- 49.03% of the world's population (7.77 billion) is active on social media in 2020

<https://ourworldindata.org/rise-of-social-media#:~:text=The%20percentage%20of%20US%20adults,to%20around%2030%25%20in%202018.>

<https://backlinko.com/social-media-users>

# Online Social Media: Some Interesting Questions

---

- ❑ What is the dynamics when one's post receive high visibility on online social media?
- ❑ How to publicise one's post in online social media?
- ❑ How to find the social media celebrities in such a vast online world?
- ❑ How to identify the prolific users in a specific domain in social media?
- ❑ What are the role of prolific users when a post becomes viral in social network?
- ❑ How to determine if two social media users are similar in terms of online activities?
- ❑ How do we know if similar users are connected in a network?
- ❑ What are the relevant quantities and how to measure these quantities?

# Network Measures: Classification

---

## Microscopic

- ❖ Degree
- ❖ Local clustering coefficient
- ❖ Node centrality

## Mesoscopic

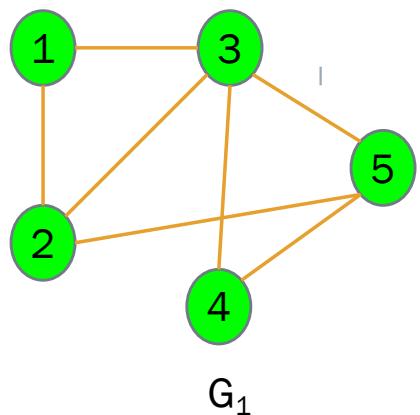
- ❖ Connected components
- ❖ Giant components
- ❖ Group centralities

## Macroscopic

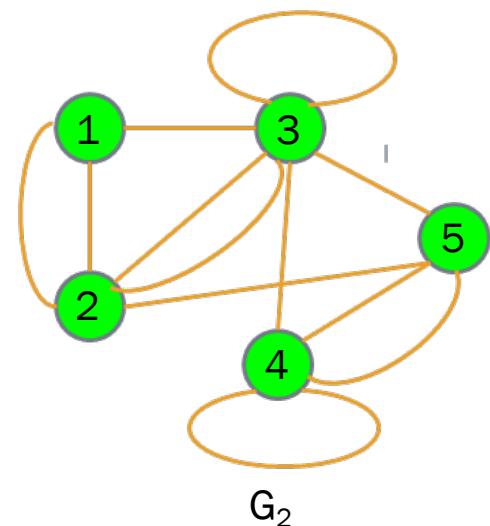
- ❖ Degree Distribution
- ❖ Path and Diameter
- ❖ Edge density
- ❖ Global clustering coefficient
- ❖ Reciprocity and Assortativity

# Degree of a Node

---

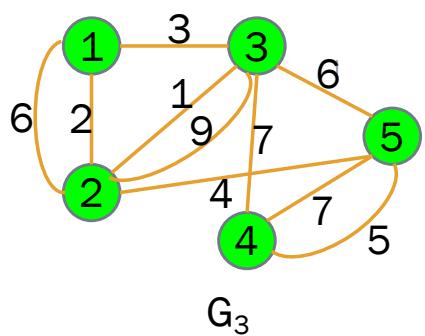


- ❑ For an undirected, unweighted network, **degree** of a node  $v$  is defined as the number of nodes in the network to which there is an edge from the node  $v$ .
- ❑ In other words, for an undirected, unweighted network, degree of a node  $v$  is the number of edges of the network that are incident on the node  $v$ .
- ❑ Putting differently, for an undirected, unweighted network, degree of a node  $v$  is the number of neighbours of the node  $v$ .
- ❑ In graph  $G_1$ , degrees of the nodes 1 through 5 are 2, 3, 4, 2, 3.
- ❑ In graph  $G_2$ , degrees of the nodes 1 through 5 are 3, 5, 7, 5, 4.
- ❑ Note: A self-loop is counted twice in evaluating degree of a node.



# Weighted Degree of a Node

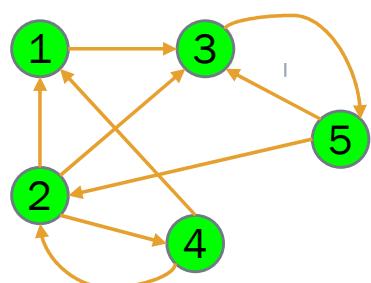
---



- ❑ For an undirected, weighted network, the weighted degree of a node is defined as the sum of weights of the edges incidents on that node
- ❑ For the weighted undirected graph  $G_3$ , the weighted degrees of the nodes are as follows:
  - Weighted degree of node 1 is 11
  - Weighted degree of node 2 is 22
  - Weighted degree of node 3 is 26
  - Weighted degree of node 4 is 16
  - Weighted degree of node 5 is 22

# Indegree and Outdegree of a Node

---



$G_4$

- ❑ In a directed network, the indegree of a node is defined as the number of incoming edges to the node
- ❑ In a directed network, the outdegree of a node is defined as the number of outgoing edges from the node
- ❑ For the directed graph  $G_4$ , the indegrees and outdegrees of the nodes are as follows:
  - Indegrees of the nodes 1 through 5 are 2, 2, 3, 1, 1
  - Outdegrees of the nodes 1 through 5 are 1, 3, 1, 2, 2

# Sum of the Degrees...

---

- ❑ For an unweighted, undirected network, the sum of the degrees of the nodes in a graph is twice the number of edges in graph
- ❑ Proof the
  - ✓ When we add an edge  $e$  to graph, it joins a pair of vertices  $v_i$  and  $v_j$  of the graph.
  - ✓ Prior to the addition of the edge  $e$  to graph, let the degrees of the nodes  $v_i$  and  $v_j$  be  $d_i$  and  $d_j$ .
  - ✓ After addition of the edge  $e$  to graph, the revised degrees of the nodes  $v_i$  and  $v_j$  be  $d_i + 1$  and  $d_j + 1$ .  
The degrees of the other nodes remain unaffected.
  - ✓ Then, on addition of an edge  $e$ , the sum of degrees of the nodes in  $G$  is incremented by 2 from its previous value. The fact is true for the addition of any edge to the graph.
  - ✓ If we add  $|E|$  number of edges to the graph one-by-one, the sum of the degrees is enhanced by  $2 \times |E|$ .
  - ✓ If a graph has no edges, all the nodes have degree zero, and so, the sum of the degrees is zero.
  - ✓ Thus, a graph with  $|E|$  edges has its sum of the degrees of the nodes as  $2 \times |E|$ .

# Sum of the Weighted Degrees...

---

- ❑ Sum of the weighted degrees of the nodes in an undirected weighted graph is twice the sum of weights of the edges in the graph
- ❖ Proof: Proved following the same line of approach

# Sum of Indegrees and Outdegrees

---

- ❑ In a directed network, the sum of indegrees is same as the sum of outdegrees.
- ❑ Proof. Proved following the same line of approach

# Number of Odd-degree Nodes...

---

- ❑ Number of odd-degree nodes in an undirected network is always even.
- ❑ Proof.
  - ✓ If possible, let the number of odd degree nodes of the graph  $G(V, E)$  be an odd integer.
  - ✓ Then, the sum of the degrees of these odd-degree nodes is an odd integer, say  $N_{odd}$ .
  - ✓ All the remaining nodes of the graph have even degrees.
  - ✓ Clearly, the sum of the degrees of these even-degree nodes is an even integer, say  $N_{even}$ .
  - ✓ Then, the sum of the degrees of all the nodes of the graph is  $N_{odd} + N_{even}$ , which is odd integer.
  - ✓ However, the sum of the degrees of all the nodes is  $2 \times |E|$
  - ✓ So,  $N_{odd} + N_{even} = 2 \times |E|$ , which is a contradiction, as the LHS is odd and RHS is even!
  - ✓ Hence the result.

# Degree Distribution

---

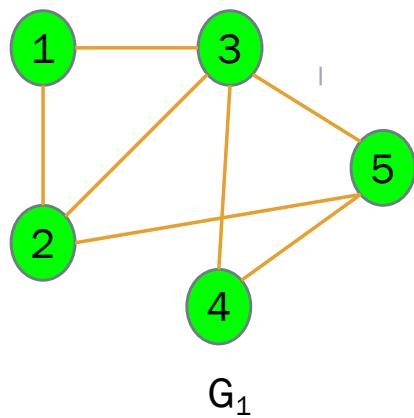
- ❑ **Degree distribution** of a network is the (probability) distribution of the degrees of nodes over the whole network.
- ❑ Let a network has  $N = |V|$  nodes.
- ❑ Let  $P_k$  denotes the probability that a randomly node has degree  $k$ .
- ❑ Then,  $P_k = \frac{N_k}{N}$ , where  $N_k$  refers to the number of nodes of degree  $k$  in the network.
- ❑ The distribution  $(k, P_k)$  represents the degree distribution of the concerned graph,
- ❑ chosen
- ❑ The mean degree, denoted  $\langle k \rangle$ , is given by  $\langle k \rangle = \sum_k k \cdot P_k$ .

# Cumulative Degree Distribution

---

- ❑ Cumulative degree distribution (**CDD**) is given by the fraction of nodes with degree smaller than  $k$ .
- ❑ In other words, it is the distribution( $k, C_k$ ), where  $C_k = \frac{\sum_{k' < k} N_{k'}}{N}$
  
- ❑ Complementary cumulative degree distribution (**CCDD**) is given by the fraction of nodes with degree greater than or equal to  $k$ .
- ❑ In other words, it is the distribution( $k, CC_k$ ), where  $CC_k = 1 - C_k$

# Degree Distribution: Example



❑ For the graph  $G_1$ , we have the following:

$N = 5$ , and  $N_1 = 0, N_2 = 2, N_3 = 2, N_4 = 1$ .

❑ The above implies,

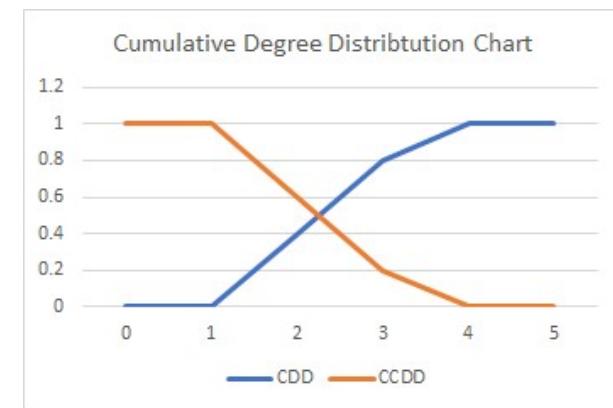
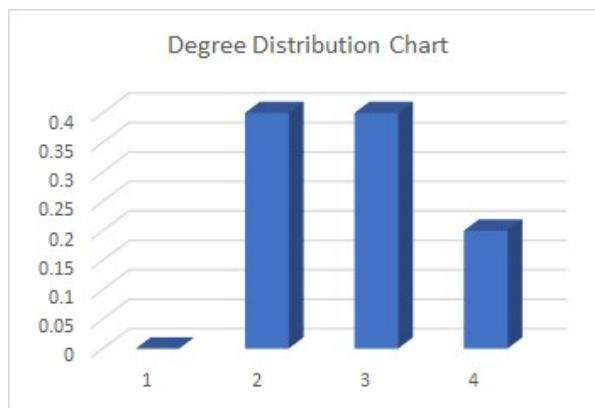
$P_1 = 0, P_2 = 0.4, P_3 = 0.4, P_4 = 0.2$ ,

$C_1 = 0, C_2 = 0.4, C_3 = 0.8, C_4 = 1.0$ ,

and

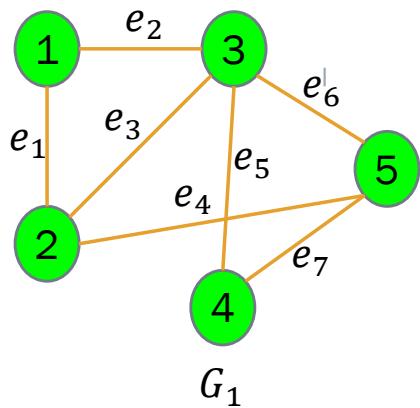
$CC_1 = 1.0, CC_2 = 0.6, CC_3 = 0.2, CC_4 = 0.0$ .

❑ Then the degree distribution, Cumulative degree distribution and Complementary cumulative degree distribution are as follows:



# Some Graph Preliminaries...

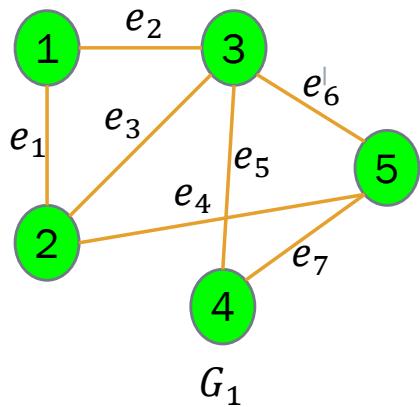
---



- ❑ In an undirected network,
  - ❑ Two nodes are called **adjacent** if they are linked by an edge.
  - ❑ Two edges are called **incident** if they share a common end-node.
- ❑ In graph  $G_1$ , the nodes **1 and 2** are adjacent, **1 and 3** are adjacent, and so on.
- ❑ In graph  $G_1$ , the edges  **$e_1$  and  $e_2$**  are incident,  **$e_1$  and  $e_3$**  are incident, and so on
- ❑ A **walk** in a network is an alternating sequence of nodes and edges, where every consecutive node pair is adjacent, and every consecutive edge pair is incident.
- ❑ A walk may pass through a node or an edge more than once. **Length** of a walk is the number of edges in the sequence.
- ❑ In graph  $G_1$ , the sequence **{3,  $e_3$ , 2,  $e_4$ , 5,  $e_6$ , 3,  $e_5$ , 4,  $e_7$ , 5,  $e_4$ , 2}** is a **walk** of length **6**.
- ❑ For a simple graph, the edges from the above sequence may be omitted.

# Some Graph Preliminaries...

---



❑ A walk in a network is called

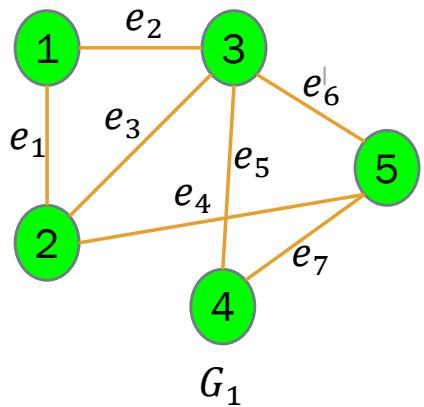
- a **closed walk** if the last node in the sequence is same as the first node; else it is called an **open walk**.
- a **trail** if the sequence has no repeated edge.
- a **path** if the sequence has neither a repeated edge nor a repeated node. In other words, a path is an open trail having no repeated nodes.
- a **cycle** if the sequence has all the edges distinct, and all the nodes, except the first and the last nodes, are also distinct. In other words, a cycle is a closed path with the only repetition of the first and the last nodes in the sequence.

❑ In graph  $G_1$ ,

- the sequence  $\{2, 5, 4, 3, 2, 1, 3, 4, 5, 2\}$  is a **closed walk**.
- the sequence  $\{5, 4, 3, 2, 1, 3\}$  is a **trail**.
- the sequence  $\{5, 4, 3, 2, 1\}$  is a **path**.
- the sequence  $\{5, 4, 3, 2, 5\}$  is a **cycle**.

# Some Graph Preliminaries...

---



- ❑ The **distance** between nodes  $v_i$  and  $v_j$  in a graph is defined as the length of the shortest path between the nodes  $v_i$  and  $v_j$ .
- ❑ In graph  $G_1$ , the distance between 1 and 4 is 2, the same between 1 and 5 is also 2.
- ❑ The **diameter** of a network is defined as the maximum distance between any pair of nodes in the network.
- ❑ The diameter of the graph  $G_1$  is 2.
- ❑ For a graph  $G$  with  $n$  nodes, the **average path length**  $l_G$  is defined as the average number of steps along the shortest paths for all possible pairs of nodes in the network.

$$l_G = \frac{\sum_{i \neq j} d_{ij}}{n(n-1)}, \text{ where } d_{ij} \text{ is distance between nodes } v_i \text{ and } v_j$$

# Some Graph Preliminaries...

---

- ❑ The **density** of a graph  $G(V, E)$ , denoted  $\rho(G)$ , is defined as the ratio of the number of edges in the graph to the total number of possible edges in the network. Mathematically,

$$\rho(G) = \frac{2 \times |E|}{|V| \times (|V| - 1)}$$

- ❑ For the graph  $G_1$ , the **average path length** is:

$$\frac{2 \times (1 + 1 + 2 + 2 + 1 + 2 + 1 + 1 + 1 + 1)}{5 \times 4} = \frac{26}{20} = 1.3$$

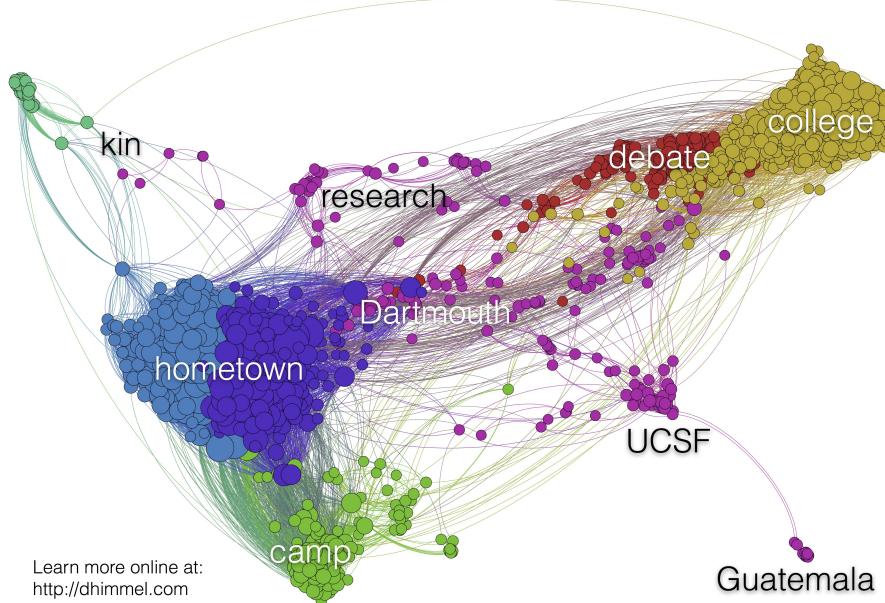
- ❑ For the graph  $G_1$ , the **network density** is:

$$\frac{2 \times 7}{5 \times 4} = 0.7$$

# Clusters in Social Networks

---

The Friendship Network of Daniel Himmelstein



A Facebook Friendship Network Example

<https://blog.dhimmel.com/friendship-network/>

- In social networks, we often find
  - tightly-knit groups here and there
  - less dense ties away from these groups
- Indicative of friendship structures in social media
- Measure used to capture these phenomena
  - Local clustering coefficient
  - Global clustering coefficient

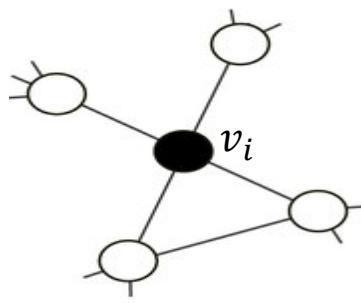
# Local Clustering Coefficient

---

The local clustering coefficient  $C_i$  for a vertex  $v_i$  in a network  $G(V, E)$  is given by the proportion of edges between the vertices within its neighborhood divided by the number of links that could possibly exist between them.

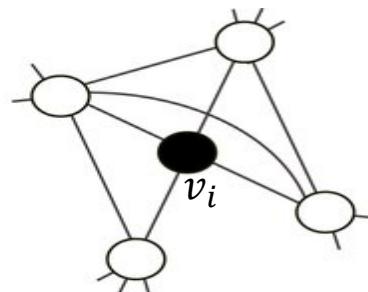
$$C_i = \frac{2 \times |\{e_{jk} \mid v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i \cdot (k_i - 1)}$$

Where  $N_i$  is the neighbourhood of the vertex  $v_i$ , and  $k_i = |N_i|$ .



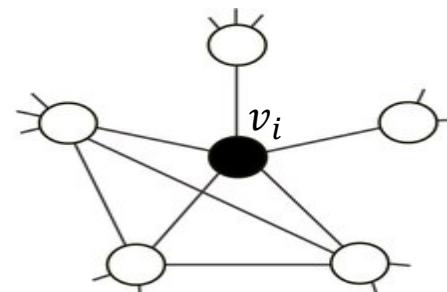
$$k_i = 4$$

$$C_i = \frac{1}{6}$$



$$k_i = 4$$

$$C_i = \frac{2}{3}$$



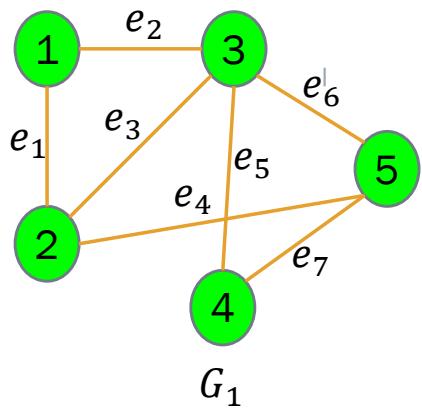
$$k_i = 5$$

$$C_i = \frac{3}{10}$$

[https://www.researchgate.net/publication/236604411\\_Suicide\\_Ideation\\_of\\_Individuals\\_in\\_Online\\_Social\\_Networks/figures?lo=1](https://www.researchgate.net/publication/236604411_Suicide_Ideation_of_Individuals_in_Online_Social_Networks/figures?lo=1)

# Local Clustering Coefficient

---

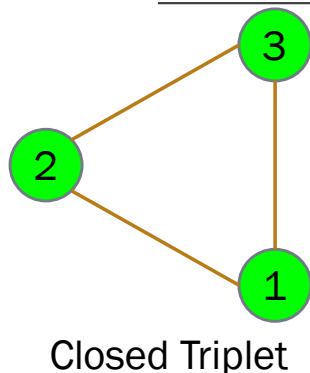


- In a network  $G(V, E)$ , the **local clustering coefficient** of node  $v_i \in V$ , denoted  $C_i$ , is defined as

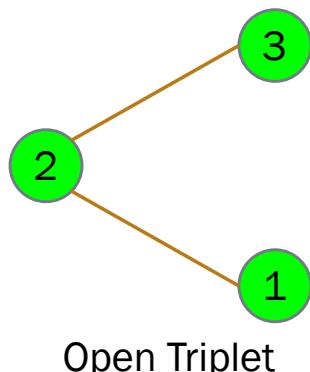
$$C_i = \frac{\text{Number of edges between neighbors of } v_i}{\text{Number of maximum possible edges between neighbors of } v_i}$$

- In graph  $G_1$ ,
  - the local clustering coefficient of node 2 is  $\frac{2}{3}$
  - the local clustering coefficient of node 3 is  $\frac{3}{6}$  i.e.  $\frac{1}{2}$
  - and so on...

# Global Clustering Coefficient



Closed Triplet

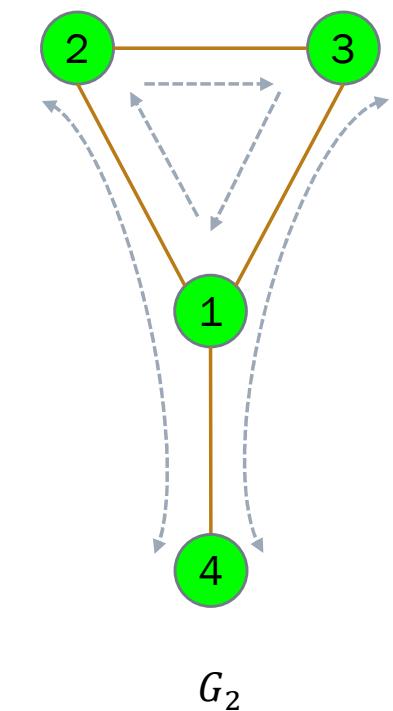


Open Triplet

- The global clustering coefficient  $C$  of a network  $G$  is defined as

$$C = \frac{\text{Total number of closed triplets in } G}{\text{Total number of triplets (open & closed) in } G}$$

- In the graph  $G_2$ , there are three closed triplets viz., [1,2,3], [2,3,1], and [3,1,2].
- In the graph  $G_2$ , there are five open and closed triplets, viz., (1,2,3), (2,3,1), (3,1,2), (2,1,4), and (3,1,4).
- Thus, the global clustering coefficient of the graph  $G_2$  is  $3/5$ .



closed mate consider karo ke ketla loko thi ae triangle pass thai 6e

# Global Clustering Coefficient

---

- The global clustering coefficient may also be written as

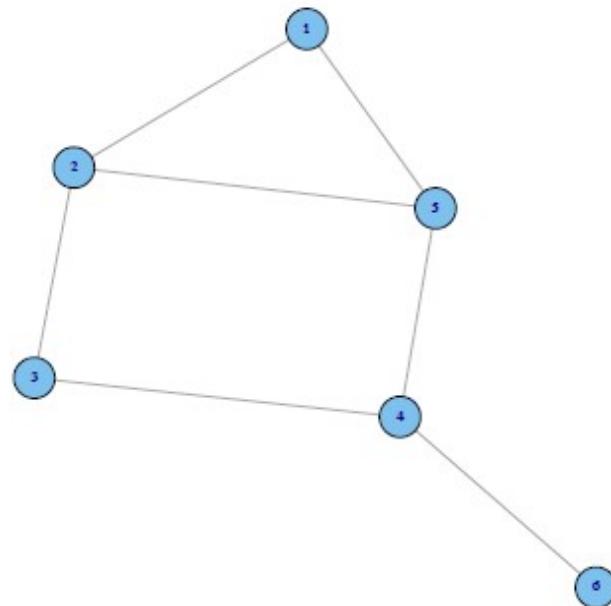
$$C = \frac{3 \times \text{Total number of triangles in } G}{\text{Total number of triplets (open \& closed) in } G}$$

- In other words, if  $A = (A_{ij})$  is the adjacency matrix of the graph  $G$ , then

$$C = \frac{\sum_{i,j,k} (A_{ij}A_{jk}A_{ki})}{\sum_i k_i(k_i - 1)} \text{ where } k_i = \sum_j A_{ij}$$

# Global Clustering Coefficient

---



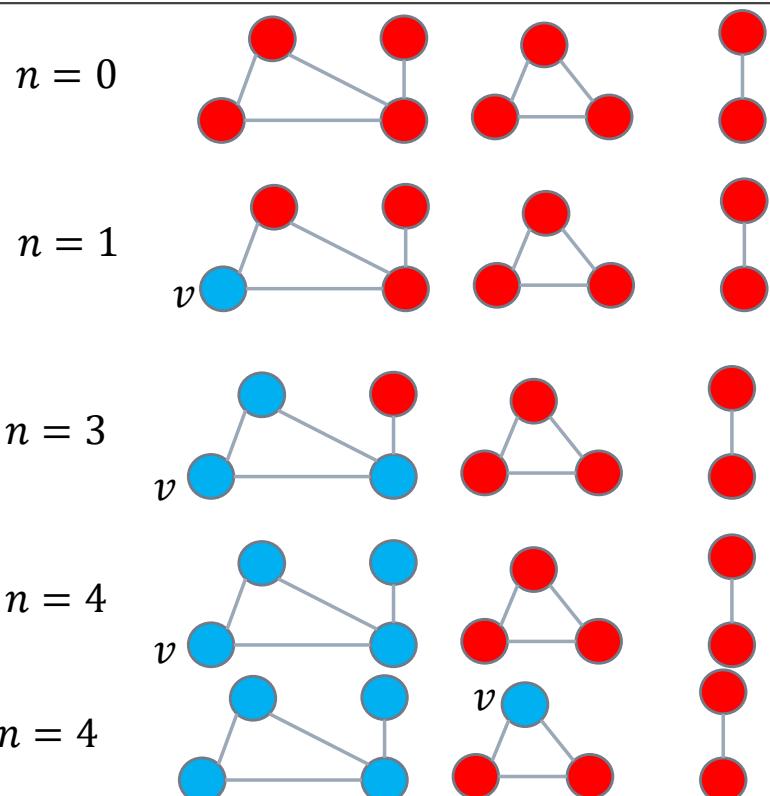
- 1) Degree distribution
- 2) Average Degree
- 3) Average path length
- 4) Density
- 5) Local Clustering coefficient
- 6) Global Clustering Coefficient

# Connected Components

---

- ❑ In a typical social network, there are loose links that **connects** the tightly-knit clusters
- ❑ In an undirected network  $G$ , two nodes  $v_i$  and  $v_j$  are said to be **connected** if there exists a path between  $v_i$  and  $v_j$ .
- ❑ An **entire network** is said to be **connected** if any pair of nodes in the network is connected.
- ❑ Connected subnetworks of a network, if exist, are called **components** of the network.
- ❑ In real-world networks, there often exist one **giant component** (consuming major chunk of nodes) and many smaller components.
- ❑ In a network, connectedness shows resilience to link breakdowns.

# Finding Connected Components



The network  $G$  with all nodes coloured red

Choose a random node  $v$ , colour it blue, and set  $n$  to 1

Apply BFS from node  $v$ , and colour with blue all the nodes reached thereof, and increment  $n$  each time

No more node can be reached from  $v$  using BFS. We get a component in blue.

Since  $n \neq 9$ , we choose a red node as  $v$ , repeat the steps above to find other components

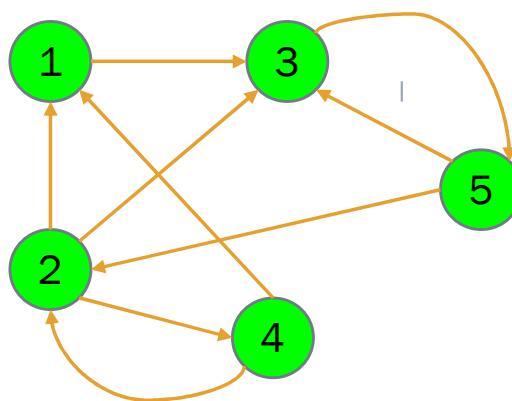
# Connectedness in Directed Networks

---

- ❑ A directed network  $G$  is **strongly connected** if there exists a **(directed) path** between every pair of nodes in  $G$ .
- ❑ If we replace all the directed edges of a directed network  $G$  with undirected edges, then the resultant network is called an **undirected version** of the directed network  $G$ .
- ❑ A directed network  $G$  is said to be **weakly connected** if its undirected version is connected.

Can you say the below graph  $G_4$  is  
strongly connected or weakly connected?

---



$G_4$

# Centrality in a Network

---

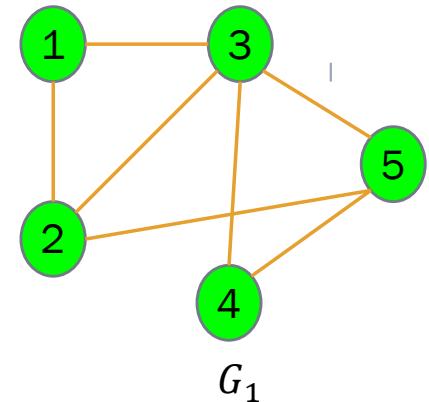
- ❑ Influential players often play central roles in a network
- ❑ Defining/Identifying influential players always remain hard
  - Some players attract limelight
  - Some others play behind the scene
  - Many others do important linkage
  - and so on...
- ❑ To identify influential players, we require
  - to define a notion of influence
  - to device measure that can capture that influence

mari ketli degree 6e/ sauthi highest degree su 6e

# Degree Centrality

---

- ❑ Centrality of the simplest kind
- ❑ In a sense, captures the **popularity** of a player within a network
- ❑ Quantifies the direct influence of a node on its local neighbourhood
- ❑ The **degree centrality**  $C_d(v)$  of a node  $v$  in a network  $G(V, E)$  is defined as:  
$$C_d(v) = \frac{\deg(v)}{\max_{u \in V} \deg(u)}$$
- ❑ Particularly useful for marketing scenarios, wherein the detected influential user can promote a product/service across her followers
- ❑ Degree centrality of the nodes 1 through 5 in network  $G_1$  are  $2/4$ ,  $3/4$ ,  $4/4$ ,  $2/4$ , and  $3/4$ , respectively; i.e., 0.5, 0.75, 1.0, 0.5, and 0.75, respectively. So, node 3 is most central according to degree centrality measure.



mare ketla loko ne information pohchadvani 6e(V-1)

/

mara thi badha nodes atla distance par connected 6e

# Closeness Centrality

---

- A means for detecting nodes that can spread information very efficiently through a graph
- The measure is useful in
  - Examining/restricting the spread of fake news/misinformation in social media
  - Examining/restricting the spread of a disease in epidemic modelling
  - Controlling/restricting the flow of vital information and resources within an organization (a terrorist network, for example)
- The **closeness centrality**  $C(v)$  of a node  $v$  in a network  $G(V, E)$  is defined as

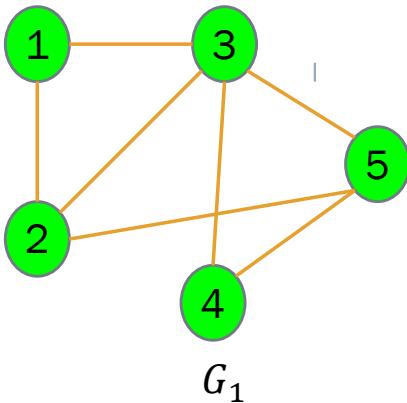
$$C(v) = \frac{|V| - 1}{\sum_{u \in V \setminus \{v\}} d(u, v)}$$

Where  $d(u, v)$  denotes the distance of node  $u$  from node  $v$

- The measure indicates how close a node from the rest of the network

# Closeness Centrality

---



□ In graph  $G_1$ , the closeness centrality for the nodes are as follows

$$C(1) = \frac{5 - 1}{1 + 1 + 2 + 2} = \frac{4}{6} = 0.67$$

$$C(2) = \frac{5 - 1}{1 + 1 + 2 + 1} = \frac{4}{5} = 0.80$$

$$C(3) = \frac{5 - 1}{1 + 1 + 1 + 1} = \frac{4}{4} = 1.0$$

$$C(4) = \frac{5 - 1}{2 + 2 + 1 + 1} = \frac{4}{6} = 0.67$$

$$C(5) = \frac{5 - 1}{2 + 1 + 1 + 1} = \frac{4}{5} = 0.80$$

□ Clearly, node 3 is most central according to closeness centrality measure

# Betweenness Centrality

---

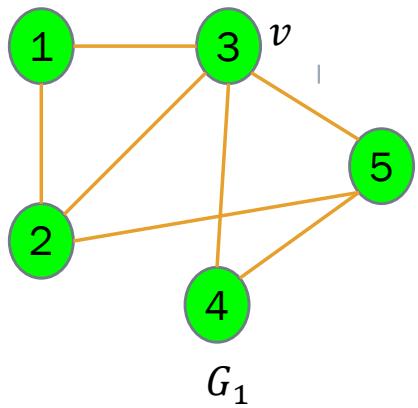
- ❑ A measure to compute how central a node is in *between* paths of the network
- ❑ A measure to compute how many (shortest) paths of the network pass through the node
- ❑ Useful in identifying
  - ❑ the *articulation points*, i.e., the points in a network which, if removed, may disconnect the network
  - ❑ The *super spreaders* in analyzing disease spreading in epidemiology
  - ❑ the *suspected spies* in security networks
- ❑ The *betweenness centrality*  $C_B(v)$  of a node  $v$  in a network  $G(x, y)$  is defined as

$$C_B(v) = \sum_{x,y \in V \setminus \{v\}} \frac{\sigma_{xy}(v)}{\sigma_{xy}}$$

where  $\sigma_{xy}$  denotes the number of shortest paths between nodes  $x$  and  $y$  in the network,  $\sigma_{xy}(v)$  denotes the same passing though  $v$ . If  $x = y$ , then  $\sigma_{xy} = 1$ .

# Betweenness Centrality

---



□ To find the betweenness centrality of node  $v = 3$  in graph  $G_1$

□ The following matrix is of the form  $\sigma_{xy}(v) | \sigma_{xy}$

$\sigma_{xy}(v)   \sigma_{xy}$	1	2	3	4	5
1	0 1	0 1	--	1 1	1 2
2	0 1	0 1	--	1 2	0 1
3	--	--	--	--	--
4	1 1	1 2	--	0 1	0 1
5	1 2	0 1	--	0 1	0 1

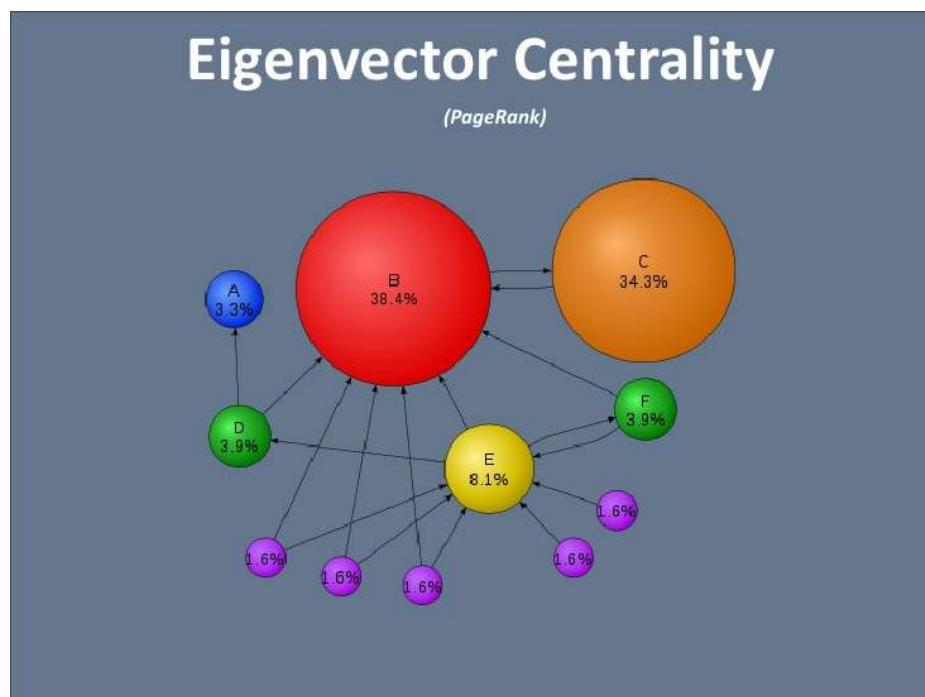
□ Thus the betweenness centrality of node 3 =  $\frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{1} + \frac{1}{2} + \frac{1}{2} = 4$

# Betweenness Centrality: Variants

---

- ❑ The **edge betweenness centrality** refers to the fraction of all pairs of shortest paths of the network that pass through a given edge.
- ❑ Computation is more-or-less similar to that of betweenness centrality
- ❑ The **flow betweenness centrality** the fraction of all paths (not necessarily the shortest paths) of the network that passes through a given edge.
- ❑ Clearly, flow betweenness centrality measure is computationally expensive than betweenness or edge betweenness centrality measures.

# Eigenvector Centrality



- Measures a node's importance by taking into consideration the preference of its neighbors
- Uses a recursive approach
- A node has a higher eigenvector centrality, if it is directly connected to other nodes having high eigenvector centrality
- Generally applied on directed networks

[https://www.slideshare.net/mdeiters/you-might-also-like-implementing-user-recommendations-in-rails/63-Eigenvector\\_Centrality\\_PageRankThursday\\_June\\_10](https://www.slideshare.net/mdeiters/you-might-also-like-implementing-user-recommendations-in-rails/63-Eigenvector_Centrality_PageRankThursday_June_10)

# Eigenvector Centrality

---

- The eigen vector centrality  $x_v$  of a node  $v$  in a network  $G(V, E)$  is given by

$$x_v = \frac{1}{\lambda_1} \sum_{t \in N(v)} x_t = \frac{1}{\lambda_1} \sum_{t \in V} (a_{vt} \times x_t)$$

where  $\lambda_1$  is the largest eigen value of the matrix  $A = (a_{ij})$ , the adjacency matrix of the network  $G$

- The largest eigen value  $\lambda_1$  is obtained by solving the equation

$$A \cdot X = \lambda_1 \cdot X$$

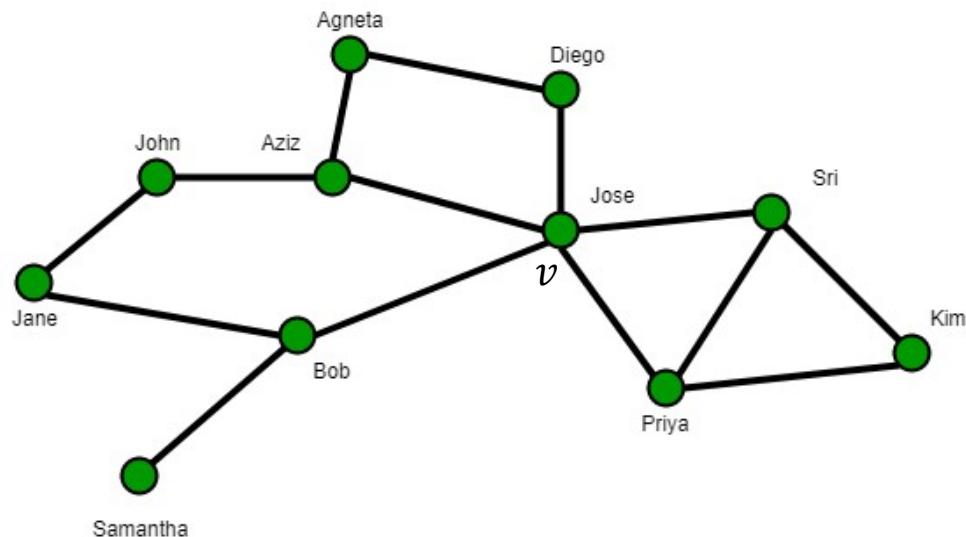
- $X$  above is a column vector, whose  $v^{th}$  entry is  $x_v$ , the eigen vector centrality of the node  $v$

# Katz Centrality

---

- ❑ An extension of eigenvector centrality
- ❑ Can be used to compute centrality in directed networks such as citation networks and the World Wide Web
- ❑ Mostly suitable in the analysis of directed acyclic graphs
- ❑ Computes the relative influence of a node in a network by considering all immediate neighbors and all further nodes connected to the node
- ❑ Connections with distant neighbors are, however, penalized by an attenuation factor

# Katz Centrality: Attenuation Factor



<https://www.geeksforgeeks.org/katz-centrality-centrality-measure/>

- Let us consider the influence of *Jose* in the network, and also let the attenuation factor be  $\alpha$ ,  $0 < \alpha < 1$
- Immediate neighbours of Jose are *Diego, Aziz, Bob, Priya, and Sri*. Influence of these neighbours on Jose would be attenuated at a factor of  $\alpha$
- Second order neighbours of Jose are *Agneta, John, Samantha, and Kim*. Influence of these neighbours on Jose would be attenuated at a factor of  $\alpha^2$
- The (only) third order neighbour of Jose is *Jane*. Influence of these neighbours on Jose would be attenuated at a factor of  $\alpha^3$

# Katz Centrality

---

- The Katz centrality of a node  $v_i$  in a network  $G(V, E)$ , denoted  $C_{Katz}(i)$ , is defined as

$$C_{Katz}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{|V|} \alpha^k \times A_{ji}^k$$

where  $A$  is the adjacency matrix of  $G$

- Matrix  $A^k$  indicates the presence/absence of a path of length  $k$  between a node-pair
- The entry  $A_{ji}^k$  in  $A^k$  matrix indicates the total number of  $k$ -hop walks between node  $j$  and node  $i$

# PageRank

---

- ❑ Devised by Larry Page and Sergey Brin in 1998
- ❑ Devised as a part of a ~~research project~~ about a new kind of search engine
- ❑ Based upon the concepts of eigenvector centrality and Katz centrality measures
- ❑ Used to rate the importance of web pages on the web
- ❑ A page's importance is determined by the importance of the web pages linked to the page
- ❑ The algorithm is inherently recursive because the page further contributes to the importance of the web pages linked to it

# PageRank

---

- The PageRank for a network node  $v_i$  in a network  $G(V, E)$ , denoted  $PG(v_i)$ , is defined as

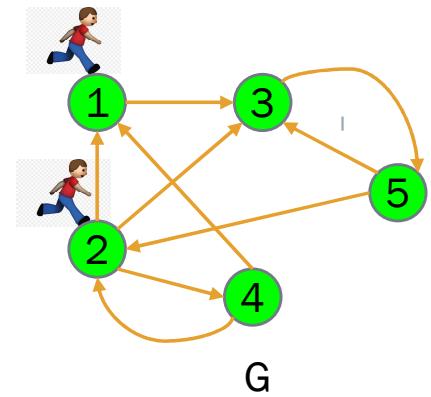
$$PG(v_i) = \frac{1-d}{|V|} + d \sum_{\substack{t=1 \\ t \neq i}}^{|V|} \frac{PG(v_t)}{outdeg(v_t)}$$

where  $d$  is constant, called the damping factor

- Though there are many works to determine the optimal value for  $d$ , it is usually set as  $d = 0.85$

# PageRank: The Random Surfer model

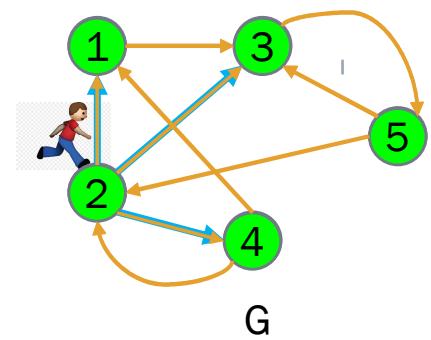
---



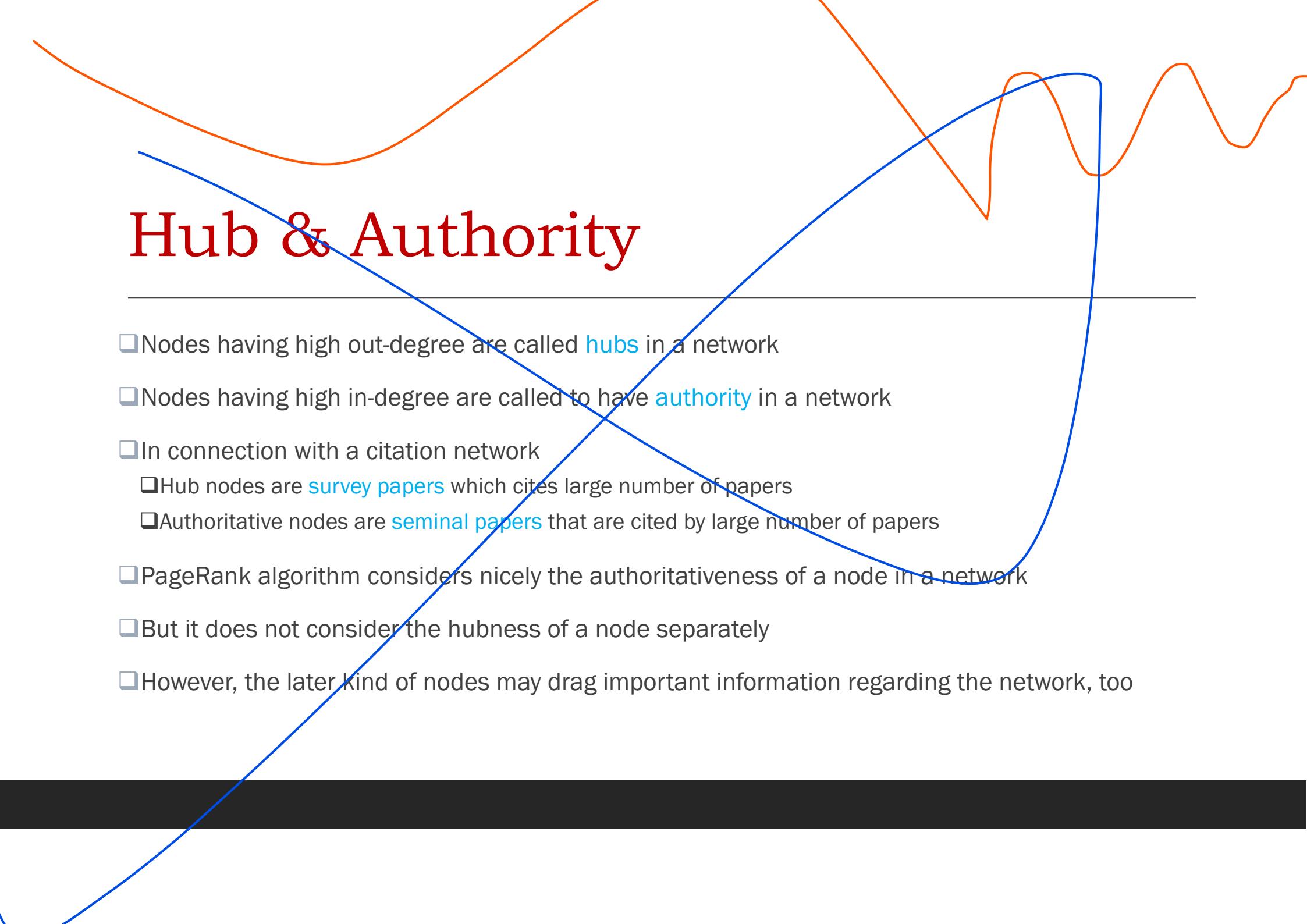
- ❑ A random surfer surfing through the Internet by
  - a. opening a webpage at random, and
  - b. moving across webpages by randomly clicking hyperlinks in the page he is in
  - c. repeating the steps (a) and (b) at random
- ❑ The surfer follows hyperlinks to surf with probability  $d$
- ❑ The surfer jumps to pages to surf with probability  $(1 - d)$
- ❑ Since there are  $|V|$  number of vertices in the network, the probability of choosing a random webpage is  $\frac{1-d}{|V|}$
- ❑ Hence, we have the **First term** of the PageRank equation

# PageRank: The Random Surfer model

---



- ❑ The surfer is in a page  $v_t$  and he decides to follow a hyperlink
- ❑ The probability that he decides to follow hyperlink than random jump is  $d$
- ❑ At node  $v_t$ , he has  $\text{outdeg}(v_t)$  number of options
- ❑ The PageRank contribution of the page  $v_t$  is  $PG(v_t)$
- ❑ The above contribution is divided across the available hyperlinks (outward links)
- ❑ However, the surfer could be anywhere in network
- ❑ Hence the total possible contribution with this choice  $d \sum_{\substack{t=1 \\ t \neq i}}^{|V|} \frac{PG(v_t)}{\text{outdeg}(v_t)}$
- ❑ Hence, we have the **Second term** of the PageRank equation

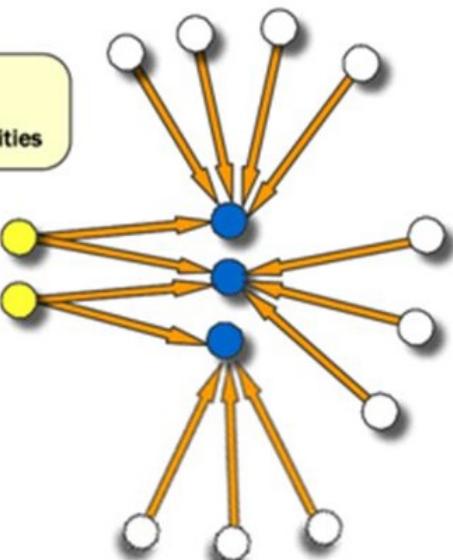


# Hub & Authority

---

- ❑ Nodes having high out-degree are called **hubs** in a network
- ❑ Nodes having high in-degree are called to have **authority** in a network
- ❑ In connection with a citation network
  - ❑ Hub nodes are **survey papers** which cites large number of papers
  - ❑ Authoritative nodes are **seminal papers** that are cited by large number of papers
- ❑ PageRank algorithm considers nicely the **authoritativeness** of a node in a network
- ❑ But it does not consider the **hubness** of a node separately
- ❑ However, the later kind of nodes may drag important information regarding the network, too

# Hub & Authority



- For node  $v$ , its **hubness** is determined by the cumulative **authoritativeness** of nodes that  $v$  points to.

$$hub(v) = \sum_{u \in out(v)} auth(u)$$

where  $out(v)$  denotes the set of nodes pointed by  $v$

- On the other hand, its **authoritativeness** is computed by the cumulative **hubness** of the nodes pointing to  $v$ ,

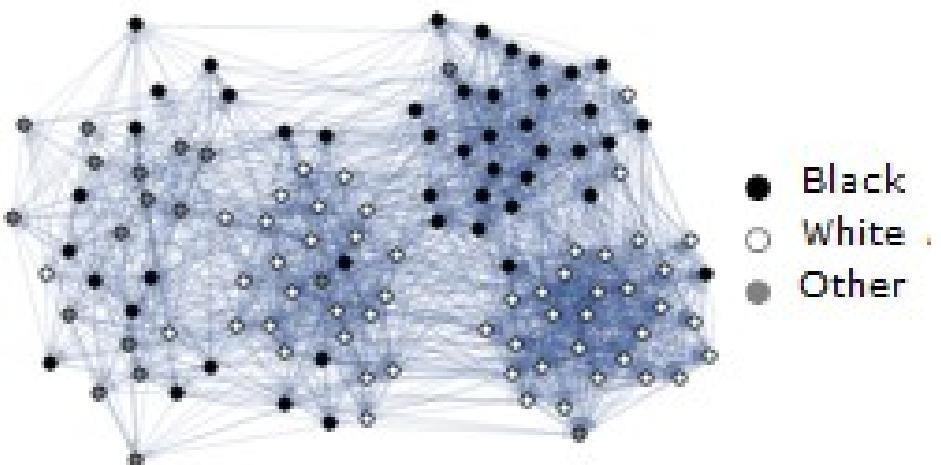
$$auth(v) = \sum_{u \in in(v)} hub(u)$$

where  $in(v)$  denotes the set of nodes pointing to  $v$

- Kleinberg proposed **Hyperlink-Induced Topic Search (HITS)** algorithm exploiting these concepts

# Assortative Mixing

---



<https://www.wolfram.com/mathematica/new-in-9/social-network-analysis/homophily-and-assortativity-mixing.html>

- ❑ In friendship kind of social networks,
  - ❑ individuals often choose to associate with others having similar characteristics
  - ❑ age, nationality, location, race, income, educational level, religion, or language are common characteristics
  - ❑ **Homophily**
- ❑ In intimate relationship kind of network,
  - ❑ mixing is also disassortative by gender
  - ❑ most people prefer to have affair with opposite sex
  - ❑ **Heterophily**
- ❑ **Assortativity** or **assortative mixing** is a measure to gauge these mixing tendencies

# Assortative Mixing

---

- ❑ The phenomenon of particular interest is the **assortative mixing by degree**
  - ❑ High degree nodes often prefers to connect other high degree nodes
  - ❑ Low degree nodes seen to connect other low degree nodes
- ❑ Assortative mixing can have impact, for example, on the spread of diseases
  - ❑ Many diseases are known to have differing prevalence in different population groups
- ❑ Such behaviors are observed in non-social types of networks, too
  - ❑ biochemical networks in the cell
  - ❑ computer and information networks

# Assortative Mixing

---

- ❑ A common practice to find similarity between nodes is to use a correlation coefficient
- ❑ The **Pearson correlation coefficient** is a good choice if we want degree-based assortativity
- ❑ For two data (degree) distribution  $x$  and  $y$ , the **Pearson correlation coefficient**  $r_{xy}$  is given by

$$r_{xy} = \frac{N \sum xy - \sum x \sum y}{\sqrt{(N \sum x^2 - (\sum x)^2)(N \sum y^2 - (\sum y)^2)}}$$

- ❑ If  $r_{xy} = 1$  , then nodes  $x$  and  $y$  are perfectly assortative (homophily)
- ❑ If  $r_{xy} = -1$  , then nodes  $x$  and  $y$  are perfectly disassortative (heterophily)
- ❑ If  $r_{xy} = 0$  , then nodes  $x$  and  $y$  are non-assortative

# Transitivity

- ❑ A metric to determine the linkage between a pair of nodes
- ❑ Very important in social networks, and to a lesser degree in other networks
- ❑ In abstract mathematics, if entity  $x$  is related to entity  $y$ , and also entity  $y$  is related to entity  $z$ , then the transitivity of the relation ensures that entity  $x$  is related to entity  $z$ .
- ❑ In social networks, a complete transitivity may yield: “**Friends of my friends are my friends**”
  - ❑ Utterly Absurd in real networks!
- ❑ In fact, a complete transitivity would imply that each component of a network is a **clique!!**
- ❑ However, partial transitivity is useful: “**Friends of my friend are more likely my friend than some randomly chosen member from the population**”

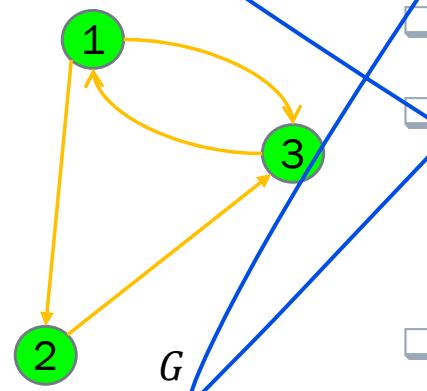
# Transitivity

- ❑ A complete graph is surely transitive
- ❑ A measure of transitivity intends to capture how close a network is to a complete graph
- ❑ A network with higher transitivity are likely to form dense clusters
- ❑ Two ways to capture this tendency
  - ❑ Local clustering coefficient
  - ❑ Global clustering coefficient

# Reciprocity

- Relevant for directed networks
- A measure of the likelihood of vertices in a directed network to be mutually linked.
- Networks that transport information or material, mutual links facilitate the transportation process
- An important phenomenon for such applications
- Informally, reciprocity refers to: “**If you would follow me, most likely I shall follow you back**”
- May be considered **a simplified version of transitivity**

# Reciprocity



❑ Reciprocity counts the closed loops of length 2

❑ The reciprocity  $R$  of a network  $G$  is defined as

$$C = \frac{\text{Total number of reciprocal pairs in } G}{\text{Total number of pairs (reciprocal \& nonreciprocal) in } G}$$

❑ For graph  $G$ , the reciprocity is  $\frac{1}{3}$

# Reciprocity

□ The reciprocity  $R$  for a graph  $G(V, E)$  having adjacency matrix  $A = (a_{ij})$  is given by

$$R = \frac{2}{|E|} \sum_{i < j} (a_{ij} \cdot a_{ji})$$

□ On simplification,

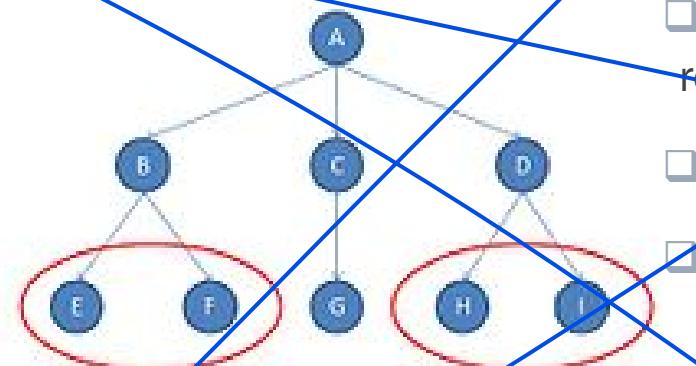
$$R = \frac{2}{|E|} \times \frac{1}{2} \text{Trace}(A^2) = \frac{\text{Trace}(A^2)}{|E|}$$

□ In the above expression,  $\text{Trace}(\cdot)$  function denotes the sum of the diagonal elements of its argument square matrix

# Similarity

- A measure to check whether a pair of nodes in a network are part of the **same equivalence class**
- An abstract ways of making sense of the patterns of relations among social actors
- Three broad classes of equivalence classes
  - Structural equivalence
  - Automorphic equivalence
  - Regular equivalence
- There is a hierarchy of these three equivalence concepts
  - Any set of structural equivalences are also automorphic and regular equivalences
  - Any set of automorphic equivalences are also regular equivalences
  - Not all regular equivalences are necessarily automorphic or structural
  - Not all automorphic equivalences are necessarily structural

# Structural Equivalence



[https://en.wikipedia.org/wiki/Similarity\\_\(network\\_science\)#:~:text=Similarity%20in%20network%20analysis%20occurs,automorphic%20equivalence%2C%20and%20regular%20equivalence.](https://en.wikipedia.org/wiki/Similarity_(network_science)#:~:text=Similarity%20in%20network%20analysis%20occurs,automorphic%20equivalence%2C%20and%20regular%20equivalence.)

- ❑ Two nodes are said to be exactly **structurally equivalent** if they have the same relationships to all other nodes
- ❑ Two actors must be **exactly substitutable** in order to be structurally equivalent
- ❑ In the attached network,
  - ❑ nodes *E* and *F* are structurally equivalent, since these two nodes have same pattern ties (viz. a single tie) with the node *B*
  - ❑ Also, nodes *H* and *I* are structurally equivalent, since these two nodes have same pattern ties (viz. a single tie) with the node *D*
- ❑ Exact structural equivalence is likely to be rare (particularly in large networks)
- ❑ the degree of structural equivalence is what interests us the most

# Measuring Structural Equivalence

## Common Neighbors

number of common neighbors shared in the neighborhoods of the nodes  $a$  and  $b$

$$\sigma_{CN}(a, b) = |N(a) \cap N(b)|$$

## Jaccard Similarity

Normalizes the common neighbors by the combined size of the neighborhoods of the two nodes

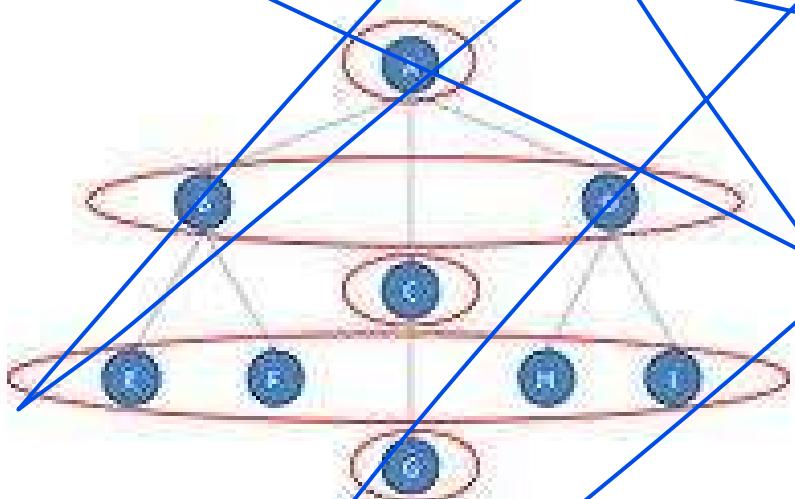
$$\sigma_{CN}(a, b) = \frac{|N(a) \cap N(b)|}{|N(a) \cup N(b)|}$$

## Cosine Similarity

normalizes the common neighbors by the individual sizes of the neighborhoods

$$\sigma_{CN}(a, b) = \frac{|N(a) \cap N(b)|}{\sqrt{|N(a)||N(b)|}}$$

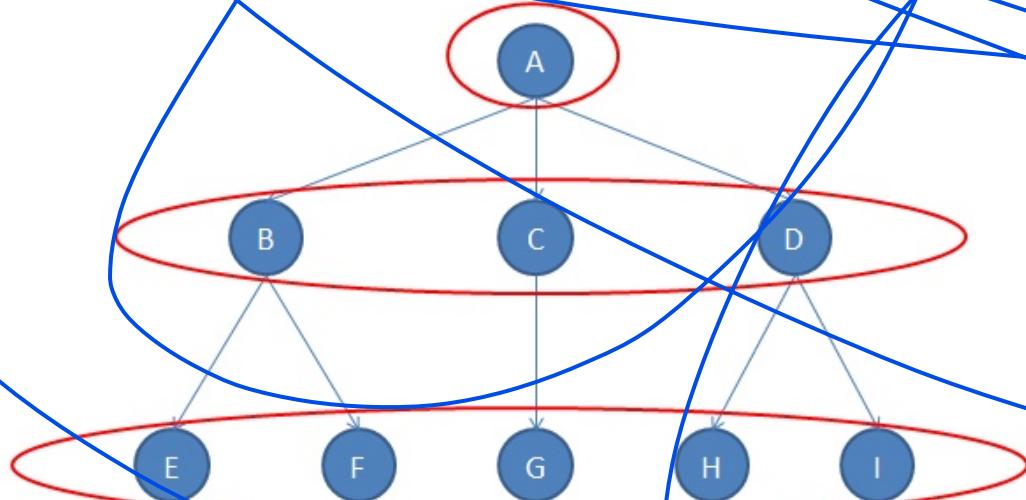
# Automorphic Equivalence



[https://en.wikipedia.org/wiki/Similarity\\_\(network\\_science\)#:~:text=Similarity%20in%20network%20analysis%20occurs,automorphic%20equivalence%20and%20regular%20equivalence.](https://en.wikipedia.org/wiki/Similarity_(network_science)#:~:text=Similarity%20in%20network%20analysis%20occurs,automorphic%20equivalence%20and%20regular%20equivalence.)

- Let us interpret the network as follows: the network describe a **franchise group of a restaurant chain**
  - A is the general manager at central headquarters
  - B, C, and D are the managers of three different stores.
  - E and F are workers at one store; G is the lone worker at a second store; H and I are workers at the third store
- B and D are **equivalent** in the following sense
  - B and D report to a boss (same boss here)
  - Each has exactly two workers
- Similarly, E, F, H, and I are also **equivalent** in the following sense
  - They report to a store manager (different boss here)
  - Nobody report to these persons
- The above approach of equivalence is **automorphic equivalence**

# Regular Equivalence



- Two actors are **regularly equivalent** if they are equally related to equivalent others
- Two mothers are **regularly equivalent**, since
  - each has a similar pattern of connections with a husband,
  - with their children,
  - with their in-laws, etc.
- The store managers are **regularly equivalent**, since
  - each has a similar pattern of connections with their employees at their stores, and
  - with the general manager at the central headquarter

[https://en.wikipedia.org/wiki/Similarity\\_\(network\\_science\)#:%~:text=Similarity%20in%20network%20analysis%20occurs,automorphic%20equivalence%20and%20regular%20equivalence.](https://en.wikipedia.org/wiki/Similarity_(network_science)#:%~:text=Similarity%20in%20network%20analysis%20occurs,automorphic%20equivalence%20and%20regular%20equivalence.)

# Measuring Regular Equivalence

- The regular equivalence between nodes  $v_i$  and  $v_j$  in network  $G(V, E)$  having adjacency matrix  $A = (A_{ij})$  is defined as

$$\sigma_{reg}(v_i, v_j) = \alpha \sum A_{ik} A_{jl} \sigma_{reg}(v_k, v_l)$$

- We may relax the equation as

$$\sigma_{reg}(v_i, v_j) = \alpha \sum_k A_{ik} \sigma_{reg}(v_k, v_j)$$

- Rewrite the above as

$$\sigma_{reg} = \alpha A \sigma_{reg}$$

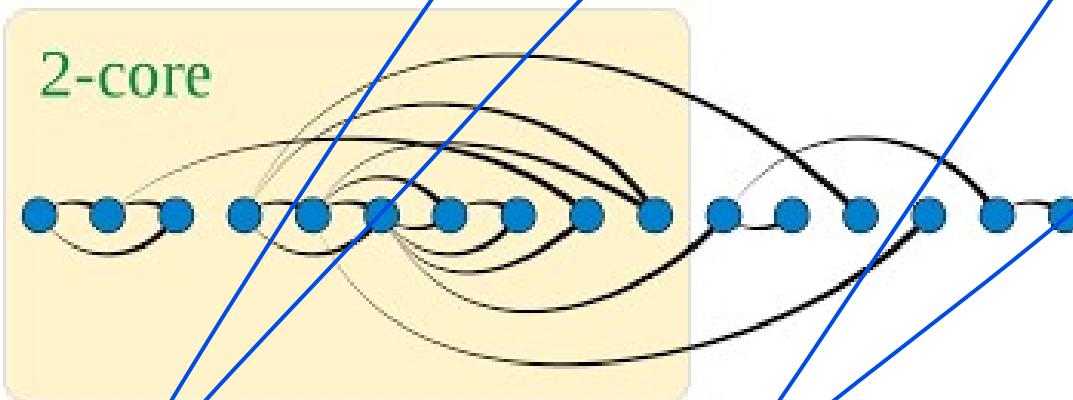
- The above imply

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

- For convergence of the above,  $\alpha < \frac{1}{\lambda_1}$ , where  $\lambda_1$  is the largest eigen value of  $A$

# Degeneracy

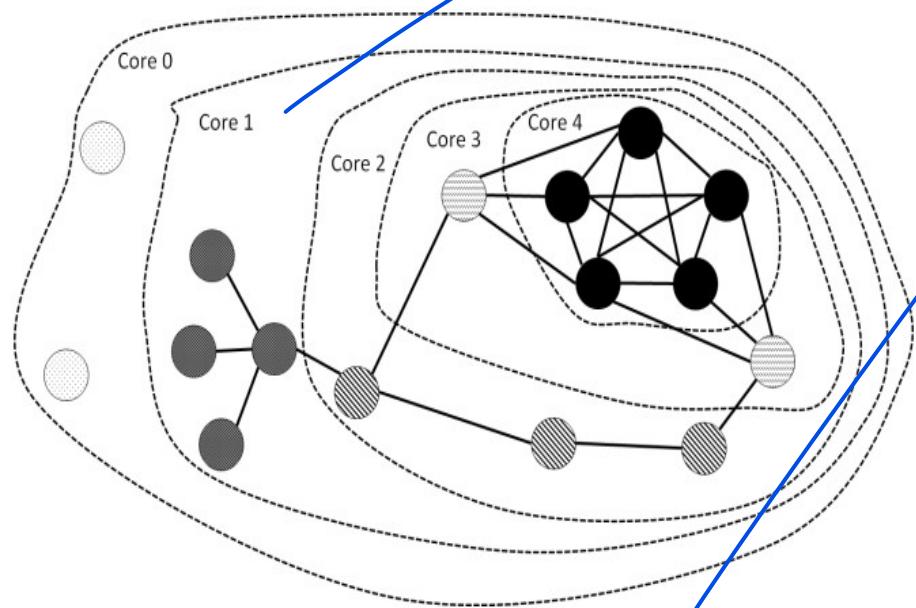
- ❑ A  $k$ -degenerate graph is an undirected graph in which every subgraph has a vertex of degree at most  $k$
- ❑ The degeneracy of a graph is the smallest value of  $k$  for which it is  $k$ -degenerate
- ❑ A  $k$ -core of a graph  $G$  is a maximal connected subgraph of  $G$  in which all vertices have degree at least  $k$



A 2-degenerate graph with one of its 2-core highlighted

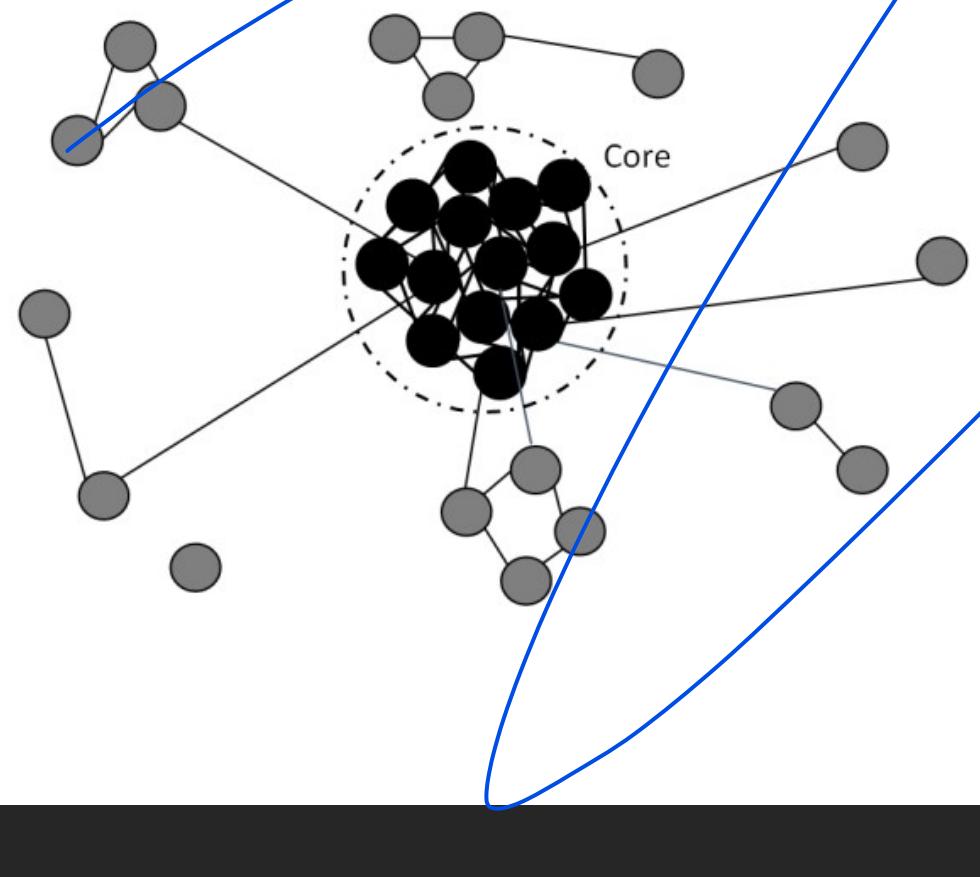
[https://en.wikipedia.org/wiki/Degeneracy\\_\(graph\\_theory\)](https://en.wikipedia.org/wiki/Degeneracy_(graph_theory))

# Degeneracy: Core Number



- ❑ The **coreness** or **core number** of a node is the order of the highest-order core that the node belongs to
- ❑ A node has a **core number  $k$**  in network  $G$  if
  - ❑ It belongs to the  $k$ -core subgraph, but
  - ❑ does not belong to the  $(k + 1)$ -core subgraph of  $G$
- ❑ In the example network, nodes inside the central-most 4-core subgraph have **core number 4**
- ❑ Similar to centrality, core number is a measure of **prestige** of a node in a network

# Degeneracy: Core-Periphery



- ❑ Real-world networks often consist of
  - ❑ a dense and connected core, and
  - ❑ surrounding the core by disconnected and scrambled periphery
- ❑ The structure above is termed as the **core-periphery structure** of the network

